Predicting, Detecting and Explaining the Occurrence of Vocal Activity in Multi-Party Conversation

Kornel Laskowski

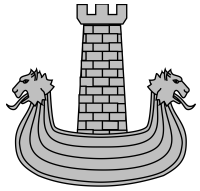
CMU-LTI-11-001

February 2011

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Richard Stern (chair)
Alan Black
Alex Waibel
Anton Batliner, *Friedrich-Alexander University*

Submitted in partial fulfullment of the requirements for the degree of Doctor of Philosophy



In memory of
my grandfather

Tadeusz Laskowski

Abstract

Understanding a conversation involves many challenges that understanding the speech in that conversation does not. An important source of this discrepancy is the form of the conversation, which emerges from tactical decisions that participants make in how, and precisely when, they choose to participate. An offline conversation understanding system, beyond understanding the spoken sequence of words, must be able to account for that form. In addition, an online system may need to evaluate its competing next-action alternatives, instant by instant, and to adjust its strategy based on the felicity of its past decisions. In circumscribed transient settings, understanding the spoken sequence of words may not be necessary for either type of system.

This dissertation explores tactical conversational conduct. It adopts an existing laconic representation of conversational form known as the vocal interaction chronogram, which effectively elides a conversation's text-dependent attributes, such as the words spoken, the language used, the sequence of topics, and any explicit or implicit goals. Chronograms are treated as Markov random fields, and probability density models are developed that characterize the joint behavior of participants. Importantly, these models are independent of a conversation's duration and the number of its participants. They render overtly dissimilar conversations directly comparable, and enable the training of a single model of conversational form using a large number of dissimilar human-human conversations. The resulting statistical framework is shown to provide a computational counterpart to the qualitative field of conversation analysis, corroborating and elaborating on several sociolinguistic observations. It extends the quantitative treatment of two-party dialogue, as found in anthropology, social psychology, and telecommunications research, to the general multi-party setting.

Experimental results indicate that the proposed computational theory benefits the detection and participant-attribution of speech activity. Furthermore, the theory is shown to enable the inference of illocutionary intent, emotional state, and social status, independently of linguistic analysis. Taken together, for conversations of arbitrary duration, participant number, and text-dependent attributes, the results demonstrate a degree of characterization and understanding of the nature of conversational interaction that has not been shown previously.

Acknowledgements

I am indebted to very many people for making the research described in this document, and the document itself, possible. In the first place, I would like to thank the members of my committee, Richard Stern, Anton Batliner, Alan Black, and Alex Waibel, for their patience, indulgence, availability, criticism, and constant encouragement. I would also like to thank Tanja Schultz, who advised me for the majority of my time as a CMU student. Special thanks are due to Mari Ostendorf, with whom working repeatedly resulted in new formulations, depth of understanding, and egress from impasse.

In the course of conducting the research presented here, I have been lucky to have worked with and learned from Susi Burger, Jens Edlund, Christian Fügen, Mattias Heldner, and Liz Shriberg. I am also very grateful to Qin Jin, Rob Malkin, Florian Metze, Matthias Paulik, Thomas Schaaf, Ashish Venugopal, Stephan Vogel, and Matthias Wölfel for the many occasions of discussion and argument which kept me on-track.

Funding and computational resources for the work described were provided by a Research Assistantship at the Institut für Theoretische Informatik at the University of Karlsruhe in Karlsruhe, Germany, a Research Assistantship at the Language Technologies Institute at CMU, a Teaching Assistantship at the Institut für Kognitive Systeme at the University of Karlsruhe, and through several collaborative efforts with the Speech, Music and Hearing department at KTH in Stockholm, Sweden. I am grateful to Alex Waibel, Tanja Schultz, Rolf Carlson, and Mattias Heldner for making these possible.

Likely the hardest part of this research for me was reporting it. If not for the sustained pressure from (sequentially) Anton Batliner, Alan Black, Richard Stern, Tanja Schultz, Susi Burger, and Florian Metze, this document would still be in its infancy. In addition to the attention it received in its entirety from my committee, several chapters have benefitted from extensive comments by Florian Metze (§2 & §3), Susi Burger (§4), Mattias Heldner (§5), Jens Edlund (§10), Khiet Truong (§13), and Tanja Schultz (§18 & §19). I take full responsibility for any omissions, inaccuracies, and errors that remain.

Contents

I OPENING MATERIAL **10**

1 Introduction **12**
- 1.1 Communication as a Form of Interaction . . . 12
- 1.2 Speech Exchange as a Form of Communication . . . 12
- 1.3 Conversation as a Form of Speech Exchange . . . 13
- 1.4 What This Thesis Is About . . . 13

2 Explicit Goals of this Research **14**
- 2.1 A Computational Theory . . . 14
- 2.2 Model Form Alternatives . . . 15
- 2.3 Invariable in Time, Invariable across Participants . . . 15
- 2.4 Invariable in Time, Variable across Participants . . . 17
- 2.5 Variable in Time, Invariable across Participants . . . 18

3 How to Read This Document **19**
- 3.1 Document Structure . . . 19
- 3.2 A Text-Independent Speech Understanding System . . . 21

4 Multi-Party Conversation Corpora **23**
- 4.1 ICSI Meeting Corpus . . . 23
- 4.2 ISL Meeting Corpus . . . 27
- 4.3 AMI Meeting Corpus . . . 28
- 4.4 NIST Rich Transcription Meeting Recognition Corpora . . . 28

5 Survey of Related Work **30**
- 5.1 Modeling Vocal Interaction . . . 30
- 5.2 Important Use Cases . . . 33
- 5.3 Modeling Emergent Chronogram Features . . . 34
- 5.4 Modeling the Chronogram as a Process . . . 37
- 5.5 Model-Driven Synthesis of Chronograms . . . 40
- 5.6 Summary . . . 40

II FACETS OF A COMPUTATIONAL THEORY **44**

6 Non-Parametric State-Space Multi-Participant Models **46**
- 6.1 Introduction . . . 46

6.2 Symbols and Definitions . . . 46
6.3 Direct Compositional Model Estimation . . . 50
6.4 Ergodic Binomial-Participant Models . . . 53
6.5 Non-Ergodic and Multinomial-Participant Models . . . 59
6.6 Relevance to Other Chapters . . . 61
6.7 Summary . . . 61

7 Parametric State-Space Multi-Participant Models 63
7.1 Introduction . . . 63
7.2 Revisiting the Compositional Conditionally Independent Model . . . 64
7.3 Ancillary Observations on Overlap . . . 64
7.4 Logistic Regression . . . 66
7.5 A One-Layer Feed-Forward Neural Network . . . 68
7.6 The Ising Anti-Ferromagnet . . . 69
7.7 Pseudo-Temperature . . . 71
7.8 An Example . . . 72
7.9 Relevance to Other Chapters . . . 74
7.10 Summary . . . 74

8 Parametric Feature-Space Multi-Participant Models 76
8.1 Introduction . . . 76
8.2 Rotating and Windowing Snapshots of Neighborhood . . . 77
8.3 Computing Durations to Observable Landmarks . . . 78
8.4 Representing the Complete Neighborhood Snapshot . . . 80
8.5 Relevance to Other Chapters . . . 82
8.6 Summary . . . 83

9 Parametric Feature-Space Multi-Channel Models 84
9.1 Introduction . . . 84
9.2 NT-Norm Maximum Cross-Channel Correlation . . . 85
9.3 "Model-Free" Multichannel Label Assignment . . . 87
9.4 Training Multi-Participant Acoustic Models on Small Data . . . 88
9.5 Fundamental Frequency Variation Spectrum . . . 91
9.6 A Filterbank for FFV Spectrum Compression . . . 95
9.7 Relevance to Other Chapters . . . 95
9.8 Summary . . . 95

III EMPIRICAL VALIDATION 99

10 Quantitative Analysis of Turn-Taking 101
10.1 Introduction . . . 101
10.2 Related Work . . . 102
10.3 Dataset Use . . . 102
10.4 Assessment of Performance . . . 102
10.5 Baseline . . . 103

10.6 Compositional Multi-Participant Modeling . . . 104
10.7 Exploiting the Systematics of Speech Overlap Dynamics . . . 106
10.8 Generalization to Unseen Data . . . 108
10.9 Potential Impact . . . 109
10.10 Relevance to Other Chapters . . . 109
10.11 Summary . . . 110
10.12 Future Directions . . . 110

11 Automatic Speech Activity Detection 111
11.1 Introduction . . . 111
11.2 Related Work . . . 112
11.3 Dataset Use . . . 113
11.4 Assessment of Performance . . . 114
11.5 Baseline . . . 116
11.6 Increasing the Frame Step . . . 118
11.7 NT-Norm Cross-Correlation Maxima . . . 120
11.8 Modeling Interlocutors . . . 123
11.9 Decorrelating Energy Features . . . 126
11.10 Minimum Duration Constraints . . . 132
11.11 Generalization to Unseen Data . . . 134
11.12 Alternative Metrics . . . 136
11.13 Potential Impact . . . 142
11.14 Relevance to Other Chapters . . . 146
11.15 Summary . . . 146
11.16 Future Directions . . . 146

12 Quantitative Analysis of Turn-Sharing 149
12.1 Introduction . . . 149
12.2 Related Work . . . 150
12.3 Dataset Use . . . 150
12.4 Constructing a Laughter Segmentation . . . 150
12.5 Analysis of the Occurrence of Laughter . . . 152
12.6 Potential Impact . . . 158
12.7 Relevance to Other Chapters . . . 159
12.8 Summary . . . 159
12.9 Future Directions . . . 159

13 Automatic Laughter Activity Detection 160
13.1 Introduction . . . 160
13.2 Related Work . . . 160
13.3 Dataset Use . . . 161
13.4 Assessment of Performance . . . 162
13.5 Baseline . . . 163
13.6 Imposing Minimum Duration Constraints . . . 163
13.7 Recognizing Voiced Laughter . . . 164

13.8 Recognizing Voiced and Unvoiced Laughter . . . 165
13.9 Modeling Interlocutors . . . 167
13.10 Generalization to Unseen Data . . . 168
13.11 Potential Impact . . . 168
13.12 Relevance to Other Chapters . . . 169
13.13 Summary . . . 169
13.14 Future Directions . . . 169

14 Text-Independent Dialog Act Recognition 171
14.1 Introduction . . . 171
14.2 Related Work . . . 173
14.3 Dataset Use . . . 174
14.4 Assessment of Performance . . . 175
14.5 Baseline . . . 175
14.6 A Single-Participant HMM Topology . . . 176
14.7 Modeling Interlocutors . . . 180
14.8 Locally Versus Globally Most Talkative . . . 182
14.9 Analysis of Performance for Specific DA Types . . . 183
14.10 Generalization to Unseen Data . . . 186
14.11 Contrasting ■/□ Context with Prosodic Information . . . 187
14.12 Contrasting ■/□ Context with Lexical Information . . . 190
14.13 Potential Impact . . . 192
14.14 Relevance to Other Chapters . . . 193
14.15 Summary . . . 194
14.16 Future Directions . . . 194

15 Text-Independent Emotional Epiphenomenon Recognition 195
15.1 Introduction . . . 195
15.2 Related Work . . . 196
15.3 Observations from Manual Annotation . . . 197
15.4 Classifying Emotional Valence . . . 206
15.5 Detecting Involvement Hotspots . . . 219
15.6 Detecting Attempts to Amuse . . . 233
15.7 Potential Impact . . . 240
15.8 Relevance to Other Chapters . . . 241
15.9 Summary . . . 241
15.10 Future Directions . . . 243

16 Text-Independent Conversation Characterization 244
16.1 Introduction . . . 244
16.2 Related Work . . . 245
16.3 Dataset Use . . . 245
16.4 Assessment of Performance . . . 246
16.5 Baseline . . . 246
16.6 Assuming Participants to Be Identical . . . 246

16.7 Assuming Participants to Be Different . . . 247
16.8 Combining Model Scores . . . 252
16.9 Potential Impact . . . 254
16.10 Relevance to Other Chapters . . . 254
16.11 Summary . . . 254
16.12 Future Directions . . . 254

17 Text-Independent Participant Characterization 256
17.1 Introduction . . . 256
17.2 Related Work . . . 256
17.3 Dataset Use . . . 257
17.4 Assessment of Performance . . . 257
17.5 Holistically Classifying Participant Groups . . . 257
17.6 Inferring Assigned Role . . . 261
17.7 Finding the Assigned Leader . . . 261
17.8 Inferring Gender . . . 262
17.9 Inferring Seniority . . . 263
17.10 Inferring Identity . . . 265
17.11 Potential Impact . . . 268
17.12 Relevance to Other Chapters . . . 268
17.13 Summary . . . 268
17.14 Future Directions . . . 269

IV CLOSING MATERIAL 271

18 Summary 273

19 Future Research Enabled by This Thesis 275

Bibliography 275

Part I

OPENING MATERIAL

Chapter 1

Introduction

1.1 Communication as a Form of Interaction

Interaction is a ubiquitous phenomenon. Virtually all entities, down from the smallest sub-atomic particle to the largest galaxy, through the variety of biological systems on earth, are enmeshed in some form of space shared with other entities with which they meaningfully interact. Over the past 100 years, the quantitative modeling of N-body systems has shed much light on the astounding complexity which can emerge from a very limited number of degrees of freedom available to each individual entity alone.

A common example of an N-body system characterized by interaction is a volume of galactic matter. The stars which comprise it continuously exert a gravitational force — a function of their mass and proximity — on all other stars, thus leading to perturbations in the position and velocity of all elements of the system. A large number of systems, particularly in the physical sciences, consist of elements whose influence is "always on" in this way, of magnitude mediated only by proximity.

In the biological sciences, influence among entities can take on a volitional aspect. A single individual may choose to drink a water hole dry, leaving nothing for others; alternately, it may choose to share. The mere proximity of a predator does not predetermine the outcome of a hunt, particularly if the would-be prey chooses to be cautious. By choosing to cut short a mating call, for example, the prey may increase its chances of survival, and simultaneously decrease its chances of attracting a co-present mate.

When the members of a species possess even minimal volitional control of expression, and a sensitivity to the expressions of their conspecifics, expression is likely to be co-opted for communication. Communication is a form of interaction, a characteristic of which is that it requires the expenditure of energy by both sender and receiver.

1.2 Speech Exchange as a Form of Communication

In humans, the volitional control of and sensitivity to expression is markedly more advanced than in other species, particularly through spoken language and the accompanying para-language. While technology to record the expression of a message has produced a range of "memory-full" channels of communication (e.g., pen and paper, email, answering machine, etc.), which allow receivers to desynchronize their effort of attention from a sender's effort, the delay associated with these modalities is generally unacceptable in co-present settings. The *memory-less* acoustic, visual, haptic, and olfactory channels are simply more efficient.

In co-present (and co-tele-present) settings, the acoustic memory-less channel is unique in one additional respect: while capable of carrying significantly more information than the olfactory channel, it can nominally support only one expression at a time (unlike the visual and haptic channels). Because acoustic expressions bearing dissimilar information are co-corruptive, the channel acts like a resource whose capacity is independent of the number of people present. Access to it must be ratified, if the efficiency of information exchange is to not deviate severly from the potential efficiency of the channel.

1.3 Conversation as a Form of Speech Exchange

It has been argued that access ratification is so crucial in speech exchange that it defines the type of speech exchange system in operation (§7 in [194], the most frequently cited publication in *Language*). Multi-party conversation, the particular type of speech exchange system studied in this thesis, is broadly characterized as

locally managed channel access (turn-order) is not predetermined;

party-administered channel access is not ratified by a specialized individual; and

interactionally controlled channel use (turn-length) is not dictated by a specialized individual.

Other types of speech exchange systems[1], differing from multi-party conversation on at least one of the above criteria, include: debates (two-party, role-undifferentiated, turns pre-allocated into "pro" and "con" alternations of fixed duration); (mediated) meetings (arbitrary-multi-party, role-undifferentiated, turns pre-allocated into rotations of fixed duration); press conferences (arbitrary-multi-party, single specialized role which unilaterally but locally allocates turns to others); seminars (arbitrary-multi-party, single specialized role which holds the turn from beginning to end); interviews (two-party, two specialized roles, turns allocated by interviewer); ceremonies (arbitrary-multi-party, some specialized roles, turns pre-allocated by custom, non-specialized parties share turns in overlap); and trials (fixed-multi-party, specialized roles, turns both pre-allocated by custom and locally allocated by a judge).

1.4 What This Thesis Is About

This thesis treats conversations, of arbitrary duration and with arbitrary numbers of participants, as N-body systems (or, as this document will prefer, K-body systems with K the number of participants) in which participants exert influence on one another via vocal activity, most often speech. The behavior of participants, with respect to the timing of the deployment of that activity, will be modeled as a stochastic process. The models developed are posited to implement a computational theory of interaction in conversation, enabling the prediction of future vocal activity within a conversation, the acoustic detection of conversational vocal activity, and the inference of explanations for the specific patterns of vocal activity observed.

Importantly, this thesis abstracts away from the words uttered during conversation. Instead, in modeling the timing of participatory events, it attempts to model the *context* within which words are sequenced. As such, it may contribute to meaning directly. But more importantly, it is likely to contribute to meaning in conversation in ways for which words alone are not well suited. Since the choice to participate vocally, in a specific context, constitutes a behavior, modeling participants' interlocking choices is likely to cast light on precisely those stationary characteristics, transient intentions, and transient internal states which participants would not normally communicate in a linguistically explicit manner.

The successful inference of meaning from the distribution of conversational vocal activity, in time and across participants, will help machines achieve the same degree of conversational competence that humans exhibit when the walk in on or overhear a conversation in an unknown language. Despite the obvious lexical handicap, we are still likely to correctly guess whether the listeners are in fact being attentive, how lively the conversation is, what would happen if we interrupted at some particular point, how the conversation might thereafter recover, why a laughing participant is doing so at precisely the point the that he is, and what the relative social status is of the individual conversants. We appear to retain these abilities even when we don't see the conversational group. At the current time, surprisingly, such competence is beyond the reach of ASR-, parser- and web-access- enabled machines.

[1] There exist also other exchange systems, which do not employ speech but which are instructive of the pecularities of multi-party conversation. One of the best examples is the IEEE 802.3 Medium Access Control (MAC) protocol for local area networks. The medium is shared, and parties (computers) requiring network access, to one another or to the wider world, cannot all signal at once. Message collision is possible, and the protocol provides contingencies for its occurrence; similar to what occurs in conversation, those contingencies implement a technique called *exponential backoff*.

Chapter 2

Explicit Goals of this Research

This thesis treats a conversation $\mathcal{C}$ as a single, contiguous interval of time, whose first and last discrete instants have indices of unity and $T \in \mathbb{N}$, respectively.

In this interval, $\mathcal{C}$ is a social occasion involving $K \in \mathbb{N}$ participants; this thesis considers all $K > 2$. Individual participants may repeatedly leave and rejoin the conversation before it is over, and new participants may join at any time. For simplicity and without loss of generality, this thesis assumes that K is the *total* number of distinct participants who vocalize at *any* instant in $\{1, 2, \ldots, T\}$, and that all of those participants are actually present or telepresent (but not necessarily vocalizing) over the interval's entirety.

What is observed over the course of $\mathcal{C}$ is $\mathbf{X}$. It is a $K \times T$ matrix of acoustic feature vectors $\mathbf{x}_t[k]$, with $1 \leq k \leq K$ and $1 \leq t \leq T$, describing the continuous acoustic state at instant t of the microphone worn by participant k.

Not observed is the matrix $\mathbf{Q}$, also of size $K \times T$, whose columns are $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_T\}$. $\mathbf{q}_t[k]$ represents the discrete vocal activity state of participant k at instant t. An example of $\mathbf{Q}$, with $\mathbf{q}_t[k] \in \{\square, \blacksquare\} \equiv \{\textit{non-speech}, \textit{speech}\}$ is shown in Figure 2.1.

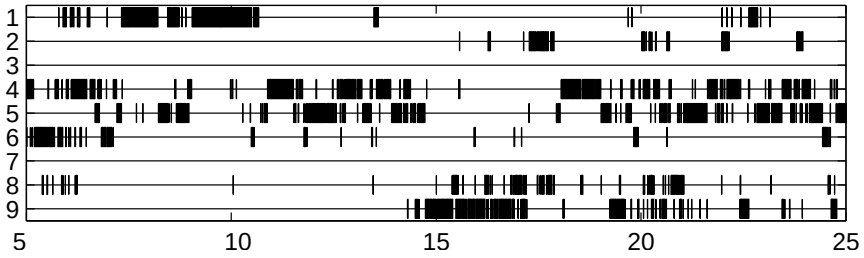


Figure 2.1: Vocal interaction chronogram $\mathbf{Q}$ for meeting `Bmr024` in ICSI Meeting Corpus, in the interval 5 to 25 minutes (along the x-axis). Speech activity shown in black, attributed to the nine participants (along the y-axis).

Finally, also not observed is the matrix $\mathbf{Y} \equiv \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T\}$, with $\mathbf{y}_t[k]$ representing the discrete intent/state of participant k at instant t.

2.1 A Computational Theory

Figure 2.2 depicts assumed dependencies among the observable $\mathbf{X}$ and the unobservable $\mathbf{Q}$ and $\mathbf{Y}$.

The goal of this thesis is to develop a *computational* theory of the distribution of $\mathbf{Q}$. The theory is embodied by a model, whose parameters can be empirically inferred from labeled training data. Once inferred, the model can provide the likelihood of hypothetical or actual $\mathbf{Q}$ in previously unseen data.

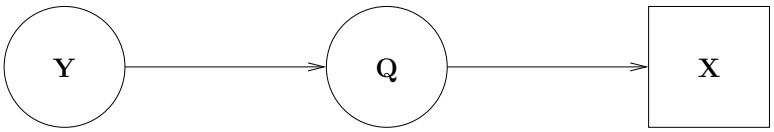


Figure 2.2: Unobservable intents/states $\mathbf{Y}$, unobservable vocal activity states $\mathbf{Q}$, and observed acoustics $\mathbf{X}$; all variables are matrices of size $K \times T$.

To achieve this goal, this thesis sets out to:

1. Propose an appropriate form for a probability density model $\mathbf{\Theta}$ over $\mathbf{Q}$, of arbitrary T and $K > 2$; and
2. Develop a tractable technique for estimating the parameters of $\mathbf{\Theta}$, using a manageable number of labeled training conversations of arbitrary and heterogenous T and $K > 2$.
3. Evaluate the joint felicity of (1) and (2), by demonstrating that the theory successfully enables previously inaccessible inference, and/or improves upon existing solutions.

2.2 Model Form Alternatives

The sample $\mathbf{Q}$ depicted in Figure 2.1 makes it clear that density varies both in time and across participants. A most general form of the model $\mathbf{\Theta}$ would explicitly account for this variability as function of t and k, for $1 \leq t \leq T$ and $1 \leq k \leq K$. This thesis does not treat the general form.

Instead, this thesis considers constrained forms of the general model, in which variability is licenced in at most one of the t or k dimensions. In the former case, the model parameters are not time-independent, but all participants are assumed to be identical. In the latter, the model parameters are fixed in time, but participants are not assumed identical. These two specialized model forms are depicted in Figure 2.3; also shown at the top of the figure is the general model which is neither invariable in time nor invariable across participants.

Finally, this thesis considers a form which eliminates dependence on both time t and participant index k. The model parameters are both fixed in time and assume identical participants; this third, least general form is also shown in Figure 2.3.

Requirements and application of the three constrained forms are described separately in the following, beginning with the least general. In considering these forms, it should be noted that the two indices of $\mathbf{X}$, $\mathbf{Q}$, and $\mathbf{Y}$ are of different scale; t is ordinal, while k is nominal. In particular, arbitrary row rotation of the indices k does not yield a linguistically different conversation, while arbitrary column rotation of the indices t does.

2.3 Invariable in Time, Invariable across Participants

2.3.1 Proposed Requirements

1. A requirement for this model form is that it be T-independent: the model parameters must be inferrable from conversations of unequal T_{train}, and the resulting model must be applicable to unseen conversations with T_{test} which may be different from any of the seen T_{train}. To achieve this, it is proposed that the model be decomposable in time, with fixed order. Here, the first-order Markov property is assumed:

$$P(\mathbf{Q} \mid \mathbf{\Theta}) \;=\; \prod_{t=1}^{T} P(\mathbf{q}_t \mid \mathbf{q}_{t-1}, \mathbf{\Theta}) \;.$$

Participant states $\mathbf{q}_t[k]$ may be conditionally dependent or independent, given the joint state $\mathbf{q}_{t-1}$ of the previous instant. Decomposing the likelihood $P(\mathbf{Q} \mid \mathbf{\Theta})$ in this way renders the model deployable as a transition probability

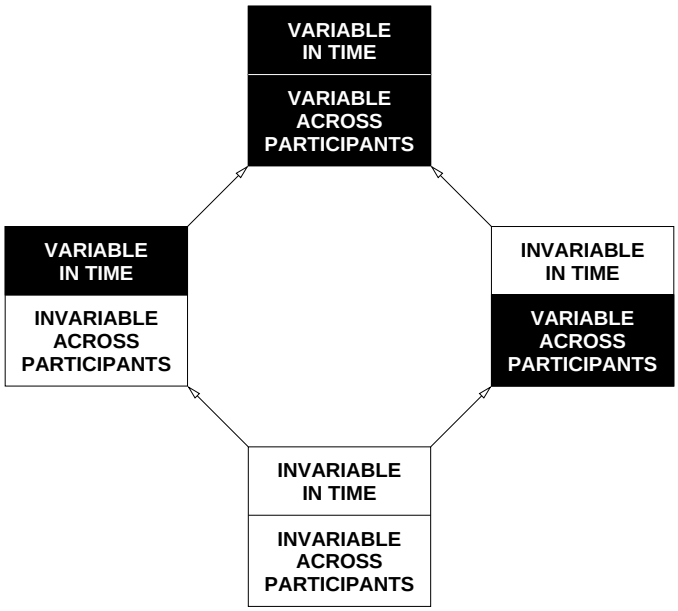


Figure 2.3: Four model forms, from most general (at the top) to most constrained (at the bottom). Arrows connecting any two forms point towards the more general.

model within a decoder. Transition probabilities are assumed to be time-independent, i.e.,

$$\begin{aligned} P\left(\mathbf{q}_t \,|\, \mathbf{q}_{t-1}, \mathbf{\Theta}\right) &= P\left(\mathbf{q}_t = \mathbf{S}_j \,|\, \mathbf{q}_{t-1} = \mathbf{S}_i, \mathbf{\Theta}\right) \\ &= a_{i,j} \ . \end{aligned}$$

2. The model must be K-independent: the model parameters must be inferrable from conversations of unequal K_{train}, and the resulting model must be applicable to unseen conversations with K_{test} which may be different from any of the seen K_{train}.

3. The model must be $\mathbf{R}$-independent: the model parameters must be invariant to arbitrary row rotation $\mathbf{R}$ in $\mathbf{Q}$ for the training conversations and the unseen conversations, i.e.,

$$P\left(\mathbf{Q} \,|\, \mathbf{\Theta}\right) = P\left(\mathbf{R} \cdot \mathbf{Q} \,|\, \mathbf{\Theta}\right) \ .$$

2.3.2 Application

Fulfillment of the above requirements yields a model $\mathbf{\Theta}$ which can provide the likelihood $P\left(\mathbf{Q} \,|\, \mathbf{\Theta}\right)$ for $\mathbf{Q}$ in at least three situations:

1. *Comparing any two conversations.* When $\mathbf{\Theta}_{train}$ is available and two conversations $\mathbf{Q}_{test}^{(1)}$ and $\mathbf{Q}_{test}^{(2)}$, of unequal dimensions $K^{(1)} \times T^{(1)}$ and $K^{(2)} \times T^{(2)}$, respectively, are to be compared, this can be achieved in the mapped space

$$\zeta\left(\mathbf{\Theta}_{train}; \mathbf{Q}_{test}^{(i)}\right) = \frac{1}{K^{(i)} \cdot T^{(i)}} \log P\left(\mathbf{Q}_{test}^{(i)} \,|\, \mathbf{\Theta}_{train}\right) \ .$$

An example of this is found in Chapter 10.

2. *Characterizing conversations.* When $\mathbf{\Theta}$ is not available and conversations $\mathbf{Q}_{test}^{(i)}$ of arbitrary dimensions $K^{(i)} \times T^{(i)}$ are to be characterized, a model $\mathbf{\Theta}_{test}^{(i)}$ may be inferred from each conversation and the parameters of that model may be used as T-, K-, and $\mathbf{R}$-independent descriptors for segmentation, classification, or regression. An example of this if found in Chapter 16.

3. *Inferring vocal activity.* When $\mathbf{\Theta}_{train}$ is available and $\mathbf{Q}_{test}$ is not observed but $\mathbf{X}_{test}$ is, alternative hypotheses $\mathbf{Q}_{test}^{(i)}$, of equal dimensions $K \times T$, may be compared implicitly to infer the most probable $\mathbf{Q}_{test}^{*}$:

$$\begin{aligned} \mathbf{Q}_{test}^{*} &= \arg\max_{\mathbf{Q}} P\left(\mathbf{Q} \,|\, \mathbf{X}_{test}, \mathbf{\Theta}_{train} \right) \\ &= \arg\max_{\mathbf{Q}} P\left(\mathbf{X}_{test} \,|\, \mathbf{Q} \right) \cdot P\left(\mathbf{Q} \,|\, \mathbf{\Theta}_{train} \right) \ . \end{aligned}$$

   Examples of this are found in Chapters 11 and 13.

2.4 Invariable in Time, Variable across Participants

2.4.1 Proposed Requirements

The requirements are assumed to be the same as those in Section 2.3, with two qualifications.

1. Participant states $\mathbf{q}_t\,[k]$ must be conditionally independent, given the joint state $\mathbf{q}_{t-1}$ of the previous instant,

$$P\left(\mathbf{q}_t \,|\, \mathbf{q}_{t-1}, \mathbf{\Theta} \right) = \prod_{k=1}^{K} P\left(\mathbf{q}_t\,[k] \,|\, \mathbf{q}_{t-1}, \mathbf{\Theta}_k \right) \ .$$

   The resulting model $\{\mathbf{\Theta}_k\}$ is an ordered composition of K models, one for each participant.

2. The model must *not* be $\mathbf{R}$-independent, unless all participants are identical.

2.4.2 Application

In this thesis, it will be assumed that when studying parameter variability across participants, or more correctly across participant profiles, $\mathbf{Q}$ has already been inferred. In the contrary case, it can first be inferred from $\mathbf{X}$ using the invariable-in-time and invariable-across-participants form of Section 2.3.

The variable-across-participants form finds application in:

1. *Comparing any two known-group, known-permutation conversations.* When $\{\mathbf{\Theta}_k\}$ is available for a known group of K participant profiles, with unknown but fixed index assignment, and two longitudinal conversations $\mathbf{Q}^{(1)}$ and $\mathbf{Q}^{(2)}$, of sizes $K \times T^{(1)}$ and $K \times T^{(2)}$, respectively, held under those conditions are to be compared, this can be achieved in the mapped space

$$\zeta\left(\{\mathbf{\Theta}_k\}; \mathbf{Q}^{(i)} \right) = \frac{1}{T^{(i)}} \log P\left(\mathbf{Q}^{(i)} \,|\, \{\mathbf{\Theta}_k\} \right) \ .$$

2. *Inferring group profile permutation in known-group conversations.* When $\{\mathbf{\Theta}_k\}$ is available for a fixed set of K participant profiles, with known and fixed index assignment, the participant profile index assignment $\mathbf{R}$ of a new converstion $\mathbf{Q}$ may be inferred via

$$\mathbf{R}^{*} = \arg\max_{\mathbf{R}} \ P\left(\mathbf{Q} \,|\, \mathbf{R} \cdot \{\mathbf{\Theta}_k\} \right) \ ,$$

   if all permutations can be assumed to be equally likely; there are $K!$ alternative permutations $\mathbf{R}$. An example of this is found in Chapter 17.

3. *Inferring group profile in arbitrary-group conversations.* When $\boldsymbol{\Theta}_j$ is available for all $j \in \{1, 2, \ldots, J\}$, where J is the number of different participant profiles under consideration, and $\mathbf{Q}$ is a conversation of known $K \leq J$ but unknown set of participant profiles and participant profile index assignment, the ordered participant profile index assignment $\mathbf{g}$ is a particular permutation of K items drawn from a set of J. It can be inferred via

$$\mathbf{g}^* \quad = \quad \arg\max_{\mathbf{g}} \; P\left(\mathbf{Q} \,|\, \mathbf{g} \cdot \{\boldsymbol{\Theta}_j\} \right) \; ,$$

if all permutations can be assumed to be equally likely. There are ${}^{J!}/_{(J-K)!}$ alternative permutations $\mathbf{g}$. An example of this is found in Chapter 17.

2.5 Variable in Time, Invariable across Participants

2.5.1 Proposed Requirements

The requirements for this form are assumed to be the same as those in Section 2.3, with one qualification.

1. The model must be T-independent. In contrast to the two variable-in-time forms, it is proposed that this property be achieved by windowing $\mathbf{Q}$, with a fixed-size window. Without subsequent factoring, the windowed region may be described by a $\mathbf{Y}$-conditioned emission probability model $\{\boldsymbol{\Theta}_y\}$.

2.5.2 Application

As in Section 2.4, it is assumed than when studying parameter variability in time, $\mathbf{Q}$ has already been inferred. When it is not, it can first be inferred from $\mathbf{X}$ using the invariable-in-time and invariable-across-participants form of Section 2.3.

The variable-in-time form finds application in:

1. *Inference of group-attributed intent/state.* When $\{\boldsymbol{\Theta}_y\}$ is available for all participant-unattributed intents/states y, and $\mathbf{Q}$ is a conversation with arbitrary dimensions $K \times T$, an optimal trajectory through participant-unattributed intents/states $Y \equiv \{y_1, y_2, \ldots, y_T\}$, with $y_t = \mathbf{y}_t[k] = \mathbf{y}_t[j]$ for all $k \neq j$ and any t, may be inferred using

$$\begin{aligned} Y^* \quad &= \quad \arg\max_{Y} P\left(Y \,|\, \mathbf{Q}, \{\boldsymbol{\Theta}_y\} \right) \\ &= \quad \arg\max_{Y} P\left(\mathbf{Q} \,|\, Y, \{\boldsymbol{\Theta}_y\} \right) \cdot P\left(Y \right) \; . \end{aligned}$$

Here, Y represents a single horizontal "ribbon" of $\mathbf{Y}$, of dimensions $1 \times T$, identically representing all participants. An example of this is found in Chapter 15, where Y represents the presence of an emotionally involved dialog act produced by *any* of the K participants.

2. *Inference of participant-attributed intent/state.* When $\{\boldsymbol{\Theta}_y\}$ is available for all participant-attributed intents/states y, and $\mathbf{Q}$ is a conversation with arbitrary dimensions $K \times T$, an optimal trajectory through participant-attributed intents/states $\mathbf{Y} \equiv \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T\}$, with potentially $\mathbf{y}_t[k] \neq \mathbf{y}_t[j]$ for all t, may be inferred using

$$\begin{aligned} \mathbf{Y}^*[k] \quad &= \quad \arg\max_{Y} P\left(Y \,|\, \mathbf{Q}[k], \{\mathbf{Q}[j \neq k]\}, \{\boldsymbol{\Theta}_y\} \right) \\ &= \quad \arg\max_{Y} P\left(\mathbf{Q}[k], \{\mathbf{Q}[j \neq k]\} \,|\, Y, \{\boldsymbol{\Theta}_y\} \right) \cdot P\left(Y \right) \; . \end{aligned}$$

Participant intents/states are assumed to be conditionally independent, given all K participants' $\mathbf{Q} \equiv \{\mathbf{Q}[k], \{\mathbf{Q}[j \neq k]\}\}$. An example of this is found in Chapters 14 and 15.

Chapter 3

How to Read This Document

3.1 Document Structure

This thesis is divided into 4 parts.

The first part contains an introduction to the problem of modeling emergent behaviors of groups of interacting entities. It presents the goals of this thesis, in the context of understanding multi-party conversational scenes; the current chapter; and a description of the data which is used throughout this thesis in examples and experiments. It also reviews past research which is relevant to the problems faced here.

The third and fourth parts of this document contain the technical contributions of this thesis. A targeted subset of model types is designed in Part II, *Facets of a Computational Theory*, which satisfy the desiderata listed in Chapter 2. These are shown in Figure 3.1. Those model forms which were referred to in that chapter as *invariable-in-time* are proposed to be implemented as state-space transition probability models, shown in the left branch of the top node in the figure. This thesis will explore both non-parametric, categorical models of this type, which assume that participants' vocal activity states are conditionally dependent, and parametric models which assume conditional independence of participants' vocal activity states, given all participants' vocal activity states at the previous instant. As shown at the bottom of Figure 3.1, these two types will implement the two forms of models of Chapter 2 (shown also in Figure 2.3) whose parameters are invariant with time.

The right branch of Figure 3.1 is an alternate approach, in which all participants' vocal activity is described using a joint emission probability model, conditioned on a single participant's intent or (higher-level) state. This approach is referred to as a feature-space approach, relying on a standard parametric probability distribution.

The three types of models, corresponding to the three forms of Chapter 2, are treated in Chapter 6 *Non-Parametric State-Space Multi-Participant Models*, Chapter 7 *Parametric State-Space Multi-Participant Models*, and Chapter 8 *Parametric Feature-Space Multi-Participant Models*, respectively. Part III also contains a Chapter 9 *Parametric Feature-Space Multi-Channel Models*, which presents several signal processing innovations which complement and facilitate the empirical work in this thesis.

Part IV presents an empirical validation of the algorithmic proposals of Part III. It is divided into 8 chapters, each treating one or more tasks which exercise complementary aspects of the computational theory. The 8 chapters are:

§10 (ATT) *Quantitative Analysis of Turn-Taking* (in speech)

§12 (ATS) *Quantitative Analysis of Turn-Sharing* (in laughter)

§11 (SAD) *Automatic Speech Activity Detection*

§13 (LAD) *Automatic Laughter Activity Detection*

§14 (TIDAR) *Text-Independent Dialog Act Recognition*

§15 (TIEER) *Text-Independent Emotional Epiphenomenon Recognition*

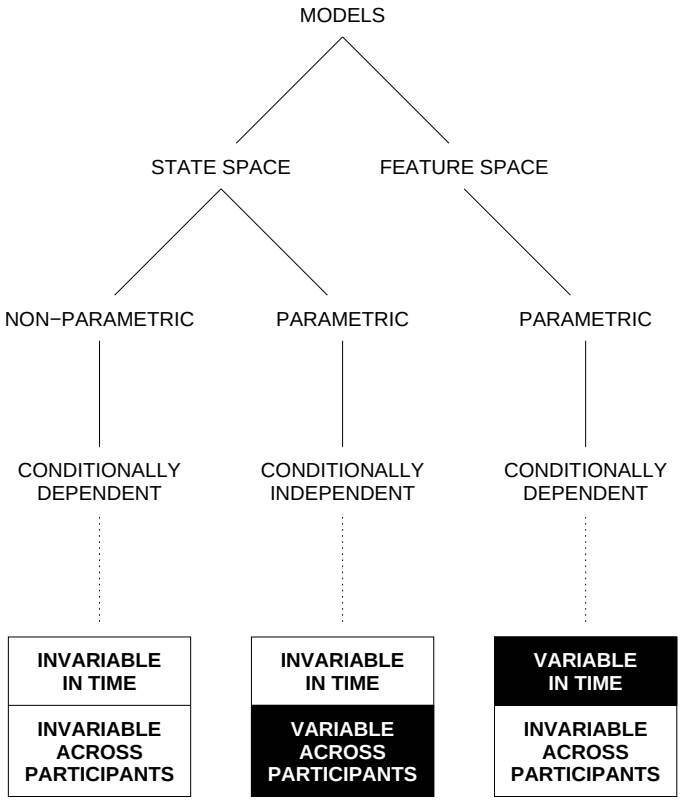


Figure 3.1: A characterization of different models of vocal interaction chronograms, found in Part II of this thesis. Objects in boxes as in Figure 2.3.

§16 (TICC) *Text-Independent Conversation Characterization*

§17 (TIPC) *Text-Independent Participant Characterization*

How these application areas employ the models of Part III is depicted in Figure 3.2. The first four applications, found in Chapter 10 through Chapter 13, exercise the invariable-in-time and invariable-across-participants form of Chapter 2. In these cases, nothing is known a priori about a conversation, and a model constrained to treat participants as identical and transition probability parameters as invariable in time is most appropriate. These four chapters break down in terms of focus on vocal activity type, with speech being the focus of Chapters 10 and 11, shown in the top two boxes, and laughter being the focus of Chapters 12 and 13, shown in the bottom two. They also break down in terms of use case: Chapters 10 and 12 (in the first column of boxes) apply models to observed $\mathbf{Q}$, while Chapters 11 and 13 (in the second column of boxes) apply very similar models to the acoustic detection of unobserved $\mathbf{Q}$.

Of the remaining four application chapters, those shown in the third column of boxes in Figure 3.2 rely on the variable-in-time but invariable-across-participants model form. Chapter 14 explores the recognition of DA type (or illocutionary intent $\mathbf{Y}$) using only the distribution of speech activity $\mathbf{Q}$. Chapter 15 in part extends this work to consider a conversational behavior which is epi-emotional in nature, namely the recognition of speech intented to amuse. It also considers two other emotion-related quantities: (1) the classification of emotional valence, in segmented user-attributed utterances; and (2) the detection of conversational intervals containing speech with marked emotional activation (involvement). Chapter 14 relies exclusively on speech activity chronograms, while Chapter 15 relies mostly on laughter activity chronograms.

The two final chapters of Part IV, shown in the last column of boxes, exercise the invariable-in-time but variable-across-participants model form. Chapter 16 shows that longitudinally obtained $\mathbf{Q}$ statistics for groups meeting regularly are sufficient to identity which group held a previously unseen meeting. Meanwhile, Chapter 17 demonstrates that $\mathbf{Q}$

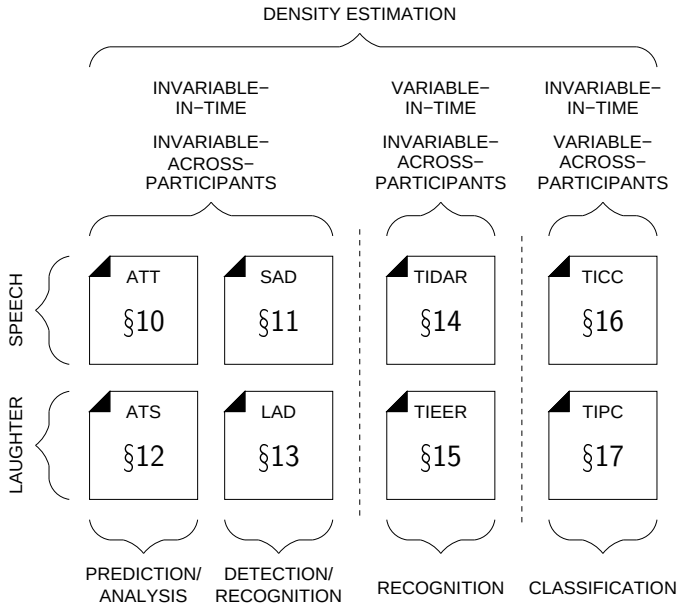


Figure 3.2: Eight applications of the algorithmic contributions of Part III, and the ensuing chapters of Part IV in which they are described. Acronyms as in the text.

statistics discriminate among assigned roles in unscripted meetings, and among diffuse social characteristics (notably seniority) in naturally occurring meetings.

The final part of this thesis, Part V, collects and summarizes the research and discusses some avenues for future directions which this thesis enables.

3.2 A Text-Independent Speech Understanding System

Although the above 8 tasks are explored in order to validate aspects of the modeling approach(es) proposed in Part III, they can conceivably be pieced together into a broader system for understanding conversation. A conceptual system is shown in Figure 3.3, with 8 components each implementing the techniques of each of the 8 chapters. Those described in Chapters 11 and 13 are seen to provide inference of $\mathbf{Q}$ given $\mathbf{X}$. Once $\mathbf{Q}$ is available, the techniques of Chapters 14, 15, 16, and 17 extract higher-level information related to intended speech function (dialog act type), expressed emotional state or intent to affect the emotional state of others, conversation type and/or group style, and assigned role or exhibited social traits of individual participants, respectively.

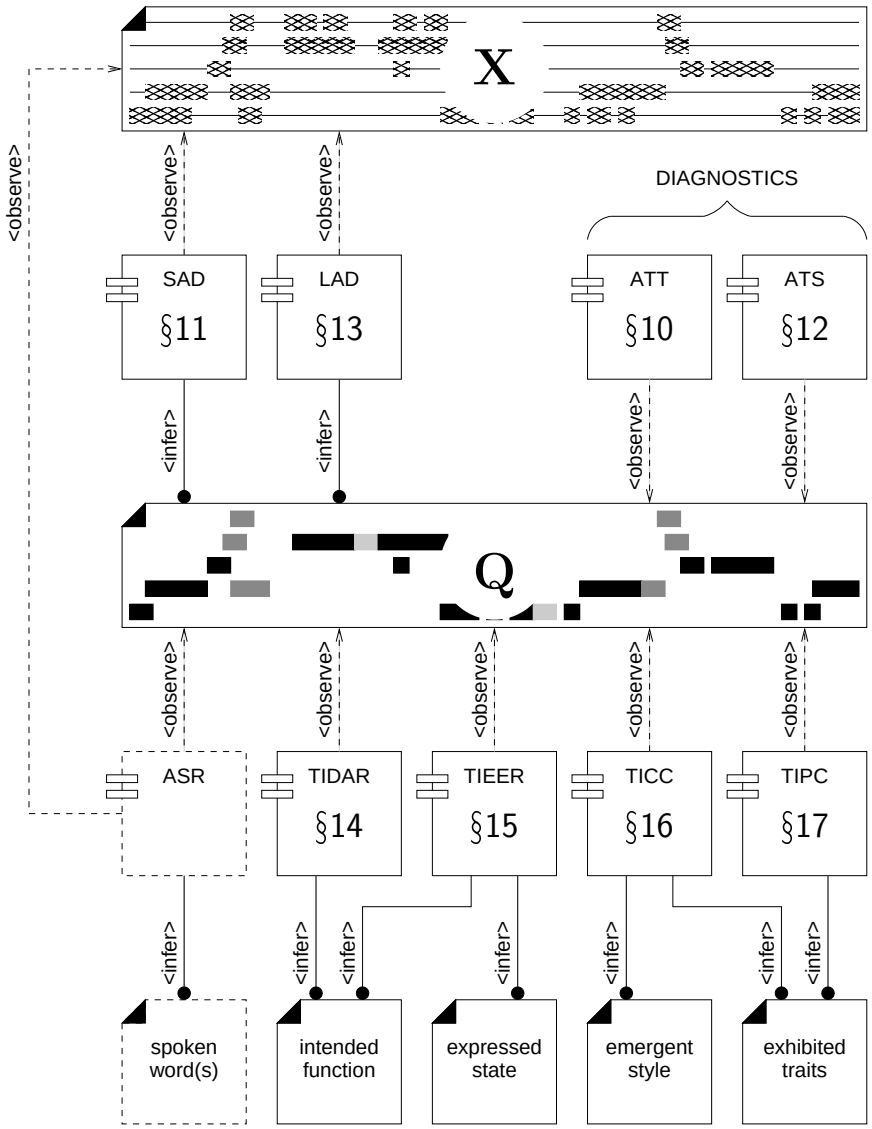


Figure 3.3: A hypothetical multi-party conversation understanding system with chapters in Part IV implemented as components. Also depicted are the artifacts which components observe or infer. Automatic speech recognition (ASR) is shown for completion, but is beyond the scope of this thesis. Not all dependencies shown. Acronyms as in the text.

Chapter 4

Multi-Party Conversation Corpora

4.1 ICSI Meeting Corpus

The ICSI Meeting Corpus was collected at the International Computer Science Institute in 2000-2002, and was first described in 2003 [108]. An important aspect of this data is that it is *naturally occurring* rather than *scenario-based*; the meetings would have occurred anyway, even if they had not been recorded. For practical purposes, the meetings are all recorded in one room at ICSI; the overwhelming majority of the corpus consists of longitudinal recordings of 3 project groups which met regularly.

4.1.1 Audio

Audio recordings of the corpus are available through the Linguistic Data Consortion (LDC) as "ICSI Meeting Speech", catalog number `LDC2004S02`. They consist of both nearfield, close-talk microphone channels for each meeting and farfield channels recorded with table-top microphones of various type. This thesis treats exclusively the nearfield recordings, from which the cleanest "oracle" transcriptions and annotations were obtained.

The sampling rate of the audio is 16 kHz, with a resolution of 16 bits; the compression applied to the recordings was lossless. The channels were recorded with a single sound card, and were meant to be sample-synchronous (post-mortem analysis had revealed that there was an unknown but constant lag between the channels due to a sound card firmware or driver bug[1]). An example of the audio from all of the close-talk microphones in an excerpt of meeting `Bed010` is shown in Figure 4.1.

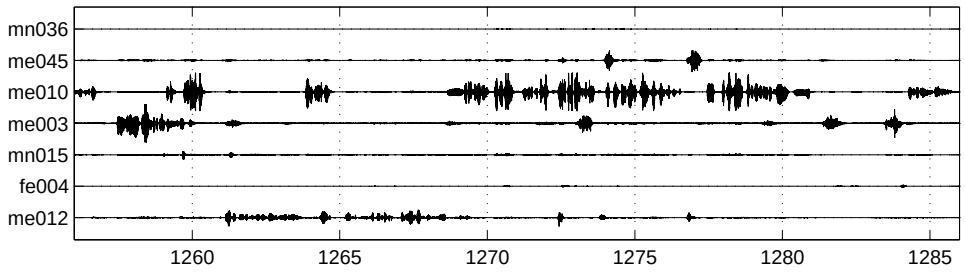


Figure 4.1: Sample-synchronous multi-channel close-talk-microphone audio of an excerpt of meeting `Bed010`. Speaker identifiers along the y-axis; time shown from left to right, in seconds.

[1] As described in `http://www.icsi.berkeley.edu/~dpwe/research/mtgrcdr/chanskew.html`.

4.1.2 Orthographic Transcription

The audio is accompanied by orthographic transcription, available through the LDC as "ICSI Meeting Transcripts", catalog number `LDC2004T04`. The transcriptions were produced by manually segmenting each close-talk channel into *utterance* or *speaker contribution* units. The endpoints of these contributions are stored in the XML-formatted `.mrt` files for each meeting. An example of the temporal extent of speaker contributions, for the same excerpt of `Bed010` as in Figure 4.1, is shown in Figure 4.2.

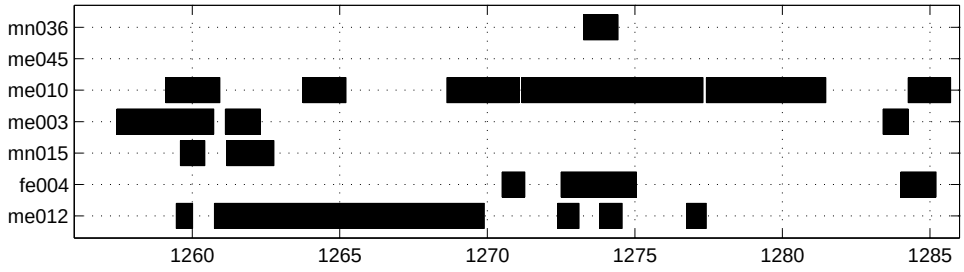


Figure 4.2: Manually produced speaker-attributed utterance-level segmentation of the meeting excerpt shown in Figure 4.1. Intervals in black indicate within-utterance instants. Speaker identifiers along the y-axis; time shown from left to right, in seconds.

The transcription is shown in Figure 4.3. To make the interaction among participants more apparent, it has been cast here into a format used frequently in conversation analysis (CA) work [99]; however, certain aspects were left unmodified, which violate the CA guidelines[2].

4.1.3 Auxillary Annotation

The ICSI Meeting Corpus has also been labeled in other ways, and the resulting annotations, known collectively as the "Meeting Recorder Dialog Act (MRDA) Corpus" [202], are not available through the LDC. Instead, they are distributed as a single `.tar.gz` file by request from ICSI. The version of that file is encoded in the file name; the version used in this thesis comes from `icsi_mrda+hs_corpus_050512.tar.gz`.

The first level of auxillary annotation of concern here is the lexical forced-alignment of all transcribed words and word fragments, an example of which is shown in Figure 4.4. As can be seen, utterances (in gray) contain significant stretches of non-speech interspersed among words (in black). This makes the lexical forced-alignment segmentation a more correct representation of when speech versus non-speech is occurring, and therefore more suitable at least for acoustic modeling purposes.

[2]The following mappings were made:

`O_K`	↦	`Okay`
`H_U_G_I_N`	↦	`H U G I N`
`<VocalSound Description="sniff"/>`	↦	`((sniff))`
`<VocalSound Description="laugh"/>`	↦	`((laugh))`
`<NonVocalSound Description="mike noise"/>`	↦	`((mike noise))`
`<Comment Description="while laughing"/>`	↦	`((laughing))`
`<Comment Description="whispered"/>`	↦	`((whispered))`
`<Uncertain>`*word*`</Uncertain>`	↦	(*word*)
`<Emphasis>`*word*`</Emphasis>`	↦	$\underline{WORD}$

With the exception of the above, lexical forms were retained as in the original ICSI transcription; neither duration nor phonetic variability are shown, as would normally be the case in a CA transcript.

```
me003:  Okay, so then I'll go back and look at the ones
        [on the l]ist [that -     ]
me010:  [Okay.   ]    [And you can] ASK Kevin.
me012:                Y[eah.       ]
mn015:                 [But -      ]
             (0.3)
me012:  Yeah, the [one that] uh people seem to use =
me003:            [M[mm.   ]
mn015:              [But - ]
me012:  = is uh Hugin or whatever?        [How exp-         ] =
me010:                              Hugin, [yeah that's free.]
me012:  = I don't think it's - Is it free? Because I've seen it
        ADVERTISED in places so I - it [seems] [to -  ]
me010:                                 U[h it ] [may be] free to
        academics. Like I - [I don't know.]
fe004:                      [((sniff))    ]

me010:  I have a co- ((laughing))
        [I [have a CO][PY ((laughing)) that [I   ] l- I] downloaded.] =
me012:  [((laugh))   ]
fe004:     [((laugh))                                              ]
mn036:                [((mike noise))                 ]
me012:                                      O[kay.]
me010:  = So, at ONE point it was free
me012:  Okay.
             (0.3)
me010:  Uh but yo- I noticed people DO use Hugin so um,
             (2.4)
me003:  How do you spell t[hat  ]?
fe004:                    [(Why)] ((whispered))
me010:  H U G I N.
```

Figure 4.3: A quasi-CA-style rendition of the manual speaker-attributed orthographic transcription for the meeting excerpt shown in Figure 4.2. "[" and "]" indicate intervals of overlap; "=" indicates speaker continuity across line-breaks. Participant labels in light gray are for ease of reading only, and do not denote a new utterance.

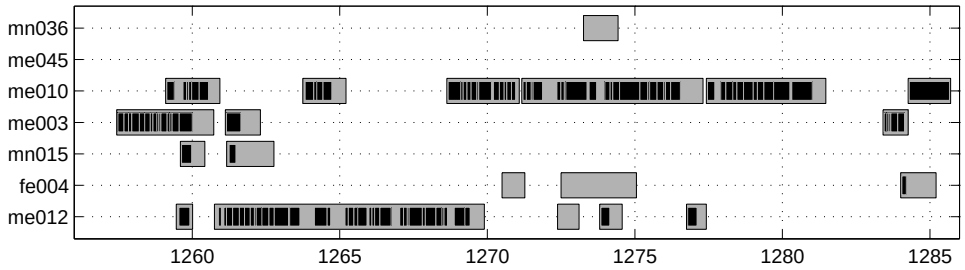


Figure 4.4: Forced-alignment-mediated speaker-attributed word-level segmentation of the meeting excerpt shown in Figure 4.2. Intervals in black indicate within-word instants, those in gray indicate within-utterance instants of Figure 4.2. Speaker identifiers along the y-axis; time shown from left to right, in seconds.

The second level of annotation of interest is the marshalling of words into dialog acts (DAs). The difference between utterances and dialog acts is that the former ignore speech function, and may therefore consist of several consecutive dialog

acts. The construction of dialog acts on top of words involves: (1) segmentation by deciding whether inter-word locations are DA boundaries; and (2) classification of inter-DA-boundary intervals into one of several DA types. The annotation scheme produced by ICSI is quite complex, involving required mutually exclusive base forms and a large number of optional tags; it is explained in detail in [54], and inter-labeler agreement is discussed in files which are part of the data release. An example of DA annotation, on top of a lexical forced-alignment segmentation, is shown in Figure 4.5.

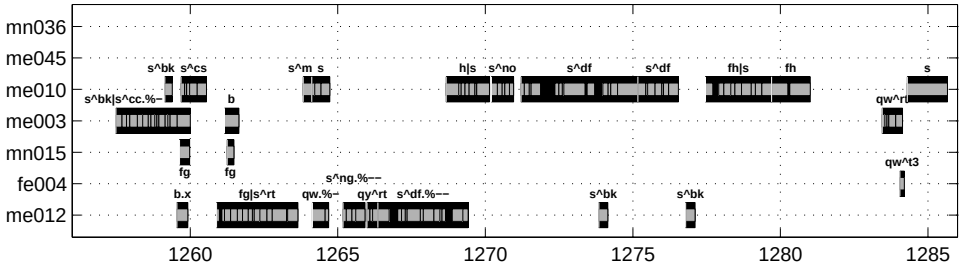


Figure 4.5: Manual segmentation of speaker-attributed words shown in Figure 4.4 into dialogue acts (DAs), and manual annotation of the latter into types. Intervals in black indicate within-DA instants, those in gray indicate within-word instants of Figure 4.4. Speaker identifiers along the y-axis; time shown from left to right, in seconds.

The ICSI MRDA Corpus also contains the annotation of adjacency pairs, which link pairs of dialog acts produced by two different speakers into pragmatically defined first-part and second-part pairs. This annotation level is also described in [54], with inter-labeler agreement statistics included in the data release. An example of the annotation of adjacency pairs is provided in Figure 4.6.

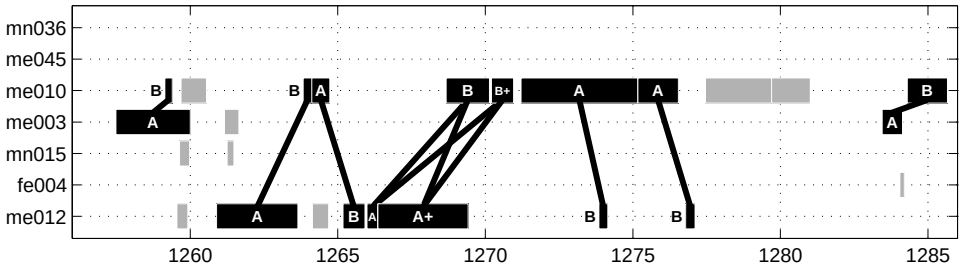


Figure 4.6: Manual structuring of speaker-attributed dialogue acts (DAs) shown in Figure 4.5 into adjacency pair parts, with first parts labeled as "A" (and "A+"), and second parts labeled as "B" (and "B+"). Intervals in black indicate DAs which are part of an adjacency pair, those in gray indicate DAs which are not. Speaker identifiers along the y-axis; time shown from left to right, in seconds.

Finally, the ICSI MRDA Corpus includes the annotation of "hot spots", or intervals of meetings "that stand out from the rest of the conversation in that

- the participants are more involved (emotionally or "interactively")
- there is a higher degree of interaction by participants who are trying to get the floor [.]" [223]

The published "hot spot" annotation deviates significantly from earlier work on "hot spots", published by the same authors in [225, 224]; "hot spots" in this thesis are therefore referred to as VERSION2 hotspots. In contrast to the first version, VERSION2 hotspots have internal structure; they are annotated on top of dialog acts, and nominally include a "trigger" and a "closure", which themselves need not be involved. Only one DA between the trigger and the closure must be involved. The annotation of VERSION2 hotspots is described in [223], and interlabeler agreement statistics can be found in [226]. An example of a hotspot is shown in Figure 4.7.

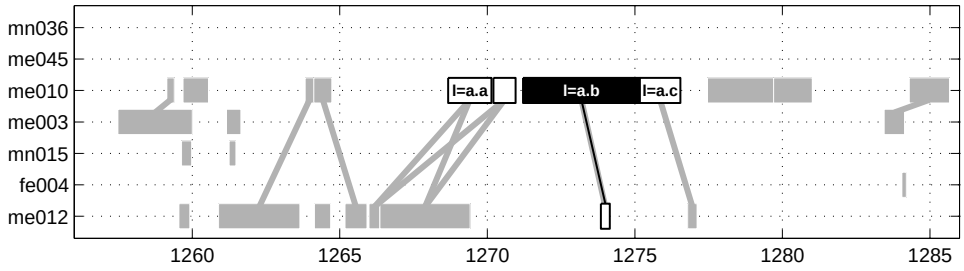


Figure 4.7: Manual structuring of speaker-attributed dialogue acts (DAs) shown in Figure 4.5 into "hot spots" (VERSION2); a single "hot spot" is shown. Intervals in black indicate DAs which are involved; those in white indicate non-involved DAs which are part of the "hot spot"; other DAs shown in gray. Links indicate adjacency pair structuring as in Figure 4.6. Speaker identifiers along the y-axis; time shown from left to right, in seconds.

4.1.4 Use in This Thesis

The speech/non-speech segmentation for meeting `Bmr025` is used as an example in Chapter 7. The speech/non-speech segmentation for the entire corpus is used in: Chapter 10 for perplexity computation (with exerpts of meeting `Bmr024` in the graphical examples); Chapter 11 to train models for SAD decoding; Chapter 14 as the input to text-independent dialog act recognition; Chapter 15 as the input to text-independent detection of emotionally involved speech and of attempts to amuse; Chapter 16 as the input to text-independent characterization of conversation type; and Chapter 17 as the input to text-independent characterization of participant gender, seniority, and identity.

In addition, this thesis describes the construction of a laughter segmentation from the audio and orthographic transcription, and an analysis of the segmentation, in Chapter 12. The laughter segmentation is used in: Chapter 13 in acoustic detection of laughter; Chapter 15 as the input to text-independent detection of emotionally involved speech and of attempts to amuse; and Chapter 17 as the input to text-independent characterization of participant identity.

4.2 ISL Meeting Corpus

The ISL Meeting Corpus was collected at the Interactive Systems Lab in 1999-2002, and has been described in [30, 31]. Although it consists of over 100 meetings, only 18, comprising a "Part 1" release, have been made available to date for research use. In contrast to the ICSI Meeting Corpus, the ISL corpus does not consist of many longitudinal recordings of the same handful of project groups. Also, only a subset of ISL meetings, those of type *work planning meeting*, *project meeting* and *military block party*, can be said to be naturally occurring and of a work-oriented nature. The remainder are of type: *game*, where a collaborative goal was *a priori* assigned and exectuted under clock pressure; *discussion*, where a competitive goal was *a priori* assigned (without role assignment); and *chatting*, where no *a priori* goal was assigned. Interaction, even under these unnaturally occurring conditions, was not scripted. Participants had various degrees of familiarity with one another, which correlated broadly with meeting type, and some participants reoccur in more than one meeting and in different meeting types.

4.2.1 Audio

The publicly released portion of the ISL Meeting Corpus is available from the LDC as "ISL Meeting Speech Part 1", catalog number `LDC2004S05`. The majority of the audio was collected using wired lapel microphones, and only occasionally using wireless and farfield table-top microphones.

Audio was sampled at a rate of 16 kHz, with 16 bits per sample; compression was not applied. All channels were recorded sample-synchronously on a single 8-channel sound card. Visual sample alignment of multiple channels leads to figures such as Figure 4.1; it is generally assumed and observed that acoustic coupling, or crosstalk, is more significant when lapel rather than head-mounted microphones are used.

4.2.2 Orthographic Transcription

The transcription for the corpus is available from the LDC as "ISL Meeting Transcripts Part 1", catalog number `LDC2004T10`. Channel audio was segmented into utterances, whose endpoints are available in one-line-per-utterance format, in one `.mar` file per meeting. The lexical transcription, also in one-line-per-utterance format, is available in each meeting's corresponding `.trl` file. It follows the well-established Verbmobil conventions [28, 219] (yielding what is known as the `TRL` tier [221] of the BAS Partitur Format [199]).

The `.mar` utterance segmentation yields chronograms similar to that shown in Figure 4.2, which, together with the `.trl` transcription and word boundary information (which is not distributed with the ISL Meeting Corpus), could be used to construct quasi-CA-style transcripts such as that shown in Figure 4.3.

4.2.3 Use in This Thesis

The utterance segmentation, orthographic transcription, and audio for the entire corpus are used in Chapter 15 to develop an annotation scheme for behavior thought to be emotionally relevant, and to evaluate the classification of emotional valence in manually segmented utterances.

4.3 AMI Meeting Corpus

The AMI Meeting Corpus was collected at several sites across Europe as part of the Augmented Multi-party Interaction project (`http://corpus.amiproject.org/`), during 2004-2005. The corpus consists of two parts, one of 33 "real" or naturally occurring meetings, and another of 138 scenario-driven meetings. This thesis is concerned exclusively with the latter half. The scenario is fixed, consisting of a remote-control design task. These meetings are attended by exactly four participants, each with an assigned, unique role (*project manager*, *marketing expert*, *user interface designer*, and *industrial designer*). People not connected with the AMI project were invited to participate in the meetings; they had neither prior experience in their assigned role nor were they professionally trained for design.

The corpus is described in detail in [36, 170, 35]. It is accompanied by a rich set of multimodal annotations, most of which were not used in this thesis. It can be accessed free of charge from `http://corpus.amiproject.org/`.

4.3.1 Audio

The audio channels that are a part of the AMI Meeting Corpus were not used in this thesis.

4.3.2 Orthographic Transcription

The orthographic transcriptions of each meeting, delivered together with the audio channels, are stored in an XML format. Each lexical item has attributes encoding that item's start and end time. It is therefore extremely easy to extract a picture like that in Figure 4.4 for each meeting.

4.3.3 Use in This Thesis

The speech/non-speech segmentation for the entire corpus is used in Chapter 17 as the input to text-independent participant role classification.

4.4 NIST Rich Transcription Meeting Recognition Corpora

The NIST Rich Transcription Meeting Recognition Corpora are sets of excerpts of multi-party meetings collected at different sites for different purposes and projects. They have been re-segmented and re-transcribed by NIST for use in the annual or biannual Rich Transcription technology evaluations held in the spring of 2004, 2005, 2006, 2007, and 2009. This thesis is concerned only with those sets distributed as `rt05s_eval` in 2005 and, for the `confmtg` task (as opposed to the `lectmtg` task) in 2006, `rt06s_eval`.

A description of the Rich Transcription task can be found at `http://www.itl.nist.gov/iad/mig//tests/rt/`. The corpora are available from NIST upon request.

4.4.1 Audio

The audio has been delivered in various formats over the years, and includes nearfield (head-mounted microphone) channels, farfield table-top channels, farfield microphone array channels, and a beamformed microphone array channel. This thesis is concerned exclusively with the nearfield channels. For the purposes of experiments, the audio was converted to 16 kHz, 16-bit quality.

In the `rt05s_eval` subset, one meeting involved a participant which was not wearing a nearfield microphone. This excerpt has been excluded from `rt05s_eval` to yield the ablated `rt05s_eval*`; this practice has become standard in recent years when working with this data (eg. [148, 19]).

4.4.2 Orthographic Transcription

Meeting excerpt endpoints are available in `.uem` files, utterance segmentation and orthographic transcription are available in `.stm` files, and auxillary speech/non-speech activity segmentation is available in `.rttm` files. The formats of these files were homogenized by NIST across the data-contributing sites.

4.4.3 Use in This Thesis

Both `rt05s_eval*` and `rt06s_eval` are used to evaluate SAD performance in Chapter 11.

Chapter 5

Survey of Related Work

This chapter sets out to describe research in general-purpose modeling of vocal interaction. For better or for worse, conversational work related to this thesis is historically scattered over a large number of disciplines, which vary in terms of the number of conversants, the purpose of modeling, and whether what is modeled are features of an interaction chronogram or the chronogram itself, as a generative process.

5.1 Modeling Vocal Interaction

5.1.1 Contributing Fields of Inquiry

It is generally assumed that multi-party conversation involves at least *some* words, rendering it an emergent product of participants' *verbal behavior*, and thus a concern of *linguistics*. However, verbal behavior and *vocal behavior*, as used in this thesis, are descriptors of neither the same nor mutually exclusive phenomena. Figure 5.1 attempts to make this clear; it is meant to be only coarsely accurate.

While the two tiles on the right depict subfields of linguistics, devoted to the study of verbal behavior, those on the left comprise a domain frequently referred to as *nonverbal behavior* [121, 89, 180, 79]. The latter includes the study of communicative manipulation, both conscious and subconscious, of the human body (i.e., *kinesics* [17] and *haptics*), eye gaze (i.e., *oculesics*), space (i.e., *proxemics* [86]), time (i.e., *chronemics*), and voice (i.e., *vocalics*). The manipulation of voice is clearly a vocal behavior, while that of body and gaze clearly are not. Chronemics and proxemics, however, are relevant to both non-vocal and vocal behavior; the former because all behaviors are extensive in time, and may begin and end at discriminative instants, and the latter because distance between speakers affects communication efficiency (for vocal behavior, it constrains voice loudness). Research in nonverbal behavior tends to treat the six modalities in the two tiles on the left of Figure 5.1 as cues, not necessarily possessing deep structure, and a central concern is their multi-modal *integration* [204].

The various subfields of linguistics, on the other hand, have a more vertical relationship, owing in part to the presence of relatively deep structure in language. Although multiple modalities are possible (speech, sign language, written text, etc.), with subfields devoted to their pecularities (*phonetics*, *cheremics*, *graphemics*, respectively), they are not normally assumed to be simultaneously present. Word construction and sequencing (treated by *syntactics*), the meaning of word sequences (*semantics*), and the grounding of that meaning (*pragmatics*) are applicable regardless of modality.

As Figure 5.1 shows, despite the division of labor between fields devoted to verbal and nonverbal communication, there exists considerable overlap. Vocalics treats non-phonemic aspects of vocalization, which include both non-speech vocalizations such as laughter and throat clearing, and non-phonemic aspects of speech. The latter, together with chronemic concerns such as pause duration and the rate of vocalization, are known as *prosody* in linguistics. Similarly, proxemics and chronemics — when applied to ground meaning in situated speech — are formally considered by linguists to belong in the domain of pragmatics.

In modeling *vocal interaction* as the multi-party production of vocal behaviors, shown in the bottom two tiles of Figure 5.1, this thesis can be seen as an attempt to quantify certain aspects of pragmatics, prosody, vocalics, chronemics,

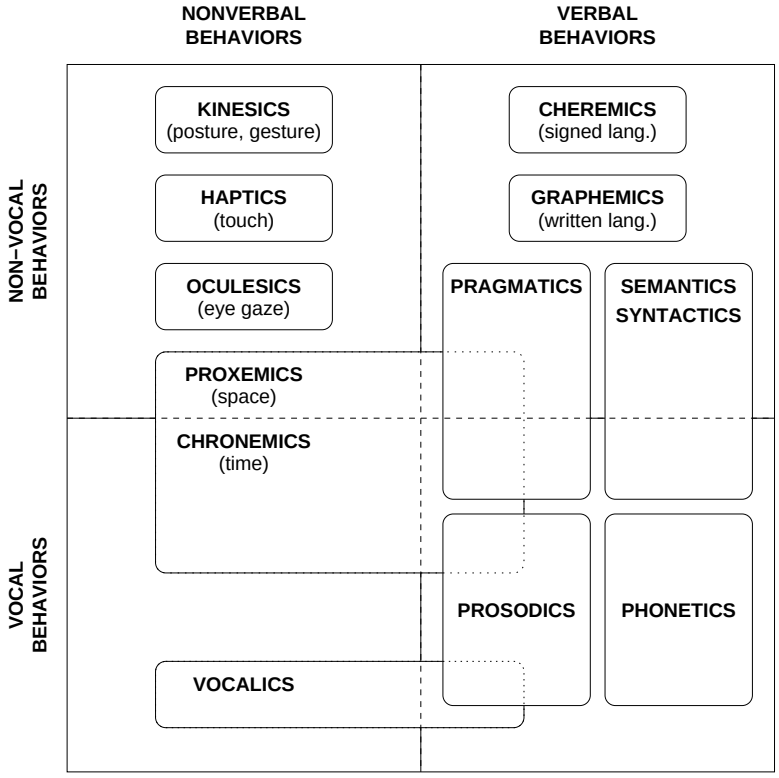


Figure 5.1: Four human communicative behavior groups, on a two-by-two grid; areas of science devoted to their study in arc-corner rectangles, with some examples in parentheses.

and, to a very limited and indirect degree, proxemics. It is not directly concerned with phonemic distinctions in speech, and therefore, by extension, with either syntax or semantics.

5.1.2 Number of Participants

There is considerable variation across the above fields of inquiry in terms of the size of the conversational group which has been studied.

Monologue has almost always formed the starting point of inquiry, for two reasons. First, an understanding of pragmatic, prosodic, vocalic, chronemic and proxemic aspects of single-participant vocalization is necessary to identify any effect that the presence of interlocutors may have. Second, modeling one source is less complex than modeling multiple sources, and has therefore been historically more attractive; associated with this is the fact that single-participant modeling does not require multi-source synchronization.

Unfortunately, collections of monologue are frequently unnatural, precisely because interlocutors are absent. While read or prepared speech may not deviate markedly from spontaneous speech at the segment level, the same is not true for those of its aspects which are relevant to this thesis. The likely only reason for the occurrence of monologue is its recording. *Dialogue* comprises a much more applicable starting point. The majority of computational models of the occurrence of situated vocalization have been developed for the two-party conversation scenario, in both social psychology and telecommunications (cf. Section 5.4).

The focus of this thesis is on *multi-party conversation*, of which the two-party case is a particular specialization. Although "multi" is intended to denote any number of participants greater than two, there is a general bias towards groups of less than approximately ten. The interaction in these groups is the focus of *small group research* [8, 107, 87, 171]. Groups which are much larger than this tend not to engage in conversation per se; they are generally not cohesive, with preference

towards forming multiple and largely independent conversational subgroups (the extreme case of which is disintegration into "coctail party effect" dyads). Alternately, when large groups must talk together they require mediation, which eliminates much of the uncertainty inherent in small-group turn-taking, leading to not conversation proper (as defined by conversation analysis) but a different type of speech exchange system.

There are some subtle but important variations reported for group behavior even within small group research; notably, the smaller the group, the less unstructured is the interaction [60]. Groups of four or more participants engender the potential for schism [194, 197]; groups larger than five or six participants tend to exhibit more marked variation in total vocalizing time across participants.

Modeling of interaction in groups is typically achieved at the level of the number of transactions, rather than of their temporal extent [78].

5.1.3 The Functional Role of Modeling

Modeling *content-free* or *text-independent* aspects of conversation may have different purposes, which may be inherently different from a functional, mathematical perspective. Several alternatives are shown in Figure 5.2.

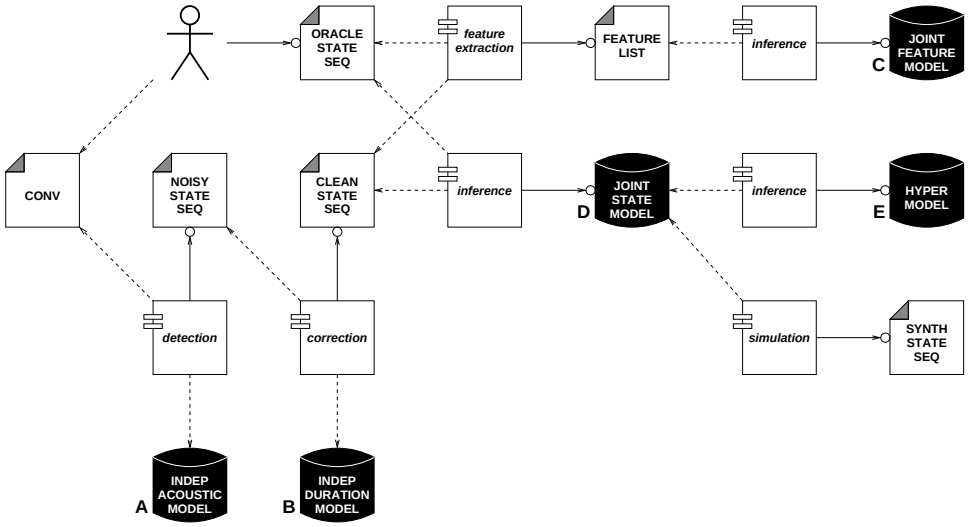


Figure 5.2: Functional contexts of different types of models, shown in black. Artifacts denoted as UML documents (in upper-case) and processes as UML components (in italicized lower-case). Dependency, including "uses" and "observes", denoted with dashed lines, with arrows pointing towards the used or observed artifact; creation denoted with solid cirlce-terminated arrows, pointing towards the created artifact. Acronyms as in the text.

Given a conversation, shown in the left of the figure, a human may transcribe the sequence to produce a reference, or "oracle" state sequence of vocal activity, for each participant. This is often performed by transcribing the spoken words and any accompanying human noises, and then either timing their start- and end- points with a stopwatch, employing some speech analysis software, or running a speech recognizer in forced-alignment mode. Implicit time alignment across participant streams yields a vocal interaction chronogram $\mathbf{Q}$ of the conversation. However, activity types may also be directly encoded by observers without reference to linguistic content.

Alternately, the *detection* of vocal activity can be attempted automatically. Historically, this has been achieved using an energy threshold, which implements a simple model of single-participant vocal activity on a participant-attributed channel. The functional role of this model is denoted with "A" in Figure 5.2. Examples of this type of model are the frame detector of [24] and the AVTA device of [104]. Since frames are treated independently in these detectors, their output can be noisy. A post-processing step which mandates minimum duration constraints for each participant stream independently, such as the conversion of spurts and gaps into perceptually valid talkspurts and pauses in [24], is an example

of this kind of model, denoted "B" in the figure. The joint statistical modeling of frame acoustics and interval duration was not explored for detectors until [1].

Modeling of most interest in this thesis is identified by the letters "C", "D", and "E" in Figure 5.2. Beginning either with the oracle state sequence or the post-processed automatically detected state sequence, "C" denotes models whose parameters describe the statistical distribution of *features* computed from the chronogram. These may include durations of single- or multiple- participant phenomena; an example of the former is the duration of a talkspurt (as in [24]), while that of the latter is the duration of an interval of mutual silence. Other examples include total vocalizing time per participant, number of talkspurts per participant, number and average duration of intervals of multi-participant overlap, etc.

In contrast to these, models of type "D" attempt to describe the stochastic process which *generates* chronograms. Rather than describing emergent patterns, these models encode the instantaneous transition probabilities inferrable from an observed state sequence. They yield a prior for unobserved state sequences, and may be used either to guide automatic detection, to compare the likelihood of alternate detected sequences, or to synthesize novel state sequences.

Finally, the class of models denoted "E" in Figure 5.2 defines hyper-models of model parameters. Given models of type "D" inferred for different conversations, which differ in group, task, mood, participant number, participant type, presence of specific participants, etc., the parameters — as features — may allow for inference of these characteristics in unseen conversations.

5.2 Important Use Cases in Social Science, Engineering, and Business Practice

Content-free or text-independent modeling of interaction chronograms or their aspects has several important applications, for social science, applied science, business practice, and the public at large. A selection is shown in Figure 5.3.

The figure depicts 3 main pursuits in the social sciences, in white hexagons, whose central theme is an understanding of how conversations, and successful conversations in particular, are achieved. From left to right, the first is improvement in human-human interaction, where success is defined as the accomplishment of a well-defined goal. Specific conversational behaviors can be tested for correlation with success rate, and those deemed most successful can be stressed in conversant training. An application of this (shown in a white circle), across and within organizations, is the development of collaborative or negotiating skills; studies show for example that laughter can be strategically deployed by team leaders to improve group task performance [118].

The second application of modeling vocal interaction is the automatic diagnosis of personality; again, specific text-independent conversational behaviors can be tested for correlation with clinically obtained assessments. Objective measurement of personality traits has been studied to improve hiring practices [45], as depicted in the figure. However, such measurement also promises to aid in the diagnosis of personality disorders and in the management of recovery. Finally, text-independent measures may also help in diagnosing and resolving disfunctional dyadic relationships.

Figure 5.3 also contains technological use cases, in black hexagons. The first, from left to right, is the improvement of human-machine dialogue. At the current time, spoken dialogue systems are already deployed in a large number of call centers, to perform relatively simple end-to-end tasks otherwise necessitating human operator attention. As the interactive behaviors of such systems improve, presumably by simulating human behavior [80], they may replace the need for operator attention in increasingly sensitive sectors. Automatic dialogue systems may also benefit organization-internal practices, by providing a spoken interface to information resources.

A second technological use case is the automatic indexing of human-human interaction. Organizations whose practices include significant internal collaboration may wish to maintain historical records of interactions. To be useful in a time-efficient manner, such records need to be browsable and searchable. This requires that their segmentation, transcription, and structural inference be automated. Models of vocal interaction have a significant role to play, particularly in correctly attributing vocal activity to specific participants. An organization's ability to digitally access its history is likely to lead to improved communication and to facilitate the desynchronization of human effort.

Finally, the modeling of vocal interaction holds one key to the improvement of channel usage in telepresent communication. Bandwidth is only necessary during vocal production, and not during pauses in activity; the prediction of necessary bandwidth calls for a model of vocal activity sequences from all parties. Telephony, both analog and digital, has greatly simplified our lives in the past century, enabling professional and personal contact with others despite geographic distance.

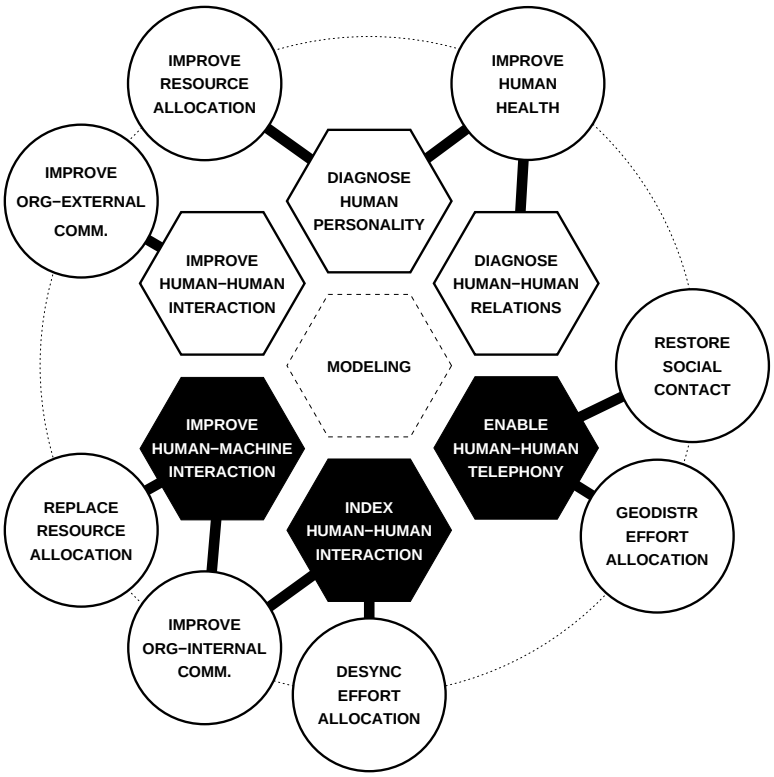


Figure 5.3: Use cases in the social sciences (white hexagons), the applied sciences (black hexagons), in business practice and society at large (white circles). Links, where some dependency can be thought to exist, as discussed in the text.

5.3 Modeling Emergent Chronogram Features

5.3.1 Towards Digital Telephony, 1930s-

Likely the earliest quantitative consideration of the chronemics of conversation is provided by [174], which defined the "time pattern of [dyadic] conversation" in terms of three temporal elements observable in an oscillographic record, or oscillogram[1] — the "pictorial record of the conversational interchanges":

talkspurt "speech by one party, including his pauses, which is preceded and followed, with or without intervening pauses, by speech from the other party perceptible to the one producing the talkspurt." ([174], p. 282)

resumption time "the length of the pause intervening between two periods of speech within a talkspurt" ([174], p. 282)

response time "the length of the interval between the beginning of a pause as heard by the listener and the beginning of his reply. It may be positive or negative." ([174], p. 282)

A consequence of the above definition of "talkspurts" is that they contain intra-talkspurt silence. [174] also defined *double talking* as the condition in which "one party is speaking and at the same time hears speech from the other" ([174], p. 282). In current practice, the preferred term for this phenomenon is *overlap*[2].

The initial motivation for [174] was to study toll circuit lockout [97]; its characterization of the durations of talk and silence was extended to several more two-party phenomena in [24, 25]. There, the acoustic units

[1] This usage has not become popular.

[2] It should be noted that there is a distinction between "two parties speaking simultaneously" and "one party speaking while hearing another", from a telecommunications perspective, because of the potential for transmission delay.

spurt "an unbroken sequence of on intervals", ("5-msec intervals during which the speech energy exceeds the threshold at some time") ([24], p.3)

gap "an unbroken sequence of off intervals", (5-msec intervals during which "the energy remains below the theshold during the entire interval") ([24], p.3)

were mapped to the perceptual units

talkspurt "a time period which is judged by a listener to contain a sequence of speech sounds unbroken by a pause" ([24], p.4)

pause "a time period which is judged by a listener to be a period of nontalking, other than one caused by a stop consonant, a slight hesitation, or a short breath" ([24], p.4)

The purpose of defining and describing these terms was to: (1) assess the impact of parameters guiding a speech/non-speech detector; (2) guide the design of voice-operated devices in telephony circuits; (3) study conversational behavior over particular circuit types. The study of emergent two-party patterns was then extended to more specific types of phenomena in [26], including "double talk", "mutual silence", "alternation silence", "pause in isolation", "solitary talkspurt", "interruption", "speech after interruption", and "speech before interruption". Crucially, [26] introduced a model of two-party conversation as a stochastic Markov process, rather than merely proposing a model of features of conversation, providing a precedent for subsequent work in telecommunications; the models proposed in that work are discussed in Section 5.4.

5.3.2 Interaction Chronography, 1930s-

The chronemics of interaction were perhaps first applied to the development of an "index of personality" in [212], for children. Human observers coded overt child behavior into categories such as PhysicalActivity, Talking, Laughing, Crying, and PhysicalContact (among others) at a granularity of 5 minutes with paper and pencil. The methodology was extended to two-party, adult conversation in [43], which replaced the cumbersomeness of paper and pencil methods by introducing a typewriter-like device with two keys, one for each party. Keys were pressed when "a person started to act or to respond, whether by talking, smiling, or nodding his head, and [released] when the action ended" [45]. The keys drew lines on moving tape; [44] describes a collection of measurement techniques for analyzing the resulting artifacts.

The term "interaction chronograph" appears to have been coined in [45], and denotes an evolution of the typewriter-like device into an instrument which makes the requisite measurements automatically given the tape, but still requires human observers to man the keys. The device's output includes measures of tempo, activity, inter-subject adjustment, initiative, dominance, and inter-subject synchronization. These measures are reported to be stable across multiple conversations with the same people [196]. Interaction chronography has been successfully applied in the hiring of department store salespeople [47]. It has also been used in psychiatric interviews [46, 81, 196, 168, 159].

5.3.3 Interaction Process Analysis, 1950-

A much richer set of categories for the actions executed by participants to conversation has been proposed by Bales [7], as part of his Interaction Process Analysis (IPA) coding system. The categories include 6 "task area" categories: neutral GivesSuggestion, GivesOpinion, GivesOrientation, AsksForOrientation, AsksForOpinion, AsksForSuggestion; 3 positive "socio-emotional area" categories: ShowsSolidarity, ShowsTensionRelease, Agrees; and 3 negative "socio-emotional area" categories: Disagrees, ShowsTension, ShowsAntagonism.

The IPA coding system, while successful in the tasks for which it was envisioned, suffers from two main limitations with respect to the description of vocal interaction. First, it cannot reliably code actions at a time granularity which is commensurate with conversational behavior; its creator reports that actions can be coded at a rate of approximately one per minute. Second, it cannot code the behavior of more than one individual simultaneously.

5.3.4 Conversation Analysis, late 1960s-

An important source of inspiration for this thesis is the body of observations known as *conversation analysis*. Founded in the late 1960s by Harvey Sacks and his collegues, it is a direct application of ethnomethodological analysis [67] to talk

as it occurs in practice. Conversation analysis departed from mainstream linguistic thought in attempting to account, empirically, for what participants to conversation achieve by talking. It has since become an important thread of sociolinguistics and linguistic pragmatics [158]. Conversation analysis has identified conversation as one possible type among speech exchange systems (cf. [194]), and has made several seminal contributions.

A main contribution of conversation analysis is the algorithmic description of how speaker change recurs in multiparticipant conversation [194], shown here in abridged form as Algorithm 1. The description assumes that *a speaker turn* consists of at least one *turn-constructional unit* — a word, phrase, sentence — whose endpoint, known as a *transition-relevance place* (TRP), is recognizable and predictable in advance by both the speaker and his/her interlocutors. The point of turn-taking appears to be the avoidance of overlap (as well as mutual silence) [197], which, when it occurs, must be resolved [197]. Overlap in naturally occurring conversation is frequent, but — owing to the efficiency with which it is normally resolved — quite short-lived. Algorithm 1 is *normative*, and its formulation is meant to aid in identifying and describing departure from the norm.

Algorithm 1 Effectuate recurrence of speaker change (after [194])

Require: $K \geq 2$
Require: $t \in \Re$, $\delta t > 0$
Require: $C \in \{1, \cdots, K\}$
loop
  if t is a TRP **then**
    if C has selected next-speaker $N \in \{1, \cdots, K\}, N \neq C$ **then**
      $C \leftarrow N$
    else
      for all $k \in \{1, \cdots, K\}, k \neq C$ **do**
        if k self-selects at $t + \Delta t_k$ **then**
          $\mathbf{N}[k] \leftarrow \Delta t_k$
        else
          $\mathbf{N}[k] \leftarrow +\infty$
        end if
      end for
      if $\min_{k \neq C} \mathbf{N}[k] < +\infty$ **then**
        $C \leftarrow \arg\min_{k \neq C} \mathbf{N}[k]$
      else
        $C \leftarrow C$
      end if
    end if
  end if
  $t \leftarrow t + \delta t$
end loop

A second important contribution of conversation analysis is the *adjacency pair*. This construct groups contributions, a part A and a part B, from two consecutively speaking participants. Examples include question-answer and greeting-greeting sequences. Because candidate part B types are severely constrained by the part A type, adjacency pairs can play an important role in forming prior expectation models of chronogram state sequences. Adjacency pairs are not directly treated in this thesis, as type inference nominally requires lexical access and is therefore not text-independent.

Finally, conversation analysis has provided a uniform means of treating conversational laughter [115, 110, 112, 111], which is often ignored in other linguistic accounts of conversation. As for speech, laughter has been found to frequently exhibit a normative sequencing, including initiating shared laughter and extending shared laughter. An informative account of laughter in interaction can be found in [76].

5.4 Modeling the Chronogram as a Process

5.4.1 One-Party Scenarios, 1964-

Markov modeling of vocal interaction chronograms was pioneered by Jaffe, Cassotta & Feldstein in [103], where it was shown that, in monologue, speech and non-speech intervals follow an exponential distribution [88]. The data used was collected from 25 subjects, asked to speak spontaneously about a topic. The authors argued that this resulted in a useful sample of unconstrained speech, to which they successfully fit the finite state model shown in Figure 5.4 (with a sampling frame step of 300 ms, or 200 frames per minute). The transition probabilities in the model were assumed to be 1st-order Markovian, and a 2nd-order model predicted monologue no better than did the 1st-order model.

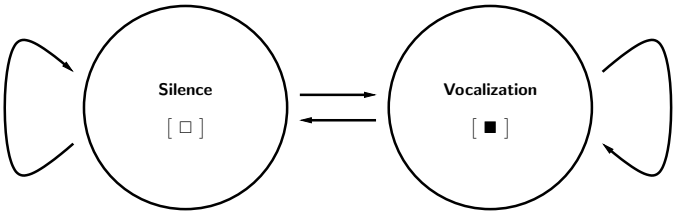


Figure 5.4: A 2-state fully-connected Markov model, proposed by Jaffe, Cassotta & Feldstein in 1964 [103] to account for the temporal distribution of speech in monologue; the frame step is 300 ms.

5.4.2 Two-Party Scenarios, 1967-

Social psychology also provided the first extension of the above model to dialogue [106], as a means of studying psychiatric interviews [105]. Arguing that the durations of concatenated two-participant states of dialogue are also exponentially distributed, Jaffe, Feldstein & Cassotta proposed the model shown in Figure 5.5. Its transition probabilities were assumed to be 1st-order Markovian, at a frame step of 300 ms. Evidence from 200 experimental interviews demonstrated that transition probabilities were at least approximately stationary.

As Figure 5.5 shows, the 4 states are fully connected, allowing the model to find itself in any state within one increment of 300 ms. [106] discussed not only the joint transition probability model, of $2^2 \cdot \left(2^2 - 1\right) = 12$ free parameters, but also showed that assuming participants to behave independently of one another, with dependence on the joint two-participant state from the previous instant, yields a model which predicts empirical data almost as well as the joint model does. The conditionally independent model, called the ""independent decision" hypothesis"[3], consists of only $2 \cdot 2^2 = 8$ free parameters [102].

In 1969, Brady considered a simplification of the topology in Figure 5.5, in which only one participant at a time could change state [26]. This appears justified because the model of [26], shown in Figure 5.6, uses a frame step of 5 ms. The transition probabilities in the investigated conditionally independent transition probability model are not truly stationary; for the first 15 ms of speech and the first 200 ms of non-speech, the probability of leaving the just-entered state is zero.

[26] found that this 4-state model does not adequately account for empirical evidence, particularly in the vicinity of joint vocalization. It was postulated that the fit could be improved if transition probabilities were conditioned not only on the previous state, but also on the identity of the distinct penultimate state. This was achieved by splitting all four states in Figure 5.6 into two, as shown in Figure 5.7.

[3] In subsequent work, various authors have used a variety of terms to denote the distinction between conditionally dependent and conditionally independent transition probability models. In [26], Brady referred to the conditionally dependent model as the "multi-port model" and to the conditionally independent model as the "single-port model". In [104], Jaffe & Feldstein referred to the conditionally dependent model as the "single source model" and to the conditionally independent model as the "separate source model". It should be noted that the word "single" occurs in descriptions of both model types, across different authors.

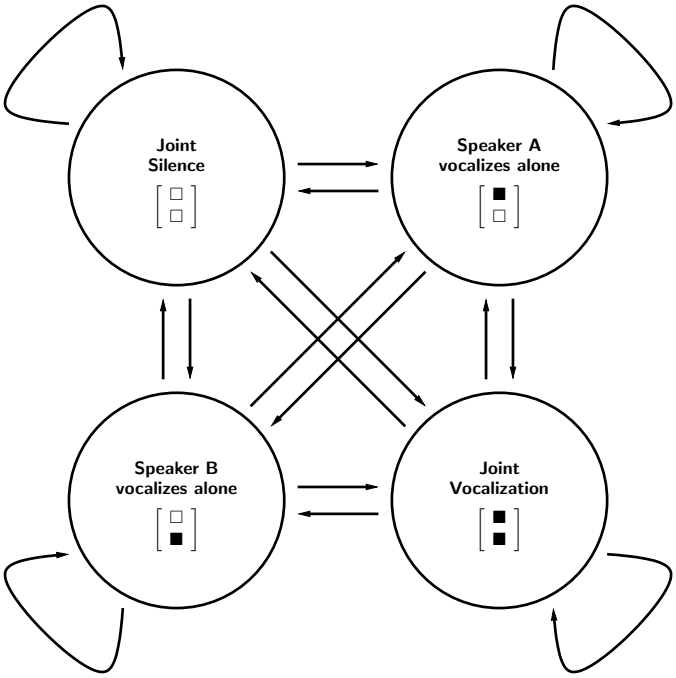


Figure 5.5: A 4-state fully-connected Markov model, proposed by Jaffe, Feldstein & Cassotta in 1967 [106] to account for the temporal distribution of speech in dialogue; the frame step is 300 ms.

Brady then argued that it is sufficient to split only two of the four states [26], leading to the model of Figure 5.8. This is interesting, because at approximately the same time, Jaffe & Feldstein proposed a nearly identical 6-state model [104]. They posited the following "requirements for a comprehensive model of the rhythm of unstructured dialogue":

1. Separate vocalization parameters for the two speakers are required.
2. Separate silence parameters for the two speakers are required.
3. Separate simultaneous speech parameters for the two speakers are required.
4. Each of these six parameters must describe a separate exponential distribution of "waiting times" in the respective state.
5. The model must distinguish which speaker has the floor at any given time (implied in 2 and 3).
6. The frequency of speaker switches and mean duration of "floor time" must be derivable from the model.
7. In addition to making these distinctions of the descriptive classification which were found to be essential for unstructured dialogue, the model should make intelligible the findings of the previous Markov models.
8. Parameters must be derivable from the model which can be combined to predict the patterns of reconstituted dyads. In short, the model must provide a theory of *rhythmic interaction* which generates every essential feature of the temporal sequence.

Their proposed model is shown in Figure 5.9 (based on the transition matrix on page 101 in [104]).

The main difference between Brady's and Jaffe's & Feldstein's 6-state models is the frame step: Brady's model increments at 5 ms steps, while Jaffe's & Feldstein's model increments at 300 ms. This difference allows Brady to license at most one participant to switch state at a time.

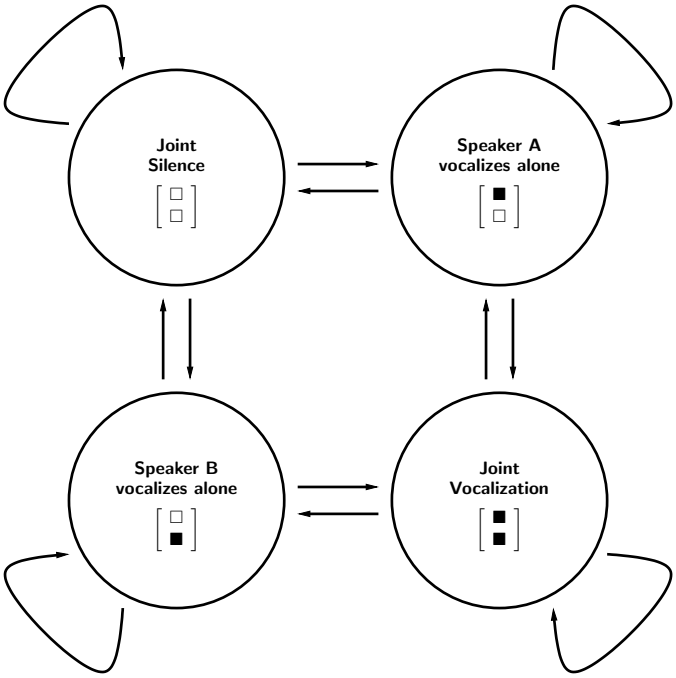


Figure 5.6: A 4-state Markov model, proposed by Brady in 1969 [26] to account for the temporal distribution of speech and non-speech in dialogue; the frame step is 5 ms.

Jaffe & Feldstein justified their 6-state rather than their original 4-state model (Figure 5.5) by the observation that it matters who has the floor. Figure 5.9 demarcates this aspect with a dashed line.

[104] has become somewhat standard in speech/non-speech modeling in the social psychology community. Its main intent has been the automation of interaction chronography (e.g., [33]) as proposed by Chapple. This was achieved by: (1) automatically detecting speech in consecutive frames, assumed to be indendent; (2) inferring the model parameters for both participants; and (3) comparing or otherwise assessing inferred parameter values with known values. It has been argued that, across conversations, model parameters for individuals are stable [40]; that they correlate with self-attributed personality characeristics [62]; that they may inform the formation of interpersonal impression [61]; and that they are subject to mutual adaptation [32]. There is also some evidence that modeling vocal interaction can be exploited in lexical prediction [222].

5.4.3 Multi-Party Scenarios, 1987-

Modeling vocal interaction has been attempted for triads, but only rarely [166].

Work most similar to that of [104], for conversations of arbitrary numbers of participants, is the "Grouptalk" model proposed by [52]. In addition to the three states per participant found in Figure 5.9, it introduces three more states, corresponding to "group turns" rather than "individual turns". These states, together with one set of "individual turn" states, is shown in Figure 5.10. The "group turn" states account for any overlap observed in the conversation, of any degree. However, because of this, it is not known which participants are vocalizing, when these states are visited. Unlike the models of Brady and Jaffe & Feldstein, the model of [52, 51] cannot be used directly to synthesize a complete conversation. However, parameters, as in the two-party models, are correlated with overt differences among groups; [50] showed that vocal activity patterns are different in conversations among males and among females.

Although multi-party modeling has received significantly less attention than two-party modeling, there is some evidence of resurging interest, particularly in teleconferencing [201, 211], dialogue systems [22, 23], and multi-player gaming [179]. These works fall short of proposing stochastic models of conversation, but elucidate an emergent need for better predictive

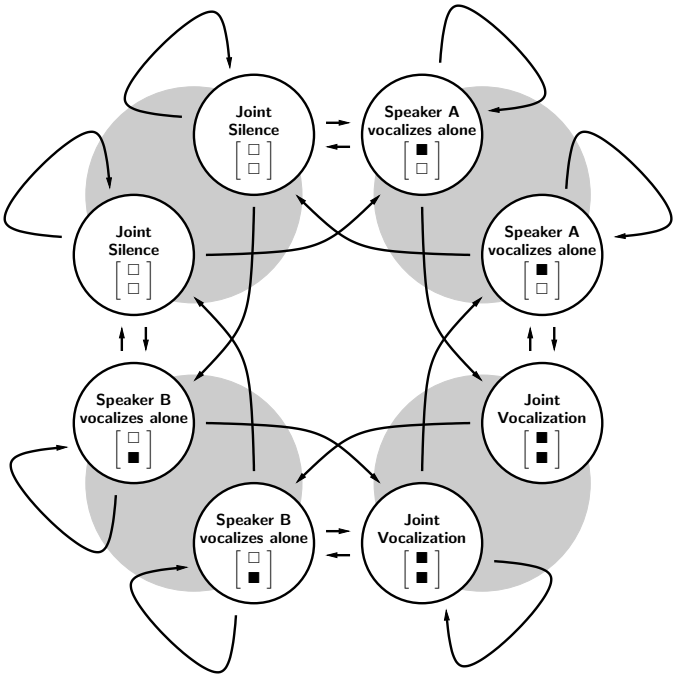


Figure 5.7: An 8-state Markov model, proposed by Brady in 1969 [26] to account for the temporal distribution of speech and non-speech in dialogue; the frame step is 5 ms. All states in Figure 5.6 (depicted in gray) are replicated for reference.

control.

5.5 Model-Driven Synthesis of Chronograms

Modeling chronograms as a process makes it possible to sample novel sequences from models. Sampling has been used frequently to verify model correctness, by comparing emergent phenomena in synthesized chronograms to those from human-human chronograms (e.g. duration of overlap at speaker change). Synthesis (and, by extension, prediction) are currently being investigated for spoken dialogue systems [100, 188].

In multi-party conversation, the absence of viable model architectures has led to hybrid models for synthesis, with a learnable component but also with a deterministic component which must be designed by hand. Examples of this approach are [206] and [178]. In the latter case, synthetic agents visit the nodes of a manually constructed decision tree, at the leaves taking a particular action with a potentially automatically inferred probability. Hybrid models also seem to be the norm in the design of dialogue systems which may face multiple interlocutors [22, 23].

5.6 Summary

At the time that work on this thesis was initiated, no general, comprehensive density model over conversational sequences of speech/non-speech activity was available. Mathematical models had been exhaustively studied for two-party settings, and a significant amount of relevant dialogue research had been accumulated. However, for general multi-party settings, those models which existed cannot be deployed with the same facility, particularly when the group size in the conversations to which the model is to be applied is not known during model training. The models either capture emergent features of chronograms, and are thereby not suitable to the synthesis of novel sequences, or they can achieve the latter but require extensive design by hand. In the first case, models can be inferred from evidence but do not encode the probabilities of

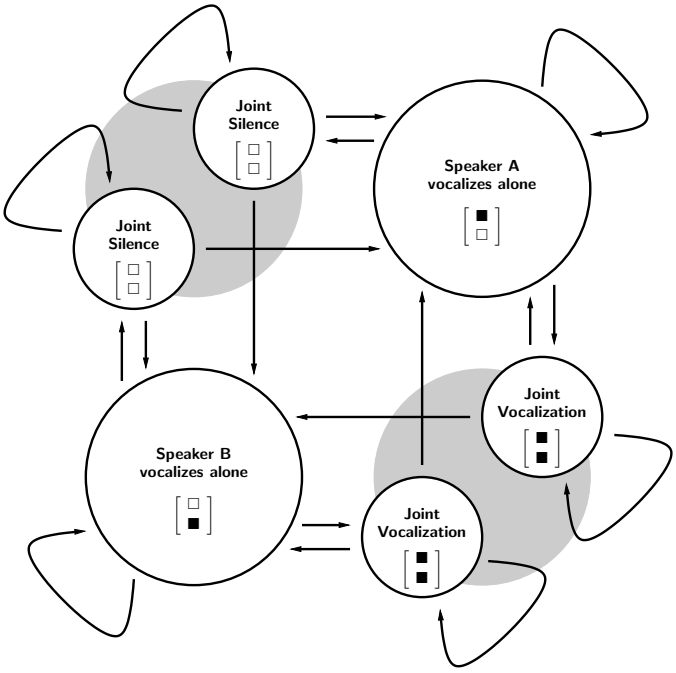


Figure 5.8: A 6-state Markov model, proposed by Brady in 1969 [26] to account for the temporal distribution of speech and non-speech in dialogue; the frame step is 5 ms. Only two of the states in Figure 5.6 (depicted in gray) are duplicated.

candidate future events conditioned on the recent past. In the second case, models cannot be fully automatically inferred from evidence.

This thesis is concerned with bridging that impasse. The intent is to propose tractable density models over sequences of conversational vocal activity, of arbitrary length and for arbitrary conversational group sizes. A joint density model, factorable in time as a stochastic process, can be used to compare conversation, to provide prior knowledge during vocal activity detection (by making available the likelihood of competing hypotheses), and to synthesize future outcomes (by sampling from the conditional distribution of participant activity during the next, and each conditionally subsequent, instant).

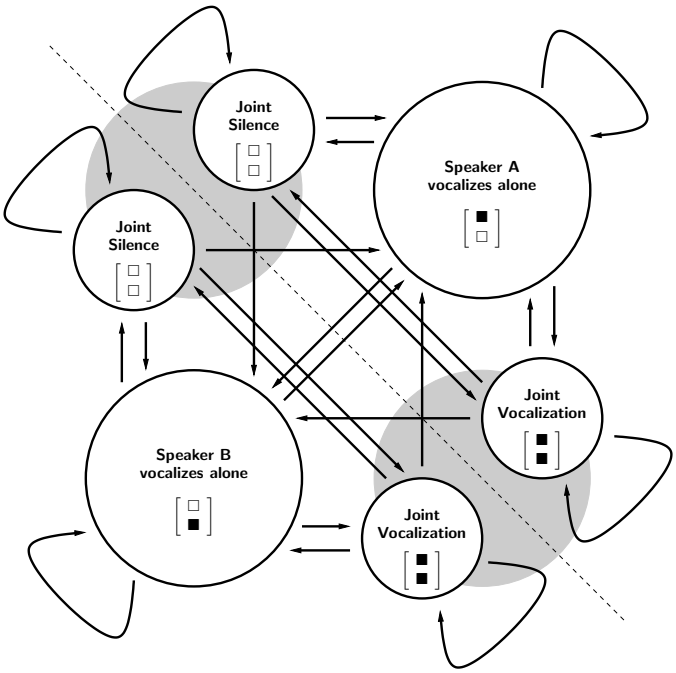


Figure 5.9: A 6-state Markov model, proposed by Jaffe & Feldstein in 1970 [104] to account for the temporal distribution of speech and non-speech in dialogue; the frame step is 300 ms. The dashed line separates instants during which speaker A holds the floor (upper right) from those during which speaker B does (lower left).

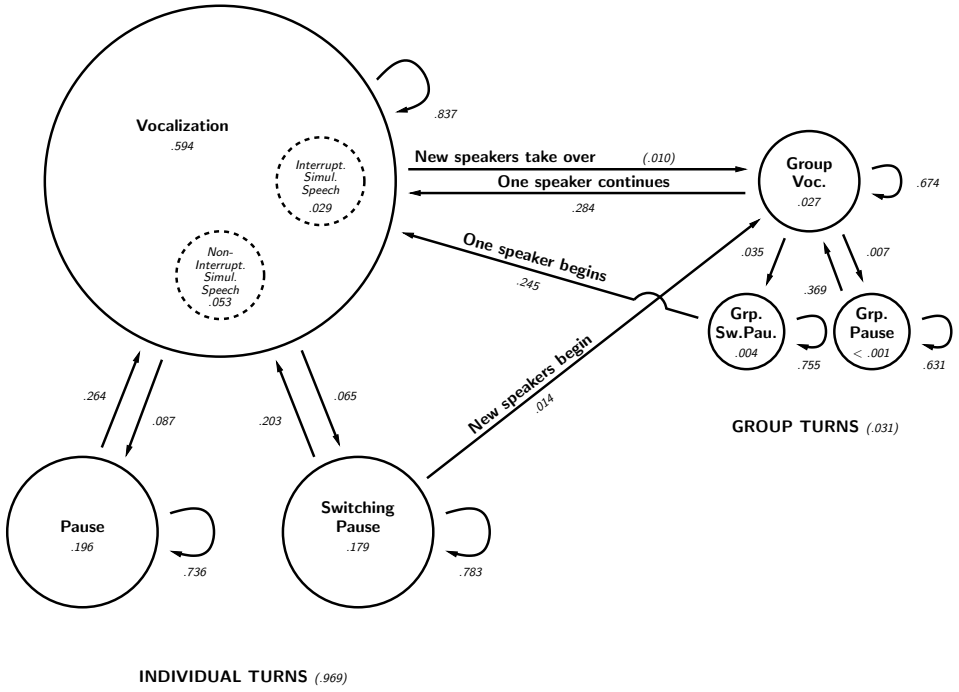


Figure 5.10: Diagram of a subset of states in the Grouptalk model proposed by Dabbs, Ruback & Evans in 1987 (rendition based on Figure 10.1 in [52]). "Solid circles constitute a set of mutually exclusive and exhaustive states. Dashed circles are additional events that may occur during the vocalizing state. Numbers inside circles are simple probabilities of the occurrence of each state or event. Numbers beside arrows are transitional probabilities of each state occurring at the next point in time. The data are summarized from 60 five-person group conversations, lasting an average of 15–20 minutes each, in which the conversational state was coded ever[y] quarter-second." ([52], p. 505).

Part II

FACETS OF A COMPUTATIONAL THEORY OF THE DISTRIBUTION OF VOCAL ACTIVITY IN CONVERSATION

Chapter 6

Non-Parametric State-Space Multi-Participant Models*

6.1 Introduction

An important artifact in the study of multi-party conversation is a model which can automatically assign a likelihood to any state sequence. As argued in Chapter 2, such functionality would make it possible to compare true state sequences of arbitrary conversations, and to compare hypothesized state sequence alternatives for any single conversation.

This chapter describes just such a model. A descriptor of the conversation as a whole, at any instant, is conceived of as the composition of all participants' descriptors, allowing conversations to be regarded as a vector-valued Markov process. This in turn makes it possible to decompose the likelihood of observing the entire conversation as a product of the conditional likelihoods of observing each successive instant given the immediate, truncated history. These conditional likelihoods, or transition probabilities, together comprise the proposed model.

Since transition probabilities describe entry into and egress out of discrete states, the techniques discussed in this chapter are a "*state-space*" modeling approach, in contrast to "feature-space" approaches (of which an example is described in Chapter 8). The transition probabilities are inferred from labeled observations under no assumption, or parametrization, of their form, rendering the models properly *non-parametric*[1].

6.2 Symbols and Definitions

The *number* of participants to a conversation $\mathcal{C}$ is denoted K; throughout this thesis, K is assumed to be known *a priori* and fixed, for the entire duration of any given $\mathcal{C}$. K may of course be different for different conversations.

6.2.1 Participant Behavior

Each participant k, $1 \leq k \leq K$, is at every continuous-time instant $\tilde{t} \in \mathbb{R}$ engaged in some behavior $\psi_k\left(\tilde{t}\right)$. The latter is referred to as the *continuous-time single-participant behavior* of participant k at instant $\tilde{t}$, and is drawn from the discrete-element set

$$\mathbb{B} \equiv \{B_0, B_1, \cdots, B_{N_B-1}\} , \tag{6.1}$$

where $N_B \equiv |\mathbb{B}|$ is the number of elements in $\mathbb{B}$. Except that its elements are intended to be exhaustive, the precise definition of $\mathbb{B}$ in Equation 6.1 depends on the task under consideration, for example:

- speech ($\mathcal{S}$) versus non-speech ($\neg\mathcal{S}$);

*The work in this chapter was conducted in collaboration with Mari Ostendorf and Tanja Schultz.

[1]It should be noted that the term "non-parametric" possesses a large number of meanings in the statistics literature.

- laughter ($\mathcal{L}$) versus non-laughter ($\neg\mathcal{L}$);
- $\mathcal{S}$ versus $\mathcal{L}$ versus silence ($\mathcal{N}$, a shorthand for $\neg\left(\mathcal{S}\cup\mathcal{L}\right)$);
- $\mathcal{S}$ versus voiced laughter ($\mathcal{L}_V$) versus unvoiced laughter ($\mathcal{L}_U$) versus $\mathcal{N}$;
- backchannels ($\mathcal{S}_{\mathtt{b}}$) versus acknowledgments ($\mathcal{S}_{\mathtt{bk}}$) versus accepts ($\mathcal{S}_{\mathtt{aa}}$) versus floor grabbers ($\mathcal{S}_{\mathtt{fg}}$) versus floor holders ($\mathcal{S}_{\mathtt{fh}}$) versus holds ($\mathcal{S}_{\mathtt{h}}$) versus questions ($\mathcal{S}_{\mathtt{q}}$) versus statements ($\mathcal{S}_{\mathtt{s}}$) versus $\neg\mathcal{S}$; or
- $\mathcal{S}_{\mathtt{b}}$ versus $\mathcal{S}_{\mathtt{bk}}$ versus $\mathcal{S}_{\mathtt{aa}}$ versus $\mathcal{S}_{\mathtt{fg}}$ versus $\mathcal{S}_{\mathtt{fh}}$ versus $\mathcal{S}_{\mathtt{h}}$ versus humor-bearing speech $\mathcal{S}_{\mathtt{j}}$ versus non-humor-bearing questions ($\mathcal{S}_{\mathtt{q}}$) versus non-humor-bearing statements ($\mathcal{S}_{\mathtt{s}}$) versus $\neg\mathcal{S}$.

Tasks involving these distinctions are treated in later parts of this thesis; the current discussion is intended to be general in nature.

6.2.2 Discretizing Time

This thesis studies conversational dynamics in discrete time. ψ_k, for each participant k, is sampled at a fixed frame step Δ_s, yielding the *discrete-time single-participant behavior* $\psi_{k,t}$, $t\in\mathbb{N}$ and $1{\leq}t{\leq}T$. T is the total number of Δ_s-separated frames in $\mathcal{C}$.

The instants at which $\psi_k\left(\tilde{t}\right)$ is sampled are the same for all K participants, and are given by

$$\tilde{t} \;=\; \tilde{t}_0 + t\cdot\Delta_s\;, \tag{6.2}$$

with $\tilde{t}_0$ representing the continuous-time start time of $\mathcal{C}$.

Although $\psi_{k,t}$ can be simply assigned the value of $\psi_k\left(\tilde{t}\right)$, the temporal support of a given frame and any given participant, for non-vanishing Δ_s, may contain multiple elements of $\mathbb{B}$. To avoid spurious phenomena, $\psi_{k,t}$ is instead assigned the value

$$\psi_{k,t} \;=\; B_{m^*_{k,t}}\;,\;\text{ with} \tag{6.3}$$

$$m^*_{k,t} \;=\; \arg\max_m \int_{\tilde{t}_0+t\cdot\Delta_s-\Delta_w/2}^{\tilde{t}_0+t\cdot\Delta_s+\Delta_w/2} \delta\left(B_m,\psi_k\left(\tilde{t}\right)\right)\,d\tilde{t}\;. \tag{6.4}$$

Here, $\delta\left(\cdot\right)$ is the Kronecker delta and Δ_w is the width of the frame centered on $\tilde{t}_0+t\cdot\Delta_s$. Equations 6.3 and 6.4 identify that phenomenon in $\mathbb{B}$ which characterizes participant k for the largest proportion of the frame[2] starting at $\tilde{t}_0+t\cdot\Delta_s-\Delta_w/2$ and ending at $\tilde{t}_0+t\cdot\Delta_s+\Delta_w/2$.

6.2.3 Modeling Participant Trajectories

The evolution of $\psi_{k,t}$ may be serviceably modeled by a discrete-time model exploring a discrete-element state space $\mathbb{S}$. In the simplest possible case, each state in $\mathbb{S}$ corresponds to exactly one behavior in $\mathbb{B}$, i.e., $\mathbb{S}\equiv\mathbb{B}$ and the number of states $N_S\equiv|\mathbb{S}|=|\mathbb{B}|\equiv N_B$. Two examples of such models, for two different values of N_B, are depicted in panels (a) and (b) of Figure 6.1.

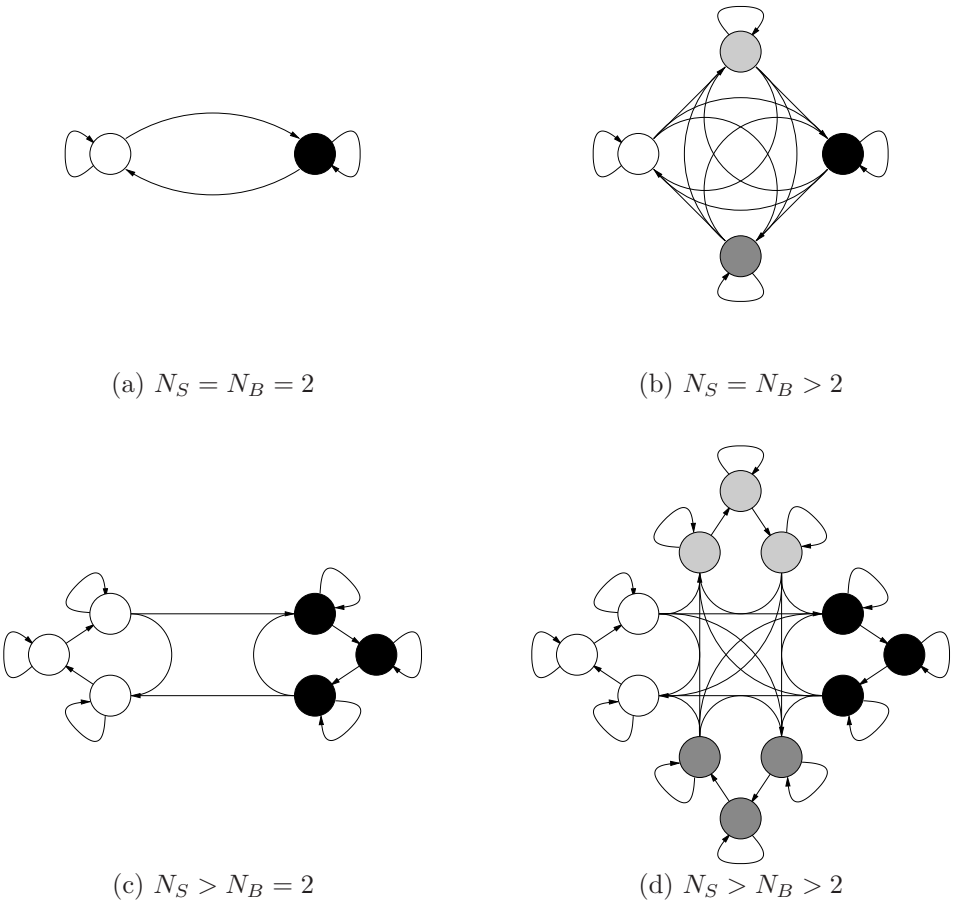
However, in the general case, there may be reason for $\mathbb{S}$ to contain many more elements than $\mathbb{B}$ does. One common reason is the desire to impose lower bounds on the durations for which participants occupy certain behaviors $\{B_m\}$ for one or more m. This is achieved by initially allowing $\mathbb{S}=\mathbb{B}$ but then by replicating the states which correspond to $\{B_m\}$. Examples of the resulting single-participant topologies, for occupation times at least $3\cdot\Delta_s$ rather than at least Δ_s, are shown in panels (c) and (d) of Figure 6.1.

Because of this general $\mathbb{S}\neq\mathbb{B}$ case, it is convenient to independently re-enumerate the states of $\mathbb{S}$,

$$\mathbb{S} \;\equiv\; \{S_0,S_1,\cdots,S_{N_S-1}\}\;, \tag{6.5}$$

The *discrete-time single-participant state* of participant k, at instant t, is henceforth denoted $q_{k,t}\in\mathbb{S}$.

[2]This is a form of low-pass filtering which ensures that the continous-time single-participant behavior $\psi_k\left(\tilde{t}\right)$ is bandlimited to below the implicit Nyquist frequency imposed by the selection of Δ_s, prior to sampling.



(a) $N_S = N_B = 2$ (b) $N_S = N_B > 2$

(c) $N_S > N_B = 2$ (d) $N_S > N_B > 2$

Figure 6.1: Example single-participant state topologies ($K = 1$), with $N_B = 2$ in the two panels on the left and $N_B = 4$ in the two panels on the right. The figure also contrasts fully-connected, or ergodic, topologies with non-ergodic, minimum-duration-enforcing topologies, in the upper and lower panels, respectively.

Transitions of single-participant state, from $q_{k,t-1} = S_m$ to $q_{k,t} = S_n$, are denoted (S_m, S_n). When a model licenses all conceivable transitions (S_m, S_n), $0 \leq m < N_S$, $0 \leq n < N_S$, as in panels (a) and (b) of Figure 6.1, the model is referred to as "fully-connected" or, somewhat less correctly but more commonly, "ergodic". In the contrary case, as in panels (c) and (d), the model will be referred to as "non-ergodic".

6.2.4 The Multi-Participant Hypercube

Multi-participant representations of the conversation $\mathcal{C}$ as a whole are obtained by concatenating, for any instant t, the single-participant representations sampled synchronously at that t, of all K participants.

The *discrete-time multi-participant behavior* $\boldsymbol{\psi}_t$ at instant t is a K-length column vector whose value is drawn from the Cartesian product

$$\mathbb{B}^K \equiv \mathbb{B} \times \mathbb{B} \times \cdots \times \mathbb{B} \tag{6.6}$$

$$= \{\mathbf{B}_0, \mathbf{B}_1, \cdots, \mathbf{B}_{N_B^K - 1}\} . \tag{6.7}$$

Equation 6.7 merely re-enumerates the elements of $\mathbb{B}^K$ for convenience. The entire conversation $\mathcal{C}$ can be represented by the matrix

$$\boldsymbol{\Psi} \equiv [\, \boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \cdots, \boldsymbol{\psi}_T \,] \in \mathbb{B}^{K \times T}, \tag{6.8}$$

with $\boldsymbol{\psi}_t[k] = \psi_{k,t}$.

Similarly, the *discrete-time multi-participant state* $\mathbf{q}_t$ at instant t, of all K participants to conversation $\mathcal{C}$, is a K-length column vector whose values are drawn from

$$\mathbb{S}^K \equiv \mathbb{S} \times \mathbb{S} \times \cdots \times \mathbb{S} \tag{6.9}$$

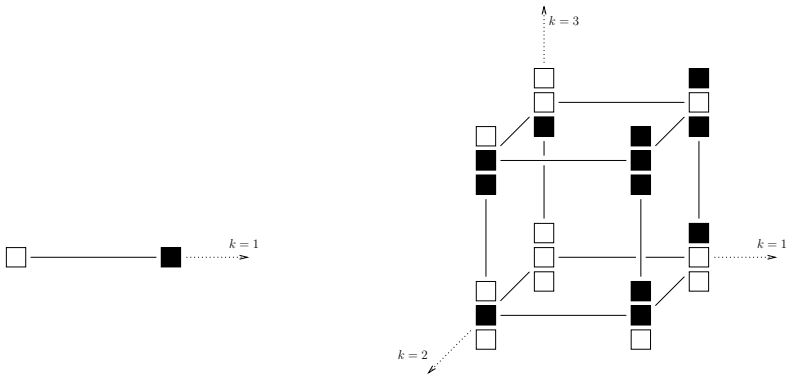
$$= \{\mathbf{S}_0, \mathbf{S}_1, \cdots, \mathbf{S}_{N_S^K - 1}\}. \tag{6.10}$$

The *discrete-time vocal interaction chronogram* is denoted as

$$\mathbf{Q} \equiv [\, \mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_T \,] \in \mathbb{S}^{K \times T} \tag{6.11}$$

with $\mathbf{q}_t[k] = q_{k,t}$.

As the multi-participant state $\mathbf{q}_t$ evolves in $\mathbb{S}^K$, each single-pariticipant state $\mathbf{q}_t[k]$ is seen to evolve in the projection of $\mathbb{S}^K$ onto the kth axis. This constrains $\mathbf{q}_t$ evolution to the vertices of a hypercube, each edge of which is a discrete axis of N_S ordinates. A graphical depiction of $\mathbb{S}^K$, with $\mathbb{S} = \{\square, \blacksquare\}$, is shown in Figure 6.2.



(a) K = 1 (b) K = 3

Figure 6.2: Ergodic state space topologies with $N_B = N_S = 2$, for the single-participant case ($K = 1$), and a multi-participant state space ($K = 3$) in panels (a) and (b), respectively. States are represented by a K-element vector with values drawn from $\{\square, \blacksquare\}^K$; transitions, connecting every state to every other state and including self-loops, are not shown for readability.

Transitions of multi-participant state, from $\mathbf{q}_{t-1} = \mathbf{S}_i$ to $\mathbf{q}_t = \mathbf{S}_j$, are denoted $(\mathbf{S}_i, \mathbf{S}_j)$. When single-participant states are ergodic in $\mathbb{S}$, the multi-participant state is ergodic in $\mathbb{S}^K$, i.e., all transitions $(\mathbf{S}_i, \mathbf{S}_j)$, $0 \leq i < N_S^K$, $0 \leq i < N_S^K$, are licensed. It is easy to conceive of the implicit multi-participant topology as consisting of connections among all pairs of hypercube vertices. However, when $\mathbf{q}_t[k]$ is not ergodic, $\mathbf{q}_t$ is also not ergodic. In such cases, $(\mathbf{S}_i, \mathbf{S}_j)$ is licensed over the state space $\mathbb{S}^K$ if and only if $(\mathbf{S}_i[k], \mathbf{S}_j[k])$ is licensed over the state space $\mathbb{S}$, for all $1 \leq k \leq K$.

6.2.5 A First-Order Markov Model Framework

A time-independent, first-order Markov model $\boldsymbol{\Theta}$ describing the evolution of an observable $\mathbf{q}_t$ is characterized by the following[3]:

[3]The proposed notation is an extension to vector-valued states of the notation in [187].

1. $N = N_S^K \in \mathbb{N}$, the number of states in the model. Since, for the purposes of this thesis, K is known *a priori* given any conversation $\mathcal{C}$, N is a known number requiring no estimation.

2. $\boldsymbol{\pi} = \{\pi_i\} \in \mathbb{P}^N$, the initial state probability distribution, where

$$\pi_i = P(\mathbf{q}_1 = \mathbf{S}_i) \ . \tag{6.12}$$

   As will be discussed subsequently, π plays a very negligible role in this thesis.

3. $\mathbf{A} = \{a_{ij}\} \in \mathbb{P}^{N \times N}$, the time-independent conditional state transition probability distribution, where

$$a_{i,j} = P(\mathbf{q}_t = \mathbf{S}_j | \mathbf{q}_{t-1} = \mathbf{S}_i, \boldsymbol{\Theta}) \ ; \tag{6.13}$$

   In contrast to N and $\boldsymbol{\pi}$, $\mathbf{A} = \{a_{i,j}\}$ will in general require estimation, for those transitions $(\mathbf{S}_i, \mathbf{S}_j)$ which the topology licenses (the remaining, unlicensed transitions have $a_{i,j} = 0$).

6.3 Direct Compositional Model Estimation

Estimation of $\{a_{i,j}\}$ in this thesis always assumes that labeled data exists. For this reason, expectation-maximization via the Forward-Backward algorithm is not used. Instead, maximum likelihood estimates are computed directly by forming the ratio of model state bigram counts to model state (unigram) counts,

$$a_{i,j} = \frac{\sum_{t=1}^{T} \delta_K(\mathbf{q}_{t-1}, \mathbf{S}_i)\, \delta_K(\mathbf{q}_t, \mathbf{S}_j)}{\sum_{t=1}^{T} \delta_K(\mathbf{q}_{t-1}, \mathbf{S}_i)} \ , \tag{6.14}$$

where $\delta_K(\cdot)$ is the Kronecker delta extended to vector arguments in a natural manner, namely

$$\delta_K(\mathbf{u}, \mathbf{v}) = \prod_{k=1}^{K} \delta(\mathbf{u}[k], \mathbf{v}[k]) \ . \tag{6.15}$$

Equation 6.14 requires that each frame $\mathbf{q}_t$ of the training data be labeled with a single (model-space) value drawn from $\mathbb{S}^K$.

When $\mathbb{S} = \mathbb{B}$, $\mathbb{S}^K = \mathbb{B}^K$ and the labels $\mathbf{q}_t = \boldsymbol{\psi}_t$ are simply the concatenation of each manually annotated and discretized $\psi_{k,t}$. In those cases for which $\mathbb{S} \neq \mathbb{B}$, however, model-space labels $\mathbf{q}_t$ are obtained via Viterbi forced-alignment, prior to parameter estimation using Equation 6.14.

6.3.1 Conditional Dependence Among Participants

The most general case occurs when participants must be assumed to behave in a manner which is conditionally dependent (CD). $\boldsymbol{\Theta}^{CD} = \left\{a_{i,j}^{CD}\right\}$, $0 \leq i < N_S^K$, $0 \leq j < N_S^K$, implements a joint, time-independent model of vocal interaction. Given a labeled corpus of R conversations, its parameters may be estimated directly using a variant of Equation 6.14,

$$a_{i,j}^{CD} = \frac{\sum_{r=1}^{R} \sum_{t=1}^{T_r} \delta_K(\mathbf{q}_{r,t-1}, \mathbf{S}_i)\, \delta_K(\mathbf{q}_{r,t}, \mathbf{S}_j)}{\sum_{r=1}^{R} \sum_{t=1}^{T_r} \delta_K(\mathbf{q}_{r,t-1}, \mathbf{S}_i)} \ , \tag{6.16}$$

where the subscript r identifies individual conversations in the corpus. Equation 6.16 requires that K be identical for all R conversations. The number of parameters $a_{i,j}^{CD}$ in $\boldsymbol{\Theta}^{CD}$ is $N_S^K \cdot N_S^K$, with $N_S^K \cdot \left(N_S^K - 1\right)$ degrees of freedom.

6.3.2 Conditional Independence Among Participants

This huge state space, for non-trivial N_S and/or K, may render ML estimation unlikely to produce useful probability estimates. However, if each single-participant state at time t can be assumed to be conditionally independent (CI), given the multi-participant state at time $t-1$, then

$$P\left(\mathbf{q}_t = \mathbf{S}_j | \mathbf{q}_{t-1} = \mathbf{S}_i, \mathbf{\Theta}^{CD}\right) = \prod_{k=1}^{K} P\left(\mathbf{q}_t\left[k\right] = \mathbf{S}_j\left[k\right] | \mathbf{q}_{t-1} = \mathbf{S}_i, \mathbf{\Theta}_k^{CI}\right) , \tag{6.17}$$

where $\mathbf{\Theta}^{CI} \equiv \{\mathbf{\Theta}_k^{CI}\} = \{\{a_{k;i,n}^{CI}\}\}$, $0 \leq i < N_S^K$, $0 \leq n < N_S$, implements K conditionally independent models, one for each participant.

Given a training corpus of R conversations, maximum likelihood estimation of the parameters of $\mathbf{\Theta}^{CI}$ can be achieved using

$$a_{k;i,n}^{CI} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\delta_K\left(\mathbf{q}_{r,t-1}, \mathbf{S}_i\right)\delta\left(\mathbf{q}_{r,t}\left[k\right], \mathbf{S}_n\right)}{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\delta_K\left(\mathbf{q}_{r,t-1}, \mathbf{S}_i\right)} , \tag{6.18}$$

As in Equation 6.16, Equation 6.18 requires that all R conversations share the same K. If conditional independence may indeed be assumed for conversations, then given an infinite amount of training data,

$$a_{i,j}^{CD} = \prod_{k=1}^{K} a_{k;i,n}^{CI} , \text{ with} \tag{6.19}$$

$$n = n\left(j,k\right) = \arg\max_{\nu}\ \delta\left(\mathbf{S}_j\left[k\right], S_\nu\right) . \tag{6.20}$$

The number of parameters $a_{k;i,n}^{CI}$ in $\mathbf{\Theta}^{CI}$ is $K \cdot N_S^K \cdot N_S$, with $K \cdot N_S^K \cdot (N_S - 1)$ degrees of freedom; in the general case, this represents a significant reduction in complexity over $\mathbf{\Theta}^{CD}$.

Model complexity can be further reduced, by a factor of K, if participants are assumed to be identical. This leads to a single model $\mathbf{\Theta}_{any}^{CI}$, rather than K models $\left\{\mathbf{\Theta}_k^{CI}\right\}$, and

$$P\left(\mathbf{q}_t = \mathbf{S}_j | \mathbf{q}_{t-1} = \mathbf{S}_i, \mathbf{\Theta}^{CD}\right) = \prod_{k=1}^{K} P\left(\mathbf{q}_t\left[k\right] = \mathbf{S}_j\left[k\right] | \mathbf{q}_{t-1} = \mathbf{S}_i, \mathbf{\Theta}_{any}^{CI}\right) , \tag{6.21}$$

with $\mathbf{\Theta}^{CI} \equiv \mathbf{\Theta}_{any}^{CI} = \{a_{i,n}^{CI}\}$. The parameters may be estimated using

$$a_{i,n}^{CI} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{k=1}^{K_r}\delta_{K_r}\left(\mathbf{q}_{r,t-1}, \mathbf{S}_i\right)\delta\left(\mathbf{q}_{r,t}\left[k\right], S_n\right)}{\sum_{r=1}^{R}\sum_{t=1}^{T_r} K_r \cdot \delta_{K_r}\left(\mathbf{q}_{r,t-1}, \mathbf{S}_i\right)} . \tag{6.22}$$

6.3.3 Unconditional Independence Among Participants

Finally, further reductions in model complexity can be achieved if it is assumed that participant behavior is unconditionally independent (UI), namely that

$$P\left(\mathbf{q}_t = \mathbf{S}_j | \mathbf{q}_{t-1} = \mathbf{S}_i, \mathbf{\Theta}^{CD}\right) = \prod_{k=1}^{K} P\left(\mathbf{q}_t\left[k\right] = \mathbf{S}_j\left[k\right], \mathbf{q}_{t-1}\left[k\right] = \mathbf{S}_i\left[k\right], \mathbf{\Theta}_k^{UI}\right) . \tag{6.23}$$

Here, $\mathbf{\Theta}^{UI} \equiv \{\mathbf{\Theta}_k^{UI}\} = \{\{a_{k;m,n}^{UI}\}\}$, $0 \leq m < N_S$, $0 \leq n < N_S$, implements K unconditionally independent models.

The parameters of $\mathbf{\Theta}^{UI}$ may be trained using

$$a^{UI}_{k;m,n} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\delta\left(\mathbf{q}_{r,t-1}\left[k\right],S_m\right)\delta\left(\mathbf{q}_{r,t}\left[k\right],S_n\right)}{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\delta\left(\mathbf{q}_{r,t-1}\left[k\right],S_m\right)}\ , \tag{6.24}$$

In contrast to Equations 6.16, 6.18 and 6.22, Equation 6.29 in no way requires that training conversations have the same number K of participants. If unconditional independence may be assumed, then

$$a^{CD}_{i,j} = \prod_{k=1}^{K} a^{UI}_{k;m,n}\ , \text{ with} \tag{6.25}$$

$$m = m\left(i,k\right) = \arg\max_{\mu}\ \delta\left(\mathbf{S}_i\left[k\right],S_\mu\right) \tag{6.26}$$

$$n = n\left(j,k\right) = \arg\max_{\nu}\ \delta\left(\mathbf{S}_j\left[k\right],S_\nu\right) \tag{6.27}$$

under an infinite amount of training material.

$\mathbf{\Theta}^{UI}$ consists of $K\cdot N_S\cdot N_S$ parameters, with only $K\cdot N_S\cdot\left(N_S-1\right)$ parameters. This complexity may be reduced, by a factor of K as in the conditionally independent case, by assuming participants to be identical. Then

$$P\left(\mathbf{q}_t=\mathbf{S}_j|\mathbf{q}_{t-1}=\mathbf{S}_i,\mathbf{\Theta}^{CD}\right) = \prod_{k=1}^{K} P\left(\mathbf{q}_t\left[k\right]=\mathbf{S}_j\left[k\right],\mathbf{q}_{t-1}\left[k\right]=\mathbf{S}_i\left[k\right],\mathbf{\Theta}^{UI}_{any}\right)\ , \tag{6.28}$$

with $\mathbf{\Theta}^{UI}\equiv\mathbf{\Theta}^{UI}_{any}=\{a^{UI}_{m,n}\}$. Maximum likelihood parameter estimation is performed as before,

$$a^{UI}_{m,n} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{k=1}^{K_r}\delta\left(\mathbf{q}_{r,t-1}\left[k\right],S_m\right)\delta\left(\mathbf{q}_{r,t}\left[k\right],S_n\right)}{\sum_{r=1}^{R}\sum_{t=1}^{T_r}K_r\cdot\delta\left(\mathbf{q}_{r,t-1}\left[k\right],S_m\right)}\ . \tag{6.29}$$

6.3.4 Some Limitations

Representing the multi-participant state of a conversation as a composition of the single-participant states of each of its participants has three rather obvious limitations, linked to the resulting size of the multi-participant state space $\mathbb{S}^K$.

First, direct models exhibit a property henceforth referred to as K-*specificity*: they have an explicit dependence on the number of conversation participants K. For the conditionally dependent and conditionally independent variants ($\mathbf{\Theta}^{CD}$, $\{\mathbf{\Theta}^{CI}_k\}$, and $\{\mathbf{\Theta}^{CI}_{any}\}$), that dependence is reflected both in the size N_S^K of the conditioning state space $\mathbb{S}^K$ and in the length K of vectors representing the states in that space. For the conditionally independent and the unconditionally independent variants which allow participants to behave differently ($\{\mathbf{\Theta}^{CI}_k\}$ and $\{\mathbf{\Theta}^{UI}_k\}$), the dependence is separately reflected in the fact that these models contain K distinct submodels. The only direct model which does not exhibit K-specificity is $\{\mathbf{\Theta}^{UI}_{any}\}$.

As alluded to in the preceding sections, K-specificity requires that all conversations in the training corpus have the same K_{train}. The models may only be applied to test conversations with $K_{test}=K_{train}$, requiring that the training corpus be striated by K into sub-corpora, limiting the amount of data available for the training of any model. This is unfortunate, particularly because conversations with large K, requiring exponentially larger direct models, are likely to be harder to arrange, to record, and to transcribe than conversations with smaller K.

An unfelicitous solution to this aspect of the problem is to "rectify" conversations by extending vocal interaction records $\mathbf{Q}$, for both training and test conversations, with inactive, "phantom" participants up to some ceiling K_{max}. This approach exacerbates other aspects of this problem, as described next.

Second, direct models suffer from $\mathbf{R}$-*specificity*: they are sensitive to the particular assignment of participants to row indices in $\mathbf{Q}$, making them poor at generalization. Were a participant at index k swapped with that at index k', the resulting altered $\mathbf{Q}'=\mathbf{R}_{kk'}\cdot\mathbf{Q}$ would lead to a different model $\mathbf{\Theta}'$ of the same conversation, which in general would quite poorly predict $\mathbf{Q}$. (Here, $\mathbf{R}_{kk'}$ is a matrix rotation operator formed by swapping columns k and k' of the $K\times K$ identity matrix $\mathbf{I}$.) This is an undesired property, since index assignment is an essentially arbitrary function, subject to

the vagaries of seating arrangement, microphone channel selection, the preferences of the recording engineer, or other factors which are irrelevant to conversational conduct. $\mathbf{Q}$ could of course be rotated exhaustively during training, but for K participants such an operation has complexity $K!$.

A mathematically less severe but scientifically disappointing aspect of direct models, even if K-specificity and $\mathbf{R}$-specificity were to be addressed by rectifying interaction records up to $K_{max} \geq K$ and then exhaustively permuting participant indices $K_{max}!$ times, is that the resulting direct model is theoretically vacuous. General conversational behavior cannot be readily understood by studying what such a model has learned.

These three constraints circumscribe the serviceability of most direct models to only a handful of situations, characterized by small K, small N_S, limited domain variability (in terms of specific participants, specific index assignments, etc.), and easy access to large quantities of training data. The one direct model type which exhibits neither K-specificity nor $\mathbf{R}$-specificity, as Table 6.1 shows, is $\mathbf{\Theta}^{UI}_{any}$, namely that which treats participants as unconditionally independent. This model type, of course, also has nothing to say about participant interaction.

Model Type		DoF	K-spec	$\mathbf{R}$-spec	interact
conditionally dependent	$\mathbf{\Theta}^{CD}$	$N_S^K \cdot \left(N_S^K - 1\right)$	✓	✓	✓
conditionally	$\{\mathbf{\Theta}^{CI}_k\}$	$K \cdot N_S^K \cdot (N_S - 1)$	✓	✓	✓
independent	$\mathbf{\Theta}^{CI}_{any}$	$N_S^K \cdot (N_S - 1)$	✓	✓	✓
unconditionally	$\{\mathbf{\Theta}^{UI}_k\}$	$K \cdot N_S \cdot (N_S - 1)$	✓	✓	
independent	$\mathbf{\Theta}^{UI}_{any}$	$N_S \cdot (N_S - 1)$			

Table 6.1: Direct compositional models described in this section, the number of their degrees of freedom (DoF), and three characteristics: (1) whether they exhibit K-specificity ("K-spec"); (2) whether they exhibit $\mathbf{R}$-specificity ("$\mathbf{R}$-spec"); and (3) whether they are able to model participant interaction ("interact").

6.4 Ergodic Binomial-Participant Models

The description of the distribution of speech involves a somewhat special setting, in which each participant may be in one of only *two* distinct behaviors, speech and non-speech. A large number of other problems relevant to modeling vocal interaction can also be cast into a similar setting (e.g., laughter versus non-laughter, possessing the floor versus not possessing the floor, etc.).

In these cases, the behavior of each participant k, $1 \leq k \leq K$, is drawn from the set

$$\psi_k(t) \in \mathbb{B} \equiv \{\square, \blacksquare\} . \tag{6.30}$$

For convenience, $\square$ and $\blacksquare$ may be equated with 0 and 1, respectively.

When the intended single-participant topology is ergodic, as in panel (a) of Figure 6.1, the number of single-participant states is also 2,

$$\mathbf{q}_t[k] \in \mathbb{S} \equiv \{\square, \blacksquare\} . \tag{6.31}$$

It is particularly straightforward to enumerate and manage multi-participant states in the ergodic setting with binary-valued participants. The number of multi-participant states is $N = 2^K$, and the index of each element of

$$\mathbb{S}^K \equiv \mathbb{S} \times \mathbb{S} \times \cdots \times \mathbb{S} \tag{6.32}$$

$$= \{\mathbf{S}_0, \mathbf{S}_1, \cdots, \mathbf{S}_{N-1}\} , \tag{6.33}$$

in base 2, can be interpreted as a binary string whose kth digit corresponds to the state of participant k, under the mapping $\{\square, \blacksquare\} \equiv \{0, 1\}$.

6.4.1 Degree-of-Overlap as an Organizing Principle

An alternative to conditioning the behavior of all participants on their immediately precedent, compositional behavior description $\mathbf{q}_{t-1}$ is to condition it instead on some emergent discrete characteristic $\chi(\mathbf{q}_{t-1})$, which is not necessarily attributable to any specific participant or participants. To fully address the limitations of direct compositional models,

1. $\chi : \mathbf{q}_t \mapsto \chi(\mathbf{q}_t)$ must be a many-to-one mapping.
2. χ must be K-*invariant*, i.e., not exhibit K-specificity. One way to achieve this is to require that appending additional single-participant states $q_{k+1,t}$ to $\mathbf{q}_t$ affects the value of $\chi(\mathbf{q}_t)$ only if the additional participants are in ■,

$$\begin{aligned} \chi(\mathbf{q}_t \oplus \square) &= \chi(\mathbf{q}_t) && (6.34) \\ \chi(\mathbf{q}_t \oplus \blacksquare) &\neq \chi(\mathbf{q}_t) \ , && (6.35) \end{aligned}$$

   where $\oplus$ represents vector concatenation.
3. χ must be $\mathbf{R}$-*invariant*, i.e., not exhibit $\mathbf{R}$-specificity:

$$\chi(\mathbf{R} \cdot \mathbf{q}_t) = \chi(\mathbf{q}_t) \quad (6.36)$$

   for arbitrary index rotation $\mathbf{R}$, but

$$\chi(\mathbf{q}_t) \neq \chi(\mathbf{q}'_t) \quad (6.37)$$

   if no $\mathbf{R}$ exists such that $\mathbf{R} \cdot \mathbf{q}_t = \mathbf{q}'_t$.

The conditionally dependent transition probability to be modeled can be rewritten as

$$\begin{aligned} P(\mathbf{q}_t \,|\, \mathbf{q}_{t-1}) &\doteq P(\mathbf{q}_t \,|\, \chi(\mathbf{q}_{t-1})) && (6.38) \\ &= P(\mathbf{q}_t, \chi(\mathbf{q}_t) \,|\, \chi(\mathbf{q}_{t-1})) && (6.39) \\ &= P(\mathbf{q}_t \,|\, \chi(\mathbf{q}_t), \chi(\mathbf{q}_{t-1})) \cdot P(\chi(\mathbf{q}_t) \,|\, \chi(\mathbf{q}_{t-1})) && (6.40) \\ &\doteq P(\mathbf{q}_t \,|\, \chi(\mathbf{q}_t)) \cdot P(\chi(\mathbf{q}_t) \,|\, \chi(\mathbf{q}_{t-1})) \ . && (6.41) \end{aligned}$$

Equation 6.39 is a strict equality since $\chi(\mathbf{q}_t)$ is a deterministic function of $\mathbf{q}_t$; Equation 6.41 assumes that $\mathbf{q}_t$ is conditionally independent of $\chi(\mathbf{q}_{t-1})$ given $\chi(\mathbf{q}_t)$.

The first term in Equation 6.41 allows for arbitrary distribution of probability mass over the possible alternatives $\mathbf{q}_t$ given $\chi(\mathbf{q}_t)$. Since χ is required to be many-to-one, its inverse is one-to-many. Furthermore, R-invariance requires that the model treat participants identically, and this means that, by extension, the first term in Equation 6.41 should be a uniform distribution. Therefore,

$$P(\mathbf{q}_t \,|\, \mathbf{q}_{t-1}) \propto P(\chi(\mathbf{q}_t) \,|\, \chi(\mathbf{q}_{t-1})) \ . \quad (6.42)$$

It remains only to find a χ which satisfies the aforementioned requirements. A sociolinguistically meaningful characteristic of $\mathbf{q}_t$ which is successful in this regard is the *degree of overlap* (DO), or the number of participants in the single-participant state ■ simultaneously,

$$\begin{aligned} \|\mathbf{q}_t\| &= \sum_{k=1}^{K} \delta(\mathbf{q}_t[k], \blacksquare) && (6.43) \\ &\in \{0, 1, \cdots, K\} \ . \end{aligned}$$

Given this definition, the conditional probability of "observing state $\mathbf{S}_j$ at time t given that state $\mathbf{S}_i$ was observed at time $t-1$" is taken to be proportional to the conditional probability of "observing n_j participants speaking at time t given that n_i were speaking at time $t-1$". Formally,

$$P(\mathbf{q}_t = \mathbf{S}_j \,|\, \mathbf{q}_{t-1} = \mathbf{S}_i) \propto P(\|\mathbf{q}_t\| = n_j \,|\, \|\mathbf{q}_{t-1}\| = n_i) \ , \quad (6.44)$$

where $n_i \equiv \|\mathbf{S}_i\|$, $0 \leq i < 2^K$, serves as notational shorthand.

The effect of Equation 6.44 is to tie the $2^K \times 2^K$ probabilities governing the evolution of $\mathbf{Q}$. The resulting much smaller parameter space consists of only $(K+1) \times (K+1)$ probabilities. This has significant implication for estimated parameter robustness.

Since Equation 6.43 is applied to every vector entity on the left-hand side of Equation 6.44, the latter can be said to completely mitigate the problems of $\mathbf{R}$- and K- specificity. $\mathbf{R}$-specificity is addressed because addition, on the right-hand side of Equation 6.43, is a commutative and therefore rotation-invariant operation. K-specificity is addressed because Equation 6.43 sums over only those participants who are in state ■, ignoring those participants who are in □ and therefore the total number K of participants (this assumes that adding silent participants does not modify conversational dynamics).

As argued, the proportionality constant implicit in Equation 6.44 ensures that probability mass is uniformly distributed across the possible $\mathbf{q}_t$ states of all like transitions emanating from $\mathbf{q}_{t-1}$. The number of like transitions, from a state $\mathbf{S}_i$ whose $\|\mathbf{S}_i\| = n_i$ to a state $\mathbf{S}_j$ whose $\|\mathbf{S}_j\| = n_j$ is henceforth referred to as its *multiplicity* m_{ij}; since the model is ergodic, m_{ij} is independent of i, i.e., $m_{ij} \equiv m_j$. The proportionality constant in Equation 6.44 is therefore taken to be $1/Z_i \cdot m_j$, where Z_i is a constant which ensures that the probability of transition is unity,

$$Z_i \equiv \sum_{j=0}^{2^K-1} \frac{1}{m_j} P\left(\|\mathbf{q}_t\| = n_j \,|\, \|\mathbf{q}_{t-1}\| = n_i\right) . \tag{6.45}$$

An example may most directly demonstrate how transition probability mass is shared among target states $\mathbf{q}_t = \mathbf{S}_j$; $K = 3$ is selected for simplicity. Only one state with $n_j = 0$ exists, namely $\mathbf{S}_0 = [\square, \square, \square]^T$, and only one with $n_j = 3$, namely $\mathbf{S}_7 = [■, ■, ■]^T$; therefore, for these states, $m_j = 1$. There are three states each with $n_j = 1$ and $n_j = 2$, and for these six states $m_j = 3$. It is readily seen that, for general K,

$$m_j = \frac{K!}{(K-n_j)! \cdot n_j!} . \tag{6.46}$$

The DO model implementing these concepts, $\mathbf{\Theta}_{DO}$, implicitly assumes conditional dependence among participants. It consists of the parameters a_{n_i,n_j}, which are estimated via

$$a^{CD}_{DO;n_i,n_j} = \frac{\sum_{r=1}^{R} \sum_{t=1}^{T_r} \delta\left(\|\mathbf{q}_{r,t-1}\|, n_i\right) \delta\left(\|\mathbf{q}_{r,t}\|, n_j\right)}{\sum_{r=1}^{R} \sum_{t=1}^{T_r} \delta\left(\|\mathbf{q}_{r,t-1}\|, n_i\right)} \tag{6.47}$$

using a corpus of R conversations. Here, $\delta(\cdot)$ is the standard Kronecker delta defined over scalars; it should be noted that K does not figure anywhere in Equation 6.47. This suggests that training can be performed using a corpus of K-heterogenous conversations, and that the DO formalism is a K-invariant representation of conversational dynamics.

During testing, the model provides the probabilities

$$P\left(\mathbf{q}_t = \mathbf{S}_j \,|\, \mathbf{q}_{t-1} = \mathbf{S}_i, \mathbf{\Theta}^{CD}_{DO}\right) = \frac{a_{n_i,n_j}}{Z_i \cdot m_j} , \tag{6.48}$$

where the number K of participants in the test conversation does not figure explicitly. K is of course implicitly present in the constants Z_i and m_j. Normalization by $Z_i \cdot m_j$ can therefore be seen as a mapping from the K-invariant DO formalism into the K-specific setting of any particular conversation for which it is expected to provide transition probability estimates.

Where conditional participant independence may be assumed within the DO formalism, each product in Equation 6.17 can be expressed as

$$\begin{aligned} P\left(\mathbf{q}_t[k] = \mathbf{S}_j[k] \,|\, \mathbf{q}_{t-1} = \mathbf{S}_i, \mathbf{\Theta}^{CI}_k\right) &= \\ P\left(\mathbf{q}_t[k] = \mathbf{S}_j[k] \,|\, \|\mathbf{q}_{t-1}\| = \|\mathbf{S}_i\|, \mathbf{\Theta}^{CI}_{DO,k}\right) &. \end{aligned} \tag{6.49}$$

The proportionality is an equality because the single-participant state space from which $\mathbf{q}_t[k]$ is drawn is identical to the domain of $\|\mathbf{q}_t[k]\|$. $\mathbf{\Theta}^{CI}_{DO} \equiv \{\mathbf{\Theta}^{CI}_{DO,k}\} = \{\{a^{UI}_{DO,k;n_i,n}\}\}$, $0 \leq n_i \leq K$, $0 \leq n < 2$, implements K conditionally independent submodels.

The parameters of $\mathbf{\Theta}^{CI}_{DO}$ may be estimated using

$$a^{CI}_{DO,k;n_i,n} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\delta\left(\|\mathbf{q}_{r,t-1}\|,n_i\right)\delta\left(\mathbf{q}_{r,t}\left[k\right],n\right)}{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\delta\left(\|\mathbf{q}_{r,t-1}\|,n_i\right)}\,, \tag{6.50}$$

or, if the model is to consist of a single model $\mathbf{\Theta}^{CI}_{DO,any}$,

$$a^{CI}_{DO;n_i,n} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{k=1}^{K_r}\delta\left(\|\mathbf{q}_{r,t-1}\|,n_i\right)\delta\left(\mathbf{q}_{r,t}\left[k\right],n\right)}{\sum_{r=1}^{R}\sum_{t=1}^{T_r}K_r\cdot\delta\left(\|\mathbf{q}_{r,t-1}\|,n_i\right)}\,. \tag{6.51}$$

Unconditional independence may of course also be considered within the DO formalism, but in that case the model is identical to the direct, compositional model $\mathbf{\Theta}^{UI}$. This is because the single-participant state space $\mathbb{S}\equiv\{\square,\blacksquare\}\equiv\{0,1\}$ is homomorphic with its χ mapping $\{\|\left[\square\right]\|,\|\left[\blacksquare\right]\|\}=\{0,1\}\equiv\{\square,\blacksquare\}$.

6.4.2 Within-Bigram Index Continuity

Despite its simplicity and seeming felicity, the DO approach exhibits an important shortcoming which has bearing on conversation. Namely, it loses track of whether the *same* or *different* participants spoke at instants $t-1$ and t. As a result, for example, it predicts that the two transitions

$$\left[\blacksquare,\square,\square,\square\right]^T \longrightarrow \left[\square,\square,\blacksquare,\square\right]^T$$

and

$$\left[\blacksquare,\square,\square,\square\right]^T \longrightarrow \left[\blacksquare,\square,\square,\square\right]^T$$

are equally likely, since both are $(n_i,n_j)=(1,1)$ transitions. This is not borne out empirically; intervals of speech typically vary widely in duration, and frame steps and sizes for discretizing $\mathbf{Q}$ are chosen to be at least as small as the shortest such intervals (to avoid missing them). Under these circumstances, each participant's behavior at time t is more strongly correlated with their *own* behavior at $t-1$ than with that of other participants at either $t-1$ or t. As a result, the second of the transitions in the above example is generally significantly more likely than the first. The DO model fails to account for this readily observable characteristic of $\mathbf{Q}$, by confounding the two transition types.

The DO model can be extended to implement the requisite sensitivity to participant index within each transition bigram, yet still retain global $\mathbf{R}$-invariance. The resulting *extended degree-of-overlap* (EDO) model is implemented by letting

$$\begin{aligned} &P\left(\mathbf{q}_t=\mathbf{S}_j\,|\,\mathbf{q}_{t-1}=\mathbf{S}_i\right) \propto \\ &\qquad P\left(\|\mathbf{q}_t\|=n_j,\|\mathbf{q}_t\cdot\mathbf{q}_{t-1}\|=o_{ij}\,|\,\|\mathbf{q}_{t-1}\|=n_i\right)\,, \end{aligned} \tag{6.52}$$

where $o_{ij}\equiv\|\mathbf{S}_i\cdot\mathbf{S}_j\|\leq\min\left(n_i,n_j\right)$, is adopted for notational convenience. The binary operator $\mathbf{S}_i\cdot\mathbf{S}_j$ is defined over binary-valued K-length vectors $\mathbf{S}_i$ and $\mathbf{S}_j$ as

$$\left(\mathbf{S}_i\cdot\mathbf{S}_j\right)\left[k\right] = \begin{cases}\blacksquare, & \text{if } \mathbf{S}_i\left[k\right]=\blacksquare \text{ and } \mathbf{S}_j\left[k\right]=\blacksquare\\ \square & \text{otherwise}\end{cases}\,. \tag{6.53}$$

The resulting $(\mathbf{S}_i\cdot\mathbf{S}_j)\in\{\square,\blacksquare\}^K$.

It can be seen from Equation 6.52 that the proposed EDO model describes transition into a joint space of two elements, one of which, n_j, is the same as for the DO model. The other element, o_{ij}, is a deterministic descriptor of the similarity between the two states $\mathbf{q}_{t-1}=\mathbf{S}_i$ and $\mathbf{q}_t=\mathbf{S}_j$. In particular, it is the number of participants who are speaking at both instants $t-1$ and t. The evolution of $\mathbf{q}$ under the EDO model is depicted in Figure 6.3.

Like the DO model, the EDO model mitigates $\mathbf{R}$- and K- specificity; all the symbols on the right-hand side of Equation 6.52 are additions. Furthermore, the term o_{ij}, computed using Equation 6.53, only counts those participants who were speaking at both $t-1$ and t, and thereby ignores at least as many participants as are ignored in computing n_i or n_j.

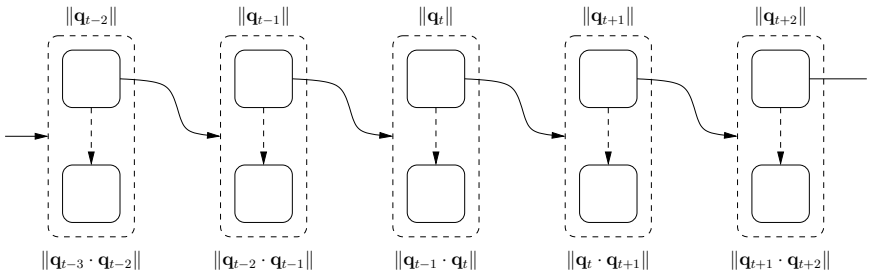


Figure 6.3: Evolution of $\mathbf{q}$ under the EDO model. At each t, the variable $\|\mathbf{q}_t\|$ gives rise to a two-element vector (shown using dashed lines) consisting of $\|\mathbf{q}_{t+1}\|$ as well as $\|\mathbf{q}_t \cdot \mathbf{q}_{t+1}\|$, whose value is bound by $\|\mathbf{q}_{t+1}\|$ from above (a constraint shown using dashed arrows).

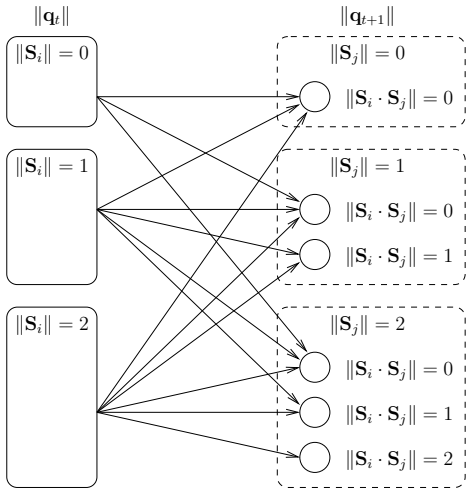


Figure 6.4: Unique transition types in the EDO model space with at most 2 participants in ■ simultaneously; higher degrees of overlap not shown.

Although $(n_i, [n_j, o_{ij}])$ more correctly describes EDO transitions, the notation (n_i, o_{ij}, n_j) will be used instead for simplicity. n_i and n_j are both bounded as in the DO model ($0 \leq n_i \leq K$ and $0 \leq n_j \leq K$), but o_{ij} must be no bigger than their minimum. More counter-intuitively, o_{ij} also has a lower bound, namely $o_{ij} > n_i + n_j - K$, for a particular conversation with K participants. To see why this is so, consider the transition $(3, 0, 3)$ when $K = 4$. This implies that $n_i = 3$ participants were in ■ at instant $t-1$, and that $n_j = 3$ *other* participants (because $o_{ij} = 0$) are in ■ at instant t. This is clearly not possible with only 4 participants. The total number of unique transition types, therefore, is given by

$$\sum_{n_i=0}^{K} \sum_{n_j=0}^{K} \left[\min(n_i, n_j) - \max(0, n_i + n_j - K) + 1 \right] . \tag{6.54}$$

A subset of these transition types is shown in Figure 6.4.

As for the DO model, use of Equation 6.52 requires the specification of its implicit proportionality constant. Identical treatment of participants requires that this constant be again equal to $1/z_{i \cdot m_{ij}}$, but for the EDO model $m_{ij} \neq m_j$ is not independent of i. This is because in considering the next state $\mathbf{S}_j$, the EDO model accounts for whether the *same* participants will be speaking as were speaking at time $t-1$, in state $\mathbf{S}_i$.

To understand the sharing of transition probability mass across like transitions in the EDO model, the example with $K = 3$ from the previous section is again considered. Suppose at time $t-1$ the conversation is in state $\mathbf{S}_3 = [■, ■, □]^T$.

Then transition into any of the 8 possible "to" states is described with the EDO triplet notation shown in Table 6.2; also shown is the multiplicity of each transition. As this example shows, what were only 4 unique DO transitions are 6 unique EDO transitions. The EDO model discriminates, given $\mathbf{q}_{t-1} = [\blacksquare, \blacksquare, \square]^T$, between transitions which the DO model does not.

"to" state, $\mathbf{q}_t$	DO transition		EDO transition	
	(n_i,n_j)	m_j	(n_i,o_{ij},n_j)	m_{ij}
$\mathbf{S}_0 = [\square, \square, \square]^T$	(2 , 0)	1	(2 , 0 , 0)	1
$\mathbf{S}_1 = [\blacksquare, \square, \square]^T$	(2 , 1)	3†	(2 , 1 , 1)	2•
$\mathbf{S}_2 = [\square, \blacksquare, \square]^T$	(2 , 1)	3†	(2 , 1 , 1)	2•
$\mathbf{S}_3 = [\blacksquare, \blacksquare, \square]^T$	(2 , 2)	3‡	(2 , 2 , 2)	1
$\mathbf{S}_4 = [\square, \square, \blacksquare]^T$	(2 , 1)	3†	(2 , 0 , 1)	1
$\mathbf{S}_5 = [\blacksquare, \square, \blacksquare]^T$	(2 , 2)	3‡	(2 , 1 , 2)	2∗
$\mathbf{S}_6 = [\square, \blacksquare, \blacksquare]^T$	(2 , 2)	3‡	(2 , 1 , 2)	2∗
$\mathbf{S}_7 = [\blacksquare, \blacksquare, \blacksquare]^T$	(2 , 3)	1	(2 , 2 , 3)	1

Table 6.2: Transitions from state $\mathbf{q}_{t-1} = \mathbf{S}_3 = [\blacksquare, \blacksquare, \square]^T$, with $K = 3$. Types with non-unity multiplicity m_j or m_{ij} are identified with unique symbols (†, ‡, •, ∗).

Most generally,

$$m_{ij} \;=\; \frac{n_i!}{o_{ij}!\,(n_i - o_{ij})!} \cdot \frac{(K - n_i)!}{(n_j - o_{ij})!\,(K - n_i - n_j + o_{ij})!} \;. \tag{6.55}$$

As for the DO model, m_{ij} is a deterministic constant, here additionally dependent on the number K of participants in the test conversation. The parameters of the EDO model which need to be estimated from data are $a^{CD}_{EDO;n_i,o_{ij},n_j}$; estimation is achieved via

$$a^{CD}_{EDO;n_i,o_{ij},n_j} \;= \tag{6.56}$$
$$\frac{\displaystyle\sum_{r=1}^{R}\sum_{t=1}^{T_r}{}_{\substack{\\ n_i+n_j-o_{ij}<K_{max}}} \delta\left(\|\mathbf{q}_{r,t-1}\|, n_i\right)\, \delta\left(\|\mathbf{q}_{r,t-1} \cdot \mathbf{q}_{r,t}\|, o_{ij}\right)\, \delta\left(\|\mathbf{q}_{r,t}\|, n_j\right)}{\displaystyle\sum_{r=1}^{R}\sum_{t=1}^{T_r}{}_{\substack{\\ n_i+n_j-o_{ij}<K_{max}}} \delta\left(\|\mathbf{q}_{r,t-1}\|, n_i\right)} \;.$$

Some comments are in order regarding the value of K_{max} in Equation 6.56. When the model is trained using conversations of homogenous number K_{train} of participants, and is then applied to only those unseen conversations for which the number of participants is $K_{test} = K_{train}$, K_{max} in Equation 6.56 is simply K_{train} and identically K_{test}. However, when the training conversations very in their number of participants, or when K_{test} cannot be assumed *a priori* (i.e., a model is required which will be applicable for any K_{test}), K_{max} is a ceiling on the anticipated degree of overlap. Imposing a K_{max} which is not equal to K_{train}, for a training conversation which is being used to accumulate statistics for parameter inference, or one which is not equal to K_{test} when applying to an unseen conversation, requires a redefinition of the $\|\cdot\|$ operator:

$$n_i \;=\; \min\left(\,\|\mathbf{S}_i\|,\, K_{max}\,\right) \tag{6.57}$$
$$n_j \;=\; \min\left(\,\|\mathbf{S}_j\|,\, K_{max}\,\right) \tag{6.58}$$
$$o_{ij} \;=\; \min\left(\,\|\mathbf{S}_i \cdot \mathbf{S}_j\|,\, n_i,\, n_j\right) \;, \tag{6.59}$$

This renders the EDO model sensitive to all degrees of overlap of K_{max} or less, by collapsing overlap of more than K_{max} participants vocalizing simultaneously to be indistinguishable from that of exactly K_{max} such participants.

To close this section, it should be noted that both conditionally independent and unconditionally independent variants of the EDO model are possible [140], with the unconditionally independent variant functionally identical to a directly estimated compositional model. Those variants are not used in this thesis.

6.5 Non-Ergodic and Multinomial-Participant Models

This section extends several of the ideas of the previous section to circumstances in which

1. the number of single-participant behaviors is greater than two; and/or
2. the intended model topology is not ergodic but explicitly prohibits certain transitions, with or without state duplication for mimimum duration enforcement.

These are the settings in panels (b), (c), and (d) of Figure 6.1, of which panel (d) depicts the most general. The modified EDO model is henceforth referred to as the generalized EDO (gEDO) model.

6.5.1 Topology Construction

Increasing the number of single-participant behaviors $\mathbb{B}$ or merely duplicating states in $\mathbb{S}$ identically results in $N_S > 2$. Even for seemingly small N_S, N may be so prohibitively large as to render the mere enumeration of states unduly time-consuming. State pruning, which can be performed after complete state space construction in the ergodic $N_S = 2$ case, must therefore be performed in the general case *during* state space construction.

This can be achieved via the following proposed sequence of steps:

Step 1 Enumerate and prune multi-participant behaviors $\mathbb{B}^K$, yielding $\mathbb{B}^K_*$.

Step 2 Expand the surviving $\mathbb{B}^K_*$ into multi-participant states and prune, yielding $\mathbb{S}^K_*$.

Step 3 License only those multi-pariticipant state transitions, among states in $\mathbb{S}^K_*$, which are licensed by single-participant state transitions.

Pruning is achieved by enforcing maxima on the number of participants who are simultaneously licensed to be in a particular single-participant behavior B_m. Given N_B single-participant behaviors $\mathbb{B}$, a vector $\mathbf{K}^{\mathbb{B}}_{max} \in \mathbb{N}^{N_B}$ is formed, with $\mathbf{K}^{\mathbb{B}}_{max}[m]$ indicating the maximum number of participants which can simultaneously be in behavior B_m. The pruned $\mathbb{B}^K_*$ is determined using Algorithm 2, where $valid(\mathbf{B}, \mathbf{K}, m)$ returns TRUE if multiparticipant behavior $\mathbf{B}$ contains at most $\mathbf{K}[m]$ participants in single-participant behavior B_m.

Algorithm 2 Prune multi-participant behaviors $\mathbb{B}^K$ using constraints $\mathbf{K}^{\mathbb{B}}_{max}$

Require: $N_B > 1$
Require: $\mathbf{K}^{\mathbb{B}}_{max} \in \mathbb{N}^{N_B}$
Require: $|\mathbb{B}^K| > 1$
Require: $|\mathbb{B}^K_*| = 0$
for all $\mathbf{B} \in \mathbb{B}^K$ **do**
  $VALID \leftarrow$ TRUE
  for all $m \in \{1, \cdots, N_B\}$ **do**
    $VALID \leftarrow VALID$ AND $valid\left(\mathbf{B}, \mathbf{K}^{\mathbb{B}}_{max}, m\right)$.
  end for
  if $VALID ==$ TRUE **then**
    Append $\mathbf{B}$ to $\mathbb{B}^K_*$.
  end if
end for

Following application of Algorithm 2,

$$\mathbb{B}_*^K = \{\mathbf{B}_{*0}, \mathbf{B}_{*1}, \cdots, \mathbf{B}_{*\left(N_B^*-1\right)}\} \tag{6.60}$$

$$\subset \mathbb{B}^K \tag{6.61}$$

with $N_B^* \equiv |\mathbb{B}_*^K| \leq N_B^K$. Expanding these pruned multiparticipant behaviors, as per Step 2 above, is achieved using Algorithm 3.

Algorithm 3 Construct multi-participant states $\mathbb{S}_*^K$ from $\mathbb{B}_*^K$, pruning using constraints $\mathbf{K}_{max}^{\mathbb{S}}$

Require: $N_S > 1$
Require: $\mathbf{K}_{max}^{\mathbb{S}} \in \mathbb{N}^{N_S}$
Require: $|\mathbb{B}_*^K| > 0$
Require: $|\mathbb{S}_*^K| = 0$
for all $\mathbf{B} \in \mathbb{B}_*^K$ **do**
  Build all multi-participant states $\{\mathbf{S_B}\}$ corresponding to $\mathbf{B}$.
  for all $\mathbf{S} \in \{\mathbf{S_B}\}$ **do**
    $VALID \leftarrow$ TRUE
    for all $m \in \{1, \cdots, N_S\}$ **do**
      $VALID \leftarrow VALID$ AND $valid\left(\mathbf{S}, \mathbf{K}_{max}^{\mathbb{S}}, m\right)$.
    end for
    if $VALID ==$ TRUE **then**
      Append $\mathbf{S}$ to $\mathbf{S}_*^K$.
    end if
  end for
end for

Finally, once the pruned set of multi-participant states $\mathbf{S}_*^K$ is enumerated, states must be connected via transitions to implement the model topology. As mentioned towards the end of Subsection 6.2.4, this is achieved by considering all potential multi-participant transitions $(\mathbf{S}_i, \mathbf{S}_j)$, $\mathbf{S}_i \in \mathbf{S}_*^K$, $\mathbf{S}_j \in \mathbf{S}_*^K$, and licensing each such transition if and only if $(\mathbf{S}_i[k], \mathbf{S}_j[k])$ is an allowed transition in the single-participant topology, for all $1 \leq k \leq K$.

6.5.2 Transition Probability Modeling

Once a multi-participant topology is constructed, it is possible to force-align an observed $\mathbf{Q}$ via a Viterbi decoding pass, and to accumulate transition counts. Doing so requires a specification of what constitutes a unique transition type.

This definition is a direct extension of what was proposed for the ergodic EDO model. Rather than merely counting the number of times a specific compositional state $\mathbf{S}_i$ egresses to a specific compositional state $\mathbf{S}_j$, which would yield a $\mathbf{R}$-specific model, both $\mathbf{S}_i$ and $\mathbf{S}_j$ are first characterized by the *number* of participants in particular single-participant states in each of these multi-participant states. A fixed ordering on the N_S single-participant states in $\mathbb{S}$ allows each multi-participant state $\mathbf{S}_i$ to be characterized by a N_S-length vector $\|\mathbf{S}_i\|$, whose entries indicate the number of participants in each single-participant state. The count for the inactive rest state (defined as that in which all participants are silent) is ignored, allowing this characterization to be independent of the total number K of participants in any conversation.

Such a formulation yields an analogue of the DO model of Subsection 6.4.1, which suffers from the problem that, given $\|\mathbf{S}_i\|$ and $\|\mathbf{S}_j\|$ for a transition from $\mathbf{S}_i$ to $\mathbf{S}_j$, it is not known which single-participant transitions were taken. Extending this to the analogue of the EDO model requires that unique transition types be characterized by a tree structure, rather than a simple sum, which consists of $\|\mathbf{S}_i\|$, the number of participants in each single-participant state within the compositonal state $\mathbf{S}_i$, and also a separate vector of sums for each element in $\|\mathbf{S}_i\|$. The latter indicates how those participants in each *specific* single-participant state in $\mathbf{S}_i$ subsequently behave[4].

[4]One implementation of this, used in this thesis, is to simply build a key-generating function for each pair $(\mathbf{S}_i, \mathbf{S}_j)$ that produces a $\mathbf{R}$-invariant ASCII string, and then to use that key when enumerating licensed $(\mathbf{S}_i, \mathbf{S}_j)$ transitions in $\mathbb{S}_*^K \times \mathbb{S}_*^K$ to build a transition probability table a_{ij}.

6.6 Relevance to Other Chapters

The transition probability modeling framework proposed in this chapter is empirically validated in many chapters of Part IV.

In Chapter 10, the ergodic EDO model of Section 6.4, which assumes that each participant is in only one of two states, is applied to the problem of estimating the likelihood of true $\mathbf{Q}$. This is achieved using a measure of turn-taking perplexity, analogous to that used in language modeling. The EDO model is shown to generlize well to unseen conversations, and to more successfully predict speech activity distribution than do models which ignore interaction.

Chapter 12 employs the EDO model to contrast the distribution of speech to that of laughter. It successfully reveals the broadly accepted wisdom that participants take turns speaking but tend to laugh together. Although this can be measured merely by observing static overlap descriptions, the EDO model sheds light on how overlap arises, and is resolved, for the two types of vocal activity.

The non-ergodic, generalized EDO (gEDO) model of Section 6.5 is used in Chapters 11 and 13, which treat the detection of vocal activity. Chapter 11 imposes minimum duration constraints on the two alternative behaviors characterizing each participant. Chapter 13 considers ergodic models, but participants may be characterized by one of three or one of four alternative behaviors. In both cases, the number of topological states in the HMM decoder proposed for each task is greater than 2.

In all four of Chapter 10, 11, 12, and 13, the performance of the EDO and/or gEDO model is contrasted with the unconditionally independent model $\boldsymbol{\Theta}^{UI}$, which ignores interaction.

Finally, Chapter 7 proposes a parametric alternative to the conditionally independent EDO model $\boldsymbol{\Theta}^{CI}_{EDO}$, which was briefly mentioned but not described in this chapter. The complete space of non-parametric models considered in this thesis is shown in Figure 6.5.

6.7 Summary

Direct compositional models of vocal interaction suffer from two problems which this chapter has solved. First, they are not K-invariant, making them irrelevant to conversations whose number of participants is K if the training corpus does not exclusively contain conversations of K participants. Second, they are not $\mathbf{R}$-invariant, making them sensitive to the arbitrary index assignment of specific participants to the rows of $\mathbf{Q}$.

The extended-degree-of-overlap (EDO) and generalized extended-degree-of-overlap (gEDO) models, described in Sections 6.4 and 6.5, respectively, eliminate these problems by ignoring those participants which are inactive at instants $t-1$ and t, and merely counting the number of participants in all other states at those instants. The former enforces K-invariance while the latter renders transition representations $\mathbf{R}$-invariant. These techniques collapse the number of free model parameters to a small integer, ensuring that even with small amounts of training data robust models may still be inferred. They also make it possible to infer models for heterogenous K.

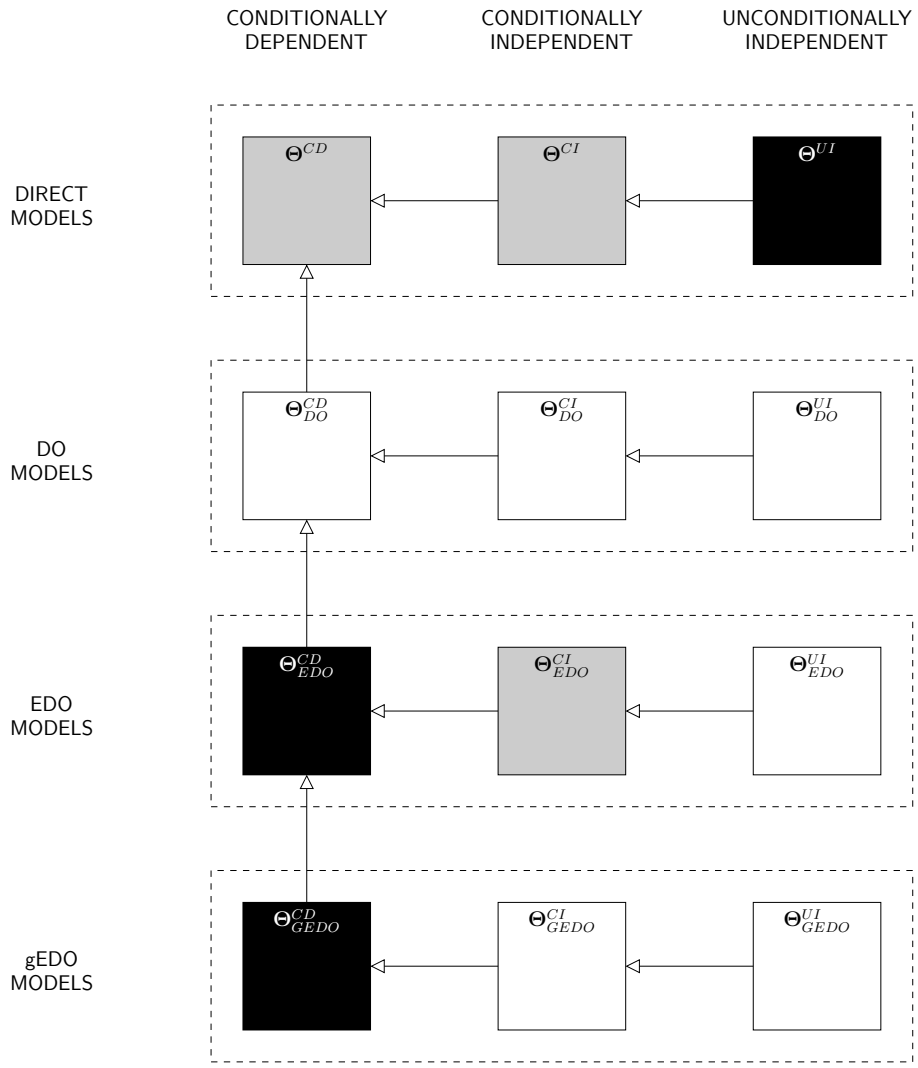


Figure 6.5: The space of multi-participant models described in this chapter. Arrows point towards more general variants. Gray indicates models discussed in other chapters of Part III; black indicates models exercised in Part IV.

Chapter 7

Parametric State-Space Multi-Participant Models

7.1 Introduction

As Chapter 6 has argued, a major limitation of compositional models of interaction is the size of their state space. The generalized extended-degree-of-overlap model significantly reduces the number of parameters which must be estimated in the conditionally independent multi-participant model. Given an ergodic topology in which each participant can be in only one of two states, the number of free parameters was shown to be $2^K \cdot \left(2^K - 1\right)$.

The EDO model achieves this reduction by expressing the joint state of a conversation in terms of the number of participants which are simultaneously vocalizing; transitions are therefore participant-independent, since probability mass is distributed uniformly. For example, the EDO transition probability $P\left(\, \|\mathbf{q}_t\| = n_j = 2, \|\mathbf{q}_t \cdot \mathbf{q}_{t-1} = o_{ij} = 1 \,|\, \|\mathbf{q}_{t-1}\| = n_i = 1 \,\right)$, in a conversation of $K = 5$ participants, is uniformly distributed over the four participants who were not vocalizing at $t - 1$. This rotation-invariance is desirable in some cases, for example during the detection of vocal activity. However, in other cases — such as the inference of participant characteristics once vocal activity is detected — rotation invariance is undesirable since it does not discriminate among participants.

So far, the only solution to this problem has been the imposition of an additional constraint for compositional models, namely that participant states at instant t are conditionally independent, given all participants' states at instant $t - 1$. This reduces the number of free parameters to be estimated, to 2^K; in many cases, however, this may be insufficient and may lead to overfitting to the training data. A technique which could solve this problem is to tie the 2^K parameters, which follow a multinomial distribution, by further *parametrizing* them.

The current chapter explores this possibility, first analyzing the multinomially distributed parameters of the conditionally independent multi-participant model in terms of the degree of overlap of the conditioning state. It is shown that the log-probability of overlap is inversely proportional to the degree of overlap; in particular, it appears that logistic regression provides a good fit to the data. This yields a parsimonious model representation with only $K \cdot (K + 1)$ free parameters, which can be inferred using a variety of well-known techniques. Most notably, the parameters can be conceived of as the weights in a single-layer feed-forward neural network, as in the case of the Ising-Glauber model in statistical physics. Furthermore, the concept of the pseudo-temperature offers a convenient paradigm for expressing temporally local departures in a turn-taking entropy. The chapter concludes with a worked example demonstrating that, on average, manually annotated involvement hotspots exhibit higher turn-taking pseudo-temperatures than do other intervals of conversation.

7.2 Revisiting the Compositional Conditionally Independent Model

The compositional model which assumes conditional independence among participant states was given in Equation 6.17,

$$\begin{aligned} P(\mathbf{Q} \mid \mathbf{\Theta}) &= P_0 \prod_{t=1}^{T} P(\mathbf{q}_t \mid \mathbf{q}_{t-1}, \mathbf{\Theta}) \\ &= P_0 \prod_{t=1}^{T} \prod_{k=1}^{K} P(\mathbf{q}_t[k] \mid \mathbf{q}_{t-1}, k, \mathbf{\Theta}) \quad . \end{aligned} \tag{7.1}$$

As in Chapter 6, $\mathbf{q}_t[k] \in \mathbb{S} \equiv \{\square, \blacksquare\}$ and $\mathbf{q}_t \in \mathbb{S}^K \equiv \mathbb{S} \times \mathbb{S} \times \cdots \times \mathbb{S}$. To learn a model $\mathbf{\Theta}$ implies the inference of the parameters of $\mathbf{\Theta}$ given a $\mathbf{Q} \equiv \{\mathbf{q}_t\} \in \{\square, \blacksquare\}^{K \times T}$, for which all unigram and bigram counts can be easily computed.

To facilitate discussion, the number of times the bigram $(\mathbf{q}_{t-1} = \mathbf{S}_i, \mathbf{q}_t[k] = \blacksquare)$ occurs in $\mathbf{Q}$ is denoted as

$$n_k^i = \sum_{t=1}^{T} \delta(\mathbf{q}_{t-1}, \mathbf{S}_i)\, \delta(\mathbf{q}_t[k], \blacksquare) \quad , \tag{7.2}$$

where δ is the Kronecker delta extended to vector arguments. The number of times each unique conditioning unigram occurs is

$$n^i = \sum_{t=1}^{T} \delta(\mathbf{q}_{t-1}, \mathbf{S}_i) \quad . \tag{7.3}$$

Consequently, given only two alternatives for the value of $\mathbf{q}_t[k]$, the number of times the bigram $(\mathbf{q}_{t-1} = \mathbf{S}_i, \mathbf{q}_t[k] = \square)$ occurs is equal to $n^i - n_k^i$.

Allowing

$$y_k^i = P(\blacksquare \mid \mathbf{S}_i, k, \mathbf{\Theta}) \tag{7.4}$$

$$\equiv 1 - P(\square \mid \mathbf{S}_i, k, \mathbf{\Theta}) \tag{7.5}$$

permits rewriting Equation 7.1 as

$$\begin{aligned} P(\mathbf{Q} \mid \mathbf{\Theta}) &= P_0 \prod_{t=1}^{T} \prod_{k=1}^{K} P(\mathbf{q}_t[k] \mid \mathbf{q}_{t-1}, k, \mathbf{\Theta}) \\ &= P_0 \prod_{i=0}^{2^K-1} \prod_{k=1}^{K} \left(y_k^i\right)^{n_k^i} \left(1 - y_k^i\right)^{n^i - n_k^i} \\ &= P_0 \prod_{k=1}^{K} \prod_{i=0}^{2^K-1} \left[\left(y_k^i\right)^{p_k^i} \left(1 - y_k^i\right)^{1-p_k^i}\right]^{n_i} \quad , \end{aligned} \tag{7.6}$$

where $p_k^i = {}^{n_k^i}/_{n^i}$.

Equation 7.6 prescribes an obvious method for estimating $\mathbf{\Theta} \equiv \{y_k^i\}$. Since each argument in square brackets is binomially distributed, $P(\mathbf{Q} \mid \mathbf{\Theta})$ has a single supremum if all y_k^i, for $1 \leq k \leq K$ and $0 \leq i < 2^K$, are unconditionally independent. It is easily seen by differentiating $\log P(\mathbf{Q} \mid \mathbf{\Theta})$ that this supremum is attained when $y_k^i = p_k^i$, namely at the maximum likelihood values of the 2^K desired conditionally independent transition probabilities.

7.3 Ancillary Observations on Overlap

As argued in Chapter 6, 2^K may still be too many parameters to robustly estimate from $\mathbf{Q}$. For models assuming conditional dependence among participants, Chapter 6 considered overlap, or the degree of simultaneous vocalization, as an organizing principle for tractably collapsing the state space. A similar proposal is made here, with respect to conditionally independent participants.

In "Observations on Overlap [...]" [203], Shriberg and collegues reported that, in the subset of the ICSI Meeting Corpus which they studied, the proportion of talkspurts which contained some overlap was between 31.4% and 54.4% [203]. On a smaller time-scale, they found that the proportion of words exhibiting some overlap was between 8.8% and 17.0%. In subsequent work [42], Cetin & Shriberg showed that in a similar multi-party conversation corpus, collected at multiple sites, 11.6% of speech time was deployed in overlap, and that of all overlap time, 92.2% consisted of simultaneous vocalization by only 2 participants.

This section attempts to complement the latter finding, by first exploring proportions such as

$$P\left(\,\|\mathbf{q}_t\| = v\,\right) \quad = \quad \frac{\sum_{t=1}^{T} \delta\left(\|\mathbf{q}_t\|, v\right)}{T} \tag{7.7}$$

for any degree of simultaneous vocalization v. Figure 7.1 shows the distribution of these probabilities for all observed states in a single meeting on a logarithmic scale, i.e., $-\log_2 P\left(\mathbf{S}_i\right)$, $0 \leq i < 2^K$, as a function of the number of participants $v_i = \|\mathbf{S}_i\|$ which speak simultaneously in each state $\mathbf{S}_i$. The figure also shows $-\log_2 P\left(\|\mathbf{S}_i\| = v\right)$ as a function of v, the negative log-likelihood of being in *any* state with degree of overlap v, with broader, wider horizontal lines. What is surprising about this figure is that there appears to exist a strong *linear* relationship between v and $-\log_2 P\left(\|\mathbf{S}_i\| = v\right)$. This property suggests that $-\log_2 P\left(\|\mathbf{S}_i\| = v\right)$ can be predicted for a particular conversation, for any $v > 0$, from its values at just two points.

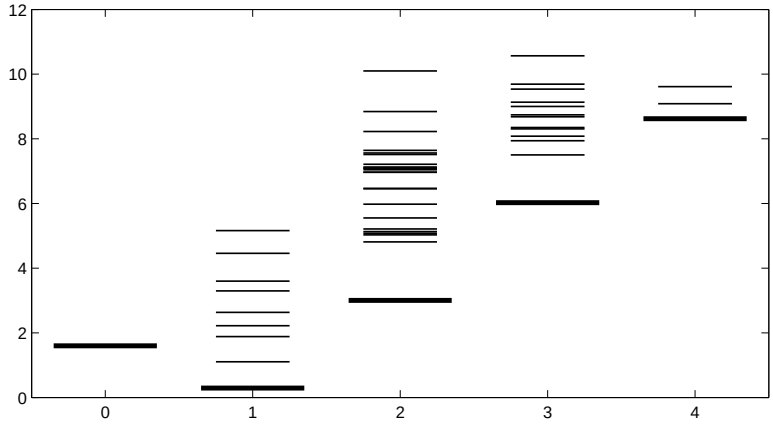


Figure 7.1: Probability of occupation of multiparticipant state $\mathbf{S}_i$ in meeting `Bmr025`, as a function of $v_i = \|\mathbf{S}_i\|$, the number of participants vocalizing in $\mathbf{S}_i$. Narrow lines represent the probability of each $\mathbf{S}_i$ individually; broad lines represent the probability of occupation of any state with v, i.e., $\sum_{i=1 \|\mathbf{S}_i\|=v} P\left(\mathbf{S}_i\right)$. Overlap degrees of $v > 4$ did not occur in the sample, and are not depicted.

The approximately linear relationship between $-\log_2 P\left(\|\mathbf{S}_i\| = v\right)$ on v begs the question of how overlap might be initiated

$$P\left(\,\mathbf{q}_t\left[k\right] = \blacksquare \,|\, \mathbf{q}_{t-1}\left[k\right] = \square, \|\mathbf{q}_{t-1}\| = v\,\right)$$

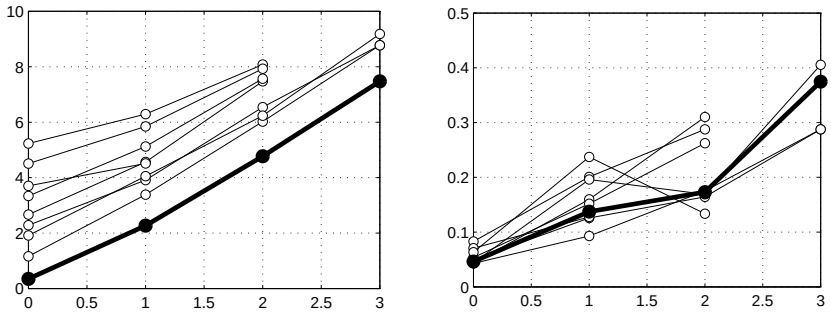
and terminated

$$P\left(\,\mathbf{q}_t\left[k\right] = \square \,|\, \mathbf{q}_{t-1}\left[k\right] = \blacksquare, \|\mathbf{q}_{t-1}\| = v\,\right) \quad .$$

Among the qualitative claims of conversation analysis is that "Talk by MORE than two at a time seems to be reduced to two (or to one) even more effectively than talk by two is reduced to one" [198]. This suggests that

$$\begin{aligned} &P\left(\,\mathbf{q}_t\left[k\right] = \blacksquare \,|\, \mathbf{q}_{t-1}\left[k\right] = \blacksquare, \|\mathbf{q}_{t-1}\| = v+1\,\right) \\ &\quad < \quad P\left(\,\mathbf{q}_t\left[k\right] = \blacksquare \,|\, \mathbf{q}_{t-1}\left[k\right] = \blacksquare, \|\mathbf{q}_{t-1}\| = v\,\right) \end{aligned} \tag{7.8}$$

for all k, $1 \leq k \leq K$, and all $v > 1$. Figure 7.2 shows that this is also borne out for the meeting depicted in Figure 7.1. The joint probability of someone continuing to talk, in the presence of a variable number v of other participants talking at the same time, is very nearly log-linear in v. The log-probability of interest, i.e. the *conditional* log-probability, is also quite close to linear in v.



(a) (b)

Figure 7.2: Negative log-probabilities (along the y-axis) of ■ at instant t following ■ at instant $t-1$, as a function of v, the number of other participants speaking at instant $t-1$ (along the x-axis). (a) depicts the joint bigram probability; (b) depicts the conditional bigram probability. Probabilities estimated from the same snippet as those shown in Figure 7.1.

Conversation analysis is somewhat less prescriptive about the converse [198], namely whether participants are less likely to initiate talkspurts in higher degrees of overlap,

$$\begin{aligned} & P\left(\mathbf{q}_t\left[k\right] = \blacksquare \,|\, \mathbf{q}_{t-1}\left[k\right] = \square, \|\mathbf{q}_{t-1}\| = v+1, \boldsymbol{\Theta}\right) \\ & \quad < \quad P\left(\mathbf{q}_t\left[k\right] = \blacksquare \,|\, \mathbf{q}_{t-1}\left[k\right] = \square, \|\mathbf{q}_{t-1}\| = v, \boldsymbol{\Theta}\right) \quad . \end{aligned} \tag{7.9}$$

Perhaps only coincidentally, the data suggests that the conditional log-probability of intending to increase the amount of overlap is not linear in v, as shown in Figure 7.3. Although beginning to speak at t when only one other participant has been speaking at instant $t-1$ is clearly less likely than doing so when exactly zero other participants have been speaking, that likelihood appears to increase as more and more participants are discovered to be already talking. This seems to indicate that in situations in which two or more persons are speaking, other participants perceive the floor to be "up for grabs". However, it is also possible that the lower neg-log-probabilities for $v > 1$ are indicative of "perceived" overlap rather than actual overlap, such as when occurs when two participants competing for the floor are overlapping dialog acts but not words.

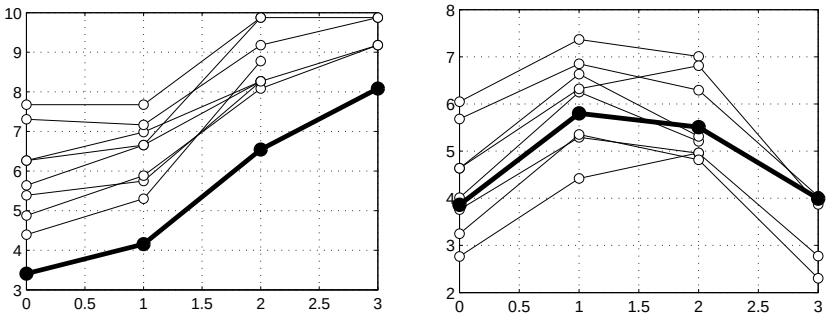
It should be kept in mind that overlap is an emergent phenomenon, measured only after it happens. However, the probabilities shown in Figures 7.2 and 7.3 only measure the degree of overlap at $t-1$. If multiple participants are allowed to simultaneously change state, initiation of overlap has the same representation in this framework as precisely timed speaker change.

7.4 Logistic Regression

Figures 7.2(b) and 7.3(b) depict probabilities on a logarithmic scale, which renders the slopes of any potential linear fit of different magnitude. A linear fit on the log-scale may also yield probabilities which are outside of the $[0, 1]$ interval. A more felicitous mapping, which leads to slopes of similar magnitude and ensures that the inverse mapping produces elements in $[0, 1]$, is the logistic function. Figure 7.4 depicts the logit-probabilities of Figures 7.2(b) and 7.3(b).

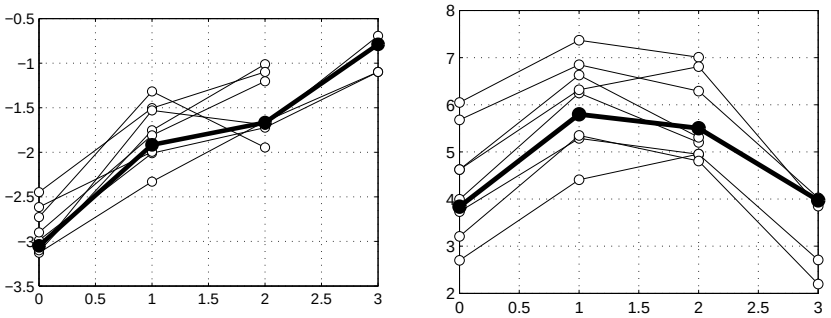
The slopes in the two panels of Figure 7.4 are seen to be much more similar in magnitude than those in Figure 7.2(b) and Figure 7.3(b)[1]. If in fact they are sufficiently similar for a weighted linear fit to warrant a single slope parameter w_{k-},

[1] A linear fit to the dark lines in both panels is overwhelmingly governed by the two points at $v = 0$ and $v = 1$, for which there is much more weight. Higher levels of overlap are much less frequent.



(a) (b)

Figure 7.3: Negative log-probabilities (along the y-axis) of ■ at instant t following □ at instant $t-1$, as a function of v, the number of other participants speaking at instant $t-1$ (along the x-axis). (a) depicts the joint bigram probability; (b) depicts the conditional bigram probability. Probabilities estimated from the same snippet as shown in Figure 7.1.



(a) (b)

Figure 7.4: Negative logit-probabilities (along the y-axis) of ■ at instant t following □ at instant $t-1$, in panel (a), and following ■ at instant $t-1$, in panel (b), as a function of v, the number of other participants speaking at instant $t-1$ (along the x-axis). Probabilities estimated from the same snippet as shown in Figure 7.1.

linear approximation may be achieved by

$$
\begin{aligned}
&-\log \frac{P\left(\mathbf{q}_t\left[k\right]=\blacksquare \,|\, \mathbf{q}_{t-1}, k, \boldsymbol{\Theta}\right)}{1-P\left(\mathbf{q}_t\left[k\right]=\blacksquare \,|\, \mathbf{q}_{t-1}, k, \boldsymbol{\Theta}\right)} \\
&\qquad = \begin{cases} b_k + w_{k-}\|\mathbf{q}_{t-1}\| & \text{if } \mathbf{q}_{t-1}\left[k\right]=\square \\ b'_k + w_{k-}\|\mathbf{q}_{t-1}\| & \text{otherwise} \end{cases} ,
\end{aligned} \tag{7.10}
$$

where b_k and b'_k are the y-intercept values for the curves describing talkspurt initiation and continuation, respectively. If □ is equated with zero, and ■ with unity, then Equation 7.10 can be rewritten as

$$
\begin{aligned}
&-\log \frac{P\left(\mathbf{q}_t\left[k\right]=\blacksquare \,|\, \mathbf{q}_{t-1}, k, \boldsymbol{\Theta}\right)}{1-P\left(\mathbf{q}_t\left[k\right]=\blacksquare \,|\, \mathbf{q}_{t-1}, k, \boldsymbol{\Theta}\right)} \\
&\qquad = \; b_k + w_{k+}\mathbf{q}_{t-1}\left[k\right] + \sum_{j\neq k}^{K} w_{k-}\mathbf{q}_{t-1}\left[j\right]
\end{aligned} \tag{7.11}
$$

if, additionally, $b'_k = b_k + w_{k+}$. The right-hand side is easily expressed, using matrix notation, as the kth entry of the

vector

$$\mathbf{h}_t \;=\; \mathbf{b} + \mathbf{W} \cdot \mathbf{q}_{t-1} \;\;, \tag{7.12}$$

where

$$\mathbf{b} \;=\; \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_K \end{bmatrix} \quad \text{and} \quad \mathbf{W} \;=\; \begin{bmatrix} w_{1+} & w_{1-} & \cdots & w_{1-} \\ w_{2-} & w_{2+} & \cdots & w_{2-} \\ \vdots & \vdots & \ddots & \vdots \\ w_{K-} & w_{K-} & \cdots & w_{K+} \end{bmatrix} \;\;. \tag{7.13}$$

The converse of Equation 7.10 given these notational simplifications is given by

$$\begin{aligned} P\left(\mathbf{q}_t\left[k\right] = \blacksquare \,|\, \mathbf{q}_{t-1}, k, \boldsymbol{\Theta}\right) \;&=\; \frac{1}{1 + e^{-\mathbf{h}_t[k]/\mathcal{T}}} \\ &=\; \text{sigmoid}\left(-\mathbf{h}_t\left[k\right]/\mathcal{T}\right) \\ &=\; \text{logit}^{-1}\left(-\mathbf{h}_t\left[k\right]/\mathcal{T}\right) \;\;. \end{aligned} \tag{7.14}$$

In this section, $\mathcal{T}$ — known as the *pseudo-temperature* — is a constant whose value is unity.

A widely used algorithm which can be used to estimate $\mathbf{b}$ and $\mathbf{W}$ as required in Equation 7.12 is known as *reweighted least squares* [220].

7.5 A One-Layer Feed-Forward Neural Network

An alternative conceptualization of the problem of inferring the parameters $\mathbf{b}$ and $\mathbf{W}$ in Equation 7.12 is to treat them as the biases and weights, respectively, in a one-layer perceptron whose activation function is the sigmoid. Any standard neural network training technique can be used, including those relying on gradient descent. This is most often accomplished in the logarithm of Equation 7.6,

$$\begin{aligned} &\log P\left(\mathbf{Q} \,|\, \boldsymbol{\Theta}\right) \;= \\ &\qquad \log P_0 + \sum_{k=1}^{K} \sum_{i=0}^{2^K-1} n_i \left[p_k^i \log\left(y_k^i\right) + \left(1 - p_k^i\right) \log\left(1 - y_k^i\right)\right] \end{aligned} \tag{7.15}$$

by differentiating with respect to any parameter $\theta \in \{b_k, w_{k+}, w_{k-}\}$,

$$\frac{\partial \log P\left(\mathbf{Q} \,|\, \boldsymbol{\Theta}\right)}{\partial \theta} \;=\; \sum_{k=1}^{K} \sum_{i=0}^{2^K-1} n_i \left[\frac{p_k^i}{y_k^i} - \frac{1 - p_k^i}{1 - y_k^i}\right] \frac{\partial y_k^i}{\partial \theta} \;\;. \tag{7.16}$$

Equation 7.15 is known as the *cross-entropy error*, applicable for feed-forward neural networks whose output layer represents posterior likelihoods of multiple independent attributes [18]. The attributes are of course only conditionally independent, given the patterns presented at the input layer.

Parameters $\mathbf{b}$ and $\mathbf{W}$ define the sought-after model $\boldsymbol{\Theta}$. Equation 7.13 represents a form of parameter tying; in the general case,

$$\mathbf{b} \;=\; \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_K \end{bmatrix} \quad \text{and} \quad \mathbf{W} \;=\; \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1K} \\ w_{21} & w_{22} & \cdots & w_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ w_{K1} & w_{K2} & \cdots & w_{KK} \end{bmatrix} \;\;. \tag{7.17}$$

In this case, the model $\boldsymbol{\Theta} \equiv \{\mathbf{b}, \mathbf{W}\}$ consists of $K \cdot (K+1)$ free parameters, as opposed to $3 \cdot K$ free parameters given the more constrained form in Equation 7.13. In contrast, the number of free parameters in the non-parametric compositional model is 2^K.

Alternately, when the amount of data is small, the parameters of the model can be further tied to yield an even more constrained (and participant-independent) model,

$$\mathbf{b} = \begin{bmatrix} b \\ b \\ \vdots \\ b \end{bmatrix} \quad \text{and} \quad \mathbf{W} = \begin{bmatrix} w_+ & w_- & \cdots & w_- \\ w_- & w_+ & \cdots & w_- \\ \vdots & \vdots & \ddots & \vdots \\ w_- & w_- & \cdots & w_+ \end{bmatrix} . \tag{7.18}$$

In this case, the model $\mathbf{\Theta} \equiv \{\mathbf{b}, \mathbf{W}\} = \{b, w_+, w_-\}$ consists of only 3 free parameters.

7.6 The Ising Anti-Ferromagnet

The neural network formulation suggests a strong similarity of the proposed model of conversational dynamics with models used to study physical ensembles of interacting particles [167]. The original application of the latter, due to Lenz and Ising [101], was the description of magnetism. Figure 7.5 shows a 2-dimensional lattice of atoms with spin either "up" or "down". Connections, to nearest neighbors only, have a weight w; there are no self-connections. An external magentic field is shown with strength b.

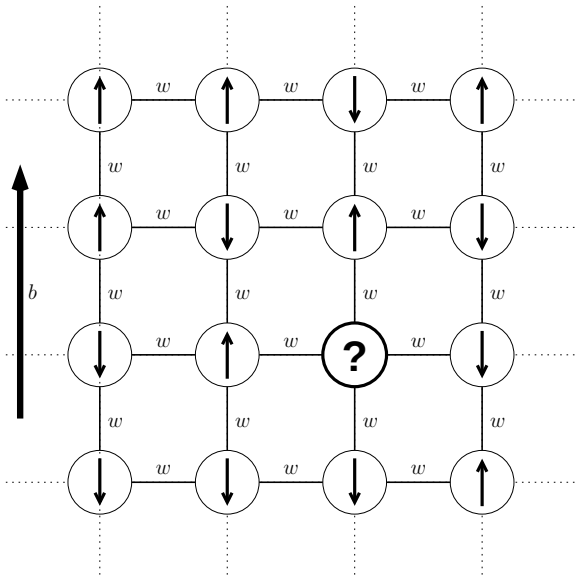


Figure 7.5: A depiction of a two-dimensional lattice, with spins either only up or down. Bi-directional connections, with weights w, among nearest neighbors only; as a result, the site identified as "?" will be updated, at the next instant, based only on the spin of its four neighbors and the external magentic field shown pointing upwards with magnitude b.

In contrast to the discussion in Section 7.5, the connection weights w are given by the known physical properties of the lattice. $\mathbf{W}$ has zeros along the diagonal as well as for any row k and column j when site j is not an immediate neighbor of site k. To infer a *steady state* for such a lattice, the spins of all sites are initialized with random values $\in \{\downarrow, \uparrow\}$, and a solution is reached asymptotically by iteratively visiting sites and sampling their future state according to Equation 7.14.

It should be stressed that the notion of "iteration" does not represent evolution in time, but rather incremental numerical improvement in the estimation of a steady state. The symmetry of $\mathbf{W}$ in such problems guarantees the existence of a *static* energy surface. Optionally *quenching* the system refers to lowering the pseudo-temperature from one iteration to the next according to some prescribed policy, which may be viewed as either incremental estimate improvement or a temporally extensive physical process. Ising models have been used to successfully predict physical macro-properties such as the temperature at which a ferromagnet loses its magnetism.

Statistical physics has also had an impact on many computational problems outside of the physical domain; a famous example is autoassociative memory. A *Hopfield network* is an artifact which can memorize a fixed number of reference images, and then retrieve the nearest reference image when presented with a noisy image [95]. In such systems, $\mathbf{b} \equiv \mathbf{0}$, and $\mathbf{W}$ must be inferred from the set of reference images, for example via *Hebbian learning* [90, 92]. The diagonal entries of $\mathbf{W}$ are zero, reflecting the absence of self-connections, and $w_{kj} = w_{jk}$ to guarantee a static energy surface. Asides from these constraints, however, $\mathbf{W}$ can have any form. Retrieval is achieved by initializing the trained network with a noisy image, and allowing it to relax via Equation 7.14 to a steady state by updating one pixel at a time (referred to as *Hopfield dynamics*) or all pixels at each iteration (referred to as *Little dynamics* [163, 164, 165]). Temperature quenching, also known as *simulated annealing*, is used to avoid entrapment in spurious states which do not correspond to any reference image attractor. An example of a 2-dimensional image with horizontally, vertically, and diagonally connected neighbors, is shown in Figure 7.6.

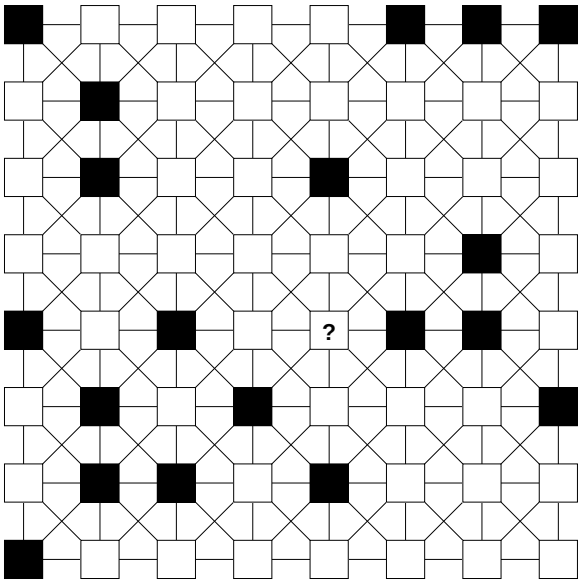


Figure 7.6: A depiction of a 2-dimensional image, with pixels either on or off. Bi-directional connections, with learned weights, among nearest horizontal, vertical, and diagonal neighbors only; as a result, the pixel identified as "?" will be updated, at the next instant, based only on the value of its 8 neighbors.

In contrast to these models of static configurations, a dynamic model which is intended to describe a sequence in time need not have symmetric $\mathbf{W}$. Physically, an asymmetric $\mathbf{W}$ corresponds to the absence of a static energy surface with static minima. Furthermore, $\mathbf{W}$ need not have zeros along the diagonal (these are referred to as "slow connections" in sequence modeling). A representation of the proposed model of conversational dynamics is shown in Figure 7.7.

As Equations 7.12 and 7.14 show, the probability that participant k speak at time t tends to zero as $\mathbf{h}_t[k] \to -\infty$, and to unity as $\mathbf{h}_t[k] \to +\infty$. When $\mathbf{h}_t[k] = 0$, the probability that participant k speak is $1/2$.

Given these considerations, b_k represents the *inhibition to break silence* by participant k; it is likely to be negative for sub-second frame rates, and is correlated with how likely participant k is to terminate the state in which all participants are silent. Participants who more frequently terminate mutual silence are expected to have negative b_k values which are closer to zero.

Parameters w_{kk} can be interpreted as each participant's *continuation potential*, or inertia. These parameters are likely to be positive, and larger in absolute magnitude than b_k. The higher the value of w_{kk}, the longer are continuous intervals in which only participant k vocalizes.

In contrast, w_{kj}, $j \neq k$, represent the *competive inhibition* exercised by participant j on the incipient behavior of participant k. These values are expected to be negative. Larger values yield lower likelihoods that participant k will vocalize at instant t given that participant j is vocalizing at instant $t-1$.

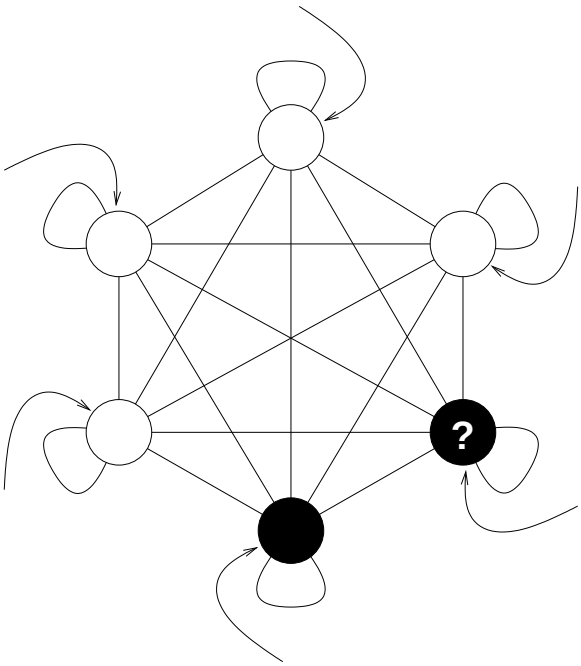


Figure 7.7: A depiction of a conversation of 6 participants, with participants either vocalizing ■ or not □. Bidirectional connections, with asymmetrical weights in each direction, among all participants; as a result, the state of the participant identified as "?" will be updated, at the next instant, based on their own past state, the states of all other participants, and their own "inhibition to break silence" shown with a source-less arrow terminating at each participant's state.

$\mathbf{h}_t$, given by Equation 7.12, can be taken to represent the *vocalization potential* of each participant. As Equation 7.12 expresses, this quantity is the sum of each participant's inhibition to break silence, their own continuation potential, or inertia, if they are already vocalizing, and the competitive inhibition exerted by other participants who are already vocalizing.

7.7 Pseudo-Temperature

Viewing the proposed parametric form $\boldsymbol{\Theta}$ as a neural network with an interpretation borrowed from statistical mechanics has an important additional implication. The preceding discussion has assumed that the pseudo-temperature $\mathcal{T}$ in Equation 7.12 is a constant, yielding "thermal equilibrium" for the duration of an observed conversation.

A potentially useful characterization of the temporal evolution of a conversation, following the inference of a time-independent model $\boldsymbol{\Theta} = \{\mathbf{b}, \mathbf{W}\}$ under the assumption of fixed $\mathcal{T} = 1$, is to subsequently estimate a time-dependent $\mathcal{T}_t$ under the assumption of fixed $\mathbf{b}$ and $\mathbf{W}$. Equation 7.12 suggests that as the pseudo-temperature drops, i.e. $\mathcal{T} \to 0$, the probability that participant k vocalizes at instant t moves closer to its extrema of zero or unity (whichever is closer at $\mathcal{T} = 1$). In contrast, for high temperatures $\mathcal{T} \to +\infty$, that probability tends to $1/2$. In this sense, for $\mathcal{T} > 1$, the probability that participant k vocalizes at instant t becomes a non-linear interpolation between what Equation 7.12 predicts at $\mathcal{T} = 1$ and a fair Bernoulli trial $P(\mathbf{q}_t[k]) = 1/2$ (i.e. randomness). $\mathcal{T}$ is therefore a measure of deviation from an uninterpolated model norm.

7.8 An Example

To illustrate some of the techniques proposed in this chapter, time-independent $\mathbf{b}$ and $\mathbf{W}$ and time-dependent $\mathcal{T}$ are inferred from a single conversation in the ICSI Meeting Corpus, namely `Bmr025`. This meeting lasts approximately 30 minutes and involves 8 participants. The first half of its speech interaction chronogram $\mathbf{Q}$ is shown in Figure 7.8. For the example of this section, it is discretized at a frame step of 10 ms.

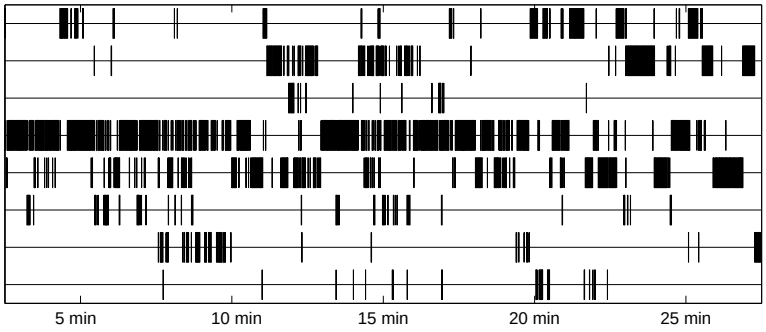


Figure 7.8: Speech interaction chronograph $\mathbf{Q}$ for approximately half of meeting `Bmr025` in the ICSI Meeting Corpus; its 8 participant tracks are shown along the y-axis.

To infer $\mathbf{b}$ and $\mathbf{W}$, the variables are initialized to Gaussian-distributed random values. Gradient descent in Equation 7.15 is performed using the delta-bar-delta technique; this is a standard first-order method which employs a momentum term, independently for each parameter, to accelerate descent [18]. The biases are found to be

$$\mathbf{b} = \begin{bmatrix} -6.63 \\ -6.61 \\ -8.59 \\ -5.25 \\ -5.77 \\ -7.38 \\ -7.22 \\ -8.09 \end{bmatrix} . \tag{7.19}$$

All b_k are seen to be negative, as expected. Also, as suggested in the preceding section, the smallest-magnitude b_k, for $k \in \{4, 5\}$, are found for those participants which begin speaking in silence most frequently (this is not evident for participant $k = 4$ in Figure 7.8, who speaks the most, because of the temporal scale of the diagram). The largest-magnitude b_k, for $k \in \{3, 8\}$, is found for the two participants who appear to speak the least frequently in Figure 7.8. The amount of speech produced by a participant is strongly correlated with the number of times that participant breaks mutual silence, which also occurs if they pause and continue without interlocutor interjection.

The neural network weights $\mathbf{W}$ are found to be

$$\begin{bmatrix} 12.15 & -1.04 & -2.73 & -1.24 & -1.35 & -1.38 & -2.03 & -1.40 \\ -1.20 & 11.69 & -0.44 & -1.19 & -0.57 & -0.84 & -0.85 & -0.70 \\ -3.34 & -0.03 & 13.43 & -0.86 & -0.96 & -0.84 & -2.73 & -5.36 \\ -1.09 & -1.23 & -0.54 & 10.83 & -1.16 & -0.76 & -1.07 & -0.76 \\ -1.14 & -1.12 & -0.42 & -1.60 & 11.29 & -0.78 & -1.45 & -1.83 \\ -1.00 & -0.70 & -0.82 & -0.66 & -1.12 & 12.52 & -0.85 & -0.21 \\ -2.11 & -1.13 & -2.19 & -0.96 & -1.22 & -1.42 & 12.54 & -0.53 \\ -0.60 & -1.82 & \mathbf{0.93} & -1.24 & -1.00 & \mathbf{0.47} & \mathbf{0.17} & 12.99 \end{bmatrix} \tag{7.20}$$

As expected, all diagonal entries are positive, while off-diagonal entries are negative: speaking entails an inertia by the speaker to continue speech once having started; already speaking interlocutors have an inhibitory effect on both initiating

and continuing speech. However, three off-diagonal entries of $\mathbf{W}$ violate this property, and are identified in bold. Their positive sign, and relatively small magnitude, indicate that either there is not enough data to infer these values robustly in this single meeting — and in fact participant $k = 8$ speaks the least — and/or that the 8th participant is actually more likely to speak when interlocutors $k \in \{3, 6, 7\}$ are speaking. The latter may be the case if the 8th participant tends to predominantly backchannel with those 3 interlocutors.

To address the problem of poor parameter estimation for rarely speaking individuals, it may be preferable to first infer a participant-independent model as in Equation 7.18, and then to initialize $\mathbf{b}$ and $\mathbf{W}$ to the resulting participant-independent values, rather than random values, during inference of a participant-dependent model. Positive off-diagonal $\mathbf{W}$ entries should then appear only if a participant in fact speaks predominantly in overlap with others.

To illustrate the variability in time of the pseudo-temperature, $\mathcal{T}$ is estimated every 15 seconds within a 60-second window. The biases $\mathbf{b}$ and weights $\mathbf{W}$ are clamped to their time-independent values, as inferred above. This yields a trajectory of perturbation about the value $\mathcal{T} = 1$, shown in Figure 7.9.

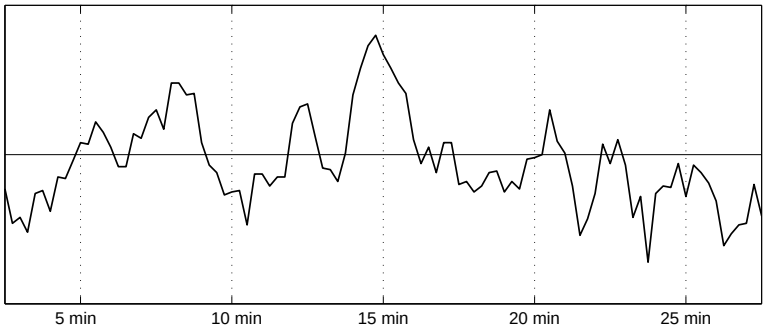


Figure 7.9: The trajectory of the pseudotemperature, re-estimated every 15 seconds, for the same snippet of conversation as shown in Figure 7.8.

It is possible to qualitatively assess the utility of measuring these perturbations, by appealing to manually annotated phenomena. The ICSI Meeting Corpus transcriptions are accompanied by a labeling of "hot spots" (cf. Chapter 4), defined as intervals in which speech is prosodically marked as involved, or — of interest here — there is floor contention. The temperature trajectory of Figure 7.9 is overlaid with these intervals, depicted in gray, in Figure 7.10.

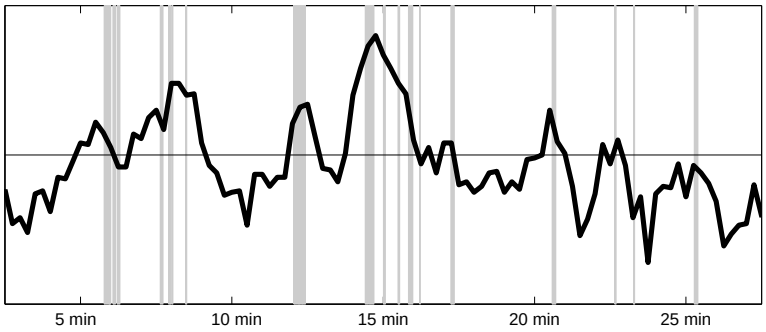


Figure 7.10: The trajectory of the pseudotemperature as in Figure 7.9, with intervals annotated as involvement hotspots depicted in gray.

It is apparent, at least for this meeting, that local maxima in the pseudo-temperature trajectory are correlated with intervals containing speech annotated as belonging to hotspots. Meeting `Bmr025` contains 127 equi-spaced 60-second intervals, 15 seconds apart. Of these, 62 contain some speech which lies within an annotated hotspot (an unusually

high ratio for the corpus as a whole). Training a one-Gaussian model for the pseudo-temperature for all 60-second hotspot-containing frames, and another for all 60-second non-hotspot-containing frames, results in the diagram shown in Figure 7.11.

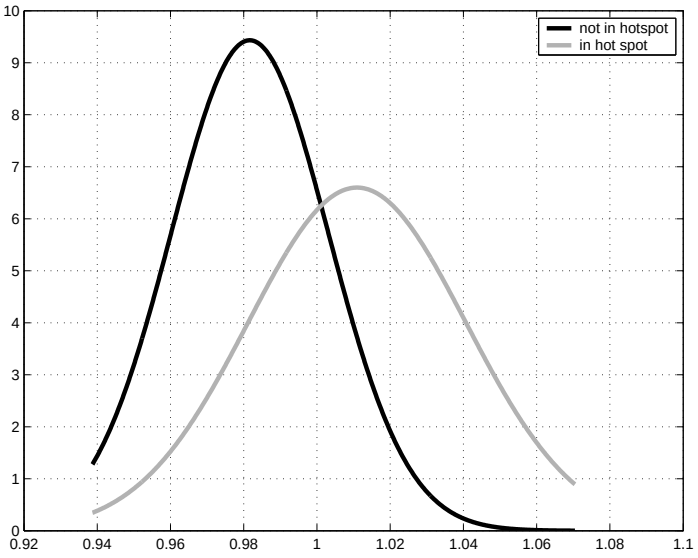


Figure 7.11: Single-Gaussian models over the temperature of intervals containing an annotated hotspot or hotspot fragment and over that of intervals not containing an annotated hotspot or hotspot fragment, for `Bmr025`.

As Figure 7.11 shows, the means of the temperatures of the two subpopulations of intervals are sufficiently far apart to allow for some discrimination of hotspots using these methods. However, for the majority of meetings in the corpus, the prior over the two interval types is skewed such that the hotspot interval curve in Figure 7.11 lies below the non-hotspot interval curve, even when the means are distinct.

7.9 Relevance to Other Chapters

The ideas presented here are directly related to those of the previous chapter. Parametrizing the $\boldsymbol{\Theta}$ transition probabilities yields much more compact models than in Chapter 6, which may be trained using fewer data than direct compositional models. It is also much more tractable to extend these compact models beyond the 1st-order Markov assumption or to multiple types of vocal activity simultaneously.

The probabilities of initializing talkspurts in silence and in the context of other participants speaking, as well as continuing talkspurts, are used as features for classifying participants in Chapters 16 and 17. Because of the way they are estimated, the entries of $\mathbf{b}$ and $\mathbf{W}$ yield potentially more robust values than maximum likelihood estimates, and the experiments of Chapters 16 and 17 suggest that for some applications this may be more appropriate.

The experiments in Chapter 15 which aim to detect conversational hotspots make use of the pseudo-temperature measure.

7.10 Summary

This chapter has explored a means of modeling the dynamics of specific conversations with even fewer parameters than the degree-of-overlap (EDO) model, with conditionally dependent participant states, of the previous chapter. The innovation consists of two operations. First, participant states are assumed to be conditionally independent, given the joint multi-participant states of the previous instant. The implicit extension of the EDO model consists of conditioning each

participant's binary state not just on that participant's earlier binary state (as in unconditionally independent models), but also on the number of other participants previously in the vocalizing state. Second, this chapter has shown that the prior likelihood of multi-participant states differing in the specific degree of overlap has an approximately log-linear relationship with the degree of overlap. This makes it felicitous to model the $K + 1$ free parameters characterizing each participant with only 3 parameters. The resulting model turns out to be related to one which has been extensively applied to study ferromagnetism over the last 8 decades. The concept of the pseudo-temperature offers a simple time-dependence to the otherwise time-independent model, and appears to elegantly quantify the concept of re-occurring metaphorical (turn-taking) "hotness" in multi-party conversation.

Chapter 8

Parametric Feature-Space Multi-Participant Models*

8.1 Introduction

Modeling interaction among conversation participants in state space, under the assumption that all participants' future states are either conditionally dependent,

$$P(\mathbf{Q}) = P_0 \prod_{t=1}^{T} P(\mathbf{q}_t \,|\, \mathbf{q}_{t-1}) \tag{8.1}$$

or conditionally independent

$$P(\mathbf{Q}) = P_0 \prod_{t=1}^{T} \prod_{k=1}^{K} P(\mathbf{q}_t[k] \,|\, \mathbf{q}_{t-1}) \ , \tag{8.2}$$

given all participants' past states, requires the enumeration of multi-participant states, as discussed in Chapter 6. The number of such states grows geometrically with the number of participants K. This property circumscribes the deployment of state-space models to only those settings in which there are few participants or few degrees of freedom per participant.

In the general case of arbitrarily large numbers of participants to a conversation, and potentially large numbers of degrees of freedom per participant, Equations 8.1 and 8.2 may not be implementable. The alternative assumption that all participants' future states are unconditionally independent,

$$P(\mathbf{Q}) = P_0 \prod_{t=1}^{T} \prod_{k=1}^{K} P(\mathbf{q}_t[k] \,|\, \mathbf{q}_{t-1}[k]) \tag{8.3}$$

is implementable, but of course completely ignores interaction among participants.

It is thereby seen that multi-participant state-space models are either potentially intractable or interaction-insensitive. Modeling interaction in large per-participant degree-of-freedom settings calls for a different approach. In this chapter, the proposed alternative consists of of conditioning the multi-participant state trajectory $\mathbf{Q}$ — which will be denoted $\mathbf{Y}$ to signal too many degrees of freedom per participant to model as in previous chapters — on an auxillary multi-participant trajectory $\mathbf{Q}$, of same dimensions as $\mathbf{Y}$ but of fewer degrees of freedom. If $\mathbf{Q}$ is a deterministic many-to-one mapping of $\mathbf{Y}$, i.e., $P(\mathbf{Q} \,|\, \mathbf{Y}) \equiv 1$, then

$$P(\mathbf{Y}) = P(\mathbf{Y}, \mathbf{Q}) \tag{8.4}$$

$$= P(\mathbf{Y} \,|\, \mathbf{Q})\, P(\mathbf{Q}) \ . \tag{8.5}$$

*The work in this chapter was conducted in collaboration with Liz Shriberg.

Furthermore, if the collapsed multi-participant state space allows only a small number of degrees of freedom per participant, then $P(\mathbf{Q})$ may be estimated using state-space modeling techniques without ignoring interaction, as was done in Equations 8.1 and 8.2.

Provided that the salient interactive aspects of $\mathbf{Y}$ are adequately accounted for through $P(\mathbf{Q})$, the first term in Equation 8.5 may assume that participants' future states in $\mathbf{Y}$ are unconditionally independent of one another as in Equation 8.3, but conditioned on $\mathbf{Q}$, namely that

$$\begin{aligned} P(\mathbf{Y} \mid \mathbf{Q}) &= P(\mathbf{Y}[1], \mathbf{Y}[2], \cdots, \mathbf{Y}[K] \mid \mathbf{Q}) && (8.6) \\ &= \prod_{k=1}^{K} P(\mathbf{Y}[k] \mid \mathbf{Q}) && (8.7) \\ &= \prod_{k=1}^{K} P_0 \prod_{t=1}^{T} P(\mathbf{y}_t[k] \mid \mathbf{y}_{t-1}[k], \mathbf{Q}) \ . && (8.8) \end{aligned}$$

Given these assumptions, the likelihood of an observed $\mathbf{Y}$ can be obtained by forming $\mathbf{Q}$, computing $P(\mathbf{Q})$ via any of the means described in Chapter 6, and then evaluating Equation 8.5 using a suitable form of Equation 8.8. If $\mathbf{Y}$ is not observed, then it may be inferred from $\mathbf{Q}$ via

$$\begin{aligned} \mathbf{Y}^* &= \arg\max_{\mathbf{Y}} P(\mathbf{Y} \mid \mathbf{Q}) && (8.9) \\ &= \arg\max_{\mathbf{Y}} P(\mathbf{Q} \mid \mathbf{Y}) \ P(\mathbf{Y}) \ . && (8.10) \end{aligned}$$

This chapter is concerned with the identification of suitable forms for $P(\mathbf{Q} \mid \mathbf{Y})$, needed for the inference task. Since $\mathbf{Q}$, the multi-participant collapsed-state version of $\mathbf{Y}$, is assumed observable, models of $P(\mathbf{Q} \mid \mathbf{Y})$ may approximate it using $P(\mathbf{f}(\mathbf{Q}) \mid \mathbf{Y})$, where $\mathbf{f}(\mathbf{Q})$ is a feature vector computed from $\mathbf{Q}$. Furthermore, if the values of $\mathbf{f}(\cdot)$ are observed to follow a known distribution, a parametric model of $P(\mathbf{f}(\mathbf{Q}) \mid \mathbf{Y})$, consisting only of the parameters of that distribution, may be much easier to learn than the alternative non-parametric model.

8.2 Rotating and Windowing Snapshots of Neighborhood

As proposed, the techniques of this chapter will treat participants as unconditionally independent of one another in $\mathbf{Y}$, but conditioned on one another in $\mathbf{Q}$. When inferring the $\mathbf{y}_t[k]$ trajectory of a participant k, that participant will be referred to as the *target participant*; other participants will be alternately referred to as *non-target participants* or *interlocutors* (of the target participant). When inferring the trajectory of all K participants, in sequential order, each participant k will become the target participant in turn.

Similarly to but more formally than in Equation 8.8, under a model $\mathbf{\Theta}$,

$$P(\mathbf{Y} \mid \mathbf{Q}, \mathbf{\Theta}) = \prod_{k=1}^{K} P(\mathbf{Y}[k] \mid \mathbf{R}_k \cdot \mathbf{Q}, \mathbf{\Theta}) \ . \quad (8.11)$$

The rotation operator $\mathbf{R}_k$, which is a $K \times K$ identity matrix with columns 1 and k exchanged, rotates that row of $\mathbf{Q}$ which corresponds to the target participant k into the first row. In the absence of this operator, the correspondence between $\mathbf{Y}[k]$ and the kth row of $\mathbf{Q}$ (as in Equation 8.8, where the operator was dropped for readability) would be lost.

Equation 8.11 may be decomposed using Bayes' rule,

$$\begin{aligned} P(\mathbf{Y} \mid \mathbf{Q}, \mathbf{\Theta}) &= \prod_{k=1}^{K} P(\mathbf{Y}[k] \mid \mathbf{R}_k \cdot \mathbf{Q}, \mathbf{\Theta}) \\ &= \prod_{k=1}^{K} \frac{P(\mathbf{R}_k \cdot \mathbf{Q} \mid \mathbf{Y}[k], \mathbf{\Theta}_q) \, P(\mathbf{Y}[k] \mid \mathbf{\Theta}_y)}{\sum_{\mathbf{Y}[k]} P(\mathbf{R}_k \cdot \mathbf{Q} \mid \mathbf{Y}[k], \mathbf{\Theta}_q) \, P(\mathbf{Y}[k] \mid \mathbf{\Theta}_y)} && (8.12) \\ &\equiv \prod_{k=1}^{K} \frac{1}{Z_k} \cdot P(\mathbf{R}_k \cdot \mathbf{Q} \mid \mathbf{Y}[k], \mathbf{\Theta}_q) \cdot P(\mathbf{Y}[k] \mid \mathbf{\Theta}_y) \ . && (8.13) \end{aligned}$$

$\boldsymbol{\Theta}_y$ and $\boldsymbol{\Theta}_q$ are implicitly defined as those parts of $\boldsymbol{\Theta}$ which model transition and emission probabilities, respectively. Equation 8.13 replaces the denominator in Equation 8.12 with a k-dependent but $\mathbf{Y}$-independent constant Z_k, for notational convenience.

The second probability factor in Equation 8.13 is of course the unconditionally independent, single-participant transition model of Subsection 6.3.3. It can be further decomposed, as earlier,

$$P\left(\mathbf{Y}\left[k\right]\mid\boldsymbol{\Theta}_y\right) \doteq P_0\prod_{t=1}^{T}P\left(\mathbf{y}_t\left[k\right]\mid\mathbf{y}_{t-1}\left[k\right],\boldsymbol{\Theta}_y\right) \tag{8.14}$$

The first probability factor in Equation 8.13 is previously not discussed in this thesis, and is the subject of the current chapter. It is the likelihood of the multi-participant observations $\mathbf{Q}$ given the state trajectory of one participant k. It too can be decomposed in time,

$$\begin{aligned} P\left(\mathbf{R}_k\cdot\mathbf{Q}\mid\mathbf{Y}\left[k\right],\boldsymbol{\Theta}_q\right) &= \prod_{t=1}^{T}P\left(\mathbf{R}_k\cdot\mathbf{Q}\cdot\mathbf{W}_t\mid\mathbf{y}_1\left[k\right],\mathbf{y}_2\left[k\right],\cdots,\mathbf{y}_T\left[k\right],\boldsymbol{\Theta}_q\right) \\ &= \prod_{t=1}^{T}P\left(\mathbf{R}_k\cdot\mathbf{Q}\cdot\mathbf{W}_t\mid\mathbf{y}_t\left[k\right],\boldsymbol{\Theta}_q\right) \end{aligned} \tag{8.15}$$

$\mathbf{W}_t$ is a generalized windowing operator which selects a neighborhood around instant t. For example, it may select a neighborhood which consists only of the column vector of $\mathbf{Q}$ at instant t, $\mathbf{q}_t$. Equation 8.15 assumes that, whatever the definition of that neighborhood, snapshots of it at consecutive instants t are conditionally independent of the target speaker's state trajectory given only her state at instant t. While most convenient, this assumption is violated when the size of the neighborhood implemented in $\mathbf{W}_t$ is large.

Figure 8.1 depicts a short interval of a $K = 6$ participant conversation. The interval identically demonstrates the effect of applying a $\mathbf{W}_t$ whose size of neighborhood is 7 frames before and after the current instant t. For simplicity in the figure, $\mathbf{q}_t\left[k\right] \in \{0,1\} \equiv \{\square,\blacksquare\}$; these states may be considered to represent the obvious binary distinction between speech and non-speech.

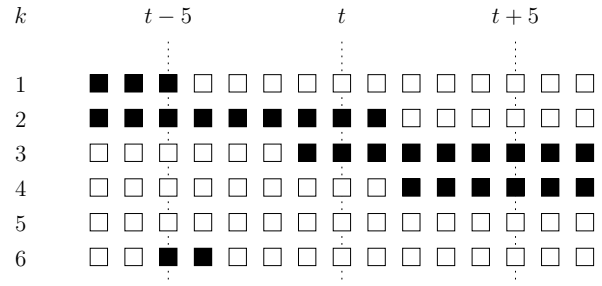


Figure 8.1: A fictitious example of $\mathbf{Q}\cdot\mathbf{W}_t$, a windowed snippet of a conversation with collapsed multi-participant state trajectory $\mathbf{Q}$. $\mathbf{Q}$ may be the observed speech/non-speech posterior, of dimensions of time along the x-axis and participant index k along the y-axis. $\mathbf{W}_t$ selects a symmetric window of 7 frames before and after the decoding instant t.

Figure 8.2 shows how the operator $\mathbf{R}_k$ modifies $\mathbf{Q}\cdot\mathbf{T}_k$, when preparing to compute features characterizing $\mathbf{y}_t\left[k\right]$ for each target participant k. The observables pertaining to the target participant are rotated into the first row; interlocutor observables are then found in rows 2 through K.

8.3 Computing Durations to Observable Landmarks

A most direct means of modeling the local neighborhood around t, for target participant k, is to measure the difference between t and temporal locations of unambiguously identifiable landmarks in $\mathbf{Q}$. Landmarks may include specific behaviors

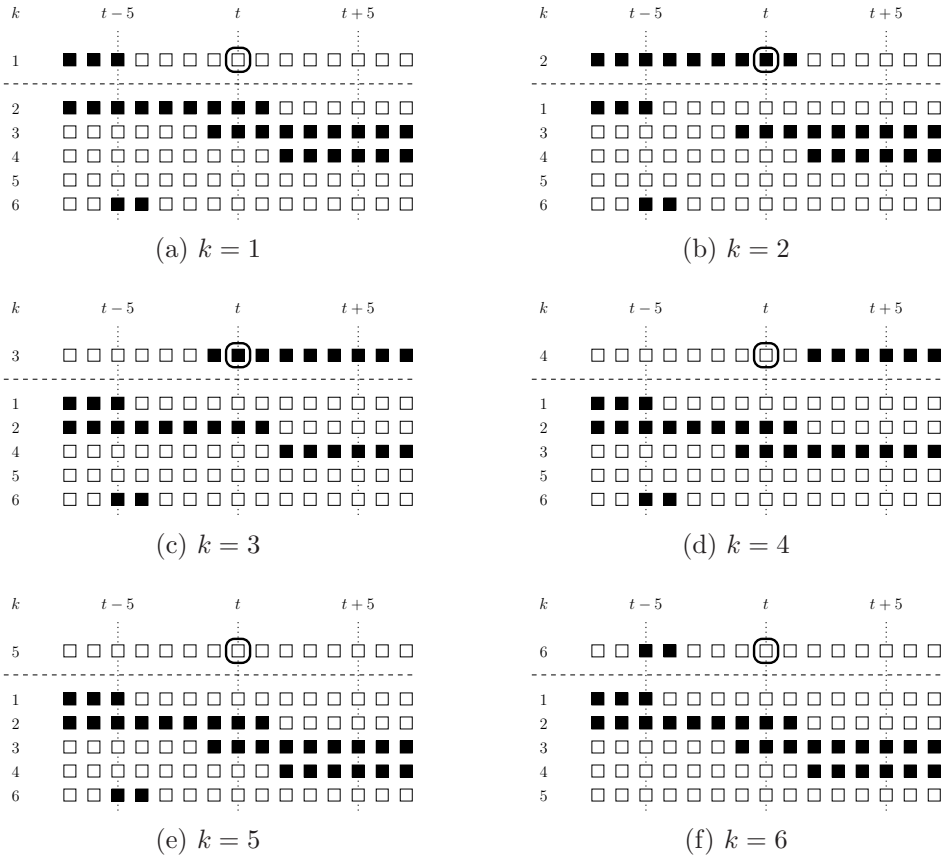


Figure 8.2: Application of $\mathbf{R}_k$ to the fictitious $\mathbf{Q} \cdot \mathbf{W}_t$ in Figure 8.1, yielding $\mathbf{R}_k^T \cdot \mathbf{Q} \cdot \mathbf{W}_t$. Time shown along the x-axis, participant index k along the y-axis.

on the part of the target participant or an interlocutor, or specific transitions among behaviors. In the ensuing discussion, as in the previous section, observables $\mathbf{q}_t[k]$ are assumed to be drawn from $\{0, 1\} \equiv \{\square, \blacksquare\}$, and may be assumed to represent a posterior which has been compared to a threshold. The generalization to any univariate or multivariate observable is straightforward.

A reasonable set of features which can be computed from the target participant's observables $\mathbf{Q}[k]$ includes:

1. the number of frames to the temporally most proximate $\square$ at $t' < t$;
2. the number of frames to the temporally most proximate $\square$ at $t' > t$;
3. the number of frames to the temporally most proximate $\blacksquare$ at $t' < t$; and
4. the number of frames to the temporally most proximate $\blacksquare$ at $t' > t$.

Similarly, reasonable features to compute for the target participant's context include:

1. the number of *other* participants in $\blacksquare$ at t;
2. the number of frames to the temporally most proximate $\blacksquare$ from *any other* participant at $t' < t$; and
3. the number of frames to the temporally most proximate $\blacksquare$ from *any other* participant at $t' > t$.

For interlocutors in multi-party conversation, it is not generally of interest to model the number of frames to the temporally most proximate □ from any participant. This is because it may be assumed, particularly when ■ represents speech activity, that there are always some participants in □.

Features such as the above are used ubiquitously in conversation understanding tasks. Their convenience lies in the fact that feature vectors are fixed in length, as they are independent of the number K of participants in the conversation under study. However, they suffer from three rather serious drawbacks.

First, they are not robust to errors in the estimated observable $\mathbf{Q}$. If, in the middle of a long interval of □, a single frame of ■ appears — nominally an error of small consequence — it will have a large impact on the numerical value of those features which represent duration to next or previous ■, over a long interval of t.

Second, the features described above are not sensitive to the *duration* of intervals of □ or ■. Since they only measure the duration to the *nearest* □ or ■, they confound snapshots which are intrinsically quite different, particularly when participants exhibit a lot of transitions between □ and ■. This can of course be remedied by extending the proposed feature set to encode not only the most subsequent (or precedent) □, but also the most subsequent (or precedent) □ occurring *after* (or *before*) the most subsequent (or precendent, respectively) ■ alternative, in any given participant's row of $\mathbf{Q}$. However, since participants may generally exhibit different rates of transition among □ and ■, modeling a fixed number of boundaries per participant may actually need to consider neighborhoods of markedly different extent across participants.

Finally, considering *any other* participants when computing features from interlocutor rows of $\mathbf{Q}$, in the way proposed above, confounds interlocutors and eliminates edges exhibited by any particular interlocutor. Characterizing interlocutors individually is of course problematic, because a conversation may involve any number K of participants which would lead to a variable feature vector size. $\mathbf{Q}$ can be extended by any number of virtual, non-existent participants to comply with a fixed maximum K_{max}, but computing features for individual interlocutors would still remain sensitive to interlocutor rotation in the same way that direct state-space models were argued to be in Chapter 6.

In summary, features representing durations to landmarks, such as specific observable values or specific differences in observable values, may not be robust enough for inference of $\mathbf{Y}$ and lead to seemingly incomplete descriptions of the local neighborhood. Systematic extension to complete those descriptions appears to lead to new sources of variability which would require additional extension.

8.4 Representing the Complete Neighborhood Snapshot

An alternative which addresses these three concerns consists of modeling the entire snapshot, such as shown in Figure 8.1, as a two-dimensional image. The x-axis of the image, representing the temporal extent of the snapshot, is given by a time-independent width of analysis window licensed by the windowing operator $\mathbf{W}_t$ (the output of the operator is not time-independent, since for different t it selects different partial snapshots of $\mathbf{Q}$, but the width of those partial snapshots is independent of time).

Representing the complete neighborhood snapshot is, however, at odds with the requirement to be both K-invariant and $\mathbf{R}$-invariant, as defined in Chapter 2. In particular, conversations with different K will entail feature vectors of different sizes, since the number of interlocutors — the extent of the neighborhood snapshot along the y-axis, less one — is given by $K-1$. Additionally, snapshots will of course be significantly different when any two participants' indexes in $\mathbf{Q}$ are exchanged. Even after the rotation operator R_k, which places the target participant in row 1, the multiplicity in representation for an identical conversational phenomenon is still $(K-1)!$.

8.4.1 Interlocutor Rotation via Ranking

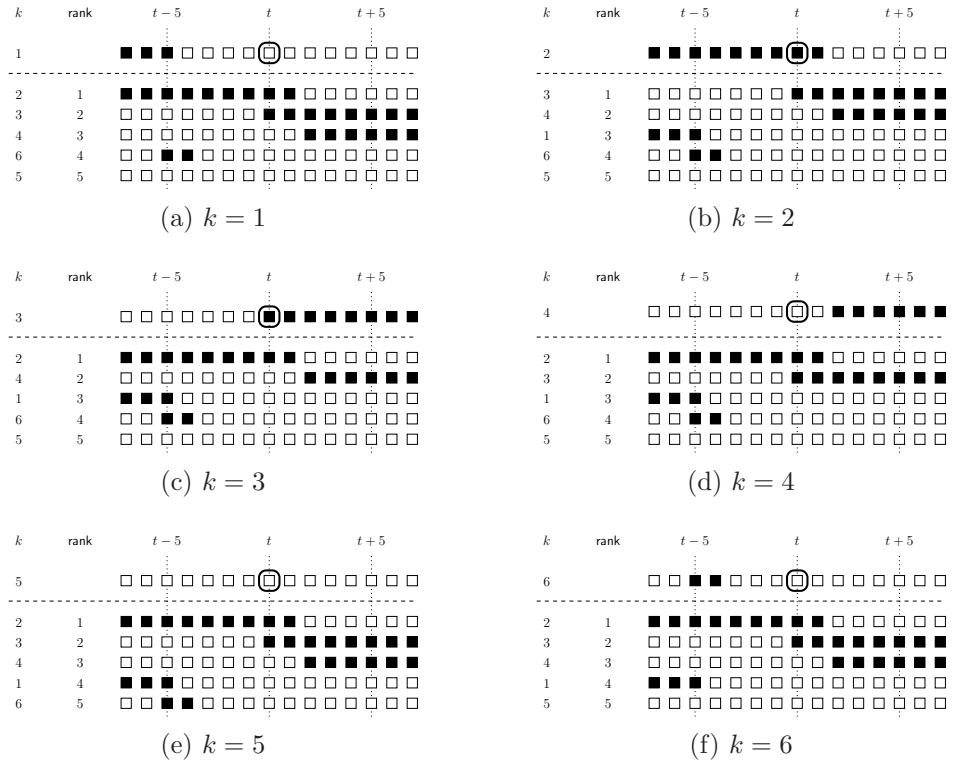
The solution proposed in this thesis, to the problem of $\mathbf{R}$-dependence for feature-space modeling of interaction, is to apply a second rotation $\mathbf{R}_q$ which rotates participant tracks in $\mathbf{Q}$ given a deterministic ranking. The function performing the ranking remains to be specified, but it must meet the following two requirements:

1. The snapshot characterizing instant t, when inferring the state trajectory $\mathbf{y}_t[k]$ of participant k in a conversation whose observable is $\mathbf{Q}$, will be invariant under arbitrary interlocutor-index rotation of $\mathbf{Q}$.

2. The ranking function will make use of no information beyond $\mathbf{Q}$ or its dimensions K and T.

A variety of ranking functions can be immediately envisioned. Complying with the second requirement, a simple option is to rank participants by the sum of their ■ posteriors (whether $\in \{0, 1\}$ or $\in [0, 1]$), over some interval defined by an auxillary generalized trunctation operator $\mathbf{W}_t^{rank}$. $\mathbf{W}_t^{rank}$ may be identical to $\mathbf{W}_t$, in which case participants are ranked by their local amount of ■, but it need not be. Allowing $\mathbf{W}_t^{rank}$ to include the conversation of interest in its entirety would entail ranking participant by their global amount of ■.

Ranking only applies to interlocutors, since the target participant is rotated by $\mathbf{R}_k$ into the first row of $\mathbf{Q}$. If $\mathbf{W}_t^{rank}$ is not global, then the resulting $\mathbf{R}_q$ is a function of t; it will generally be denoted $\mathbf{R}_q\left(\mathbf{W}_t^{rank}\right)$. The transformation applied to $\mathbf{Q}$, then, to yield a $\mathbf{R}$-invariant neighborhood snapshop is $\mathbf{R}_q\left(\mathbf{W}_t^{rank}\right) \cdot \mathbf{R}_k \cdot \mathbf{Q} \cdot \mathbf{W}_t$. Figure 8.3 shows the same conversational snippet as does Figure 8.2, following interlocutor ranking by the proposed method with $\mathbf{W}_t^{rank} = \mathbf{W}_t$.



(a) $k = 1$ (b) $k = 2$

(c) $k = 3$ (d) $k = 4$

(e) $k = 5$ (f) $k = 6$

Figure 8.3: Application of $\mathbf{R}_q\left(\mathbf{W}_t^{rank}\right)$ to the fictitious $\mathbf{R}_k^T \cdot \mathbf{Q} \cdot \mathbf{W}_t$ in Figure 8.2, yielding $\left(\mathbf{R}_q\left(\mathbf{W}_t^{rank}\right)\right)^T \cdot \mathbf{R}_k^T \cdot \mathbf{Q} \cdot \mathbf{W}_t$. Time shown along the x-axis, participant index k along the y-axis.

8.4.2 Interlocutor Truncation or Padding

The rotation operator $\mathbf{R}_q$ defined in the previous subsection induces a ranking of interlocutors along the y-dimension of the neighborhood snapshot. Assuming that the posterior available in $\mathbf{Q}$, within the temporal extent licensed by $\mathbf{W}_t^{rank}$, is related to the degree of influence that an interlocutor may have on the target participant's state $\mathbf{y}_t[k]$, the rotation effectively "bubbles up" influential interlocutors towards the top of the snapshot. The requirement of K-invariance can then be met by either: (1) truncating those rows which are found past a minimum K_{max} of interest; or (2) padding $\mathbf{Q}$ with □ entries for virtual, non-existent participants up to some maximum K_{max}, representing the anticipated true K of the most populous conversation groups.

8.4.3 Extracting Features from the Neighborhood Snapshot

Given a fixed-size neighborhood snapshot, which is both K- and $\mathbf{R}$-invariant, feature extraction leading to a fixed-size feature vector is relatively straightforward, and requires only the specification of feature granularity. Features can of course be the individual posteriors found within the temporal support licensed by $\mathbf{W}_t$, but they can also be general functions over groups of those posteriors. This thesis considers only *averages* over tiles, encompassing one participant at a time but with temporal extent between 1, corresponding to single frames of observables in $\mathbf{Q}$, to the full temporal extent of the neighborhood snapshot.

Tiles consisting of more than one frame allow for a controllable measure of plesiochrony in models constructed over such features. However, because it is expected that the precise degree of overlap (when ■ corresponds to speech activity) may be important at t, the decoding instant, frames at t for all interlocutors are allocated their own tiles. A graphical depiction of this arrangement, for tiles of duration 0.3 s when the frame step is 0.1 s, is shown in Figure 8.4.

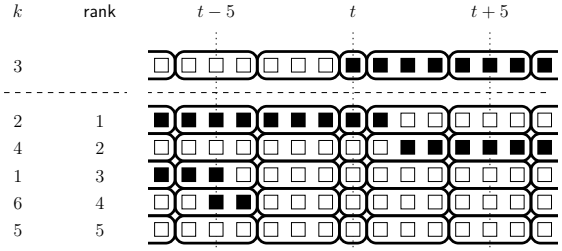


Figure 8.4: 0.3-second regions used to tile the neighborhood snapshot corresponding to panel (c) in Figure 8.3 for which $k = 3$; frames occurring at instant t are not grouped with other frames. Each tile yields one feature, the average ■ posterior in this thesis. Time shown along the x-axis, participant index k along the y-axis.

8.4.4 Modeling Feature Density

Given a neighborhood snapshot which is fixed in temporal extent and independent of the number K of participants, each instant t can be characterized by a fixed-length feature vector as described in Subsection 8.4.3. It remains to propose a convenient and useful generative statistical model of the distribution of such vectors.

In this thesis, the chosen form is the multivariate Gaussian mixture model (GMM). It may be argued, since $\mathbf{Q}$ consists of posteriors $\in \{0, 1\}$ or $\in [0, 1]$, that the appropriate form for each element of the feature vector has a univariate Binomial distribution, and that therefore what most concisely describes the complete vector is a multivariate Binomial mixture model. In spite of this, the GMM is widely applied, well understood, and has been shown to capably approximate any distribution, given a sufficient number of mixtures. Furthermore, Gaussian mixture modeling is typically applied after some form of globally inferred rotation of feature vectors, inducing a weighted linear combination of the features. Such combination has the effect of gaussianizing the input space.

8.5 Relevance to Other Chapters

The modeling techniques proposed in this chapter find application in scenarios in which the complete multi-participant state space from which $\mathbf{y}_t$ are drawn is too large to tractably enumerate. An example of such a scenario is the recognition not just of speech versus non-speech activity, but the simultaneous characterization of contrastive types of speech activity.

Chapter 14 explores the applicability of the proposed techniques to the problem of recognizing dialog acts. This task entails the simultaneous segmentation of participant verbal behavior into dialog acts and its classification into several classes of illocutionary intent. The problem is not directly tractable using state-space modeling due to the large number of classes which each participant can in principle occupy, as well as the various minimum duration constraints which discriminate among classes. In Chapter 14, the collapsed-state multi-participant trajectory $\mathbf{Q}$, assumed to be observable (i.e., inferred in an earlier processing step), is the multi-participant speech versus non-speech chronogram.

The same techniques are also applied to the inference of attempts at humor or sarcasm, in Chapter 15. There, two distinct binary-valued $\mathbf{Q}$ are compared and combined. One is the multi-participant speech versus non-speech chronogram, as in Chapter 14; the other is a chronogram of laughter versus non-laughter.

8.6 Summary

This chapter has proposed a solution to the general problem of inferring multi-participant behavior trajectories, in conversations of arbitrary participant number and arbitrary number of degrees of freedom per participant. It has been argued that trajectory inference in this potentially intractable state space may be achieved by treating single-participant trajectories as mutually independent but conditioning them on an auxillary trajectory drawn from a collapsed state space in which participants are not independent.

This proposed solution requires that the auxillary trajectory $\mathbf{Q}$ be treated as observable. Emission probability models of the likelihood of $\mathbf{Q}$, conditioned on the unobserved trajectory $\mathbf{Y}$ which is to be inferred, may then be trained using standard statistical methods and applied in the general hidden Markov modeling paradigm.

For the inferred $\mathbf{\Theta}$ to be independent of the index assignment of participants in $\mathbf{Q}$ once trained, and for the models to be deployable in conversations of arbitrary participant number, feature extraction from the observable $\mathbf{Q}$ must be preceded by a pre-processing algorithm. This chapter has proposed an algorithm controlled by several parameters which may be empirically estimated beforehand for each task of interest. In the case where the entries of $\mathbf{Q}$ are posteriors, drawn from $\{0,1\} \equiv \{\square, \blacksquare\}$, the pre-processing steps are summarized in Algorithm 4.

Algorithm 4 Compute emission probability $P(\mathbf{Q} \,|\, \mathbf{Y}, \mathbf{\Theta}_q)$.

Require: A t-independent ranking window $\mathbf{W}_t^{rank}$.
Require: A t-independent feature extraction window $\mathbf{W}_t$.
Require: K_{max} the maximum number of interlocutors to be models.
Require: A tiling policy $\mathcal{T}$.
Require: K, the number of participants.
Require: T, the number of frames.
for all $k \in \{1, 2, \ldots, K\}$ **do**
  Rotate participant k into row 1 of $\mathbf{Q}$, yielding $\mathbf{R}_k^T \cdot \mathbf{Q}$.
  for all $t \in \{1, 2, \cdots, T\}$ **do**
    Construct window operator $\mathbf{W}_t^{rank}$ for instant t.
    for all $j \in \{2, \cdots, K\}$ **do**
      Compute the average posterior in row j of $\mathbf{R}_k^T \cdot \mathbf{Q} \cdot \mathbf{W}_t^{rank}$
    end for
    Rank interlocutors j in order of decreasing average posterior magnitude.
    Define a rotation $\mathbf{R}_q$ based on average posterior rank.
    Rotate interlocutors in $\mathbf{R}_k^T \cdot \mathbf{Q} \cdot \mathbf{W}_t$, yielding $\mathbf{R}_q \cdot \left(\mathbf{R}_k^T \cdot \mathbf{Q} \cdot \mathbf{W}_t\right)$.
    Tile $\mathbf{R}_q \cdot \left(\mathbf{R}_k^T \cdot \mathbf{Q} \cdot \mathbf{W}_t\right)$ according to $\mathcal{T}$.
    Form feature vector of average posteriors for each tile.
    Optionally rotate feature vector.
    Compute rotated feature vector likelihood given parametric $\mathbf{\Theta}_q$.
  end for
end for

The algorithm as shown is equally applicable to scenarios in which the entries of $\mathbf{Q}$ are probabilistic, drawn from $[0, 1]$; in particular, "hard" inference of $\mathbf{Q}$ is not a requirement.

Chapter 9

Parametric Feature-Space Multi-Channel Models*

9.1 Introduction

While preceding chapters in this part of the thesis have been concerned with modeling the states of multiple participants, nothing has been said about modeling the acoustics of multiple channels which correspond to those participants. This is the role of the current chapter. Whether participant states are modeled independently or jointly, there is a need during recognition for expressing the likelihood of the observed multi-channel acoustics as a function of the multi-participant state, i.e. $P(\mathbf{X} \mid \mathbf{Q}, \mathbf{\Theta})$, where $\mathbf{X}$, $\mathbf{Q}$, and $\mathbf{\Theta}$ are the observable sequence, the (hidden) state sequence, and the model, respectively.

As can be imagined given the discussion in earlier chapters, modeling multi-channel acoustics is complicated by the fact that different conversations can involve different numbers of participants, leading to the problem of an essentially variable feature vector (across training and testing, for example). A possible solution to this problem is to factor the acoustic model, by assuming that

$$\begin{aligned} P(\mathbf{X} \mid \mathbf{Q}, \mathbf{\Theta}) &= \prod_{t=1}^{T} P(\mathbf{x}_t \mid \mathbf{q}_t, \mathbf{\Theta}) \\ &\doteq \prod_{t=1}^{T} \prod_{k=1}^{K} P(\mathbf{x}_t[k] \mid \mathbf{q}_t[k]) \end{aligned} \tag{9.1}$$

This assumption of unconditional independence in the acoustic model can be combined with transition models which do not make that assumption.

Unfortunately, the assumption of independence is inappropriate, because features across channels are correlated. Crosstalk, the main problem for vocal activity detection in multiple head-mounted microphone settings, is merely the name for a particularly strong correlation: the phenomenon of the same speech appearing on another participant's microphone. As a result, acoustic likelihoods in these settings should ideally rely on explicitly or implicitly decorrelated channels. Furthermore, multi-channel acoustic models have to address the possibility that the gains of each microphone may be different, and that the ambient acoustics of enclosures in which conversations take place may affect microphones in different and unpredictable ways.

In this thesis, the proposed solution to this problem is to train multichannel acoustic models on data which is maximally similar to the test data, including an identical number of participants, identical gains, and identical room response characteristics. This can only be tractably achieved by training on the test data itself. Since the test data is unlabeled, models trained using this data must rely on some preliminary pass which uses weaker models. A potential problem with this approach is that the test data may be quite short. Even with a near-perfect first pass, models trained using the test

*The work in this chapter was conducted in collaboration with Jens Edlund, Mattias Heldner, Tanja Schultz, and Matthias Wölfel.

data exclusively would then be severly undertrained, and may produce a final pass hypothesis which is worse than the first pass.

To address these concerns, this chapter describes a technique for assigning labels in an initial pass, which is completely model-free. The technique first estimates a matrix of participant-microphone distances, for all participants and all microphones, using a particular normalization of the cross-channel correlation spectrum maximum. It is then shown that these estimates can be used to decide whether a particular participant is currently speaking. The chapter also proposes several ways to train a multi-channel acoustic model following such initial label assignment, even when the data is short. As Chapter 11 will show, models trained in this way significantly outperform standard factored acoustic models.

A final acoustic modeling issue dealt with in this chapter is the computation of prosodic features from multi-channel audio. Modeling multi-channel acoustics requires that features be extracted synchronously from all channels. This presents some complications, particularly for pitch, since the latter is normally computed only after speech is detected and segmented. Since they are anchored to word boundaries, pitch features are unlikely to be synchronously extractable across channels. This makes standard pitch-derived features cumbersome for some applications, such as segmentation. Furthermore, in settings in which participants are not known but need not be recognized, pitch needs to be speaker-normalized. To simplify multichannel processing, this chapter describes a means of estimating variation in fundamental frequency, but bypassing the estimation of fundamental frequency itself. The resulting feature vector, which in contrast to pitch is always defined, can be modeled in much the same way as spectral envelope features, either independently per speaker channel or in models over channel-concatenated feature vectors.

9.2 Non-Target-Normalized Maximum Cross-Channel Correlation

A hypothetical arrangement of $K = 3$ participants is shown in Figure 9.1. Also shown in this figure are the $K = 3$ head-mounted microphones worn by these participants. The locations of the microphones M_k, $1 \leq k \leq K$, are denoted $\mathbf{m}_k \in \mathbb{R}^3$, and their response is denoted $m_k(t)$. The mouth of each participant is assumed to be a point source S_k, and is found at location $\mathbf{s}_k \in \mathbb{R}^3$.

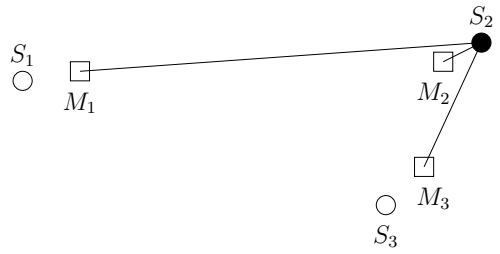


Figure 9.1: Hypothetical depiction of 3 participants' mouths S_k, $1 \leq k \leq J$, and the three head-mounted microphones M_k worn by them. Lines indicate distances from one of the sources, S_2, to each of the 3 microphones.

Four rather standard assumptions are made about the acoustic conditions shown in the figure:

Assumption 1 *Every source S_j is omnidirectional.*

Assumption 2 *Every microphone M_k is omnidirectional.*

Assumption 3 *The positions of all sources S_j and all microphones M_k are stationary over an observation interval Ω.*

Assumption 4 *Sound propagation is spherical and all surfaces have zero impedance.*

Given these assumptions, the response of any of the microphones can be approximated [15] by

$$m_k(t) \doteq G_k \left(\sum_{j=1}^{J} \frac{1}{d_{jk}} s_j \left(t - \frac{d_{jk}}{c} \right) + n_k(t) \right) , \tag{9.2}$$

where G_k is the stationary, omnidirectional gain of microphone M_k, $d_{jk} \equiv \|\mathbf{s}_j - \mathbf{m}_k\|$ is the distance between source S_j and microphone M_k, c is the speed of sound, and $n_k(t)$ is a lumped-term, additive, Gaussian zero-mean noise.

In the remainder of this derivation, it will also be assumed that:

Assumption 5 *During any observation interval Ω, there exists one and only one $j = j^*$ for which $s_j(t) \neq 0$ for some $t \in \Omega$.*

This assumption is justified because both multi-participant silence and speech overlap account for less than 10% of time in the data treated in this thesis (cf. Chapter 10); laughter, which accounts for just under 10% of vocalization effort (cf. Chapter 12) frequently occurs in overlap, making its proportion of meeting time much smaller than 10%. Other vocalization types are much rarer still, and breathing (which may be audible) is here considered to be noise, lumped in with the noise term $n_k(t)$. This allows Equation 9.2 to be rewritten as

$$m_k(t) \doteq G_k \left[\frac{1}{d_{j^*k}} s_{j^*} \left(t - \frac{d_{j^*k}}{c} \right) + n_k(t) \right] . \tag{9.3}$$

The cross-channel correlation for any pair of microphones M_k and $M_{k'}$ is given by

$$\begin{aligned}
\phi_{kk'}(\tau) &\doteq \int_\Omega m_k(t)\, m_{k'}(t-\tau)\, dt && (9.4)\\
&= \int_\Omega G_k \left(\frac{1}{d_{j^*k}} s_{j^*} \left(t - \frac{d_{j^*k}}{c} \right) + n_k(t) \right) \\
&\qquad \cdot\ G_{k'} \left(\frac{1}{d_{j^*k'}} s_{j^*} \left(t - \frac{d_{j^*k'}}{c} - \tau \right) + n_{k'}(t) \right) dt && (9.5)\\
&= \frac{G_k G_{k'}}{d_{j^*k} d_{j^*k'}} \int_\Omega s_{j^*} \left(t - \frac{d_{j^*k}}{c} \right) s_{j^*} \left(t - \frac{d_{j^*k'}}{c} - \tau \right) dt \\
&\qquad +\ G_k G_{k'} \int_\Omega n_k(t)\, n_{k'}(t)\, dt\ . && (9.6)
\end{aligned}$$

Only the supremum that $\phi_{kk'}(\tau)$ attains is of interest. Letting

$$\mathscr{S}_{j^*} \equiv \int_\Omega s_{j^*}^2(t)\, dt\ , \tag{9.7}$$

$$\mathscr{N}_k \equiv \int_\Omega n_k^2(t)\, dt \tag{9.8}$$

represent the signal and noise power, respectively, the desired supremum is given by

$$\max_\tau \phi_{kk'}(\tau) = \begin{cases} G_k^2 \left(\frac{1}{d_{j^*k}^2} \mathscr{S}_{j^*} + \mathscr{N}_k \right) & \text{if } k = k' \\ \frac{G_k G_{k'}}{d_{j^*k} d_{j^*k'}} \mathscr{S}_{j^*} & \text{otherwise} \end{cases} , \tag{9.9}$$

since

$$\arg\max_\tau \phi_{kk'}(\tau) = \begin{cases} 0 & \text{if } k = k' \\ \frac{d_{j^*k} - d_{j^*k'}}{c} & \text{otherwise} \end{cases} . \tag{9.10}$$

It should be noted that the quantities G_k, $G_{k'}$, d_{j^*k}, $d_{j^*k'}$, $\mathscr{S}_{j^*}$, $\mathscr{N}_k$, and $\mathscr{N}_{k'}$ are not directly observable; however, estimates of $\max_\tau \phi_{kk'}(\tau)$ can be readily obtained via the inner product in the discrete Fourier domain, for both $k = k'$ and for every pair (k, k') when $k \neq k'$.

The proposed non-target-normalized maximum cross-channel correlation $\xi_{kk'}$ is defined as

$$\begin{aligned}
\xi_{kk'} &= \frac{\max_\tau \phi_{kk'}(\tau)}{\max_\tau \phi_{k'k'}(\tau)} \\
&= \frac{\frac{G_k G_{k'}}{d_{j^*k} d_{j^*k'}} \mathscr{S}_{j^*}}{G_{k'}^2 \left(\frac{1}{d_{j^*k'}^2} \mathscr{S}_{j^*} + \mathscr{N}_{k'} \right)} \\
&= \frac{d_{j^*k'}}{d_{j^*k}} \cdot \frac{G_k}{G_{k'}} \cdot \frac{\mathscr{S}_{j^*}}{\mathscr{S}_{j^*} + d_{j^*k'}^2 \mathscr{N}_{k'}} \\
&= \frac{d_{j^*k'}}{d_{j^*k}} \cdot \frac{G_k}{G_{k'}} \cdot \left[1 - \frac{\mathscr{N}_{k'}}{\frac{1}{d_{j^*k'}^2} \mathscr{S}_{j^*} + \mathscr{N}_{k'}} \right] .
\end{aligned} \tag{9.11}$$

It is a descriptor of microphone k, obtained by normalizing the peak in the *cross-channel correlation* spectrum between microphone channel k and some other microphone channel $k' \neq k$ by the peak in the *autocorrelation* spectrum of the non-target channel k'.

It can be seen that if the second factor in Equation 9.11

$$\frac{G_k}{G_{k'}} \approx 1 , \tag{9.12}$$

namely that the gains of microphones M_k and $M_{k'}$ are approximately equal, and if

$$\frac{1}{d_{j^*k'}^2} \mathscr{S}_{j^*} \gg \mathscr{N}_{k'} , \tag{9.13}$$

namely that the power from S_{j^*} observed at microphone $M_{k'}$ is large compared to the power of the Gaussian noise component at that microphone (leading to a value for the third factor in Equation 9.11 of approximately unity), then

$$\xi_{kk'} = \frac{\max_\tau \phi_{kk'}(\tau)}{\max_\tau \phi_{k'k'}(\tau)} \approx \frac{d_{j^*k'}}{d_{j^*k}} . \tag{9.14}$$

This is a somewhat surprising result. It states that, even though the locations of both microphones and of the source are not known, the ratio between the two microphones' distances to the source can be approximated using the cross-correlation spectrum maximum.

9.3 "Model-Free" Multichannel Label Assignment

The non-target-normalized maximum cross-channel correlation $\xi_{kk'}$ can be used for initial label assignment in multichannel audio; two methods are presented.

In a first variant, it is assumed that — because microphones are head-mounted — the distance $\|\mathbf{s}_k - \mathbf{m}_k\|$ from the (only) speaker's mouth S_k to her own microphone M_k is smaller than the distance from that speaker's mouth to any other microphone $M_{k'}$, for $k' \neq k$,

$$d_{kk} < d_{kk'} , \quad \forall\, k' \neq k \tag{9.15}$$

This entails

$$d_{kk} < \min_{k' \neq k} d_{kk'} \tag{9.16}$$

or, identically, that

$$\begin{aligned}
\frac{\min_{k' \neq k} d_{kk'}}{d_{kk}} &= \min_{k' \neq k} \frac{d_{kk'}}{d_{kk}} \\
&= \min_{k' \neq k} \xi_{kk'} \\
&> 1 ,
\end{aligned} \tag{9.17}$$

since distances are non-negative.

This allows for the estimation of a label $\mathbf{q}[k]$, describing the speech activity state of participant k, during the analyzed interval Ω:

$$\mathbf{q}[k] \;=\; \begin{cases} \blacksquare & \text{if } \min\limits_{k' \neq k} \log \xi_{kk'} \;>\; 0 \\ \square & \text{otherwise} \end{cases} \tag{9.18}$$

This initial label assignment variant is henceforth in this thesis referred to as ILA(XMIN).

An alternative variant relies on a less restrictive assumption, namely that the distance $\|\mathbf{s}_k - \mathbf{m}_k\|$ from the speaker's mouth S_k to their own microphone M_k is smaller than the geometric mean of the distances from that speaker's mouth to all of the other microphones $M_{k'}$, for $k' \neq k$,

$$d_{kk} \;<\; \sqrt[K-1]{\prod_{k' \neq k} d_{kk'}} \tag{9.19}$$

This in turn entails

$$d_{kk}^{K-1} \;<\; \prod_{k' \neq k} d_{kk'} \tag{9.20}$$

or, identically, that

$$\begin{aligned} \frac{\prod_{k' \neq k} d_{kk'}}{d_{kk}^{K-1}} \;&=\; \prod_{k' \neq k} \frac{d_{kk'}}{d_{kk}} \\ &=\; \prod_{k' \neq k} \xi_{kk'} \\ &>\; 1\,. \end{aligned} \tag{9.21}$$

A label $\mathbf{q}[k]$ for participant k can then be estimated using

$$\mathbf{q}[k] \;=\; \begin{cases} \blacksquare & \text{if } \sum\limits_{k' \neq k} \log \xi_{kk'} \;>\; 0 \\ \square & \text{otherwise} \end{cases} \tag{9.22}$$

This initial label assignment variant is henceforth referred to as ILA(XAVE).

It should be noted that averages or extrema of $\xi_{kk'}$ over k' can be modeled as features for channel k in a supervised setting as well, and this has been explored by others [20] following the initial proposal of non-target normalization of cross-correlation maxima in [144] (where it was referred to as "NMXC"). The "non-target normalization" nomenclature is due to [20].

9.4 Training Multi-Participant Acoustic Models on Small Data

This section treats the problem of inferring the parameters of a model $\boldsymbol{\Theta}$ required to provide the likelihood $P\left(\mathbf{F}\left(\mathbf{X}\right) \mid \mathbf{Q}, \boldsymbol{\Theta}\right)$ of features $\mathbf{F}$ of multi-channel observations $\mathbf{X}$ conditioned on a multi-participant state $\mathbf{Q}$. $\mathbf{F}\left(\mathbf{X}\right)$ is a matrix of concatenated channel-specific vectors, each of F features; the size of $\mathbf{F}\left(\mathbf{X}\right)$ is therefore $KF \times T$, while $\mathbf{X}$ consists of K rows, but of $N_S \times T$ samples (given N_S audio samples per frame size). As is usual, it is assumed that the likelihood is provided by a Gaussian mixture model (GMM),

$$P\left(\mathbf{F}\left(\mathbf{X}\right) \mid \mathbf{Q}, \boldsymbol{\Theta}\right) \;\sim\; \sum_{m=1}^{M} \alpha_m \, N\left(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m^{diag}\right) \tag{9.23}$$

where $N\left(\cdot, \cdot\right)$ is a Gaussian distribution, M is the number of mixture components, $\alpha_m \in [0, 1]$ is the prior weight of the mth component, $\boldsymbol{\mu} \in \mathbb{R}^{KF}$ is the vector mean of that component, and $\boldsymbol{\Sigma}_m^{diag} \in \mathbb{R}^{KF \times KF}$ is its covariance matrix.

Furthermore, $\mathbf{\Sigma}_m^{diag}$ is diagonal, with zeros in the off-diagonal entries. Diagonality is justified because, in spite of it, the GMM is still capable of modeling feature correlation provided that M is large. Finally, as elsewhere, $\sum_{m=1}^{M} \alpha_m \equiv 1$.

When there is sufficient training data for a state $\mathbf{S}_i$, $0 \leq i < N_S^K$, where K is the number of participants, the parameters α_m, $\boldsymbol{\mu}_m$ and $\mathbf{\Sigma}_m^{diag}$, $1 \leq m \leq M$, can be inferred using the maximum likelihood criterion via the expectation-maximization algorithm. However, when there is insufficient data, as is likely to be the case for large K, and few or short conversations of small T, maximum likelihood estimation may lead to undertrained models.

The proposed solution for small data conditions, when the available and labeled $\mathbf{Q} \in \{\square, \blacksquare\}^{K \times T}$ is deemed too small for GMM training via maximum likelihood, begins by letting the number of GMM components M be unity. This has the added benefit of significantly simplifying training, since there are no priors α_m and the estimation of the mean and covariance does not require an iterative algorithm. However, because features across channels are expected to be correlated (as evidenced by the frequency of crosstalk), the assumption of a diagonal covariance matrix must be relaxed. This results in a single-component model $\Theta_i \equiv \{\boldsymbol{\mu}_i, \mathbf{\Sigma}_i\}$ for each state $\mathbf{S}_i$.

The maximum likelihood estimates of $\boldsymbol{\mu}_i$ and $\mathbf{\Sigma}_i$ can be obtained by accumulating the zeroth-, first-, and second- order cumulants $C_{i,0} \in \mathbb{R}$, $C_{i,1} \in \mathbb{R}^{KF}$, and $C_{i,2} \in \mathbb{R}^{KF \times KF}$, respectively. Then

$$\boldsymbol{\mu}_i = \frac{C_{i,1}}{C_{i,0}} \tag{9.24}$$

$$\mathbf{\Sigma}_i = \frac{C_{i,2} - C_{i,1}^2}{C_{i,0}} \tag{9.25}$$

Since there are fewer parameters to estimate for $\boldsymbol{\mu}_i$ than for $\mathbf{\Sigma}_i$, the modifications proposed below allow for different cumulants to be used for their estimation. Cumulants used for estimating the mean will be denoted $C_{i,\mu,0}$ and $C_{i,\mu,1}$; those used for estimating the covariance will be denoted $C_{i,\Sigma,0}$, $C_{i,\Sigma,1}$, and $C_{i,\Sigma,2}$. For completeness, maximum likelihood estimation yields:

$$C_{i,\mu,0;ML} = \sum_{t=1}^{T} \delta\left(\mathbf{q}_t, \mathbf{S}_i\right) \tag{9.26}$$

$$C_{i,\mu,1;ML} = \sum_{t=1}^{T} \delta\left(\mathbf{q}_t, \mathbf{S}_i\right) \cdot \mathbf{f}\left(\mathbf{x}_t\right) \tag{9.27}$$

$$\boldsymbol{\mu}_i^{ML} = \frac{C_{i,\mu,1}}{C_{i,\mu,0}} \tag{9.28}$$

$$C_{i,\Sigma,0;ML} = \sum_{t=1}^{T} \delta\left(\mathbf{q}_t, \mathbf{S}_i\right) \tag{9.29}$$

$$C_{i,\Sigma,1;ML} = \sum_{t=1}^{T} \delta\left(\mathbf{q}_t, \mathbf{S}_i\right) \cdot \mathbf{f}\left(\mathbf{x}_t\right) \tag{9.30}$$

$$C_{i,\Sigma,1;ML} = \sum_{t=1}^{T} \delta\left(\mathbf{q}_t, \mathbf{S}_i\right) \cdot \mathbf{f}\left(\mathbf{x}_t\right) \cdot \mathbf{f}^*\left(\mathbf{x}_t\right) \tag{9.31}$$

$$\mathbf{\Sigma}_i^{ML} = \frac{C_{i,\Sigma,2} - C_{i,\Sigma,1}^2}{C_{i,\Sigma,0}} \tag{9.32}$$

where * denotes matrix or vector transpose, to not confuse with the number of frames T, and δ is the Kronecker delta extended to vector arguments.

To improve the performance of a Viterbi decoder which makes use of the models $\mathbf{\Theta}_i$, three different techniques are proposed.

9.4.1 Global Covariance Sharing

The first technique is quite common, consisting of interpolating state-conditioned model parameters with the global, unconditioned model parameters. An auxillary set of cumulants, denoted "G" for global, is accumulated here for covariance estimation only:

$$C_{i,\Sigma,\gamma;G} = \sum_{t=1}^{T} \left(1 - \delta\left(\mathbf{q}_t, \mathbf{S}_i\right)\right) \cdot \mathbf{f}^{\gamma}\left(\mathbf{x}_t\right) \tag{9.33}$$

for the γth cumulant. As can be seen, only the other, non-$\mathbf{S}_i$-labeled frames are used.

9.4.2 Channel Rotation

A second technique consists of rotating channels. To understand the motivation behind this, consider an example with a training set $\mathbf{Q}$ in which only one of the K participants, namely $k = 2$, ever spoke, and that there are $K = 4$ participants. While this may yield ample data for training the state $S_2 \equiv [\square, \blacksquare, \square, \square]^*$, it results in no data for $S_1 \equiv [\blacksquare, \square, \square, \square]^*$, $S_4 \equiv [\square, \square, \blacksquare, \square]^*$, or $S_8 \equiv [\square, \square, \square, \blacksquare]^*$.

It is quite normal to train single-participant models completely ignoring channel number, and here the same idea is used except that the multi-channel $\mathbf{F}\left(\mathbf{X}\right)$ is retained, with a particular subset of channels rotated. This yields accurate covariance estimates for at least some channel pairs; for those channels for which entries of $\boldsymbol{\mu}$ and those channel pairs for which entries of $\boldsymbol{\Sigma}$ are noisy, the technique results in at least some estimate.

Formally,

$$C_{i,\cdot,\gamma;R} = \sum_{t=1}^{T} \sum_{\mathbf{R}} \delta\left(\mathbf{R}\mathbf{q}_t, \mathbf{S}_i\right) \cdot \mathbf{f}\left(\mathbf{x}_t\right)^{\gamma} \tag{9.34}$$

where "R" denotes rotation and $\mathbf{R}$ is a row-rotation matrix operator. The number of unique row-rotation operators is $K!$. It should be noted that this leads to sharing of count mass across states other than $\mathbf{S}_i$, in which exactly the same number of participants are in the $\blacksquare$ state.

9.4.3 Overlap Synthesis

Finally a third technique involves enhancing those models for which data is expected to be particularly sparse. These are models of states in which more than one participant is speaking simultaneously; overlapped speech is much more rare than unoverlapped speech.

To extend the example from the previous subsection, consider that there is ample training data for $\mathbf{S}_2 \equiv [\square, \blacksquare, \square, \square]^*$ and for $\mathbf{S}_4 \equiv [\square, \square, \blacksquare, \square]^*$, but none for the state where both $k = 2$ and $k = 3$, and only they, are speaking, $\mathbf{S}_6 \equiv [\square, \blacksquare, \blacksquare, \square]^*$. It is reasonable that some model (better than no model) can be estimated by combining the models for $\mathbf{S}_2$ and $\mathbf{S}_4$.

Early empirical observations indicated that such a model is of very little benefit. However, in contrast to a model which achieves combination *over features*, a model which is instead trained over features extracted from audio synthesized *at the sample level* is almost as good as one trained over unsynthesized two-participant speech states; the difference, somewhat surprisingly, appears to be mainly due to the fact that when actually overlapping, two participants don't speak as loudly as they do out of overlap — one of the participants tends to speak less loudly.

Auxillary "S" cumulants, short for "synthesis", are therefore trained using

$$C_{i,\cdot,\gamma;S} = \sum_{t=1}^{T} \sum_{u=1}^{T} \delta\left(\mathbf{q}_t \oplus \mathbf{q}_u, \mathbf{S}_i\right) \cdot \mathbf{f}\left(\mathbf{x}_t + \mathbf{x}_u\right)^{\gamma} \tag{9.35}$$

where "$\oplus$" is logical element-wise OR. Where Equation 9.35 would entail too many pair-wise channel combinations, the total number can be randomly downsampled.

9.4.4 Combining Techniques

Empirical evidence, from automatic speech activity detection in meetings (cf. Channel 11), indicates that all three methods are useful. When used together, they lead to cumulants

$$C_{i,\mu,\gamma} = C_{i,\mu,\gamma;ML} + \lambda_R C_{i,\mu,\gamma;R} + \lambda_S C_{i,\mu,\gamma;S} \quad (9.36)$$

$$C_{i,\Sigma,\gamma} = C_{i,\Sigma,\gamma;ML} + \lambda_G C_{i,\Sigma,\gamma;G} + \lambda_R C_{i,\Sigma,\gamma;R} + \lambda_S C_{i,\mu,\gamma;S} \quad (9.37)$$

which are then inserted into Equations 9.24 and 9.25 to replace the maximum likelihood estimates in Equations 9.28 and 9.32.

The parameters λ_G, λ_R, λ_S must be estimated empirically, based on the target task, the target data type, and the target acoustic conditions.

9.5 Fundamental Frequency Variation Spectrum

Modeling multiple participants simultaneously, in the context of multiple microphone channels, calls for the ability to extract features from those channels in a frame-synchronous manner. While the most common acoustic features, including log energy and any of a variety of "spectral" features, are easily extractable from any channel independently of what is happening on that channel, prosodic features are most commonly extracted only after words have been transcribed. This is particularly true of fundamental frequency, whose instantaneous estimates tend to be so noisy that all currently available pitch trackers temper their estimates using long-time continuity constraints.

While useful if the desired product is a trajectory of absolute pitch in pre-segmented utterances, the aforementioned aspect of pitch tracker operation makes it difficult to use estimates of *variation in fundamental frequency* for automatic segmentation. Furthermore, because words are not produced in a synchronous manner across channels, obtaining a multi-channel representation of what all participants are doing at once is complicated. Finally, pitch trackers estimate absolute fundamental frequency, whereas for many tasks, particularly speaker-independent tasks, it is *variation* in fundamental frequency which is called for. First estimating absolute fundamental frequency is a round-about means of obtaining estimates of variation.

This section proposes a novel, alternate representation of fundamental frequency variation (FFV). It is based on the observation that such variation, between two temporally adjacent voiced frames from the same speaker, can be inferred by finding the dilation factor required to optimally align the harmonic spacing in their respective magnitude frequency spectra. This can be achieved without any knowledge of the frequency scale, or precisely which inter-harmonic interval corresponds to the fundamental frequency.

9.5.1 Vanishing Point Product

Finding a dilation factor nominally entails applying a series of alternate dilation factors to one of two adjacent frames and computing a measure of similarity with the other. A common measure of similarity is the normalized dot-product, identically the cosine distance between two same-size spectra. In this thesis, rather than dilating one magnitude frequency spectrum and then taking the dot product with the other, the dot product operation is redefined based on a technique from architectural drawing and grounded in perspective geometry. The argument is depicted graphically in Figure 9.2.

What the figure shows are two magnitude frequency spectra, computed at instants $-T_0$ and $+T_0$. The standard dot product entails a sum over pair-wise products, where any pair is a grouping defined by the points at which a single ray bisects each spectrum. The sum is computed where all rays meet; it can be seen that the standard dot product requires that the rays meet only at infinity.

The proposed *vanishing point product* is defined in the same way, except that the point at which all rays are required to meet is not at infinity but at some finite distance τ to the left of the origin. The vanishing point product, on the right of Figure 9.2, clearly implements a compression of $F_L(\omega, -T_0)$ relative to $F_R(\omega, +T_0)$, by a factor which is given by the geometry of the figure.

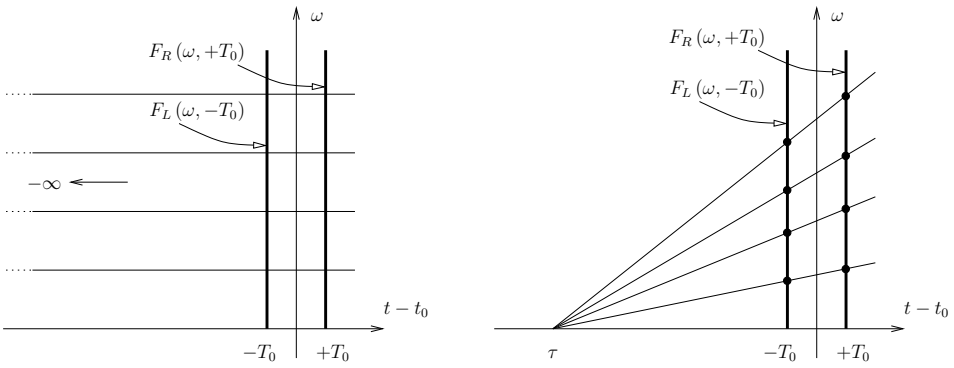


Figure 9.2: The standard dot-product shown as an orthonormal projection onto a point at infinity, on the left, and the proposed vanishing-point product on the right, which approaches the former when $\tau \to \infty$. The vertical lines at $-T_0$ and $+T_0$ are two temporally adjacent magnitude frequency spectra, with frequency along the y-axis and time along the x-axis, as in a spectrogram.

9.5.2 Derivation of the Continuous Spectrum

It can be seen in Figure 9.2 that τ has the disjoint domain $(-\infty, -T_0) \cup (+T_0, +\infty)$. As $\tau \longrightarrow -\infty$ and $\tau \longrightarrow +\infty$, identically, the vanishing dot product approaches the standard dot product, implying identical inter-harmonic spacing between the two spectra. To compute the spectrum for increasing fundamental frequency (shown on the right of Figure 9.2), $\tau \in (-\infty, -T_0)$ must be considered. This means keeping F_R fixed and (conceptually) compressing F_L. To compute the other side of the spectrum, for decreasing fundamental frequency (not shown in the figure), $\tau \in (+T_0, +\infty)$ must be considered instead. This means keeping F_L fixed and (conceptually) compressing F_R.

The precise expression for the continuous spectrum, denoted g^τ, at some point τ, can be obtained by annotating Figure 9.2, as shown in Figure 9.3.

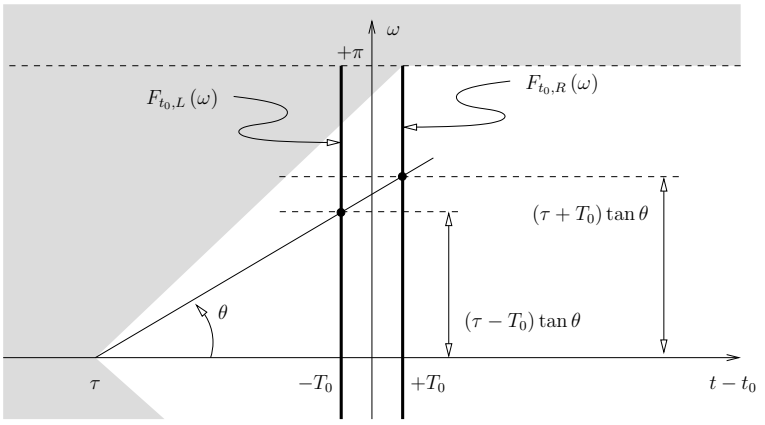


Figure 9.3: Computation of $g^\tau(\tau)$ for $\tau < -T_0$; F_L and F_R are at $-T_0$ and $+T_0$, respectively, with frequency along the y-axis. $\theta_0 = \arctan f_s/2\,(|\tau| + T_0)$.

It can be seen that

$$g^{\tau}(\tau) \tag{9.38}$$

$$= \begin{cases} \int_{-\theta_0}^{+\theta_0} F_L\left((-\tau - T_0)\tan\theta\right) \\ \qquad \cdot\ F_R^*\left((-\tau + T_0)\tan\theta\right) & \tau < -T_0 \\ \qquad \cdot\ (-\tau \cdot \sec\theta)\,(+d\theta) \\ \int_{\pi-\theta_0}^{\pi+\theta_0} F_L\left((+\tau + T_0)\tan(\pi-\theta)\right) \\ \qquad \cdot\ F_R^*\left((+\tau - T_0)\tan(\pi-\theta)\right) & \tau > +T_0 \\ \qquad \cdot\ (+\tau \cdot \sec(\pi-\theta))\,(-d\theta) \end{cases} \tag{9.39}$$

$$= \int_{-\arctan\left(\frac{f_s}{2(|\tau|+T_0)}\right)}^{+\arctan\left(\frac{f_s}{2(|\tau|+T_0)}\right)} \left[\begin{array}{l} F_{t_0,L}\left((-\tau - T_0)\tan\theta\right) \\ \quad \cdot\ F_{t_0,R}^*\left((-\tau + T_0)\tan\theta\right) \\ \quad \cdot\ (|-\tau|\sec\theta)\,d\theta \end{array}\right] \tag{9.40}$$

$$= \begin{cases} \int_{-f_s/2}^{+f_s/2} F_{t_0,L}\left(\left(\frac{-\tau-T_0}{-\tau+T_0}\right)f\right) F_{t_0,R}^*(f)\ df & \tau < 0 \\ \int_{-f_s/2}^{+f_s/2} F_{t_0,L}(f)\, F_{t_0,R}^*\left(\left(\frac{-\tau+T_0}{-\tau-T_0}\right)f\right)\ df & \tau \geq 0 \end{cases} \tag{9.41}$$

$$= \begin{cases} \int_{-f_s/2}^{+f_s/2} F_{t_0,L}\left(f\left(\frac{|\tau|-T_0}{|\tau|+T_0}\right)\right) F_{t_0,R}^*(f)\ df & \tau < 0 \\ \int_{-f_s/2}^{+f_s/2} F_{t_0,L}(f)\, F_{t_0,R}^*\left(f\left(\frac{|\tau|-T_0}{|\tau|+T_0}\right)\right)\ df & \tau \geq 0 \end{cases} \tag{9.42}$$

$$= \begin{cases} \int_{-f_s/2}^{+f_s/2} F_{t_0,L}\left(f\left(1-\frac{2T_0}{|\tau|+T_0}\right)\right) F_{t_0,R}^*(f)\ df & \tau < 0 \\ \int_{-f_s/2}^{+f_s/2} F_{t_0,L}(f)\, F_{t_0,R}^*\left(f\left(1-\frac{2T_0}{|\tau|+T_0}\right)\right)\ df & \tau \geq 0 \end{cases} \tag{9.43}$$

$$= \begin{cases} \int_{-f_s/2}^{+f_s/2} F_L\left(\left(\frac{-\tau-T_0}{-\tau+T_0}\right)f\right) F_R^*(f)\ df & \tau < -T_0 \\ \int_{-f_s/2}^{+f_s/2} F_L(f)\, F_R^*\left(\left(\frac{+\tau-T_0}{+\tau+T_0}\right)f\right)\ df & \tau > +T_0 \end{cases}$$

where "*" represents complex conjugation.

The expression can be simplified by defining the conformal mapping α,

$$\alpha : \tau \mapsto \rho \quad = \quad \begin{cases} -\log_2\left(\frac{-\tau-T_0}{-\tau+T_0}\right) & \tau < -T_0 \\ +\log_2\left(\frac{+\tau-T_0}{+\tau+T_0}\right) & \tau > +T_0 \end{cases} \tag{9.44}$$

It is observed that ρ and τ are conveniently of opposite sign (at $\arg\max_\tau g^\tau(\tau) < -T_0$, F_0 is increasing, and vice versa). Also, Equation 9.44 offers the additional advantage that while $2^{\pm\rho}$ represents change in *linear frequency* per separation lag $2{\cdot}T_0$, ρ itself represents the same change in *octaves* per $2{\cdot}T_0$ seconds.

The mapping α permits rewriting Equation 9.38,

$$g^{\rho}(\rho) \quad = \quad \begin{cases} \int_{-f_s/2}^{+f_s/2} F_L(f)\, F_R^*(2^{+\rho}f)\ df & \rho < 0 \\ \int_{-f_s/2}^{+f_s/2} F_L(2^{-\rho}f)\, F_R^*(f)\ df & \rho \geq 0 \end{cases} \tag{9.45}$$

Conveniently, $g^{\rho}(\rho)$ is analytic in ρ.

9.5.3 Derivation of the Discrete Spectrum

To be serviceable within the prosodic component of a speech processing system, Equation 9.45 must be modified in two ways. First, the frequency spectra F_L and F_R are typically discrete outputs of a fast Fourier transform (FFT) routine. This complicates the evaluation of $F_L(\omega)$ and $F_R(\omega)$, for general frequencies ω. Second, the FFV spectrum $g^{\rho}(\rho)$ itself can be computed only at a finite number of values of ρ.

In this thesis, $F_L(\cdot)$ and $F_R(\cdot)$ are actually discrete magnitude spectra $|F_L[\cdot]|$ and $|F_R[\cdot]|$, representing discrete transforms computed using 32 ms windows over audio sampled at 16 kHz. This makes the spectra, containing both

negative and positive frequency halves, of length $N = 512$. The proposed implementation of Equation 9.45 relies on linear interpolation, using the two closest points, in the (conceptually) compressed spectrum (either F_L or F_R, depending on the sign of ρ). Given discrete transforms $F_L[k]$ and $F_R[k]$, for $k \in \{-N/2+1, -N/2+2, \cdots, -1, 0, +1, \ldots, N/2\}$, the interpolated values are defined as

$$\begin{aligned} \tilde{F}_L\left(2^{-\rho}k\right) &= \left|\lceil 2^{-\rho}k\rceil - 2^{-\rho}k\right| F_L\left[\lfloor 2^{-\rho}k\rfloor\right] \\ &\quad + \left(1 - \left|\lceil 2^{-\rho}k\rceil - 2^{-\rho}k\right|\right) F_L\left[\lceil 2^{-\rho}k\rceil\right] \end{aligned} \tag{9.46}$$

$$\begin{aligned} \tilde{F}_R\left(2^{+\rho}k\right) &= \left|\lceil 2^{+\rho}k\rceil - 2^{+\rho}k\right| F_R\left[\lfloor 2^{+\rho}k\rfloor\right] \\ &\quad + \left(1 - \left|\lceil 2^{+\rho}k\rceil - 2^{+\rho}k\right|\right) F_R\left[\lceil 2^{+\rho}k\rceil\right] \end{aligned} \tag{9.47}$$

Second, Equation 9.45 can only be sampled at a finite number of values of ρ. These values are taken to be equispaced, at $\rho = 4r/N$, $-N/2 \le r < N/2$, representing a range of $[-2, +2)$ octaves.

These two modifications yield the following approximation to Equation 9.45:

$$g^{\rho}[r] = \begin{cases} \sum_{k=-N/2+1}^{N/2} \left|\tilde{F}_L\left(2^{-4r/N}k\right)\right| \left|F_R^*[k]\right| & r \ge 0 \\ \sum_{k=-N/2+1}^{N/2} \left|F_L[k]\right| \left|\tilde{F}_R^*\left(2^{+4r/N}k\right)\right| & r < 0 \end{cases} \tag{9.48}$$

For subsequent modeling, Equation 9.48 is spherically normalized by the square root of $\sum|F_L|^2 \cdot \sum|F_R|^2$, with either $|F_L|$ or $|F_R|$ dilated as in Equation 9.48, to yield an energy-independent vector representation,

$$g_N^{\rho}(\rho) = \begin{cases} \frac{\sum|F_L[k]|\left|\tilde{F}_R^*\left(2^{+\rho}k\right)\right|}{\sqrt{\sum|F_L[k]|^2 \sum\left|\tilde{F}_R^*\left(2^{+\rho}k\right)\right|^2}}, & \rho < 0 \\ \frac{\sum\left|\tilde{F}_L\left(2^{-\rho}k\right)\right||F_R^*[k]|}{\sqrt{\sum\left|\tilde{F}_L\left(2^{-\rho}k\right)\right|^2 \sum|F_R^*[k]|^2}}, & \rho \ge 0 \end{cases}. \tag{9.49}$$

9.5.4 Framing Policy

Given Equation 9.49, the only remaining piece of the definition of the FFV spectrum is a description of how $F_L[\cdot]$ and $F_R[\cdot]$ are computed. Because the FFV spectrum is intended to characterize the instantaneous variation in fundamental frequency within a frame, and preferably in the center of the frame, it is desirable to not use the previous and current frames for computing F_L and F_R, respectively.

Instead, in this thesis, both F_L and F_R are estimated from audio contained within the current frame, using two windows placed symmetrically about the frame's center. The depiction of these two windows, h_L and h_R, which need not be symmetrical but are mirror images of one another, is given in Figure 9.4.

As can be seen, the inside edge of both windows has a Hann profile, while the outside edge has a Hamming profile. This is done to maximize the amount of information used to estimate F_L and F_R, and to simultaneously render them maximally disjoint. Each window's inside edge has size t_{int}, while the outside edge has size t_{ext}. t_{sep} is the distance between the maxima of the windows, and is equal to T_0 in the earlier figures and equations in this section. The width of the audio frame containing both windows is $t_{wid} = 2 \cdot t_{ext} + t_{sep}$.

The parameter t_{sep} is used in expressing the rate of change in fundamental frequency, with the magnitude of change given by the arg max of $g^{\rho}(\rho)$.

9.5.5 A Simple Example

An example of the computation of the FFV spectrum is shown in Figure 9.5, for five values of ρ. The original $|F_L|$ and $|F_R|$, undilated, are shown in the third of the 5 rows. As is shown in the fourth row, the highest value of $g^{\rho}(\rho)$ for the five values of ρ is obtained when ρ is slightly larger than zero, indicating that $|F_L|$ must be dilated to match the interharmonic spacing of $|F_R|$. This indicates that fundamental frequency is smaller in $|F_L|$ than in $|F_R|$, i.e. that it is rising during the 32 ms frame depicted.

When evaluated for a sequence of discrete values of ρ, in the same audio frame as in Figure 9.5, the vanishing point dot product spectrum shown in Figure 9.6 is obtained. As can be seen, the maximum is obtained at a value of ρ which is slighly higher than zero. The location of the maximum is nominally assumed to be the rate of F_0 change in octaves per second.

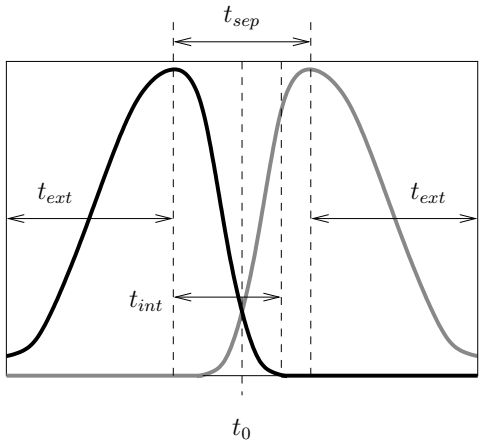


Figure 9.4: Nominal shape and location of the h_L (black) and h_R (gray) windows within a single analysis frame of width t_{wid}, centered on t_0; symbols as in the text.

9.6 A Filterbank for FFV Spectrum Compression

To use the FFV spectrum within a decoder, the information contained in it must be compressed. While it is common to simply identify the arg max in pitch-related work, here it is proposed instead to compress the spectrum to some small number of coefficients. The technique is similar to passing an FFT through the Mel filterbank in speech recognition, and modeling the filter outputs, rather than identifying local maxima in the spectral envelope (i.e., formants).

The filterbank proposed for instantaneous prosodic modeling in this thesis contains 5 filters, shown in Figure 9.7. These 5 filters represent, from top to botton in the figure: (1) quickly falling F_0; (2) slowly falling F_0; (3) flat F_0; (4) slowly rising F_0; and (5) quickly rising F_0.

The filterbank contains two additional filters, which have empirically been found to be useful in circumstances in which the complete 7-filter-output feature vector is subsequently rotated via a principal component analysis rotation or a linear discriminant analysis rotation. They are shown in Figure 9.8. They represent the magnitude of the spectrum at locations which correspond to ± 1 octave errors.

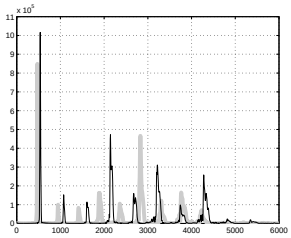
9.7 Relevance to Other Chapters

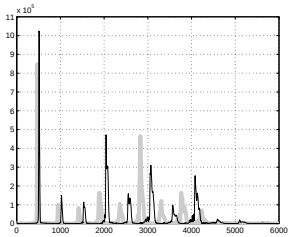
The multi-channel acoustic modeling techniques described in this chapter are exercised in two chapters in Part IV. Non-target-normalized maximum cross-channel correlation is used both as a supervised model feature and as the basis for model-free initial label assignment in Chapter 11. Joint multi-channel models, defined over K channel energies, are then used in a two-pass speech activity detection system in that same chapter. The fundamental frequency variation spectrum, along with the filterbank proposed in Section 9.6, are used in conjunction with a likelihood model of local vocal interaction in Chapter 14, to segment and classify dialog acts without relying on words.

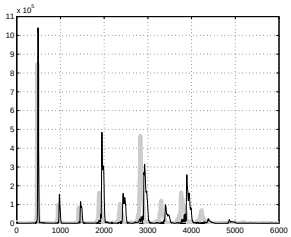
9.8 Summary

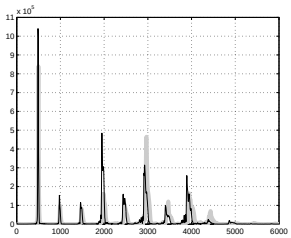
This chapter has presented 5 innovations, related to joint multi-channel acoustic modeling, which fall into two distinct areas of work.

The first area relates to the training of joint multi-channel models for arbitrary conversations with arbitrary numbers of participants in arbitrary rooms. It was argued that given this amount of variability, the most likely tractable approach to speech activity detection involves training models on the test data itself. To that end, this chapter has presented an initial label assignment algorithm, which does not rely on models but on mouth-to-microphone distances for all participant

$$2^{\rho} = 2^{-0.0683} = 0.9538$$  $$g^{\rho}(\rho) = 0.0707$$

$$2^{\rho} = 2^{-0.342} = 0.9766$$  $$g^{\rho}(\rho) = 0.0261$$

$$2^{\rho} = 2^{0} = 1$$  $$g^{\rho}(\rho) = 0.2299$$

$$2^{\rho} = 2^{+0.0342} = 1.0240$$  $$g^{\rho}(\rho) = 0.6877$$

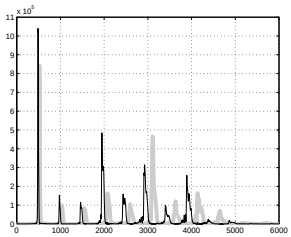
$$2^{\rho} = 2^{+0.0683} = 1.0485$$  $$g^{\rho}(\rho) = 0.0386$$

Figure 9.5: A real example of the computation of $g^{\rho}[\rho]$, for 5 values (one per row) of ρ, shown in the first column. The second column shown $|F_L$ and $|F_R|$, following dilation of one of them (or neither, in middle row), in wide gray or thin black, respectively. The third column gives the computed vanishing point product value.

pairs. A novel signal processing technique, dubbed *non-target-normalized maximum cross-channel correlation*, was derived to estimate ratios between these distances, using audio only. To complete the end-to-end unsupervised system, a number of techniques were presented to train multi-channel models using the initial labels, for situations in which the test data is too short to yield robust maximum likelihood multi-participant models.

The second area dealt with the extraction of a prosodic descriptor, namely variation in fundamental frequency, in a frame-synchronous manner across all channels and independently of prior knowledge of speech activity in any channel. The presented solution involved computing the *fundamental frequency variation spectrum*, an analytic measure of similarity in harmonic spacing between adjacent spectra. Rather than identifying the maximum in this spectrum, it was proposed to

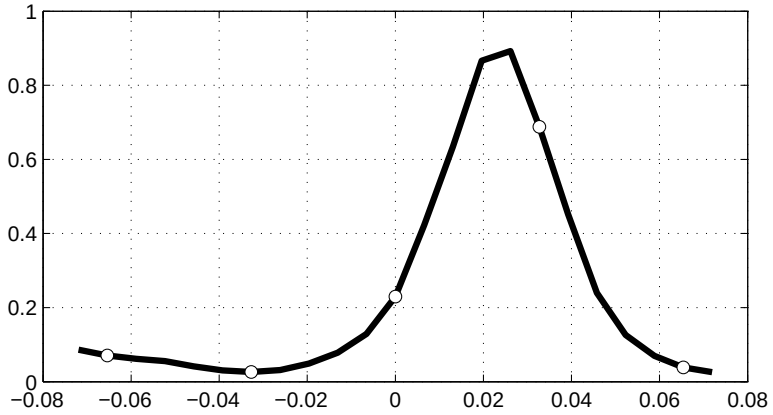


Figure 9.6: A real example of $g^{\rho}(\rho)$, with octaves per second (ρ/t_{sep}) shown along the x-axis. The five values computed in Figure 9.5, from top to bottom, are depicted with white circles, from left to right. Only a small subinterval of the complete discrete domain is shown.

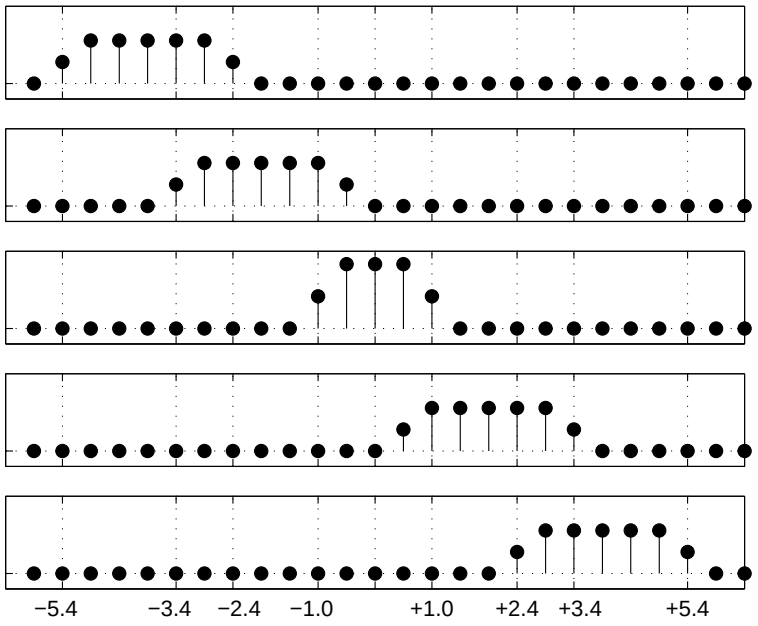


Figure 9.7: The 5 central filters in the FFV filterbank. The x-axis is in semitones per 8 ms. The area under each filter is unity.

compress this spectrum by passing it through a specially designed filterbank, functionally similar to the application of the Mel filterbank to energy spectra when modeling the spectral envelope. The resulting filterbank outputs, it was argued, could then be modeled in a frame-synchronous manner, across channels if desired, using standard acoustic modeling techniques. Most promising in this approach is the potential to use variation in fundamental frequency for acoustic segmentation of speech.

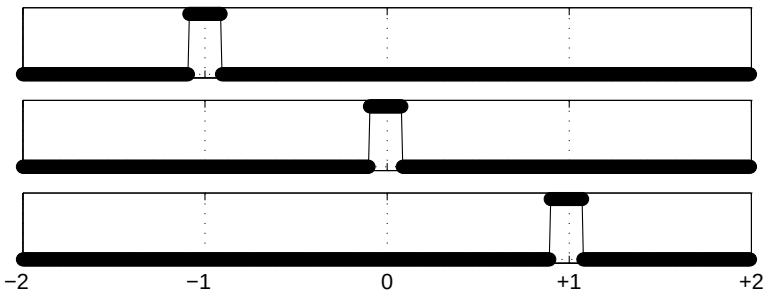


Figure 9.8: Left extremity filter, temporal support of the 5 central filters, and right extremity filter of the proposed filterbank. The x-axis is in octaves per 8 ms.

Part III

EMPIRICAL VALIDATION

Chapter 10

Quantitative Analysis of Turn-Taking, or the Co-Inhibitory Occurrence of Speech

10.1 Introduction

The distribution of speech activity in time and across participants to a conversation has qualitatively been described as the outcome of a process of *turn-taking* [194]. This insight, and its consequences, have proven to be quite useful. And yet the absence of a quantitative model of that distribution, for conversation in general, continues to circumscribe efforts at understanding and reproducing natural interaction. A main concern is that it is not possible to compare conversations with respect to their emergent turn-taking patterns.

To argue by analogy, the ability to compare *written documents* has had a crucial influence on the development of computational models of language and language use. Language models (cf. [116]), which capture the statistics of word sequences, have not only made large vocabulary speech recognition and translation systems possible, but have also significantly facilitated system construction. It is possible, using language modeling techniques alone, to predict a word error rate without running a system. Language models have also made it possible to predict lexical punctuation where it is missing [208], classify texts [41], styles and authors [5], and synthesize "grammatically correct" material [126].

Similar lexical functionality is possible for transcribed *spoken documents*, by extension, but only under one of two particularly felicitous conditions: (1) the transcription describes a single source monologue; or (2) the multi-source transcription has been linearized using a strict turn-taking assumption. This makes it possible to estimate the density of word sequences within turns. However, conversations exhibit two additional degrees of freedom, namely that speech is deployed by participants at variable rates, entailing variable alternate durations of speech and non-speech, and that participants may overlap in time, sometimes dramatically. These degrees of freedom, important to understanding participation patterns, are not captured by language models. As a result, conversation similarity can only be assessed once conversations are lexically transcribed, and then only in a limited way.

This chapter implements the techniques proposed in Chapter 6 and defines a perplexity-based framework for comparing not what is said in conversations, but how those conversations are conducted — specifically, how speech activity is distributed in time and across participants. It is argued that this functionality makes it possible, as language models do for text, to not only compare conversations of arbitrary length and participant number, but also to predict boundaries in conversational phases, to classify conversations, conversational styles, and conversation groups, select subsequences of interest based on arbitrary criteria, and synthesize seemingly natural multi-participant speech deployment patterns.

The research presented here first appeared in [132] and [140].

10.2 Related Work

Work related to that of this chapter is described in Chapter 5.

10.3 Dataset Use

All of the experiments presented in this chapter were conducted using the ICSI Meeting Corpus [108]. As described in Chapter 4, the corpus contains 75 meetings, held by various research groups at ICSI, which would have occurred even if they had not been recorded. This is important for studying naturally occurring interaction, since any form of intervention (including occurrence staging solely for the purpose of obtaining a record) may have an unknown but consistent impact on the emergence of turn-taking behaviors. Each meeting was attended by 3 to 9 participants, which is instrumental for the current purposes as it highlights the proposed models' ability to generalize beyond participant configurations observed in training data.

Only the start and end times of lexical items which accompany the corpus are used in this chapter. They are discretized using a frame step of 100 ms, as explained in Subsection 6.2.2, yielding a vocal interaction chronogram $\mathbf{Q}$. In this chapter, $\mathbf{Q}$ is *observed*.

The meetings in the corpus are used in two distinct ways. First, the chapter explores conversation-specific settings, in which the task is to predict $\mathbf{q}_t$ given $\mathbf{q}_{t-1}$ in a meeting $\mathcal{C}$, having seen (and trained models using) only meeting $\mathcal{C}$. To this end, each meeting $\mathcal{C}$ is split in two complementary ways:

- In the temporally distant A+B split, all instants up to the midpoint of $\mathcal{C}$ are placed in the A half, while all those instants past the midpoint are placed in the B half.
- In the temporally proximate C+D split, all odd-numbered bigrams (with t odd, in $(\mathbf{q}_{t-1}, \mathbf{q}_t)$) are placed in the C half, while all even-numbered bigrams are placed in the D half.

It is anticipated that halves C and D are much more similar to one another than are halves A and B.

Second, the chapter explores conversation-independent settings. In this case, the task is to predict $\mathbf{q}_t$ given $\mathbf{q}_{t-1}$ in all of a meeting $\mathcal{C}$, having seen (and trained models using) the remaining meetings in the corpus. In this setting, meetings are used in their entirety.

10.4 Assessment of Performance

Successful prediction of the true $\mathbf{q}_t$ given the true $\mathbf{q}_{t-1}$ is achieved by maximizing $P(\mathbf{q}_t \,|\, \mathbf{q}_{t-1})$. Maximizing the probability of a sequence of $\mathbf{q}_t$, $t_0 \leq t \leq t_1$, involves maximizing the product of such probabilities. For an entire conversation $\mathcal{C}$, whose vocal interaction chronogram is $\mathbf{Q} \equiv \{\mathbf{q}_t\}$, $1 \leq t \leq T$, maximizing the product is tantamount to maximizing the likelihood of $\mathbf{Q}$ given some model $\mathbf{\Theta}$,

$$P(\mathbf{Q} \,|\, \mathbf{\Theta}) \;=\; P_0 \cdot \prod_{t=1}^{T} P(\mathbf{q}_t \,|\, \mathbf{q}_{t-1}, \mathbf{\Theta}) \;. \tag{10.1}$$

Here, $P_0 \equiv P(\mathbf{q}_0)$ is the unconditional prior probability of observing $\mathbf{q}_0 = \mathbf{S}_0 \equiv [\square, \square, \cdots, \square]$, defined to represent the state in which all participants are silent. $\mathbf{S}_0$ is prepended to every $\mathbf{Q}$ for notational convenience, and does not impact interpretation in any meaningful way[1].

The likelihood of observing $P(\mathbf{Q} \,|\, \mathbf{\Theta})$ is an appropriate metric for comparing models $\mathbf{\Theta}$. However, its numerical value is cumbersome. To simplify comparison, this chapter borrows from language modeling practice, where, for a word sequence

[1] In reality, the instant $t = 0$ refers to the beginning of *the recording* of a conversation rather than the beginning of the conversation, and the recording in general may start at a time when it is unsafe to assume that all participants are silent. However, because $\mathbf{S}_0$ is prepended to *all* $\mathbf{Q}$ when scoring *any* model, comparison among models — the focus of this chapter — remains largely unaffected.

$\mathbf{w}$ of length $\|\mathbf{w}\|$, the *negative log-likelihood* (NLL) and *perplexity* (PPL), defined as

$$\text{NLL} = -\frac{1}{\|\mathbf{w}\|} \log_e P(\mathbf{w} \mid \mathbf{\Theta}) \tag{10.2}$$

$$\text{PPL} = 10^{\text{NLL}}, \tag{10.3}$$

are preferred [116]. PPL in particular is ubiquitously used to compare different word sequences (or corpora) $\mathbf{w}$ and $\mathbf{w}'$ under the same model $\mathbf{\Theta}$, or to compare different models $\mathbf{\Theta}$ and $\mathbf{\Theta}'$ using the same word sequence $\mathbf{w}$.

The corresponding metrics, to be used for similar purposes, are proposed here for the record $\mathbf{Q}$ and the model $\mathbf{\Theta}$.

$$\text{NLL} = -\frac{1}{KT} \log_2 P(\mathbf{Q} \mid \mathbf{\Theta}) \tag{10.4}$$

$$\begin{aligned} \text{PPL} &= 2^{\text{NLL}} \\ &= (P(\mathbf{Q} \mid \mathbf{\Theta}))^{-1/KT} \end{aligned} \tag{10.5}$$

are defined as measures of *turn-taking perplexity*. As can be seen in Equation 10.4, the negative log-likelihood is normalized by the number K of participants and the number T of frames in $\mathbf{Q}$; the latter renders the measure useful for making duration-independent comparisons. The normalization by K does not *per se* suggest that turn-taking in conversations with different K is necessarily similar; it merely provides similar bounds on the magnitudes of these metrics.

Model performance will also be expressed using a relative measure of performance, referred to as Δ, with the baseline model assigned zero and the oracle model assigned unity, or 100%. This will express how much of the gap between the proposed baseline model, and oracle performance, a particular model closes.

10.5 Baseline

The baseline system in this work is a model which ignores participant interaction, i.e., one which treats participants as unconditionally independent. This factors Equation 10.1 into

$$P(\mathbf{Q} \mid \mathbf{\Theta}^{UI}_{any}) = P_0 \cdot \prod_{t=1}^{T} P(\mathbf{q}_t[k] \mid \mathbf{q}_{t-1}[k], \mathbf{\Theta}^{UI}_{any}) \tag{10.6}$$

This model is the only one of the direct models of Section 6.3 which can be applied to any conversation, once it is trained on the same or any other conversation or conversations.

Topline performance, under a first-order Markov assumption, is that of the conditionally dependent model trained on exactly the same data on which it is evaluated. It yields a maximum likelihood estimate of $\mathbf{Q}$, which in turn yields minimum perplexity as defined in Equation 10.5. This model, under matched conditions, is henceforth referred to as the "oracle" system.

On all of the ICSI Meeting Corpus, the baseline model has the performance shown in Table 10.1. When comparing two models, relative performance is most appropriately quantified by subtracting unity from the observed perplexities, since unity represents zero entropy. Given this definition, it can be seen that ignoring interlocutors, even when evaluating models on the same data as they were trained on, yields a perplexity increase of 15.6%rel for the A/B split, or 14.3%rel for the easier C/D split. This is the case when perplexity is computed for the entire half A, B, C, or D. The table also shows perplexities computed for only those bigrams for which $\mathbf{q}_{t-1} \neq \mathbf{q}_t$, denoted "subset", representing instants during which a change occurs (this includes one participant stopping to speak yielding silence from all participants). This is an interesting subset because if the proposed models also modeled duration, then they would need to be able to predict only the $\mathbf{q}_{t-1} \neq \mathbf{q}_t$ transitions. It can be seen that ignoring interlocutors, and still evaluating only on the same data which is used for training models, leads to 24.9%rel and 10.9%rel increases in perplexity for the A/B and C/D splits, respectively, when only this $\mathbf{q}_{t-1} \neq \mathbf{q}_t$ subset is considered.

What can be said about $\mathbf{\Theta}^{UI}_{any}$, as is evident in the table, is that in mismatched conditions, when models are trained on half A and evaluated on half B, and vice versa, and when they are trained on half C and evaluated on half D, and vice versa, perplexities are almost identical to those of the matched condition. In other words, the model which assumes unconditional independence among participants, and which consists of only 2 free parameters as explained in Subsection 6.3.3, generalizes extremely well to unseen portions of the same meeting on which it was trained.

Model	C/D split (easy)		A/B split (hard)	
	all	subset	all	subset
oracle ($\boldsymbol{\Theta}^{CD}$)	1.0915	1.6555	1.0905	1.6444
$\boldsymbol{\Theta}^{UI}_{any}$, matched	1.1046	1.8050	1.1046	1.8047
$\boldsymbol{\Theta}^{UI}_{any}$, mismatched	1.1046	1.8052	1.1047	1.8059

Table 10.1: Perplexities for the oracle and the baseline turn-taking models, on the entire ICSI Meeting Corpus. Both the temporally proximate C/D and the temporally distant A/B splits are shown, with performance for all of $\mathbf{Q}$ in each case as well as for the subset for which $\mathbf{q}_t \neq \mathbf{q}_{t-1}$. "matched" performance is that obtained when scoring a model on the same data which was used to train it; "mismatched" performance refers to the scoring of the complementary half.

10.6 Compositional Multi-Participant Modeling

Conversation-specific settings, in which K is fixed and participants retain their indices in $\mathbf{Q}$, lend themselves to analysis using the direct, compositional multi-participant models of Section 6.3. The current section incrementally relaxes the unconditional independence constraint of the baseline model $\boldsymbol{\Theta}^{UI}_{any}$.

First, separate unconditonally independent submodels can be trained for each participant k, $1 \leq k \leq K$, leading to the model $\{\boldsymbol{\Theta}^{UI}_k\}$ of Equation 6.23. This model continues to ignore interaction, but is able to better represent individual participants' durational preferences of □ and ■ occupation.

Second, separate conditionally independent submodels $\{\boldsymbol{\Theta}^{CI}_k\}$ can be trained for each participant. In constrast to $\{\boldsymbol{\Theta}^{UI}_k\}$, these models condition each participant's behavior at instant t on an explicit representation of the state of *all* participants at the instant $t-1$.

Finally, the conditionally dependent model $\boldsymbol{\Theta}^{CD}$, can be evaluated not only on the portions of conversations used for its training (leading to the "oracle" condition) but also on other portions of the same conversation.

10.6.1 Temporally Proximate Conversation Intervals

The performance of these three models on the easy C/D split of the ICSI Meeting Corpus is shown in Table 10.2. Results are presented for the (cheating) matched condition, in which the models are evaluated on their training material, as well as on the mismatched condtion.

Model	matched				mismatched			
	all		subset		all		subset	
	PPL	Δ	PPL	Δ	PPL	Δ	PPL	Δ
$\boldsymbol{\Theta}^{UI}_{any}$	1.105	0	1.805	0	1.105	0	1.805	+0
$\{\boldsymbol{\Theta}^{UI}_k\}$	1.099	−42	1.738	−45	1.099	−41	1.740	−43
$\{\boldsymbol{\Theta}^{CI}_k\}$	1.093	−92	1.670	−91	*1.096	−69	*1.709	−64
$\boldsymbol{\Theta}^{CD}$	*1.092	−100	*1.656	−100	1.099	−42	1.741	−43

Table 10.2: Temporally proximate C/D split perplexities and Δ measures using direct compositional models, in the matched and mismatched conditions, for all bigrams in each split and for only the subset for which $\mathbf{q}_t \neq \mathbf{q}_{t-1}$. Best performance in each perplexity column identified with a "*".

The table demonstrates that, in matched conditions, the $\boldsymbol{\Theta}^{CD}$ model exhibits the best performance — this is of course the oracle condition. As model complexity increases (downwards in the table), perplexities decrease. In particular, it can be seen that the conditionally indpendent model is almost as good as the conditionally dependent model, suggesting that participants may in fact behave in a conditionally independent manner. In the mismatched condition, however, the $\boldsymbol{\Theta}^{CD}$ fails to generalize as well as the $\boldsymbol{\Theta}^{CI}$ model does, even though the C/D split represents a near-optimally similar but still temporally disjoint split of the data. All models outperform the baseline on all conditions.

10.6.2 Temporally Distant Conversation Intervals

Table 10.3 shows the performance of the same models, but on the less unnatural A/B split of the data, which represents the much more realistic problem of predicting the future in a conversation having seen its first half (and vice versa).

Model	matched				mismatched			
	all		subset		all		subset	
	PPL	Δ	PPL	Δ	PPL	Δ	PPL	Δ
$\boldsymbol{\Theta}^{UI}_{any}$	1.105	0	1.805	0	*1.105	+1	1.806	+8
$\{\boldsymbol{\Theta}^{UI}_{k}\}$	1.098	−48	1.724	−51	1.109	+28	1.783	−14
$\{\boldsymbol{\Theta}^{CI}_{k}\}$	1.092	−93	1.658	−92	1.115	+71	*1.775	−19
$\boldsymbol{\Theta}^{CD}$	*1.091	−100	*1.644	−100	1.122	+121	1.828	+14

Table 10.3: Temporally distant A/B split perplexities and Δ measures using direct compositional models, in the matched and mismatched conditions, for all bigrams in each split and for only the subset for which $\mathbf{q}_t \neq \mathbf{q}_{t-1}$. Best performance in each perplexity column identified with a "*".

What can be seen in the table is that, in the matched condition, as expected, increasing model complexity reduces perplexity; the trends are quite similar to those in Table 10.2. However, in the more interesting mismatched condition, and when the entirety of each half is evaluated, no model beats the baseline unconditionally independent model $\boldsymbol{\Theta}^{UI}_{any}$. The conditionally dependent model $\boldsymbol{\Theta}^{CD}$, for example, is worse than the baseline by more than the baseline is worse than the matched conditionally dependent "oracle" model.

When only the $\mathbf{q}_t \neq \mathbf{q}_{t-1}$ subset is evaluated, the baseline is outperformed by both the $\{\boldsymbol{\Theta}^{UI}_{k}\}$ and $\{\boldsymbol{\Theta}^{CI}_{k}\}$ models. Since this subset contains all instances of change in $\mathbf{Q}$, including those bigrams in which more than one participant is talking, it is not particularly surprising that the baseline model can be beaten. The $\{\boldsymbol{\Theta}^{CI}_{k}\}$ model, it should be noted, was also the best-performing model for this subset in the mismatched C/D condition (cf. Table 10.2).

10.6.3 Analysis of Perplexity Departures in Time

The perplexities shown so far have been computed over entire (halves of) conversations, and therefore obfuscate temporal departures in perplexity from their global mean. Due to the nature of most naturally occurring conversations, those departures can be quite dramatic.

Figure 10.1 shows an interval of a single meeting, `Bmr024` from the ICSI corpus, starting 5 minutes into the conversation and ending 20 minutes later (the meeting is actually over 50 minutes long but only a snippet is shown to better appreciate small time-scale variation). The curves are not unweighted global averages as in Equation 10.5, but local averages computed using a 60-second Hamming window centered on each instant. Three curves are shown.

The dashed curve represents the performance of the "oracle" $\boldsymbol{\Theta}^{CD}$ model trained on the entire 50+ minutes of the meeting, while the thin black line shows the performance of the same model trained only on the first A half, of which the depicted snippet is a proper subinterval. It can be seen that at some instants, the latter model yields slightly better performance (because it overfits the depicted interval), but that in general the two models are quite similar. As the curves demonstrate, there is significant variation in turn-taking perplexity over time.

The third curve in the figure, depicted with a thick gray line, is the perplexity of the model trained on the second B half of the same meeting. As can be seen, it follows the "oracle" model only in the interval $[11, 15]$ minutes. Over the remainder of this 20-minute temporal window, the mismatched model is unable to predict the observed distribution of speech.

Figure 10.2 shows similar trajectories for the other direct compositional models, retaining in each panel the full-duration conditionally dependent "oracle" model for comparison.

Perplexities achieved by the conditionally independent model $\{\boldsymbol{\Theta}^{CI}_{k}\}$ are shown in panel (a) of the figure. As can be seen, in the matched condition the model appears to follow approximately the same trajectory as the matched $\boldsymbol{\Theta}^{CD}$ model. This suggests again that participant behavior may in fact be largely conditionally independent. It is also seen that failures in the mismatched condition are of smaller magnitude than those of the $\boldsymbol{\Theta}^{CD}$ model.

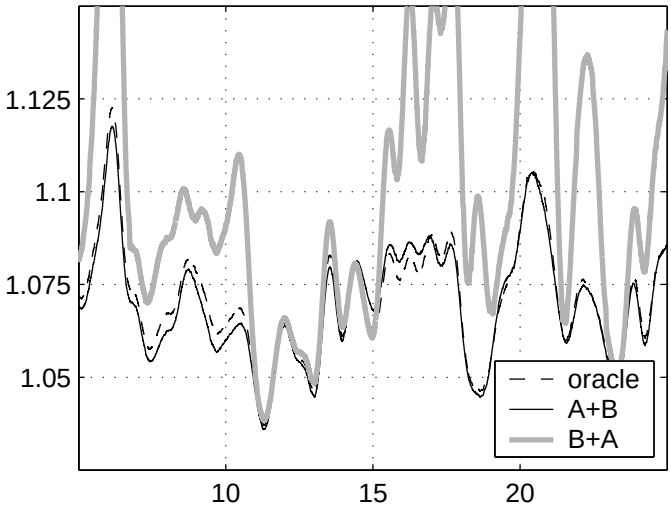


Figure 10.1: Perplexity (along y-axis) in time (along x-axis, in minutes) for meeting `Bmr024` under a conditionally dependent global oracle model, two "matched-half" models (A+B), and two "mismatched-half" models (B+A).

Panel (b) demonstrates the trivial fact that a conditionally independent model $\mathbf{\Theta}_{any}^{CI}$, described in Section 6.3.2, is useless. This is of course because it cannot predict the next state of a generic participant for which the index k in $\mathbf{q}_{t-1}$ has been lost.

Unconditionally independent models are shown in panels (c) and (d). In the first of these, it can be noted that, in the matched condition, the $\{\mathbf{\Theta}_k^{UI}\}$ model, with K distinct submodels, visibly departs from the "oracle" trajectory, particularly in the interval $[15, 18]$ minutes. Elsewhere it yields perplexities which are systematically higher. In the mismatched condition, however, the $\{\mathbf{\Theta}_k^{UI}\}$ model trajectory tracks the oracle trajectory far more closely than any of the models which model interlocutor context.

This trend is taken even further in panel (d), which shows that the $\mathbf{\Theta}_{any}^{UI}$ model yields nearly identical trajectories in the matched and mismatched conditions (the gray line is superimposed on the solid black line). However, it is consistently more surprised than any of the more complex models. It should be stressed that because the the depicted perplexities are smoothed over 60-second windows, they underestimate poor performance at specific points in time which are frequent but short in duration.

10.7 Exploiting the Systematics of Speech Overlap Dynamics

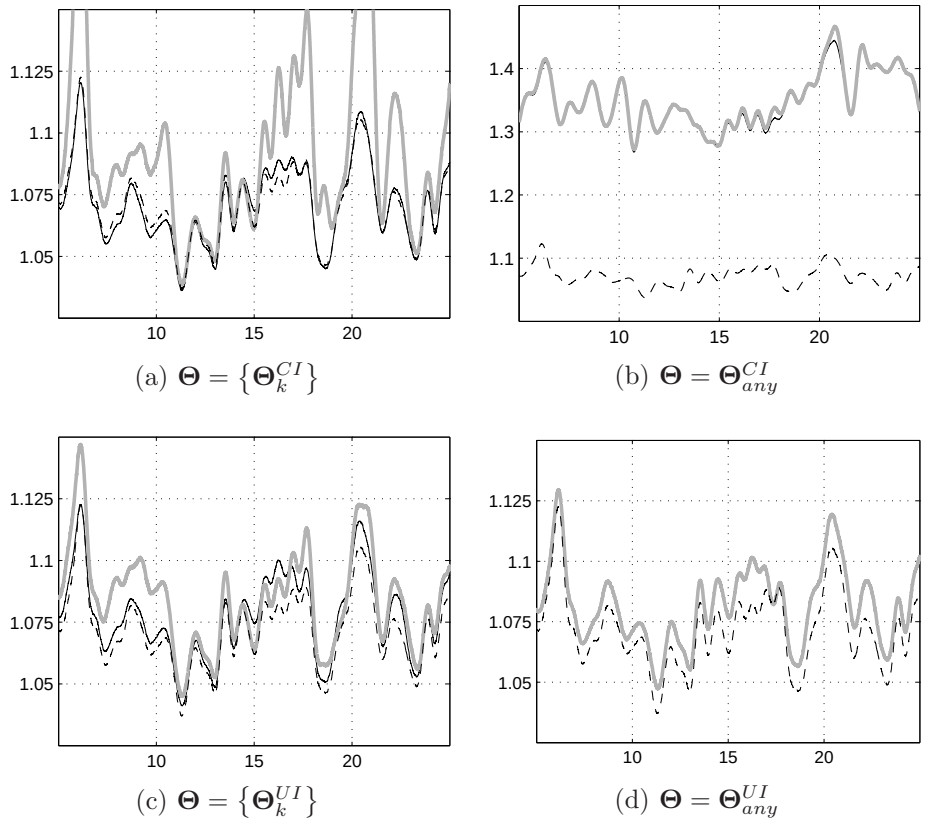
The compositional dependent-participant models of the previous section suffer from the problem of a very large state space, as argued in Chapter 6. In that chapter, it was proposed that the EDO model may be more successful at predicting unseen instants in conversation, precisely because its state space is far smaller. This section provides empirical evidence for that claim.

The EDO model, in this section, is trained with $K_{max} = K$ in Equation 6.56, since it is applied only to those meetings using which it was trained.

10.7.1 Temporally Proximate Conversation Intervals

For the C/D split of the ICSI Meeting Corpus, EDO perplexities are shown in Table 10.4. The performance of the best-performing direct, compositional model from Table 10.2 in each column is included for reference as $\mathbf{\Theta}^{OPT}$.

As Table 10.4 reveals, the EDO model is outperformed by the best direct models $\mathbf{\Theta}^{OPT}$, on the C/D split. In the matched condition, it comes within approximately $\Delta = 50\%$ of the "oracle" model. It also comes to within approximately



(a) $\boldsymbol{\Theta} = \left\{\boldsymbol{\Theta}_k^{CI}\right\}$ (b) $\boldsymbol{\Theta} = \boldsymbol{\Theta}_{any}^{CI}$

(c) $\boldsymbol{\Theta} = \left\{\boldsymbol{\Theta}_k^{UI}\right\}$ (d) $\boldsymbol{\Theta} = \boldsymbol{\Theta}_{any}^{UI}$

Figure 10.2: Perplexity (along y-axis) in time (along x-axis, in minutes) for meeting `Bmr024` under a conditionally dependent global oracle model, and various matched (A+B) and mismatched (B+A) model pairs with relaxed dependence assumptions. Legend as in Figure 10.1.

Model	matched				mismatched			
	all		subset		all		subset	
	PPL	Δ	PPL	Δ	PPL	Δ	PPL	Δ
$\boldsymbol{\Theta}_{any}^{UI}$	1.105	0	1.805	0	1.105	0	1.805	+0
$\boldsymbol{\Theta}^{OPT}$	*1.092	−100	*1.656	−100	*1.096	−69	*1.703	−68
$\boldsymbol{\Theta}^{EDO}$	1.098	−53	1.726	−52	1.098	−49	1.730	−49

Table 10.4: Temporally proximate C/D split perplexities and Δ measures using the EDO model, in the matched and mismatched conditions, for all bigrams in each split and for only the subset for which $\mathbf{q}_t \neq \mathbf{q}_{t-1}$. Also shown, for each matched/mismatched condition and all/subset data separately, is the lowest perplexity and Δ measure achieved by the direct compositional models in Table 10.2, denoted here as $\boldsymbol{\Theta}^{OPT}$. Best performance in each perplexity column identified with a "*".

$\Delta = 50\%$ of the "oracle" model in the mismatched condition, suggesting that its generalization ability is similar to that of the baseline (i.e., it is high relative to the other models explored). As Table 10.2 also showed, direct compositional models exhibit a significant performance difference for matched and mismatched conditions.

10.7.2 Temporally Distant Conversation Intervals

When the training and evaluation material is not as temporally proximate as in the C/D split, the EDO models are far more competitive. The perplexities for $\mathbf{\Theta}^{EDO}$, along with the baseline and the best performing direct compositional model from Table 10.3, are shown in Table 10.5.

Model	matched				mismatched			
	all		subset		all		subset	
	PPL	Δ	PPL	Δ	PPL	Δ	PPL	Δ
$\mathbf{\Theta}^{UI}_{any}$	1.105	0	1.805	0	1.105	+1	1.806	+8
$\mathbf{\Theta}^{OPT}$	*1.091	−100	*1.644	−100	1.105	+1	1.781	−15
$\mathbf{\Theta}^{EDO}$	1.098	−49	1.725	−49	*1.098	−43	*1.731	−45

Table 10.5: Temporally distant A/B split perplexities and Δ measures using the EDO model, in the matched and mismatched conditions, for all bigrams in each split and for only the subset for which $\mathbf{q}_t \neq \mathbf{q}_{t-1}$. Also shown, for each matched/mismatched condition and all/subset data separately, is the lowest perplexity and Δ measure achieved by the direct compositional models in Table 10.3, denoted here as $\mathbf{\Theta}^{OPT}$. Best performance in each perplexity column identified with a "*".

It can be seen in column 2 through 5 of the table that EDO model performance comes within approximately $\Delta = 50\%$ of the performance of the "oracle" $\mathbf{\Theta}^{CD}$ model in the matched condition. This is similar to its relative performance on the C/D split. In the mismatched conditon, however, the EDO models come within approximately $\Delta = 44\%$ of the "oracle" model's matched performance, far exceeding the performance of any of the direct compositional models in this condition.

10.8 Generalization to Unseen Data

The ability to predict future behavior in conversations already underway is important, and the experiments in the precedings sections demonstrate that model choice, guided by minimizing perplexity, may be conditioned on scenario specifics. This includes the duration of the conversation which is available for training. Ideally, models would be trainable using a corpus of conversations, not necessarily involving the same participants or even the same number of participants. As was argued in Subsection 6.3.4, however, compositional models which do not ignore participant interaction are both K-specific and $\mathbf{R}$-specific. This largely eliminates the possibility of robustly training them on other conversations.

In contrast, the EDO model exhibits neither K-specificity nor $\mathbf{R}$-specificity. This makes it possible to augment training sets consisting of already observed portions of a conversation with large amounts of material from other conversations, and then apply it to predict the future as attempted in the preceding section.

Furthermore, this flexibility allows for the EDO model to be used in circumstances in which there is *no* material available from the ongoing conversation, such as at the beginning of the latter. This section considers precisely that task. The ICSI Meeting Corpus is evaluated one meeting at a time, in a round robin fashion, using all of the remaining meetings for training. Of the direct compositional models, the only model which can be applied in this way is the $\mathbf{\Theta}^{UI}_{any}$ model, which treats participants as unconditionally independent. As in the previous sections, this model also serves as the baseline for the experiments presented in Table 10.6.

The table lists several variants of the $\mathbf{\Theta}^{EDO}$ model, for several values of K_{max}. Unlike in the experiments of Section 10.7, a K_{max} must be selected during training because the model is subsequently applied to a conversation whose K is potentially different from that of any K in the training data.

Table 10.6 shows that for all explored $K_{max} > 3$, EDO model performance is far better than that of the only alternative, the baseline. For $K_{max} = 6$, the EDO model closes the gap between the baseline system and the "oracle" model (which has been trained on only the one meeting which is being evaluated in each meeting's round robin turn) by 51% when all bigrams are scored, and by 54% when only the $\mathbf{q}_t \neq \mathbf{q}_{t-1}$ bigrams are.

EDO performance for $K_{max} \leq 3$ is worse than that of the baseline, because transition probability mass is shared uniformly into all states with degrees of overlap of 3 or more. Evidently, overlap of degree 3 occurs frequently enough

Model	all		subset	
	PPL	Δ	PPL	Δ
Θ^{UI}_{any}	1.105	0	1.817	0
Θ^{EDO}_{6}	1.099	−51	1.733	−54
Θ^{EDO}_{5}	1.099	−51	1.733	−54
Θ^{EDO}_{4}	1.099	−50	1.733	−54
Θ^{EDO}_{3}	*1.100	−42	*1.737	−52
oracle (Θ^{CD})	1.092	−100	1.662	−100

Table 10.6: Perplexities and Δ measures for several EDO models when applied to entirely unseen conversations, for all bigrams in each split and for only the subset for which $\mathbf{q}_t \neq \mathbf{q}_{t-1}$. Subscripts on Θ^{EDO} indicate the K_{max} value with which they were trained. Also shown is baseline model performance and that of the "oracle" Θ^{CD}, the latter trained only on each evaluation conversation. Best (non-oracle) performance in each perplexity column identified with a "*".

that taking probability mass away from transitions to it, and sharing it evenly with the often more numerous (but less frequently occurring) states of overlap of degree of 4 or more, is deleterious.

10.9 Potential Impact

This chapter has proposed a novel means of analyzing conversations, by providing a framework for computing turn-taking perplexity. This measure may be used to compare conversations, and to compare models of conversations, in the same way that lexical perplexity is used to compare documents, or to compare language models. As such, the techniques presented may have very broad appeal in both the conversation understanding and the dialogue system communities.

In conversation understanding, and in meeting understanding specifically, turn-taking perplexity may be used to quantify the difficulty of an understanding task. For example, it may be the case that errors in inferring argument structure [85] are more numerous in meetings whose turn-taking perplexity is higher. At the current time, means of automatically predicting performance on this task do not formally exist. Turn-taking perplexity may also serve as an indicator of phenomena *within* individual conversations, such as conversational hot spots [225]. Correct identification of the latter may help in qualifying the propositional content found in their vicinity. Finally, it is possible that turn-taking perplexity is correlated, positively, negatively, or both, at different potentially predictable instants, with lexical perplexity. As such, there exists the prospect of conditioning speech recognizers on the temporally local distribution of speech across participants.

In spoken dialogue systems, models of turn-taking can be expected to successfully participate in assessing the naturalness of the timing of system contributions[2], in both two-party and general multi-party settings.

10.10 Relevance to Other Chapters

The experiments in this chapter have demonstrated that the EDO model described in Chapter 6 may be used to predict the speech chronogram $\mathbf{Q}$. As such, they suggest that the model may be used in speech activity detection, where the distribution of speech in time and across participants is not known *a priori*. The application of a non-ergodic variant of the EDO model, the generalized EDO model (gEDO), to precisely that task is presented in the ensuing chapter.

In Chapter 12, the EDO model is used to contrast the distribution of speech with that of laughter, in the same corpus of meetings as used here. The two types of vocal activity are shown to exhibit dramatically different distributions, primarily because people generally await their turn to speak but, when intending to laugh, do not wait until other laughers are finished.

Finally, Chapter 16 demonstrates the application of the EDO model to the task of characterizing conversation type. Competing models are trained for meetings held by different project groups. Unseen meetings are then analyzed to infer

[2]The author is grateful to an anonymous ACL 2010 reviewer for suggesting this application.

the most likely group to have held them. It is shown that EDO models can contribute to this task, but that they are outperformed by other approaches which account for observed differences among unlabeled participants.

10.11 Summary

This chapter has proposed a framework for computing turn-taking perplexity of a multi-party conversation, defined as being proportional to the log-likelihood of the distribution of speech in time and across participants. The measure can be computed for any conversation, regardless of duration and participant number, and makes possible the comparison of conversations given a model and of models given a conversation.

Within this framework, it was shown that independent-participant models, as commonly used in acoustic speech processing and inference, are frequently surprised by observations, yielding perplexities which are much higher than those obtained with oracle joint-participant models which have been trained on the test data. In contrast, the ergodic EDO model described in Chapter 6 performs quite well, since it is able to predict participants' future behaviors on the past behaviors of all participants. On unseen conversations, it reduces the relative gap between the perplexity of an ergodic independent-participant model and the ergodic oracle model by 50%.

10.12 Future Directions

Over and above the possible applications of computing turn-taking perplexities, some of which are listed in Section 10.9, there is significant scope for improving the modeling techniques described in this chapter.

One obvious extension of the proposed models is to Markov orders beyond the first. This can be achieved by applying the generalized EDO model of Section 6.5. It is expected that future vocal production is conditioned on its past duration, and on that of interlocutors' vocal productions' past duration; truncating dependency to only the most recent frame, as in the models of this chapter, circumscribes exploration in this direction. Temporal dependency extension may also be achieved by considering larger frame steps within the EDO (or other ergodic) model, and it is possible that combinations of systems with different frame steps can be deployed to predict the future with complementary effect.

A particularly interesting future endeavor is the application of the parametric state-space models of vocal interaction proposed in Chapter 7. These models may be easily extended to consider instants significantly further back than the most recent, their generalization being arbitrary-structure feed-forward multi-layer perceptrons. This inherent flexibility allows for the modeling of inter-dependence among vocal activity types. It is quite probable that speech may be better predicted by also using laughter activity context, and vice versa, than is either using only one vocal activity type alone.

Finally, parametric models of turn-taking can be easily extended to allow inclusion of other types of features, beyond the mere occurrence of a particular vocal activity type. For example, the parametric models of Chapter 7 can easily accommodate per-frame, participant-attributed prosodic or spectral features. That turn boundaries can be predicted from prosodic contours is well-argued in the literature; at the current time, a general multi-party, joint-participant model which can benefit from prosody in predicting turn-taking phenomena is not available.

Chapter 11

Automatic Speech Activity Detection

11.1 Introduction

Speech activity detection (SAD) comprises the task of detecting, per frame, the occurrence of speech. In this sense, it is a binary classification problem.

In scenarios in which speakers are individually instrumented with microphones, SAD is typically implicitly also a speaker diarization task; that is, speech must not only be detected, it must also be correctly attributed to its source or sources. It is still a binary classification task, but rather than discriminating between speech and non-speech, systems must discriminate between nearfield-speech and non-nearfield-speech (which includes non-speech as well as farfield speech) on each channel. This makes SAD a challenging task, as observed for multi-party meetings by [181].

The main challenge lies in the fact that, because participants share an acoustic space and because instrumentation is most often arranged ad hoc, farfield speech from others appears on microphone channels worn by participants who are not speaking. Such farfield speech is known as *crosstalk*. When misclassified as nearfield speech, crosstalk leads to the erroneous hypothesis of *overlap*, or the condition of more than one participant talking simultaneously.

The fact that crosstalk can be quite frequent, but true overlap is rare, makes SAD the testbed application *par excellence* for models of vocal interaction. Since overlap is a multi-participant effect, models of multi-participant behavior have the opportunity to temper acoustic misclassification of crosstalk by providing a prior on the statistical likelihood of overlap.

SAD output is most often an intermediate product in the task of speech/non-speech segmentation for automatic speech recognition (ASR) systems. Segmenters fulfill ASR requirements by marshalling consecutive SAD frames hypothesized as speech into much longer intervals, referred to as utterances or talkspurts, often closing short non-speech gaps, pruning out short speech spurts, and pre- and post-padding the resulting intervals. Successful exclusion of crosstalk is quite important for speech recognition and understanding systems, for three main reasons:

1. Utterances should be *speaker-homogenous*, that is they should include productions from a single source, for acoustic modeling reasons. Failure to do so circumscribes the benefit of speaker adaptation passes in multi-pass recognition systems, which assume a single source.

2. Utterances should be *syntactically contiguous*, that is they should exclude concommittant or interspersed productions from other sources, for language modeling reasons. Failure to do so circumscribes the benefit of language modeling in ASR, which assumes conditional dependence of each word on its same-source predecessor.

3. Mis-attributed speech, even when correctly recognized in spite of failure to meet the above two criteria, may cause semantic or pragmatic errors in downstream processing.

This chapter applies the models of Chapter 6 to demonstrate the role that models of vocal interaction can play in speech activity detection. Multi-participant modeling in SAD requires first and foremost a reduction in the number of degrees of freedom of each participant to be modeled. This enables the joint modeling of the multi-channel observable space, particularly channel correlation, without requiring a rule-based post-processing step. As the experiments in this

chapter show, both multi-participant state-space modeling and multi-channel feature-space modeling can significantly limit the deleterious misclassification of crosstalk as nearfield speech.

The work presented in this chapter is based on early SAD experiments described in [144] in which no transition modeling was used. It was deployed in the interACT submission to the NIST Rich Transcription 2004 evaluation [172, 117]. Refinements involving transition modeling were published in [147, 64, 148, 149, 141]. The supervised independent-participant acoustic model was described in [150].

11.2 Related Work

The overwhelming majority of close-talk speech activity detectors in current use are based on the HMM decoder described in [1], but typically rely on mixtures rather than single Gaussians. However, since the advent of large meeting corpora earlier in this decade, it has been known that the basic design is deficient in multi-party settings, due in large part to its high false alarm rates caused by crosstalk [181].

The solution proposed in [181] has been to continue to decode the speech activity of all K conversation participants independently, but after normalizing channel features by substracting their channel-specific extrema. Importantly, for log-energies $\mathbf{e}_t[k]$, this normalization is accompanied by rotation $\mathbf{R}$ into a slightly less correlated space,

$$\mathbf{e}'_t = (\mathbf{R} \cdot \mathbf{e}_t) - \min_t \mathbf{e}_t \tag{11.1}$$

$$\mathbf{R} = \frac{1}{K}(K\mathbf{I} - 1) \ , \tag{11.2}$$

where $\mathbf{I}$ is the $K \times K$ identity matrix. After indepenedent-participant decoding, a subset of false alarms is eliminated via a post-processing step based on spherically-normalized cross-channel-correlation maxima. For any two channels hypothesized to contain speech at the same time, the speech hypothesis in the channel with the lower energy is discarded if the cross-channel-correlation maximum exceeds a threshold indicating a single-source signal.

A descendent of this approach was described in [207], where the cross-channel-correlation-based post-processing was replaced. The new approach consisted of intersecting independent-participant-decoder output $\mathbf{q}_t$ with a mask $\mathbf{m}_t$ of same dimensionality, with $\mathbf{m}_t[k] = 1$ if $\mathbf{e}'_t > 0$ and zero otherwise. $\mathbf{e}'_t$ was defined as

$$\mathbf{e}'_t = \mathbf{R} \cdot \left(\mathbf{e}_t - \min_t \mathbf{e}_t\right) \tag{11.3}$$

$$\mathbf{R} = \frac{1}{K-1}(K\mathbf{I} - 1) \ . \tag{11.4}$$

In a further descendent described in [20], the post-processing step of [207] was eliminated altogether, and instead the quantities in Equations 11.1 and 11.2, after modification, were modeled alongside other features during Viterbi decoding (as originally in [181]). The modifications consist of

$$\begin{aligned} \mathbf{e}'_t &= \mathbf{R} \cdot \left(\mathbf{e}_t - \min_t \mathbf{e}_t\right) \\ \mathbf{R} &= \mathbf{I} - \boldsymbol{\delta} \ , \end{aligned} \tag{11.5}$$

where $\boldsymbol{\delta}$ is a $K \times K$ matrix of entries (δ_{ij}) and

$$\delta_{ij} = \begin{cases} 1 & \text{if } \arg\max_{k \neq i} (\mathbf{e}_t - \min_t \mathbf{e}_t)[k] = j \\ 0 & \text{otherwise} \end{cases} \ . \tag{11.6}$$

In contrast to Equation 11.4, where the rotation merely subtracts from each channel's floor-subtracted log-energy a constant fraction of every other channel's floor-subtracted log-energy, Equation 11.6 defines a data-dependent rotation. It subtracts from each channel's floor-substracted log-energy the maximum of the remaining channels' floor-subtracted log-energies. The resulting K-length feature vector consists of each channel's NLED(min) feature; a similar operation involving $\arg\min$ in place of $\arg\max$ in Equation 11.6 yields a vector of NLED(max) features. These two features, comprising the NLED feature pair [20], are modeled alongside the original $\mathbf{e}_t$.

These very successful acoustic feature developments, proposed by ICSI, are intended for use within an independent-participant HMM decoder; in particular, emission probability models describe a single participant, and discard the $\mathbf{e}_t[j]$

and $\mathbf{e}'_t[j]$ for all $j \neq k$ when decoding participant k. An alternative proposed in [98, 227] involves not decoding *participants* independently, but instead decoding participants' *microphones* independently. This creates an opportunity to train constrastive models not just of *nearfield speech* versus *nearfield non-speech*, but for example *nearfield speech* versus *farfield speech* versus *non-speech* as in [98], or *nearfield speech without farfield speech* versus *nearfield speech with farfield speech* versus *farfield speech without nearfield speech* versus *non-speech* as in [227]. When viewed jointly across participants (even when decoding is performed independently for each participant), this yields a state space of 3^K and 4^K acoustically distinct states, respectively, rather than 2^K acoustically distinct states. Because only 2^K of these microphone states correspond to valid participant states (e.g., all microphones cannot be in the *farfield speech* state), rule-based voting is applied in a post-processing stage.

The experiments in this chapter were conducted against the background of the work cited above [1, 181, 98, 227, 207, 20]. It should be noted that, in addition to the NLED features of [20], the majority of the systems described in these works make use of auxillary features, either in the emission probability models in their HMM state spaces or in post-processing suites. Experimenting with such features is beyond the scope of the current effort, which aims to compare independent-participant versus joint-participant decoding.

Auxillary features and alternative acoustic modeling approaches appear to have dominated development in close-talk SAD technology since 2006. Of note are [83, 56, 84], which have made use of an echo-cancellation approach, those auxillary features deemed most successful in [227], and a neural network classifier with large temporal context. It is particularly interesting that in [84], the largest contribution of error came from missing speech rather than from inserting it. As in [1, 181, 98, 227, 207, 20], however, these systems have all treated participants as conditionally independent, given the multi-channel acoustic space.

In pursuing joint-participant decoding, the experiments described in what follows differ significantly from the cited work. With respect to modeling energy in a rotated, arguably more decorrelated space, it will be shown that jointly modeling the acoustics of all close-talk channels present allows for true energy decorrelation. This is performed implicitly, by training a full-covariance single-Gaussian log-energy model. Since the number of participants may vary across conversations, and since acoustic coupling is likely to be different even if the number of participants stays constant, such a model can only be trained on the *testing* data, relying on an initial label assignment from a prior decoding pass. The model is single-Gaussian because conversations can be short (10 minutes in this work), providing very little evidence for model training.

Modeling participant *states* jointly, as proposed in this chapter, eliminates the need for a voting scheme in a post-processing pass. Rather than decoding microphone states individually in a space whose Cartesian product may not be licensed, as in [98, 227], decoding proceeds in a Cartesian product space which already contains only those 2^K multi-microphone configurations which correspond to valid multi-participant configurations. In this way, the rule-based voting algorithms which cannot be trained statistically in [98, 227] are eliminated.

11.3 Dataset Use

This chapter evaluates speech activity detection on datasets which have been used extensively for evaluating speech recognition systems in multi-party meeting scenarios.

The DevSet is the NIST Rich Transcription Evaluation evaluation corpus from 2005 (`rt05s_eval`), described in some detail in Chapter 4. Briefly, it consists of a 10-minute excerpt from each of 10 meetings held at 5 locations. The number of participants varies from meeting to meeting, in the range $\{4, 9\}$. Each participant is instrumented with a head-mounted microphone; one meeting, known as `NIST_20050412-1303`, was excluded from DevSet precisely because one of the participants is not so instrumented[1].

The EvalSet is the NIST Rich Transcription Evaluation evaluation corpus from 2006 (`rt06s_eval`). It consists of a 10-minute excerpt from each of 9 meetings held at 5 locations. Most of the EvalSet collection locations are the same as for DevSet, and the number of participants falls in the same range. No meetings from the corpus were excluded.

Emission and transition probability models in this chapter are trained using TrainSet, drawn from the ICSI Meeting Corpus. Like the NIST Rich Transcription Evaluation corpora, it is also described in detail in Chapter 4. 51 meetings,

[1] `rt05s_eval` less the one excerpt was referred to as *confDEV* in [64], `rt05s_eval*` in [141, 149], and *Eval05** in [19].

		Hypothesis	
		$\mathcal{S}$ (■)	$\neg\mathcal{S}$ (□)
Reference	$\mathcal{S}$ (■)	true positive (TP)	false negative (FN)
	$\neg\mathcal{S}$ (□)	false positive (FP)	true negative (TN)

Table 11.1: Names of discrete events in a 2×2 confusion matrix for the detection of an event; here, the detection is of nearfield speech activity frames.

of the total of 75 available, were included. These are the same 51 meetings commonly used by many authors as training material for dialog act segmentation and classification work[2]

11.4 Assessment of Performance

Automatic speech activity detectors hypothesize, for each frame of audio, a binary label of speech ($\mathcal{S} \equiv$ ■) or non-speech ($\neg\mathcal{S} \equiv$ □). For close-talk-microphone-instrumented participants, as is the case here, these two labels represent *nearfield-speech* and *non-nearfield-speech*, respectively. Since each frame of audio is also associated with a reference nearfield speech activity label, the performance of an automatic system can be expressed concisely using a 2×2 confusion matrix; the standard names for the cells of this matrix are shown in Table 11.1.

11.4.1 Metrics

As is customary and convenient, from any given confusion matrix, a *scalar* objective function is computed to facilitate system comparison. There are a number of objective functions currently in use.

A popular metric is the *error rate* (ER). It involves the frame false alarm rate (FA),

$$\text{FA} \quad = \quad \frac{\text{FP}}{\text{FP} + \text{TN}} \,, \tag{11.7}$$

and the frame miss rate (MS),

$$\text{MS} \quad = \quad \frac{\text{FN}}{\text{TP} + \text{FN}} \,. \tag{11.8}$$

The frame error rate is then given by their sum,

$$\text{ER} \quad = \quad \text{FA} + \text{MS} \,. \tag{11.9}$$

System improvement aims to minimize ER, to zero.

Minimization of FA is often at odds with that of MS. It is typically possible to vary a single parameter to construct systems with a range of increasing FAs corresponding to a range of decreasing MSs. The locus of the resulting alternative operating points is known as a *receiver operating curve*. The point on this curve at which FA=MS is known as the equal error rate (EER). When defined in this way, the ER at the EER point is twice the EER.

A different measure of performance is afforded by *classification error* (CER), which is given by

$$\text{CER} \quad = \quad \frac{\text{FN} + \text{FP}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}}. \tag{11.10}$$

[2]During a late stage in the writing of this chapter, it was discovered that 4 of the 51 meetings were actually not included. These are `Bed009`, `Bed011`, `Bns001`, and `Bro007`. Feature extraction failed to parse the meta-files for these meetings, because they involve participants which said nothing.

Alternately, systems may be characterized by their precision and recall. The former is given by

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}\,, \tag{11.11}$$

and the latter by

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\,. \tag{11.12}$$

Their unweighted harmonic mean is known as the F-score,

$$\begin{aligned}\mathrm{F} &= \frac{2}{\frac{1}{\mathrm{P}}+\frac{1}{\mathrm{R}}} \\ &= \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}\end{aligned} \tag{11.13}$$

The above metrics are used ubiquitously, and somewhat interchangeably, in many pattern classification and signal detection domains. A metric introduced by NIST, specifically for assessing speaker diarization performance, is known as the *NIST diarization error*, which will be referred to as "NISTERR" in this thesis. It is the sum of the NIST-defined false alarm rate (NISTFA) and the NIST-defined miss rate (NISTMS),

$$\mathrm{DER} = \mathrm{NISTFA}+\mathrm{NISTMS}\,; \tag{11.14}$$

it represents the fraction of speaker time which is not correctly attributed to a speaker. NISTERRs are computed in this chapter using a scoring script provided by NIST (`md-eval-v18a.pl`), with a forgiveness collar of 0.25 s around speech/non-speech transition boundaries.

This chapter will use EER as its primary metric[3]. Additionally, Section 11.12 reoptimizes systems for the alternative ER, CER, F-score, and NISTERR metrics.

11.4.2 ROC Construction

ROC curves are constructed by varying a parameter or a set of parameters which controls the amount of positives that a system posits. Since ROC curve construction is implicit in the computation of the equal error rate, this subsection outlines the process used throughout this chapter.

ROC curves and EERs were computed by varying the *smoothing policy* applied to SAD output following decoding. It should be noted that in classification scenarios in which items are classified independently (in constrast to consecutive frames in an audio stream), ROC curves are most often constructed by varying the threshold used in likelihood comparison. However, in a sequence decoding scenario, modifying parameters internal to Viterbi search would have a significant impact on time complexity. Creating ROC curves via smoothing policy perturbations was therefore chosen for expediency reasons.

The smoothing policy σ used in the experiments of this chapter is parametrized using two values, the pre-padding parameter and the post-padding parameter. They are applied to extend the duration of contiguous intervals of speech, fore and aft of where it is actually detected, respectively. The values are drawn from the ranges $[0.0, 1.0]$, with a granularity of 0.02 s. An exhaustive search for an optimal σ therefore includes $51 \times 51 = 2601$ trials.

[3] [56] observed that the sum of false alarms and misses was correlated with word error rates (WERs); however, in that work false alarms and misses were normalized differently than here, rendering the observation true for what are here termed CERs. [216, 19] observed a correlation between WER and the SAD error rate:

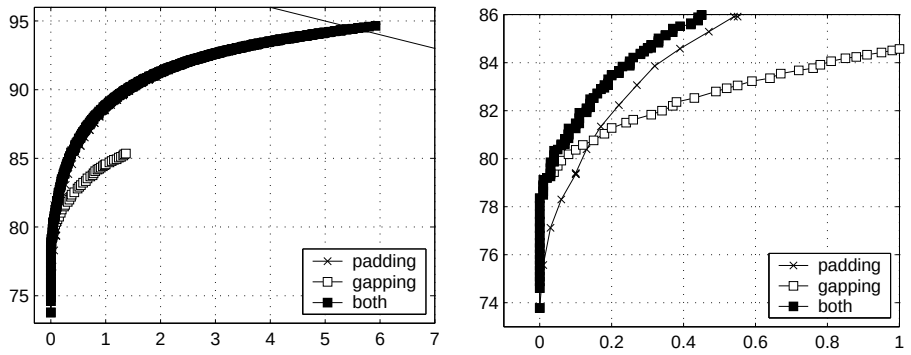
$$e_{\mathrm{SAD}} = \frac{\mathrm{FP}+\mathrm{FN}}{\mathrm{TP}+\mathrm{FN}}\,, \tag{11.15}$$

also known as the speech diarization error rate (SDER).

11.4.3 References

Both DevSet and EvalSet are accompanied by a reference speech/non-speech segmentation Υ^U; this is referred to as the *utterance-level segmentation* in what follows. The corpora are also accomanied by word-level forced-alignment, making possible the construction of an alternate *word-level segmentation* Υ^W. This was done by marshalling temporally adjacent words and word fragments into contiguous speech intervals, retaining all inter-word gaps as non-speech intervals. A comparison of Υ^W, as the output of a hypothetical, perfect acoustic SAD system, with Υ^U on DevSet yields an ER = MS = 26.21% (CERR = 5.54%, F-score = 84.92%). This means that over a quarter of the time during which participants are nominally producing an utterance, they are actually not producing any speech.

An ROC for Υ^W, when scored against Υ^U, is shown in Figure 11.1. The figure shows not only curves plotted using the 2-parameter "padding" smoothing policy σ_1, but also an alternative 2-parameter "gapping" policy σ_2. Instead of padding, the latter bridges gaps shorter than its first parameter and then eliminates spurts shorter than its second parameter. It can be seen that neither policy is able to achieve the desired mapping $\sigma_* : \Upsilon^W \mapsto \Upsilon^U$.



(a) (b) close-up of (a)

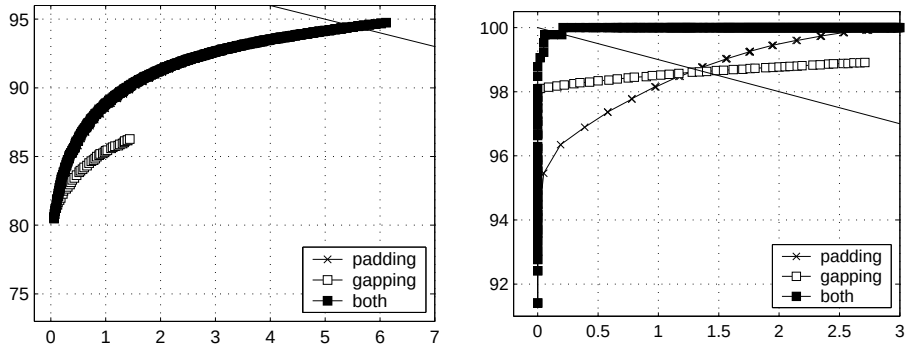
Figure 11.1: DevSet ROC curve with word-level references Υ^W plotted as the hypothesized output against utterance-level references Υ^U.

As a result of these observations, an intermediate segmentation was formed against which performance is computed in the remainder of this chapter. That segmentation is the *talkspurt segmentation* Υ^T. It was formed early in the writing of this thesis, by starting with Υ^W and then selecting a $\sigma_T = \sigma_1 \circledast \sigma_2$, where $\circledast$ denotes composition, which minimizes the WER on DevSet (an ASR metric). Manual optimization identified a σ_T which: (1) bridges all non-speech gaps shorter than 0.375 s; (2) eliminates no speech intervals, regardless of their duration; (3) pre-pads all speech intervals with 0.015 s; and (4) post-pads all speech intervals with 0.025 s. First-pass DevSet WERs, using the meeting ASR system described in [149], were 1.2%abs lower for the resulting Υ^T than for the manually created utterance reference segmentation Υ^U.

For completion, ROC curves for Υ^T scored against Υ^U, and for Υ^W scored against Υ^T, are shown in Figure 11.2. As panel (a) shows (in comparison to Figure 11.1), Υ^T does not appear to differ much from Υ^W when scored agasinst Υ^U. Panel (b) indicates that a policy which first performs "gapping" and then "padding" is able to reproduce Υ^W from Υ^T with high fidelity; the reason that the curve does not include the point $(0, 100)$ is because the parameters governing σ_T, described in the preceding paragraph, have a granularity of 0.005 s. Gapping by itself yields higher recall than does padding by itself, at the same FAs, but only there where FA < MS. For FA > MS, the trend is reversed.

11.5 Baseline

A standard HMM decoder is the SAD baseline for this chapter. A second, alternative baseline is also used, and is formed by appending crosstalk suppression features to the feature vector of the first baseline.



(a) Υ^T (hyp) against Υ^U (ref) (b) Υ^W (hyp) against Υ^T (ref)

Figure 11.2: DevSet ROC curves for pairs of reference segmentations, one of which is treated as hypothetical ("hyp") system output.

11.5.1 The 16-ms HMM Decoder

HMM decoding of speech activity was described in [1]. The following is a description of an internal implementation of that decoder, which operates at a frame step of 16 ms. This is a ubiquitous frame step in many speech processing applications.

The HMM topology is shown in Figure 11.3; it consists of two acoustically unique states, one for nearfield-speech ($\mathcal{S} \equiv \blacksquare$) and one for non-nearfield-speech ($\neg\mathcal{S} \equiv \square$). These states, and their acoustic models, are replicated to enforce minimum sojourn times T_{min} of 0.5 s, identically for both nearfield-speech and nearfield-non-speech.

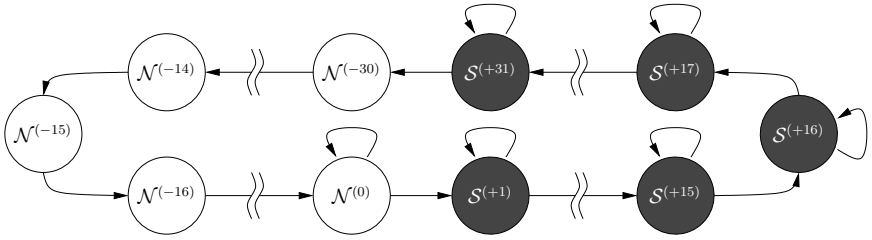


Figure 11.3: A single-participant HMM topology for SAD decoding.

Channel audio is pre-emphasized using a first-order FIR filter, $1 - 0.97z^{-1}$, and framed at increments of 16 ms into 32 ms Hamming window frames. The pre-emphasis filter, Hamming window shape, and the 50% frame overlap are common parameter settings in speech processing. For each resulting frame, log-energy (LE) and MFCC coefficients are computed. Only the first through to the twelth MFCC coefficients are retained; the zeroth coefficient is discarded. To the resulting 13-element feature vector are appended its 13 first- and second-order differences, for 39 features in total. This is the standard ASR feature vector, and will be referred to in subsequent sections as LE+MFCC+Δ+$\Delta\Delta$.

Emissions of the feature vector are modeled using a 256-element GMM, trained using EM to maximize the log-likelihood of TrainSet. Training stops once the first-order difference in log-likelihood falls below 0.001% of its previous value. Early investigations, using different data and a different error function, had revealed that performance improves monotonically at least to GMMs of 1024 components, but only minimally past 256 components, and at much higher time-complexity.

This baseline system achieves a false alarm rate of 4.97% and a miss rate of 9.96%, for an ER of 14.93%. After smoothing, the lowest error rate observed along the equal error line is 12.71% (this implies an equal error rate, as defined in this thesis, of 6.35%).

11.5.2 Inclusion of Features for Crosstalk Suppression

This chapter also considers a second baseline, which is like the first but additionally models a pair of crosstalk suppression feature known as Normalized Log-Energy Differences (NLEDs). These features, proposed in [20], are the minimum and maximum over inter-channel differences. They are among the best-performing crosstalk suppression features as used in GMM-based independent-participant SAD decoding.

This baseline, referred to as LE+MFCC+Δ+ΔΔ+NLED or simply +NLED, exhibits a DevSet false alarm rate of 2.81%, a miss rate of 8.30%, and a resulting ER of 11.12% at the decoder output. The EER is 4.49%. The errors committed by this system, in all respects, are lower than those committed by the first baseline.

11.6 Increasing the Frame Step

The ultimate aim of this chapter is to demonstrate that modeling interlocutors has a significant impact on speech activity detection in multi-party conversations. This will be attempted, as described in subsequent sections, by decoding the joint, multiparticipant state of a conversation. However, if every participant can be in one of the N_S states in Figure 11.3, then a conversation of K participants can be in one of N_S^K states, a number which is intractable even for reasonable K. The purpose of this section is to limit the size of the resulting multi-participant topology.

A simple means of reducing the number of degrees of freedom of any given participant is to increase the frame step of the decoder, but to retain the topological minimum durations constraints (of 0.5 s in the baseline). This is explored below, followed by a re-optimization of the minimum duration parameters and then of the frame size. To not miss audio, a larger frame step will require a larger frame size; very little is currently known about whether MFCC features extracted from large frame sizes retain their ability to discriminate between different vocal activity states.

11.6.1 Optimizing the Frame Step

The experimental suite in this subsection retains the baseline decoder structure, but the frame step is increased to one of 32 ms, 50 ms, 75 ms, 100 ms, and 200 ms. In each case, the 50% frame overlap is also retained, yielding frame sizes of 64 ms, 100 ms, 150 ms, 200 ms, and 400 ms, respectively. No other parameters are modified; the suite is exectuted twice, for the LE+MFCC+Δ+ΔΔ baseline and for the LE+MFCC+Δ+ΔΔ+NLED baseline. The results are shown in Figure 11.4.

As can be seen in this figure, there appears to be no evidence that small frame sizes yield better SAD performance. The lowest EERs are achieved, for both baselines, with a frame step of 50 ms. For the purposes of this work, however, in order to facilitate multi-participant modeling, a frame step of 100 ms is henceforth adopted. When NLED features are not used, this choice entails an EER increase of approximately 0.5%abs. However, for the much better baseline which does rely on NLED features, a frame step of 100 ms achieves an EER which is approximately 0.1%abs lower than the corresponding 16 ms baseline.

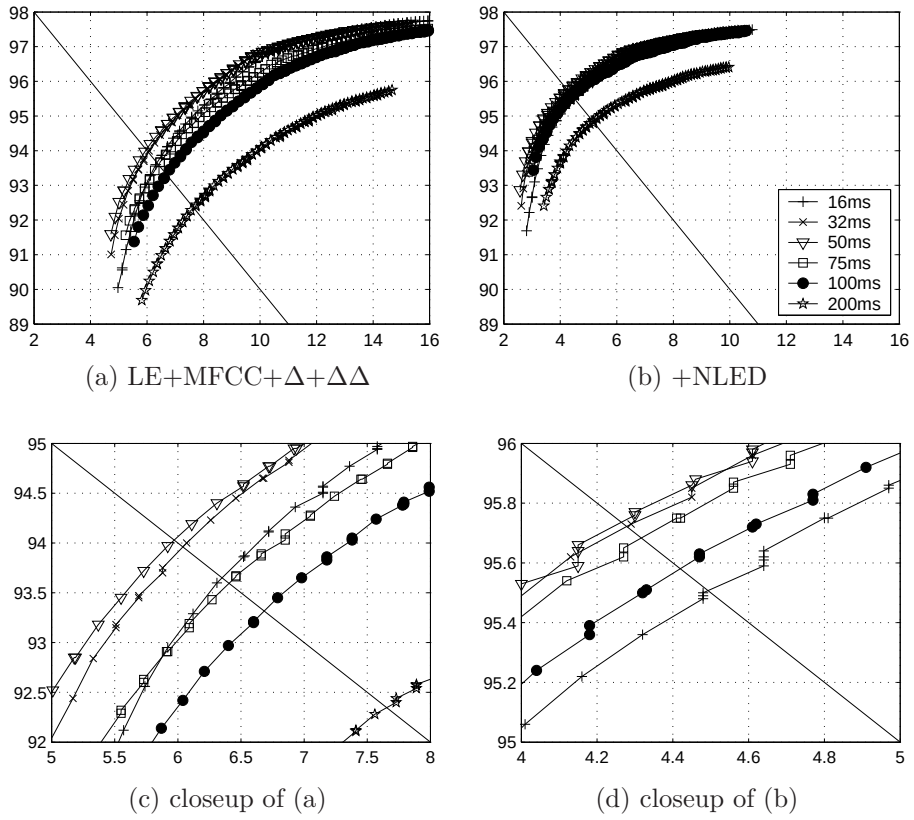
The aforementioned results have two implications. First, they suggest that MFCC features computed using large frame sizes are no worse for SAD than those computed using traditional frame sizes of 16 ms. Second, the loss of granularity at edges, due to a coarser frame step, appears to have at worst only minimal impact on performance.

11.6.2 Optimizing Topological Minimum Duration Constraints

Given the chosen frame step of 100 ms, and the corresponding frame size of 200 ms, this section investigates the effect of minimum duration constraints on speech and non-speech sojourn times in the HMM topology. 400 alternate topologies are exhaustively constructed with constraints

$$\begin{aligned} T_{min}^{\mathcal{S}} \quad &\in \quad [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, \\ &\qquad 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9, 2.0] \ , \end{aligned} \tag{11.16}$$

$$\begin{aligned} T_{min}^{\neg\mathcal{S}} \quad &\in \quad [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, \\ &\qquad 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9, 2.0] \ , \end{aligned} \tag{11.17}$$

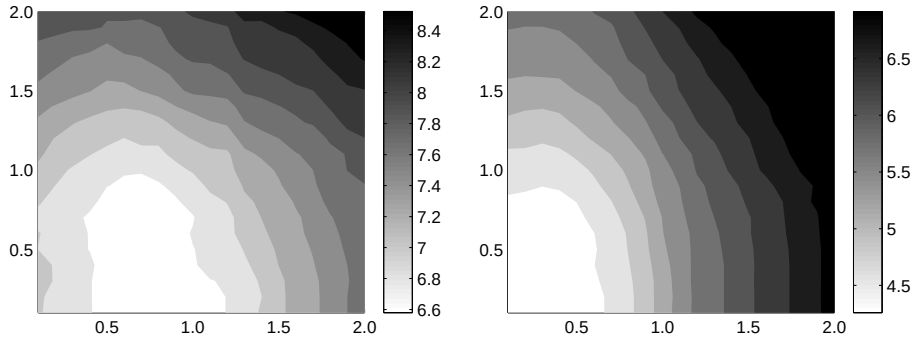


(a) LE+MFCC+Δ+ΔΔ (b) +NLED

(c) closeup of (a) (d) closeup of (b)

Figure 11.4: DevSet ROC curves for the two baseline systems, with several frame steps and 50% frame overlap.

for speech and non-speech, respectively. Transition probabilities are retrained using TrainSet for every $(T^{\mathcal{S}}_{min}, T^{\neg\mathcal{S}}_{min})$ pair, prior to decoding DevSet. The results are shown in Figure 11.5 for both systems from the preceding section; lighter shades of gray indicate better performance.



(a) LE+MFCC+Δ+ΔΔ (b) +NLED

Figure 11.5: DevSet EER surfaces for various combinations of mimimum duration constraints placed on speech (along the x-axis) and on non-speech (along the y-axis). Ligher shades of gray indicate lower EERs. Both systems shown operate at a frame step of 100 ms, with 50% overlapping frames.

Figure 11.5 shows that minimum duration constraints play a role in suppressing crosstalk, since their optima are different when crosstalk suppression features are employed. Without NLED features, in panel (a), the optimal minimum duration of speech is at 0.8 s; with NLED features, the optimum is at 0.4 s. The optimal constrains for minimum duration of non-speech are approximately the same for both systems. In order to use exactly the same topology for descendants of both system types, minimum duration constraints of 0.5 s on both speech and non-speech are retained throughout the remainder of this chapter, until Section 11.10 where they are re-optimized individually for all systems.

11.6.3 Optimizing the Frame Size

Finally, as a third experiment in this section, a frame size is selected for the new frame step and minimum duration constraints determined in previous subsections. This potentially loosens the 50% frame overlap stipulation made earlier. SAD performance is explored for frame sizes of 100 ms, 150 ms, 200 ms, and 250 ms. The results are depicted in Figure 11.6.

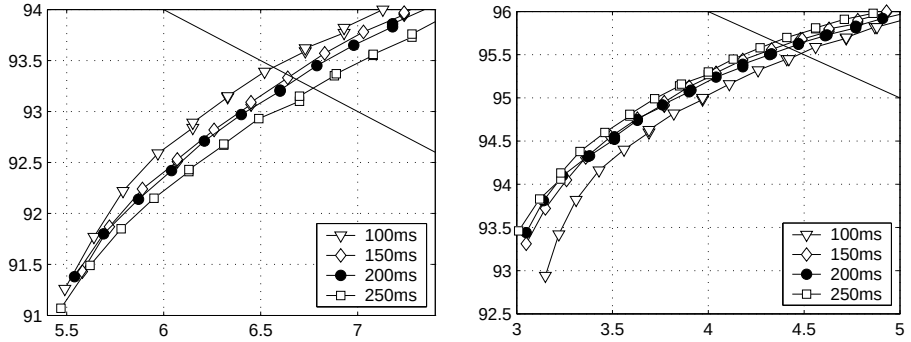


Figure 11.6: DevSet ROC curves for the two baseline systems, with several frame sizes. All systems employ a frame step of 100 ms, and minimum duration constraints of 0.5 s for speech and non-speech, identically.

The figure indicates that, in the absence of crosstalk suppression features, smaller frame sizes yield lower EERs, while the opposite is true when crosstalk features are used. In fact, the EER ranks of the explored frame sizes are mirror images of each other for the two system types. As a compromise, to facilitate comparison for descendants of the two system types, a frame size of 200 ms is used in subsequent sections, which offers intermediate performance for both.

11.7 NT-Norm Cross-Correlation Maxima

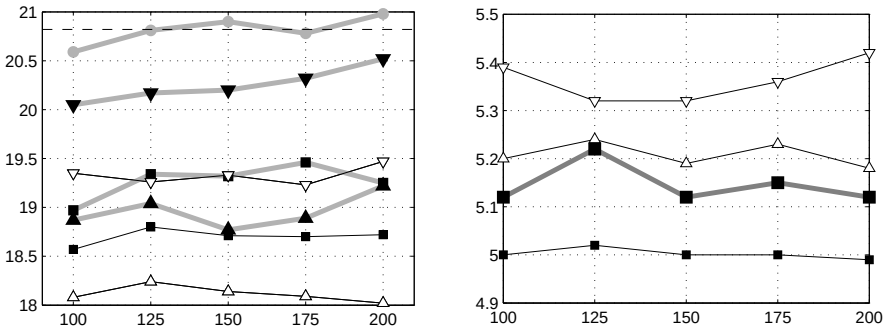
As a final set of experiments prior to modeling other participants, this section assesses the impact of employing additional crosstalk suppression features. This is important because one of the expected benefits of modeling interlocutors is that of discouraging crosstalk (which is relatively frequent) from being recognized as overlap (which is relatively infrequent). To verify that benefit in subsequent sections, the ability to model crosstalk acoustically should be exhausted insofar as possible given current techniques. The alternative crosstalk suppression feature, described below, is based on the non-target normalized cross-correlation maxima of Section 9.2.

11.7.1 Using a Rectangular Window

Section 9.2 described how, given channels corresponding to the microphones worn by participants k and j, a feature describing k can be computed by normalizing the maximum of the cross-correlation between channels k and j by the non-target signal j. What remains to be determined is the duration of the window over which the cross-correlation spectrum is computed, and the specific features extracted. The number of the latter should be independent of the number

of participants to the conversation, thereby yielding a fixed-length feature vector, in order to be able to train speech and non-speech models deployable in conversations of arbitrary participant number.

Given a conversation of K independently instrumented participants, the number of NT-norm cross-correlation maxima for each channel is $K-1$. As explained in Section 9.2, the minimum and the average over these $K-1$ quantities have direct geometric interpretations, under reasonable assumptions. On the other hand, the maximum over these $K-1$ quantities is not known to offer a similar interpretation at the current time. To verify the geometric motivation, all seven combinations of summary statistics are explored: average (XAVE); minimum (XMIN); maximum (XMAX); average and minimum (XAVE+XMIN); average and maximum (XAVE+XMAX); minimum and maximum (XMIN+XMAX); and average, minimum, and maximum (XAVE+XMIN+XMAX). These features are combined in feature space with the 39 LE+MFCC+Δ+$\Delta\Delta$ features, as was done for NLED features. Cross-correlations are computed over a series of alternative window sizes, namely 100 ms, 125 ms, 150 ms, 175 ms, and 200 ms. The results, when the computation is performed using a rectangular window, are shown in Figure 11.7.



(a) ER (b) EER

Figure 11.7: DevSet performance, as a function of rectangular window size used for computing cross-correlation (along the x-axis), and several combinations of statistics computed over $K-1$ NT-Norm features per channel. Error rates at the decoder output, prior to smoothing, are shown in panel (a), while equal error rates are shown for a subset of statistic combinations in panel (b). —▽— : XAVE; —△— : XMIN; —●— : XMAX; —■— : XAVE+XMIN; —▼— : XAVE+XMAX; —▲— : XMIN+XMAX; —■— : XAVE+XMIN+XAVE.

It appears from panel (a) in the figure that the error at the decoder output (ER) is not very sensitive to the window duration over which cross-correlations are computed. However, there are relatively large differences among the combinations. The best combination is XMIN alone, followed by XAVE+XMIN; the difference between these two is approximately 0.5%abs over the entire window range from 100 ms to 200 ms. On average, combinations involving XMAX are weaker than those not involving the maximum.

Panel (b) shows the equal error rates, achieved by optimizing a smoothing function σ independently for each combination. Only the four combinations achieving the lowest EERs, in the range 100 ms to 200 ms, are shown. These are XAVE+XMIN, XAVE+XMIN+XMAX, XMIN, and XAVE; the remaining three combinations of NT-Norm features are not considered further in this chapter. The optimal window length for computing cross-correlations, given these four curves, appears to be approximately 150 ms.

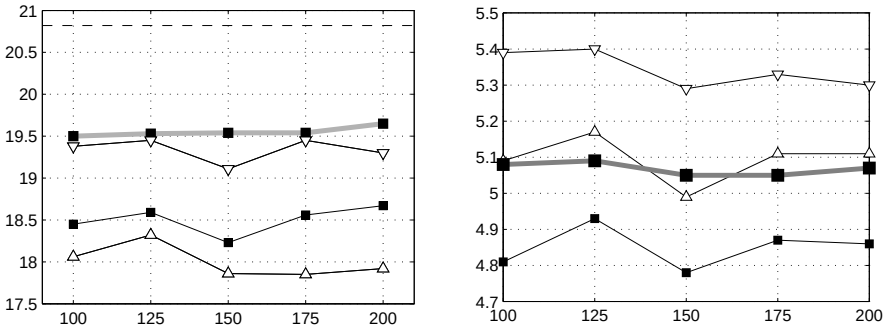
11.7.2 Using a Hamming Window with Correlation Spectrum Envelope Compensation

The above experiments are now repeated, but with a Hamming envelope rather than a rectangular envelope. It should be noted that the maximum possible value for the cross-correlation maximum is a function of the lag at which it is found. This occurs for two reasons. First, the spectrum is the result of a convolution; at non-zero lags, there is less overlap between the supports of the two signals than there is at zero lag. Second, and more importantly, the effect is due to the taper of the Hamming window itself. It is therefore important to compensate for this by element-wise multiplication of the observed cross-correlation spectrum with a lag-specific rectifying factor which reflects the maximum possible value at

NT-Norm Features	w/o NLED	with NLED
none	6.68	4.42
XAVE	5.29	4.38
XMIN	4.99	4.35
XAVE+XMIN	4.78	4.43
XAVE+XMIN+XMAX	5.05	4.47

Table 11.2: DevSet equal error rates (EERs) for various combinations of crosstalk suppression features, in addition to the baseline LE+MFCC+Δ+$\Delta\Delta$ feature vector. All NT-Norm features are computed using a window length of 150 ms; the window length used for computing all other features is 200 ms.

each lag. The results are shown in Figure 11.8.



(a) ER (b) EER

Figure 11.8: DevSet performance, as a function of Hamming window size used for computing cross-correlations (along the x-axis) with envelope compensation, and several combinations of statistics computed over $K-1$ NT-Norm features per channel. Error rates at the decoder output, prior to smoothing, are shown in panel (a), while equal error rates are shown for a subset of statistic combinations in panel (b). —▽— : XAVE; —△— : XMIN; —■— : XAVE+XMIN; ▬■▬ : XAVE+XMIN+XAVE.

The figure illustrates that the Hamming window, in combination with envelope compensation, reduces both ERRs and EERs for the best performing NT-Norm feature combination. That combination continues to be XAVE+XMIN. The EER reduction observed for it is approximately 0.2%abs, compared to panel (b) in Figure 11.7. The optimal window width for this combination is 150 ms; for simplicity, this width will be retained when computing any NT-Norm features in the remainder of this chapter.

11.7.3 Combination of NT-Norm and NLED Features

Finally for this section, the four best NT-Norm features are combined with NLED features in the context of the LE+MFCC+Δ+$\Delta\Delta$ feature vector. The results are shown in Table 11.2.

As can be seen, adding XAVE features to XMIN features reduces EERs, by 0.21%abs, to 4.78%. This is the best performance achievable without NLED features. However, when NLED features are used, including more than one NT-Norm features leads to worse performance, while including only one of either XAVE and XMIN helps (the improvements due to NT-Norm feature inclusion, on top of NLED features, are quite small and likely not statistically significant).

For these reasons, the remainder of this chapter will only consider 6 feature combinations: LE+MFCC+Δ+$\Delta\Delta$ (39 features), the baseline feature vector with no crosstalk suppression features; +XAVE ($39+1$ features); +XAVE+XMIN ($39+2$ features); +NLED ($39+2$ features); +NLED+XAVE ($39+3$ features); and +NLED+XMIN ($39+3$ features).

11.8 Modeling Interlocutors

As argued throughout this thesis, modeling the temporally proximate speech activity of interlocutors stands to benefit speech activity detection, because participants are less likely to continue speaking in overlap with other speakers. This mechanism, of avoiding overlap, runs precisely counter to the effect that unmitigated crosstalk has, since the latter leads to hypotheses of multiple participants talking at once when they are not.

11.8.1 HMM Topology

The joint multi-participant topology is a Cartesian product of the single-participant topology shown in Figure 11.3, reproduced in Figure 11.9 with additional annotation (to be described below); given a decoder frame step of 100 ms, and two acoustically distinct states per participant whose minimum occupation time is 500 ms each, $N_S = 10$ and the complete multi-participant topology contains $N = |\mathbb{S}^K| = 10^K$ states. To render search tractable, the topology is pruned during construction, as described in Subsection 6.5.1, using 5 alternate sets of constraints denoted J_a^b. a indicates the maximum number of participants which, at the same instant, can be in any of the single-participant states corresponding to behavior ■, denoted in dark gray in Figure 11.9. b indicates the maximum number of participants which, at the same instant, can be in any state other than the "long-time inactive (non-speech)" state, identified with a double cirle in the figure as $\mathcal{N}^{(0)}$. (Using this J_a^b notation, the complete multi-participant topology, without pruning, would be denoted J_K^K.)

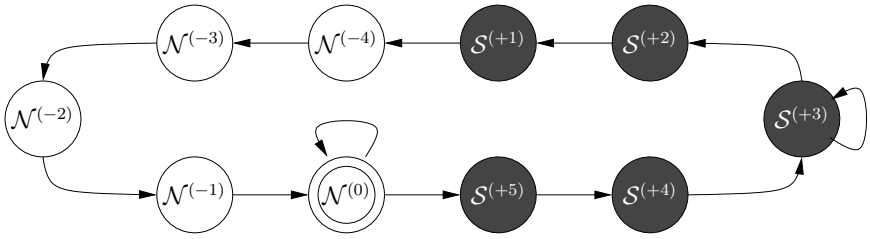


Figure 11.9: Projection of the multi-participant HMM topology onto the state sub-space of a single participant; the frame step is 100 ms. Shown are two acoustically distinct behaviors, each duplicated 5 times to enforce a minimum duration of 500 ms on both behavior types. $\mathcal{N}^{(0)}$ represents the default, "long-time inactive" state.

The five sets of constraints considered, leading to five multi-participant topologies which will be identified using the same notation, are J_1^1, J_1^2, J_2^2, J_2^3, and J_3^3. The last of these topologies, J_3^3, is of course the most permissive; J_1^1, on the other hand, allows at most one participant to be speaking (or to be about to start speaking or to be just finished speaking) at a time. The effect that these pruning constraints have on the number of surviving multi-participant states $|\mathbb{S}_*^K|$ is shown in Figure 11.10. As can be seen in this figure, the number of states in the most permissive J_3^3 topology never exceeds 10^5, for $K \leq 10$. It is also evident that allowing one more participant to be in single-participant non-speech states just after or just prior to single-participant speech states increases the topology complexity by almost as much as placing no constraints at all on that participant.

11.8.2 Transition Probability Model

Transition probabilities, $P(\mathbf{q}_t = \mathbf{S}_j \,|\, \mathbf{q}_{t-1} = \mathbf{S}_i)$ for $\mathbf{S}_i \in \mathbb{S}_*^K$ and $\mathbf{S}_j \in \mathbb{S}_*^K$, for each of the 5 topologies proposed, are provided by the generalized EDO model of Section 6.5. The model is trained using TrainSet, exactly the same corpus as was used for training the single-participant transition probability model of Section 11.5.

11.8.3 Multimicrophone Acoustic Model with 2 States per Microphone

To complete the definition of the multi-participant HMM decoder for the current task, an acoustic model is needed which provides the emission probabilities $P(\mathbf{X}_t \,|\, \mathbf{S}_i)$. Here, $\mathbf{X}_t$ is a feature *matrix*, consisting of K column feature vectors, each

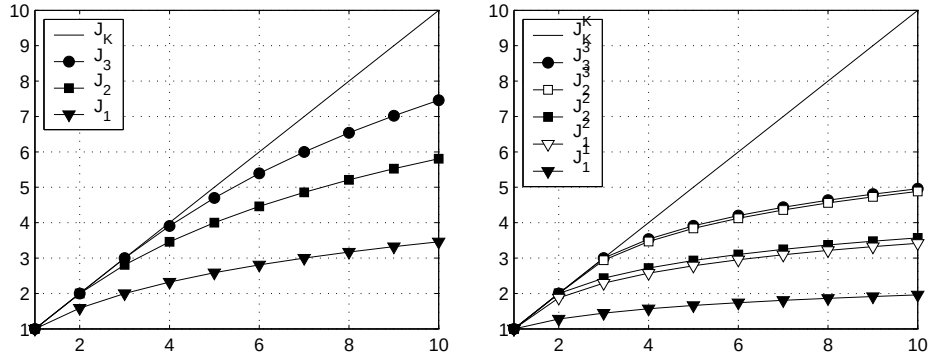


Figure 11.10: Number of multi-participant states, along the logarithmically scaled y-axis, as a function of the topology type J_a^b and the number K of participants, along the x-axis. (a) shows $\log_2 |\mathbb{B}_*^K|$, the number of multiparticipant behaviors after pruning, while (b) shows $\log_{10} |\mathbb{S}_*^K|$, the number of multiparticipant states after pruning. Logarithm base corresponds to $N_B = 2$ and $N_S = 10$, in panels (a) and (b), respectively.

corresponding to an identical set of features computed for each channel at instant t. This likelihood can be factored, to make use of the single-channel acoustic models of the previous section,

$$P\left(\mathbf{X}_t \,|\, \mathbf{S}_i\right) \;\doteq\; \prod_{k=1}^{K} P\left(\mathbf{X}_t\,[k] \,|\, \Theta_{\zeta(\mathbf{S}_i,k)}\right) \tag{11.18}$$

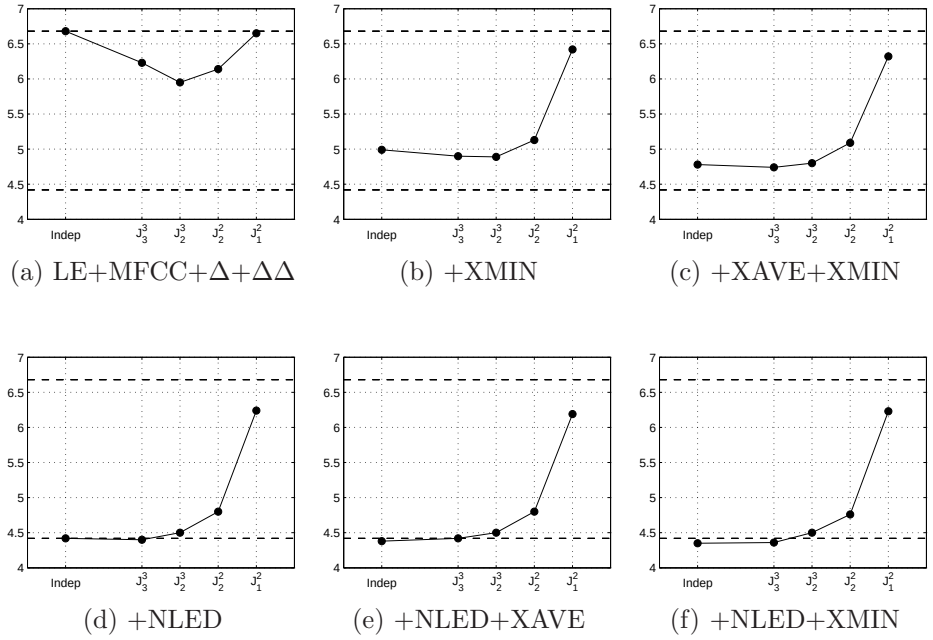
where

$$\zeta\,(i,k) \;\equiv\; \begin{cases} \square, & \text{if } \mathbf{S}_i\,[k] = \square \\ \blacksquare, & \text{if } \mathbf{S}_i\,[k] = \blacksquare \end{cases} . \tag{11.19}$$

In this manner, although there are $|\mathbb{S}_*^K|$ multi-participant states, only two emission probability models are needed[4], those conditioned on each of the two single-participant behaviors speech ($\blacksquare$) and non-speech ($\square$). More correctly, the two models are conditioned on (close-talk) microphone state, namely "presence of nearfield speech" and "absence of nearfield speech", respectively (and these are inferred from the speech activity behavior of their wearers). This 2-state microphone model variant is henceforth referred to as "$\square\blacksquare$".

EERs for the system described so far, for 6 feature set combinations, $\square\blacksquare$ factorable acoustic models, and all 5 multi-participant topology types, are shown in Figure 11.11. Also shown (using dashed lines) are the EERs for comparable independent-participant decoders using (1) standard log-energy, MFCC, and first- and second-order difference features, as well as (2) that same feature set augmented with NLED features. What is obvious in panel (a) is that when no crosstalk suppression features are used, decoding participants jointly lowers EERs, with a minimum for the J_2^3 topology. This is because allowing at most 2 participants to speak decreases the false alarm rate by an amount which is larger than the corresponding miss rate increase. When NT-Norm features are used but NLED features are not, joint decoding lowers EERs but by a much smaller amount than when no such features are used. Finally, when NLED features are modeled, joint decoding appears to have negligible impact, provided the most permissive J_3^3 of the attempted topologies is used. For less permissive topologies, performance is worse than in the independent-participant decoder setting.

[4]This *factored* emission model type, together with the J_K^K topology (if it were not prohibitive to construct), with the gEDO transition probability model trained under an assumption of unconditional participant independence, should lead to results identical to those obtained with the single-participant decoder of the previous section. It is therefore the extent to which participants are *not* independent, in the context of topologies J_a^b for $a < K$ and $b < K$, which explains why the results in this and subsequent sections deviate from that of the previous section.



(a) LE+MFCC+Δ+ΔΔ (b) +XMIN (c) +XAVE+XMIN

(d) +NLED (e) +NLED+XAVE (f) +NLED+XMIN

Figure 11.11: DevSet equal error rates (EERs, along the y-axis) for systems with multi-participant emission probabilities factorable into products conditioned on 2-state microphone states (denoted "□■" in the text). Dashed lines indicate the performance of the 100-ms independent-participant systems with and without NLED features (lower and upper lines, respectively).

11.8.4 Multichannel Acoustic Models with More Than 2 States per Microphone

In the joint-participant decoder setting, microphone state models other than the one in Equation 11.19 are possible[5]. Each microphone state may be conditioned not just on the speech activity state of its wearer, but also on the speech activity of other participants. Two such variants have been explored in previous work.

First, a model which considers 3 mutually exclusive categories ("nearfield (wearer) non-speech in the absence of farfield speech", "nearfield non-speech in the presence of farfield speech", and "neafield speech") was considered in [98]. This model will henceforth be referred to as the □◪■ microphone state model; it involves the same Equation 11.18 but with

$$\zeta(i,k) \equiv \begin{cases} \square, & \text{if } \mathbf{S}_i[k] = \square \text{ and } \mathbf{S}_i[j] = \square \ \forall j \neq k \\ \blacksquare, & \text{if } \mathbf{S}_i[k] = \square \text{ and } \exists j \neq k \text{ with } \mathbf{S}_i[j] = \blacksquare \\ \blacksquare, & \text{if } \mathbf{S}_i[k] = \blacksquare \end{cases} \tag{11.20}$$

Second, a model which considers 4 mutually exclusive categories was proposed in [227]. This model is identical to the □◪■ model, but breaks out the ■ microphone state into two states, representing "nearfield (wearer) speech in the absence of farfield speech", and "nearfield speech in the presence of farfield speech"; it will be henceforth referred to as the □◪■■ microphone state model. Formally,

$$\zeta(i,k) \equiv \begin{cases} \square, & \text{if } \mathbf{S}_i[k] = \square \text{ and } \mathbf{S}_i[j] = \square \ \forall j \neq k \\ \blacksquare, & \text{if } \mathbf{S}_i[k] = \square \text{ and } \exists j \neq k \text{ with } \mathbf{S}_i[j] = \blacksquare \\ \blacksquare, & \text{if } \mathbf{S}_i[k] = \blacksquare \text{ and } \mathbf{S}_i[j] = \square \ \forall j \neq k \\ \blacksquare, & \text{if } \mathbf{S}_i[k] = \blacksquare \text{ and } \exists j \neq k \text{ with } \mathbf{S}_i[j] = \blacksquare \end{cases} \tag{11.21}$$

[5]They are also possible in independent-participant decoder settings, of course, but then require consolidation after decoding. Consolidation is often based on hand-crafted rules [98, 227].

Finally, for completion, a mutually exclusive microphone space which has not been considered in previous work is the 3-state space □■■, whose precise formulation is

$$\zeta(i,k) \equiv \begin{cases} \square, & \text{if } \mathbf{S}_i[k] = \square \\ \blacksquare, & \text{if } \mathbf{S}_i[k] = \blacksquare \text{ and } \mathbf{S}_i[j] = \square \ \forall j \neq k \\ \blacksquare, & \text{if } \mathbf{S}_i[k] = \blacksquare \text{ and } \exists j \neq k \text{ with } \mathbf{S}_i[j] = \blacksquare \end{cases} \tag{11.22}$$

The three alternative microphone state spaces □□■, □□■■, and □■■ are related to the □■ microphone state space as shown in Figure 11.12.

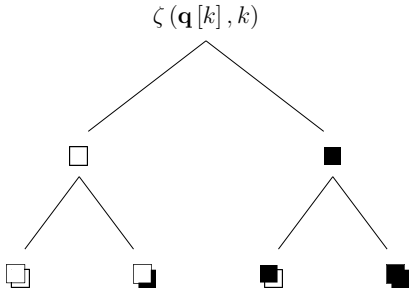


Figure 11.12: Alternative microphone state values for multi-participant decoding with only two values of single-participant behavior. The "□■" space involves only the microphone states □ and ■. The "□□■" space involves the leaves of the left branch, the "□■■" space involves the leaves of the right branch, and the "□□■■" space involves the leaves of both branches.

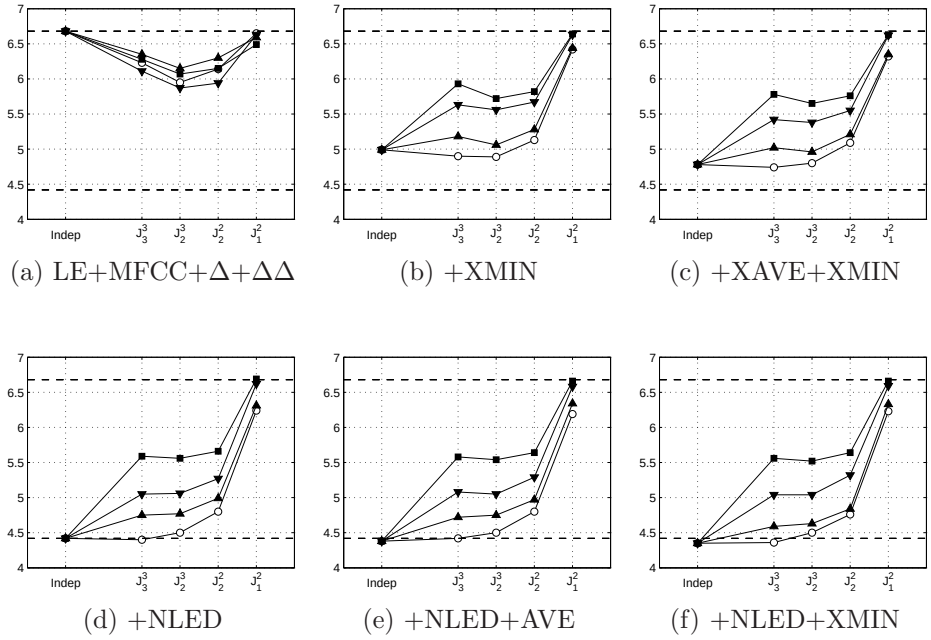
EERs for all four microphone state models, with all 6 feature sets considered earlier and all 5 joint-participant topologies, are provided in Figure 11.13. For feature sets which include crosstalk suppression features, the best microphone model is the □■ model already explored in Subsection 11.8.3; more complex models bring no benefit for any of the explored topology types. For the system without crosstalk suppression features, depicted in panel (a), the □□■ is consistently better than □■, for all topology variants.

A conclusion from these experiments is that considering more that two states per microphone improves DevSet performance only in the absence of crosstalk suppression features. To investigate why this might be the case, Figure 11.14 shows the distribution of standard log-energy features and that of NLED minimum features, for the 4 microphone states □, □, ■, and ■. What appears to be the case is that □ and ■ are further apart in log-energy than are □ and ■. For the NLED-MIN feature, □ and ■ occupy intermediate values between □ and ■. This suggests that when NLED features are used, the intent of which is to ensure that crosstalk (□) not be mistaken for overlap (■), the □□■■ acoustic model confuses ■ with □.

11.9 Decorrelating Energy Features

The preceeding sections in this chapter have demonstrated that cross-channel energy features improve the performance of a speech activity detector. The underlying problem is of course that energy features are correlated across microphones deployed in a shared acoustic space. Cross-channel features, when appended to standard single-channel MFCC features and associated with a particular channel, appear to go some distance to disambiguate the source. However, the factored multiparticipant emission probability models introduced in Section 11.8 still ignore correlations *across* channels. The proposed decoding framework offers broad scope for fielding acoustic models which model the channels jointly, in addition to modeling participant state transitions jointly. Joint multichannel modeling enables a comprehensive account of energy correlation across channels.

Unfortunately, the estimation of joint multichannel acoustic models is generally intractable, in the current supervised decoding paradigm. A first problem is that conversations are not fixed in participant number. A feature vector, consisting



(a) LE+MFCC+Δ+ΔΔ (b) +XMIN (c) +XAVE+XMIN

(d) +NLED (e) +NLED+AVE (f) +NLED+XMIN

Figure 11.13: DevSet equal error rates (EERs, along the y-axis) for systems with multi-participant emission probabilities factorable into products conditioned on multi-state microphone states, as a function of topology type (along the x-axis); —○— : □■; —▽— : □□■; —△— : □■■; —■— : □□■■. Dashed lines indicate the performance of the 100-ms independent-participant systems with and without NLED features (lower and upper lines, respectively).

of concatenated features from K participants, is of a size which is naturally a function of K; this eliminates the potential for using a conversation involving 6 participants as training material when the system is required to decode the speech activity in a conversation of 5 participants. This seems absurd, and significantly ablates the amount of training material for any system by stratifying training data into K-participant subsets.

Second, even if a sufficient amount of K-participant training data were available, the feature correlations observed across channels are a function of the acoustic characteristics of the recording space, as well as of the relative locations and orientations of the participants. These sources of variation are in addition to the variation normally observed on single channels.

Finally, assuming the first two issues are adressed, standard GMMs would require a significant upgrade in complexity (in terms of number of components) to deal with the arbitrary sequence in which participant features, for any particular conversation, are marshalled into the joint feature vector. This sequence, in terms of participant talkativity, participants' microphone characteristics, and each participant's acoustic coupling with other channels, is arbitrary. Rotating participant indices is tantamount to arbitrarily shuffling MFCC coefficients in a standard acoustic vector; in principle, a GMM can handle this operation if its number of components is replicated for each possible unique permutation of features. That number of index permutations, for K participants, is $K!$.

These concerns recommend the application of *unsupervised* techniques to the acquisition of a joint multichannel acoustic model, which is specific to a particular test conversation. The approach in this section will be to:

1. Provide an initial label assignment to all frames of the test conversation.

2. Train a joint multichannel model using the multichannel feature vectors of the *test* conversation audio and the labels from (1).

3. Apply the multichannel model from (2) to relabel all frames of the test conversation.

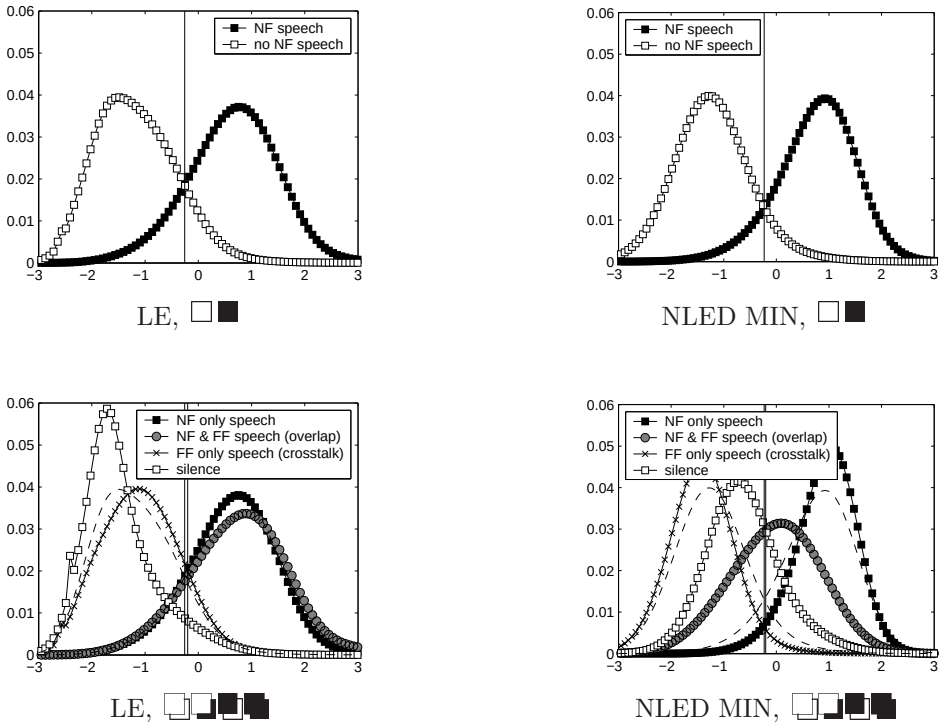


Figure 11.14: Distribution of log-energy and NLED minimum features (globally Z-normalized), for the □■ and the □□■■■ microphone state models.

The estimation of a model using the test data is subject to the duration of the test data. Since the data used in these experiments consists of 10-minute meeting snippets from the NIST RT evaluations, joint multichannel single-Gaussian models are estimated for LogEnergy only.

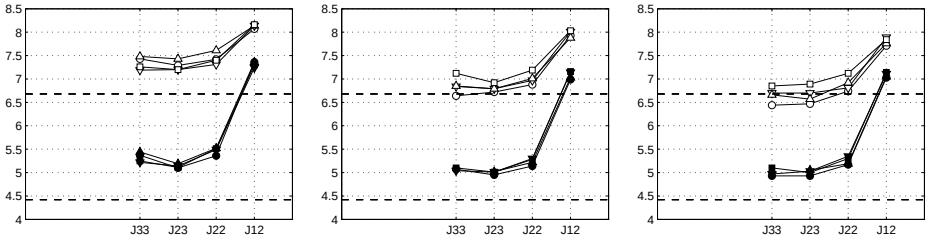
11.9.1 Initial Label Assignment Using Supervised Models

An obvious means of achieving step (1) above is to use the automatic labels obtained using the supervised factored acoustic model of Section 11.8. It should be noted that the above 3 steps lead to a joint multichannel model, where only a factored multichannel model existed; strictly speaking, they implement neither an adaptation nor a re-estimation of parameters (as these two terms are frequently used).

Figure 11.15 shows the EERs for systems whose initial label assignment is provided by the factored supervised acoustic models of Section 11.8, over 6 feature sets. Unfilled (white) markers indicate performance when the single multichannel model estimated in step (3) has only a diagonal covariance; filled markers, in constrast, indicate performance of models which have a full covariance matrix. The difference in performance between the diagonal- and the full- covariance system variants is due to conceptually decorrelating the log-energies observed across all channels. As can be seen in all 6 panels of Figure 11.15, that difference is approximately 1.5-2.0%abs in equal error rate. The full-covariance log-energy-only systems whose initial assignment is based on Viterbi passes employing crosstalk suppression features achieve EERs in the range 4.70-4.95%. Systems not using such features during initial assignment (panel (a)) achieve EERs slighly in excess of 5%.

11.9.2 Alternative Initial Label Assignment

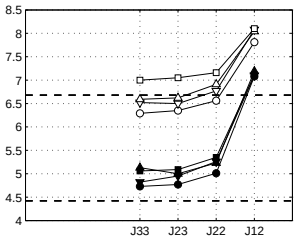
Alternately, step (1) in the preceding discussion can be implemented using a novel interpretation of the XAVE or XMIN features, as explained in Section 9.2. This leads to two competing algorithms, henceforth referred to as ILA(XAVE) and



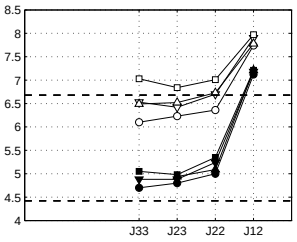
(a) LE+MFCC+Δ+ΔΔ (5.10 at J_2^3)

(b) +XMIN (4.95 at J_2^3)

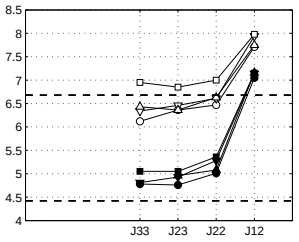
(c) +XAVE (4.93 at J_2^3/J_3^3)



(d) +NLED (4.73 at J_3^3)



(e) +NLED+XAVE (4.70 at J_3^3)



(f) +NLED+XMIN (4.76 at J_2^3)

Figure 11.15: DevSet equal error rates (along the y-axis) for systems relying exclusively on multichannel log-energy models, obtained following a first Viterbi pass relying on supervised models of 6 different feature combinations, as a function of joint topology type (along the x-axis). Unfilled (filled) markers represent diagonal (full) covariance models of multichannel log-energy; —●—, —▼—, —■—, and —▲— correspond to □■, □◪■, □◪■■, and □■■ microphone state models, respectively. Absolute minima in each panel were achieved exclusively by the □■ variant. Dashed lines indicate the performance of the (first-pass) 100-ms independent-participant systems with and without NLED features (lower and upper lines, respectively).

ILA(XMIN). The initial labels provided by either algorithm can be used to train models of both observed multiparticipant states, and also of unobserved multiparticipant states (cf. Section 9.2). This approach, if successful for this task, would not require supervised acoustic models, and would therefore eliminate the need to train acoustic models if used without subsequent model interpolation.

The approach does however require the estimation of 4 parameters, governing the ways in which the labeled data is shared among the models for both observed and unobserved states. Estimation of these parameters is costly; in this thesis, it has been achieved by a partly manual, adaptive 4-dimensional grid search. The fourth parameter, θ_N, is that proportion of all unlabeled frames, ranked by their energy across all K channels, which are used to train the all-silent model. The parameter vector which minimizes EER on DevSet was found to consist of $\lambda_G = 0.01$, $\lambda_R = 0.30$, $\lambda_S = 0.30$, and $\theta_N = 0.35$ for ILA(XAVE). For ILA(XMIN), the corresponding optimal values were found to be $\lambda_G = 0.01$, $\lambda_R = 0.40$, $\lambda_S = 0.30$, and $\theta_N = 0.20$.

The EER performance of a system which uses only log-energy, yielding hypotheses as described in step (3) above and obtaining the initial assignment labels from either ILA(XAVE) or ILA(XMIN) is shown in Figure 11.16. Only full-covariance variants are depicted, since in the previous subsection diagonal-covariance systems were shown to yield inferior performance.

The figure indicates that ILA(XMIN) is slightly better than ILA(XAVE), achieving an EER of 4.71. This is essentially the same performance as that obstained in Figure 11.15, panel (e), in which the first pass is provided by LE+MFCC+Δ+ΔΔ and +NLED+XAVE features. In contrast, the ILA(XMIN) (and the ILA(XAVE)) algorithm relies only on LE and the maximum cross-channel correlations for all channel pairs. To set this finding in context, Table 11.3

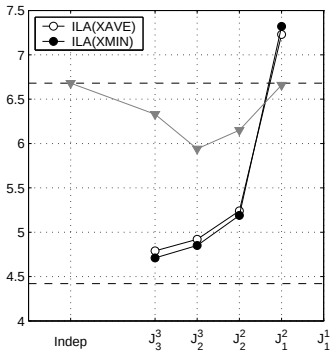


Figure 11.16: DevSet equal error rates (along the y-axis) for two systems relying exclusively on multichannel log-energy models, obtained using the ILA(XMIN) and ILA(XAVE) algorithms, as a function of joint topology type (along the x-axis). For comparison, the lowest-EER multi-participant system with supervised LE+MFCC+Δ+$\Delta\Delta$ models is depicted as ▼. Dashed lines indicate the performance of the 100-ms independent-participant systems with and without NLED features (lower and upper lines, respectively).

shows the EERs achieved individually by only the energy-related features in each feature combination explored so far. These numbers were obtained without model retraining; each line corresponds to a model which was originally trained on all features in a given combination, but then the MFCC, Δ, and $\Delta\Delta$ were excluded from likelihood computation. As such, Table 11.3 provides only a preliminary analysis of the relative strength of energy-like features.

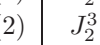
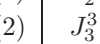
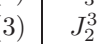
Other Features	(#)	without LE			with LE (1)		
		topo	mic	EER	topo	mic	EER
none	(0)	—	—	—	J_2^3		11.42
XMIN	(1)	J_2^3		7.24	J_2^3		7.92
XAVE+XMIN	(2)	J_2^3		6.50	J_2^3		6.81
NLED	(2)	J_3^3		6.54	J_3^3		6.56
NLED+XAVE	(3)	J_2^3		6.55	J_3^3		6.48
NLED+XMIN	(3)	J_2^3		6.46	J_3^3		6.44
MFCC'	(38)	J_2^2		6.11	J_2^3		5.87

Table 11.3: DevSet equal error rates (EERs) for several joint-participant decoders relying on combinations of crosstalk suppression features, with MFCC, Δ and $\Delta\Delta$ features excluded, with and without log-energy (LE) features. Only the best system variant, achieved with the topology ("topo") and microphone state model ("mic"), is shown. "MFCC"' represents the LE+MFCC+Δ+$\Delta\Delta$ feature vector with LE excluded.

As is evident in the table, log-energy by itself yields quite poor joint-participant decoder performance, with an EER of 11.42%. All crosstalk suppression features, by themselves, outperform log-energy by approximately 4-5%abs; their performance alongside log-energy is not very different, and often worse. In contrast, MFCC features together with MFCC and LE first- and second-order differences ("MFCC"' in the table) achieve the best EER of 6.11, when LE is excluded; including log-energy improves this number by 0.24%abs. This is interesting, because it indicates that log-energy is complementary with MFCC' features, but not (consistently) with the energy-based crosstalk suppression features.

Against this background (4.70% EER in Figure 11.15(e)), the 4.71% EER in Figure 11.16 of the two-pass system employing an unsupervised log-energy-only acoustic model appears quite remarkable. Given that the gap between diagonal- and full-covariance single-Gaussian models was shown to be approximately 1.5% in Figure 11.15, merely re-estimating models on the test data, using the ILA(XMIN) algorithm, appears to be responsible for the majority of the gap between 11.42% in Table 11.3 and 4.71%. It appears that a diagonal-covariance single-Gaussian model of only log-energy, whose EER may be estimated as $4.71 + 1.5 = 6.21$%abs, would perform better than a supervised system employing LE, NLED, and XMIN features, whose estimated EER is 6.44 in the table.

11.9.3 Combination with Supervised Features

Given the observed DevSet performance of the decorrelated, single-Gaussian, full-covariance, log-energy-only model, it is of interest to determine whether it is complementary to the much more complex supervised acoustic models. This introduces a degree of complexity into the systems as presented so far, since inference of the full covariance matrices require a preliminary label assignment pass which the supervised acoustic models do not require.

It is assumed in this suite of experiments that the initial label assignment provided by the heuristic algorithms ILA(XMIN) and ILA(XAVE) will better combine with the supervised models, than when ILA is provided by the supervised models themselves, since it relies on additional techniques which are not available to the supervised systems. In contrast, initial label assignment provided by a first Viterbi pass using the supervised models is likely to be less complimentary. These hypotheses were not tested in this thesis.

The experiments in this subsection consists of linearly interpolating the unsupervised full-covariance models with supervised models, using a global interpolation weight $\lambda \in [0.0, 1.0]$ with a 0.1 increment. This is performed for all 6 feature combinations (LE+MFCC+Δ+$\Delta\Delta$, +XMIN, +XAVE+XMIN, +NLED, +NLED+XAVE, and +NLED+XMIN), and for all four microphone models (□■, □□■, □□□■, and □□■■). The EERs for the combined systems are shown in Figure 11.17.

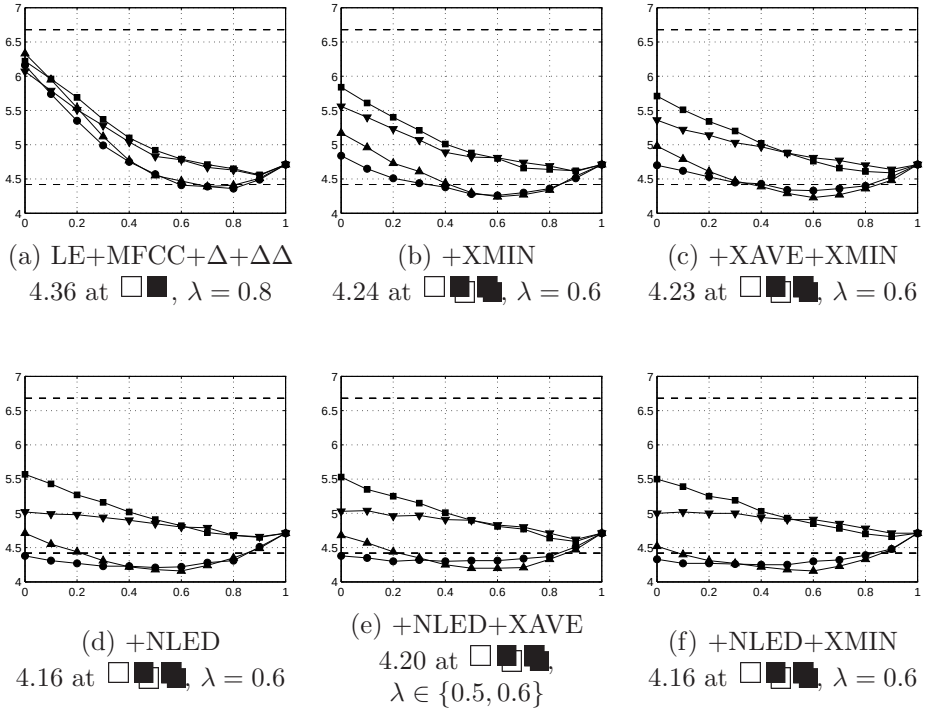


Figure 11.17: DevSet equal error rates (EERs, along the y-axis) as a function of the linear interpolation weight between supervised and unsupervised acoustic models, at the left and right edges of each panel, respectively. Results are shown for 4 different types of microphone model; EERs for □■, □□■, □□□■, □□■■ systems are depicted using —●—, —▼—, —■—, and —▲—, respectively. Supervised models trained and depicted for 6 different feature sets. Dashed horizontal lines indicate the error rates of the baseline independent-participant decoders, at a frame step of 100 ms; the upper line corresponds to the baseline without crosstalk suppression features.

As can be seen in the figure, linear interpolation always leads to better performance than for the corresponding supervised acoustic model alone. Given supervised modeling of any crosstalk suppression features (panels (b) through (f)), the linear combination outperforms the independent-participant 100-ms decoder with NLED features. The interpolation weight always favors the unsupervised acoustic model. It also tends to prefer the □□■■ microphone-state modeling

approach over the standard □■ alternative.

11.10 Minimum Duration Constraints

The adoption of joint-participant decoding, with multi-participant transition modeling and multi-participant acoustic modeling, warrants the re-optimization of minimum duration constraints in the topology. This is achieved via an exhaustive grid search, with constraints of $\{100, 200, 300, 400, 500\}$ ms for both speech and non-speech. Figure 11.18 shows EERs for the independent-participant decoder. As can be seen, when no crosstalk suppression features are used (panel (a)), the best-performing minimum constraints from among those tried are 500 ms for both speech and non-speech. The inclusion of supervised models of XMIN and XAVE features lowers the optimal duration for non-speech, but the minimum duration for speech remains at 400-500 ms. In contrast, the inclusion of supervised models of NLED features reduces the minimum duration constraint for speech, to 300 ms. This suggests that many potential false alarms, eliminated only via longer minimum duration constraints when crosstalk suppression features are not used, simply do not arise when crosstalk suppression is employed, leading to less stringent requirements for minimum duration constraints in the decoding topology.

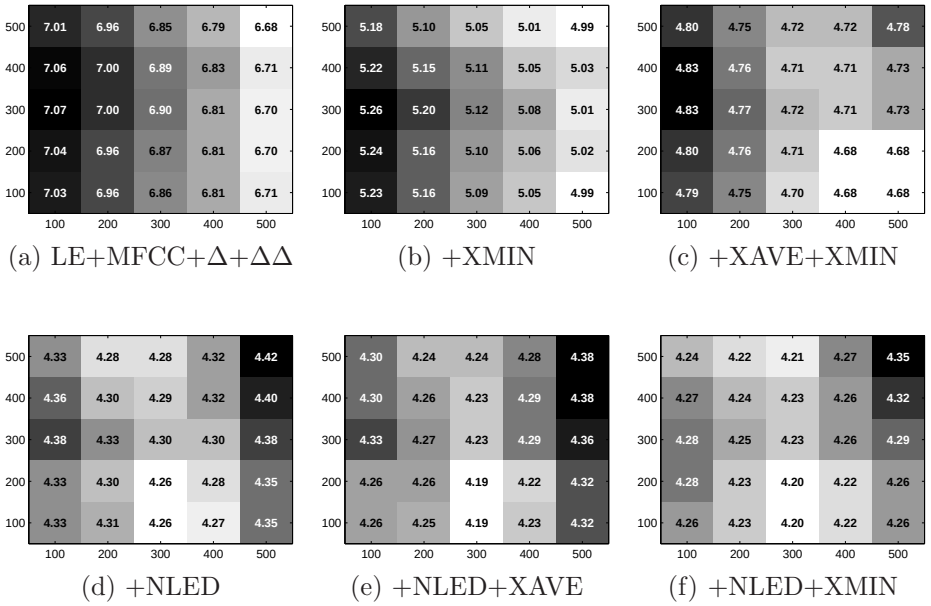


Figure 11.18: DevSet equal error rates achieved for specific minimum duration constraints for speech intervals (along the x-axis) and non-speech (along the y-axis), by independent-participant decoders. Lighter areas indicate better performing combinations.

Figure 11.19 shows similar figures, this time for joint-participant decoding. As is evident, decoding participants jointly reduces both the EERs observed and the required constraints on minimum duration. The lowest EERs, when crosstalk suppression features are not used (panel (a)), are achieved when the minimum duration constraints are 200 ms for speech and 100-200 ms for non-speech. They remain in this range despite inclusion of crosstalk suppression features (panels (b) through (f)).

The post-processing policy used throughout this chapter, in order to render the miss and false alarm rates equal, is one which merely pads hypothesized speech intervals at the beginning and at the end. As argued in Subsection 11.4.2, better performance can be expected if the policy also eliminates short intervals, of both speech and non-speech. Optimizing such a policy individually for each system on DevSet yields the EERs depicted in Figure 11.20. What can be seen is that although EERs are lower, the optimal minimum duration constraints for all feature combinations explored, do not differ markedly from those in Figure 11.19.

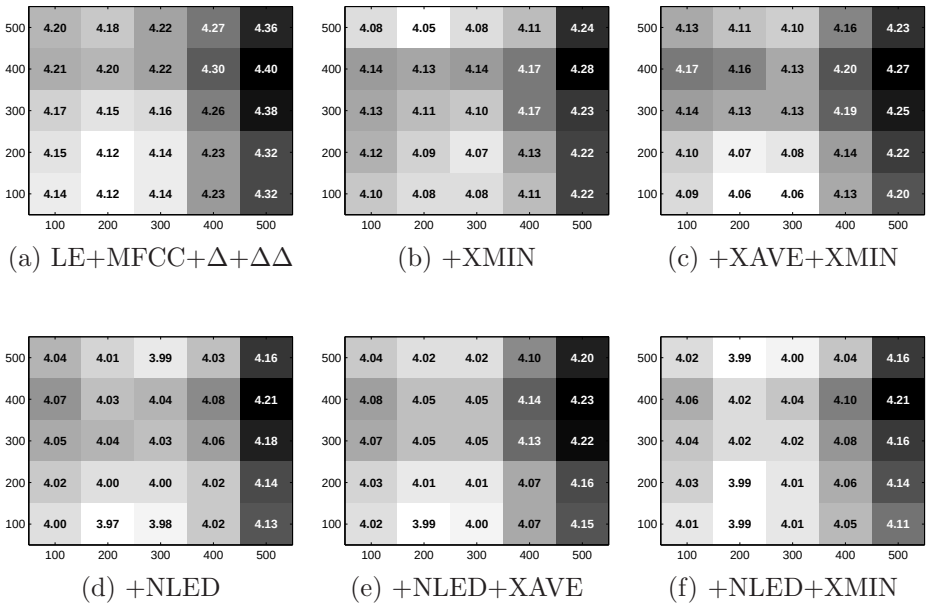


Figure 11.19: DevSet equal error rates achieved for specific minimum duration constraints for speech intervals (along the x-axis) and non-speech (along the y-axis), by joint-participant decoders. Lighter areas indicate better performing combinations.

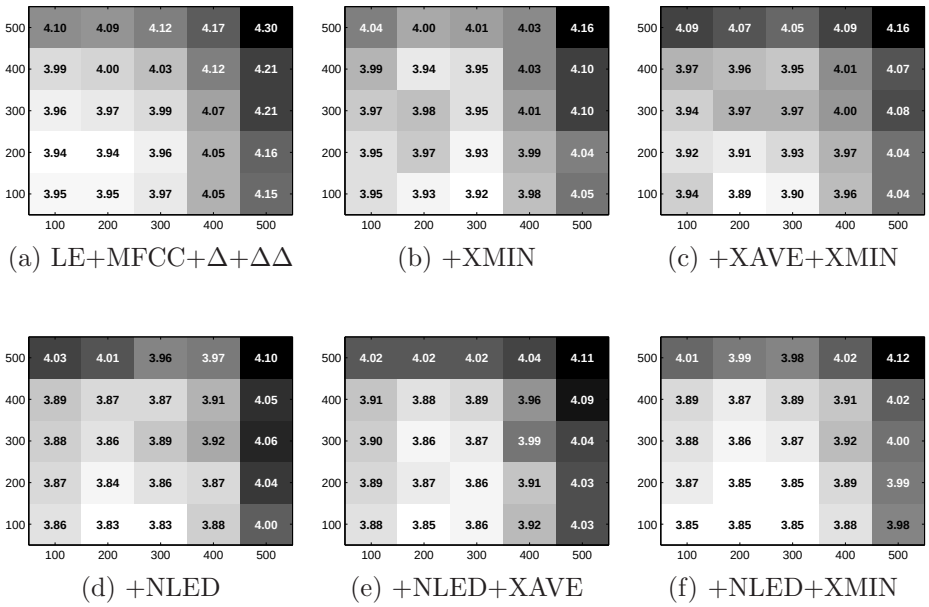


Figure 11.20: DevSet equal error rates achieved for specific minimum duration constraints for speech intervals (along the x-axis) and non-speech (along the y-axis), by joint-participant decoders. In contrast to Figure 11.19, the employed smoothing policy additionally prunes short intervals of speech and non-speech; parameters governing the smoothing are selected to minimize DevSet EERs. Lighter areas indicate better performing combinations.

11.11 Generalization to Unseen Data

This section summarizes cumulative performance improvements due to the techniques described in this chapter. Application of the said techniques has been segmented into 8 major "stages", as listed in Table 11.4; their sequence follows their descriptions in Sections 11.5 through 11.10. Performance improvements continue to be characterized as reductions in the equal error rate (EER); competing metrics are explored in the following section.

Symbol	Section(s)	Brief Description
16ms	11.5	independent-participant decoding frame step of 16 ms minimum duration constraints of 500 ms
100ms	11.6	extension of frame step to 100 ms
J_3^3	11.8.1 – 11.8.3	joint-participant decoding with topology J_3^3
$\mathsf{J}_{\mathsf{opt}}$	11.8.1 – 11.8.3	optimization of topology, $\in \{J_3^3, J_2^3, J_2^2, J_1^2, J_1^1\}$
$\mathsf{AM}_{\mathsf{opt}}^{\mathsf{S}}$	11.8.4	optimization of supervised acoustic model type, $\in \{$  $\}$
$+\mathsf{AM}^{\mathsf{U}}$	11.9	linear model-space combination with unsupervised log-energy model
$\mathsf{AM}_{\mathsf{opt}}$	11.9	re-optimization of supervised acoustic model type $\in \{$  $\}$
DUR	11.10	optimization of minimum duration constraints
$\mathsf{DUR}_{\mathsf{G}}$	11.10	re-optimization, with alternate σ

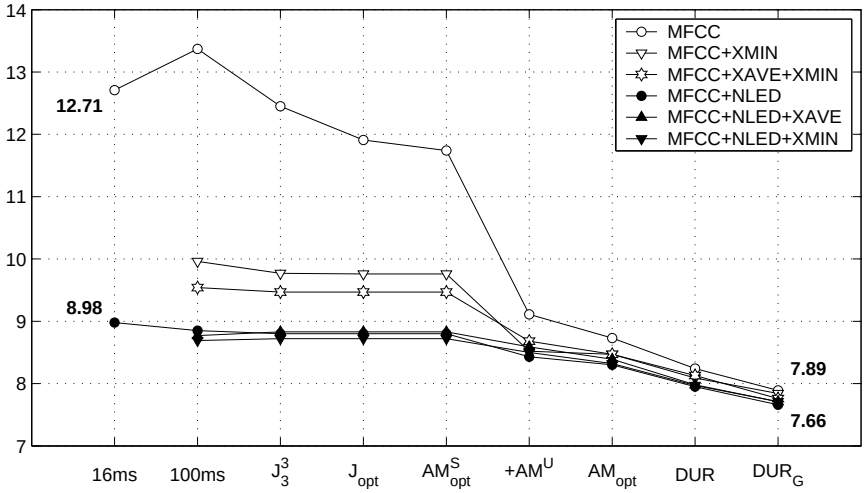
Table 11.4: Grouping of the techniques in this chapter into a sequence of 8 stages.

Consideration is given to 2 system lineages, namely those of the two baselines. They are identical at every stage of development, except with regard to inclusion of NLED features. As will be seen, supervised models of the other crosstalk suppression features, the XAVE and XMIN features, are inferior to those of NLED features, and they offer no significant additional improvement over the best NLED system. For this reason, systems employing supervised models of XAVE and XMIN features receive no further comment, but are included in the figures below for completion. Absolute and relative EER reductions are first presented for DevSet, to better appreciate the meaning of those reductions on the unseen EvalSet.

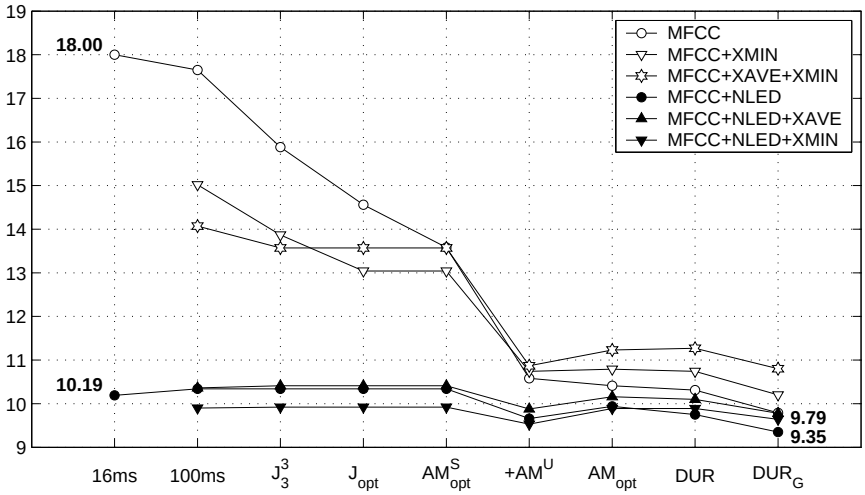
As Figure 11.23(a) shows, for DevSet, the cumulative improvement due to the modifications described in this chapter is a reduction of the equal error rate from 12.71% for the first baseline to 7.89%. This represents a difference of 4.82%abs or, identically, 38%rel.

The techniques of this chapter also yield a 1.32%abs (from 9.89% to 7.66%) EER reduction, or 15%rel, over the alternative baseline which relies on NLED features. The techniques appear to be superior to the use of NLED features by itself, by 1.09%abs (8.98% vs 7.89%) or 12%rel. However, NLED features are complimentary to some extent, offering an additional 0.23%abs, or 3%rel, improvement over a system which does not compute them but is otherwise identical in design.

Similar cumulative results are observed for EvalSet, which was unseen during development, in Figure 11.21(b). The standard baseline, without NLED features, yields an EER of 18.00%, indicating that this dataset may be more difficult than DevSet. The techniques of this chapter cumulatively lead to an EER of 9.79%, which is 8.21%abs or 46%rel lower. They also yield improved performance when crosstalk suppression features are used; they lower the EER of 10.19% to 9.35%, by



(a) DevSet



(b) EvalSet

Figure 11.21: Equal error rates (EERs, multiplied by 2 along the y-axis to facilitate comparison with ERs) at 9 stages of system development (along the x-axis and as described in Table 11.4). Trajectories shown for 6 systems differing in the types of features for which supervised acoustic models are trained; points are joined by lines for illustration purposes only. Each point represents the EER of decoder output followed by post-processing optimized for that point alone, on DevSet. EERs for both baselines, and those of the ultimate descendents of both baselines, explicitly annotated.

0.84%abs or 8%rel. As for DevSet, the contributions of this chapter appear to outperform crosstalk suppression features when the latter are used alone, by 0.40%abs or 4%rel, but NLED features are complementary and yield an additional 0.44%abs or 4%rel over the best identical-design system without NLED features.

11.12 Alternative Metrics

To more thoroughly validate the effectiveness of the techniques in this chapter, systems were tuned to optimize the other performance metrics described in Section 11.4. For simplicity, no modifications were made to acoustic feature computation; as a result, it is possible that performance on these alternative metrics is in actuality better than as described in what follows.

11.12.1 Error Rate at Decoder Output

The primary EER metric requires that misses and false alarms be balanced, and this has been systematically achieved via pre- and post-padding of the decoder output. This subsection explores the "raw" ER metric, at the output of the decoder directly (without padding of any sort). This is relevant, for example in situations where multiple decoding passes are applied. Because post-processing is not applied, the last, ninth stage in Table 11.4 has no bearing on the results of this subsection.

The most time-consuming step in re-optimizing systems to achieve a minimum ER, from among those in Table 11.4, is linear interpolation with the unsupervised acoustic model of multichannel log-energy. This requires inference of new values for the parameters λ_G, λ_R, λ_S, and θ_N; as for EER systems, this is achieved via a partially automated 4-dimensional adaptive grid search, using the most permissive J_3^3 joint-participant topology. The results of this search, performed separately for the two alternative initial label assignment algorithms ILA(XMIN) and ILA(XAVE), are the systems shown in Figure 11.22.

The 4 parameters were not re-optimized for the other topologies, for both simplicity and tractability. This may explain why the J_3^3 topology yields the best performance, and leaves open the possibility that better performance can be achieved by jointly optimizing the ILA algorithm parameters and the topology. As can be seen in the figure, the unsupervised system seeded with ILA(XAVE) is better than that seeded with ILA(XMIN); all subsequent systems in this section employ only the ILA(XAVE) variant. It achieves, by itself, an ER which is almost as good as the ER achieved by a supervised system which models not only spectral features but also state-of-the-art crosstalk suppression features.

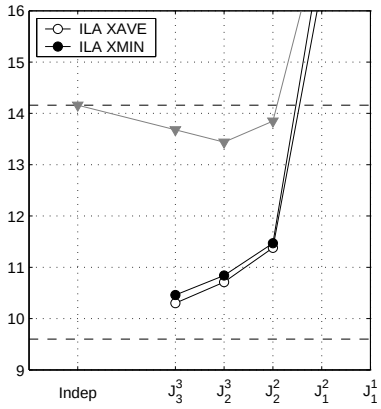


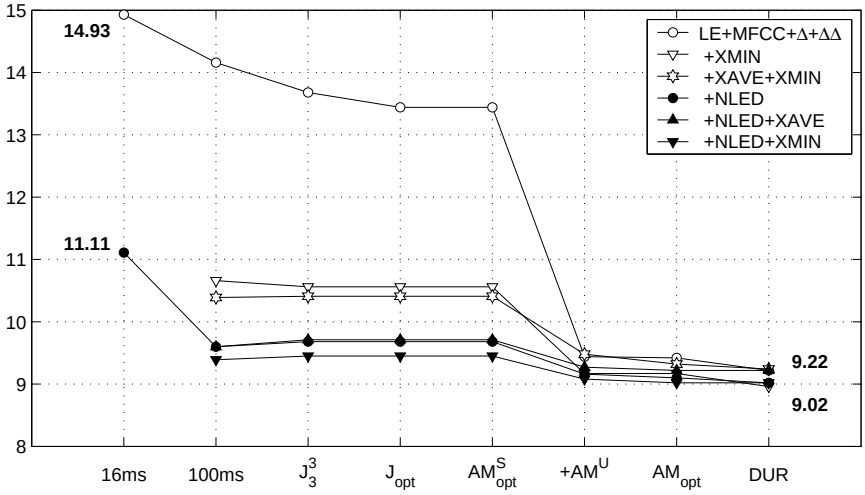
Figure 11.22: DevSet error rates (along the y-axis) for two unsupervised log-energy-only systems, as a function of transition probability model type. For comparison, the lowest-ER system with supervised LE+MFCC+Δ+ΔΔ models is depicted as ▼. Dashed lines indicate the performance of the 100-ms independent-participant systems with and without NLED features (lower and upper lines, respectively).

Interpolation of supervised acoustic models with this unsupervised model always leads to lower DevSet ERs, in ways analogous to what is shown in Figure 11.17. The best supervised acoustic models, in the combination, are the □■■ models (except for the LE+MFCC+Δ+ΔΔ+XMIN feature combination for which □■ is best). Optimizing topological minimum duration constraints, in the range $100, 200, 300, 400, 500$ milliseconds, tends to favor 500 ms for non-speech and $200 - 400$ ms for speech.

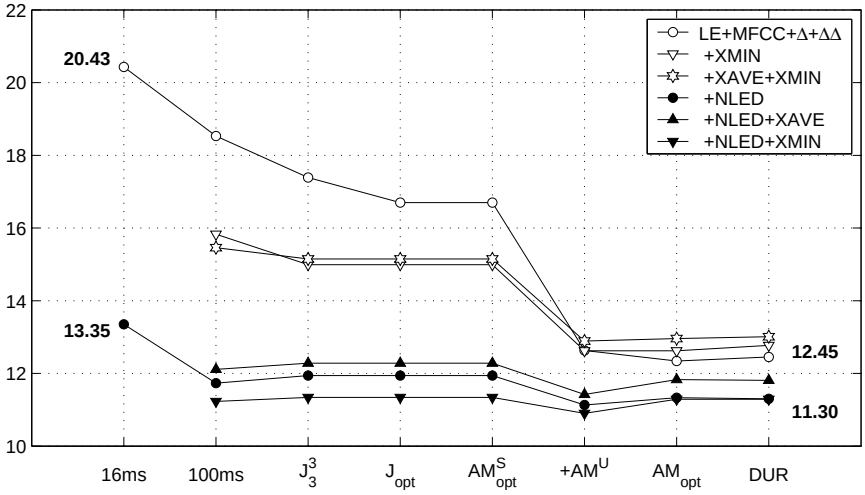
When optimized to minimize ERs, the techniques described in this chapter yield performance improvements as shown in Figure 11.23. In panel (a), it can be seen that the cumulative improvement on DevSet, for the baseline without crosstalk suppression features, consists of a 5.71%abs (38%rel) reduction. This is 1.89%abs better than the improvement due to only adding crosstalk suppression features, exhibited by the second baseline. Adding crosstalk suppression features to the system achieving an ER of 9.22% yields only a small additional improvement of 0.20%abs (2%rel). In contrast,

the techniques of this chapter, when applied to the baseline with crosstalk suppression features, yield an ER reduction of 2.09%abs (19%rel).

On EvalSet, the techniques of this chapter yield an ER reduction for the baseline without crosstalk suppression features of 7.98%abs (39%rel). The resulting system outperforms the second baseline, which includes crosstalk suppression features, by 0.90%abs (7%rel). Inclusion of these features in the system achieving an ER of 12.45% leads to a further 1.15%abs (9%rel) reduction. The techniques of this chapter improve the performance of the second baseline by 2.05%abs (15%rel).



(a) DevSet



(b) EvalSet

Figure 11.23: Error rates (ERs, along the y-axis) at 8 stages of system development (along the x-axis and as described in Table 11.4). Trajectories shown for 6 systems differing in the types of features for which supervised acoustic models are trained; points are joined by lines for illustration purposes only. Systems were optimized for smallest ER on DevSet; each point represents the classification error rate of decoder output only, without EER post-processing. ERs for both baselines, and those of the ultimate descendents of both baselines, explicitly annotated.

Closer analysis of Figure 11.23, particularly panel (b), indicates that a larger frame step, of 100 ms rather than the ubiquitously used 16 ms, leads to lower ERs. Joint-participant decoding using the J_3^3 topology is better than independent-participant decoding when no crosstalk suppression features are used; otherwise, there is no appreciable difference in performance. The single modification yielding the largest absolute reduction of ER is linear interpolation with an unsupervised acoustic model of log-energy only, using the XAVE heuristic initial label assignment algorithm. As for EERs, optimization of the factored supervised acoustic model type on DevSet favors the [□■■] variant, by a very small amount, but hurts performance on EvalSet by a larger absolute amount. Without this step, final cumulative performance on EvalSet is likely to be even better than observed, especially when crosstalk suppression features are used.

11.12.2 Classification Error at Decoder Output

The second explored alternative performance metric, described in Section 11.4, is the classification error (CERR). As for ER, the metric is computed at the output of the decoder, rather than following the application of a smoothing policy. Such policies can only increase the amount of speech, and the CERR metric already penalizes false alarms more than the ER and EER metrics described so far.

Inference of optimal parameters governing unsupervised model performance is achieved as for the minimization of other performance metrics, and is shown in Figure 11.24. In contrast to Section 11.12.1, joint-participant decoding leads to supervised model performance, without crosstalk suppression features, which is almost as good as the performance of the supervised independent-participant model with crosstalk suppression features (indicated by the lower dashed line in the figure). This is because joint-participant decoding as implemented in this thesis limits the number of participants which can be simultaneously hypothesized as speaking. Since CER penalizes insertions much more than ER does, the best performance by the supervised system without crosstalk suppression features is achieved by the J_1^2 topology, which allows at most one person to be hypothesized as speaking. Against this context, both of the unsupervised log-energy-only systems achieve lower CERs, using the J_3^3 topology, than any supervised system without crosstalk suppression features. That performance is very close to the performance of the independent-participant system which uses crosstalk suppression features in addition to spectral features.

Linear interpolation of the ILA(XMIN) unsupervised acoustic model of log-energy with the supervised acoustic models in each system yields improvements as in earlier sections; optimal model interpolation weights always favor the unsupervised model. Except for the system relying only on LE+MFCC+Δ+ΔΔ features (for which the best supervised acoustic model structure following combination is [□□■]), all interpolated systems favor the [□■] supervised model variant. Optimal minimum duration constraints for speech are 200–300 ms, while those for non-speech are 500 ms.

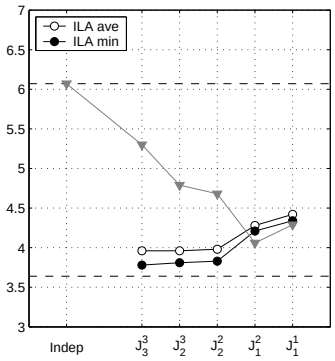
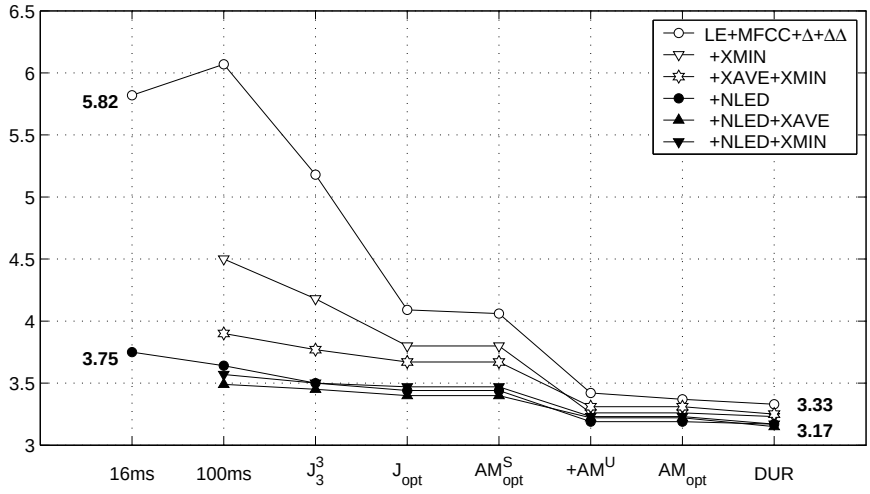


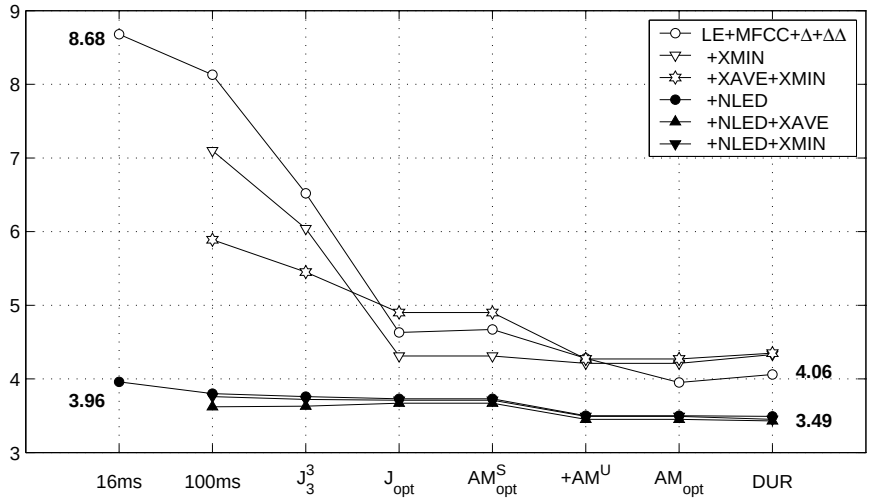
Figure 11.24: DevSet classification error rates (CERs; along the y-axis) for two unsupervised log-energy-only systems, as a function of transition probability model type. For comparison, the lowest-CER system with supervised LE+MFCC+Δ+ΔΔ models is depicted as ▼. Dashed lines indicate the performance of the 100-ms independent-participant systems with and without NLED features (lower and upper lines, respectively).

Panel (a) in Figure 11.25 shows the cumulative impact on the DevSet classification error rate of the modifications argued for in this chapter. The latter, when applied to the baseline without crosstalk suppression features, yields a CER reduction of 2.49%abs (43%rel). The resulting system outperforms the baseline with crosstalk suppression features, by 0.42%abs (11%rel). Inclusion of NLED features improves performance further, if only slightly, by 0.16%abs (5%rel). The techniques have the cumulative effect of reducing the CER of the baseline with crosstalk suppression features by 0.58%abs (15%rel).

Performance is similar on EvalSet and is shown in panel (b) of the same figure. CER reduction for the baseline without crosstalk suppression features is 4.62%abs (53%rel), a larger effect than for DevSet. In constrast to DevSet, supervised models of crosstalk suppression features perform better than the improved system without them, by a small 0.10%abs (2%rel). The combination of the techniques of this chapter and NLED crosstalk suppression features yields an additional CER reduction of 0.47-0.57%abs (12-14%rel).



(a) DevSet



(b) EvalSet

Figure 11.25: Classification error rates (CERs, along the y-axis) at 8 stages of system development (along the x-axis and as described in Table 11.4). Trajectories shown for 6 systems differing in the types of features for which supervised acoustic models are trained; points are joined by lines for illustration purposes only. Systems were optimized for smallest CER, rather than ER, on DevSet; each point represents the classification error rate of decoder output only, without EER post-processing. CERs for both baselines, and those of the ultimate descendents of both baselines, explicitly annotated.

11.12.3 F-Score at Decoder Output

A third performance metric explored in this section is the F-score, or the unweighted harmonic mean of recall and precision. Optimization of unsupervised model parameters shows trends which are similar to those observed during system optimization for minimum CER. As Figure 11.26 reveals, the techniques described in this chapter have a significant effect on systems without crosstalk suppression features, because joint-participant decoding allows for the successful exclusion of much crosstalk by prohibiting multiple participants from being hypothesized as speaking simultaneously. This makes the system depicted with ▼ quite competitive with that depicted with the dashed line, at the top of the figure, representing an independent-participant decoder with models for crosstalk suppression features. Unsupervised log-energy modeling, however, is able to outperform even the best ▼ system in the figure, in spite of completely ignoring spectral MFCC features or their first- and second-order differences. The unsupervised system seeded using ILA(XMIN), the better of the two unsupervised systems in the figure, is retained for the remainder of the experiments in this section.

As observed for CER, interpolation of supervised and unsupervised acoustic models leads to improvements, for all 6 investigated feature combinations, and the supervised □■ acoustic model variant is always preferred. Optimum minimum duration constraints, for DevSet, continue to be at or around 300 ms for speech and exactly 500 ms for non-speech.

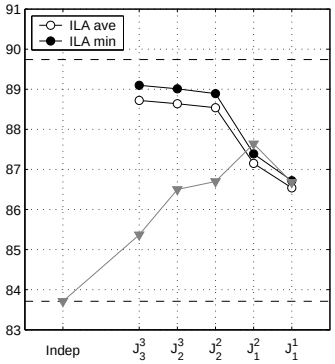


Figure 11.26: DevSet F-scores (along the y-axis) for two unsupervised log-energy-only systems, as a function of transition probability model type. For comparison, the highest-F-score system with supervised LE+MFCC+Δ+ΔΔ models is depicted as ▼. Dashed lines indicate the performance of the 100-ms independent-participant systems with and without NLED features (upper and lower, respectively).

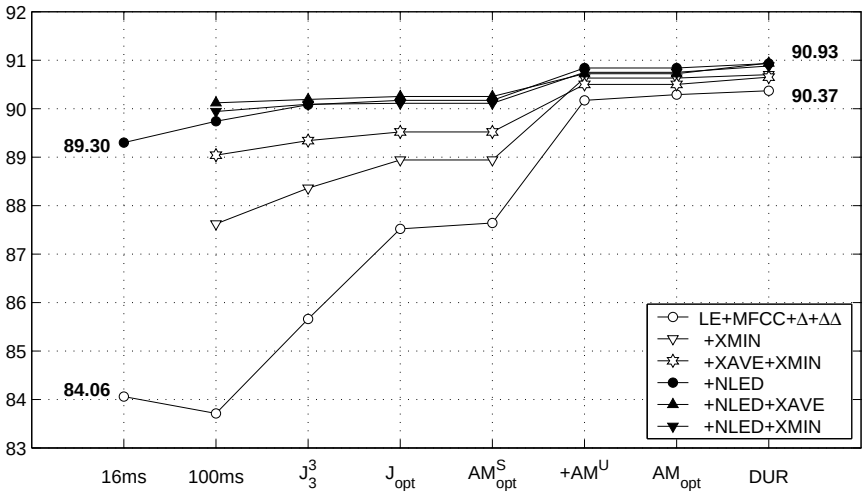
The cumulative gains due to the techniques listed in Table 11.4, on DevSet, are shown in panel (a) of Figure 11.27. When starting from the baseline without crosstalk suppression features, they yield an F-score improvement of 6.31%abs. The resulting system is better than the baseline with crosstalk suppression features, by 1.07%abs, but crosstalk suppression features help by an additional 0.56%abs. The techniques, when applied to the baseline which already models crosstalk suppression features, yield and improvement of 1.63%abs.

The improvements on EvalSet are larger in magnitude, because the baseline (without crosstalk suppression features) performs significantly worse on EvalSet than on DevSet. Correspondingly, the techniques of this chapter yield a 10.53%abs increase in F-score. However, their impact is smaller than that of crosstalk suppression features, and the second baseline achieves a performance which is 0.31%abs higher. The two approaches are somewhat complementary, as observed for other metrics: modeling crosstalk suppression features yields an additional F-score increase of 1.66%abs, while applying the techniques of this chapter to the baseline which already models crosstalk suppression features results in a 1.35%abs improvement.

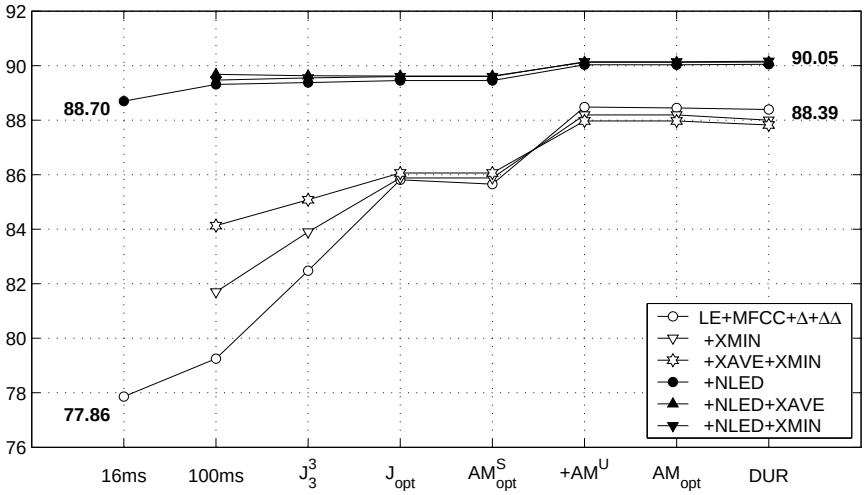
11.12.4 NIST Error Rate at Decoder Output

The final performance metric explored in this chapter is the NIST error. Optimization of the unsupervised systems towards minimal values (when the J_3^3 topology is used) results in the performance shown in Figure 11.28. The curves are somewhat different from the same curves for other metrics, in that the unsupervised systems (for topologies J_3^3, J_2^3, and J_2^2) outperform an independent-participant decoder. The systems using ILA(XMIN) appear to do slighly better that those based on ILA(XAVE).

Linear interpolation with supervised models favors the [□■] model variants for all 6 feature types explored, except for the system without any crosstalk suppression features, where [□▟■] is preferred. Optimal interpolation weights strongly



(a) DevSet



(b) EvalSet

Figure 11.27: F-scores (along the y-axis) at 8 stages of system development (along the x-axis and as described in Table 11.4). Trajectories shown for 6 systems differing in the types of features for which supervised acoustic models are trained; points are joined by lines for illustration purposes only. Systems were optimized for largest F-score, rather than smallest ER, on DevSet; each point represents the F-score of decoder output only, without EER post-processing. F-scores for both baselines, and those of the ultimate descendents of both baselines, explicitly annotated.

favor the unsupervised model ($\lambda \geq 0.7$). For all feature set combinations, the preferred minimum duration constraints are 0.500 ms for both speech and non-speech. This contrasts with the parameter values found for other metrics, where slightly shorter constraints around 300 ms were preferred for speech intervals; it also suggests that better performance can be achieved for even longer mimimum durations. This has not been attempted in this thesis, due to the exponential growth of the topology as the minimum duration constraints grow.

The cumulative reduction in NIST error given the techniques proposed in this chapter are shown in Figure 11.29. For DevSet, in panel (a), it can be seen that they lead to a 33.14%abs (77%rel) reduction, and outperform merely including

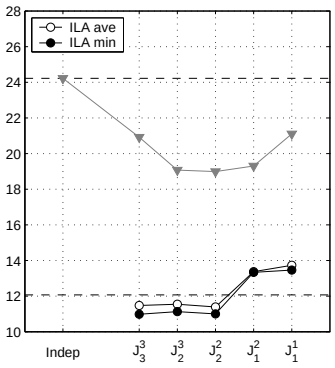


Figure 11.28: DevSet NIST error rates (along the y-axis) for two unsupervised log-energy-only systems, as a function of transition probability model type. For comparison, the lowest-NIST-error-rate system with supervised LE+MFCC+Δ+$\Delta\Delta$ models is depicted as ▼. Dashed lines indicate the performance of the 100-ms independent-participant systems with and without NLED features (lower and upper lines, respectively).

crosstalk suppression features in a participant-independent decoder by 4.84%abs (33%rel). Inclusion of these features leads to a 0.92%abs (9%rel) additional gain. For the baseline with crosstalk suppression features already implemented, the techniques yield a NIST error reduction of 5.76%abs (39%rel).

On EvalSet, the impact is similar. The techniques of this chapter yield a 35.67%abs (74%rel) reduction of NIST error for the baseline which does not employ crosstalk suppression features. This is an improvement over using crosstalk suppression features instead, of 1.50%abs (11%rel). Crosstalk suppression features, however, yield an additional 2.25%abs (18%rel) reduction of NIST error. The techniques lead to a reduction of 3.75%abs (26%rel) when applied to the baseline modeling crosstalk suppression features.

The above error rates, as well as all those in this chapter section up to this point, were computed against the forced-alignment-mediated references Υ^T whose construction was described in Section 11.4.3. They are recomputed below using Υ^U, for the purposes of comparability with existing work by other researchers, or those whose talkspurt reference construction may deviate from that adopted in this work. Figure 11.30 is identical to Figure 11.29 in all other respects. It is important to note that changing the references used for scoring effectively results in a different metric, and the systems shown in Figure 11.29 were not re-optimized to minimize that metric; they are merely rescored using the alternative reference segmentation. As such, there is the potential that better performance is possible.

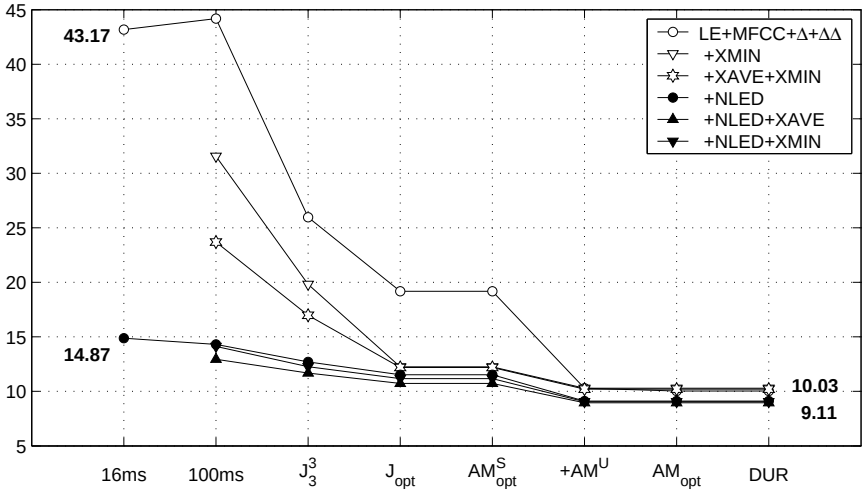
As Figure 11.30 shows, the techniques of this chapter result in a 12.64%abs (51%rel) reduction of NIST error, when scored against the original segmentations provided by NIST during the RT06s evaluation. The improvement is higher than that observed by including crosstalk suppression features in the baseline, by 1.26%abs (9%rel). Crosstalk suppression features improve on this by 1.88%abs (15%rel). The techniques of this chapter reduce the NIST error of the baseline which already includes crosstalk suppression features by 3.14%abs (23%rel).

On the EvalSet, the improvements are similar. The techniques of this chapter cumulatively lead to an 18.43%abs (41%rel) reduction of the NIST error. In this case, however, they are not as good as merely including crosstalk suppression features, which are better by 0.91%abs. Adding crosstalk suppression features to the system achieving 27.01% yield an additional 2.28%abs (8%rel) reduction of error. The baseline which includes crosstalk suppression feature modeling benefits from the techniques of this chapter by a 1.37%abs (5%rel) NIST error reduction. Errors are bigger than when scoring is against Υ^T, due to misses of within-utterance non-speech.

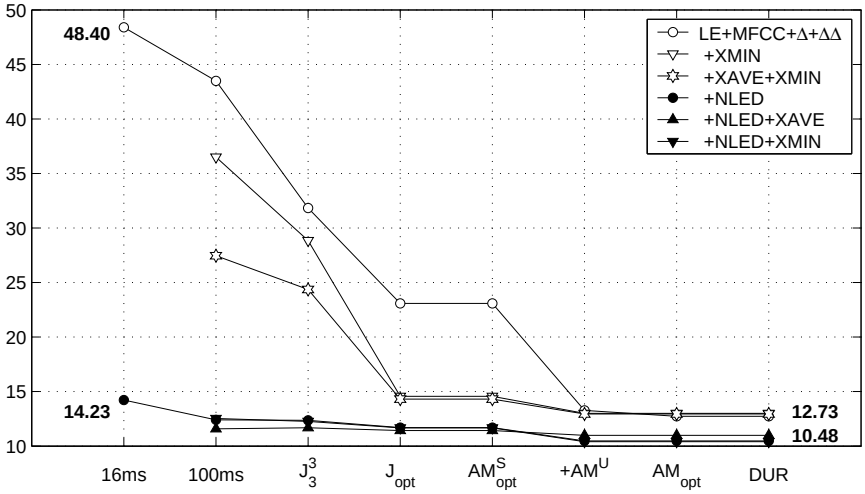
As noted above, there is some scope for improved performance against the original NIST references, by optimizing system design decisions to minimize the corresponding NIST error. As can be seen in panel (a) of Figure 11.30, several decisions led to poorer performance. This is true of the selection of optimal topology (step $\mathsf{J_{opt}}$); while all explored feature combinations except that without any crosstalk suppression features exhibited higher errors, this is particularly true for the systems relying on XAVE and/or XMIN features. Similarly, the step labeled $\mathsf{AM_{opt}}$ resulted in poorer DevSet performance for the system not relying on feature-space crosstalk suppression.

11.13 Potential Impact

This chapter has described several technological contributions to the task of automatic speech activity detection, and it may benefit the research community in two additional ways. The technological contributions consist of two findings



(a) DevSet



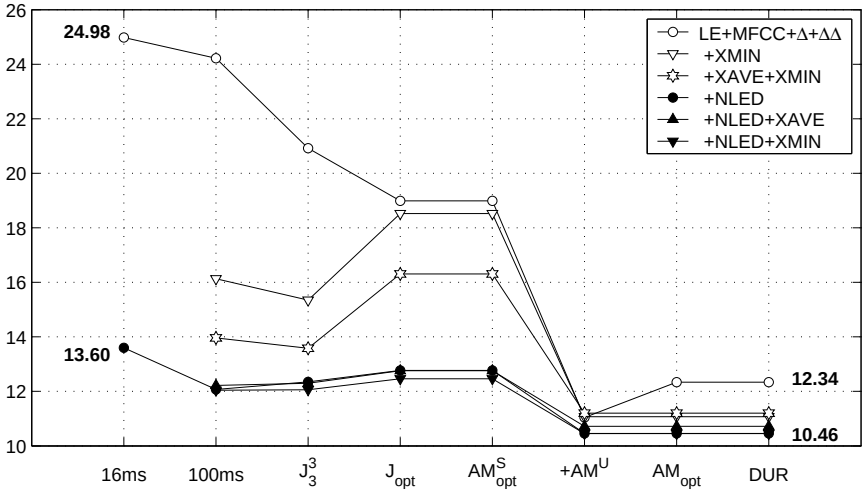
(b) EvalSet

Figure 11.29: NIST error rates (along the y-axis) at 8 stages of system development (along the x-axis and as described in Table 11.4). Trajectories shown for 6 systems differing in the types of features for which supervised acoustic models are trained; points are joined by lines for illustration purposes only. Systems were optimized for smallest NIST error rate, rather than ER, on DevSet; each point represents the NIST error rate of decoder output only, without EER post-processing. NIST error rates for both baselines, and those of the ultimate descendents of both baselines, explicitly annotated.

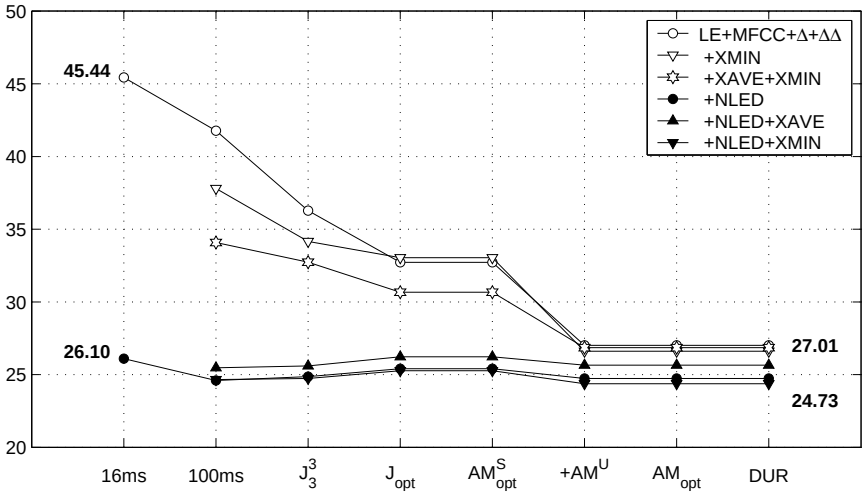
relevant to both independent-participant and joint-participant decoding, and five findings relevant to joint-participant decoding.

11.13.1 General Clost-Talk SAD Decoding

First, the experiments presented provide a strong indication that small decoding frame steps (and as a result small frame sizes) that are typically favored are suboptimal for speech activity detection. That this should be the case is not surprising, since, on average, phonemes are much shorter than contiguous intervals of talk. However there appears to be



(a) DevSet



(b) EvalSet

Figure 11.30: NIST error rates (along the y-axis) against reference utterance segmentation Υ^U at 8 stages of system development (along the x-axis and as described in Table 11.4). Trajectories shown for 6 systems differing in the types of features for which supervised acoustic models are trained; points are joined by lines for illustration purposes only. Systems were optimized for smallest NIST error rate, rather than ER, on DevSet; each point represents the NIST error rate of decoder output only, without EER post-processing. NIST error rates for both baselines, and those of the ultimate descendents of both baselines, explicitly annotated.

some resistance in the speech processing community regarding the utility of MFCC features computed for much larger frame sizes. The results in this thesis are not unequivocal; Figures 11.4 and 11.6 indicate that in some cases, the move towards a larger frame step and size is deleterious. However, as shown in Figure 11.4, ROC curves suggest that a frame step of 50 ms may be optimal.

Larger frame sizes, even when not offering improved performance, offer additional benefits. These consist of increased decoding speed for applications requiring real-time performance, the possibility of decoding participants jointly, as well as

the potential to explore more complicated topologies. An example of the latter is a topology with different sub-topologies for significantly different speech types, some of which may require relatively long minimum duration constraints.

Second, this section has provided independent verification of the very good performance of NLED features. Although the techniques of this chapter improve upon NLED performance regardless of the metric used, in several cases NLED features alone provide nearly as good performance. This may be important when the effort to implement joint-participant decoding is deemed too high.

11.13.2 Joint-Participant SAD Decoding

The single most important finding in this chapter is the degree to which simple, unsupervised multi-participant acoustic models outperform significantly more complex, supervised acoustic models. The unsupervised models are simple in that they are single-Gaussian, and model only the log-energy in each channel. In contrast, the supervised acoustic models to which they are compared in this work are 256-component GMMs, modeling not only log-energy, but also the 12 lowest-order MFCCs, as well as first- and second-order derivatives. As Figures 11.15, 11.22, 11.24, 11.26, and 11.28 show, the performance of systems relying only on unsupervised models is much better than the performance of independent-participant decoders relying on no crosstalk suppression features. It is also better than that of joint-participant decoders relying on no crosstalk suppression features. Decoders with crosstalk suppression features typically outperform the unsupervised model systems, but by a far smaller absolute amount; for the NIST error, however, unsupervised model systems outperform even those independent- and joint-participant systems which do model crosstalk suppression features.

To be fair, it should be noted that the unsupervised models are deployed in a second decoding pass, following a first pass which performs an initial label assignment to the audio. Although NLED features also benefit from a preliminary pass in which the noise floor estimates are acquired for each channel, it is easier to foresee at the current time an implementation of NLED computation in a single-pass decoder, via a running minimum estimate, than it is the implementation of the unsupervised architecture under similar constraints.

Second, the success of the unsupervised model, which relies on two-pass decoding as noted above, indicates that acoustic model adaptation for all features, not just for log-energy, is likely to have a significant impact on SAD performance in future systems. At the current time, close-talk multichannel SAD systems do not explicitly re-estimate model parameters or transform feature vectors.

Third, the described experiments have shown that the initial label assignment performed in the first pass can be achieved using a heuristic based on only channel energies and cross-channel correlation maxima. This is important, as it posits the decoder based only on the resulting unsupervised acoustic models as a privacy-sensitive system. At the current time, there is significant interest in performing speech activity detection in text-independent environments where MFCC features may not be computed for privacy reasons. As shown in Figures 11.15, 11.22, 11.24, 11.26, and 11.28, the unsupervised acoustic model, joint-participant decoder described in this work already outperforms MFCC systems, by a significant amount.

Fourth, joint-participant decoding has been shown to be not only tractable, but, given larger frame sizes than are used in speech recognition, to yield significantly improved performance over independent-particicpant decoders, particularly in the absence of crosstalk suppression features. Its tractability for conversations with arbitrary numbers of participants is facilitated by a pecularity of conversation: participants tend to talk one at a time, and the occurrence and duration of overlap is inversely proportional to its degree. Metrics which are lenient to false alarms (e.g. EERs and ERs) are not unequivocal in this regard; however, other metrics, which are less lenient to false alarms (e.g. CERs and NIST errors), exhibit improvement when joint-participant decoding is employed. The optimal topology for minimizing NIST errors, for example, when scored against talkspurt rather than utterance segmentations, was found to be J_1^2. It allows at most one participant to be speaking at any given instant. This suggests that limiting the potential for hypothesized overlap is more important at current performance levels, given this metric, than is allowing the short intervals of multi-participant overlap which may occur frequently.

It should be noted that the use of joint-participant decoding is orthogonal to acoustic model design. It is therefore independent of the features computed for each speaker or for each channel. As such, there is the potential for impact in a wide variety of situations in which conversations are instrumented in ways other than by outfitting each participant with a close-talk microphone. For example, at the current time, speaker diarization systems using farfield microphones rely on a clustering stage which may account for temporal adjacency, but which does not model duration (or duration of overlap).

The transition modeling framework proposed in this thesis may provide a useful underpinning to future systems wishing to streamline processing.

11.13.3 Ancillary Aspects

This chapter has the potential to impact future work in SAD in two additional ways.

First, although the described techniques have been published in [144, 147, 148, 149, 141], the incremental history of their development did not start with an independent-participant HMM decoder. As a result, the benefit of the modeling approaches has been difficult to assess. The presentation in this chapter corrects this problem; relationships between existing research and the sections of this chapter are shown in Figure 11.31.

Second, this thesis has compared the performance of several systems, in a rigorous manner, on a large number and variety of metrics. This is relatively important. The majority of existing work on SAD has focused on a single metric, and often demonstrated significant improvement. However, it is never clear in such cases whether the techniques described in such work are not merely well-suited to a particular metric type, and researchers using other, equally valid metrics cannot predict the magnitude of improvements. In re-optimizing parameters to minimize different error types, on a widely available dataset, this thesis has provided a correlated set of state-of-the-art baselines for further development of SAD technologies.

11.14 Relevance to Other Chapters

The experiments presented in this chapter have exercised and validated the multi-participant state-space modeling approaches described in Chapter 6 and the multi-channel feature-space modeling approaches of Chapter 9. In particular, multi-participant state-space models were employed to provide *a priori* knowledge in multi-participant speech activity detection in the same way that language models provide *a priori* knowledge in speech recognition. The chapter complements Chapter 10, where state-space models were used to score the turn-taking perplexity of multi-party conversations.

This chapter is directly related to Chapter 13, in which speech/non-speech topologies are extended to account for laughter. Chapters 14, 15, 16, and 17 are related indirectly, in that they operate at a level where speech activity is assumed to already be detected, using the techniques proposed here or elsewhere.

11.15 Summary

The joint modeling of the states of all participants to a conversation appears to be a useful technique for reducing the high false alarm rate observed for independent-participant decoding when participants are instrumented with personal microphones. To be tractable and useful, joint modeling calls for a much larger frame step than used traditionally; experiments have shown that larger frame steps and sizes may be beneficial anyway, even when participants are not jointly decoded.

Joint multi-participant modeling brings gains which are approximately as large as those observed by computing multi-channel, crosstalk suppression features; this appears to be true for a large number of metrics, including agglomerates of false alarm and miss rates, classification error rate, precision and recall, and metrics designed specifically for multi-party SAD. In all cases, on unseen data, multi-participant modeling and multi-channel crosstalk suppression are somewhat complementary, offering improved performance when used together. As a result, modeling participants in the way proposed can be expected to contribute to future multi-party speech activity detection efforts.

11.16 Future Directions

The experiments described suggest many avenues for future enquiry. They likely should be repeated using a frame step of 50 ms, rather than the 100 ms appearing throughout this thesis.

The results indicate an interesting finding, namely that the non-target cross-channel correlation features (e.g. XAVE and XMIN) do not yield large improvements over NLED features. This contrasts with what has been reported in [19]. Possible reasons for this are that, at large frame steps and sizes, NLED features are much stronger than at the small

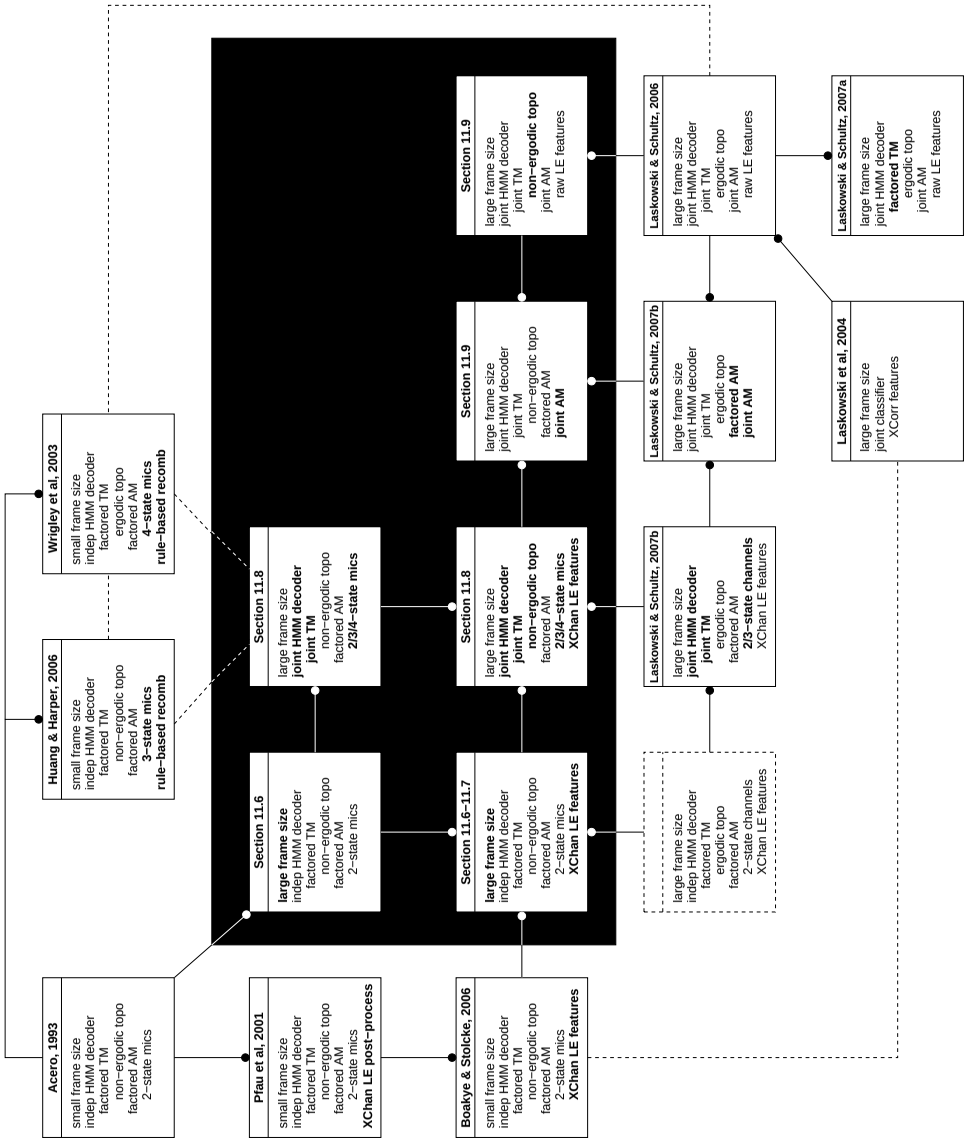


Figure 11.31: Conceptual placement of experiments of this chapter (against a black background) in the context of other published work. "TM": transition model; "AM": acoustic model; "LE": log energy; "XChan": cross-channel ("XChan LE" refers to NLED features); "XCorr": cross-correlation (refers to NT-Norm features). Solid-line object connections indicate potential offspring systems with a circle; dashed connections indicate shared components but dissimilar structure.

frame size of [19]. Also, the features were computed following more aggressive pre-emphasis ($1 - z^{-1}$) than elsewhere ($1 - 0.97z^{-1}$). This difference with [19] should be explored, since numerically the error rates of the final systems of this chapter are still quite far from zero.

Another approach which may be beneficial for speech activity detection is to condition emission probability models on the topological state, yielding (as an example) different acoustic models for speech at the beginning of a talkspurt than at the end.

Finally, as mentioned earlier, model adaptation and multi-pass speech activity decoding appear to be a useful next step to evaluate, both for systems using crosstalk suppression features and for those not using them. The unsupervised

two-pass modeling of energy described in this chapter suggests that, when only energy is modeled, large gains are observed which are likely due to what appear to be quite significant differences in training and testing acoustic conditions.

Chapter 12

Quantitative Analysis of Turn-Sharing, or the Co-Excitatory Occurrence of Laughter*

12.1 Introduction

Turn-taking is a generally acknowledged phenomenon which qualitatively accounts for the distribution of speech activity in time and across participants to a conversation; it was treated quantitatively in this thesis in Chapter 10. Whether other types of vocal production, not speech, can also be thus computationally described is not typically explored. This chapter treats one specific type of nonverbal vocal activity, laughter, and presents some preliminaries towards a general model of the distribution of laughter in conversation.

As this chapter will show, laughter in meetings, at least given the corpus used for the ensuing study, is suprisingly frequent by time. It is by far the most often transcribed nonverbal behavior. In addition to being interesting on its own merits, this aspect of its occurrence promises that it may be more tractable to model acoustically than other nonverbal behaviors. Most importantly to this thesis, it is well known that while participants tend to await their turn when wanting to speak, they do not do so when wanting to laugh. Indeed, laughter is said to be contagious [185], with participants frequently joining in simultaneous group laughter with others. "Turns" consisting of laughter are thereby *shared*. It is therefore of some importance whether the modeling approaches described in Chapter 6, which render multi-participant search tractable and felicitous by imposing constraints on the degree of simultaneous vocalization, have a role to play in the description of the occurrence of laughter.

That laughter is ubiquitously shared, as contagion, finds much counter-evidence in the literature. A significant body of research in conversational analysis [114, 109, 115, 110, 112, 111, 96, 70, 71, 72, 73, 74, 75, 77, 82, 76, 113] suggests that group laughter is in fact initiated by a "first laugh", with interlocutor laughter forming the preferred response. Interlocutors may decline to respond, with consequences relevant to first laugher. Furthermore, there exists research [193, 177] which reports that laughter may be directed strategically towards a specific subset of interlocutors, thus splitting the conversational group into participants with whom we might wish to bond, those we are laughing at, and those whose allegiance is not of relevance in the moment.

This chapter explores the potential for density modeling of the occurrence of laughter. Specifically, it contrasts the occurrence of speech with that of laughter, in part using the modeling approaches of Chapter 6. The latter are shown to provide a useful tool to studying interaction, not only when it is conceived of as consisting exclusively of speech.

The research described here first appeared in [135, 137, 136]; the occurrence of laughter in meetings was compared to that in interactive seminars in [29].

*The work in this chapter was conducted in collaboration with Susanne Burger.

12.2 Related Work

To the best of the author's knowledge, there is no computational work involving density modeling of the occurrence of laughter. The nearest such work is that involving density modeling of speech. An overview of work with bearing on that theme is provided in Chapter 5.

12.3 Dataset Use

The analysis and experiments of this chapter use the ICSI Meeting Corpus [108, 202], which is described in Chapter 4. It was chosen here because of its number of meetings (75), its number of participants (3-9) per meeting, and the fact that the meetings would have occurred even if they were not recorded. Importantly, the orthographic annotation which accompanies the recordings includes laughter in significant quantity, and it appears safe to assume that while the annotators' main task was to get the words right, they also took care to include nonverbal vocalizations. That effort, and the availability of forced-alignment timestamps for all verbal vocalization, makes it possible to automatically infer the boundaries of a significant proportion of the annotated laughter instances. All analyses are performed using the entirety of the corpus.

It should be noted that the ICSI Meeting Corpus is not homogenous in meeting type, and that there is significant variability across conversation groups, meeting towards diverse ends, in the amount and structure of multiparticipant laughter patterns. In this chapter, laughter occurrence has been quantified without taking this variability into account; the numbers presented are averages for the entire corpus.

12.4 Constructing a Laughter Segmentation

The laughter segmentation produced in this chapter treats laughter as occurring in *bouts* [6], within which vocal activity is nominally distributed in a sequence of one or more *calls*. In manually segmenting laughter, internal bout structure is ignored; the purpose is to describe each instance (equated with a bout) with only its start time, its end time, and a binary value indicating whether the laugher employs voicing *anywhere* during the bout. The starting point for producing this segmentation for the ICSI Meeting Corpus is the original orthographic transcription.

12.4.1 Identifying Laughter

Laughter had been transcribed in the ICSI corpus, by its original transcribers, in two ways. First, discrete events were annotated as `VocalSound` instances, and appear interspersed among lexical items. Their location among such items is assumed to be indicative of their temporal extent. Table 12.1 shows a subset of `VocalSound` types, namely those that are most frequent. As can be seen, the `VocalSound` type `laugh` is the most frequently annotated nonverbal vocal production.

Freq Rank	Token Count	`VocalSound` Description	Used Here
1	11515	`laugh`	√
2	7091	`breath`	
3	4589	`inbreath`	
4	2223	`mouth`	
5	970	`breath-laugh`	√
11	97	`laugh-breath`	√
46	6	`cough-laugh`	√
63	3	`laugh, "hmmph"`	√
69	3	`breath while smiling`	
75	2	`very long laugh`	√

Table 12.1: Top 5 most frequently occurring `VocalSound` types in the ICSI Meeting Corpus, and the next 5 most frequently occurring types relevant to laughter.

The second type of laughter-relevant annotation found in the corpus transcriptions, `Comment`, describes events of extended duration which were not localized between specific lexical items. In particular, this annotation appears to cover

the phenomenon of "speech-laughs" [175]. The top five most frequently occurring `Comment` descriptions pertaining to laughter are listed in Table 12.2. The `Description` attributes of both the `VocalSound` and `Comment` tags, as produced by the ICSI transcribers, appear to be largely ad hoc, and seem to reflect practical considerations during an annotation pass whose primary aim was to transcribe words.

Freq Rank	Token Count	`Comment` Description
2	980	`while laughing`
16	59	`while smiling`
44	13	`last two words while laughing`
125	4	`last word while laughing`
145	3	`vocal gesture, a mock laugh`

Table 12.2: Top 5 most frequently occurring `Comment` descriptions containing the substring "`laugh`" or "`smil`". All utterances whose transcription contained these descriptions were listened to, but portions were included in our final laugh bout segmentation only if the utterances contained laughter (in particular, intervals annotated with "`while smiling`" were not automatically included.)

In this thesis, the `Description`s were used only to identify and possibly to help segment bouts of laughter, but their content was ignored. 12635 transcribed `VocalSound` laughter instances were identified, of which 65 were attributed to farfield channels (rarely occurring side participants uninstrumented with close-talk microphones). The latter were excluded from the ensuing analysis, because the ICSI MRDA Corpus includes forced alignment timestamps for nearfield channels only. Also identified were 1108 transcribed `Comment` laughter instances, for a total of 13678 transcribed laughter instances in the original ICSI transcriptions.

12.4.2 Automatic Endpoint Determination

Of the 12570 indentified non-farfield `VocalSound` instances, 11845 are adjacent on both the left and the right to either a time-stamped utterance boundary, or a lexical item. It was thus possible to automatically deduce start and end times for 87% of the identified laughter instances. Each automatically segmented instance was then manually inspected, as described in [137]. In a small handful of cases (<3%), when there appeared to be ample evidence that no laughter in fact occured, the instance was discarded.

12.4.3 Manual Endpoint Determination

The remaining 725 identified non-farfield `VocalSound` instances were not adjacent to an available timestamp on either or both of the left and the right. These instances were segmented manually, by listening to the entire utterance, as delineated in the orthographic transcription, containing them; only the foreground channel for each laughter instance was inspected. Since the absence of a timestamp was due mostly to a transcribed, non-lexical item before and/or after the laughter instance, segmentation consisted of determining a boundary between laughter and, for example, throat-clearing. No attempts were made to internally segment contiguous stretches of laughter into multiple bouts.

All of the 1108 `Comment` instances were segmented manually. This task was more demanding than manual segmentation of the `VocalSound` laughter. Information in the `Description` field on occasion provided cues as to the location and extent of the laugh (e.g., `last two words while laughing`). Laughter start points were placed where the speaker's respiratory function was perceived to deviate from that during speech; in determining the end of laughter, the audible final recovery inhalation which often accompanies laughter [63] was included.

The final laugh bout segmentation was formed by combining the automatically segmented `VocalSound` laughter, the manually segmented `VocalSound` laughter, and the manually segmented `Comment` laughter; due to some overlap among the three sets, a small number of laugh segments were merged to yield 13259 distinct segments.

12.4.4 Annotating Voicing

Binary per-bout voicing was annotated by two annotators as reported in [137], but was thereafter corrected. The inter-annotator agreement was found to be only 0.76–0.79, far lower than expected for a binary acoustic distinction. In a second effort, an experienced phonetitian checked the voicing and boundary precision of each of the 13259 bouts. This led to a change of voicing label in 942 instances. Endpoints were modified in 306 instances, and 50 laughter instances were removed altogether. 11961 laughter instances (90% of the total) were not modified.

12.5 Analysis of the Occurrence of Laughter

Of the total number of 13209 bouts in the ICSI Meeting Corpus, 66.0% are voiced; of the total duration of all laughter (5.54 hours), 74.4% is found in voiced bouts. These 5.54 hours account for 9.4% of the total vocalizing time of 59.17 hours. This is a surprisingly large proportion, given that laughter is often ignored in the automatic processing of multiparty conversation.

12.5.1 Variability Across Participants

The average participant appears to spend 0.98% of their total meeting time in voiced laughter, and 0.35% of their total meeting time in unvoiced laughter. By contrast, the average participant spends 14.8% of their total meeting time in speech. These proportions, however, at least for speech, are a strong function of the average number of participants per meeting in the ICSI Meeting Corpus (approximately 6).

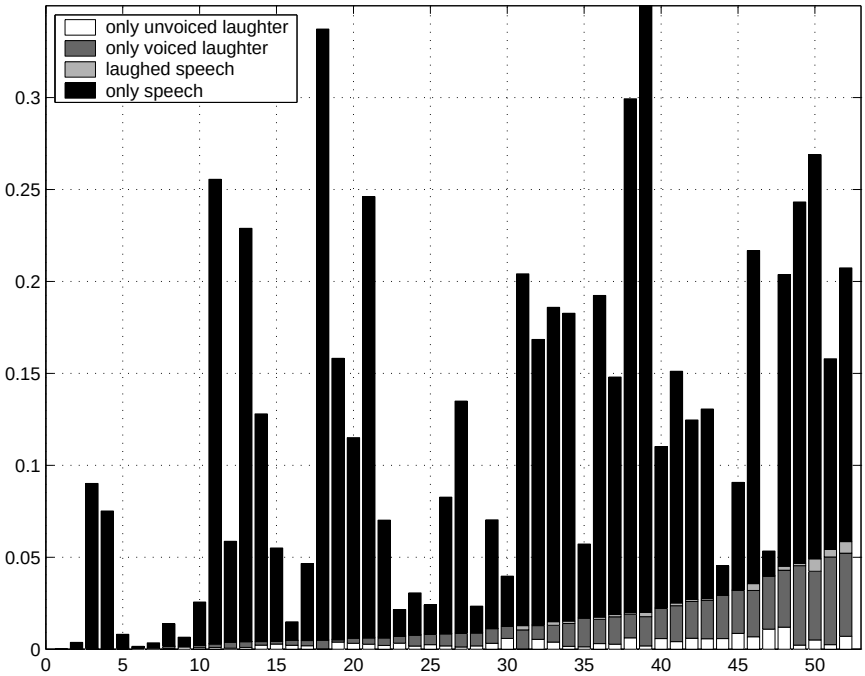


Figure 12.1: Proportion of total recorded time per participant spent in speech, in speech-laughter, in voiced laughter and in unvoiced laughter. Participants are shown in order of ascending proportion of (all) laughter $\mathcal{L} \equiv \mathcal{L}_V \cup \mathcal{L}_U$.

Both types of laughter appear to be very unevenly distributed across participants. Figure 12.1 depicts the proportion of overall time spent in voiced laughter ($\mathcal{L}_V$), unvoiced laughter ($\mathcal{L}_U$), "speech-laughter" ($\mathcal{L}_S \equiv \mathcal{L} \cap \mathcal{S}$), and speech excluding "speech-laughter" ($\mathcal{S} - \mathcal{L}_S$), for each of the 52 participants in the corpus. Although almost all participants appear capable

of producing both voiced and unvoiced laughter, it is easily seen that the proportions of voiced and unvoiced laughter, and of laughter and speech, are only weakly correlated.

12.5.2 Static Aspects

Laughter is not only unevenly distributed across participants, it is also unevenly distributed in time. The normalized distributions of bout durations are shown in panel (a) of Figure 12.2, individually for voiced and unvoiced bouts. Also shown are the normalized distributions of talk spurt durations. The durations of both types of laughter appear to be log-normally distributed, with the most likely duration of bouts, containing voicing, approximately twice as long as that of bouts not containing voicing. The distribution of the durations of talk spurts, on the other hand, appears to be defined by the sum of two log-normal curves. Laughter bouts of either type, as short as the smaller of the two talk spurt curve means (approximately 0.6 s), are quite rare.

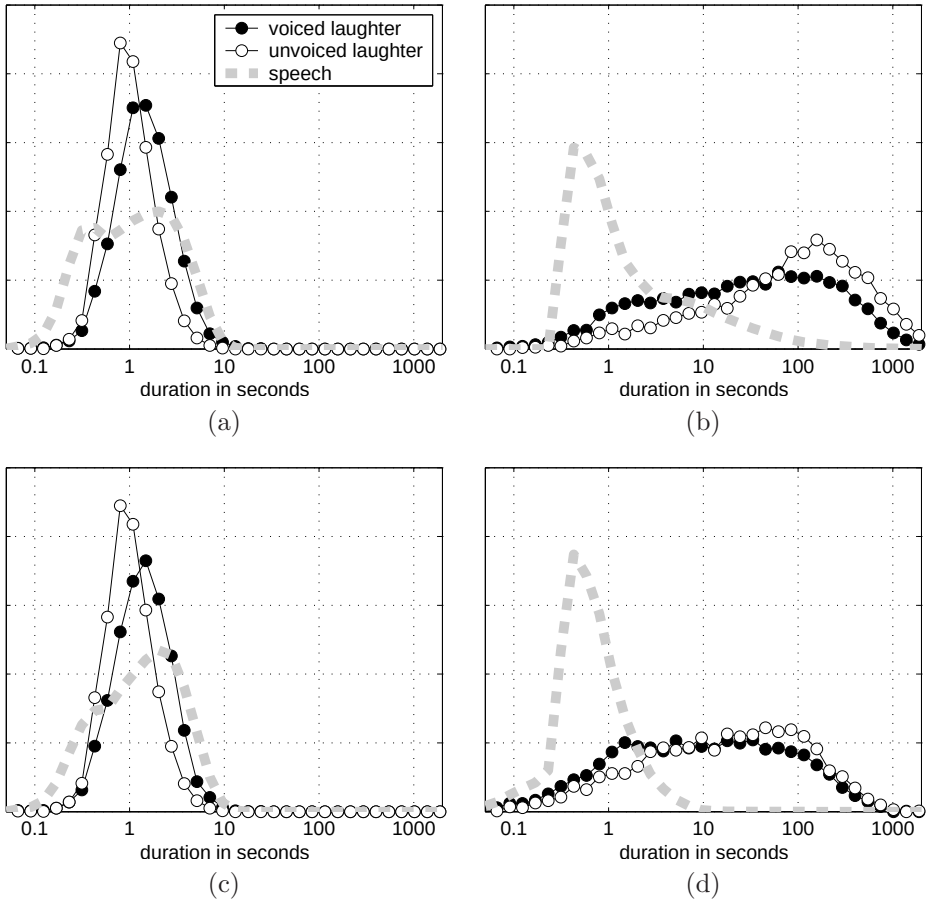


Figure 12.2: Normalized distributions of the durations (in seconds along the x-axis) of: (a) laugh bouts and talkspurts; (b) intervals between laugh bouts and intervals between talkspurts; (c) laugh bout islands and talkspurt islands; and (d) intervals between laugh bout islands and intervals between talkspurt islands. Legend for all panels as for (a).

Panel (b) depicts the normalized distributions of inter-bout interval durations, for intervals terminated by a bout produced by the same participant who produced the interval-initiating bout. The most likely interval between voiced bouts appears to be just over 1 minute long, and approximately 2 minutes between unvoiced bouts. The interval durations between two talk spurts produced by any same participant are on average two orders of magnitude shorter.

Panel (c) shows the normalized distribution of the duration of laugh bout *islands*, which are defined to be contiguous intervals of any participant laughing. If participants laughed one after another, the durations of such participant-unattributed islands could be expected to be much longer than those of participant-attributed bouts. However, as is evident in panel (c), that turns out to not be the case. The distribution of durations of bout islands, for both voiced and unvoiced laughter, appears near-identical to that of durations of individual bouts. This suggests that laughter bouts partially overlap or abut the bouts produced by others only infrequently, and/or that when they do overlap the bouts of others they do so nearly completely. That is clearly not the case for talk spurts; forming talk spurt islands de-balances the priors on the two log-normal distributions which together appear to define talk spurt durations in panel (a).

Inter-island durations are depicted in panel (d). The plot shows that participant-unattributed laughter occurs more frequently than does laughter from any single participant, for both voiced and unvoiced bouts.

These graphical results suggest that speech and laughter incur very different amounts of overlap. Measuring this explicitly, for speech $\mathcal{S}$, both types of laughter $\mathcal{L}_V$ and $\mathcal{L}_U$, and several logical combinations of $\mathcal{S}$, $\mathcal{L}$, $\mathcal{L}_V$, and $\mathcal{L}_U$, yields Table 12.3.

Segmentation	Vocaling Time (hrs)					
	per part.	per meet.	number of simultaneously vocalizing participants			
			1	2	3	≥4
$\mathcal{S}$	54.16	50.09	46.29	3.53	0.24	0.02
$\mathcal{L}$	5.54	3.25	1.96	0.71	0.31	0.27
$\mathcal{L}_V$	4.12	2.49	1.53	0.54	0.23	0.18
$\mathcal{L}_U$	1.42	1.27	1.12	0.13	0.01	0.00
$\mathcal{S} \cup \mathcal{L}$	59.17	51.25	45.35	4.60	0.84	0.46
$\mathcal{S} \cup \mathcal{L}_V$	57.77	50.97	45.65	4.30	0.70	0.32
$\mathcal{S} \cup \mathcal{L}_U$	55.56	50.60	46.11	4.07	0.38	0.04
$\mathcal{S} \cap \mathcal{L}$	0.27	0.27	0.26	0.01	0.00	0.00
$\mathcal{S} - \mathcal{S} \cap \mathcal{L}$	50.36	46.69	43.26	3.19	0.22	0.02
$\mathcal{L} - \mathcal{S} \cap \mathcal{L}$	4.88	3.02	1.90	0.65	0.27	0.19

Table 12.3: Overlap for segmentations under different combinations of speech ($\mathcal{S}$), voiced laughter ($\mathcal{L}_V$), unvoiced laughter ($\mathcal{L}_U$), and all laughter ($\mathcal{L} \equiv \mathcal{L}_V \cup \mathcal{V}_U$); speech-laughter is shown as $\mathcal{S} \cap \mathcal{L}$. The total recorded meeting time is 66.31 hours. Column 2 ("per part.") shows the total vocalization time, summed over all participants in all meetings. Column 3 ("per meet.") shows the total meeting time during which one participant was vocalizing; columns 4 through 7 break this quantity down in terms of the number of simultaneously vocalizing participants.

The table shows that the ICSI Meeting Corpus contains 54.16 hours of speech, in participant time, which are compressed into 50.09 hours of meeting time. Approximately 4 hours of speech are overlapped with the speech of others, the overwhelming majority with the speech of *one* other. For laughter this compression is much more drastic: 5.54 hours of laughter occur in 3.25 hours. Furthermore, a much larger proportion of laughter than of speech occurs in degrees of overlap of 3 or more. Even though approximately 10 times as much speech is produced as laughter, when 3 participants are observed to be vocalizing simultaneously the likelihood that they are all laughing is higher than the likelihood they are all speaking.

The lines corresponding to $\mathcal{L}_V$ and $\mathcal{L}_U$ in Table 12.3 indicate that voiced laughter is responsible for a disproportionate amount of laughter overlap. Unvoiced laughter rarely overlaps other unvoiced laughter, but, as the disparity between the last two columns for $\mathcal{L}$ and $\mathcal{L}_V$ suggests, if frequently overlaps with voiced laughter.

Laughter of both types appears to overlap with speech. For degrees of overlap ≥3, the amount of time in $\mathcal{V} \equiv \mathcal{S} \cup \mathcal{L}$ is higher than that in either $\mathcal{S}$ or $\mathcal{L}$ by more than 50%rel.

The last three lines in Table 12.3 show that "speech-laughs" $\mathcal{S} \cap \mathcal{L}$ does not overlap with "speech-laughts" from other participants, and so behaves more like speech than it does like laughter.

12.5.3 Dynamic Aspects

Static descriptions do not indicate how a conversation achieves entry or egress into individual degrees of overlap. To study dynamic aspects of the occurrence of overlap in laughter, the degree-of-overlap (DO) and extended-degree-of-overlap (EDO) models of Section 6.4 are exploited. To apply these discrete-time modeling methods, the segmentations of interest, namely $\mathcal{S}$, $\mathcal{L}$, $\mathcal{L}_V$, and $\mathcal{L}_U$, must first be discretized as described in Section 6.2.2. Two alternate frame steps were used, 100 ms and 500 ms.

Figure 12.3 shows the inferred DO model transition probabilities, for DO models trained using the four segmentations with $K_{max} = 4$ and a frame step of 500 ms. For speech, in panel (a), it can easily be seen that the state labeled "1", i.e., one participant talking at a time, is the most stable. Transitions into states "0" and "2" occur with probabilities 8% and 6%, respectively, but transitions from those states back to "1" have probabilities of almost 50%. Transitioning into "3" from "1" has a probability below 5% (not shown), but from "2" the probability is 5%. As for "2", however, the probability of leaving "3" is much higher than entering it; "3" can also egress directly into "1" with a probability (21%) which is much higher than that of entering "3".

Laughter, shown in panel (b), exhibits significant differences. First, contiguous intervals of "0" are much longer than for $\mathcal{S}$; the probability of self-transition is 99%. The probability of egressing "0", into any state, is hence quite low (and not shown). However, once "1" is entered, the probability of transitioning to "2" is higher than for speech, and transitioning back into "1" from "2" once there is much lower than for speech. The same is true for transitions between "2" and "3". Finally, occupation times of both "2" and "3" are longer than for speech.

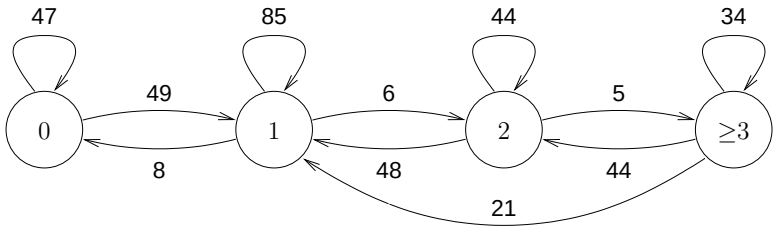
Panels (c) and (d) show similar behavior to panel (b), except that for panel (d), which describes unvoiced laughter dynamics, transitions from "1" to "2" and from "2" to "3" are not shown because their probability is $< 5\%$. As noted in the previous section, voiced laughter appears to account for most of the multi-participant overlap that occurs in laughter.

These diagrams suggest but do not explicitly demonstrate that overlap, in speech as in laughter, arises when a participant begins vocalizing while other participants are already doing so, and thereby increments the degree of overlap by one. As argued in Section 6.4.1, the DO model does not distinguish this case from that in which N participants vocalize, and then stop vocalizing while $N+1$ *other* participants begin vocalizing. The EDO model, on the other hand, does make this distinction. A subset of EDO transitions, for models trained using the four segmentations $\mathcal{S}$, $\mathcal{L}$, $\mathcal{L}_V$, and $\mathcal{L}_U$, with $K_{max} = 3$, is shown in Table 12.4.

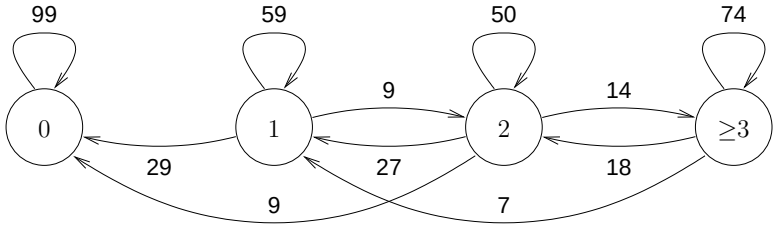
Transition			100 ms frames				500 ms frames			
n_i	o_{ij}	n_j	$\mathcal{S}$	$\mathcal{L}$	$\mathcal{L}_V$	$\mathcal{L}_U$	$\mathcal{S}$	$\mathcal{L}$	$\mathcal{L}_V$	$\mathcal{L}_U$
1	1	1	93.6	87.5	89.1	90.0	82.6	57.7	62.1	63.7
1	1	2	1.5	3.7	3.1	1.4	6.0	8.6	7.4	4.2
1	1	$\geq$3	0.0	0.2	0.2	0.0	0.4	2.3	1.9	0.3
2	1	1	20.8	10.6	9.3	12.4	46.8	26.6	24.2	32.8
2	2	2	75.9	82.8	85.3	85.3	39.3	47.0	53.1	51.1
2	2	$\geq$3	1.9	5.5	4.6	1.4	4.2	13.5	12.0	3.7
$\geq$3	1	1	3.8	0.6	0.5	1.2	20.4	6.7	5.7	13.3
$\geq$3	2	2	26.7	6.9	6.4	14.3	41.1	17.4	17.5	33.9
$\geq$3	$\geq$3	$\geq$3	68.3	92.1	92.9	84.5	27.8	70.9	72.9	48.9

Table 12.4: Select EDO transition probabilities (n_i, o_{ij}, n_j), and their values as inferred from the speech ($\mathcal{S}$), voiced laughter ($\mathcal{L}_V$), unvoiced laughter ($\mathcal{L}_U$), and all laughter ($\mathcal{L} \equiv \mathcal{L}_V \cup \mathcal{L}_U$) segmentations, for non-overlapping frames of 100 ms and 500 ms.

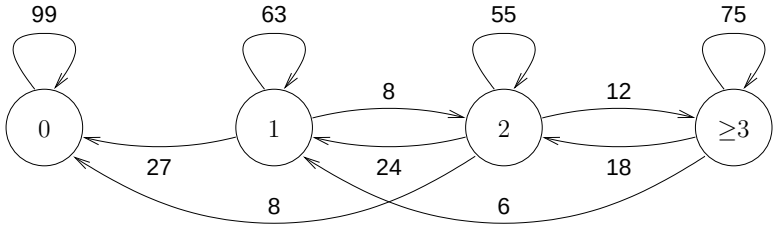
The table shows transition probabilities for $\Delta_T = 100$ ms, a frame step used ubiquitously in this thesis, as well as for $\Delta_T = 500$ ms for direct comparison with Figure 12.3. As can be seen when $\Delta_T = 500$ ms, contiguous stretches of talk are much longer than contiguous stretches of laughter, but at least for laughter in general $\mathcal{L}$ and voiced laughter $\mathcal{L}_V$, transition into overlap of degree 2 is more likely than in speech. When in $n_i = 2$, the probability of remaining in that state, with the same participants vocalizing, is approximately 50% for all three laughter segmentations, and is higher than the probability of egressing to $n_j = 1$. The opposite trend can be seen for speech.



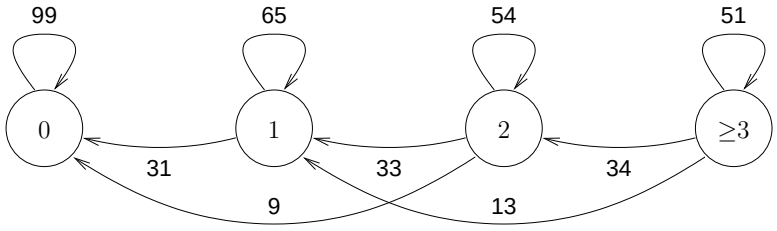
(a) $\mathcal{S}$



(b) $\mathcal{L}$



(c) $\mathcal{L}_V$



(d) $\mathcal{L}_U$

Figure 12.3: 4-state DO models of multi-participant speech, laughter, voiced laughter, and unvoiced laughter. State occupation lasts 500 ms. Transitions whose probabilities are $< 5\%$ are elided for clarity.

For $n_i = 3$, when three participants are vocalizing, the probability that the same three participants or more continue to vocalize is greater than 70% for all laughter and for voiced laughter. For speech, the most likely transition is for exactly one of the vocalizing participants to stop vocalizing, leading to $n_j = 2$.

With a smaller frame size, e.g. $\Delta_T = 100$ ms, these differences between speech and laughter are less striking because 100 ms is much shorter than most contiguous speech spurts and contiguous laugher bouts. For example, it appears that for both speech and laughter, overlap of degree 3 is most likely to persist at this frame step.

12.5.4 Voicing in Laughter

A final question explored in this chapter is whether the presence of voicing during laughter is random. Voiced and unvoiced laughter has been shown to be deployed contrastively in social situations, and therefore there is the possibility that situation type largely predetermines whether a bout is voiced or not.

In this subsection, "situation type" is equated with *local vocal interaction context*, characterized simultaneously by the three segmentations $\mathcal{S}$, $\mathcal{L}_V$, and $\mathcal{L}_U$. As the above analyses have demonstrated, when multiple participants laugh simultaneously, it is very likely that at least one of them is employing voicing.

To render analysis tractable in the event of large K, participant states are not explicitly modeled in a joint way. Instead, for each participant k, $1 \leq k \leq K$, vocalization by other participants is modeled in the feature space (cf. Chapter 8), using a collapsed version of a temporally short neighborhood snapshot. The 11 features used here, computed at each instant t at which participant k is laughing, are:

1. whether k is speaking at $t-1$;
2. whether k is speaking at $t+1$;
3. the number of other participants speaking at $t-1$;
4. the number of other participants speaking at t;
5. the number of other participants speaking at $t+1$;
6. the number of other participants laughing with voicing at $t-1$;
7. the number of other participants laughing with voicing at t;
8. the number of other participants laughing with voicing at $t+1$;
9. the number of other participants laughing without voicing at $t-1$;
10. the number of other participants laughing without voicing at t; and
11. the number of other participants laughing without voicing at $t+1$.

The first two features are $\in \{0,1\}$, while the following nine $\in \{0,1,\cdots,K-1\}$. The question posed is whether the presence of voicing, at time t for participant k, can be predicted from these 11 features alone. If so, then voicing during laughter in meetings is clearly not random, and is correlated with even this simplified a representation of context.

To answer this question, a decision tree is trained for each of three types of discretized $\mathcal{L}$ segmentation frame:

1. *initiation* frames, which are the first in a bout;
2. *termination* frames, which are the last in a bout; and
3. *continuation* frames, which are neither.

It is possible for a laughter frame to be both an initiation and a termination frame, but otherwise frames are at most one of the above three types. For each type, a C4.5 decision tree is trained using all meetings in the ICSI Meeting Corpus, with binary target values given by each meeting's $\mathcal{L}_V$ (or $\mathcal{L}_U$) segmentation. Tree construction via C4.5 is followed by χ^2 pruning. Tree nodes which survive pruning represent statistically significant partitions of context; each pruned decision trees represents a nested χ^2 test.

This experiment identifies no significant distinction between the conversational context of voiced and unvoiced laughter continuation (the third frame type). That is, there appears to be no significant difference in the kinds of interactions that occur during voiced and unvoiced laughter bouts away from bout edges.

For initiation and termination frames, on the other hand, context *is* discriminative. The two classification trees, once pruned, are quite small; they are shown in Figure 12.4. Somewhat surprisingly, they are symmetric.

In attempting to predict the voicing of a frame which initiates a bout, the most useful contextual feature, of those studied, is whether others will be laughing with voicing at $t+1$; in other words, starting a voiced bout is significantly

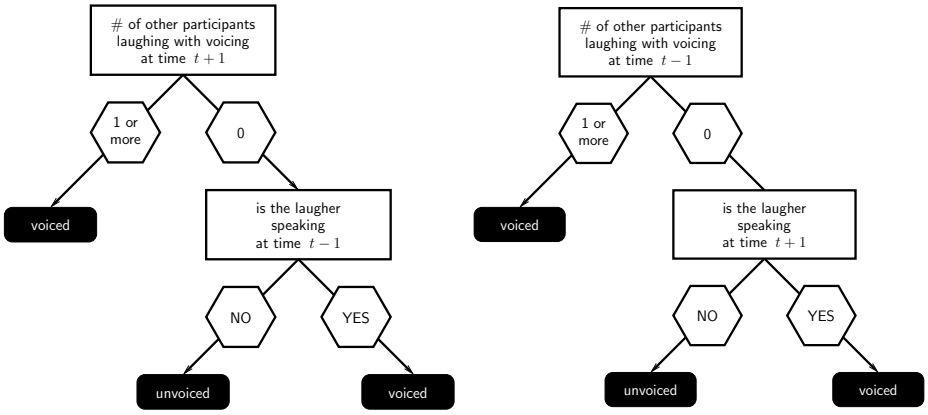


Figure 12.4: Automatically inferred and pruned decision trees for predicting voicing in laughter based on multiparticipant vocal activity context: (a) laughter initiation, and (b) laughter termination. Questions are shown in white rectangles, alternative answers in white hexagons, and leaf decisions in black rectangles.

more likely to be followed by at least one other participant laughing with voicing. In attempting to predict the voicing of a laugh bout which is terminating , the most useful feature is whether others were laughing with voicing at $t-1$. The next most useful, and only other, feature explored by the initiation (termination) tree is whether the laugher is speaking immediately before (after) laughing. Together, the two trees predict that a bout of laughter is unvoiced only if others are not laughing with voicing just following its initiation and just prior to its termination, and the laugher is not speaking immediately before or immediately after the bout.

12.6 Potential Impact

This chapter has described three achievements; each of these has the potential to be impactful in conversation processing in general.

First, this chapter has described the creation of a complete laughter segmentation for the largest currently available corpus of longitudinal recordings of the same groups in naturally occurring conversation. Although other laughter segmentations for this same corpus exist, they consist of only those laughs to which the original orthographic transcription team allocated separate speaker contributions, or utterances. As a result, those segmentations possess only approximately a third of all the laughter bouts in the segmentation described in this chapter. The more complete segmentation has considerable potential in the study of laughter in conversation, particularly of laughter near to the laugher's speech. To the author's best knowledge, no similar segmentation for any other comparable corpus exists which would meet this need.

Second, the proportion of all vocalization effort by time that laughter accounts for has been shown to be surprisingly high. For every nine seconds of speech, participants on average produce almost one second of laughter. This by itself warrants revisiting the role of laughter in conversational situations, and meetings in particular. For the most part, current speech understanding systems ignore (or identify to discard) laughter, and appear poised to continue to do so for the foreseeable future.

Third, the analysis of the dynamics of laughter overlap, presented in Subsection 12.5.3, has implemented another application of the EDO model in studying vocal interaction in multi-party conversational settings. Such analysis can be easily extended to other binary vocalization types, for example breathing, which is gaining some attention in automatic spoken dialogue system research. The same analysis can be applied to any binary behavior type, including those obtainable from the video modality. Video-only, and multi-model approaches involving both video and audio, are currently being widely explored for online speech processing applications.

12.7 Relevance to Other Chapters

This chapter is directly related to the following chapter, which treats the automatic detection of laughter in conversation. The segmentation produced here, and its annotation with voicing, is used as the reference segmentation for scoring detection results. Furthermore, the quantitative observations presented here are used to inform the construction of detector state topologies. Most importantly, it will be observed in Chapter 13 that, because laughter is generally not distributed in turns, constraining the number of simultaneously vocalizing participants does not aid in laughter activity detection to the same extent that doing so aided in speech activity detection (cf. Chapter 11).

The references produced in this chapter are also used in Chapter 15 and Chapter 17 as input. That is, tasks in those chapters assume the availability of a perfect laughter detector. Its output is $\mathcal{L}$, $\mathcal{L}_V$, or $\mathcal{L}_U$, as described here.

12.8 Summary

This chapter has presented an analysis of the occurrence of laughter in naturally occurring multi-party conversation. The effort began with the semi-automatic production of a segmentation of laughter, based on an available orthographic transcription which contained mark-up for laughter and timestamps for the edges of nearby words. It was shown that retaining the overwhelming majority of the laughter originally transcribed leads to a proportion of laughter which far exceeds expectations, of approximately 10% of vocalization effort by time. Investigation into the static aspects of laughter's occurrence showed that laughter differs substantially from speech, particularly in overlap. Dynamic analysis, using models proposed in this thesis, confirmed that laughter occurs less frequently in time and that laughter overlap also arises less frequently. However, once it does arise, laughter overlap is much more likely to be sustained than is speech overlap. This corroborates the accepted wisdom that laughers do not take turns in the same way that they take turns at talking, but also leaves room for inviting group laughter with an opening bout, as conversation analysis has shown [76].

Although the presented numerical findings may serve to re-orient current research efforts to pay closer attention to laughter in interaction, there is significant scope for extending the described techniques to produce joint speech and laughter density models. These extensions are straightforward, particularly in combination with the techniques of Chapter10.

12.9 Future Directions

A critical future avenue of research, if speech understanding systems are to leverage laughter to the extent that its frequency of occurrence suggests they should, is to determine what users mean or intend when they laugh. There appear to be many contrastive uses of laughter in dialogue, e.g., to demonstrate understanding of humor, to express amusement, to affiliate, to apologize, to hide embarassment, and others. It is not clear at this time whether these distinctions are acoustically or pragmatically communicated, and whether their function is independent of what is said. The techniques applied in this chapter can be usefully extended to explore this issue. Although perceptual annotation of laughter (cf. [136]) has been attempted, preliminary analysis suggests that annotators do not sufficiently agree. However, it may be possible to infer function without annotation, by building models to predict subsequent speech (and speech type) as well as laughter, from all participants, along the lines proposed in Chapter 10. Falling perplexity on unseen conversations, as used in that chapter, could be a useful indicator that alternative modeling approaches are successfully leveraging the presence of some multiparticipant patterns of laughter, but not others, eventually leading to a data-driven classification scheme.

Chapter 13

Automatic Laughter Activity Detection*

13.1 Introduction

Like speech activity detection, laughter activity detection involves the segmentation of a participant's vocal production sequence into laughter and non-laughter intervals. If multiple types of laughter are envisioned, the task also entails the classification of laughter intervals into types.

Detecting laughter is important in conversation, because, as shown in the previous chapter, it occurs much more rarely than speech but accounts for almost 10% of vocalization effort by time. It can therefore be assumed that, once it does occur, it significantly alters the vocal activity context for an appreciable duration. Successful inference of the underlying reasons for its occurrence, provided the latter can be detected automatically, is likely to offer important insight as to what is going on in a conversation under study. In automatic systems, the detection of laughter has been argued to be relevant to general discourse segmentation [120], topic change detection [65], meeting recording browsing [9], inference of humorous intent and emotional expression [193], classification of perceived emotional valence [134], and conversational hotspot detection [225].

The work described in this chapter was presented in [151, 128].

13.2 Related Work

Laughter detection and its computational modeling have received a large amount of attention in the last two decades. The focus has often been on acoustically discriminating between laughter and speech, and corpora for study have tended to ignore the tactical use of laughter in conversational interaction.

In meetings, automatic detection of laughter was spearheaded by [120], in which group laughter was detected without prior meeting segmentation and was not attributed to specific participants. This work was unique in that it used farfield recordings of conversation, and appeared to rely on pragmatic cues — namely the number of participants vocalizing simultaneously — at least as much as on the acoustic features of laughter. It also set the stage for future research by benchmarking results on the `Bmr` meetings of the ICSI Meeting Corpus.

Subsequent work on aspects of laughter recognition has almost exclusively used this corpus, but has been carried out using its nearfield recordings. [213, 214] focused on the discrimination of vocal activity into two classes, laughter $\mathcal{L}$ and speech $\mathcal{S}$, given pre-segmented exemplars. Laughter exemplars were selected from a segmentation of laughter consisting of instances which the original ICSI transcription team had placed into laughter-only utterances. [214] reported equal error rates of 3%, even on meetings held by groups other than `Bmr` not seen during training, when the mean instance durations of both speech and laughter were 2 seconds.

Work on segmenting laughter in meetings was treated in [215, 122, 123]. It relied on the same reference laughter segmentation as [213, 214], but aimed at automatically finding segment boundaries rather than discriminating among pre-segmented instances. Using the same reference segmentation as in [214], [215] reported equal error rates of 11% on

*The work in this chapter was conducted in collaboration with Tanja Schultz.

test Bmr meetings. [122] achieved an equal error rate of 7.9% on the same set, but excluded from their detector channels which were inactive. Extension of this detector in [123] led to equal error rates of 5.4%.

Aspect	$\mathcal{L}/\mathcal{S}$ class.		$\mathcal{L}/\neg\mathcal{L}$ segm.			this
	[213]	[214]	[215]	[122]	[120]	work
close-talk microphones	✓	✓	✓	✓		✓
farfield microphones					✓	
single channel at-a-time	✓	✓	✓	✓		
multi-channel at-a-time					✓	✓
participant attribution	✓	✓	✓	✓		✓
only group laughter					✓	
only isolated laughter	✓		✓	✓		
only clear laughter		✓				
rely on pre-segmentation	✓	✓	?			
rely on prior rebalancing	✓	✓	?			
rely on channel exclusion			?	✓		

Table 13.1: Overview of previous research on laughter/speech ($\mathcal{L}/\mathcal{S}$) classification and laughter/non-laughter ($\mathcal{L}/\neg\mathcal{L}$) segmentation, and of the current work, in terms of several differentiating aspects. From [151]; "?" indicates unspecified characteristics.

The current chapter describes an approach which differs from the above, in three main ways, while retaining the same corpus in the nearfield condition to enable some comparison. Several of the differences are shown in Table 13.1. First, the task here is the simultaneous segmentation and classification of several types of vocal activity, of which laughter and speech are two examples. This makes the proposed task more difficult from a design point of view than a collapsed-class binary task, but renders the proposed system's results potentially more informative of interaction. Second, the laughter segmentation which is used is that of all the laughter which was transcribed by the original ICSI annotators; this includes not only the utterances containing only laughter, but also all those in which laughter is either interspersed or concommittant with speech. This renders models of the classes less distinct, and therefore classification errors more likely; however, it eliminates the possibility that the laughter subset heretofore explored is not the one most useful to downstream understanding applications. Third, in scoring the proposed system's output, all channels are considered over the entire duration of each meeting (excluding the Digits calibration interval). This has the effect of increasing false alarms, but makes the proposed systems deployable without *a priori* information.

13.3 Dataset Use

The experiments of this chapter are conducted using the ICSI Meeting Corpus. 67 of the 75 meetings in the corpus are of one of three types, Bed, Bmr, and Bro, representing longitudinal recordings of three groups at ICSI. The total number of distinct participants in these three subsets is 23. The sets are largely disjoint in participants; there are 3 participants who attend meetings of both Bmr type and Bro type, and 1 participant who attends meetings of Bmr and Bed types.

The majority of existing work on laughter detection using this corpus has reported results on three Bmr meetings or on portions thereof, with no results available on held-out data. To allow for comparison with that work, these three meetings (Bmr029, Bmr030, and Bmr031) comprise the DevSet in this chapter. The remaining Bmr meetings, as elsewhere, comprise the TrainSet. All meetings of type Bed and Bro are therefore available as a held-out EvalSet. The remaining 8 meetings, of other types, are not used.

13.4 Assessment of Performance

In principle, an automatic laughter detector hypothesizes, for each frame of audio, a binary label of laughter ($\mathcal{L}$) or non-laughter ($\neg\mathcal{L}$). For close-talk-microphone-instrumented participants, automatic laughter detection systems are also expected to implicitly perform *laugher diarization* (much as nearfield speech detection systems perform speaker diarization), and thus the two labels represent *nearfield-laughter* and *non-nearfield-laughter* (the latter includes farfield laughter and non-laughter), respectively.

However, because the $\neg\mathcal{L}$ class contains many acoustically distinct phenomena, including both speech and non-speech, which may provide constrastive conditioning contexts for the occurrence of laughter, it is expected that better laughter detection performance can be achieved by considering the 3-class problem of discriminating among $\mathcal{L}$, speech $\mathcal{S}$, and neither ($\mathcal{N} \equiv \neg\mathcal{V} \equiv \neg(\mathcal{S} \cup \mathcal{L})$). For the purposes of this chapter, the phenomenon of "speech laughs" [175] in which the same participant simultaneously speaks and laughs, is mapped to speech only. It has been shown to be relatively infrequent in meetings [135], and is denoted $\mathcal{L}_S \equiv \mathcal{L} \cap \mathcal{S}$ where reference to it is made.

The target speech category for the presented recognition experiments, which includes $\mathcal{L}_S$, is simply $\mathcal{S} \cup \mathcal{L}_S \equiv \mathcal{S}$. The target laughter category exludes speech laughs, $\mathcal{L}' \equiv \mathcal{L} \cap \neg\mathcal{S}$. This chapter also considers two acoustically different types of laughter, namely voiced laughter $\mathcal{L}_V$ and unvoiced laughter $\mathcal{L}_U$, with $\mathcal{L} = \mathcal{L}_V \cup \mathcal{L}_U$. This yields a 4-class vocal activity recognition task. Recognition experiments will aim at detecting these categories with speech laughs excluded, namely $\mathcal{L}'_V \equiv \mathcal{L}_V \cap \neg\mathcal{S}$ and $\mathcal{L}'_U \equiv \mathcal{L}_U \cap \neg\mathcal{S}$, respectively.

13.4.1 Metrics

For both the 3-class and the 4-class problems, the primary metric used will be the F-score, as defined in Chapter 11. F-scores will be reported for all behaviors where relevant, including $\mathcal{L}'$, $\mathcal{L}'_V$, $\mathcal{L}'_U$, $\mathcal{S}$, and $\mathcal{V}$. F-scores are computed from confusion matrices in which data points are 100-ms frames of vocal activity.

There are some consequences to choosing this metric, which are due to the infrequency of laughter relative to speech (and to meeting participation time). It has been common to report equal error rates instead, in laughter detection work, which normalize the number of false negatives by the total amount of time participants are *not* laughing. Since the latter quantity is large, false alarm rates are low. As mentioned in Section 13.2, equal error rates of $< 10\%$ are quite commonly reported. In contrast, precision as it is used in F-score computation normalizes false negatives by the total amount of hypothesized laughter, which must be small if prior distributions are to equal the empirical distributions observed. This aspect makes it difficult to compare the performance achieved by the systems of this chapter to that of systems published in other work.

For completion, it should be noted that, given DevSet priors, a system which always guesses $\mathcal{N}$ achieves an F-score of 92.8% for $\mathcal{N}$ and of zero for all other classes; a system which always guesses $\mathcal{L}'$ achieves an F-score of 3.8% for $\mathcal{L}'$ and of zero for all other classes; and a system which always guesses $\mathcal{S}$ achieves an F-score of 20.6% for $\mathcal{S}$ and of zero for all other classes.

13.4.2 References

To compute the proposed metrics, a 3-way or 4-way reference vocal activity segmentation must be constructed first.

The 3-way segmentation is formed by intersecting a laughter segmentation and a speech segmentation. The former comes from the laugh bout segmentation described in Chapter 12, and is available for all the meetings in the ICSI Meeting Corpus. The speech segmentation is formed using the word and word-fragment start and end times from automatic forced alignment, available in the ICSI MRDA Corpus [202]. Inter-word gaps shorter than 0.3 s were bridged to yield talkspurts.

As mentioned above, where the logical AND of the speech and laughter segmentations is not empty, one or more "speech laughs" must be occurring. All such instants are mapped to speech only.

For the 4-way task which additionally discriminates between voiced and unvoiced laughter, each laugh bout in the reference laughter segmentation must be manually classified as voiced or unvoiced. This was performed in [135], and is described in the previous chapter.

13.5 Baseline

There are relatively few laughter detectors in existence, particularly in the domain of multi-party conversation where laughter appears to be used tactically and is often interspersed with speech (as opposed to joke-punctuating canned audience laughter in television shows). As a result, not much is known about what might be optimal for this task. The baseline proposed here is an independent-participant ergodic HMM decoder, with the topology as shown in Figure 13.1. The frame step is 100 ms, as in Chapter 11. Transitions are governed by bigram probabilities, learned from the best forced-alignment Viterbi pass over all the meetings in TrainSet.

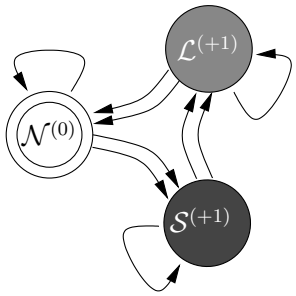


Figure 13.1: An ergodic 3-state topology for the recognition of speech activity $\mathcal{S}$, laughter activity $\mathcal{L}'$, and silence $\mathcal{N} \equiv \neg(\mathcal{S} \cup \mathcal{L})$, of any single participant.

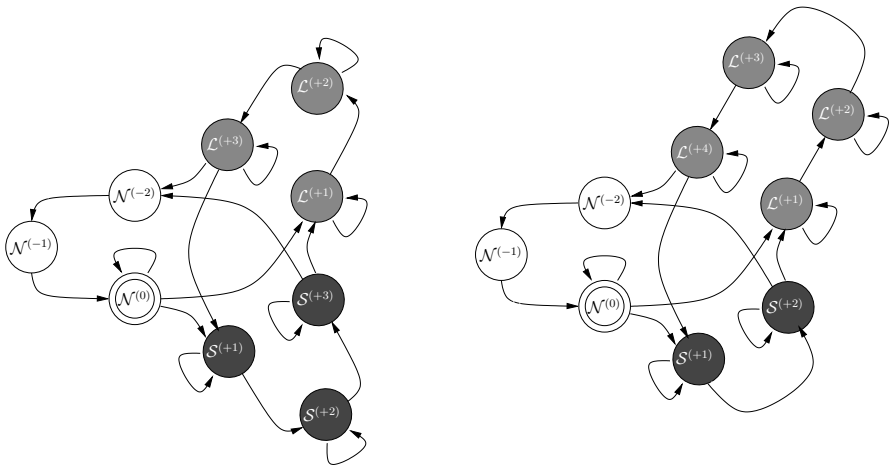
The emission probabilities of each of the $\mathcal{S}$, $\mathcal{L}'$, and $\mathcal{N}$ states are modeled by a 64-component Gaussian mixture model (GMM), defined over a feature vector of 41 elements. These include the first 13 Mel-frequency cepstral coefficients, channel-normalized using cepstral mean subtraction, their first- and second-order differences, and two features, known as minimum and maximum normalized log-energy differences (NLEDs) [20], used to mitigate the effects of crosstalk [181]. This feature vector was referred to as LE+MFCC+Δ+$\Delta\Delta$+NLED in Chapter 11.

13.6 Imposing Minimum Duration Constraints

As for speech activity detection, performance can be expected to improve when occupation of the three behaviors in Figure 13.1 is constrained to be at least as long as the shortest observed durations of the three vocal activity states; detection performance may improve even if the minimum durations are set to be longer than observed minima to eliminate spurious hypotheses. Enforcing minimum duration constraints of $\mathbf{T}_{min} = \{T^{\mathcal{S}}_{min}, T^{\mathcal{L}'}_{min}, T^{\mathcal{N}}_{min}\} = \{0.3, 0.3, 0.3\}$ seconds yields the topology of Figure 13.2(a), of 9 states.

Performance can also be expected to improve when the duration constraints are contrastive, since it has been shown that the phenomena have differently distributed durations (cf. Chapter 12). Panel (b) of Figure 13.2 shows a topology whose minimum duration constraints are $\mathbf{T}_{min} = \{0.2, 0.4, 0.3\}$ seconds, retaining the same overall number of states of 9, as in panel (a). Under the latter constraint of 9 single-participant states in total, the topology in panel (b) yields the highest F-scores for laughter from among those assessed. F-scores may of course be even higher when the overall number of states is allowed to freely exceed 9, but this would preclude direct comparison with multi-participant decoding. As is argued throughout this thesis, the construction of multi-participant topologies is rendered intractable as the number of degrees of freedom per participant grows.

The performance of both topologies, as well as of the ergodic baseline, is presented in Table 13.2. System identifiers begin with the letter "I" to denote independent-participant decoding, to constrast with joint-participant decoding ("J") described in Section 13.9.



(a) IA1 (b) IA2

Figure 13.2: Non-ergodic topologies for the recognition of speech activity $\mathcal{S}$, laughter activity $\mathcal{L}'$, and silence $\mathcal{N} \equiv \neg(\mathcal{S} \cup \mathcal{L})$. Both (a) and (b) allow 9 degrees of freedom per participant, but (b) enforces contrastive minimum duration constraints on the three vocal activity types.

Topology	$\mathbf{T}_{min}$ (s)			F-score (%)		
	$\mathcal{S}$	$\mathcal{L}'$	$\mathcal{N}$	$\mathcal{S} \cup \mathcal{L}$	$\mathcal{S}$	$\mathcal{L}'$
IA0	0.1	0.1	0.1	75.4	87.4	30.9
IA1	0.3	0.3	0.3	76.3	87.7	31.7
IA2	0.2	0.4	0.3	76.3	87.6	32.6

Table 13.2: DevSet F-scores for vocalization $\mathcal{S} \cup \mathcal{L}$, speech $\mathcal{S}$, and laughter $\mathcal{L}'$, for three independent-participant HMM decoders, including the baseline, differing in topology only. The number of states $N \leq 9$ for all systems.

13.7 Recognizing Voiced Laughter

Closer analysis of the results in Table 13.2 reveals that the main error committed by the best-performing of the depicted systems is the hypothesis of silence as laughter. A confusion matrix for all of DevSet is shown in Table 13.3; the reference label for laughter is broken out into laughed speech $\mathcal{L}_S$, voiced laughter excluding laughed speech $\mathcal{L}'_V$, and unvoiced laughter excluding laughed speech $\mathcal{L}'_U$.

As the confusion matrix shows, the total amount of laughter by time in DevSet is 16.4 minutes. However, the IA2 system hypothesizes 46.5 minutes of silence as laughter, leading to very low precision. The likely reason for this is that unvoiced laughter, which contributes approximately a quarter of the acoustic training data for the $\mathcal{L}'$ model, captures a significant proportion of breathing which is considered $\mathcal{N}$ in the references[1].

To validate this hypothesis, the current section groups unvoiced laughter with silence, and retains only voiced laughter in the laughter class. The task continues to be a three-class recognition problem, but between $\mathcal{S}$, $\mathcal{L}'_V$, and $\mathcal{N}' \equiv \mathcal{N} \cup \mathcal{L}'_U \equiv \neg(\mathcal{S} \cup \mathcal{L}_V)$. The F-scores achieved by the resulting systems IB0 and IB2, otherwise identical to systems IA0 and IA2 of

[1]It is also important to note that voiced bouts of laughter in this work are those which are not entirely unvoiced; as a result, a large portion of both calls and inter-call intervals in voiced laughter may actually be unvoiced, and thus as confusable with breathing as unvoiced bouts are.

References \ Hypotheses	$\mathcal{N}$	$\mathcal{L}'$	$\mathcal{S}$	Σ
$\mathcal{N}$	655.8	46.5	13.9	*716.2*
$\mathcal{L}'_U$	0.9	4.3	0.2	*5.4*
$\mathcal{L}'_V$	1.1	8.9	0.4	*10.4*
$\mathcal{L}_S$	0.0	0.3	0.5	*0.8*
$\mathcal{S}$	3.4	5.8	85.3	*94.4*
Σ	*661.2*	*65.7*	*100.4*	*827.2*

Table 13.3: DevSet confusion matrix obtained using system IA2. All quantities in minutes. Reference labels for laughter $\mathcal{L}$ broken out into three subcategories for analysis. Hypothesized and reference totals for each class are shown in the bottom row and column, respectively, labeled Σ.

the previous section, are shown in Table 13.4. Systems whose identifier contains the letter "B" as the second character denote those described in this section (as opposed to "A" systems which model unvoiced laughter together with voiced laughter).

Topology	$\mathbf{T}_{min}$ (s)				F-score (%)				
	$\mathcal{S}$	$\mathcal{L}'_V$	$\mathcal{L}'_U$	$\mathcal{N}$	$\mathcal{S} \cup \mathcal{L}$	$\mathcal{S} \cup \mathcal{L}_V$	$\mathcal{S}$	$\mathcal{L}'$	$\mathcal{L}'_V$
IA0	0.1	0.1		0.1	75.4	—	87.4	30.9	—
IA2	0.2	0.4		0.3	76.3	—	87.6	32.6	—
IB0	0.1	0.1		0.1	—	78.3	86.6	—	34.6
IB2	0.2	0.4		0.3	—	79.9	86.9	—	36.4
IB3	0.1	2.5		0.4	—	81.7	86.6	—	46.0

Table 13.4: DevSet F-scores for voiced vocalization $\mathcal{S} \cup \mathcal{L}_V$, speech $\mathcal{S}$, and voiced laughter $\mathcal{L}'$, for three independent-participant HMM decoders, differing in topology only. Also shown are the baseline and best systems from Table 13.2.

What the table unambiguously shows is that grouping unvoiced laughter with silence rather than with voiced laughter, while retaining the same decoder structure, improves the detection of vocalization (the union of speech and laughter in this work) by 2.9%abs with the ergodic topology and by 3.6%abs with the non-ergodic topology. This is due largely to the improved detectability of voiced laughter over all laughter, of 3.7%abs and 3.8%abs, respectively, for the two topologies. However, a drop of approximately 0.75%abs in the F-score for speech accompanies these laughter F-score gains.

Table 13.4 also shows the performance of a system which is non-ergodic, like IB2, but whose minimum duration constraints have been modified to maximize the voiced laughter F-score. The $\mathbf{T}_{min}$ which achieves this was found to be $\{0.1, 2.5, 0.4\}$, with very long constraints on laughter. This system, denoted IB3 in the table, achieves an F-score of 46.0% for $\mathcal{L}'_V$, an improvement of 11.4%abs over the ergodic IB0 system, and at no reduction to the F-score for speech.

13.8 Recognizing Voiced and Unvoiced Laughter

The errors committed by the IB systems suggest that voiced laughter is significantly easier to detect than is all laughter. Table 13.5 shows the specific DevSet confusions which the IB2 and IB3 systems commit.

Of the two systems, that in panel (a) compares directly with IA2 in Table 13.3, since the only difference is the grouping of the $\mathcal{L}'_U$ TrainSet data during training. The most visible difference, in terms of performance, is that IB2 posits 11.1 minutes less of laughter overall, thus reducing the amount of laughter hypothesized by the IA2 system for $\mathcal{N}$ by 11.2 minutes. However, IB2 also hypothesizes the majority of unvoiced laughter as laughter, which may be without consequence in practice, but was not intended (and is therefore penalized).

The IB3 system posits still less laughter, by a significant amount of 18.4 minutes, while losing only 0.2 minutes to the

References \ Hypotheses	$\mathcal{N} \cup \mathcal{L}'_U$	$\mathcal{L}'_V$	$\mathcal{S}$
$\mathcal{N}$	666.5	35.3	14.4
$\mathcal{L}'_U$	1.2	3.9	0.3
$\mathcal{L}'_V$	1.3	8.7	0.4
$\mathcal{L}_S$	0.0	0.3	0.4
$\mathcal{S}$	3.6	6.3	84.5
Σ	*672.6*	*54.6*	*100.1*

(a) IB2

Hypotheses: $\mathcal{N} \cup \mathcal{L}'_U$	$\mathcal{L}'_V$	$\mathcal{S}$	Σ
675.9	21.5	18.8	*716.2*
1.5	3.3	0.6	*5.4*
1.2	8.5	0.7	*10.4*
0.0	0.3	0.5	*0.8*
3.9	2.6	87.9	*94.4*
682.4	*36.2*	*108.6*	*827.2*

(b) IB3

Table 13.5: DevSet confusion matrix obtained using system IB3. All quantities in minutes. Reference labels for laughter $\mathcal{L}$ broken out into three subcategories for analysis. Hypothesized and reference totals for each class are shown in the bottom row and column, respectively, labeled Σ.

speech $\mathcal{S}$ model. This represents a reduction of laughter insertions during non-vocalization $\mathcal{N}'$ by 14.4 minutes, and of laughter insertions during speech by 2.7 minutes.

Given that the IB3 system shows such improved separation of voiced laughter from unvoiced laughter, when the latter is grouped with silence, it is possible that unvoiced laughter can be differentiated from silence to some extent, without hurting $\mathcal{L}'_V$ F-scores. This turns the problem into a 4-class recognition task. A sample 4-way topology of the type employed in the decoder is shown in Figure 13.3.

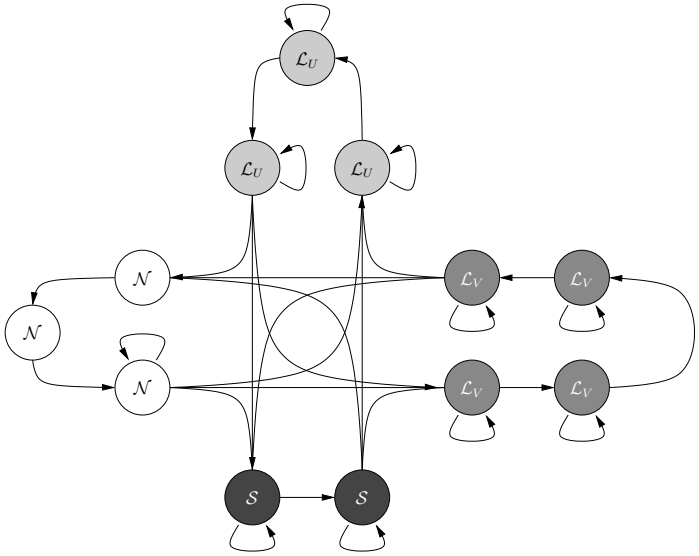


Figure 13.3: Sample non-ergodic topology for the recognition of speech $\mathcal{S}$, voiced laughter $\mathcal{L}'_V$, unvoiced laughter $\mathcal{L}'_U$, and silence $\mathcal{N} \equiv \neg(\mathcal{S} \cup \mathcal{L}_V \cup \mathcal{L}_U)$.

To compare the 4-class systems to the 3-class systems, IC topologies are constructed which are like the IB topologies, with the minimum duration constraints for both laughter types identical; "C" in system identifiers denotes the 4-class problem. The performance of these systems is shown in Table 13.6.

What can be observed from this table is that including a separate sequence of states for unvoiced laughter, and removing unvoiced laughter from the training data for the silence state, leads to systems whose F-scores for voiced laughter are slightly lower. For all three of IC0, IC2, and IC3, $\mathcal{L}'_V$ F-scores are lower by 1.7%abs, 2.0%abs, and 0.6%abs, respectively, than for IB0, IB2, and IB3. However, in all three cases the F-scores for speech are slightly higher, and those for $\mathcal{S} \cup \mathcal{L}_V$ are

Topology	$\mathbf{T}_{min}$ (s)				F-score (%)				
	$\mathcal{S}$	$\mathcal{L}'_V$	$\mathcal{L}'_U$	$\mathcal{N}$	$\mathcal{S} \cup \mathcal{L}$	$\mathcal{S} \cup \mathcal{L}_V$	$\mathcal{S}$	$\mathcal{L}'$	$\mathcal{L}'_V$
IA0	0.1	0.1		0.1	75.4	—	87.4	30.9	—
IA2	0.2	0.4		0.3	76.3	—	87.6	32.6	—
IB0	0.1	0.1		0.1	—	78.3	86.6	—	34.6
IB2	0.2	0.4		0.3	—	79.9	86.9	—	36.4
IB3	0.1	2.5		0.4	—	81.7	86.6	—	46.0
IC0	0.1	0.1	0.1	0.1	71.6	83.5	87.3	25.9	32.9
IC2	0.2	0.4	0.4	0.3	72.6	83.9	87.6	27.6	34.4
IC3	0.1	2.5	2.5	0.4	76.3	84.7	87.2	35.4	45.4
IC4	0.1	3.2	1.4	0.4	74.6	85.2	87.5	32.4	47.7

Table 13.6: DevSet F-scores for vocalization $\mathcal{S} \cup \mathcal{L}$, voiced vocalization $\mathcal{S} \cup \mathcal{L}_V$, speech $\mathcal{S}$, laughter $\mathcal{L}'$, and voiced laughter $\mathcal{L}'_V$, for four independent-participant HMM decoders (labeled ICn), differing in topology only. Also shown are the baseline and best systems from Table 13.2, and those from Table 13.4.

much higher, by 3.0-5.2%abs. This indicates that as performance improves, discrimination between speech and laughter may grow in importance.

When the minimum duration constraints are optimized to maximize $\mathcal{L}'_V$ detection performance, however, the F-score for that class climbs to 47.7%abs, which is 1.7%abs higher than for the best system in which unvoiced laughter is not separately modeled. It is worth noting that the resulting system, IC4, achieves F-scores for all laughter $\mathcal{L}'$ which are higher than those of the baseline IA0, but additionally labels the detected laughter with a voicing label.

13.9 Modeling Interlocutors

The systems described and evaluated up to this point treat participants individually by decoding their channels one at a time (albeit with the cross-channel NLED features). Those systems are now extended in the same way as was done for speech activity detection in Chapter 11. Namely, a multi-participant topology is constructed by computing the Cartesian product of the topologies shown in Figures 13.1, 13.2, and 13.3 (cf. Chapter 6). In doing so, maximum simultaneous vocalization constraints are imposed, to limit the complete topology size. In initial experiments [151], it was determined using DevSet that the optimal constraints for the 3-class A systems were $\mathbf{K}_{max} = \{K^{\mathcal{S}}_{max}, K^{\mathcal{L}'}_{max}, K^{\neg\mathcal{N}}_{max}\} = \{2, 3, 3\}$; those for the 3-class B systems were $\mathbf{K}_{max} = \{K^{\mathcal{S}}_{max}, K^{\mathcal{L}'_V}_{max}, K^{\neg\left(\mathcal{N} \cup \mathcal{L}'_U\right)}_{max}\} = \{2, 3, 3\}$; and those for the 4-class C systems were $\mathbf{K}_{max} = \{K^{\mathcal{S}}_{max}, K^{\mathcal{L}'_V}_{max}, K^{\mathcal{L}'_U}_{max}, K^{\neg\mathcal{N}}_{max}\} = \{2, 3, 1, 3\}$. Optimizing these constraints was informed by the findings of Chapter 12, particularly $K^{\mathcal{L}'_U}_{max}$, describing a behavior which is infrequently overlapped by other producers of $\mathcal{L}_U$.

Table 13.7 shows the results for topologies JA0, JA2, JB0, JB2, JC0, JC2, corresponding to single-participant topologies IA0, IA2, IB0, IB2, IC0, and IC2. Other topologies, characterized by longer minimum duration constraints, could not be trained because of the size of the resulting multi-participant topology.

As the table shows (compared with Table 13.6), systems JA0, JA2, and JB2 achieve higher F-scores for laughter and for voiced laughter than do the corresponding independent-participant systems, by 0.8%abs, 1.9%abs, and 0.9%abs, respectively. JB0 is worse than its independent-participant counterpart, by 0.4%abs. On average, the benefit of multi-participant modeling for laughter, based on these four systems, is positive but modest. The JC systems, however, appear to be significantly outperformed by their independent-participant counterparts. Furthermore, as mentioned above, the best-performing independent-participant decoders cannot be ported to a multi-participant setting.

These results suggest that multi-participant modeling for laughter detection, within the types of systems considered here, may have a place in the future. However, at the current time and with the currently available size of conversational corpora, controlling hypothesized interaction by imposing maximum simultaneous vocalization (across-participant) constraints is not as effective as imposing minimum duration (in-time) constraints on laughter. This appears to be related to the fact that while participants take turns speaking, they do not take turns laughing.

Sys	T_{min}, s				F, %				
	$\mathcal{S}$	$\mathcal{L}'_V$	$\mathcal{L}'_U$	$\mathcal{N}$	$\mathcal{S} \cup \mathcal{L}'$	$\mathcal{S} \cup \mathcal{L}'_V$	$\mathcal{S}$	$\mathcal{L}'$	$\mathcal{L}'_V$
JA0	0.1	0.1		0.1	78.1	—	86.0	31.7	—
JA2	0.2	0.4		0.3	79.5	—	86.7	34.5	—
JB0	0.1	0.1	0.1		—	79.5	84.9	—	34.2
JB2	0.2	0.4	0.3		—	80.9	85.6	—	37.3
JC0	0.1	0.1	0.1	0.1	76.0	81.7	83.7	26.2	27.3
JC2	0.2	0.4	0.4	0.3	78.9	83.3	84.5	30.4	31.2

Table 13.7: DevSet F-scores of detecting vocalization ($\mathcal{S} \cup \mathcal{L}$), speech ($\mathcal{S}$), and laughter ($\mathcal{L}'$) by VAD systems in which participants are decoded jointly; symbols as in the text.

13.10 Generalization to Unseen Data

The performance of several of the described systems on EvalSet is shown in Table 13.8.

Data		$p_{\mathcal{V}}(\mathcal{L})$, %	System			
			IA2	JA2	IC4	
			$\mathcal{L}'$	$\mathcal{L}'$	$\mathcal{L}'$	$\mathcal{L}'_V$
DevSet	`Bmr`(3)	14.94	32.6	34.5	32.4	47.7
EvalSet	`Bed`	7.53	16.7	17.0	14.5	22.0
	`Bro`	5.94	19.1	19.0	16.3	37.1

Table 13.8: Laughter ($\mathcal{L}'$) and voiced laughter ($\mathcal{L}'_V$) detection F-scores on several datasets using three different VAD systems. Also shown is the proportion $p_{\mathcal{V}}(\mathcal{L})$ of vocalization time spent in laughter.

Several observations can be made based on these numbers. First, Table 13.8 shows that performance on both EvalSet subsets is much worse than on DevSet. This appears to be due first and foremost to the amount of laughter present; DevSet contains almost twice as much laughter as the `Bed` meetings, and more than two and a half times as much laughter as the `Bro` meetings. (Although not shown, the $\mathcal{L}'$ F-score for the system labeled as JA2 on TrainSet, whose proportion of laughter is 10.91%, was only 28.4%.) It should be noted that all of the research cited in Section 13.2 has reported performance on only DevSet, as defined here, and performance on other unseen portions of the ICSI Meeting Corpus has not been reported.

Second, the proportion of laughter present does not account for the differences between the two EvalSet subsets. Although it contains more than 25% more laughter than does the `Bro` subset, the `Bed` subset exhibits lower F-scores. This may be due to the fact that there are more participants in the `Bro` meetings who also attend the `Bmr` meetings in TrainSet, suggesting that the acoustic modeling approach in the proposed systems exhibits a non-negligible degree of speaker-dependence.

Third, for EvalSet as for DevSet, F-scores for voiced laughter are much higher than they are for all laughter. This is felicitous because in subsequent chapters, focusing on several downstream applications, it will be shown that voiced laughter appears to be generally more relevant than all laughter.

13.11 Potential Impact

The experiments in this chapter describe the first attempt to detect *all* laughter present in conversation, and not just the most clearly audible or that transcribed in utterances not containing speech. This is important. As subsequent chapters will show, the utility of detecting laughter for downstream conversation processing appears to be strongly correlated with

the temporal proximity of that laughter to the laugher's speech. Laughter bouts which have been assigned their own utterances by conversation transcribers, rather than being included in utterances containing lexical productions, are likely to be distant from speech. This raises a concern whether existing results on laughter detection in meetings port to the arguably more important speech-proximate laughter.

The above experiments have also not relied on the exclusion of "dormant" channels, neither for detection nor for scoring. Such exclusion presents problems in fully automatic settings, not least because, as has been demonstrated in the experiments of this chapter, the largest number of errors is due to confusions between laughter and silence, and not those between laughter and speech as previous research has assumed. This is due in part to crosstalk (as for speech), but also because laughter bouts contain intervals of very quiet acoustic activity between calls, proportionally much more than do talk spurts.

13.12 Relevance to Other Chapters

This chapter has explored the *detection* of laughter activity in multi-party conversation, and is related to subsequent chapters which treat vocal activity segmentations as input. A laughter segmentation is employed in Chapter 15, for the inference of several phenomena which are referred to as epiemotional. Since that chapter, like all subsequent chapters, relies on reference vocal activity, the results obtained in this chapter for *detecting* vocal activity should be used to qualify the obtainability of subsequent results in fully automatic settings.

As Chapter 15 will demonstrate, the use of voicing during laughter is relevant to understanding. In particular, some of the applications described in that chapter will be shown to achieve better results if unvoiced laughter, that which has been shown to be harder to detect in this chapter, is ignored.

13.13 Summary

Laughter detection in multiparty meetings is a viable endeavor, but F-scores achieved for laughter are still much lower than those achieved for speech. Laughter is much more rare than speech, resulting in less training data, and its acoustically salient calls are also interspersed by relatively long stretches of inter-call intervals, which are acoustically similar to silence. Furthermore, approximately a quarter of all laughter by time belongs to unvoiced bouts, which are similar to regular breathing. This is believed to explain why, as this research shows, the largest source of error in detecting laughter are false alarms during reference silence.

The presented experiments suggest that state-space multi-participant modeling is of limited value in laughter activity detection, at least given the current size of conversational corpora. Although for small topologies its impact is positive in some cases, constraining the number of hypothesized simultaneous laughers is not very effective, since in situ it often occurs that participants laugh simultaneously. In contrast, because laughter contains significant stretches of silence, requiring that hypothesized laughter be long — sometimes much longer than its observed minimum durations — leads to considerable F-score improvement. Unfortunately, longer minimum duration constraints lead to geometrically larger multi-participant topologies, making it impossible at the current time to constrain the degree of simultaneous vocalization and the minimum laughter duration in the same decoding pass.

The best F-score reported in this work of 47.7%, for voiced laughter in the popularly used DevSet, is currently the only performance characterization of detection which attempts to find *all* transcribed voiced laughter in the ICSI Meeting Corpus, and which does not rely on the exclusion from consideration of those participants which are known *a priori* to be inactive.

13.14 Future Directions

Detailed analysis of the performance of the system achieving the highest F-score for voiced laughter on DevSet reveals that even when unvoiced laughter is assigned its own class and when the minimum duration constraints for voiced laughter are relatively long, the majority of voiced laughter detection errors are due to insertion during silence. The relevant confusion matrix is shown in Table 13.9.

	$\mathcal{N}$	$\mathcal{L}'_U$	$\mathcal{L}'_V$	$\mathcal{S}$	Σ
$\mathcal{N}$	**649.4**	45.1	5.0	16.7	*716.2*
$\mathcal{L}_U$	0.8	**3.6**	0.7	0.4	*5.4*
$\mathcal{L}_V$	0.9	3.7	**5.6**	1.0	*11.2*
$\mathcal{S}$	3.6	2.2	1.1	**87.6**	*94.4*
Σ	*654.5*	*54.6*	*12.4*	105.6	

Table 13.9: Confusion matrix for the IC4 independent-participant decoder on DevSet; format and symbols as in Table 13.5.

This indicates that more effort will need to be invested in discriminating between voiced laughter and silence. One possible solution is to include return loops within a bout topology which represent single calls and their subsequent inter-call intervals. This will increase the complexity of single-participant topologies, further putting them out of reach of use in multi-participant decoding.

Another solution is to experiment with longer frame sizes. The determined minimum duration constraints suggest that laughter may be safely treated with much coarser granularity than has been typically assumed for speech. It remains to be seen whether speech and laughter in conversation can be successfully decoded simultaneously.

Chapter 14

Text-Independent Dialog Act Recognition*

14.1 Introduction

The preceding four chapters of the current part of this thesis have dealt with the detection or inference of vocal behavior, in scenarios where each participant's "vocal behavior" may take on a limited number of values. Examples involving 2 classes have included *speech* $\mathcal{S}$ *versus non-speech* $\neg\mathcal{S}$ in Chapters 10 and 11 and *laughter* $\mathcal{L}$ *versus non-laughter* $\neg\mathcal{L}$ in Chapter 12; those involving 3 or 4 classes, namely *speech* $\mathcal{S}$ *versus laughter* $\mathcal{L}$ *versus other* $\mathcal{N} \equiv \neg(\mathcal{S} \cup \mathcal{L})$, *speech* $\mathcal{S}$ *versus voiced laughter* $\mathcal{L}_V$ *versus other* $\mathcal{N} \equiv \neg(\mathcal{S} \cup \mathcal{L}_V)$, and *speech* $\mathcal{S}$ *versus voiced laughter* $\mathcal{L}_V$ *versus unvoiced laughter* $\mathcal{L}_U$ *versus other* $\mathcal{N} \equiv \neg(\mathcal{S} \cup \mathcal{L}_V \cup \mathcal{L}_U)$, were discussed in Chapter 13. Recognizing these vocal activity types was shown to be tractable using a state-space modeling approach. However, when more than a handful of classes are of interest, state-space modeling in a space which grows geometrically with the number K of participants becomes prohibitive.

An example of an important problem involving arbitrarily fine subclassification of speech is the recognition of the *type* of speech, particularly if the different subtypes thus conceived have bearing on incipient interaction patterns among participants. In a general conversational scene understanding framework, it is of relevance whether a spoken contribution conveys propositional content and is information-bearing, whether it aims to take the floor from another participant or to acknowledge another participant's floor, or whether its aim is to hedge against potential floor takeover. These distinctions are to a large degree content-neutral, and as humans we frequently appear to be able to make them even for conversations conducted in a language we do not understand.

The text-independent DA recognition framework based on speech chronograms alone was first described in [153]. The prosodic feature vector used to augment that system was developed in [138, 139, 155, 91], and was evaluated on the same data as used here in [142, 143]. It was applied to the DA recognition framework in [154] and [131].

14.1.1 Dialog Acts

The formal distinction among speech types differing in tactical intent is captured by what have come to be known as dialog act (DA) ontologies. The segmentation of single-participant speech into temporally non-overlapping dialog acts and the classification of dialog acts into mutually exclusive types is known to be helpful for many downstream speech understanding applications [228, 119, 229]. Most important among such applications, for recordings of multi-party conversation, are summarization and indexing. These tasks analyze and distill the propositional content in speech, and are therefore heavily word- and often topic- centric. DA recognition systems used under these circumstances group DAs into approximately 5 types, including: (1) questions and (2) statements, broadly embodying the two sides of any interactive informational exchange; (3) backchannels, of no propositional value; (4) floor mechanisms, such as filled pauses used to retain the floor or grabbers to acquire it, also of no propositional value; and (5) disrupted forms, whose propositional content has not been delivered in its entirety and may therefore be of dubious value.

Although these 5 classes appear sufficient to describe propositional argument structure, they do not fully describe the interactive context, and in particular how speaker turns are initiated, continued, and terminated by the participants,

*The work in this chapter was conducted in collaboration with Liz Shriberg.

jointly. The broad category of floor mechanisms may thus contain different behaviors, whose placement in one group has been mainly motivated by absence of propositional content worth parsing (and the fact that they are not backchannels). This chapter proposes a finer subclassing of intent, which would allow inference of argument structure but also of turn structure, based on the ICSI Meeting Recorder Dialog Act (MRDA) [202] tagset. Three groups are considered, together comprising 8 types, which are described in the coding manual [54] (excerpts reproduced in quotes for completeness):

floor mechanisms utterances whose role is the management of one's own floor

- floor holders `fh` "A floor holder occurs mid-speech by a speaker who has the floor [...] and is used as a means to pause and continue holding the floor. [...] The duration of a floor holder is usually longer than that of the other words spoken by a speaker. Also, the energy of a floor holder is often similar to that of the surrounding speech by the same speaker. [...] [F]loor holders do not occur at the beginning of a speaker's turn, but rather occur throughout the middle and at the end of a speaker's turn." ([54], pp. 45)
- holds `h` "[A hold] is predominantly used when a speaker is responding to a question that he in particular was asked, and that speaker pauses or "holds off" prior to answering the question. [...] Holds are very similar to floor holders in the way that they sound, however holds occur at the beginning of a speaker's turn, as opposed to floor holders which occur in the middle or at the end of a speaker's turn." ([54], pp. 46)
- floor grabbers `fg` "Floor grabbers usually mark instances in which a speaker has not been speaking and wants to gain the floor so that he may commence speaking. [...] Most often, floor grabbers tend to occur at the beginning of a speaker's turn. In some cases, none of the speakers will have the floor, resulting in multiple speakers vying for the floor and consequently using floor grabbers to attain it. [...] Floor grabbers are generally louder than the surrounding speech." ([54], pp. 43)

feedback utterances whose role is the management of another's floor, a type of grounding

- backchannels `b` "Utterances which function as backchannels are not made by the speaker who has the floor. Instead, backchannels are utterances made in the background that simply indicate that a listener is following along or at least is yielding the illusion that he is paying attention. [...] [B]ackchannels are more often confused with acknowledgments and accepts than with floor grabbers, floor holders, and holds. One method in distinguishing if the [`b`], [`bk`], or [`aa`] tag is appropriate lies in the point at which the utterance occurs with regard to the speaker who has the floor's utterance. Acknowledgments generally appear after another speaker has completed a phrase or an utterace, [...]. Accepts usually occur at the end of another speaker's utterances, [...]. Backchannels, although they can occur in the same locations as acknowledgments and accepts, can also be found in the middle of another speaker's phrase." ([54], pp. 49)
- acknowledgments `bk` (MRDA: `s^bk`, a subclass of statements, below) "[Acknowledgments are] used to express a speaker's acknowledgment of a previous speaker's utterance or of a semantically significant portion of a previous speaker's utterance. [...] As opposed to backchannels, acknowledgments encode a level of direct communication between speakers." ([54], pp. 50)
- accepts `aa` (MRDA: `s^aa`, a subclass of statements, below) "[Accepts are] used for utterances which exhibit agreement to or acceptance of a previous speaker's question, proposal, or statement. [...] Generally, utterances marked with the [`aa`] tag have much more energy and are more assertive than backchannels and acknowledgments." ([54], pp. 57)

propositional content sentence-like units which bear propositional meaning

- statements `s`
- questions `q`, including yes/no questions, wh-questions, or-questions, and others.

This chapter also considers, for statements and questions, three alternative forms of DA termination:

completion the canonical form of DA termination;

interruption applied to incomplete utterances, which were interrupted by another speaker;

abandonment applied to incomplete utterances, which were abandoned by the speaker but not interrupted by someone else.

DAs other than statements and questions are only allowed to be terminated in the first of the ways listed above.

14.1.2 Text Independence

DA recognition, or the segmentation into dialog acts and their classification into types, has been shown to yield optimal performance using features characterizing words and word sequences. Since most existing applications of DA recognition aim to summarize propositional content, DA recognizers have always assumed that words are available. But the inference of turn structure and conversational interaction, in contrast to propositional content summarization, do not in principle require knowledge of word identity. It is therefore important to determine the extent to which DAs can be recognized in the absence of word information. Realistic scenarios which represent this case include: (1) privacy-sensitive settings, in which spectral features (required for automatic speech recognition) cannot be computed; (2) acoustically degraded environments, in which it is possible to hear in the recording that someone is speaking but not what they are saying; (3) acoustically uninstrumented environments, in which it is possible to see or otherwise detect that someone is speaking, but not hear that they're doing so; (4) resource-poor or high-variability domains, in which insufficient labeled data exists to train reliable speech and dialog act recognition systems; (5) resource-poor dialects or languages, characterized by similar problems; and (6) constrained real-time systems in which a speech recognition system cannot or should not be universally applied to all encountered vocal productions (in order to seize tactical opportunity before it passes). It should be noted, particularly with respect to the last item, that for the 6 non-propositional-content DAs considered here word identity is not very relevant once the DA type is known. Provided that a downstream system knows that a DA is of one of these 6 types, it need never parse it.

This chapter explores the *text-independent* dialog act recognition problem, given that only speech activity detection results are available; experiments rely on reference speech activity, as opposed to the output of an automatic SAD system such as in Chapter 11, in order to avoid being tied to any particular speech detection technology. To recognize DAs in this way, the proposed approach models interlocutor speech activity in feature space, as opposed to in state space, using the modeling techniques of Chapter 8. The resulting system is assessed and compared to a prosodic system as well as to an "oracle" lexical system which uses true words, thereby providing an approximate upper bound on expected performance. The results indicate that the feature-space models discussed are viable, and that together with prosodic information the detection of many DA types and DA boundary types approaches or even exceeds that achieved with an oracle lexical system.

14.2 Related Work

Related to text-independent DA recognition in multi-party conversation, but relying on words, is of course *text-dependent* DA recognition in multi-party conversation. As mentioned above, most if not all DA recognition systems in the literature rely on words.

The DA recognition problem is often treated as a two-stage process, the first of which consists of segmenting a word-stream into DAs by inferring the presence of DA boundaries at a subset of all word boundaries. Once segmented, DA units are classified into types, with or without prior knowledge of the precedent and subsequent word or DA type. Word-synchronous HMMs are a frequently used paradigm for the segmentation as well as the recognition task [209].

Work on DA segmentation in multi-party meetings has shown that while true word sequence information is a strong predictor of DA termination, inter-word pauses offer better performance when automatically recognized words are used [4]. Pauses are a frequently employed feature in DA segmentation and recognition systems [55, 176, 231]; text-dependent work sometimes relies on modeling the pattern of inter-word pauses, such as pause duration after the current, previous, and following words [124]. Non-zero pause duration is of course available to a text-independent system, as proposed here. In contrast, word duration, also appearing in the literature, is likely to be accessible to the methods of this chapter only in very degraded form, since words are frequently temporally adjacent, without an intervening pause. Finally, cross-speaker

pauses, in addition to within-speaker pauses, have been modeled and were found to be useful for the segmentation task [49]. The general multi-participant DA recognition problem has been explored for two-party dialogue in [192].

In DA classification, prosodic features, including within-DA pause gaps for both the target speaker and their interlocutors (as might be captured by the system proposed in this work), were shown [4] to reduce error rates by 5-10%rel. A similar improvement was observed when adding true precedent and subsequent DA type, from both the target speaker and temporally most proximate interlocutor. The duration of the pre-segmented DA, which may be available to the system proposed in this work when DAs are separated by long pauses, is also commonly modeled [4].

In HMM-based frameworks, however, segmentation and classification can be performed jointly, rather than sequentially; other frameworks have also been proposed [233, 232]. An important implicit feature of such systems is that a DA grammar, defining the inter-DA transition probabilities, is available. It is usually implemented as an N-gram model, with relatively high values for N given that DA tagsets are small [217].

14.3 Dataset Use

The data used in these experiments is the ICSI Meeting Corpus [108]. The corpus has been labeled with the Meeting Recorder Dialog Act (MRDA) annotation scheme [202, 223], which was developed with precisely this corpus in mind. General characteristics of the corpus are provided in Chapter 4.

The meetings in the corpus were divided into three portions, a TrainSet of 51 meetings[1], a DevSet of 11 meetings[2], and a EvalSet also of 11 meetings[3]. This split has been used in previous work [4] on dialog act segmentation and classification in this corpus, and is therefore retained in this work to enable at least qualitative comparison. As for any longitudinal collection of meeting groups, TrainSet, DevSet, and EvalSet contain some of the same participants.

Phenomenon	TrainSet	DevSet	EvalSet
`fh`, floor holder	2.32	2.29	3.00
`h`, hold	0.21	0.36	0.26
`fg`, floor grabber	0.55	0.58	0.62
`b`, backchannel	2.86	2.65	2.83
`bk`, acknowledgment	1.42	1.42	1.48
`aa`, accept	1.18	1.13	1.10
`s`, statement	84.93	84.11	82.79
`q`, question	6.53	7.45	7.92
completed	5.18	5.06	5.45
interrupted	0.31	0.31	0.32
abandoned	0.39	0.43	0.50

Table 14.1: The 8 DA types and 3 DA boundary types of interest , and their prior probability distribution (in %) in the three datasets used in this chapter. DA boundary priors are a strong function of the frame size; they indicate the prior probability of the proportion of *frames* containing a boundary (here, frames are spaced 100 ms apart).

All three datasets exhibit very skewed prior distributions over the 8 DA types of interest[4] . The actual numbers, once the datasets are discretized using the 100 ms frame step used throughout this thesis, are shown in Table 14.1. The 100 ms frame size has only negligible impact on the precise accuracy of these numbers; for DA boundary types, however, also shown, the absolute proportion of frames containing punctuation is more sensitive to variability in frame size.

[1] TrainSet consists of: `Bdb001`, `Bed002`, `Bed004`, `Bed005`, `Bed008`, `Bed009`, `Bed011`, `Bed013`, `Bed014`, `Bed015`, `Bed017`, `Bmr002`, `Bmr003`, `Bmr006`, `Bmr007`, `Bmr008`, `Bmr009`, `Bmr011`, `Bmr012`, `Bmr015`, `Bmr016`, `Bmr020`, `Bmr021`, `Bmr023`, `Bmr025`, `Bmr026`, `Bmr027`, `Bmr029`, `Bmr031`, `Bns001`, `Bns002`, `Bns003`, `Bro003`, `Bro005`, `Bro007`, `Bro010`, `Bro012`, `Bro013`, `Bro015`, `Bro016`, `Bro017`, `Bro019`, `Bro022`, `Bro023`, `Bro025`, `Bro026`, `Bro028`, `Bsr001`, `Btr001`, `Btr002`, and `Buw001`.

[2] DevSet consists of: `Bed003`, `Bed010`, `Bmr005`, `Bmr014`, `Bmr019`, `Bmr024`, `Bmr030`, `Bro004`, `Bro011`, `Bro018`, and `Bro024`.

[3] EvalSet consists of: `Bed006`, `Bed012`, `Bed016`, `Bmr001`, `Bmr010`, `Bmr022`, `Bmr028`, `Bro008`, `Bro014`, `Bro021`, and `Bro027`.

[4] In [153], the author mistakenly referred to `aa` as "assert" instead of "accept".

14.4 Assessment of Performance

The aim of this chapter is to explore the performance of techniques to "color" the black regions of a black and white chronogram, **Q** (which is also the **Q** of Chapters 10, 11, 12, and 13), with colors corresponding to licensed DA and DA boundary combinations. White non-speech intervals are not processed[5], and are therefore excluded from the proposed scoring; a system relying on reference or already detected speech activity cannot get these regions wrong. Since on average participants are more often silent than they are speaking in multi-party ($K > 2$) conversation, including always-correct regions in scoring would consistently but uninformatively bias performance assessment.

The heavily skewed prior, towards statements (`s`) in the dataset used, also suggests that classification accuracy is not an optimal measure, since always guessing `s` would result in accuracies of approximately 85%. Doing so makes no distinction whatsoever among the more rare, and pragmatically potentially more interesting, dialog act types.

The experiments in this chapter are therefore scored using F-score, computed individually for each dialog act type, by time (or, equivalently, by frame of speech). The DA-specific F-scores are then averaged to yield the *mean (8-class) F-score*. A system always guessing the majority class would achieve a mean F-score of 11.5%. This metric, relative to classification accuray, encourages systems to agressively "peel" off speech frames from the majority `s` class into the 7 other DA classes, while retaining as much separation among the latter as possible.

It should be noted that the proposed metric does not pay attention to DA boundary detection. To some extent, this is mitigated by the fact that boundaries are implicitly posited every time a system detects a different DA type than the one accounting for previously produced speech. As a result, a DA recognizer scored by correctly classified time is expected to also correctly hypothesize some boundaries.

To complement the mean F-score metric, experiments are additionally characterized using: (1) F-scores for individual DA types; (2) absolute classification error; (3) F-scores for the three DA boundary types individually; (4) a mean F-score over the three DA boundary types; (5) an F-score for DA boundaries when all three types are collapsed into a binary event; and (6) the NIST SU boundary metric evaluated on the collapsed DA boundary event. The latter metric in particular makes it possible to compare the proposed systems with published work; however, it should be noted that other work on DA segmentation uses knowledge of word boundaries, and thereby entertains far fewer locations for boundaries than does the proposed text-independent approach of this chapter. Segmentation performance, here, can therefore be expected to be worse than elsewhere due to higher false alarm rates.

Where F-scores are computed for individual conditions, statistical significance for F-score differences observed for EvalSet is assessed using the approximate randomization test[6]. For these tests, time is striated into intervals delineated by true DA boundaries. This is a much more stringent condition than merely striating time into frames, since adjacent frames of speech are not independent.

14.5 Baseline

This task, in its text-independent form, has not been previously attempted. As a result, no prior baseline exists. As argued in the preceding section, segmenting participant speech into 100-ms frames, and, independently for each frame, always guessing the majority `s` class yields a DevSet mean F-score of 11.5%. For completeness, a better-performing alternative baseline is proposed here which relies on the preliminary set of duration-to-landmark context features described in Section 8.3.

To recap, the duration-to-landmark features consider temporal distance to nearby speech from both the target speaker and their interlocutors. For the target speaker, 4 features are computed: the number of frames to the nearest previous speech frame, the number of frames to the nearest next speech frame, and similarly for the nearest non-speech frames. Also computed are 3 interlocutor features, namely the number of frames to the nearest previous and nearest next speech frame from *any* non-target participants, as well as the number of concurrent non-target speakers at time t.

A decision tree is trained over the resulting feature vectors using TrainSet. The performance on DevSet for the 4 target participant features alone is 16.51%; adding the 3 non-target participant features improves performance to 18.51%.

[5] □ frames are in fact implicitly classified as either intra- or inter-DA, but this information is also available in the presence of DA termination in the most recent ■ speech frame.

[6] All randomization tests are conducted using Sebastian Pado's SIGF implementation, which can be found at `http://www.nlpado.de/~sebastian/sigf.html`.

This baseline has two main limitations, namely that: (1) frames are assumed conditionally independent, as the decision tree does not leverage sequence information, and (2) the features capture only distance to the nearest edge, and not what lies beyond it.

14.6 A Single-Participant HMM Topology

The observable available to the proposed system is a multi-participant speech/non-speech (■/□) chronogram **Q**. Contiguous intervals of speech ■ may of course implement multiple consecutive DAs. When word boundaries are available, DA boundaries need only be entertained at those instants. However, in the absence of word boundary information, there is no a priori evidence for positing talkspurt-internal DA boundaries. The approach proposed here is to entertain DA boundaries every 100 ms, the duration of the shortest DAs in the ICSI corpus.

Additionally, DAs can contain intra-DA intervals of non-speech □. As a result, the proposed system must correctly split and merge talkspurts into DAs. This operation is illustrated in Figure 14.1; for clarity, the term *talkspurt fragment* (TSF) is introduced to denote a contiguous interval of speech belonging to exactly one talkspurt and to exactly one DA. Boundaries between immediately adjacent TSFs, without intervening non-speech gaps, must be DA boundaries.

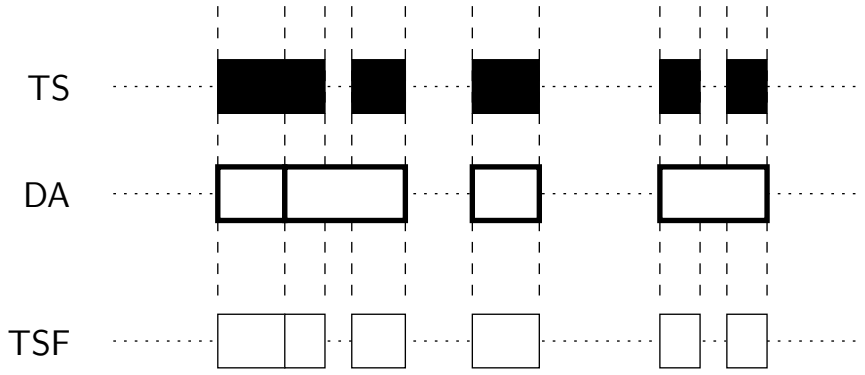


Figure 14.1: Construction of the talkspurt fragment (TSF) unit, given reference talkspurt (TS) and dialog act (DA) segmentation.

14.6.1 Basic Requirements

An interval of conversational speech of one person consists of a sequence of n DAs, separated (and terminated on both sides) by $n + 1$ inter-DA (non-speech) gaps; the latter may be of zero duration but may also be quite long (while other participants speak). Every DA is of one of the 8 types mentioned in Section 14.1. Each DA is terminated by a TSF which is DA-terminal; the termination, for `q` and `s`, may be of one of three types listed in Section 14.1. DAs are allowed to optionally consist of m non-DA-terminal TSFs, separated from each other and from the final DA-terminal TSF by m intra-DA (non-speech) gaps. These nested relationships are shown in Figure 14.2.

14.6.2 Proposed Implementation

The HMM topology proposed for labeling the intervals of conversation depicted as leaves in Figure 14.2, consists of individual subtopologies implementing each leaf. The multiplicity of these subtopologies, in a complete topology of conversational speech, is shown in Figure 14.3. Each of the 8 DA types is implemented using a separate DaTypeSubtopology; any two of these subtopologies are separated by one of 64 unique InterDaGapTypeSubtopologys. A DaTypeSubtopology contains one NonDatTsfSubtopology for non-DA-terminatal TSFs[7], one IntraDaGapSubtopology, and up to three distinct DatTsfType-

[7]Although non-DA-terminal TSFs could be further subclassed into {DA-initial, DA-next-initial, ⋯ , DA-penultimate} TSFs, these distinctions have been collapsed for simplicity.

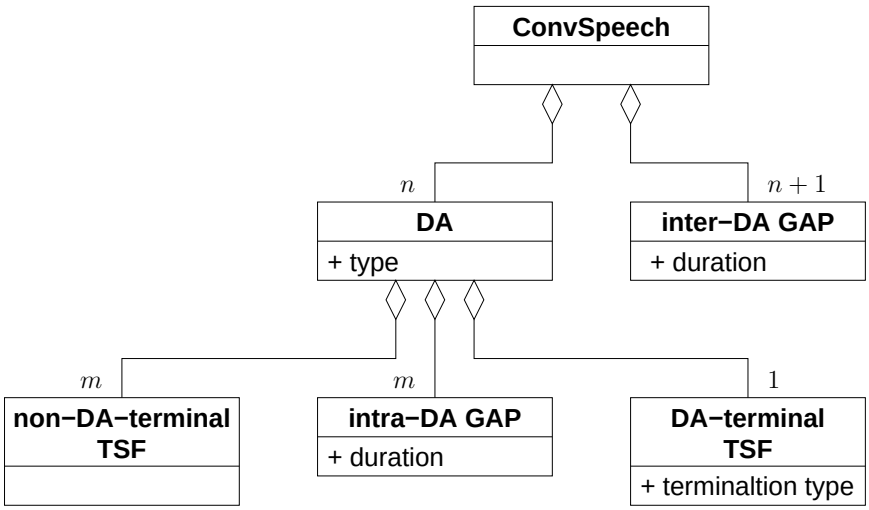


Figure 14.2: UML class diagram of relationships among the temporal units DAs, GAPs, and TSFs. "ConvSpeech" represents "conversational speech".

Subtopologys. Each DatTsfTypeSubtopology corresponds to one of completed DAs, interrupted DAs, or abandoned DAs. However, only question (q) and statement (s) DAs can terminate in all three of these ways; other DA types possess a single termination type.

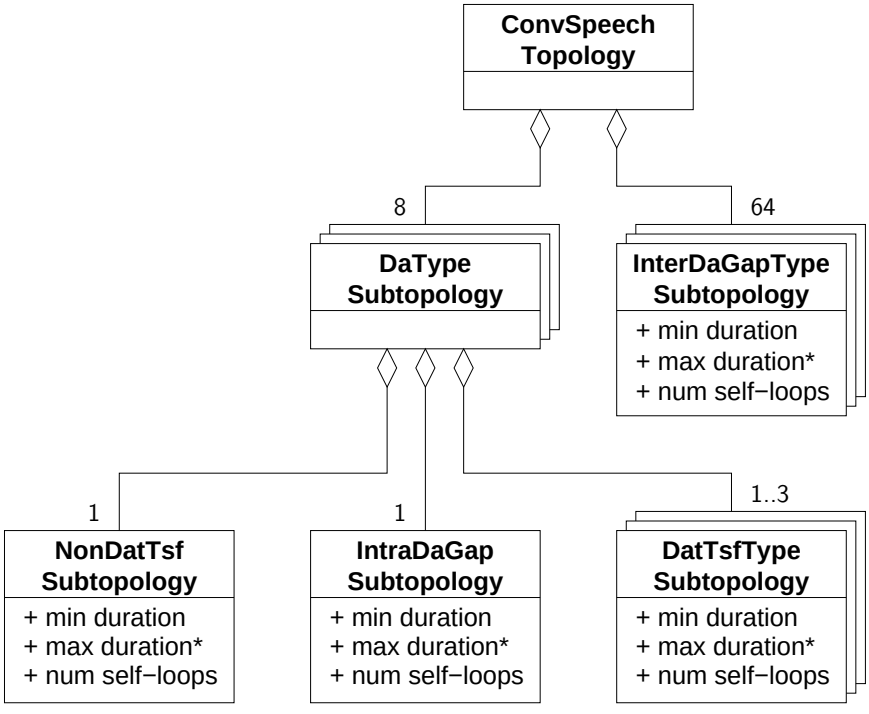


Figure 14.3: UML class diagram of relationships among the subtopologies implementing the temporal units in Figure 14.2.

14.6.3 Internal Subtopology Design

Subtopologies for each leaf in Figure 14.3 are implemented as shown in Figures 14.4. Non-DA-terminal TSF subtopologies and DA-terminal TSF subtopologies are identical except that egress states in the latter are punctuation-bearing. Inter-DA GAP subtopologies and intra-DA GAP subtopologies are identical except that inter-DA GAP subtopologies may be zero frames in duration.

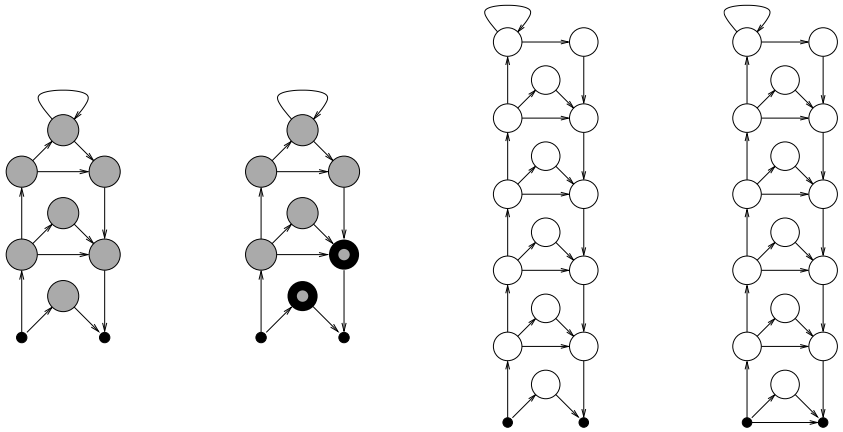


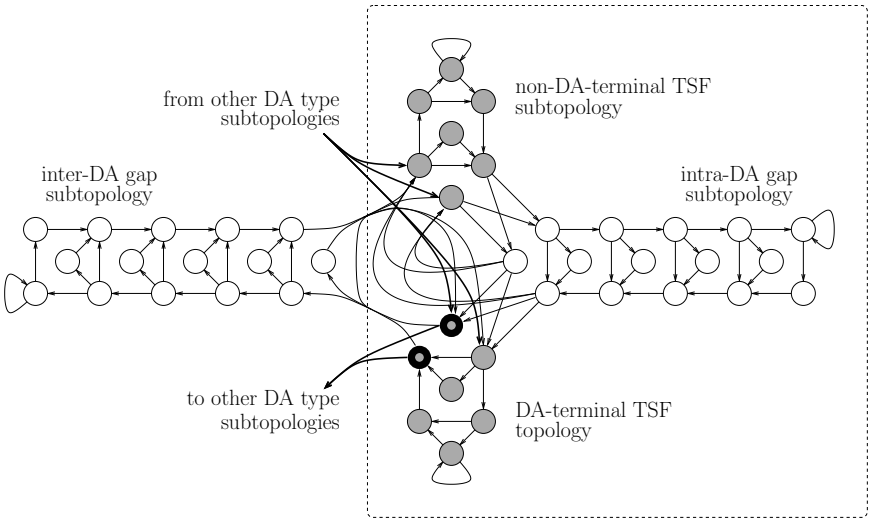
Figure 14.4: HMM subtopologies: (a) NonDatTsfSubtopology; (b) DatTsfSubtopology; (c) IntraDaGapSubtopology; and (d) InterDaGapSubtopology. Speech states shown in gray, non-speech states shown in white; states which bear punctuation are identified in black.

The same DaSubtopology structure type implements each of the 8 DA types of interest. Connectivity among the TSF and GAP subtopologies, as prescribed by Figure 14.3, is shown in Figure 14.5(a). For clarity, each type of transition among the TSF and GAP types is depicted separately in Figures 14.5(b) through (i). A DA subtopology can be entered by entering directly into its DA-terminal TSF subtopology (b), for those DAs consisting of only one TSF. A DA consisting of two or more TSFs must be entered through the non-DA-terminal TSF subtopology (c), and then must enter its intra-DA GAP subtopology (d). For a DA consisting of more than two TSFs, egress out of the intra-DA GAP subtopology is immediately followed by re-entry into the non-DA-terminal TSF subtopology (e); otherwise, the DA-terminal TSF subtopology is entered (f). Upon termination of the DA, control passes out of the DA subtopology entirely. It either passes on to inter-DA GAP subtopologies connecting the terminated DA type to other DA types (g), or returns to the terminated DA type in case of successive same-type DAs. For the latter, it may re-enter the current DA type subtopology either via the non-DA-terminal TSF subtopology (h) or via the DA-terminal TSF subtopology (i), depending on the number of TSFs in the current DA, as in (b) and (c), respectively.

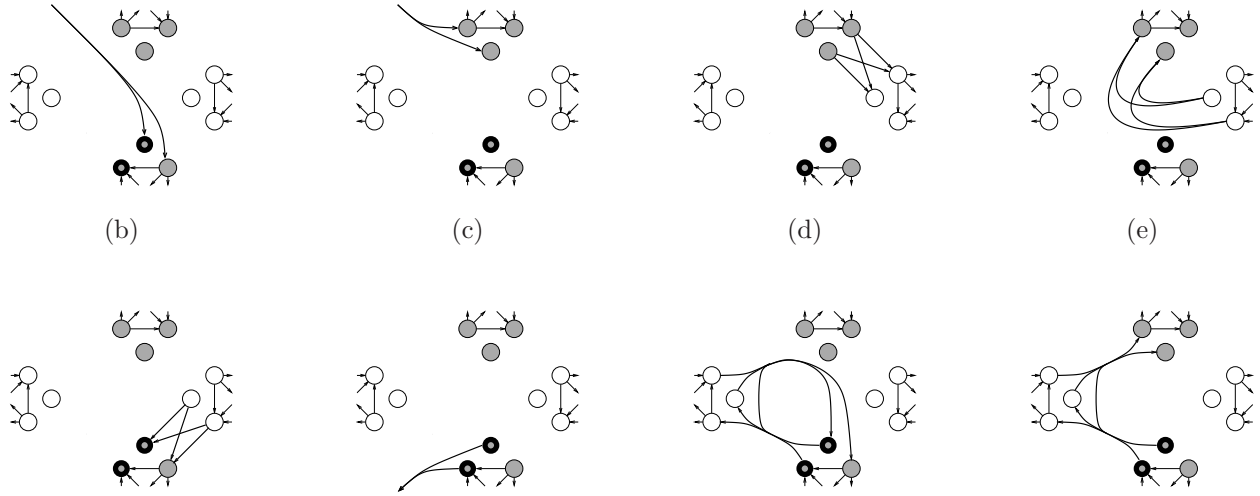
The subtopologies as shown in Figure 14.4 implement the complete topology in the proposed DA recognition system exactly, namely each TSF subtopology has a minimum duration of one 100 ms frame, a maximum unique-state duration of 5 frames, and may be arbitrarily long owing to the self-loop positioned at the state furthest away from either edge of the TSF. Similarly, each GAP subtopology in the system has a minimum duration of 1 frame (or zero for inter-DA GAP subtopologies), a maximum unique-state duration of 10 frames, and may also be arbitrarily long.

14.6.4 Transition Probability Modeling

Bigram transition probabilities from a state to any other state, subject to connectivity constraints, exhibit a categorical, or multinomial, distribution. Because the total number of degrees of freedom per participant is 1220 in the proposed single-participant topology, only independent-participant models are tractable in the general conversational case of arbitrary K. Transition probabilities are learned using maximum likelihood estimation as described in Equation 6.29, following forced Viterbi alignment to the reference segmentation of TrainSet.



(a) DA subtopology (inside the dashed region)



(b) (c) (d) (e)

(f) (g) (h) (i)

Figure 14.5: The HMM subtopology for an entire dialog act (DaSubtopology) is shown in (e); detailed connectivity among TSF type subtopologies and GAP type subtopologies are depicted in (b) through (i); explanation as in the text.

14.6.5 Topology Performance

To assess the performance of the topology alone, whose transition probabilities to some extent encode both DA-specific duration preferences and inter-DA sequencing preferences found in TrainSet, the following minimal emission probability model is proposed:

$$P\left(\mathbf{q}_t[k] \mid \mathbf{y}_t[k]\right) = \begin{cases} 1 & \text{if } \mathbf{q}_t[k] = \square \text{ and } \mathbf{y}_t[k] = \square \\ 1 & \text{if } \mathbf{q}_t[k] = \blacksquare \text{ and } \mathbf{y}_t[k] \neq \square \\ \epsilon & \text{otherwise} \end{cases} \tag{14.1}$$

where ϵ is a tiny number. In effect, this model aligns non-speech in the black-and-white chronogram $\mathbf{Q}$ to non-speech in the "colored", DA-augmented chronogram $\mathbf{Y}$, but allows the alignment of speech in $\mathbf{Q}$ to any DA type without preference.

On DevSet, a system constructed in this way, effectively without an informative emission probability model, achieves a mean F-score of 20.6%. This system will henceforth be referred to as "Topo only".

14.7 Modeling Interlocutors

14.7.1 Duration-to-Landmark Features

The baseline proposed in Section 14.5 does not explicitly account for sequence, relying only on features describing duration to speech activity landmarks. In contrast, the "Topo only" system models only state transition probabilities. To combine these systems, a generative model of the baseline features was defined to consist of a state-specific Gaussian mixture model (GMM). The duration features were transformed into the log-domain. The GMMs were trained with the standard expectation-maximization algorithm using TrainSet, with complexity and linear weight with respect to the transition probability model selected such as to maximize mean F-score on DevSet.

On DevSet, this combined system featuring the topology of Section 14.6 and the baseline duration feature set achieved a mean F-score of 21.93% when only target-participant features were used, and 24.06% when interlocutor features were also included.

14.7.2 Neighborhood Snapshot Features

An alternative to the simple duration features of the preceeding subsection are the neighborhood snapshot features of Section 8.4, which more comprehensively represent the local speech activity context. Their computation was summarized in Algorithm 4, and requires the prior specification of: (1) a window for feature extraction $\mathbf{W}_t$; (2) a window for interlocutor ranking $\mathbf{W}_t^{rank}$; (3) a neighborhood tiling policy $\mathcal{T}$; and (4) K_{max} the number of interlocutors represented in the neightborhood snapshot. The values of these parameters are chosen to maximize mean F-score on DevSet.

14.7.3 Optimizing Tiling Policy

To optimize tiling policy, $\mathbf{W}_t^{rank}$ is equated with $\mathbf{W}_t$, and both neighborhood selection operators are defined by a rectangular window (shown in Figure 14.6), centered on the instant t and of width $\Delta T = 10$ seconds. K_{max} is assigned the value 2, resulting in neighborhood snapshots consisting of the local speech activity distribution of the target participant and the two locally most talkative interlocutors.

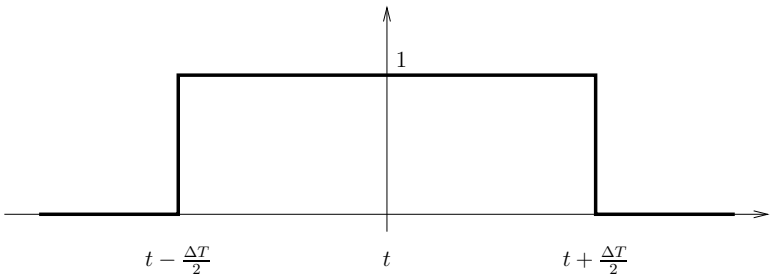


Figure 14.6: A simple definition of the window defining neighborhood selection operators $\mathbf{W}_t$ and $\mathbf{W}_t^{rank}$. ΔT is the width of the neighborhood.

Figure 14.7 shows DevSet mean F-scores as a function of the transition log-probability model weight, for tiles of fixed width $\in \{0.2, 0.3, 0.5, 1.0\}$ seconds (except at t, where tiles are 0.1 seconds wide). Given $\Delta T = 10$ seconds, $K_{max} = 2$ interlocutors, and a frame step of 100 ms, these widths entail feature vectors of 152, 104, 62, and 32 elements, respectively. To produce Figure 14.7, a global LDA transform was trained using TrainSet, with topology states as labels, to rotate these

vectors and to reduce their dimensionality to 24 elements. Each state was subsequently modeled using a 64-component GMM (also estimated using TrainSet). The results indicate that an optimal tiling policy under these fixed complexity conditions consists of tiles of 0.5 seconds in duration.

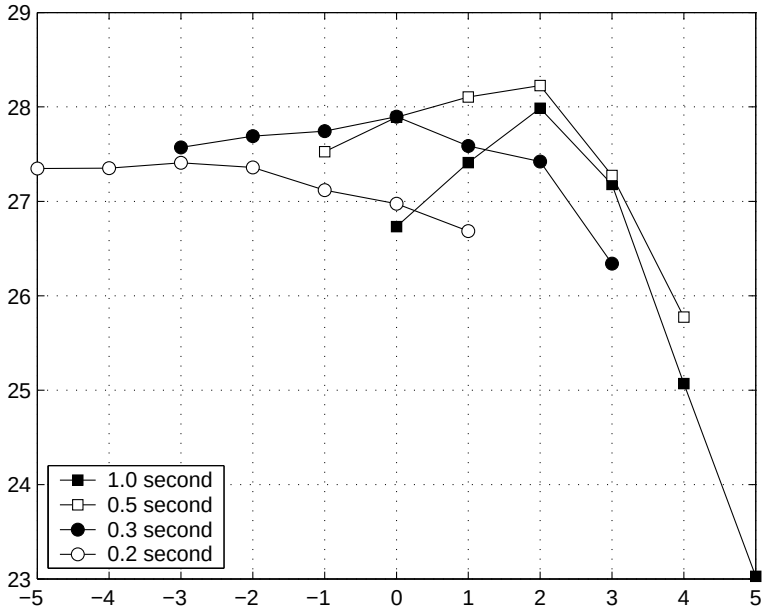


Figure 14.7: DevSet mean F-scores (along y-axis), as a function of transition log-probability model weight (along x-axis), for four systems differing in the size of tile used to tesselate the neighborhood snapshot.

14.7.4 Optimizing Windowing Policy

Assuming that tiles of 0.5 seconds in duration are unconditionally optimal, subsequent experiments aim to jointly optimize K_{max} and $\mathbf{W}_t$, the number of interlocutors and the temporal extent of the neighborhood snapshot, respectively. These values are selected to also maximize the DevSet mean F-score. This optimization was performed by exhaustively sampling a grid, defined by $K_{max} \in \{0, 1, 2, 3\}$ interlocutors[8] and $\Delta T \in \{1, 2, 5, 10, 20, 40\}$ seconds. For each resulting system, the number of GMM mixtures was 64, but the number of LDA dimensions was individually optimized to maximize the mean F-score on DevSet. The results are shown in Figure 14.8.

As can be seen in the figure, performance for $K_{max} = 0$, i.e., when no interlocutors are modeled, increases sharply as the neighborhood snapshot grows to approximately 15 seconds in duration; the same is true when any non-zero number $K_{max} > 0$ of interlocutors is modeled as the neighborhood snapshot grows to approximately 10 seconds. Although the curve for $K_{max} = 0$ continues rising past 15 seconds, and may even outpace the $K_{max} > 0$ curves for much larger values of ΔT, its value at the obverved maximum of 40 seconds is still lower than that for any $K_{max} > 0$ system at only 5 seconds.

A second observation is that differences among the $K_{max} > 0$ systems are quite small, relative to their differences with $K_{max} = 0$. This indicates that modeling more than just the single locally most talkative non-target participant offers little additional gain. While the best observed performance is achieved with $K_{max} = 2$ interlocutors at $\Delta T = 20$ seconds, both the $K_{max} = 1$ and $K_{max} = 2$ curves exhibit what appears to be noise past $\Delta T = 5$ seconds. In contrast, the $K_{max} = 3$ curve is much smoother. However, all $K_{max} > 0$ curves begin dipping at $\Delta T = 40$ seconds. This, and the occasionaly noisy behavior exhibited by some curves, may be due the rectangular profile of $\mathbf{W}_t^{rank}$ used in interlocutor ranking; it is

[8] In [153], feature vectors corresponding to $K_{max} = 0$, $K_{max} = 1$, $K_{max} = 2$, and $K_{max} = 3$ were referred to as Target, Target+Other1 ≡ VocInt1, Target + Other2 ≡ VocInt2, and Target + Other3 ≡ VocInt3, respectively; it should be noted that Other1 ⊃ Other2 ⊃ Other3.

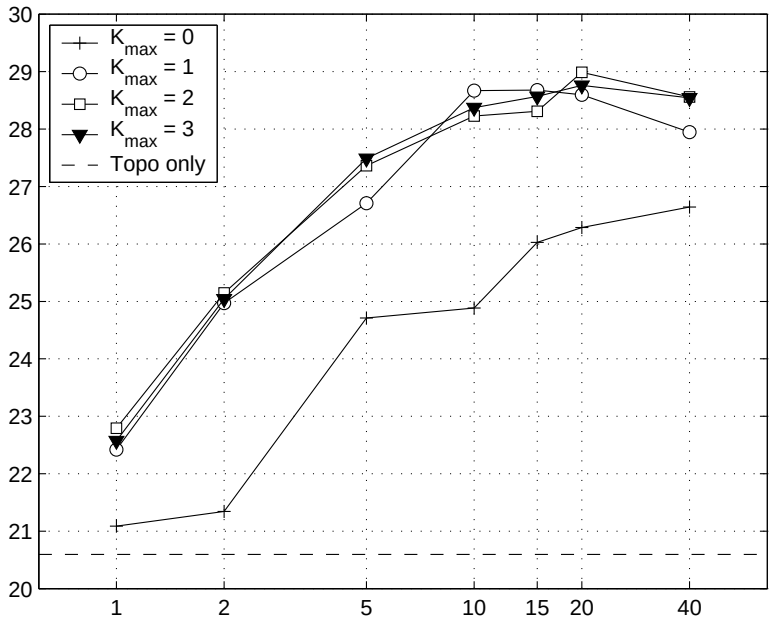


Figure 14.8: Average 8-class F-score (along the y-axis) for 5 systems, as a function of the neighborhood size ΔT, shown on a logarithmic scale in seconds along the x-axis.

possible that interlocutors who speak in close temporal proximity to the current instant t should be ranked higher than those who may speak more but also much sooner or later.

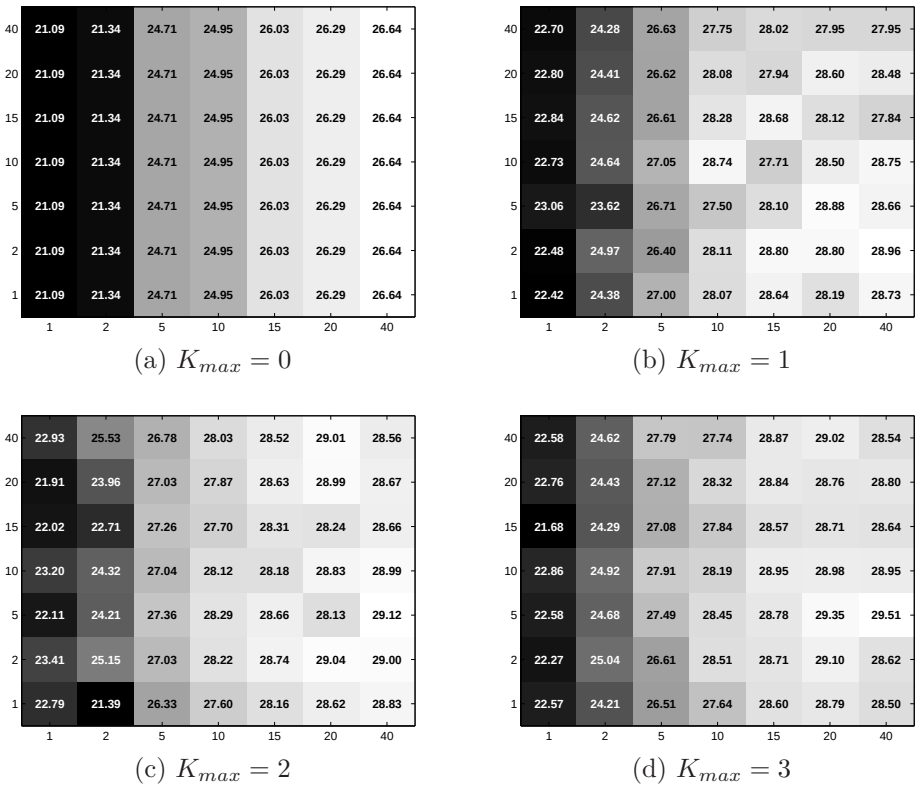
Systems selected for subsequent experiments will retain $\Delta T = 10$ seconds, where the $K_{max} > 0$ curves all begin to flatten out on the logarithmic scale in Figure 14.8. At this operating point, the DevSet mean F-score is 24.88% for $K_{max} = 0$ and 28.37% for $K_{max} = 3$, that K_{max} which yields the smoothest curve. This represents absolute improvements of 2.95% and 4.31% with respect to the combination of the duration-to-landmark features of Subsection 14.7.1 and the HMM topology.

14.8 Locally Versus Globally Most Talkative

Up to this point, the windows controlling feature extraction and interlocutor ranking have been assumed identical, i.e., $\mathbf{W}_t = \mathbf{W}_t^{rank}$. To explore the effect of the size of the temporal context used for ranking independently of that used to model the neighborhood snapshot, a series of experiments are presented in which ΔT is drawn from $\{1, 2, 5, 10, 15, 20, 40\}$ seconds, separately for both windows. It has been suggested that the most impactful participants (on others) may be those who are globally the most talkative, rather than only locally the most talkative. The results of this inquiry are presented in Figure 14.9.

What Figure 14.9 demonstrates is that the temporal extent of the rectangular and symmetric window used for ranking interlocutors has only a secondary effect on mean F-score, which appears to be determined primarily by the temporal extent of the modeled context[9]. Second, and perhaps more importantly, the higher mean F-scores generally occur in the lower right-hand corner of each panel, there where the modeled context window $\mathbf{W}_t$ is large but the interlocutor ranking window $\mathbf{W}_t^{rank}$ is relatively small. This suggests that, although the globally most talkative interlocutors probably exhibit some effect on the behavior of participants, the specific DAs produced are in fact influenced most by only the *locally* most

[9]Results along the diagonals differ slightly from those in Figure 14.8, because a negligibly different topology was used during emission probability model training. These experiments, intended to be exploratory in nature, have not been duplicated with the original topology in that role.



(a) $K_{max} = 0$ (b) $K_{max} = 1$

(c) $K_{max} = 2$ (d) $K_{max} = 3$

Figure 14.9: DevSet mean F-scores for all $K_{max} \in \{0, 1, 2, 3\}$, with width of rectangular window used for feature extraction along the x-axis and width of rectangular window used for interlocutor ranking along the y-axis, both in seconds. Lighter shades of gray indicate higher mean F-scores. (Panel (a) shows no vertical variation, since rotation involves only interlocutors, and $K_{max} = 0$ interlocutors are being modeled in that panel.)

talkative interlocutor(s).

14.9 Analysis of Performance for Specific DA Types

To complement previous sections, Table 14.2 shows the performance of the proposed techniques for individual DA types, DA boundary types, and several overall metrics. Two types of systems are described. First, "g-Opt" represents the single system which maximizes the mean F-score over 8 DA types; performance on individual DAs or boundaries, in the "g-Opt" columns, is the performance of that single globally optimized system on a specific condition. Second, "c-Opt" represents the "g-Opt" system with one parameter, namely the interpolation weight of the emission log-probability with respect to the transition log-probability, re-optimized to maximize the F-score for that DA only. As such, the condition-optimized "c-Opt" system columns present the performance of many systems optimized for different criteria; the "mean F" row for these columns computes an average over multiple systems.

As the table shows, the mean F-score increases first when the speech activity for the entire neighborhood snapshot is modeled for the target speaker (on top of "Topo only"), and then further when interlocutors are modeled. The single system with the best mean F-score ("g-Opt" for $K_{max} = 3$), yields 28.4%abs, or an improvement of 7.8%abs over the "Topo only" system; this amounts to a relative increase of 38%rel. The absolute improvement consists of a 4.3%abs increase due to modeling the target participant's speech activity, and a 3.5%abs increase due to modeling interlocutors.

It can also be seen that modeling the target participant's neighborhood snapshot leads to higher "g-Opt" F-scores for all DAs and DA boundaries except for backchannels `b` and statements `s`. These are among the most frequently

	Topo only	$K_{max}=0$		$K_{max}=3$		Δ, for c-Opt $K_{max}=3$	
		g-Opt	c-Opt	g-Opt	c-Opt		
	%	%	%	%	%	%abs	%rel
DA Types							
mean F	20.6	24.9	27.2	28.4	30.3	+9.7	+47
F, `fh`	9.4	17.9	17.9	19.0	20.6	+11.2	+119
F, `h`	0.0	0.7	1.9	8.2	8.3	+8.3	$+\infty$
F, `fg`	0.0	9.6	9.6	12.2	12.5	+12.5	$+\infty$
F, `b`	52.3	42.9	52.3	54.0	54.5	+2.2	+4
F, `bk`	2.7	10.2	10.2	15.6	15.9	+13.2	+489
F, `aa`	2.0	10.3	10.4	11.2	12.9	+10.9	+545
F, `s`	91.9	84.8	91.9	82.8	91.9	0	0
F, `q`	6.5	22.5	23.2	24.0	25.7	+19.2	+295
CER	16.3	28.3	16.4	30.4	16.2	−0.1	−1
DA Termination Types							
F, compl.	52.5	58.7	60.4	57.9	61.6	+9.1	+17
F, interr.	0.0	10.9	12.1	24.8	25.2	+25.2	$+\infty$
F, aband.	0.0	3.9	6.1	6.5	6.5	+6.5	$+\infty$
F any	53.8	66.1	66.1	65.4	67.2	+13.4	+25
NIST	65.1	61.8	58.0	68.3	56.8	−8.3	−13

Table 14.2: DevSet performance on individual DA and DA boundary types, for the "Topo only" system, and for systems modeling the neighborhood snapshot with both none or three interlocutors modeled. The two systems modeling the neighborhood snapshopt have been tuned to maximize the mean 8-class F-score, yielding the globally optimized ("g-Opt") variant in the table, but then have had the emission probability model weight re-tuned to optimize F-scores (and other metrics; appearing as rows) for individual DA and DA boundary types, yielding the condition-optimized "c-Opt" variants. Also shown is the delta ("Δ") incurred by adding the ■/□ with $K_{max}=3$ to the "Topo only" system.

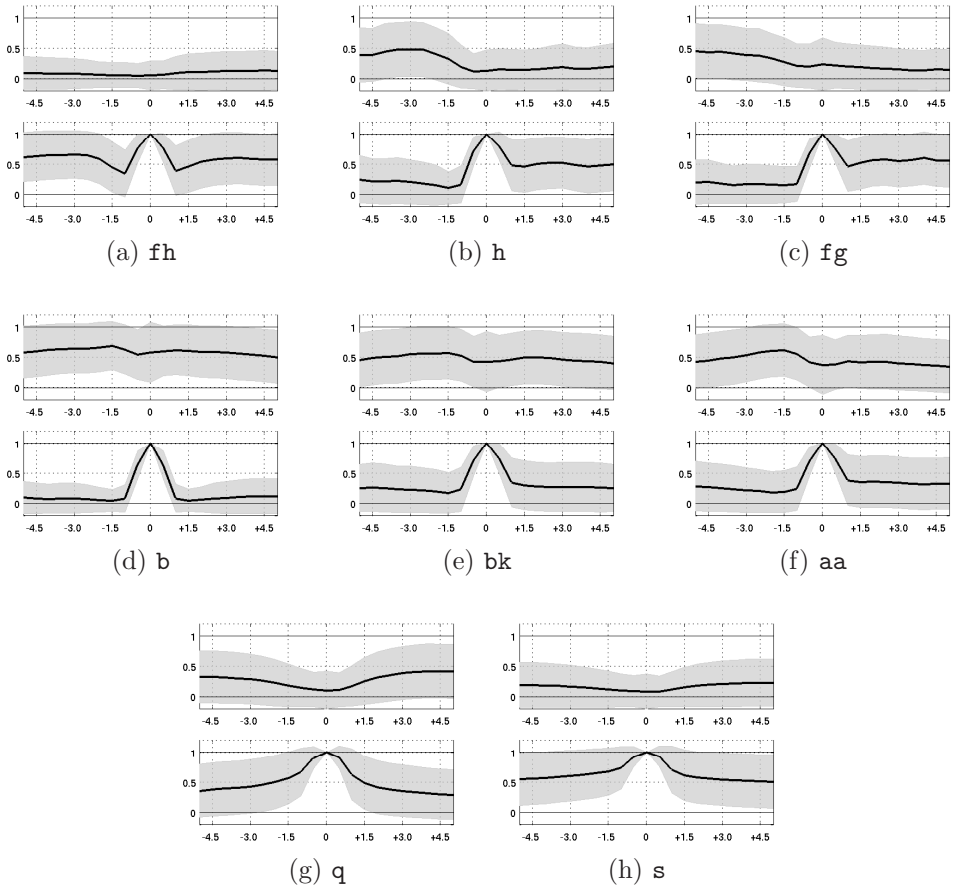
occurring DAs, and, correspondingly, the "g-Opt" classification error CER rises from 16.3%abs to 28.3%abs. Additionally modeling non-target participants leads to "g-Opt" improvement for all DAs and DA boundary F-scores except `s` and DA completion. For backchannels, the "g-Opt" system with $K_{max}=3$ outperforms the "Topo only" system, even though modeling target-participant speech activity appears to be deleterious for that DA type. The F-score for DA completion falls by 0.8%abs, but continues to be higher than the "Topo only" F-score by 5.4%abs.

Table 14.2 also shows that it is generally possible to improve performance for specific DAs and DA boundaries by re-optimizing the emission log-probability weight. In the case of statements `s`, the F-score never exceeds the value attained by the "Topo only" system, even though CER (which is sensitive to frequency of occurrence and duration), yields a drop of 0.1%abs or 1%rel. In all other cases, re-optimizing the weight leads to improved F-scores for the $K_{max}=3$ systems relative to "Topo only". Particularly large improvements in absolute terms are observed for `fh`, `fg`, `bk`, `aa`, and `q`. The largest improvement is seen in the detection of interrupted DAs, as can be expected because interruption is inherently a multi-participant event.

It should be noted that in the lower panel, relevant to DA termination types, the "g-Opt" condition is not very informative since the mean F-score metric does not explicitly penalize segmentation error, and it is the mean F-score metric that the "g-Opt" condition maximizes. For termination-related conditions, "c-Opt" performance is more informative, and improvement is consistently observed from left to right.

In summary, modeling the neighborhood snapshot brings gains in the detection of all but one DA type and all DA boundary types. To shed light on why this is so, Figure 14.10 shows what models actually learn. To make the presentation more clear, feature vectors are not rotated via LDA, and only a single-Gaussian model is trained. For brevity, the figure depicts models for the internal state (the one with the self-loop in panel (b) of Figure 14.4) of the DA-terminal talkspurt

fragment for each DA. That state is maximally distant from non-speech.



(a) `fh` (b) `h` (c) `fg`

(d) `b` (e) `bk` (f) `aa`

(g) `q` (h) `s`

Figure 14.10: The probability of speaking (along the y-axes) in the 10-second context (along the x-axes) centered on the self-loop state of the DA-terminal TSF subtopologies of 8 dialog act types. In each of the 8 panels, the bottom plot refers to the target speaker who is producing the DA in question; the top plot refers to the target speaker's locally most talkative interlocutor. The black line indicates mean values, the gray area the first standard deviation around the mean, under the expedient assumption of gaussianity.

Each panel in the figure consists of two plots; the upper plot corresponds to the single most-talkative interlocutor, while the lower plot corresponds to the target speaker. As can be seen in the first panel, floor holders occur in the middle of the target participant's speech. Just 1.5 seconds before and 1.5 seconds after the center of the TSF, the target participant is more likely to be speaking than not. The most talkative interlocutor is very unlikely to be speaking, in the entire 10-second window shown. In contrast, hold and floor grabber DA-terminal TSFs are typically followed by target participant speech, but not preceded by it; the locally most talkative interlocutor exhibits the reverse trend. For holds, interlocutors appear more likely to have spoken more recently than for floor grabbers, suggesting that floor grabbers frequently occur following all-participant silence. Both holds and floor grabbers appear to be turn-initial, as expected.

Panels (d) through (f) in Figure 14.10 depict behavior for feedback DA types. The relatively small variance for backchannels in panel (d) suggests that DA-terminal backchannel TSFs are consistently the shortest among the 8 DA types, with fewest self-loop state repetitions. The target participant does not appear likely to have been speaking in the 5 seconds preceding the center of a `b` production or to speak again in the subsequent 5 seconds. The most locally talkative interlocutor is more likely to be speaking during the entire window than not to be. This suggests that foreground speakers do not consistently pause during the production of backchannels from interlocutors. In contrast to backchannels, during

both acknowledgments and accepts, in panels (e) and (f), respectively, the locally most talkative interlocutor is less likely to be speaking than not during the center of the target participant's DA-terminal TSF. Both of these DA types seem much more likely to be followed by more target participant speech than to be preceded by it. This makes at least some DAs of these two types turn-initial.

The final two panels in the figure depict models for `q` and `s`. The main difference between these two DA types is that while `s` appears nearly as likely to be followed by target participant speech as to be preceded by it, questions appear to be more frequently turn-final; the probability that the target participant speaks 5 seconds after the `q` center is approximately 10% lower than than he or she speaks 5 seconds earlier.

14.10 Generalization to Unseen Data

Results for EvalSet, like those in Table 14.2 for DevSet, are presented in Table 14.3.

	Topo only	$K_{max}=0$		$K_{max}=3$		Δ, for c-Opt $K_{max}=3$		
		g-Opt	c-Opt	g-Opt	c-Opt			
	%	%	%	%	%	%abs	%rel	
DA Types								
mean F	21.8	25.5	28.5	29.3	31.1	+9.3	+43	
F, `fh`	11.3	21.7	21.7	24.0	25.6	+14.3	+127	•
F, `h`	0.0	2.2	1.1	8.5	6.3	+6.3	+∞	○
F, `fg`	0.0	10.4	10.4	12.5	13.7	+13.7	+∞	•
F, `b`	57.1	42.6	56.7	54.7	57.8	+0.7	+1	○
F, `bk`	3.2	12.6	12.6	15.7	14.9	+11.7	+366	•
F, `aa`	2.6	9.4	8.7	12.3	13.0	+10.4	+400	•
F, `s`	91.4	83.0	91.4	82.3	91.3	−0.1	∼0	○
F, `q`	8.8	22.1	23.4	23.9	26.3	+17.5	+199	•
CER	17.0	31.0	17.1	31.1	17.1	+0.1	+1	
DA Termination Types								
F, compl.	53.1	59.9	61.4	58.3	62.1	+9.0	+17	•
F, interr.	0.0	9.6	10.7	22.6	22.6	+22.6	+∞	•
F, aband.	0.0	4.3	7.0	6.6	6.6	+6.6	+∞	•
F, any	53.9	67.6	67.6	67.0	68.2	+14.3	+27	
NIST	64.7	59.0	56.7	66.2	56.2	−8.5	−13	

Table 14.3: EvalSet performance on individual DA and DA boundary types, for the "Topo only" system, and for systems modeling the neighborhood snapshot with both none or three interlocutors modeled. Also shown is the delta ("Δ") incurred by adding the ■/□ with $K_{max}=3$ to the "Topo only" system. Differences tested for statistical significance at the $p > 0.005$ level are shown with • or ○, indicating significance or insignificance, respectively. Other symbols as in Table 14.2.

Generally, the results are numerically similar to those for DevSet. On the primary mean F-score metric for "g-Opt" systems, modeling temporally proximate target-participant speech activity yields an increase of 3.7%abs. Additionally modeling temporally proximate interlocutor speech activity, from the three most talkative interlocutors, yields an additional improvement of 3.8%abs.

As for DevSet, systems exhibit better performance when they are tuned to specific DA types and DA boundary types, by optimizing the emission probability model weight. The largest absolute improvements due to modeling the neighborhood snapshot are observed for floor holders (`fh`), floor grabbers (`fg`), acknowledgements (`bk`), accepts (`aa`), and questions (`q`), as in DevSet. The single largest absolute improvement is also observed for interrupted DAs. On the NIST

boundary detection error, the reduction is 8.5%abs, or 13%rel. However, it appears to not be possible to improve overall classification accuracy by time.

In contrast to DevSet, it occurs occasionally that the c-Opt F-score for a particular DA type is lower than that achieved by the g-Opt counterpart. This is the case for holds (`h`) for both the $K_{max} = 0$ and $K_{max} = 3$ system, acknowledgments (`bk`) for the $K_{max} = 3$ system, and accepts (`aa`) for the $K_{max} = 0$ system. This is because the selection of the c-Opt emission probability model weight, using DevSet, led to values which overfit to the DevSet data; these DA types account for less than 2% of the EvalSet data by time.

14.11 Contrasting ■/□ Context with Prosodic Information

In encoding the temporally proximate distribution of speech from both the target speaker and the most impactful interlocutors, the neighborhood snapshot features also model information related to pause duration — a characteristic falling under the rubric of prosody. Pause duration has frequently been shown to be the most important feature in dialog act segmentation [4]. It is of interest how far other aspects of prosody improve the performance of the proposed text-independent dialog act segmenter and classifier, or, conversely, how far the proposed neighborhood snapshot modeling technique improves over an already existing text-independent prosodic recognizer.

Although prosody has been extensively studied for dialog act recognition and related tasks, the computation of prosodic features is always conditioned on the availability of word information. Even those systems which do not make use of word identity attach prosodic features to word boundaries; word boundaries themselves enable the computation of additional inter-boundary durations. It is therefore currently an open question how to construct a prosodic, text-independent DA recognizer in situations in which words are not available.

This section sets out to do so by beginning with the HMM topology of Section 14.6, and augmenting it with a new emission probability model. That model characterizes the state-conditioned distributions of a feature vector consisting of instantaneous correlates of prosody. The system is then directly extended by inserting a second emission probability model characterizing the ■/□ context, namely that described in preceding sections.

14.11.1 Instantaneous Prosody Features

Prosody is generally taken to refer to phenomena which are extensive in time; in phonetics, phonemics, and phonology, such phenomena are known as *supra-segmental* (where a "segment" coincides with the duration of phonemes). However, even such supra-segmental phenomena have segment-level, or frame-level, correlates. For example, a speaker-normalized fundamental frequency (F_0) trajectory can be reconstructed from frame-level characterizations of the direction of F_0 change. Such correlates are referred to as instantaneous prosody features in the remainder of this chapter.

The instantaneous prosody features considered here are correlates of intonation, probability of voicing, speaking rate, and loudness. The latter three characteristics are frequently modeled, as discussed subsequently. The particular means of modeling intonation, however, is a novel contribution of this thesis. It is achieved by computing the fundamental frequency variation (FFV) spectrum, as described in Section 9.5, which is then passed through a filterbank of seven filters. Five of these filters are related to the likelihood that F_0 in the current frame is quickly falling, slowly falling, flat, slowly rising, or quickly rising. The remaining two filters are used for normalization during feature space rotation.

The FFV spectrum was developed for use in spoken dialog systems [138], using a frame step of 8 ms and a frame size of 32 ms. To retain comparability, these framing parameters are unchanged in the current work. The remaining instantaneous prosody features of this section are therefore also computed using this same framing policy, with a Hamming window. They are:

1. LogEnergy, the logarithm of the total energy in the windowed frame;
2. DeltaLogEnergy, the first-order difference of LogEnergy;
3. NormAutoCorrMax, the value of the first local maximum in the autocorrelation spectrum divided by the value of the zeroeth local maximum (at $\tau = 0$);
4. MelCosDist, the first-order cosine difference of the Mel-frequency spectrum; and

5. LogMelCosDist, the first-order cosine difference of the log-Mel-frequency spectrum.

The first two features are, broadly, correlates of loudness; the third feature is correlated with the probability of voicing; and the last two features characterize spectral flux and are therefore thought to be related to speaking rate.

These 5 features, together with the 7 FFV filterbank outputs, comprise the 12-element instantaneous prosody feature vector of this section.

14.11.2 Emission Probability Model

The instantaneous prosody feature vector, computed at a frame step of 8 ms, must be explicitly or implicitly downsampled in order to be used in a HMM decoder whose states are explored at increments of 100 ms. This is achieved by aligning the two frame sequences, and then training a single state-specific GMM emission probability model using all 8-ms frames which overlap with the 100-ms duration of the state in question; there are 12.5 such 8-ms frames, on average. During decoding, the likelihood of the N observations, with $N \in \{12, 13, 14\}$ depending on the local multirate alignment, is assumed to be given by

$$P\left(\mathbf{q}_{8ms}^{(1)}, \mathbf{q}_{8ms}^{(2)}, \ldots, \mathbf{q}_{8ms}^{(N)} \,|\, \mathbf{y}_{100ms}\right) \;\doteq\; \prod_{n=1}^{N} P\left(\mathbf{q}_{8ms}^{(n)} \,|\, \mathbf{y}_{100ms}\right) . \tag{14.2}$$

The GMM is trained using features extracted from TrainSet; it should be noted that because of the frame step differential, there is effectively 12.5 times as much training material as there would be if the features were computed at a frame step of 100 ms. Prior to training, feature vectors are rotated using a linear discriminant analysis transform. The number of retained dimensions, and the number of GMM elements, is set to maximize the mean F-score on DevSet.

When decoding, the log-likelihood of the 12.5 observations (on average) is combined linearly with the transition log-probability given by the HMM topology. The value of the interpolation weight is also set such as to maximize the DevSet mean F-score.

14.11.3 Analysis of Performance for Specific DA Types

The results achieved with the proposed instantaneous prosody system are shown in Table 14.4.

The table reveals that, in general, the instantaneous prosody system outperforms the neighborhood snapshot system with $K_{max} = 3$ in Table 14.2 by 4.1%abs and 5.0%abs on the primary mean F-score metric, in the g-Opt and c-Opt conditions, respectively. However, adding the neighborhood ■/□ snapshot features, using a second linear log-probability weight (also tuned to maximize DevSet mean F-score), results in improvements on all metrics. The snapshot features are particularly beneficial (cf. the relative improvement for "c-Opt" systems) for floor grabbers `fg`, accepts `aa`, questions `q`, and interrupted or abandoned DA termination.

14.11.4 Generalization to Unseen Data

Table 14.5 shows the results of similar experiments, on the completely unseen TestData. The trends are similar to those observed for DevSet. The instantaneous prosody system is better than the $K_{max} = 3$ neighborhood snapshot system of Table 14.3 by 2.2%abs and 2.6%abs, for the g-Opt and c-Opt conditions, respectively. Adding the snapshot features to the instantaneous prosody system yields improvements for all DA and DA boundary types, which are statistically significant, except for the statement `s` "c-Opt" system (where the change is zero). As for DevSet, improvements in relative terms are particularly noteworthy for floor grabbers `fg`, accepts `aa`, questions `q`, and interrupted or anandoned DA termination. They are also relatively large for holds `h`.

14.11.5 Comparison of Knowledge Sources

A comparison of the performance of the instantaneous prosody system and the ■/□ neighborhood snapshot system, for individual DAs and DA boundaries in EvalSet, is shown in Figure 14.11. The depicted F-score is normalized such that 0% corresponds to the performance of the "Topo only" system, while 100% corresponds to the performance of an oracle

	Topo only	Inst. Prosody		Pros & ■/□		Δ, for c-Opt	
		g-Opt	c-Opt	g-Opt	c-Opt		
	%	%	%	%	%	%abs	%rel
DA Types							
mean F	20.6	32.5	35.3	38.0	39.5	+4.2	+12
F, `fh`	9.4	31.5	33.9	36.5	37.7	+3.8	+11
F, `h`	0.0	19.8	22.6	22.2	23.4	+0.8	+4
F, `fg`	0.0	12.9	12.9	18.9	19.6	+6.7	+52
F, `b`	52.3	56.0	62.9	63.5	65.2	+2.3	+4
F, `bk`	2.7	23.7	25.7	27.9	28.9	+3.2	+12
F, `aa`	2.0	12.7	12.9	17.3	19.4	+6.5	+50
F, `s`	91.9	84.2	92.2	88.2	92.3	+0.1	∼0
F, `q`	6.5	19.3	19.3	29.7	29.6	+10.3	+53
CER	16.3	28.0	15.6	22.0	—	—	—
DA Termination Types							
F, compl.	52.5	58.5	58.5	62.1	62.6	+4.1	+7
F, interr.	0.0	8.4	10.7	23.2	28.6	+17.1	+160
F, aband.	0.0	1.7	3.3	3.7	7.2	+3.9	+118
F, any	53.8	61.2	61.2	66.8	—	—	—
NIST	65.1	68.4	63.7	61.0	—	—	—

Table 14.4: DevSet performance on individual DA and DA boundary types, for the "Topo only" system, an instantaneous prosody system on top of the topology ("Inst. Pros"), and on the combined emission probability system ("Pros & ■/□") consisting of both instantaneous prosody and the ■/□ neighborhood snapshot ($K_{max} = 3$). Also shown is the delta ("Δ") incurred by adding the ■/□ features to the instantaneous prosody features. Other symbols as in Table 14.2.

lexical bigram system (described in the subsequent section). The latter represents an estimate of the best performance achievable by a system which would have access to the true, correctly recognized words.

Figure 14.11 separates DAs and DA boundaries into two groups, those for which the instantaneous prosody system yields higher F-scores (in panel (a)), and those for which the neighborhood snapshot system yields higher F-scores (in panel (b)).

Panel (a) is seen to represent DA types signalling the intent to retain the floor (`fh` and `h`), and DA types implementing two types of feedback (`b` and `bk`). These phenomena are better detected using instantaneous prosody features. For floor mechanisms, this seems to be due to the features ability to capture slower speaking rates and flatter intonation contours. It is, however, somewhat surprising that instantaneous prosody features are also better at detecting feedback speech; feedback was expected to be better detected using long intervals of non-speech, both before and after, which is modeled explicitly by the neighborhood snapshot features. On average, inclusion of instantaneous prosodic features in a ■/□-context-only system, yields F-score improvements of 39%rel.

Panel (b) of Figure 14.11 shows the opposite trend for the remaining DA types and DA termination types, namely higher F-scores achieved with contextual features than with prosodic features. This trend characterizes either beginnings of turns (`fg` and also `aa`) or ends of turns (`q`, *com*, *aba* and *int*, although the latter is not shown because it is off the scale), in which models of the multiparticipant speech/non-speech activity context, indicative of turn construction by interlocutors, were expected to yield good performance.

In summary, the combination of contextual and prosodic features delivers improved performance over using either knowledge source alone (except for statements for which contextual features do not help).

	Topo only	Inst. Prosody		Pros. & ■/□		Δ, for c-Opt		
		g-Opt	c-Opt	g-Opt	c-Opt			
	%	%	%	%	%	%abs	%rel	
DA Types								
mean F	21.8	31.5	33.7	38.4	39.8	+6.1	+18	
F, `fh`	11.3	37.7	39.5	43.5	43.7	+4.2	+11	•
F, `h`	0.0	25.0	17.1	31.8	29.2	+12.1	+71	•
F, `fg`	0.0	7.2	7.2	11.6	14.0	+6.8	+94	•
F, `b`	57.1	48.0	64.6	64.5	66.9	+2.3	+4	•
F, `bk`	3.2	19.0	20.9	24.2	25.6	+4.7	+22	•
F, `aa`	2.6	9.5	8.9	14.0	16.0	+7.1	+80	•
F, `s`	91.4	85.8	91.8	87.3	91.8	0	0	○
F, `q`	8.8	19.6	19.6	30.4	30.9	+11.3	+58	•
CER	17.0	25.9	16.2	23.2	—	—	—	
DA Termination Types								
F, compl.	53.1	59.1	59.1	63.4	63.7	+4.6	+8	•
F, interr.	0.0	10.5	11.8	26.0	28.7	+16.9	+143	•
F, aband.	0.0	2.4	3.6	5.4	7.6	+4.0	+111	•
F, any	53.9	62.6	62.6	68.6	—	—	—	
NIST	64.7	66.5	62.9	57.9	—	—	—	

Table 14.5: EvalSet performance on individual DA and DA boundary types, for the "Topo only" system, an instantaneous prosody system on top of the topology ("Inst. Pros"), and on the combined emission probability system ("Pros & ■/□") consisting of both instantaneous prosody and the ■/□ neighborhood snapshot ($K_{max} = 3$). Also shown is the delta ("Δ") incurred by adding the ■/□ to the instantaneous prosody features. Differences tested for statistical significance at the $p > 0.005$ level are shown with • or ○, indicating significance or insignificance, respectively. Other symbols as in Table 14.4.

14.12 Contrasting ■/□ Context with Lexical Information

In a spirit similar to that of the last section, in which the neighborhood snapshot features were contrasted with instantaneous prosody features, this section contrasts them with lexical features. The goal is to provide a reasonable upper bound on performance. It is known that DA classification is biased towards lexical information because of the way corpora, specifically the ICSI Meeting Corpus, are labeled. The lexical system is made even harder to beat because it relies on true, manually transcribed words, which would not be available to a fully automatic system.

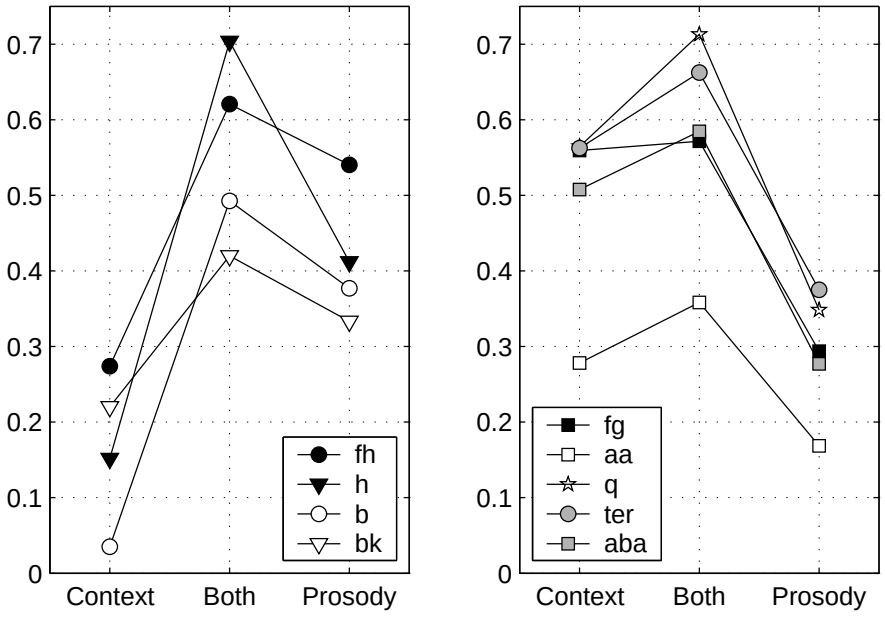
To facilitate system comparison and combination, the lexical system relies on the same HMM topology as do both the neighborhood snapshot and instantaneous prosody systems.

14.12.1 Lexical Features and Model

The features in the proposed lexical HMM system are the left and right bigram of the current word. Bigrams are constructed over a vocabulary as follows.

Where words are separated by more than 0.7 s of non-speech, a `SIL` token is inserted; 0.7 s was found to be optimal in TrainSet for unconditionally deciding whether an inter-word gap is also an inter-DA gap (cf. [4]). For each DA type d, $1 \leq d \leq 8$, the set $\mathcal{U}_d$ of unigrams is formed from TrainSet. These are sorted by frequency of occurrence, and those unigrams whose probability of occurrence exceeds 0.1% are placed in $\mathcal{U}'_d$. Unigrams not found in the union $\mathcal{U}' \equiv \cup_d \mathcal{U}'_d$ are mapped to the token `UNK1`.

Following this first mapping, the set $\mathcal{B}_d$ of bigrams over $\mathcal{U}' \cup$ `UNK1` is formed for each DA type d. Some of the bigrams may thus contain `UNK1` and/or `SIL`. As for unigrams, those bigrams in $\mathcal{B}_d$ whose probability of occurrence exceeds 0.1%



(a) Prosody > Context (b) Context > Prosody

Figure 14.11: Normalized EvalSet F-scores achieved by the neighborhood snapshot system ("Context"), the instantaneous prosody system ("Prosody"), and their weighted log-linear combination ("Both"). Normalization consists of assigning zero to the "Topo only" system and unity to an oracle lexical bigram system (described in the next section) which uses manually transcribed words. Statements (`s`) and interrupted DA termination (*int*) not shown.

are placed in $\mathcal{B}'_d$, and those not found in the resulting union $\mathcal{B}' \equiv \cup_d \mathcal{B}'_d$ are mapped to `UNK2`. $\mathcal{B}'$, together with the `UNK2` token, comprise the bigram vocabulary.

14.12.2 Emission Probability Model

The left and right bigrams are modeled using separate categorical distributions, trained using TrainSet. The returned log-likelihoods from both models are equally weighted, and combined linearly with the transition log-probability using a single interpolation parameter whose value maximizes mean F-score on DevSet.

14.12.3 Validation of Correctness

To assess the correctness of this approach, a separate system was trained on the more standard 5-class DA classification task [4]. When the above lexical bigram model is combined with the HMM topology of Section 14.6 (modified for the 5-class task), its performance slightly exceeds the performance of the language modeling (LM) approach published in [4]. This section assumes that on the 8-class task proposed here, the lexical bigram model embedded in the HMM topology provides performance which would be equally competitive with an 8-class LM system, constructed as in [4].

It should be noted, however, that removing the HMM topology[10] from this lexical bigram system reduces the mean F-score by approximately 20%abs. This makes the proposed lexical emission-only system uncompetitive with the LM system, which does not rely on any topological constraints. Since the HMM topology forms an integral part of the text-independent approach, a better constrastive system would consist of the LM system retrained for the 8-class task, without the topology of Section 14.6. Combining the LM system with the neighborhood snapshot system (which includes the topology) is then likely to lead to much better performance than that presented here.

[10]This is achieved in the decoder by allowing all states in the HMM topology to transition to all states, and assigning all transitions a probability of unity.

14.12.4 Analysis of Performance for Specific DA Types

The mean F-score achieved by the lexical HMM system on DevSet is 50.1%, which is indicative of the difficulty of this task, even when perfect word information is present. The results for individual DA and DA boundary types are shown in Table 14.6.

To assess the complimentarity of the neighborhood snapshot features, the ■/□ model is combined with this lexical baseline (using a second emission model weight). The maximum F-score achieved by the combined system is 52.3%. This represents a 2.2% absolute improvement over a system to which true words are available.

	Topo only	Lex Bigram		Lex & ■/□		Δ, for c-Opt	
		g-Opt	c-Opt	g-Opt	c-Opt		
	%	%	%	%	%	%abs	%rel
DA Types							
mean F	20.6	50.1	51.1	52.3	53.9	+2.8	+5
F, `fh`	9.4	54.1	54.9	56.1	56.3	+1.4	+3
F, `h`	0.0	26.7	29.3	36.4	36.4	+7.1	+24
F, `fg`	0.0	22.8	22.8	24.0	26.8	+4.0	+18
F, `b`	52.3	74.4	74.7	75.9	76.0	+1.3	+2
F, `bk`	2.7	54.9	55.3	54.1	56.2	+0.9	+2
F, `aa`	2.0	36.7	39.0	38.7	42.0	+3.0	+8
F, `s`	91.9	92.1	93.5	92.2	93.8	+0.3	∼0
F, `q`	6.5	39.2	39.2	41.3	43.3	+4.1	+10
CER	16.3	15.6	13.2	16.1	12.8	−0.4	−3
DA Termination Types							
F, compl.	52.5	66.8	67.9	68.1	68.9	+1.0	+1
F, interr.	0.0	24.0	24.0	35.1	36.7	+12.7	+53
F, aband.	0.0	8.8	10.4	10.4	12.1	+1.7	+16
F, any	53.8	69.8	69.8	71.8	71.8	+2.0	+3
NIST	65.1	57.1	53.1	53.4	50.7	−2.4	−5

Table 14.6: DevSet performance on individual DA and DA boundary types, for the "Topo only" system, a lexical bigram system on top of the topology ("Lex Bigram"), and on the combined emission probability system ("Lex & ■/□") consisting of both lexical bigrams and the ■/□ neighborhood snapshot ($K_{max} = 3$). Also shown is the delta ("Δ") incurred by adding the ■/□ to lexical bigram features. Other symbols as in Table 14.4.

14.12.5 Generalization to Unseen Data

On EvalSet, the mean F-score achieved by the lexical-only system is higher than on DevSet by approximately 3%abs. The improvement due to adding ■/□ features to the globally-optimized "g-Opt" system is lower, of 1.7%abs. For "c-Opt" systems, representing systems optimizing individual DA F-scores, improvement is small in absolute terms. However, for the most frequently occurring DAs by time (and accepts `aa`), the improvement is seen to be statistically significant. Larger improvements are observed for DA interruption and abandonment; for all 3 DA boundary types, the improvements are statistically significant.

14.13 Potential Impact

The experiments in this chapter validate the feature-space vocal interaction techniques described in Chapter 8. Modeling the neighborhood snapshot, consisting of the target participant's and interlocutors' temporally proximate speech activity,

	Topo only	Lex Bigram		Lex & ■/□		Δ, for c-Opt		
		g-Opt	c-Opt	g-Opt	c-Opt			
	%	%	%	%	%	%abs	%rel	
DA Types								
mean F	21.8	53.0	54.5	54.7	55.7	+1.2	+2	
F, `fh`	11.3	62.3	63.5	64.8	64.5	+1.0	+2	○
F, `h`	0.0	33.9	41.5	42.3	42.3	+0.8	+2	○
F, `fg`	0.0	24.5	24.5	27.0	27.0	+2.5	+10	○
F, `b`	57.1	77.0	77.0	78.0	77.9	+0.9	+1	○
F, `bk`	3.2	56.3	56.3	55.2	56.0	−0.3	−1	○
F, `aa`	2.6	38.1	40.0	38.1	42.0	+2.0	+5	•
F, `s`	91.4	91.9	93.3	91.8	93.5	+0.2	∼0	•
F, `q`	8.8	39.8	39.8	40.6	42.5	+2.7	+7	•
CER	17.0	15.8	13.4	15.7	13.1	−0.3	−2	
DA Termination Types								
F, compl.	53.1	68.0	69.1	69.3	69.6	+0.5	+1	•
F, interr.	0.0	21.9	21.9	34.0	34.1	+12.2	+56	•
F, aband.	0.0	11.4	13.0	13.1	14.4	+1.4	+11	•
F, any	53.9	71.3	71.3	73.2	73.2	+1.9	+3	
NIST	64.7	53.7	50.7	50.7	49.2	−1.5	−3	

Table 14.7: EvalSet performance on individual DA and DA boundary types, for the "Topo only" system, a lexical bigram system on top of the topology ("Lex Bigram"), and on the combined emission probability system ("Lex & ■/□") consisting of both lexical bigrams and the ■/□ neighborhood snapshot ($K_{max} = 3$). Also shown is the delta ("Δ") incurred by adding the ■/□ to lexical bigram features. Differences tested for statistical significance at the $p > 0.005$ level are shown with • or ○, indicating significance or insignificance, respectively. Other symbols as in Table 14.6.

appears to be a tractable means of inferring interactive behaviors in very large state spaces, with many degrees of freedom per participant. That it appears sufficient to model only the locally most talkative interlocutor, in a 10-second window centered on the current instant, and that "locally" entails a locale shorter still, is interesting, and can impact solutions to numerous conversation problems involving the inference of structure, including topic change, group humor, floor contention, schism, and adjacency pairing.

The techniques are easily extensible to larger frame sizes and steps, and may be useful in detecting conversational phases, namely long intervals characterized by unique interaction patterns rather than the short intervals of this chapter designed to leverage tactical differences in DA type deployment.

On a more pragmatic level, the near-oracle performance on some DAs of the combined instantaneous prosody and ■/□ neighborhood snapshot demonstrates that it is frequently possible to infer DA type, without relying on words or word boundaries. This may find application in many situations, including privacy-sensitive scenarios, resource-deficient domains, languages, and dialects, and acoustically degraded environments.

14.14 Relevance to Other Chapters

This chapter has exercised the feature-space approach to modeling interlocutors (and target participants) which was described in Chapter 8, and provided empirical support of the soundness of that approach.

Additional support is provided in Chapter 15, a part of which treats the detection of attempts at humor and sarcasm. Individual dialog acts in the ICSI Meeting Corpus carry an optional diacritic `j` to indicate when DAs have this function. Section 15.6 will form a ninth dialog act "type", by grouping together all `j`-bearing questions and statements.

14.15 Summary

This chapter has explored the inference of DA type and boundary detection in multi-party conversation, based only on the distribution of speech activity in time and across participants. The approach taken exercises the parametric feature-space models described in Chapter 8. It was shown that modeling a local neighborhood snapshot of both the target speaker's and the locally most talkative interlocutors' speech activity yields mean F-score improvements on unseen EvalSet data of 7.5%abs over a system not modeling that snapshot, of 6.9%abs over an instantaneous prosody system, and of 1.7%abs over a lexical bigram system which assumes 0% word error rates. In defining the neighborhood snapshot, it appears sufficient to capture a centered context of 10 seconds, and only the single interlocutor who is most talkative in an even shorter centered context. The snapshot features outperform duration-to-landmark features traditionally used when modeling context.

Text-independent DA recognition appears viable, when relying on models of both the neighborhood snapshot and instantaneous, word-independent prosody features. Relative to those achieved using the oracle lexical system, condition-optimized F-scores are as high as 78% for questions, 87% for backchannels, 92% for DA completion, and 131% for DA interruption. These results suggest that there is significant promise in modeling interaction in feature space, as proposed, which may be applicable to the inference of a variety of phenomena implicitly related to turn construction.

14.16 Future Directions

The text-independent dialog act recognizer proposed in this work is without precedent, and as such it can be expected that further work on most if not all all of its components will yield significant improvement.

There are some fundamental questions which deserve attention, for example whether models other than the hidden Markov model, such as the conditional random field, would not be better at this task. It is certainly the case that the large temporal extent of feature extraction windows renders them far from independent.

As for other chapters in this thesis, there is reason to more exhaustively sample the parameter space of the proposed techniques, if only to produce additional, more competitive baselines for future approaches to the same problems. In the case of text-independent dialog recognition, it may turn out to be the case that different techniques of ranking interlocutors exist. Rather than centering a rectangular window on the instant t, windows need not be rectangular, and may be mode discriminative of target participants behavior if they contrastively weight precedent and subsequent interlocutor behavior.

Another avenue of inquiry pertains to feature vector transformation. The currently employed LDA transform is linear, while interlocutor behavior may be most optimally combined and expressed using a non-linear transform. A reason to believe this may be the case is that most interlocutors are silent in the temporal neighborhood of the majority of speech.

Finally, it remains to be seen whether dialog act recognition can be performed in a text-independent manner when the posteriors of speech are not $\in \{0, 1\}$, but $\in [0, 1]$. There are some methodological difficulties of how to score dialog act segmentation when speech is entirely missed or falsely posited. But asides from this, true posteriors may actually improve on the results presented in this chapter. Interlocutors may audibly inhale and/or exhale, for example, at times which are plesiochronous with target-participant DA production, but this behavior is indistinguishable from silence when reference speech/non-speech posteriors are used.

Chapter 15

Text-Independent Emotional Epiphenomenon Recognition*

15.1 Introduction

A conversation can be expected to evolve through a number of phases, such as an opening, a list of topics punctuated by topic changes, and a closing; each topic nominally has its own nested opening, its own closing, and a body consisting of speaker contributions. These distinct phases may but may not have obvious correlates in the vocal chronogram of a conversation. However, inspection of real vocal interaction records — without topic annotation — does reveal distinct phases. This suggests that what is inferrable at the vocal activity level is a sequencing which may be independent of the propositionally relevant topic axis.

Examples of phenomena apparent in vocal interaction records, which are time-dependent, are intervals of systematic departure from time-independent statistics. For persistent vocal activity types whose production is governed by a turn-taking mechanism (i.e., speech), this includes intervals of varying density of speaker change, and of duration of overlap. For spurious vocal activity types whose production is not so governed (i.e., laughter), this includes intervals of varying frequency of occurrence, as well as of duration overlap.

This chapter hypothesizes that the sequencing of intervals contrasting in the above ways is associated with participants' trajectories of internal state. Internal states, not overtly observable, are often called *emotions*, and this convention will be used in the current work. The term is appropriate because emotions are said to mobilize (or demobilize) subsequent action [156]. Such action, defined in terms of observable vocal activity, can be expected to have repercussions on the interaction statistics in the immediate temporal neighborhood.

Although a direct test of the above hypothesis consists of measuring the strength and statistical significance of association between joint vocal activity and joint emotional state, this chapter — like the preceding Chapter 14 — tests the hypothesis indirectly. Alternatives to computing association metrics include the prediction of vocal activity given emotional state, or the inferrence of emotional state given vocal activity. Three inference tasks, exemplifying the second alternative and properly falling under the rubric of emotion recognition from speech, are described here.

Emotion recognition has a long tradition in speech processing, although in multi-party conversation it has been studied rarely (cf. Section 15.2 for exceptions). Research was initially motivated by attempts to recognize emotionally inflected speech [53, 183], to augment appropriateness of response in spoken dialog interfaces [11], and to provide a simulacrum of emotional intelligence in myriad applications [182]. Early work made use of observations on so-called "prototypical" emotional expressions, contrastive exemplars of which were created by employing actors (cf. [10]). This led to interesting insights, but insights which did not port well to naturally occurring spontaneous speech. Actors were often asked to say the same utterances in contrasting emotional states, whose absolutely constrained lexical choice led to compensation via exaggerated prosodic contours. Perhaps more importantly, the canonical closed set of the emotions studied does not seem

*The work in this chapter was conducted in collaboration with Anton Batliner, Susanne Burger, Kjell Elenius, Inger Karlsson, and Daniel Neiberg.

to appear in its entirety, or even significant part, in most natural application scenarios, and what is considered emotionally unmarked in practice exhibits far more variability than what actors produce when asked to be unemotional.

Research which has addressed the deficiencies inherent to acted emotional speech has turned to data collected under the same or similar conditions as those in which systems are intended to be deployed [57]. This, however, entails its own set of concerns. First, naturally occurring emotional speech is relatively rare in almost all envisioned scenarios, requiring that large amounts of data be collected to construct even the lowest-complexity models. Second, the annotation of emotional speech is difficult, and it is frequently conducted without describing to annotators the behavior of the system that is being designed. Finally, task-specific system design, as called for by these two concerns, leads to splintering of a small research community and a large curtailment of interpretability of both techniques and results across different task domains.

As a result, the number of domains for which emotion recognition from speech can be said to be maturing is small. Examples of successes are call center dialog systems [157, 3, 12], tutoring systems [2, 161], and robotic toys [14]. The focus in these areas has been on specific emotions, such as annoyance or anger in the first, confusion or enthusiasm in the second, and coaxing, controlled or frustrated command states in the second. An important aspect of these types of systems is that they treat machine-directed speech. Not only is the expression of the same emotions in human-human speech likely to be different, but the emotions appearing in human-human social interaction themselves are likely to be different. From a computational point of view, the natural occurrence of emotional speech in social interaction is poorly understood.

The work in this chapter aims to address some of these problems, in the context of multi-party conversations or unscripted and unmediated meetings. It is divided into four parts. The first part describes an attempt to annotate emotionally relevant behavior in a large meeting corpus. The findings indicate that annotators agree on the emotional status of utterances but not necessarily on the type of emotion, except in cases where utterances contain laughter. This in part corroborates studies on the functional role of laughter in professional meetings [118]. The second and third parts aim to infer the valence and activation ordinates of emotional states, the two most commonly studied dimensions in Schlosberg's now-famous multi-dimensional emotional space [200]. Both experimental suites, conducted on different corpora, point to the paramount importance of laughter to the inference of emotional state in social situations. The fourth and final part explores one reason for the occurrence of laughter, namely interlocutor attempts to amuse.

The annotation work on meetings presented here first appeared in [134]. Computational work using the resulting annotations was described in [173], and was based on earlier experiments conducted on the CEICES data set in [14]. The research on hotspot detection appeared in [127] and [130]. Research on attempts to amuse first appeared in [129].

15.2 Related Work

The recognition of emotion or affect in multi-party meetings remains an underexplored area of research.

In 2003, Wrede & Shriberg claimed [225] that utterances of marked emotional activation, or involvement, in the ICSI Meeting Corpus convey *amusement*, *disagreement*, interest, surprise, or excitement (the latter three categories are mapped to a single category of *other* in [225, 224]). Using a small portion of the corpus, it was shown that prosodic variation was correlated with involved speech, or hotspots. In [224], the same authors claimed that only 2.2% of *M-segs*[1] are emotionally involved. Finally, data in Table 4 of [224] suggests that *other* M-segs comprise the majority of emotionally involved M-segs, while *amusement* and *disagreement* M-segs account for 21% and 17%, respectively. Hotspots were shown to associate with specific DA types at rates which are above chance. These observations were made using a larger portion of the ICSI Meeting Corpus; the corpus has since been annotated using a different definition of "involvement hotspot" in its entirety [223]. In [42], it was demonstrated that hotspots also co-occurred with higher rates of speech overlap.

Annotation of emotion was also studied in the AMI Meeting Corpus [93, 189], using categorical labels as well as continuous multi-dimensional-space attributes such as activation and valence. These works focused on annotation methodology and did not treat the classification or recognition problems.

A somewhat related topic of inquiry is the detection of agreement and disagreement, which has received some attention earlier in this decade [94, 66], as well as more recently under the larger umbrella of subjectivity and sentiment analysis [205, 186, 69]. In contrast to the research mentioned above, this work does involve a computational component, which relies extensively on verbal rather than nonverbal production.

[1] An "M-seg" is the longest contiguous word sequence belonging to at most one dialog act segment (DA-seg) and at most one involvement segment (I-seg); cf. [224] for a complete description.

15.3 Observations from Manual Annotation

At the time the work described in this chapter was begun, currently-available, large corpora of multi-party conversation had either not been released yet (e.g., the AMI Meeting Corpus), or had been released but no annotation of emotionally relevant phenomena was yet available (e.g., the ICSI Meeting Corpus). This led to a large annotation effort, described in this section.

A primary goal of the annotation work was to answer the following two questions:

1. What kinds of emotional phenomena are present in meeting corpora?
2. What is the frequency of occurrence of those emotional phenomena?

These questions are indicative of the paucity of understanding in the field of meeting speech; the successful modeling (and detection) of phenomena is a strong function of the frequency of their occurrence, particularly for phenemomena at the lower end of the spectrum.

15.3.1 Dataset Use

The preliminary analysis of emotional epiphenomena in conversation was conducted using the ISL Meeting Corpus [30], described in Chapter 4. Several aspects of this corpus recommended its use in this work. First, it is a corpus of conversations, of which work-place meetings comprise but a subset; a wide range of conversation types is included [30, 31]. Second, the conversations are not scripted; although on occasion participants were collectively given a topic to discuss, they did not have the opportunity to practice or rehearse their utterances. Speech in these conversations may be characterized as spontaneous, even though the conversations themselves cannot be said to be naturally occurring. Third, the conversations are multi-party, in that each involves more than two participants, and the number of participants varies from conversation to conversation. Finally, at the time the corpus was collected, analysis of emotionally relevant phenomena was not on the researchers' agenda. It is therefore expected that if participants were concealing their emotions, they were doing so as part of a broader posture towards being recorded in the first place.

Speech in the ISL Meeting Corpus had been transcribed, and segmented into speaker contributions, or turns. Nonverbal productions were also transcribed, and in some cases comprised the entirety of a turn, if they were temporally distant from speech. Examples of nonverbal production occuring in this way in the transcriptions are laughs, coughs, and breaths. The analysis conducted in this section used all 18 meetings, referring to them jointly as CompleteDataSet. Pilot work involved a subset of 5 meetings, namely `m013`, `m031`, `m036`, `m038`, and `m042`, of meeting types Project, Work, Game, Discussion, and Chat, respectively. This subset is referred to as PilotDataSet1; it consists of 3668 distinct speaker contributions.

15.3.2 Open-Set Annotation

A data-driven, as opposed to theory-driven, approach was selected to analyze the occurrence of emotion in conversation. Three annotators were asked to identify from among all utterances those which are *emotionally marked*, namely those whose production an observer feels is emotionally inflected, or potentially informative of the speaker's emotional state in some way. Annotators had not participated in the meetings, and therefore their judgment may best simulate what users of a browsing system might want to retrieve from an arbitrary corpus. The annotators were allowed to qualify each utterance with *any text string*, based on the multichannel close-talk microphone audio, the utterance-level segmentation, and the orthographic transcriptions. Splitting of utterances which annotators felt contained multiple phenomena was allowed.

Open-set annotation was applied to only the five meetings in PilotDataSet1. Annotators listened to meetings in arbitrary order, and in several cases did not listen to all of a meeting's utterances. As a result, of the 3668 utterances, only 1649 were listened to by all three annotators. Figure 15.1 shows, per meeting type, the proportion of utterances listened to by all annotators which received an emotionally marked label from at least one annotator, from at least two annotators, and from all three annotators. As can be seen, the meetings of type Project and Work seem to contain fewer emotionally marked utterances than the other three meetings of more informal type. This suggests a strong negative correlation between formality of meeting and density of perceived emotional phenomena, which is not unexpected. The figure also indicates that agreement on whether an utterance is emotionally marked or not is a function of meeting type; a

second annotator disagrees more frequently than not for meeting types PROJECT and WORK, while for the other meeting types agreement is more frequent. Except for the CHAT meeting, three annotators agreed on emotional relevance for less than 20% of utterances.

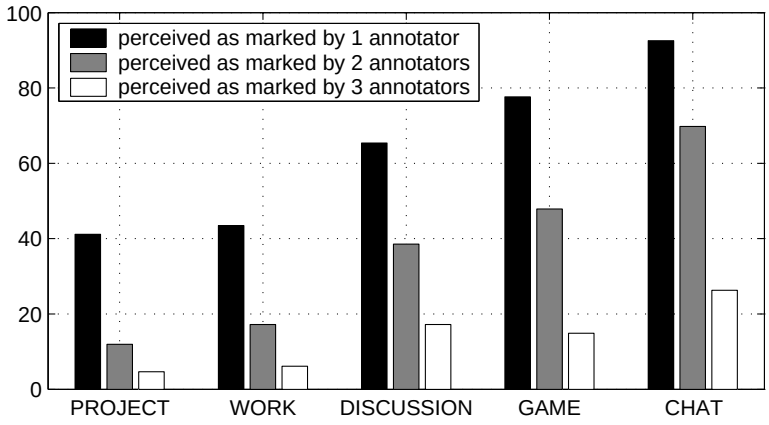


Figure 15.1: Proportion (in %) of utterances listened to by all three annotators, per meeting type in PILOTDATASET1, which were perceived as emotionally marked by $n \in \{1, 2, 3\}$ annotators.

Pooling all the open-set labels, henceforth referred to as the L0 labels, resulted in a total of 3170 assignments involving 596 unique text strings.

15.3.3 Closed-Set Annotation

While open-set annotation can provide insight into the proportion of utterances which are thought to be emotionally relevant, tractably quantifying the *kind* of emotional markedness requires either that a distance measure be defined over the open-set labels, or that annotators systematically employ a small, closed-set of labels. Since the ultimate goal was to annotate all COMPLETEDATASET meetings, it was decided that the second route be followed to further simplify subsequent annotator effort.

Annotation with L1

The process of designing a closed-set L1 annotation scheme began with the open-set L0 labels, and is shown in Figure 15.2. The process was iterative, and many details are not shown. The first step consisted of manually clustering the 596 L0 labels to yield 255 labels, which were further partitioned into 13 mutually exclusive categories (plus NOTAPPLICABLE). This scheme is referred to as L1 in Figure 15.2; L1 label names were chosen to describe the common aspects of the L0 labels they subsumed. The number of distinct L0 labels for each L1 label is shown in Figure 15.3. As can be seen, these numbers are unbalanced across L1 labels, possibly indicating the certain L1 labels exhibit a much richer variety of underlying phenomena than others. For example, PLEASED and COMPLAINING subsume 94 and 76 different L0 labels, respectively. AGREEING, on the other hand, is a fairly homogenous L1 class containing only 16 distinct L0 labels. Differences can also be observed across labelers; one labeler was responsible for generating half of all L0 labels. An important observation is that, with the exception of PLEASED and DISENGAGED, L1 labels are descriptors of behavior rather than of state. This suggests that naive observers find the description of observable action more intuitive than the inference of unobservable state.

The L1 label set was given to two of the L0 labelers, to annotate only the *main* intent or phenomenon in each utterance. As in L0 annotation, splitting of utterances was allowed. The set of meetings thus annotated was the same as in the L0 pass, but `m013` and `m031` were replaced with `m043` and `m063`. The resulting data set is referred to as PILOTDATASET2 in Figure 15.2. L1 phenomena frequently labeled ($> 5\%$) by both annotators include UNMARKED, PLEASED, DEFENDING

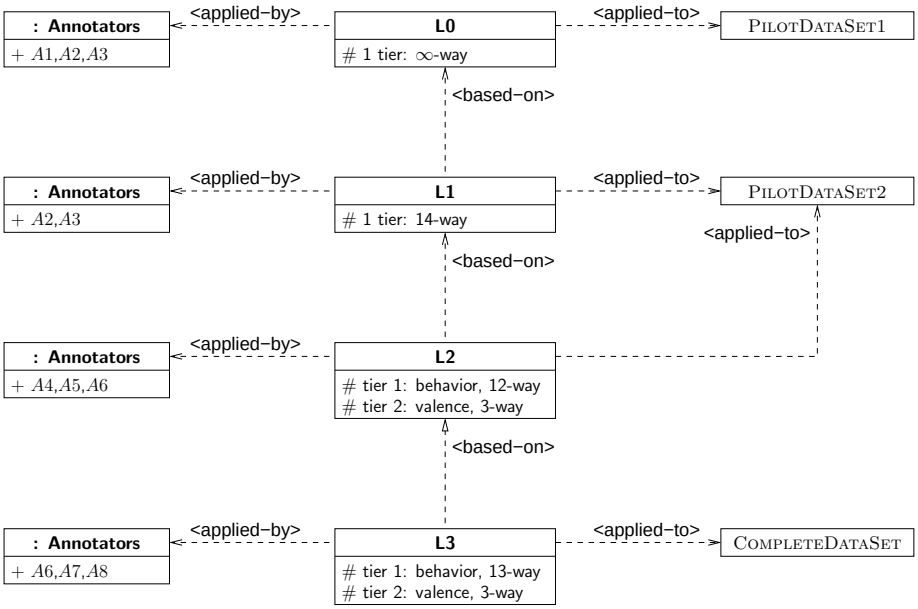


Figure 15.2: Iterative design of the final closed-set annotation scheme, applied to CompleteDataSet, starting from the open-set L0 labels.

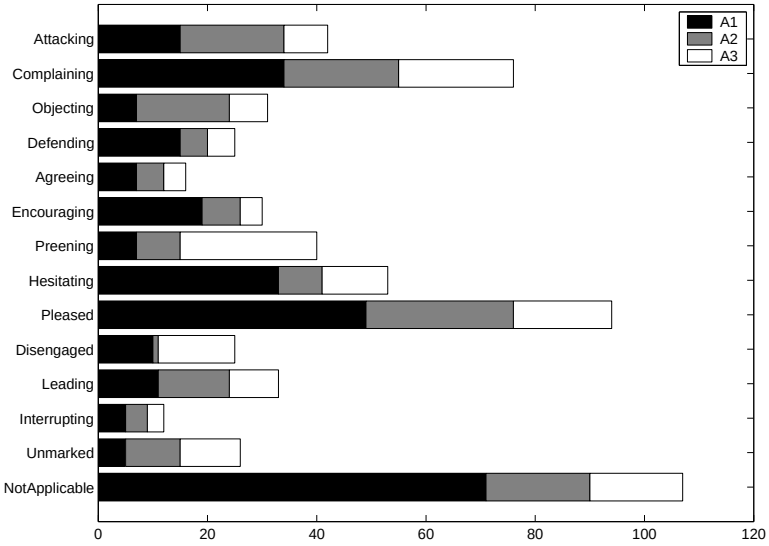


Figure 15.3: Number of L0 labels (along x-axis) applied by annotators A1, A2, and A3, eventually subsumed by the L1 labels (along y-axis).

and Agreeing, while Attacking, Disengaged, Interrupting, and Preening were assigned to fewer than 2% by both. Interlabeler kappa for all five meetings (2586 utterances in total) was 0.30, even though the annotators were already familiar with 64% of the utterances from the initial open-set annotation step. This level of agreement was deemed to be too low for computational modeling, and prompted an analysis of the L1 labels and a subsequent improvement over the L1 annotation scheme.

Annotation with L2

Two observations were made. First, it was found that the L1 labels were not sufficiently mutually exclusive; in particular, PLEASED seemed mostly orthogonal to the other labels. Second, annotators reported that a lot of time was spent reinterpreting the meaning of the L1 label names. It is not known to what extent interlabeler disagreement is due to their different between-annotator as well as within-annotator interpretations.

To address these concerns, descriptions of *emotionally relevant behavior* were separated from descriptions of state. The PLEASED label was removed, and its function was replaced by a second annotation tier of three discrete emotional valence classes (NEGATIVE, NEUTRAL, and POSITIVE), as frequently proposed elsewhere (e.g., [160]). Emotional activation was not annotated[2]; it had been studied in the context of meetings in [225].

The labels of the first annotation tier, describing emotionally relevant behavior, were nested in a decision tree, as has been done for other dialogue-level distinctions elsewhere [38]. The decision tree is shown in Figure 15.4. The aim was that instead of interpreting the suitability of a label name, annotators would arrive at a label by answering yes-or-no questions from the root of the tree. For this reason, leaf labels appeared as random letters during annotation; they are encoded here as text strings, formed by concatenating prefixes of keywords in affirmatively answered questions, for clarity of description. The structure of the tree is a deliberate attempt to focus on emotionally interesting phenomena. For example, the first question, *Express discontent?*, is meant to identify emotionally important behaviors, which are rare and which might otherwise be lost if the labeler was allowed to answer other questions first (e.g., *Providing/requesting info or opinion?*). Certain categories exist to eliminate certain behaviors prior to further questioning. In general, behaviors which were felt to be positive interaction behaviors are on the right side of the tree, those which were felt to be negative are on the left [156]. The single dashed line from the right side to the left expresses the assumption that when defending one participant from a second, a third party objects to the second participant's behavior in the same way they might object to an opinion.

The task of relabeling the five PILOTDATASET2 meetings with the resulting L2 annotation scheme was given to three new annotators, $A4$, $A5$, and $A6$, previously exposed to neither the ISL Meeting Corpus data nor the task of annotating emotionally relevant behavior. In a first pass, annotators were asked to assign to each utterance a label for emotionally relevant behavior. Then, in a second pass, during which annotators had access to their own behavior labels, they were asked to assign a label for valence (one of POSITIVE, NEUTRAL or NEGATIVE). Pairwise interlabeler agreement κ is shown in Table 15.1 for both 3-way emotional valence and 12-way emotionally relevant behavior. Even the lowest agreement κ of 0.43 is higher than that obtained using L1 and annotators familiar with the data and task. 3-way agreement on assessment of valence lies between 0.47 and 0.63.

	$A4$	$A5$	$A6$
$A4$	—	.63	.48
$A5$	.48	—	.47
$A6$	.43	.51	—

Table 15.1: Pairwise κ for interlabeler agreement using the L2 annotation scheme. 3-way emotional valence is shown above the diagonal, 12-way emotionally relevant behavior below.

Analysis of the emotionally relevant behavior labels from annotators $A4$ and $A6$, which exhibit the lowest chance-corrected agreement in Table 15.1, indicates that agreement is relatively high for `Info`, `Agr`, `AgrAck` and `Other`; these four labels were also the most frequently assigned labels for all three annotators. `AttAmuse`, `DisagrConf`, and `Discnt` also showed high agreement, but were frequently confused with `Info`. Labels which exhibited low agreement between $A1$ and $A3$ included `Disagr`, `AgrImprEst`, and `PromEgo` — all three of which were relatively rarely applied.

Figure 15.5 shows a non-metrical dimensional scaling solution for a two dimensional space [13] which attempts to account for the confusions which were observed. In this diagram, behaviors which exhibit significant positive correlation with NEUTRAL valence are shown as squares; those which exhibit significant positive correlation with POSITIVE valence are shown as circles; the remainder are shown as triangles. The diagram indicates that valence appears to be a good

[2]It has been reported [193] that for naturally occurring speech, listeners find it easier to distinguish between activation levels than they do between valence levels.

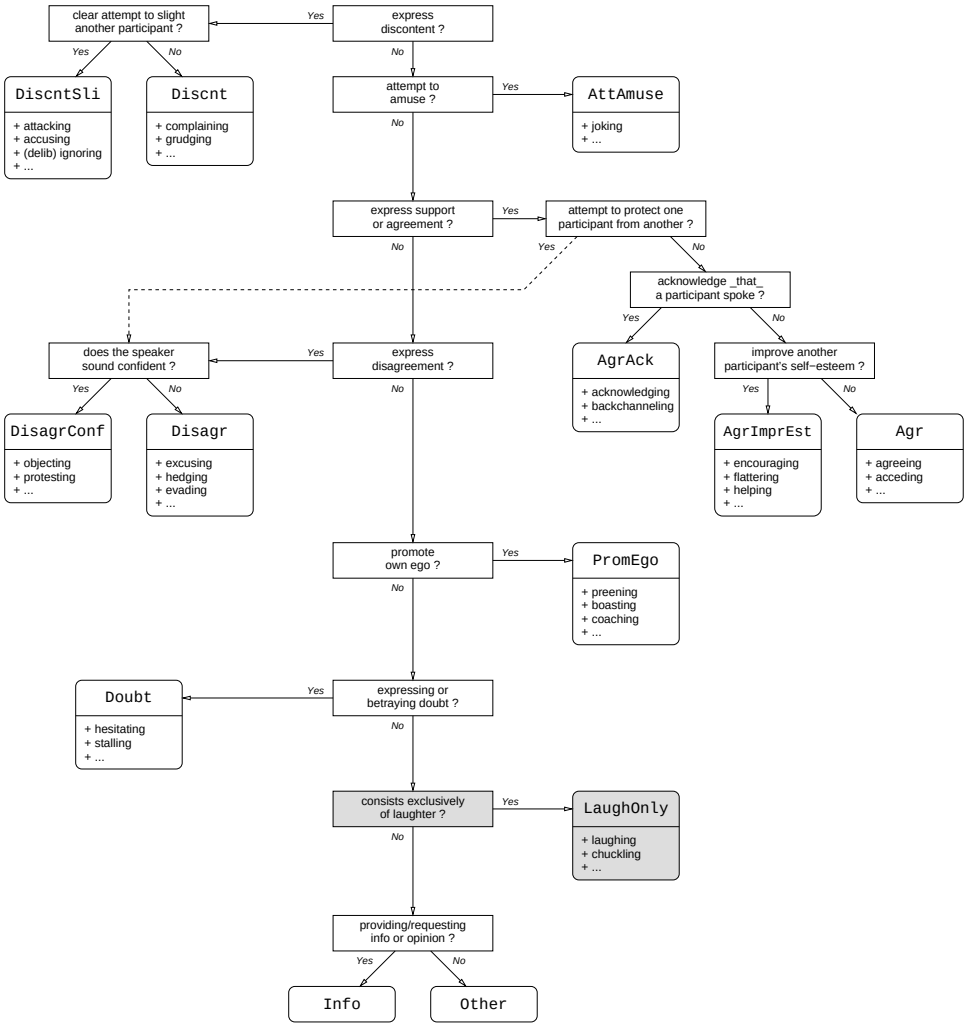


Figure 15.4: Decision tree used to annotate emotionally relevant behavior, comprising the L2 annotation scheme (when the question and leaf node in gray are excluded) and the L3 annotation scheme (when the question and leaf node in gray are included).

candidate for the horizontal axis in this inferred two-dimensional space, and its annotation may have therefore been a good choice in an independent tier.

Annotation with L3

Finally, analysis of `Other` revealed that a very large proportion of these utterances contain laughter only, and that it is on these laughter-only utterances that annotators agree the most, within `Other`. Therefore, a new scheme, denoted L3 in Figure 15.2, was proposed to annotate all of CompleteDataSet; it is identical to the L2 scheme, except that a new label, `LaughOnly`, was added to the tree (as shown in Figure 15.4 in gray). This final scheme was applied to the entire ISL Meeting Corpus by annotators $A6$, $A7$, and $A8$, two of which had not previously had experience annotating emotion.

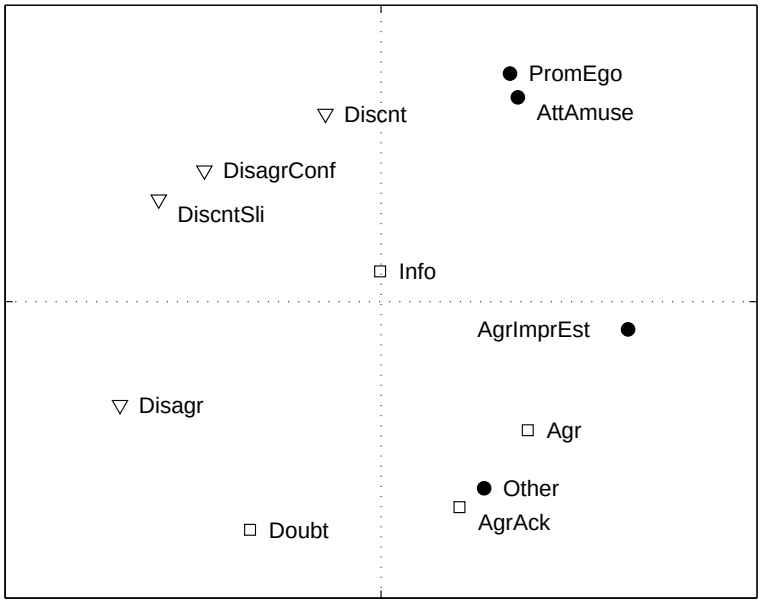


Figure 15.5: Non-metrical dimensional scaling (NMDS) solution, given the confusion matrix for L2 labels assigned by annotators $A1$ and $A3$, to the conversations in PilotDataSet2.

15.3.4 Analysis of Agreement on Emotionally Relevant Behavior

Of the 13221 speaker contributions in the corpus, 803 (6.1%) exhibit no majority (each of three labelers assigns a different category). Of the remaining 12418 speaker contributions for which a majority does exist, 7872 (63.4%, or 59.5% of the total) exhibit complete agreement among all three labelers. The majority of speaker contributions are devoid of behaviors usually associated with emotion, consisting primarily of `Info`, `AgrAck` and `Agr`; the results of majority voting are shown in Table 15.2. Annotators appear to largely agree on the exclusive presence of laughter: speaker contributions receiving the `LaughOnly` label by at least two annotators make up 8% of the corpus. Other behaviors which are interesting from an emotion point of view are more rare, together accounting for just over 7% of all speaker contributions, given majority label voting and excluding laughter. However, all categories with the exception of `DiscntSli` receive a vote at least 1% of the time. In general, when these labels are assigned by only one annotator, the other two annotators tend to select one of `Info`, `AgrAck` and `Agr`.

	DiscntSli	Discnt	DisagrConf	Disagr	Doubt	Other	Info	AgrAck	Agr	PromEgo	AgrImprEst	AttAmuse	LaughOnly	Majority Votesper behavior
DiscntSli	8	5	1	1	0	0	8	0	0	0	1	1	1	26
Discnt	1	3	0	1	1	5	12	7	0	0	0	1	0	31
DisagrConf	3	2	49	26	4	1	106	2	4	0	0	0	0	197
Disagr	0	1	12	9	6	2	42	3	3	0	0	0	0	78
Doubt	0	0	1	3	34	29	35	17	6	0	0	0	0	125
Other	0	12	0	2	26	155	35	13	1	1	5	0	6	256
Info	21	139	195	192	203	191	4973	229	477	121	103	276	4	7124
AgrAck	0	13	4	4	120	53	161	901	428	0	13	1	11	1709
Agr	0	1	3	6	9	4	337	498	687	2	13	4	2	1566
PromEgo	0	0	0	0	0	0	3	0	0	2	0	0	0	5
AgrImprEst	0	0	0	0	0	0	14	7	5	0	6	0	0	32
AttAmuse	3	1	1	1	0	1	161	1	1	1	0	41	0	212
LaughOnly	0	0	1	0	1	22	4	23	1	0	0	1	1004	1057
Minority Votes per behavior	28	174	218	236	370	308	918	800	926	125	135	284	24	

Table 15.2: Annotator majority (rows) vs annotator minority (columns) voting, emotionally relevant behavior, for the 93.9% of speaker contributions for which an annotator majority does exist. The rightmost column represents the number of instances of a majority, per behavior category; the number of instances of a strict minority, per behavior category, is shown in the bottom row. Numbers along the diagonal represent unanimity among the three annotators.

Table 15.3 shows absolute agreement, chance agreement (assuming labeler independence), and chance-corrected agreement in the form of the kappa statistic, for each annotator pair. The κ values for the 3 untrained annotators lie in a tight range of 0.56 to 0.59, which can be considered acceptable [48, 34]. Agreement is notably worse than that reported for dialog act/structure coding schemes involving practiced labelers, ie. $0.75 \leq \kappa \leq 0.86$ for 4 classes in [39] and $0.75 \leq \kappa \leq 0.82$ for 6 classes in [202], but it is on par with more subjective distinctions in meetings such as agreement/disagreement in talk-spurts, where $\kappa = 0.63$ for 4 classes [66].

Labelers	$A6/A7$	$A6/A8$	$A7/A8$
Absolute agreement	0.72	0.70	0.71
Chance agreement	0.34	0.32	0.29
κ coefficient	0.58	0.56	0.59

Table 15.3: Interlabeler agreement on the entire ISL Meeting Corpus (13221 speaker contributions), emotionally relevant behavior.

15.3.5 Analysis of Agreement on Emotional Valence

Of the 13221 speaker contributions, only 76 (0.58%) exhibit no majority. Of the remaining 13145 speaker contributions for which a majority does exist, 9526 exhibit unanimity. The distribution of the assignments of the minority annotator versus the assignments of the majority annotators is given in Table 15.4. As is shown, NEUTRAL valence accounts for 81% of speaker contributions, with an annotator majority agreeing that the proportion of NEGATIVE speaker contributions is less than 1%. However, over 16% of speaker contributions receive a POSITIVE valence label from an annotator majority, which is more than was expected.

Table 15.5 shows absolute agreement, chance agreement (assuming labeler independence), and chance-corrected agreement kappa for each annotator pair. Agreement between annotators $A6$ and $A8$ is similar to that reported elsewhere, ie. utterance-level hot spots in meetings ($0.35 \leq \kappa 0.79$, 4 classes) in [225], as well as for valence in other domains, including tutoring dialogues ($0.40 \leq \kappa \leq 0.68$, 3 classes) in [162] and automated travel planning systems (0.47, 5 classes) in [3]. However, agreement for pairs involving labeler $A7$ is close to chance. In spite of using naive annotators, significantly better than chance agreement was expected on what passes for non-NEUTRAL valence.

In the remainder of this section, we assess this annotator's behavior in the context of more annotators which had labeled a pilot subset [133] of the ISL Meeting Corpus.

Some insight into $A7$'s behavior can be gained by comparing inter-labeler kappas from $A6$, $A7$, and $A8$ during L3 annotation with that from $A4$, $A5$, and $A6$ during L2 annotation, on PILOTDATASET2 (which is a proper subset of

	NEGATIVE	NEUTRAL	POSITIVE	Majority Votesper valence
NEGATIVE	**22**	85	10	117
NEUTRAL	354	**9361**	1142	10751
POSITIVE	49	1887	**235**	2155
Minority Votes per valence	403	1972	1152	

Table 15.4: Annotator majority (rows) versus annotator minority (columns) voting, emotional valence, for the 99.4% of speaker contributions for which an annotator majority does exist. Conventions as in Table 15.2.

Labeler	*A6*/*A7*	*A6*/*A8*	*A7*/*A8*
Absolute agreement	0.77	0.89	0.79
Chance agreement	0.73	0.65	0.76
κ coefficient	0.15	0.67	0.14

Table 15.5: Inter-labeler agreement on CompleteDataSet (13221 speaker contributions), emotional valence.

CompleteDataSet). This is shown in Table 15.6. What can be seen is that *A7* has an average $\kappa = 0.10$; excluding *A7* from all averages yields average κ increases of approximately 0.10. This suggests that *A7* was in fact atypical in his valence assignments (*A7* subsequently reported that he systematically avoided assigning valence labels to `LaughOnly` utterances). It should also be noted that *A6* exhibited significant differences during L2 and L3 annotation, yielding the highest $\kappa = 0.73$ in the table.

Annotation Pass	L2			L3		
& Labeler	*A4*	*A5*	*A6*	*A6*	*A7*	*A8*
A4		0.48	0.68	0.66	0.10	0.59
A5	0.48		0.49	0.48	0.08	0.45
A6	0.68	0.49		0.73	0.11	0.64
A6	0.66	0.48	0.73		0.11	0.64
A7	0.10	0.08	0.11	0.11		0.11
A8	0.59	0.45	0.64	0.64	0.11	
mean	0.50	0.40	0.53	0.52	0.10	0.49
excl A7	0.60	0.46	0.61	0.60	–	0.58

Table 15.6: Pairwise inter-labeler agreement kappas for PilotDataSet2 (2558 speaker contributions), emotional valence. The last two rows represent average kappas, both including and excluding labeler L1.

15.3.6 Intra-Speaker State-to-Behavior Association

For emotionally relevant behavior to be in fact relevant to emotion, it is expected that there is an association between the two annotation tiers. A sample cross-tabulation analysis, between behavior as assigned by annotator *A8* and valence as assigned by *A6*, is shown in Table 15.7. In addition to absolute counts, the table reports the statistical significance of deviation from the null hypothesis of no association, at both the $p < 0.01$ and $p < 0.001$ levels.

It appears that the informational behaviors which comprise the majority in this corpus have an association with Neutral valence which is significantly higher than that expected by chance, and their association with Positive valence is significantly lower. As anticipated, `AttAmuse` and `LaughOnly` exhibit the reverse trend. It is interesting to note that while there is 50% more of `DiscntSli`, `Discnt`, `DisagrConf` and `Disagr` when the four are taken together than there is of `AttAmuse`, only the latter is perceived by observers to be co-occurring with non-Neutral valence in a large majority of cases. Co-occurrence with Negative valence of the four behaviors expressing discontent or disagreement is significantly above chance, but all four co-occur with Neutral valence more than they do with Negative valence. This suggests that meeting participants may be suppressing their expression of Negative valence more effectively than of their Positive valence, or alternately that the vocal expression of Negative valence is more recipient-specific and not perceptible to outside observers (the annotators).

Finally, it is noted that `DiscntSli` and `Discnt` exhibit significant above-chance association with Positive in addition to that with Negative valence. This is provisionally attributed to "teasing" behaviors, in which participants display discontent towards each other mixed with, or covered by, humor, or in which they enjoy complaining. Cross-tabulation analyses involving different pairings of labelers reveal a similar pattern (except those involving valence from labeler *A7*, whose valence assignments were disregarded for reasons mentioned earlier).

	NEGATIVE		NEUTRAL		POSITIVE	
DiscntSli	++	8	−−	13	++	25
Discnt	++	37	−−	132	+	69
DisagrConf	++	20		231	−	45
Disagr	+	11	+	258	−−	39
Doubt		10	++	524	−−	42
Other		5	++	218	−−	21
Info	−	81	++	5452	−−	897
AgrAck	−	10	++	1455	−−	91
Agr	−	11	++	1398	−−	218
PromEgo		5		138		24
AgrImprEst		3		174		64
AttAmuse		6	−−	66	++	360
LaughOnly	−	6	−−	56	++	998

Table 15.7: Co-occurrence of emotional valence as assigned by labeler $A6$ with the same speaker's emotionally relevant behavior as assigned by labeler $A8$ in CompleteDataSet (13221 speaker contributions), absolute counts. + and − represent rejection of the null hypothesis of no association, based on a χ^2 test. ++ and + identify counts which are significantly above that expected by chance; −− and − identify counts significantly below chance. Significance is at the $p < 0.001$ level for ++/−−, and at the $p < 0.01$ level for +/−.

15.3.7 Inter-Speaker Behavior-to-State Association

While the preceding subsection assesses the degree to which a speaker's emotional valence is associated with the same speaker's concurrent behavior, this subsection assesses the degree to which a speaker's emotional valence is associated with her/his interlocutor's previous behavior. This requires the identification of pragmatic adjacency [158]. It was shown in the ICSI Meeting Corpus [108], for which adjacency pair annotation exists, that selecting the most recent speaker yields a correct part-A identification accuracy of 79.8%. Using this simple algorithm, with minor extensions to resolve overlapping speaker contributions, results in the cross-tabulation analysis shown in Table 15.8. Speaker contributions which had been split prior to annotation, and for which no segmentation start-time or end-times are immediately available, were excluded. This results in a total of 11857 speaker contributions with an identified part-A. It is possible for a given speaker contribution to be the part-A of zero, one or more contributions from other speakers.

Table 15.8 shows a pattern similar to that of Table 15.7, in that the precedent speaker's `Info`, `AgrAck` and `Agr` show significant above chance co-occurrence with the current speaker's NEUTRAL valence, and below chance co-occurrence with the current speaker's POSITIVE valence. Similarly, `AttAmuse` and `LaughOnly` exhibit the opposite association. In contrast to Table 15.7, this crosstabulation analysis reveals that the association between NEGATIVE valence and the precedent speaker's `DiscntSli` or `Discnt` is not significantly different from chance. This suggests that complaining or criticizing behaviors, which are rare to begin with, may not lead to NEGATIVE valence in other meeting participants. However, they appear to have the same significantly above chance association with their hearers' POSITIVE valence as with their speaker's. Finally, it seems that `AgrImprEst` is effective. As in the previous section, patterns for crosstabulation analyses with different labeler pairs show similar results.

15.4 Classifying Emotional Valence

Given the annotation produced in the preceding section, the current section explores the classification of valence given conversational speech pre-segmented into utterances. Although the average κ value for emotionally relevant behavior is 0.58 (Table 15.3), not much lower than the κ achieved for valence of 0.67 when annotator $A7$ is ignored (Table 15.5), Table 15.2 indicates that only the distinction between `Info`, `AgrAck`, `Agr`, and `AgrOnly` may be computationally feasible, among the identified emotionally relevant behaviors. This appears no more useful, from the point of view of studying emotion, than is the output of laughter detection, already treated in Chapter 13.

	Negative		Neutral		Positive	
DiscntSli		3	−−	28	++	22
Discnt		9	−−	165	++	79
DisagrConf	++	14		275		71
Disagr		5	++	291	−−	45
Doubt		6	+	261	−	48
Other	+	5		68		30
Info		107	++	6001	−−	1319
AgrAck		6	++	471	−−	87
Agr		19	++	761	−−	167
PromEgo		6		120		26
AgrImprEst		3	−	135	+	64
AttAmuse		3	−−	229	++	416
LaughOnly		4	−−	200	++	288

Table 15.8: Adjacency of emotional valence as assigned by labeler 2 with precedent speaker's emotionally relevant behavior as assigned by labeler 3 for the ISL Meeting Corpus (11857 speaker contributions), absolute counts. + and − represent rejection of the null hypothesis of no association; notation as in Table 6.

The task of classifying per-utterance valence in the ISL Meeting Corpus was first conducted within a collaborative emotion recognition technology evaluation within the CHIL project in 2005[3].

15.4.1 Dataset Use

The dataset used is exactly the ISL Meeting Corpus, pre-segmented into speaker contributions and inclusive of the orthographic transcription, together with the 3-way valence labels {Negative, Neutral, Positive}, assigned during L3 annotation as described in the preceeding section.

The corpus has been divided into 3 subsets for the current purposes, TrainSet, DevSet, and EvalSet. TrainSet (of 3964 speaker contributions) consists of: m036 (999), m038 (183), m039a (143), m039b (121), m042 (382), m043 (166), m045 (536), m046 (493), m051 (490), m053 (451). DevSet (of 3849 speaker contributions) consists of: m055 (1771), m063 (758), m064 (1320). EvalSet (of 4664 speaker contributions), consists of: m035 (568), m048 (704), m052 (798), m054 (664), m057 (855), m061 (1075).

15.4.2 Assessment of Performance

Since classification is performed for each pre-segmented utterance, a natural choice of performance measure is classification accuracy, namely the proportion of utterances by number which are classified correctly. To supplement this, the unweighted F-score is computed for each valence class of interest.

15.4.3 Feature Assessment and Selection

The main impetus behind the CHIL 2005 Emotion Recognition Evaluation was the identification of discriminative features. It that pursuit, it was of paramount importance to be able to try a large number of features in different combinations. To render this tractable, a simple classifier structure was opted for, with known limitations, but whose training involved as few iterations as possible.

To meet this constraint, a linear regression classifier was chosen which requires exactly one analytic step for the inference of paramerers. The classification decision is simply $v^* = \mathbf{v}\,[c^*]$, where $\mathbf{v}$ is the ordered vector of possible valence

[3]The sites participating were UKA, under whose aegis this work was performed, and the other was KTH; the experiences of the latter have been published in [173]. A third site had worked on the visual detection of emotion, and this evaluation was therefore not applicable.

alternatives {NEGATIVE, NEUTRAL, POSITIVE} and

$$c^* \;=\; \arg\max_c \; \tilde{\mathbf{y}}\,[c] \;, \tag{15.1}$$

i.e., we selected that class $\mathbf{v}\,[c]$, $1{\leq}c{\leq}3$, whose regressed score $\tilde{\mathbf{y}}\,[c]$ was the highest. Each $\tilde{y}_c = \tilde{\mathbf{y}}\,[c]$, for a test utterance u whose feature vector is $\mathbf{x} \in \mathbb{R}^J$ of J features, and the true class is c_{ref}, was given by

$$\tilde{y}_c \;=\; \mathbf{x}^T\,\mathbf{w}_c \tag{15.2}$$

where $\mathbf{w}_c \in \mathbb{R}^J$ is the weight vector corresponding to the class $\mathbf{c}\,[c]$.

Training of the three weight vectors $\mathbf{w}_c$, $1{\leq}c{\leq}3$ is accomplished by minimizing the sum of squares of errors with respect to the encoding $y_c = \delta\,(\mathbf{v}\,[c]\,,\mathbf{v}\,[c_{ref}])$. Specifically, given a training set of I exemplars with feature vectors $\mathbf{x}_i$,

$$\mathbf{X} \;=\; \begin{bmatrix} \mathbf{x}_1\,[1] & \cdots & \mathbf{x}_1\,[j] & \cdots & \mathbf{x}_1\,[J] \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{x}_i\,[1] & \cdots & \mathbf{x}_i\,[j] & \cdots & \mathbf{x}_i\,[J] \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{x}_I\,[1] & \cdots & \mathbf{x}_I\,[j] & \cdots & \mathbf{x}_I\,[J] \end{bmatrix} \;\in\; \mathbb{R}^{I\times J} \;, \tag{15.3}$$

as well as the targets

$$\mathbf{Y} \;=\; \begin{bmatrix} \mathbf{y}_1\,[1] & \mathbf{y}_1\,[2] & \mathbf{y}_1\,[3] \\ \vdots & \vdots & \vdots \\ \mathbf{y}_i\,[1] & \mathbf{y}_i\,[2] & \mathbf{y}_i\,[3] \\ \vdots & \vdots & \vdots \\ \mathbf{y}_Y\,[1] & \mathbf{y}_I\,[2] & \mathbf{y}_I\,[3] \end{bmatrix} \;\in\; \{0,1\}^{I\times 3} \;. \tag{15.4}$$

The value of $\mathbf{W} = [\mathbf{w}_1\,\mathbf{w}_2\,\mathbf{w}_3] \in \mathbb{R}^{J\times 3}$ which yields a $\tilde{\mathbf{Y}} = \mathbf{X}\,\mathbf{W}$ closest to $\mathbf{Y}$, in a least-squares sense, is given by

$$\mathbf{W} \;=\; \left(\mathbf{X}^T\,\mathbf{X}\right)^{-1}\,\left(\mathbf{X}^T\,\mathbf{Y}\right) \;. \tag{15.5}$$

The simplicity of this operation makes it particularly easy to search for potentially useful features, using both exhaustive feature combination for small feature sets as well as sequential forward or backward one-feature-at-a-time selection. Feature selection was performed in two steps. First, a preliminary feature set was grown incrementally by maximizing the classification accuracy on TRAINSET and DEVSET pooled together, for each family of features considered. Features making it into this set are identified with a "◀" in the description which follows. Then, prior to the classification experiments, a subset of each feature set was grown incrementally by maximizing the classification accuracy on DEVSET, training classifiers on TRAINSET only. Feature set growth, in both steps, terminates when the next best feature leads to a decrease in accuracy; this occasionally leads to feature sets which are suboptimal (since a supremum may be reached later, beyond local minima during incremental search), but the resulting feature sets are smaller and potentially better at generalization.

15.4.4 Vocal Interaction Features

Unlike other places in this thesis, in which speech segmentation into words, dialog acts, utterances, or multi-frame units was not available, the utterance-pre-segmented nature of the current data makes possible the computation of duration features. For each utterance u to be classified, the following features — which describe the occurrence of vocal activity in the temporal neighborhood of u — were computed from the manual utterance segmentation of each meeting:

- `TDUR◀` utterance duration in seconds;
- `TSEPPREV` duration in seconds of own silence preceding u (max of 1 hour);
- `TSEPNEXT` duration in seconds of own silence following u (max of 1 hour);
- `TPAUSE0` duration in seconds of all-silence preceding u;

- `TPAUSE1` duration in seconds of all-silence following u;
- `NOVSPK` number of other participants producing utterances during the temporal support of u;
- `NOV` number of utterances produced by other participants during the temporal support of u;
- `TOV` total duration in seconds of overlap, with u, in utterances produced by other participants;
- `NSPKOV0` number of other participants producing utterances while u begins;
- `TDUROV0` time in seconds from beginning of u until latest end among those utterances which are being produced while u begins;
- `NSPKOV1` number of other participants producing utterances while u ends;
- `TDUROV1` time in seconds from earliest beginning among those utterances which are being produced when u ends until end of u;
- `NVICOSPK` number of other participants producing utterances in the 5 seconds preceding u (excluding those utterances exhibiting overlap with u); and
- `NVIC1SPK` number of other participants producing utterances in the 5 seconds following u (excluding those utterances exhibiting overlap with u).
- `nLAUGH` number of `<Laugh>` tokens transcribed in the utterance

15.4.5 Contrastive Spectral Features

To contrast performance using vocal interaction as an information source, features representing the spectral characteristics of each utterance are also computed. This is done at the frame level, since utterances are of arbitrary duration but a fixed-length vector is most attractive from the point of view of classifier training. The process of computing spectral features consists of: (1) framing each utterance u into a sequence of frames; (2) computing the magnitude frequency spectrum for each frame; (3) transforming each per-frame spectrum into its truncated Mel-frequency cepstral representation, used ubiquitously in speech processing; (4) transforming it back to its 40 *reconstructed* Mel filterbank outputs; and (5) computing per-utterance summary features over all frames[4]. A frame step of 10 ms was employed. Finally, summary features consist of the mean, range, variance, standard deviation, skewness, and kurtosis.

The process yields 240 features for every utterance u, shown in Table 15.9. Feature selection within this set identified 29 features, denoted with a "◀" in the table. This set includes mean and measures of spread from what appears to be uniform subset of the Mel filterbank outputs; selected measures of skewness appear to be drawn from only the low-frequency Mel filters, while those of kurtosis appear to be drawn from only the high-frequency filters.

15.4.6 Contrastive Prosodic Features

The computation of proposed prosodic features, also for contrast, is significantly more involved than that of the spectral features in the previous section. This is due in part to the fact that what is thought to be of importance prosodically are not the absolute targets which feature trajectories achieve, but their evolution in time. Because utterances may be of arbitrary length, however, the temporal units across which variation should be measured or modeled are not obvious. Second, pitch, an important prosodic variable, is discontinuous over the course of speech: much as utterances are internally fragmented due to inter-talkspurt pauses, talkspurts are internally fragmented by occlusions in the quasiperiodicity of speech.

The process by which the proposed prosodic features are computed is shown in some detail in Figure 15.6. As the diagram shows, features are computed at the frame level, at the voicing segment (VSeg) level, at the sample level, at the quasisyllable level, and at the utterance level. Features computed for units shorter than the utterance must be subsequently combined to form utterance-level descriptors. This is described, separately for each step shown in the figure, below.

STEP 1a: Framing and Computation of Frame-Level Features

[4]To perform these operations, Malcolm Slaney's AuditoryToolbox in MATLAB was used.

#	Mean	Range	Variance	StDev	Skewness	Kurtosis
0	mel00MEAN◀	mel00RNGE◀	mel00STD	mel00VAR◀	mel00SKEW◀	mel00KURT
1	mel01MEAN	mel01RNGE	mel01STD	mel01VAR	mel01SKEW◀	mel01KURT
2	mel02MEAN	mel02RNGE	mel02STD	mel02VAR	mel02SKEW	mel02KURT
3	mel03MEAN	mel03RNGE	mel03STD	mel03VAR	mel03SKEW◀	mel03KURT
4	mel04MEAN	mel04RNGE	mel04STD	mel04VAR	mel04SKEW	mel04KURT
5	mel05MEAN	mel05RNGE	mel05STD	mel05VAR	mel05SKEW	mel05KURT
6	mel06MEAN	mel06RNGE	mel06STD	mel06VAR	mel06SKEW	mel06KURT
7	mel07MEAN	mel07RNGE	mel07STD	mel07VAR	mel07SKEW◀	mel07KURT
8	mel08MEAN	mel08RNGE	mel08STD	mel08VAR	mel08SKEW	mel08KURT
9	mel09MEAN	mel09RNGE	mel09STD	mel09VAR	mel09SKEW◀	mel09KURT
10	mel10MEAN◀	mel10RNGE◀	mel10STD◀	mel10VAR◀	mel10SKEW	mel10KURT
11	mel11MEAN	mel11RNGE	mel11STD	mel11VAR	mel11SKEW	mel11KURT
12	mel12MEAN	mel12RNGE	mel12STD	mel12VAR	mel12SKEW	mel12KURT
13	mel13MEAN	mel13RNGE	mel13STD	mel13VAR	mel13SKEW	mel13KURT
14	mel14MEAN	mel14RNGE	mel14STD	mel14VAR	mel14SKEW	mel14KURT
15	mel15MEAN	mel15RNGE	mel15STD	mel15VAR	mel15SKEW	mel15KURT
16	mel16MEAN	mel16RNGE	mel16STD◀	mel16VAR	mel16SKEW	mel16KURT
17	mel17MEAN	mel17RNGE◀	mel17STD	mel17VAR◀	mel17SKEW	mel17KURT
18	mel18MEAN	mel18RNGE	mel18STD	mel18VAR	mel18SKEW	mel18KURT
19	mel19MEAN	mel19RNGE	mel19STD	mel19VAR	mel19SKEW	mel19KURT
20	mel20MEAN◀	mel20RNGE	mel20STD	mel20VAR	mel20SKEW	mel20KURT
21	mel21MEAN	mel21RNGE	mel21STD	mel21VAR	mel21SKEW	mel21KURT
22	mel22MEAN	mel22RNGE	mel22STD	mel22VAR	mel22SKEW	mel22KURT
23	mel23MEAN	mel23RNGE	mel23STD	mel23VAR	mel23SKEW	mel23KURT
24	mel24MEAN	mel24RNGE	mel24STD	mel24VAR	mel24SKEW	mel24KURT
25	mel25MEAN	mel25RNGE◀	mel25STD	mel25VAR◀	mel25SKEW	mel25KURT
26	mel26MEAN◀	mel26RNGE	mel26STD	mel26VAR	mel26SKEW	mel26KURT
27	mel27MEAN	mel27RNGE	mel27STD	mel27VAR◀	mel27SKEW	mel27KURT
28	mel28MEAN	mel28RNGE	mel28STD◀	mel28VAR	mel28SKEW	mel28KURT
29	mel29MEAN	mel29RNGE	mel29STD	mel29VAR	mel29SKEW	mel29KURT
30	mel30MEAN	mel30RNGE	mel30STD	mel30VAR◀	mel30SKEW	mel30KURT
31	mel31MEAN	mel31RNGE	mel31STD	mel31VAR	mel31SKEW	mel31KURT
32	mel32MEAN	mel32RNGE◀	mel32STD	mel32VAR	mel32SKEW	mel32KURT
33	mel33MEAN	mel33RNGE	mel33STD	mel33VAR	mel33SKEW	mel33KURT◀
34	mel34MEAN	mel34RNGE	mel34STD	mel34VAR◀	mel34SKEW	mel34KURT◀
35	mel35MEAN◀	mel35RNGE	mel35STD	mel35VAR	mel35SKEW	mel35KURT◀
36	mel36MEAN	mel36RNGE	mel36STD	mel36VAR	mel36SKEW	mel36KURT
37	mel37MEAN	mel37RNGE	mel37STD	mel37VAR◀	mel37SKEW	mel37KURT
38	mel38MEAN	mel38RNGE	mel38STD	mel38VAR	mel38SKEW	mel38KURT
39	mel39MEAN	mel39RNGE	mel39STD	mel39VAR	mel39SKEW	mel39KURT

Table 15.9: Spectral utterance-level features formed by computing moments (and range) over quantities estimated for individual frames. "StDev" is standard deviation; "◀" indicates features which were automatically selected given all spectral features.

A frequently used pitch tracker, `get_f0`[5], was used to obtain for each single-channel audio snippet, corresponding to an utterance u, a sequence of 3 estimates per frame. These consist of: (1) root-mean-square (RMS) energy (henceforth ENERGY); (2) the magnitude of the first off-axis peak in the frame's autocorrelation spectrum, normalized by the magnitude of the autocorrelation spectrum at $\tau = 0$ (henceforth ACORR); and (3) the F0 estimate in Hz (henceforth, F0), when ACORR exceeds an internal threshold and zero otherwise.

In addition to these quantities, also at a frame step of 10 ms, an attempt was made to characterize the degree of openness of the pharynx and lips by computing a family of features referred to here as *vocal volume features* (henceforth VOCVOL). Computation assumes the lossless tube model of speech production, whose per-frame parameters are estimated using linear prediction. The volume of the first 3 tube sections is taken to represent the pharynx, while that of the last 3 tube sections is taken to represent the lips. For completion, the total volume given all 18 tube sections is also computed, as well as its first- and second-order differences. These features may be more correctly characterized as spectral features, but were included in this section as a result of architectural and implementation decisions.

Computation of ENERGY, ACORR, F0, and the five VOCVOL features yields the object *FrameFeatSeq* in Figure 15.6, a temporal sequence of 8-element feature vectors.

[5]The remnant `get_f0` code found in Snack 2.2.2, available from `http://www.speech.kth.se/snack/`, was linked against in the described implementation.

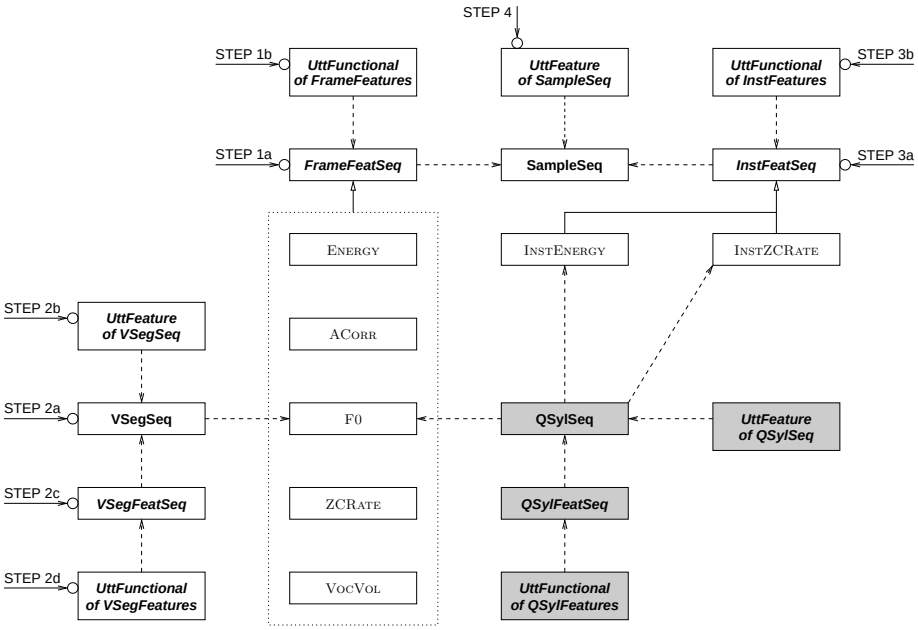


Figure 15.6: Objects involved in computing the proposed prosodic features, as in the UKA CEICES System. Acronyms as described in the text. Dashed lines connecting any two objects point towards the one depended on; solid lines terminating on circles indicate object creation; other solid lines connecting any two objects indicate composition and point from subset to superset. Gray objects not described in this thesis, as none of the features computed using them were found beneficial for the current task.

STEP 1b: Computing Utterance-Level Functionals of Frame-Level Features

As for spectral features, to produce utterance-level descriptors whose number is not a function of utterance duration, summary functionals of the *FrameFeatSeq* quantities are computed. Those for which moments (and range) were computed are shown in Table 15.10; LogF0 was computed using

$$\text{LogF0}\,[i] \quad = \quad \log_2 \text{F0}\,[i] \ , \tag{15.6}$$

where i is the frame index. Both Energy and ACorr were also broken down into two non-overlapping sets, as shown in the table, namely those which exhibited voicing and those which did not. Feature selection of the resulting 48 features identified 11 features; the selection process ignored functionals over LogF0, preferring those over the raw F0 instead, and identified only one functional over exclusively unvoiced frames (`euVAR`).

An alternative representation of distribution is that using quartile bounds, and these were applied to several of the *FrameFeatSeq* features as shown in Table 15.11. Energy was additionally normalized by the maximum observed within the utterance,

$$\text{NormEnergy}\,[i] \quad = \quad \max_j \text{Energy}\,[j] - \text{Energy}\,[i] \tag{15.7}$$

where i is the within-utterance frame index. As it turned out (for this task), none of the quartile functionals of NormEnergy were identified as useful by feature selection, nor were any of the VocVol functionals. Feature selection identified only 9 features from the 55 features in Table 15.11, two of which described the distribution of energy in unvoiced frames.

The ad hoc application of specific functional types, in Tables 15.10 and 15.11, to specific frame-level feature types is due to the fact that these functionals and features were developed at disparate times over the course of multiple projects.

The presence of voicing is used to compute several additional simple functionals, namely:

- `rVOI` ratio of the number of voiced frames to the number of frames;

Feature	Mean	Range	StDev	Variance	Skewness	Kurtosis
ENERGY	eMEAN	eRNGE	eSDEV	eVAR	eSKEW	eKURT
voiced	evMEAN	evRNGE	evSDEV	evVAR◀	evSKEW◀	evKURT◀
unvoiced	euMEAN	euRNGE	euSDEV	euVAR◀	euSKEW	euKURT
ACORR	aMEAN	aRNGE◀	aSDEV◀	aVAR◀	aSKEW◀	aKURT
voiced	avMEAN	avRNGE	avSDEV	avVAR◀	avSKEW	avKURT
unvoiced	auMEAN	auRNGE	auSDEV	auVAR	auSKEW	auKURT
F0	pMEAN	pRNGE◀	pSDEV	pVAR	pSKEW◀	pKURT
LOGF0	oMEAN	oRNGE	oSDEV	oVAR	pSKEW	oKURT

Table 15.10: Prosodic utterance-level frame-shuffle-invariant moments and range of frame-level ENERGY, ACORR, F0 (and the derived LOGF0) features, and of frame-level and voicing-striated ENERGY and ACORR features. "◀" indicates features which were automatically selected.

Feature	Minimum	P25	P50	P75	Maximum
ENERGY	eMIN	eP25	eP50	eP75	eMAX
voiced	evMIN◀	evP25	evP50	evP75	evMAX◀
unvoiced	euMIN	euP25	euP50◀	euP75◀	euMAX
NORMENERGY	enMIN	enP25	enP50	enP75	—
ACORR	aMIN◀	aP25	aP50	aP75	aMAX
F0	pMIN◀	pP25	pP50◀	pP75◀	pMAX◀
VOCVOL	volMIN	volP25	volP50	volP75	volMAX
1-st order diff.	dvolMIN	dvolP25	dvolP50	dvolP75	dvolMAX
2-nd order diff.	ddvolMIN	ddvolP25	ddvolP50	ddvolP75	ddvolMAX
at lips	mthMIN	mthP25	mthP50	mthP75	mthMAX
at pharynx	phaMIN	phaP25	phaP50	phaP75	phaMAX

Table 15.11: Prosodic utterance-level frame-shuffle-invariant quartile bounds of frame-level ENERGY, (and the derived LOGENERGY), ACORR, F0, and VOCVOL features, and of frame-level and voicing-striated ENERGY features. "◀" indicates features which were automatically selected.

- `rUNV` ratio of the number of unvoiced frames to the number of frames;
- `pDURVOI` total duration in seconds of voiced frames;
- `pDURUNV` total duration in seconds of unvoiced frames;
- `rDURVOIUNV` ratio of the duration of voiced frames to the number of unvoiced frames, or 10 (a large number) if all frames are voiced;
- `pT0◀` duration in seconds of interval from beginning of u to first voiced frame in u;
- `pT1` duration in seconds of interval from beginning of u to last voiced frame in u;

Only `pT0` is selected by the feature selection procedure.

These functionals (like those in Tables 15.10, 15.11, and 15.15) are invariant under frame reshuffling, and therefore do not capture anything about the shape of feature trajectories in time. To model feature trajectory information, the following four simple features are computed:

- `eMINt` duration in seconds of interval from beginning of u to voiced frame of lowest ENERGY in u;
- `eMAXt` duration in seconds of interval from beginning of u to voiced frame of highest ENERGY in u;
- `pMINt` duration in seconds of interval from beginning of u to voiced frame of lowest F0 in u;
- `pMAXt◀` duration in seconds of interval from beginning of u to voiced frame of highest F0 in u;

Of these, feature selection identifies only `pMAXt` as useful.

Several somewhat more involved functionals which are computed and which also capture trajectory information include linear regression for frame-level features, inter-feature correlations, and linear prediction coefficients. Linear regressors are computed for ENERGY, ACORR, and F0; both regression coefficients and mean-square errors are used as features, shown in Table 15.12, for the entire utterance, its onset, and its coda. These functionals, particularly over ACORR — a correlate of the probability of voicing — appear to be useful to the proposed task, as all 6 ACORR functionals are automatically selected. Feature selection also identifies two additional functionals as useful, namely the global mean-square errors for ENERGY and for F0.

Feature	Global		Onset		Coda	
	Slope	Error	Slope	Error	Slope	error
ENERGY	eREGR	eMSE◀	eREGR_b	eMSE_b	eREGR_e	eMSE_e
ACORR	aREGR◀	aMSE◀	aREGR_b◀	aMSE_b◀	aREGR_e◀	aMSE_e◀
F0	pREGR	pMSE◀	pREGR_b	pMSE_b	pREGR_e	pMSE_e

Table 15.12: Prosodic utterance-level frame-shuffle-non-invariant linear regression coefficients and mean-square errors for frame-level ENERGY, ACORR, and F0 features. "Global" indicates that each utterance in its entirety was used for linear regression; "Onset" indicates the first third of the total duration of the utterance, while "Coda" indicates the last third of its total duration. "◀" indicates features which were automatically selected.

Inter-feature correlations, shown in Table 15.13, were computed between ENERGY, ACORR, and F0, as well as between each of these features and the frame index. Some of these features are not independent of those in Table 15.12. Only 4 out of a total of 18 were automatically selected as beneficial.

Feature Pair			Global	Onset	Coda
TIME	vs	ENERGY	teXCORR	teXCORR_b	teXCORR_e
TIME	vs	ACORR	taXCORR	taXCORR_b◀	taXCORR_e◀
TIME	vs	F0	tpXCORR	tpXCORR_b◀	tpXCORR_e
ENERGY	vs	ACORR	eaXCORR	eaXCORR_b	eaXCORR_e
ENERGY	vs	F0	epXCORR◀	epXCORR_b	epXCORR_e
ACORR	vs	F0	paXCORR	paXCORR_b	paXCORR_e

Table 15.13: Prosodic utterance-level frame-shuffle-non-invariant linear correlation coefficients between ENERGY, ACORR, F0, and the frame index (TIME). "Global" indicates that each utterance in its entirety was used for linear regression; "Onset" indicates the first third of the total duration of the utterance, while "Coda" indicates the last third of its total duration. "◀" indicates features which were automatically selected.

Finally, 8th-order LPC coefficients were computed for ENERGY and ACORR. Of these 16 features, shown in Table 15.14, feature selection identified 13 as beneficial.

Feature	LPC Coefficients							
ENERGY	eLPC0◀	eLPC1◀	eLPC2	eLPC3◀	eLPC4◀	eLPC5	eLPC6◀	eLPC7◀
ACORR	aLPC0◀	aLPC1◀	aLPC2◀	aLPC3◀	aLPC4◀	aLPC5	aLPC6◀	aLPC7◀

Table 15.14: Prosodic utterance-level frame-shuffle-non-invariant 8th-order linear prediction coefficients for ENERGY and ACORR. "◀" indicates features which were automatically selected.

STEP 2a: Segmenting Utterances into Alternating Voicing Segments (VSegs)

The binary estimate of voicing per frame, implicit in the F0 features, allows for a segmentation of each utterance u into alternating voiced and unvoiced segments (VSeg). To simplify subsequent processing, the alternating segments must be longer than 3 frames in length.

STEP 2b: Computing Utterance-Level Features from the VSeg Sequence

Several utterance-level features are computed directly from the VSeg sequence. These include:

- `nVOI` number of voiced intervals in u;
- `nUNV` number of unvoiced intervals in u.

Neither of these features is adopted, using automatic feature selection, for the proposed task.

STEP 2c: Computing VSeg-Level Features

For each VSeg, only its duration is computed.

STEP 2d: Computing Utterance-Level Functionals of VSeg-Level Features

From the VSeg sequence of duration features, four quantities are computed:

- `pDURVOImax`◀ duration in seconds of longest voiced interval;
- `pDURUNVmax`◀ duration in seconds of longest unvoiced interval;
- `pDURVOIave` average number of frames per voiced interval;
- `pDURUNVave` average number of frames per unvoiced interval;
- `rNVOIUNV`◀ ratio of the number of voiced segments (at least 3 frames in duration) to the number of unvoiced segments (at least 3 frames in duration), or 10 (a large number) if the latter is zero.

Automatic feature selection selected three of these features, as indicated with ◀.

STEP 3a: Computing Instantaneous (Sample-Level) Features

Two features are computed which are sample-by-sample transformations of the original audio signal $x[t]$. The latter is first passed through a DC notch filter,

$$y[t] = x[t] - 0.999x[t-1] + y[t-1] \ . \tag{15.8}$$

Then, the instantaneous energy InstEnergy and instantaneous zero-crossing rate InstZCRate are computed using

$$\text{InstEnergy}[t] = y[t] \cdot y[t] \tag{15.9}$$

$$\text{InstZCRate}[t] = 1 - \delta(\text{sgn}(y[t-1]), \text{sgn}(y[t])) \ , \tag{15.10}$$

where $\text{sgn}(\cdot)$ yields the sign of its argument and $\delta(\cdot)$ is the Kronecker delta. These instantaneous signals are then smoothed with a Hamming filter whose width is 0.020 s.

STEP 3b: Computing Utterance-Level Functionals of Sample-Level Features

Computed from the smoothed InstEnergy and InstZCRate trajectories are their utterance minima (`eMINs` and `zMINs`, respectively) and their utterance maxima (`eMAXs` and `zMAXs`, respectively). Also computed is a 2-dimensional, 4×4 histogram over both features jointly, together with 1-dimensional marginals, which is shown in Table 15.15. The cell sample counts are normalized by the number of 16 kHz samples in the utterance. Feature selection identifies 7 of these features as beneficial, but ignores all four extrema `eMINs`, `zMINs`, `eMAXs`, and `zMAXs`.

InstEnergy	InstZCRate				Marg
	Bin 1 of 4	Bin 2 of 4	Bin 3 of 4	Bin 4 of 4	
Bin 1	eH14zH14s	eH14zH24s	eH14zH34s◀	eH14zH44s	eH14s
Bin 2	eH24zH14s◀	eH24zH24s◀	eH24zH34s◀	eH24zH44s	eH24s
Bin 3	eH34zH14s	eH34zH24s◀	eH34zH34s	eH34zH44s◀	eH34s
Bin 4	eH44zH14s◀	eH44zH24s	eH44zH34s	eH44zH44s	—
Marg	zH14s	zH24s	zH34s	—	

Table 15.15: Prosodic utterance-level frame-shuffle-invariant non-parametric probability distribution functionals of sample-level InstEnergy and InstZCRate features. "Marg" indicates marginals; "◀" indicates features which were automatically selected.

STEP 4: Computing Utterance-Level Features

Last of all, several features are computed which rely on all of the audio in an utterance u, but lead to a single quantity. These features comprise mainly voice quality measures which require long temporal spans for estimation, and include jitter, shimmer, and harmonicity measures. Jitter describes average departure from a smoothly evolving pitch period; the measures computed were those available within Praat[6] [21]. The five jitter variables computed are:

- `JITLOC` Praat, `PointProcess_getJitter_local()`;
- `JITLOCABS`◀ Praat, `PointProcess_getJitter_local_absolute()`;
- `JITRAP` Praat, `PointProcess_getJitter_rap()`;
- `JITPPQ5`◀ Praat, `PointProcess_getJitter_ppq5()`; and
- `JITDDP` Praat, `PointProcess_getJitter_ddp()`.

The second type of feature considered in this group are shimmer features, which describe the average departure from a smoothly evolving amplitude of the glottal peak sequence[7]. Six features are computed:

- `SHIMLOC`◀ Praat, `AmplitudeTier_getShimmer_local()`;
- `SHILOCDB` Praat, `AmplitudeTier_getShimmer_local_dB()`;
- `SHIMAPQ3`◀ Praat, `AmplitudeTier_getShimmer_apq3()`;
- `SHIMAPQ5`◀ Praat, `AmplitudeTier_getShimmer_apq5()`;
- `SHIMAPQ11` Praat, `AmplitudeTier_getShimmer_apq11()`; and
- `SHIMDDA`◀ Praat, `AmplitudeTier_getShimmer_dda()`.

Finally, three harmonicity measures are computed, which characterize the amount of energy in the harmonics of voiced speech as opposed to that at other frequencies[8]:

- `harmACORR` Praat, `Pitch_STRENGTH_UNIT_AUTOCORRELATION`;
- `harmNHR` Praat, `Pitch_STRENGTH_UNIT_NOISE_HARMONICS_RATIO`; and
- `harmHNR`◀ Praat, `Pitch_STRENGTH_UNIT_HARMONICS_NOISE_DB`.

As shown with a "◀", automatic feature selection identified 7 of these 14 voice quality features as beneficial for the proposed task.

[6]The Praat functions rely on a native `PointProcess_Struct` structure which encodes glottal peak instants. These were computed by inverse filtering the audio and then calling two ESPS functions, `mask()` and `epochs`. At the time these experiments were performed, the ESPS code was not available in its entirety and these functions were partly re-implemented using standard numerical algorithm implementations from [184].

[7]As for shimmer, the computation of these features involved a mixture of Praat code [21], ESPS code, and NRC code [184].

[8]These features are computed using the Praat function `Pitch_getMeanStrength()`, which requires population of the `Pitch_Struct` structure. Its `frequency` and `strength` members are assigned the F0 and ACorr values produced by a re-engineered implementation of ESPS `get_f0`.

15.4.7 Contrastive Lexical Features

In addition to spectral and prosodic features, a set of simple lexical features were considered for discrimination among NEGATIVE, NEUTRAL, and POSITIVE valence. These were extracted from the human-produced orthographic transcription. All utterance transcriptions were first normalized to be upper-case, with minimal mapping of disfluencies. Following this, TRAINSET data for the three classes were used to construct unigram, bigram, and trigram models for each[9]. The lexical feature vector for each test utterance consisted of 10 features:

- `NLEX◀` number of lexically transcribed events, including words, word fragments, and noises (plus markup resulting from disfluency normalization)
- `GRAM(1)_NEG◀` unigram perplexity given NEGATIVE utterances in TRAINSET
- `GRAM(1)_0◀` unigram perplexity given NEUTRAL utterances in TRAINSET
- `GRAM(1)_POS◀` unigram perplexity given POSITIVE utterances in TRAINSET
- `GRAM(2)_NEG◀` bigram perplexity given NEGATIVE utterances in TRAINSET
- `GRAM(2)_0◀` bigram perplexity given NEUTRAL utterances in TRAINSET
- `GRAM(2)_POS◀` bigram perplexity given POSITIVE utterances in TRAINSET
- `GRAM(3)_NEG◀` trigram perplexity given NEGATIVE utterances in TRAINSET
- `GRAM(3)_0◀` trigram perplexity given NEUTRAL utterances in TRAINSET
- `GRAM(3)_POS◀` trigram perplexity given POSITIVE utterances in TRAINSET

It should be noted that the lexical features are computed without discarding transcribed nonverbal vocalizations; in particular, perplexities reflect the presence of `<Laugh>`, something already captured in the vocal interaction feature set.

15.4.8 Feature Comparison and Combination

Classification experiments were performed for the meetings in DEVSET, for each of the feature sets described, in conjunction with feature selection.

The system relying only on spectral features is denoted SPEC. As shown in Table 15.9, automatic feature selection had produced a feature set of 29 features, from the total available of 240, when accuracy was maximized for TRAINSET and DEVSET pooled together. Subsequently maximizing accuracy on DEVSET, while training exclusively on TRAINSET, led to only 11 features (`mel00SKEW`, `mel20MEAN`, `mel10MEAN`, `mel00RNGE`, `mel16STD`, `mel34VAR`, `mel35MEAN`, `mel09SKEW`, `mel17VAR`, `mel35KURT`, and `mel17RNGE`, in that order). On DEVSET, this system achieved an accuracy of 82.98%, 2.19%abs higher than merely always guessing the majority class (NEUTRAL valence). Precision, recall, and F-score achieved with this system are shown in Table 15.16.

The prosodic system, denoted PROS, considered all 197 features described in Subsection 15.4.6. The first feature selection step identified 64 features. The second feature selection step, maximizing only DEVSET accuracy, produced a set of only 16 features (`evVAR`, `pMAX`, `harmHNR`, `pP50`, `evMAX`, `pDURUNVmax`, `eH24zH24`, `euP75`, `euP50`, `JITPPQ5`, `eH24zH14`, `evSKEW`, `evKURT`, `aREGR_e`, `taXCORR_e`, and `rNVOIUNV`, in that order). PROS achieves an accuracy of 82.53% on DEVSET, a 1.74%abs improvement over majority class guessing.

Feature space combination of the SPEC and PROS systems yielded a system denoted SPEC+PROS. Pooling the initially selected 29 SPEC features and 64 PROS features, and then performing the second feature selection step to maximize accuracy on DEVSET, yielded 16 features (`evVAR`, `mel00SKEW`, `pMAX`, `mel20MEAN`, `eH24zH24`, `mel10MEAN`, `mel00RNGE`, `aVAR`, `TDUR`, `mel17VAR`, `JITPPQ5`, `pDURVOImax`, `eH24zH14`, `mel35MEAN`, `eLPC3`, and `pP50`, in that order). This combined system

[9] NGram models and subsequent perplexity were computed using the `clausi` Language Modeling Toolkit, sometimes referred to as CLAUSI-LM. This toolkit was developed by Klaus Ries, but was never unambiguously versioned; the version used here is that widely in use in the year 2002 at the (then) Interactive Systems Labs. The models were trained using Katz-style Good-Turing discounting, Kneser-Ney backoff, and cutoff constraints as follows:

- unigram: `ngrammodel -n 1 -GT 1 -cutoff 1 3 X`;
- bigram: `ngrammodel -n 2 -GT 2 -GT 1 -cutoff 2 3`; and
- trigram: `ngrammodel -n 3 -GT 3 -GT 2 -GT 1 -cutoff 3 3 -cutoff 2 10`

Feature Set	Acc, %	Neutral			Positive		
		P,%	R,%	F,%	P,%	R,%	F,%
guess (uniform)	33.33	80.77	33.33	47.19	15.75	33.33	21.39
*guess (*DevSet*)*	67.85	80.78	80.78	80.78	15.70	15.70	15.70
guess (majority)	80.79	80.79	100	89.37	0	0	0
Spec	82.98	82.96	99.43	90.45	83.47	16.84	28.03
Pros	82.53	82.66	99.26	90.20	78.21	14.89	25.02
Spec+Pros	84.55	84.75	98.72	91.20	81.34	30.51	44.38
Lex	88.16	90.65	95.36	92.94	74.08	70.71	72.36
`nLAUGH` only	91.08	92.91	96.40	94.63	81.57	83.89	82.71
Lex+`nLAUGH`	91.31	92.92	96.70	94.78	82.80	83.81	83.30
All	91.18	92.92	96.53	94.69	82.09	83.89	82.98
All–`nLAUGH`	87.41	88.45	97.28	92.65	79.10	56.06	65.62

Table 15.16: DevSet results for 3-way classification of per-utterance emotional valence into one of Negative, Neutral, and Positive. Shown are overall classification accuracy ("Acc"), as well as precision ("P"), recall ("R"), and unweighted F-score ("F") for each of the three classes.

achieves an accuracy of 84.55%, 3.76%abs higher than guessing the majority class. The increase was expected to be $2.19 + 1.74 = 3.93$%abs, and suggests that the Spec and Pros feature sets are very complementary.

The lexical feature system, Lex, was exposed to 1 word count feature and 9 perplexity features, and all features were included by both the first and the second step feature selection procedures. Lex achieves a DevSet accuracy of 88.16%, which is far higher than the accuracy obtained with only acoustic features.

Only two features had been automatically selected for inclusion in the VocInt system. These achieve a DevSet accuracy of 91.12%. However, only one feature, namely the number of laughter tokens `nLAUGH`, achieves an accuracy of 91.08%. This is better than the performance achieved by `Lex`, in which language models were trained over an orthographic transcription which included `<Laugh>` tokens. A combination of Lex and the single `nLAUGH` feature, denoted Lex+`nLAUGH`, yields an accuracy of 91.31%, slightly better than `nLAUGH` alone.

The combination of all feature families (yielding the system All), in which feature selection identifies `nLAUGH`, `TDUR`, `pRNGE`, `GRAM(3)_0`, `GRAM(2)_POS`, `mel00VAR`, `NLEX`, `GRAM(3)_NEG`, and `eH24zH34`, yields 91.18%, which is slightly lower than Lex+`nLAUGH`. Excluding `nLAUGH` from this set yields an accuracy of 87.41%, lower than Lex.

These experiments suggest that, for the classification of utterance valence as annotated by "naive" labelers in the ISL Meeting Corpus, the presence of laughter is overwhelmingly the single most important feature. It reduces the error rate incurred by a system which always assigns the majority class (19.21%) by 54%rel. Other vocal interaction features, as explored here, have negligible or negative impact. The simple lexical features, when perfect transcription is available, achieve a reduction of 38%rel. In these cases, feature-space combination with acoustically derived features hurts performace. However, when lexical information is not available, acoustic features drawn from the manually pre-segmented utterances achieve a reduction of 20%rel.

15.4.9 Generalization to Unseen Data

The experiments shown in Table 15.16 for DevSet, are replicated for unseen EvalSet data, without additional feature set selection or other tuning, in Table 15.17. What is most surprising about this table is that the trends observed for DevSet and EvalSet are nearly identical, and all the observations made with regard to DevSet appear to hold for unseen data also.

15.4.10 Analysis

The above experiments indicate that a single feature, representing the number of instances of laughter found in the orthographic transcription of u, is responsible for the observed relative reduction in classification error. The results

Feature Set	Acc %	NEUTRAL			POSITIVE		
		P,%	R,%	F,%	P,%	R,%	F,%
guess (uniform)	33.33	78.70	33.33	46.83	18.14	33.41	23.51
*guess (*DEVSET*)*	66.51	78.66	80.78	79.71	18.12	15.76	16.86
guess (majority)	78.67	78.67	100	88.06	0	0	0
SPEC	81.50	81.27	99.45	89.45	87.36	18.01	29.86
PROS	81.54	81.33	99.35	89.44	86.81	18.72	30.80
SPEC+PROS	84.05	83.73	99.05	90.75	88.27	33.89	48.97
LEX	88.40	90.32	95.67	92.92	78.79	72.63	75.59
`nLAUGH` only	91.25	92.69	96.81	94.71	84.62	83.41	84.01
LEX+`nLAUGH`	91.34	92.49	97.06	94.72	85.87	82.82	84.32
ALL	91.32	92.60	96.89	94.70	85.33	83.41	84.36
ALL–`nLAUGH`	87.07	87.31	97.87	92.29	85.30	55.69	67.38

Table 15.17: EVALSET results for 3-way classification of per-utterance emotional valence into one of NEGATIVE, NEUTRAL, and POSITIVE. Shown are overall classification accuracy ("Acc"), as well as precision ("P"), recall ("R"), and unweighted F-score ("F") for each of the three classes.

Valence Class	Proportion/Number of Utterances			
	All	only `<LAUGH>`	incl `<LAUGH>`	no `<LAUGH>`
NEGATIVE (%)	3.24	0	1.32	3.65
NEUTRAL (%)	78.67	23.21	14.06	92.69
POSITIVE (%)	18.10	76.79	84.62	3.65
Total (#)	4720	56	832	3832

Table 15.18: Breakdown of utterances in EVALSET, into those consisting of only `<Laugh>`, those including `<Laugh>` but also other tokens, and those not containing `<Laugh>`.

suggest that other features, notably those not relying on human transcription (SPEC and PROS), are identifying some of the utterances which contain laugher, but not others, thereby yielding much worse performance. It is not immediately clear whether these feature sets are also identifying those POSITIVE utterances which do not contain laughter. This subsection attempts to answer that question.

Table 15.18 shows the breakdown of TRAINSET into three categories, namely those utterances which consist of `<Laugh>` only, those which contain `<Laugh>` as well as other words, and those which do not contain `<Laugh>`.

To analyze performance for these three subsets of EVALSET, a different classifier is trained (a C4.5 decision tree[10]); the same feature sets are retained. When a single classifier is trained using all utterances in TRAINSET for each feature set, performance is as shown in Table 15.19

What can be seen in Table 15.19 is that, first of all, accuracies in the second column are not identical but similar to those in Table 15.17, indicating that the linear regression classifier is competitive with standard classification paradigms on this data. More importantly, as the *guess (majority)* line shows, accuracies within each of the three groups of utterances are low, when laughter is present, because the majority class when all the data is considered is actually in the minority in these two groups. Almost every feature set does better than guessing the majority. Even feature sets that do not contain the `nLAUGH` feature, or do not model `<Laugh>` internally as the LEX features do, perform much better than guessing that valence is NEUTRAL. This indicates that these features, and the acoustic features in SPEC and PROS in particular, may actually be identifying some acoustic properties of laughter.

In a second and final analysis experiment, three different classifiers are trained using TRAINSET for every feature set,

[10]Release 8 of the popular C4.5 algorithm, courtesy of Ross Quinlan, in available from `http://www.rulequest.com/Personal/c4.5r8.tar.gz` (last downloaded 28 July 2010).

Feature Set	All		only <LAUGH>		incl <LAUGH>		no <LAUGH>	
	Acc %	$F(+)$ %	Acc %	$F(+)$ %	Acc %	$F(+)$ %	Acc %	$F(+)$ %
guess (majority)	78.67	NaN	23.21	NaN	14.06	NaN	**92.69**	NaN
SPEC	80.27	47.17	37.50	36.36	49.76	60.69	86.90	**7.84**
PROS	81.93	49.39	46.43	54.55	51.56	62.77	88.52	5.56
SPEC+PROS	80.55	47.92	53.57	60.61	51.92	63.15	86.77	8.33
LEX	89.73	79.23	73.21	84.54	82.57	90.00	91.28	4.93
VOCINT - nLAUGH	78.62	11.69	23.21	4.44	19.59	13.21	91.44	5.97
nLAUGH only	**91.25**	84.01	**76.79**	**86.87**	84.62	91.67	—	—
VOCINT	91.17	**84.19**	**76.79**	**86.87**	81.01	89.14	92.46	0.00
ALL	90.67	83.16	**76.79**	**86.87**	**85.22**	**92.18**	92.20	7.55

Table 15.19: Accuracies, and F-scores for the POSITIVE valence class, when, for each feature set, a single classifier is trained using all the utterances in TRAINSET.

one classifier for each category of utterance shown in Table 15.18. The results are shown in Table 15.20; the first two columns are the same as in Table 15.19. What can be seen in this table is that for those utterances that contain only <Laugh>, no classifier trained on those utterances outperforms simply guessing that all utterances are POSITIVE. Similarly, for those utterances which do not contain <Laugh>, no classifier trained on those utterances achieves higher accuracy than simply guessing that all utterances are NEUTRAL (although a low POSITIVE F-score when using ALL features is observed). It is only for those utterances which contain <Laugh> but also other tokens that guessing the majority class (also POSITIVE for these utterances) can be outperformed. This is achieved by relying on the TDUR feature, the duration of the utterance.

	All		only <LAUGH>		incl <LAUGH>		no <LAUGH>	
	Acc %	$F(+)$ %	Acc %	$F(+)$ %	Acc %	$F(+)$ %	Acc %	$F(+)$ %
guess (majority)	78.67	NaN	**76.79**	**86.87**	84.62	91.67	**92.69**	NaN
SPEC	80.27	47.17	69.64	80.90	79.57	88.22	**92.69**	NaN
PROS	81.93	49.39	71.43	82.61	82.93	90.72	92.22	2.80
SPEC+PROS	80.55	47.92	71.43	82.61	81.01	89.28	92.35	0.00
LEX	89.73	79.23	**76.79**	**86.87**	84.62	91.70	92.25	1.28
VOCINT - nLAUGH	78.62	11.69	**76.79**	**86.87**	**84.98**	**91.88**	91.86	0.00
nLAUGH only	**91.25**	84.01	**76.79**	**86.87**	84.62	91.67	—	—
VOCINT	91.17	**84.19**	**76.79**	**86.87**	81.01	89.14	91.86	0.00
ALL	90.67	83.16	71.43	81.82	84.50	91.52	91.94	**7.23**

Table 15.20: Accuracies, and F-scores for the POSITIVE valence class, when, for each feature set, a different classifier is trained for each subclass of utterances in TRAINSET.

These results strongly suggest that the acoustic features which were selected by the linear regression classifier are in fact finding laughter, leading to better-than-majority-class-guessing performance when the fact that laughter is present is *not known* a priori.

15.5 Detecting Involvement Hotspots

Conversational hotspots have been defined [225] as "regions in which participants are highly involved in the discussion", and, more precisely, "periods of about half a minute to one minute in the meeting where more than one participant [has] a high level of involvement". It was also argued in [225] that involvement, as used to annotate such hotspots, is closely related to emotional activation.

This section sets out to detect such hotspots, using the above definition which will henceforth be referred to as VERSION1 (an alternate definition was thereafter applied to annotate hotspots by the authors of [225]). Because hotspots can, and often do, describe the state of multiple participants, a most general hotspot detector is one which treats hotspot

occurrence as descriptors of the conversation as a whole. Given a multi-participant vocal interaction record $\mathbf{Q}$, only a one-dimensional trajectory $Y = \{y_t\} = \{y_1, \cdots, y_T\}$ must be inferred. This should satisfy

$$\begin{aligned} Y^* &= \arg\max_{Y=\{y_t\}} P(Y \mid \mathbf{Q}) \\ &= \arg\max_{Y=\{y_t\}} P(\mathbf{Q} \mid Y) \cdot P(Y) \\ &= \arg\max_{Y=\{y_t\}} \prod_{t=1}^{T} P(\mathbf{q}_t \mid y_t) \cdot \prod_{t=1}^{T} P(y_t \mid y_1, y_2, \cdots, y_{t-1}) \\ &\doteq \arg\max_{Y=\{y_t\}} \prod_{t=1}^{T} P(\mathbf{q}_t \mid y_t) \cdot \prod_{t=1}^{T} P(y_t \mid y_{t-1}) \ , \end{aligned} \tag{15.11}$$

where in the last line, Y is assumed to be 1st-order Markovian.

In this section, the size of the observation is 60 seconds, and the frame step is 15 seconds. Furthermore, the effect of the transition model $P(y_t \mid y_{t-1},)$ will be ignored, for simplicity. This renders hotspot detection a temporally independent binary classification task, in which, for a 60 second interval centered on instant t, the value of

$$y_t^* = \arg\max_{y \in \{\neg\mathcal{I}, \mathcal{I}\}} P(y \mid \mathbf{q}_t) \tag{15.12}$$

must be determined. The two mutually exclusive labels, $\neg\mathcal{I}$ and $\mathcal{I}$, represent intervals which are not-involved and involved, respectively.

15.5.1 Dataset Use

This chapter makes use of the ICSI Meeting Corpus [108], described in Chapter 4, its forced-alignment speech segmentation, its VERSION2 hotspot annotation [202], as well as the laughter segmentation produced as part of this thesis (detailed in Chapter 12).

Division of the 75 meetings into a TRAINSET, a DEVSET, and an EVALSET was motivated as follows. There are two sets of manually produced hot spot labels available for excerpts from 11 meetings. To enable comparison with human performance (cf. Subsection 15.5.8), these 11 meetings, subsequently referred to as EVALSUBSET, were placed in EVALSET; EVALSET was further augmented with 4 meetings of groups which are under-represented in EVALSUBSET relative to the rest of the corpus (`Bro010`, `Bro012`, `Bro016`, and `Bns002`). Of the remaining meetings, those whose numerical identifier ends with 1, 3, 5 or 7 were placed in DEVSET, and the remainder in TRAINSET.

VERSION2 hotspots differ from VERSION1 hotspots in duration; the duration of the former has an approximately log-normal distribution, with a most likely duration of 7 seconds and only one hotspot as long as 30 seconds. VERSION2 hotspots may contain, within their temporal support, DAs which are marked involved and those which are marked uninvolved. As suggested in the introduction, this section appeals to the VERSION1 hotspot duration guidelines (of 30-60 s) and attempts to detect not whether a dialog act contains involved speech, but whether a 60 second interval of meeting time does so. Reference labels for each interval are produced from the VERSION2 hotspot utterance tag [223]; an interval is given the label $\mathcal{I}$ only when it contains lexical productions marked as involved. Intervals are extracted from each meeting every 15 seconds. The resulting total number of intervals in the corpus is 15649; of these, 26.6% contain involved speech. The priors across the two labels $\mathcal{I}$ and $\neg\mathcal{I}$ are near-identical for all three of TRAINSET, DEVSET, and EVALSET.

15.5.2 Assessment of Performance

Since this task has been defined as one in which successive 60-second intervals are treated independently, standard classification accuracy is adopted as the primary metric. Also used in this section will be the F-score for the minority class $\mathcal{I}$ ("containing involved speech").

15.5.3 Baseline

Hotspots have been defined and studied on several occasions. In [225], it was shown for a small subset of the ICSI Meeting Corpus that annotators agree on the existence of VERSION1 hotspots significantly above chance, and that both fundamental frequency and energy show significant differences between involved and not involved speech; however, no hotspot detection system was proposed based on these findings. [224] in turn showed that, in a larger subset of the ICSI Meeting Corpus, VERSION1 hotspots are correlated with specific types of dialog acts (DAs), and that a system based on perfect DA knowledge would yield accuracies significantly above random guessing based on priors. It was not shown that these accuracies would be higher than always guessing the majority class. Finally, in [42], it was argued that although co-occurrence of VERSION2 hotspots and overlaps was much higher than could be expected by chance, the observed association was not strong enough to yield a useful classifier.

Since no previous experimental results are available, the baseline chosen here is the one not beaten by [224], namely always guessing the majority class — that no 60-second interval contains involved speech. On DEVSET and EVALSET, the resulting accuracies are 72.9% and 73.7%, respectively. Chance guessing, informed by the TRAINSET class prior, would yield accuracies of 60.9% and 61.2%, respectively.

15.5.4 Snapshots of Alternative Segmentations

The approach implied in Equation 15.12 requires that the binary label of a 60 second window be inferred from a local snapshot of vocal interaction $\mathbf{Q}$, as in Chapter 14. (However, since Y is a one-dimensional trajectory, any proposed models must be invariant under participant index rotation of $\mathbf{Q}$.) $\mathbf{Q}$ can be constructed from a segmentation of any type of vocal activity, and in this section multi-channel segmentations speech $\mathcal{S}$, laughter $\mathcal{L}$, voiced laughter $\mathcal{L}_V$, and unvoiced laughter $\mathcal{L}_U$ will be considered. These are shown in the Figure 15.7, for the same sample snippet of a single meeting.

The three binary multi-participant segmentations in Figure 15.7 lead to three alternate $\mathbf{Q}$, and it will be one of the contributions of this section to show that, in detecting involved speech, features drawn from the laughter segmentation $\mathcal{L}$ are better than those drawn from $\mathcal{S}$. Experiments will also consider various logical combinations of binary segmentations; Figure 15.8 depicts the logical OR and the logical AND between $\mathcal{S}$ and $\mathcal{L}$.

15.5.5 Static and Dynamic Features

Given some vocal activity segmentation $\mathcal{V} \in \{\mathcal{S}, \mathcal{L}, \mathcal{L}_V, \dots\}$, from which a discrete vocal interaction record $\mathbf{Q}$ can be constructed, both static and dynamic features are extracted for each 60-second interval.

The static features are of two types:

- $\{p_j^{\mathcal{V}}\}$: the proportion of interval duration for which each participant vocalizes, sorted by decreasing magnitude and padded with zeros to $K_{max} = 9$; and
- $\{o_j^{\mathcal{V}}\}$: the proportion of interval duration for which at least j participants vocalize simultaneously, for $1 \leq j \leq 9$.

Zero-padding allows for length-consistent feature vectors when meetings contain fewer than $K_{max} = 9$ participants.

The proposed dynamic features intend to compactly encode variation in the transitions from frame to frame, within each 60-second interval. They are computed using the parametric state-space model $\mathbf{\Theta}$ of Chapter 7. The model yields the likelihood of observing $\mathbf{Q}$ under an assumption of conditional participant independence,

$$P(\mathbf{Q} \,|\, \mathbf{\Theta}) \;\doteq\; \prod_{t=1}^{\mathcal{T}} \prod_{k=1}^{K} P_k(\mathbf{q}_t[k] \,|\, \mathbf{q}_{t-1}, \mathbf{\Theta}) \tag{15.13}$$

where

$$P_k(\mathbf{q}_t[k] = \blacksquare \,|\, \mathbf{q}_{t-1}) \;=\; \frac{1}{1 + e^{-\left(b_k + \sum_{j=1}^{K} w_{kj} \mathbf{q}_{t-1}[j]\right)/\mathcal{T}_k}} \tag{15.14}$$

$$P_k(\mathbf{q}_t[k] = \square \,|\, \mathbf{q}_{t-1}) \;=\; 1 - P_k(\mathbf{q}_t[k] = \blacksquare \,|\, \mathbf{q}_{t-1}) \tag{15.15}$$

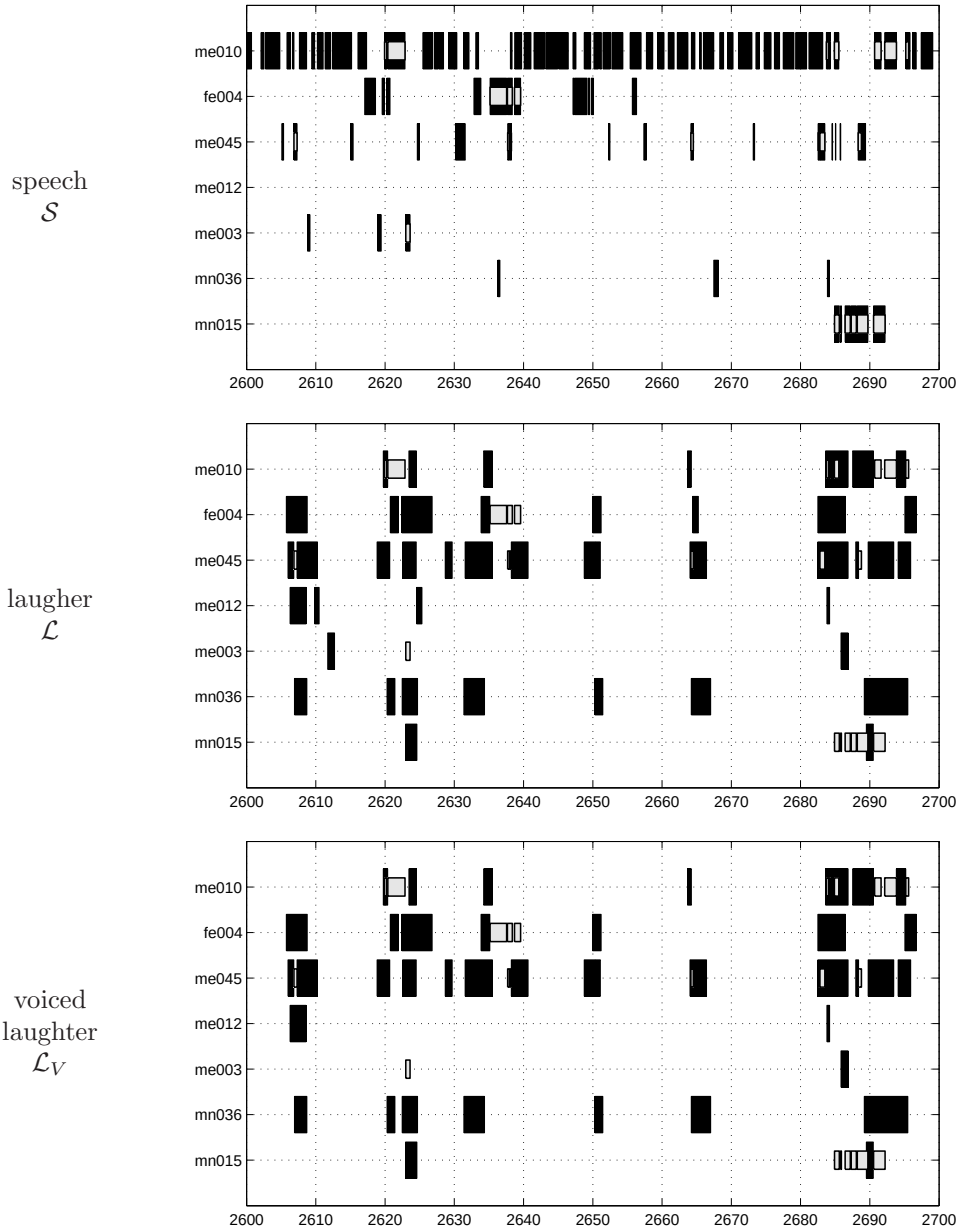


Figure 15.7: Three alternate multi-participant vocal activity segmentations for ICSI meeting `Bed010`, of 100 seconds between 2600 and 2700 seconds from the start. Black indicates the participant-attributed presence of vocal activity, one of $\mathcal{S}$, $\mathcal{L}$, or $\mathcal{L}_V$; light gray indicates the presence of participant-attributed involved speech.

To compute dynamic features, $\mathbf{W} = \{\{w_{kj}\}\} \in \mathbb{R}^{K \times K}$ and $\mathbf{b} = \{b_k\} \in \mathbf{R}^K$ are first estimated using all of $\mathbf{Q}$, assuming that $\mathcal{T}_k = 1$, $1 \leq k \leq K$. This yields a time-independent model $\mathbf{\Theta} = \{\mathbf{W}, \mathbf{b}\}$ which describes the transition statistics for the entire meeting. Then, for each 60-interval, the pseudotemperatures $\mathbf{T} = \{\mathcal{T}_k\}$ are adapted to maximize the likelihood of $\mathbf{Q}$ within that interval only. As was argued in Chapter 7, psuedotemperatures higher than unity entail a non-linear interpolation of the global model with a "random" model for which ■ and □ are equally likely.

The two specific types of dynamic features computed are:

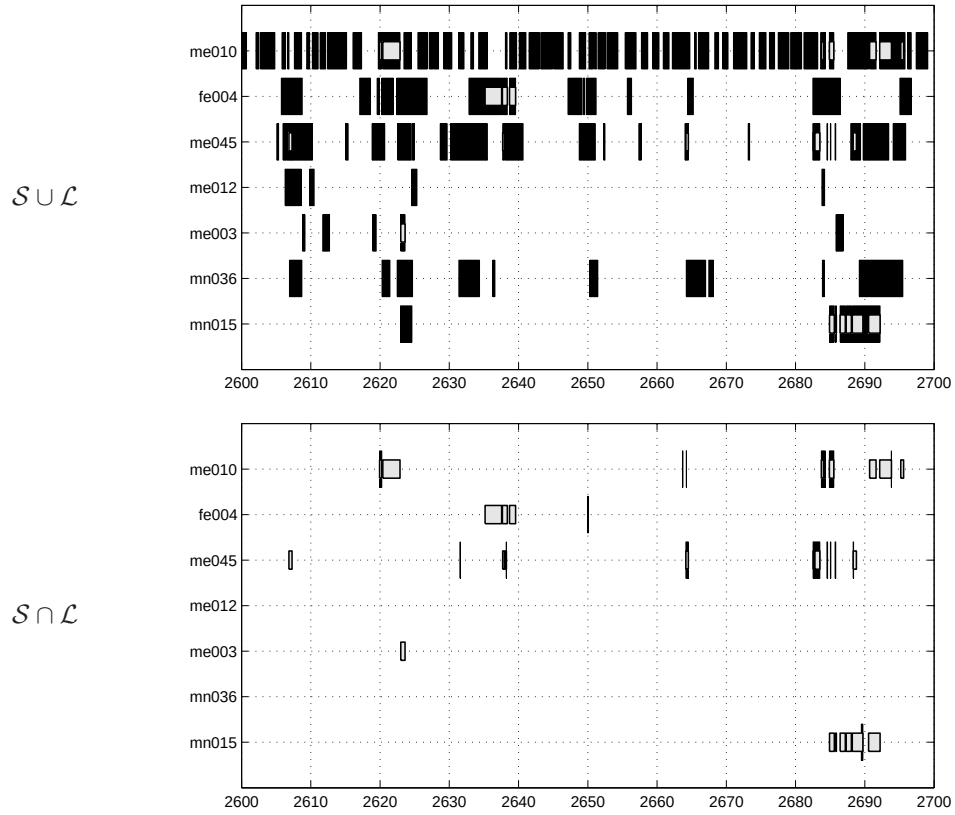


Figure 15.8: Two logical combinations of different multi-participant vocal activity segmentations, with operands as shown in Figure 15.7. Black indicates the participant-attributed presence of vocal activity, one of $\mathcal{S} \cup \mathcal{L}$ or $\mathcal{S} \cap \mathcal{L}$; light gray indicates the presence of participant-attributed involved speech.

- $\{\mathcal{T}_k^{PI}\}$: K participant-specific pseudotemperatures, or measures of departure from a participant-independent (PI) model (ie. $b_k = b$, $w_{kk} = w_+$, and $w_{kl} = w_-$, $\forall l \neq k$, for 3 degrees of freedom), sorted by decreasing magnitude and padded with zeros to $K_{max} = 9$; and

- $\{\mathcal{T}_k^{PS}\}$: K participant-specific pseudotemperatures, or measures of departure from participant-dependent (PS) norms (ie. untied $\mathbf{b}$ and $\mathbf{W}$, for $K + K^2$ degrees of freedom), sorted by decreasing magnitude and padded with zeros to $K_{max} = 9$.

15.5.6 System Performance

Experiments are conducted using a support vector machine (SVM) classifier[11]. All 4 feature types $\{p_j^{\mathcal{V}}\}$, $\{o_j^{\mathcal{V}}\}$, $\{\mathcal{T}_k^{PI}\}$, and $\{\mathcal{T}_k^{PS}\}$, of 9 features each, are computed for all 5 vocal activity segmentation types $\mathcal{S}$, $\mathcal{L}$, $\mathcal{L}_V$, $\mathcal{S} \cup \mathcal{L}$, and $\mathcal{S} \cap \mathcal{L}$. The resulting 36 cells are shown in Table 15.21, with performance measured on DevSet. In each cell, an SVM is trained on TrainSet, and single forward feature selection is performed to maximize accuracy on DevSet. All feature values are Z-normalized to facilitate SVM learning.

4 broad sets of observations can be made. First, all segmentation types and all feature types yield accuracies which

[11]The implementation used was SVM^{light}, available from Thorsten Joachims at `http://svmlight.joachims.org/` (downloaded on 5 August 2008 at 1430hrs GMT). Only a linear kernel with a biased hyperplane was explored; all other toolkit parameters were left at their default values to facilitate trend analysis.

Segm. Type	Feature Type				
	Static		Dynamic		all
	$\{p_j^{\mathcal{V}}\}$	$\{o_j^{\mathcal{V}}\}$	$\{\mathcal{T}_j^{PI}\}$	$\{\mathcal{T}_j^{PS}\}$	
$\mathcal{S}$	74.8 (3)	73.4 (3)	74.4 (1)	73.9 (4)	74.6 (7)
$\mathcal{L} \cup \mathcal{S}$	78.0 (4)	78.7 (9)	75.5 (1)	75.4 (1)	78.9 (3)
$\mathcal{L}$	80.4 (1)	79.9 (6)	80.4 (1)	79.9 (1)	80.7 (5)
$\mathcal{L}_V$	81.2 (2)	80.5 (6)	81.5 (1)	80.4 (6)	82.0 (8)
$\mathcal{L} \cap \mathcal{S}$	82.7 (2)	82.7 (6)	78.7 (1)	78.5 (4)	83.2 (7)
all	83.4 (9)	82.8 (2)	82.7 (8)	80.0 (3)	84.4 (5)

Table 15.21: Classification accuracy on DevSet using a linear-kernel SVM, for static and dynamic feature types (in columns) computed from different segmentation types ("Segm.", in rows). Each cell shows the accuracy achieved in % by an optimal feature subset identified using DevSet; the number of selected features, out of a total of $K_{max} = 9$ available in each non-"all" cell, is shown in parentheses. Majority class guessing yields 72.9%.

exceed majority class guessing; chance-corrected accuracies [224], given by

$$A' = \frac{A - A_E}{100 - A_E} \tag{15.16}$$

where A is the absolute accuracy, and $A_E = 60.9\%$ is the accuracy obtained by guessing with prior probabilities from TrainSet, lie between 32.0% and 60.1%. [224] applied this measure to detecting speaker-attributed involved speech (rather than 60-second intervals containing involved speech) and reported chance-corrected accuracies of 37% at best, using true dialog act information.

Second, looking at the static feature types only, there exists a clear progression in accuracy towards increasingly smaller subsets of the laughter segmentation $\mathcal{L}$; it should be noted that $(\mathcal{L} \cap \mathcal{S}) \subseteq \mathcal{L}_V \subseteq \mathcal{L} \subseteq (\mathcal{L} \cup \mathcal{S}) \supseteq \mathcal{S}$. The best single-segmentation-type, single-feature-type accuracy of 83.4%, for $\{p_j^{\mathcal{V}}\}$ from $\mathcal{L} \cap \mathcal{S}$, decreases as supersets of the $\mathcal{L} \cap \mathcal{S}$ segmentation are considered.

Third, dynamic feature types outperform static feature types only infrequently, and only by small amounts. In particular, the $\{\mathcal{T}_j^{PS}\}$ feature type appears to be uncompetitive as a whole. This may be due to the fact that the most informative segmentations (judging using static features) are also more sparse, making it difficult to estimate dynamics. However, as is shown in later subsections, dynamic features can be complementary even alongside the best-performing static features.

Finally, feature type combination frequently results in improved DevSet performance. Table 15.21 shows the effect of combining across feature types in the rightmost column, and across segmentation types in the bottom row; feature selection in these cells is performed over 36 and 45 features, respectively. Cases in which feature combination results in a degradation are instances where single feature forward selection gets trapped in a local maximum. Feature selection performed on all 180 features yields an accuracy of 84.4%, which is 1.7%abs better than the best single-segmentation single-feature-type accuracy in the table.

15.5.7 Generalization to Unseen Data

Results from experiments repeated on EvalSet, with features as selected for DevSet, are presented in Table 15.22. Trends are quite similar to those observed for DevSet. The accuracy achieved using all segmentations and all feature types is slightly lower, 84.0%, despite the fact that guessing the majority class yields accuracies slightly higher than for DevSet. 84.0% is only 1.0%abs higher than the best single-segmentation, single-feature-type accuracy obtained for $\{p_j^{\mathcal{V}}\}$ when $\mathcal{V} = \mathcal{L} \cap \mathcal{S}$. It is also lower, by 0.2%abs, than the accuracy achieved with a combination of all feature types using $\mathcal{L} \cap \mathcal{S}$ alone, indicating a degree of overfitting to DevSet during feature selection.

The accuracy of 84.0% is treated as the final performance measure on unseen data; it represents a 39.2% relative reduction of error over guessing the majority class, and a 58% relative error reduction over chance informed by TrainSet

Segm. Type	Feature Type				
	Static		Dynamic		all
	$\{p_j^{\mathcal{V}}\}$	$\{o_j^{\mathcal{V}}\}$	$\{\mathcal{T}_j^{PI}\}$	$\{\mathcal{T}_j^{PS}\}$	
$\mathcal{S}$	75.2 (3)	73.9 (3)	75.3 (1)	73.5 (4)	75.5 (7)
$\mathcal{L} \cup \mathcal{S}$	77.7 (4)	80.1 (9)	77.1 (1)	76.5 (1)	80.0 (3)
$\mathcal{L}$	80.6 (1)	81.2 (6)	80.8 (1)	75.5 (1)	80.0 (5)
$\mathcal{L}_V$	81.4 (2)	82.1 (6)	81.6 (1)	75.9 (6)	81.9 (8)
$\mathcal{L} \cap \mathcal{S}$	83.0 (2)	82.1 (6)	78.1 (1)	79.0 (4)	84.2 (7)
all	83.4 (9)	82.6 (2)	82.7 (8)	75.4 (3)	84.0 (5)

Table 15.22: Classification accuracy on EvalSet using a linear-kernel SVM, for static and dynamic feature types (in columns) computed from different segmentation types ("Segm.", in rows). Each cell shows the accuracy achieved in % by an optimal feature subset identified using DevSet; the number of selected features, out of a total of $K_{max} = 9$ available in each non-"all" cell, is shown in parentheses. Majority class guessing yields 73.7%.

priors; the corresponding chance-corrected accuracy is 59%.

15.5.8 Comparison with Human Performance

The detection of involvement is known to be a difficult and subjective task, as shown in an analysis of 13 meetings in which the majority of speech was contributed by 6 same participants [225]. Utterance-level agreement between any two native English-speaking labelers (out of 6) who were familiar with the meeting participants was shown to be $\kappa = 0.63$; non-native labelers, also familiar with the participants, appeared to agree at only $\kappa = 0.52$.

Subsequent analysis on EvalSubSet (a more varied subset of the corpus than used in [225]) using two labelers showed that per-utterance agreement on involvement is $\kappa = 0.63$, while that for "grown" hotspot intervals [223] is $\kappa = 0.67$. In this section, agreement is explored between those same two labelers (here, A and B) and on the same data as [226], on whether a 60 s interval is $\mathcal{I}$ or $\neg\mathcal{I}$. For each interval in EvalSubSet, A and B labels are extracted exactly as for the final consensus labels; also computed are $A \cup B$ and $A \cap B$ to gain insight into consensus creation on this task. Pair-wise agreement for all four sets of labels, the final consensus labels, and those produced by the final system whose accuracy was demonstrated to be 84.0% in the preceding subsection are shown in Table 15.23.

	B	$A \cup B$	$A \cap B$	`ref`	`hyp`
A	0.68	0.91	0.77	0.84	0.59
B		0.78	0.90	0.83	0.57
$A \cup B$			0.69	0.85	0.58
$A \cap B$				0.81	0.57
`ref`					0.54

Table 15.23: Pair-wise inter-labeler agreement measures (κ) on EvalSubSet between two human judges (A and B), their logical combinations ($A \cup B$ and $A \cap B$), the final consensus labels (`ref`) used as reference, and labels of the automatic system (`hyp`) achieving 84.0% in Table 15.22.

As Table 15.23 shows, inter-labeler agreement is 0.68, similar to that reported for DA-level involvement [225]. Because agreement between B and $A \cap B$ is near unity, and that between A and $A \cup B$ is near unity, B's involvement judgments appear to be a subset of A's. However, comparison between $A \cup B$, $A \cap B$, and the consensus labels `ref` indicates that the latter are a relatively complex combination of the two annotators' labels.

Table 15.23 also shows that agreement between automatic labels and the human-produced consensus labels is $\kappa = 0.54$, and that between the automatic system and either human taken alone is $\kappa \in [0.57, 0.59]$, slighly higher. This is only 0.10

lower than the agreement observed between the two human labelers.

15.5.9 Excluding Speech-Laughter

As is clear from Tables 15.21 and 15.22, the best classification accuracy is obtained with static features drawn from the $\mathcal{L} \cap \mathcal{S}$ segmentation, namely the segmentation of speech-laughter when the same participant is simultaneously speaking and laughing. This is somewhat suboptimal. Speech-laughter may be hard to detect, since it is rare (cf. Chapter 12), and since it is likely to be acoustically similar to both speech and laughter, which are much more frequent. This would make the performance observed here out of reach for fully automatic systems.

It is quite likely that speech which is produced simultaneously with laughter by the same speaker would be labeled as prosodically involved by annotators, and therefore it can be expected that speech-laughter is indicative of hotspots as defined here. The much more interesting question regarding hotspot detection is the extent to which it is robust to the undetectability of speech-laughter. Tables 15.21 and 15.22 also show that features computed using voiced laughter alone (which includes speech-laughter) achieve accuracies which are close to those achieved using speech-laughter. Even more realistically, for vocal activity detection systems which do not attempt to discriminate between speech and laughter, but do discriminate between vocalization $\mathcal{L} \cup \mathcal{S}$ and non-vocalization, hotspot detection performance is also quite good (80.1% using the $\{o_j^{\mathcal{V}}\}$ features for EvalSet).

The effect of excluding speech-laughter can be investigated by forming the logical And between a laughter segmentation $\mathcal{L}$ and the complement of the speech segmentation $\mathcal{S}$, yielding $\mathcal{L} \cap \neg\mathcal{S}$. Table 15.24 presents classification accuracies, as well as F-scores, for SVM system relying only on the $p_j^{\mathcal{V}}$ feature type; some numbers are duplicated from Tables 15.21 and 15.22. For completion, systems are also constructed which exclude speech-laughter from the $\mathcal{L}_V$ and $\mathcal{L}_U$ segmentations.

Feature Set	Accuracy, %		F-Score, %	
	dev	eval	dev	eval
guess, priors	60.9	61.2	—	—
guess, major.	72.9	73.7	—	—
$\mathcal{S}$	74.8	75.2	34.4	28.0
$\mathcal{L}$	80.4	81.2	64.6	64.8
$\mathcal{L} \cap \mathcal{S}$	82.7	83.0	68.9	70.6
$\mathcal{L} \cap \neg\mathcal{S}$	80.4	80.8	64.6	64.8
$\mathcal{L}_V$	81.2	81.4	65.1	64.3
$\mathcal{L}_V \cap \mathcal{S}$	82.9	85.6	69.5	67.1
$\mathcal{L}_V \cap \neg\mathcal{S}$	81.5	81.4	65.0	67.1
$\mathcal{L}_U$	76.4	77.4	56.3	55.1
$\mathcal{L}_U \cap \mathcal{S}$	73.7	72.6	27.6	21.9
$\mathcal{L}_U \cap \neg\mathcal{S}$	76.4	77.4	56.3	55.1

Table 15.24: Accuracies and F-scores for SVM classifiers relying on $\{p_j^{\mathcal{V}}\}$ features only, with $\mathcal{V}$ defined as given in each row. "dev" and "eval" refer to DevSet and EvalSet, respectively.

As the table shows, features drawn from all laughter segmentations, and from logical intersections with laughter segmentations, lead to F-scores exceeding 50% (except the intersection of unvoiced laughter with speech, $\mathcal{L}_U \cap \mathcal{S}$, which is very rare). Unvoiced laughter appears to be much less relevant to the detection of involved speech than voiced laughter, and features drawn from the latter often outperform those drawn from all laughter. This is felicitous from the point of view of fully automatic systems, since voiced laughter is acoustically easier to detect (cf. Chapter 13) than unvoiced laughter (and hence than all laughter $\mathcal{L} \equiv \mathcal{L}_V \cup \mathcal{L}_U$). Most importantly, excluding speech-laughter from voiced laughter, $\mathcal{L}_V \cap \neg\mathcal{S}$ yields accuracies and F-scores which are only 1.6%abs and 3.5% lower than all speech-laughter, respectively. This may be because, immediately following speech-laughter, laughter continues, and/or that non-speech-laughter precedes speech-laughter, for physiological reasons. It is therefore of interest just how far away, from the laugher's temporally most proximate speech, laughter must lie in order to be indicative of involved speech, if in fact laughter identifies only the involved speaker.

15.5.10 Selecting Voiced Laughter Based on Proximity to Laugher's Speech

To explore how temporally proximate laughter must be, to the laugher's own speech, in order to be indicative of the presence of involved speech from any participant, a series of *masks* are created. The masks are achored to the speech segmentation; they are then intersected with the laughter segmentation via logical AND. The process of mask creation is depicted in Figure 15.9; to facilitate understanding, the operator $\sigma(\tau_L, \Upsilon, \tau_R)$ is defined to implement a padding of τ_L and τ_R seconds on the left and right, respectively, of every ■ interval in a segmentation Υ.

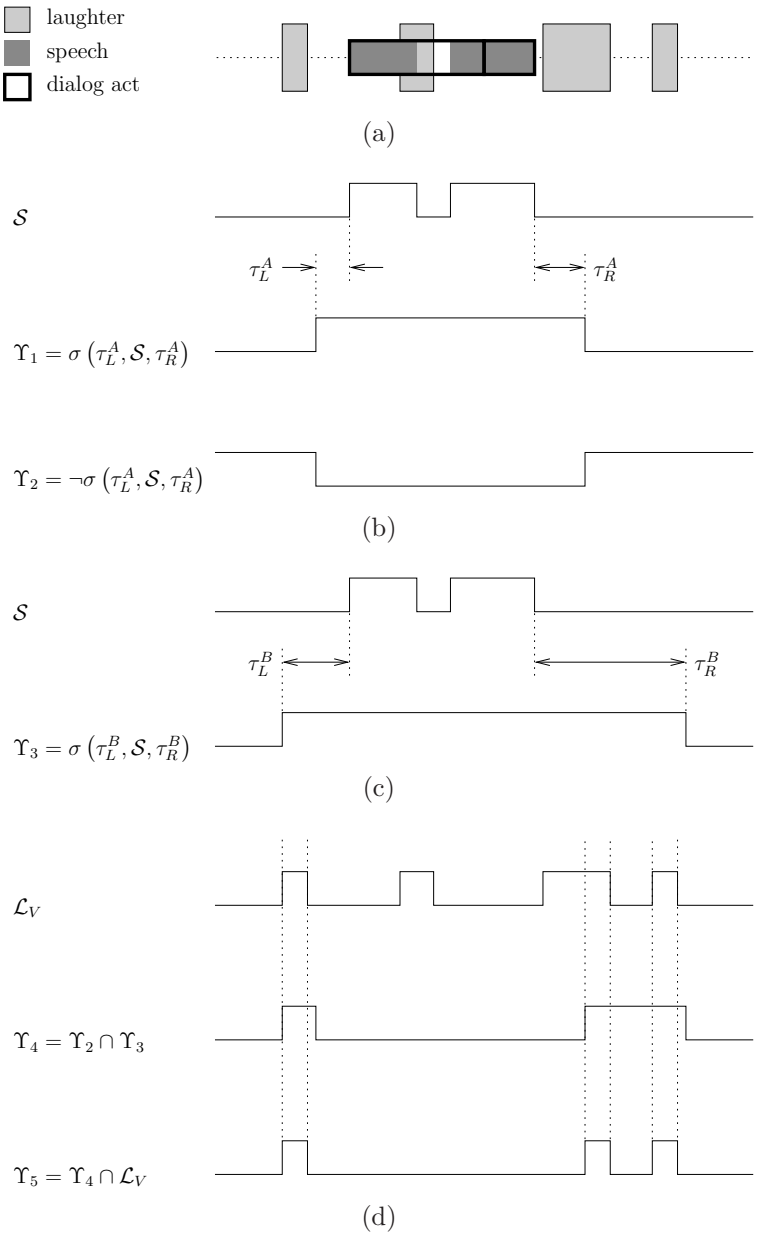


Figure 15.9: Masking voiced laughter $\mathcal{L}_V$ with a mask constructed using speech $\mathcal{S}$; time τ is shown from left to right. The process of arriving at the final binary trajectory in panel (d), from the given speech and laughter segmentations for a particular participant shown in panel (a), is as described in the text.

Panel (a) of the figure depicts simultaneously the speech activity and the laughter activity segmentations of a single participant, with time shown from left to right; also depicted is the dialog act segmentation over the speech activity (two dialog acts are shown). In Panel (b), the speech segmentation $\mathcal{S}$ is prepadded with τ_L^A and postpadded with τ_R^A seconds, and then complemented, to yield $\Upsilon_2 = \neg\sigma\left(\tau_L^A, \mathcal{S}, \tau_R^A\right)$. This mask identifies all those instants which are at least τ_L^A seconds prior to any speech and at least τ_R^A seconds following any speech. Panel (c) depicts prepadding by τ_L^B and postpadding by τ_R^B, also of the speech segmentation $\mathcal{S}$. The resulting mask, $\Upsilon_3 = \sigma\left(\tau_L^B, \mathcal{S}, \tau_R^B\right)$, identifies all instants which are at most τ_L^B seconds before any speech and at least τ_R^B seconds following any speech. When the two masks Υ_2 and Υ_3 are intersected, shown as Υ_4 in the figure, and then applied to the voiced laughter segmentation $\mathcal{L}_V$, the result is only those instants during which the participant is laughing, which are between $\tau_L^B - \tau_L^A$ seconds before any speech and $\tau_R^B - \tau_R^A$ after any speech.

Using this technique, a family of masks which are at most 1 second in duration are defined; it consists of three sub-families of masks:

- *pre-talkspurt masks*, $\Upsilon_{pre}^{slice}(\tau) = \neg\sigma(\tau - 1, \mathcal{S}, \tau) \cap \sigma(\tau, \mathcal{S}, 0)$, consisting of slices of up to 1 second in duration, at least $\tau - 1$ seconds before subsequent speech, at most τ seconds before subsequent speech, and at least τ seconds after precedent speech;

- *post-talkspurt masks*, $\Upsilon_{post}^{slice}(\tau) = \neg\sigma(\tau, \mathcal{S}, \tau - 1) \cap \sigma(0, \mathcal{S}, \tau)$, consisting of slices of up to 1 second in duration, at least $\tau - 1$ seconds after precedent speech, at most τ seconds after precedent speech, and at least τ seconds before subsequent speech; and

- *inter-talkspurt masks*, $\Upsilon_{inter}^{slice}(\tau) = \neg\sigma(\tau - 1, \mathcal{S}, \tau - 1) \cap \sigma(\tau, \mathcal{S}, 0) \cap \sigma(0, \mathcal{S}, \tau)$, consisting of slices of 1 second in duration, at least $\tau - 1$ seconds after precedent speech, at most τ seconds after precedent speech, at least $\tau - 1$ seconds before subsequent speech, and at most τ seconds before subsequent speech. The latter category consists of all those slices that are equally proximate to precedent and subsequent talkspurts.

Figure 15.10 shows the proportion of laughter found in each mask, indexed by τ, for all three 1-second mask sub-families. As can be seen, voiced laughter occurs predominantly near laugher's own speech, and its occurrence (within any 1-second slice) decreases exponentially with temporal distance away from that speech. There appears to be far more voiced laughter following speech than preceeding it. Although it is not shown, it is assumed that for $\tau > 10$ seconds the exponential trend observed in Figure 15.12(a) for all three of pre-talkspurt, post-talkspurt, and inter-talkspurt voiced laughter continues, making the amount of voiced non-speech laughter beyond $\tau = 10$ too sparse for modeling. In the remainder of this section, only the slices for which $\tau \in [1, 10]$ are considered; together, they account for 68% of all voiced non-speech laughter by time.

The $\mathcal{L}_V$ experiments of Subsection 15.5.9 are repeated, by masking $\Upsilon = \mathcal{L}_V$ with each mask. A similar suite of experiments is conducted with two additional families of masks. The first of these are τ-second masks, also of three sub-families; the masks of each sub-family are formed by taking the union of all 1-second masks of the same family,

- *pre-talkspurt masks*, $\Upsilon_{pre}^{prox}(\tau) = \bigcup_{\tau'=1}^{\tau} \Upsilon_{pre}^{slice}(\tau')$;

- *post-talkspurt masks*, $\Upsilon_{post}^{prox}(\tau) = \bigcup_{\tau'=1}^{\tau} \Upsilon_{post}^{slice}(\tau')$; and

- *inter-talkspurt masks*, $\Upsilon_{inter}^{prox}(\tau) = \bigcup_{\tau'=1}^{\tau} \Upsilon_{inter}^{slice}(\tau')$.

These masks identify all those instants which are *at most* τ seconds away from speech.

The second derived set of masks consists of $(10 - \tau)$-second masks,

- *pre-talkspurt masks*, $\Upsilon_{pre}^{dist}(\tau) = \bigcup_{\tau'=\tau}^{10} \Upsilon_{pre}^{slice}(\tau')$;

- *post-talkspurt masks*, $\Upsilon_{post}^{dist}(\tau) = \bigcup_{\tau'=\tau}^{10} \Upsilon_{post}^{slice}(\tau')$; and

- *inter-talkspurt masks*, $\Upsilon_{inter}^{dist}(\tau) = \bigcup_{\tau'=\tau}^{10} \Upsilon_{inter}^{slice}(\tau')$.

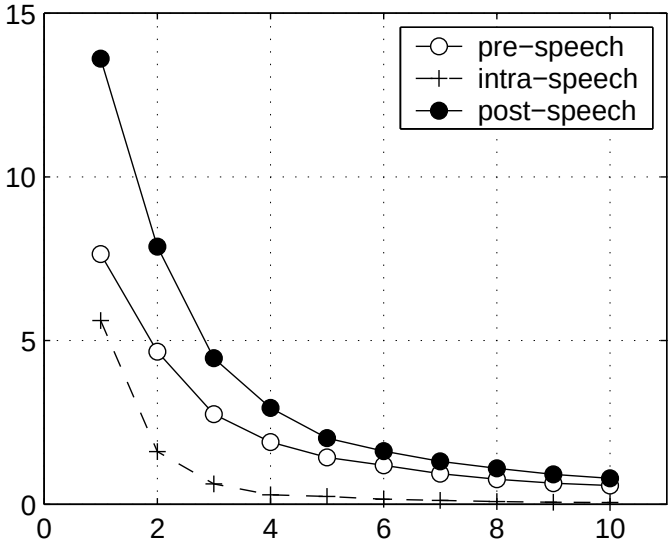


Figure 15.10: Proportion in % of overall laughter by time, per mask $\Upsilon_\alpha(\tau)$; τ shown in seconds along the x-axis.

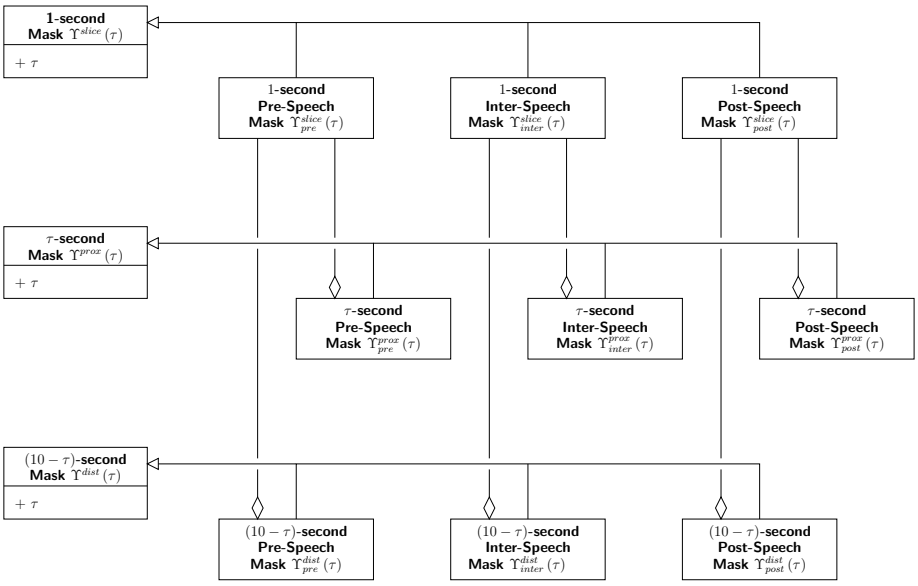


Figure 15.11: Relationships among the three types of mask families, "slice", "prox", and "dist", and their three subfamilies *pre*, *inter*, and *post*. "$\triangleleft$" and "$\diamondsuit$" denote (standard UML) generalization and composition, respectively.

These masks identify all those instants which are *at least* τ seconds away from speech (and at most 10 seconds). The relationships among the families and subfamilies of masks considered here are shown in Figure 15.11.

Repetition of the experiments described in Section 15.5.9, but using only the subset of voiced laughter given by $\Upsilon' = \Upsilon^{slice}_{\alpha}(\tau) \cap \mathcal{L}_V$, for each of ten masks indexed by $\tau \in [1, 2, \cdots, 10]$ seconds and each of the three subfamilies of masks $\alpha \in \{pre, post, inter\}$, yields the results shown in Figure 15.12. Classification accuracy and F-score for EvalSet are shown separately, together with the performance achieved by not masking out any voiced laughter. It is evident that

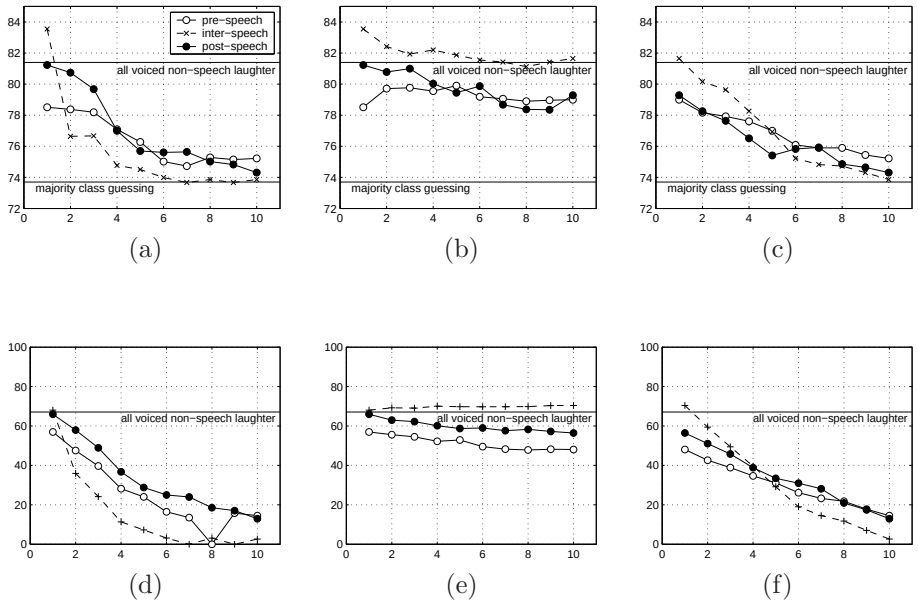


Figure 15.12: Classification accuracy (in %) for $\{p_j^{\mathcal{L}_V}\}$ features extracted from: (a) laughter in 1-second slices τ seconds away from speech, using $\Upsilon_\alpha(\tau) \cap \mathcal{L}_V$; (b) laughter in τ-second slices immediately proximate to speech, using $\Upsilon_\alpha^{prox}(\tau) \cap \mathcal{L}_V$; and (c) laughter in $10 - \tau$-second slices τ seconds away from speech, using $\Upsilon_\alpha^{dist}(\tau) \cap \mathcal{L}_V$. F-scores curves (in %) corresponding to (a), (b), and (c) are shown in (d), (e), and (f), respectively. The legend in (a) relates to all panels.

the most relevant voiced non-speech laughter is that found in immediate proximity to laugher's speech; F-scores which lie above 50% are to be found for all three of pre-talkspurt, post-talkspurt, and inter-talkspurt contexts for $\tau = 1$, and only for the post-talkspurt context for $\tau = 2$. As τ approaches 10 seconds, the classification accuracies for voiced laughter in all three contexts approach that obtained with majority class guessing, and F-scores approach zero.

It also appears that, at least for $\tau < 4$, post-talkspurt voiced laughter is more relevant than pre-talkspurt voiced laughter, and that for $\tau = 1$ inter-talkspurt voiced laughter is most relevant. The latter may be to be due to the fact that inter-talkspurt laughter is much more likely to "bleed" from or into the laugher's ongoing verbal production, making that production involved and thereby marking the current interval as containing involved speech. Other laughter is arguably less likely to have affected the production of that speech.

System performance using voiced laughter which is found at most τ seconds away from the laugher's closest talkspurt, identified by the second family of masks (Υ^{prox}), is shown in panels (b) and (e) of Figure 15.12. There are three observations to be made with regard to these results. First, classification accuracy for pre-talkspurt and post-talkspurt voiced laughter never exceeds that for all voiced non-speech laughter, at any value of the threshold τ past which all voiced laughter is discarded. The same is true for the F-score curve in panel (e). Post-talkspurt laughter appears to be more relevant than pre-talkspurt laughter. Second, except for voiced pre-talkspurt laughter at small values of τ, the accuracy actually decreases as more and more distant voiced non-speech laughter is considered. This is true of the F-score curves over the entire range $\tau \in [1, 10]$. Third, voiced inter-talkspurt laughter yields higher classification accuracies and higher F-scores than does all voiced non-speech laughter, over the majority of the interval $\tau \in [1, 10]$. Interestingly, the accuracy and F-score curves for this class of voiced laughter observe opposite slope trends with increasing τ. Closer inspection reveals that feature selection results in a different optimal set when maximizing the two performance measures using DevSet; this precludes a direct comparison of features in the accuracy and F-score plots.

Panels (c) and (f) of Figure 15.12 show results for voiced laughter which is found at least τ seconds away from the laugher's closest talkspurt, given by intersection with the third family of masks (Υ^{dist}). Classification accuracies can be seen to fall steeply towards the accuracy achieved by majority class guessing, and F-scores fall equally steeply towards zero, for increasing values of τ. The contrast between performance obtained by retaining laughter close to laugher's

speech (panels (b) and (e)) and that obtained by ablating laughter close to laugher's speech (panels (c) and (f)) offers the strongest evidence that laughter proximate to laugher's speech is especially relevant to the considered task.

15.5.11 Annotating Voiced Laughter Based on Proximity to Laugher's Speech

This subsection explores the performance achieved by using all three subfamilies of the 1-second masks, by combining features. The feature selection scheme is exposed to features drawn from all of the subsegmentations $\Upsilon_{\alpha}^{slice}(\tau) \cap \mathcal{L}_V$, which are orthogonal in the interval $\tau \in [1, 10]$. This amounts to a total of $9 \times 3 \times 10$ features, rather than only 9 features, namely

$$\mathbf{F} \equiv \bigcup_{\alpha} \bigcup_{\tau=1}^{10} \mathbf{f}\left(\Upsilon_{\alpha}^{slice}(\tau) \cap \mathcal{L}_V\right) . \tag{15.17}$$

where $\mathbf{f}$ is a feature vector of 9 features. To a certain degree, this approach is tantamount to annotating slices of laughter of up to 1 second in duration with temporal proximity to and co-orientation with the laugher's nearest talkspurt. Performance is shown in Table 15.25.

Feature Set	Accuracy, %		F-Score, %	
	DevSet	EvalSet	DevSet	EvalSet
$\mathbf{f}(\mathcal{L}_V)$	81.5	81.6	65.1	64.3
$\mathbf{f}(\mathcal{L}_V \cap \mathcal{S})$	82.9	85.6	69.5	67.1
$\mathbf{f}(\mathcal{L}_V \cap \neg\mathcal{S})$	81.5	81.4	65.0	67.1
$\mathbf{F}$	83.2	84.4	71.2	70.3

Table 15.25: Classification accuracy and F-score, both in %, for retrieval of 60-second meeting intervals containing involved speech, based on the proximity-annotated feature set $\mathbf{F}$. Also shown are several comparable systems from Table 15.24.

As can be seen, on held-out EvalSet data, classification performance improves by 3% absolute over the unannotated, complete voiced non-speech laughter segmentation $\mathcal{L}_V$, representing a 16% relative reduction of classification error. It also approaches the performance of the speech-laugh segmentation ($\mathcal{L}_V \cap \mathcal{S}$) which was claimed to be likely much more difficult to produce automatically. The corresponding increase in F-score over the unannotated voiced non-speech laughter segmentation $\mathcal{L}_V$ is 3.2% absolute; the same absolute increase in F-score is observed over the speech-laugh segmentation ($\mathcal{L}_V \cap \mathcal{S}$) that produces the highest classification accuracy.

15.5.12 The Effect of Retrieved Interval Duration

Experiments in this section aim to characterize performance as a function of the interval size; the choice thus far has been for 60-second intervals. However, for any particular application, an interval of 60 seconds may turn out to be inconveniently long or short. One would therefore like to be able to estimate ahead of time the impact of adopting a different interval size.

As a basis for such estimation, the experiment from the previous subsection, involving the proximity-annotated feature set $\mathbf{F}$, is repeated for each of 30 s, 40 s, 50 s, 60 s, 70 s, 80 s, 90 s, and 100 s. In each case, the oversampling ratio of 4 is retained, such that the interval step is always a quarter of the interval size. With smaller interval sizes, the temporal support of all intervals containing involved speech shrinks, further skewing the prior from $[1/2, 1/2]$; in tying the interval step to the interval size, it is hoped that these effects are mitigated by increasing the absolute number of intervals containing involved speech.

The results of these experiments are shown in Figure 15.13. Panel (a) gives the EvalSet classification accuracy using the best feature subset selected out of $\mathbf{F}$ using DevSet, as well as the accuracy obtained by guessing the majority class. As predicted, decreasing the interval size leads to higher majority guessing accuracies, while increasing the interval size leads to the opposite phenomenon. However, in the latter case, the best classification accuracy does not drop as quickly, leading to higher accuracies relative to majority class guessing as the interval size is increased (cf. panel (b)). An almost identical

trend is observed in panel (c) of Figure 15.13, which shows the EvalSet F-score. In summary, over the explored interval of window sizes between 30 seconds and 100 seconds, the proposed retrieval system exhibits accuracies that are 30-50% relative above majority class guessing and F-scores in the range 65-76%. Furthermore, in this range both performance measures exhibit a near-linear relationship with interval size.

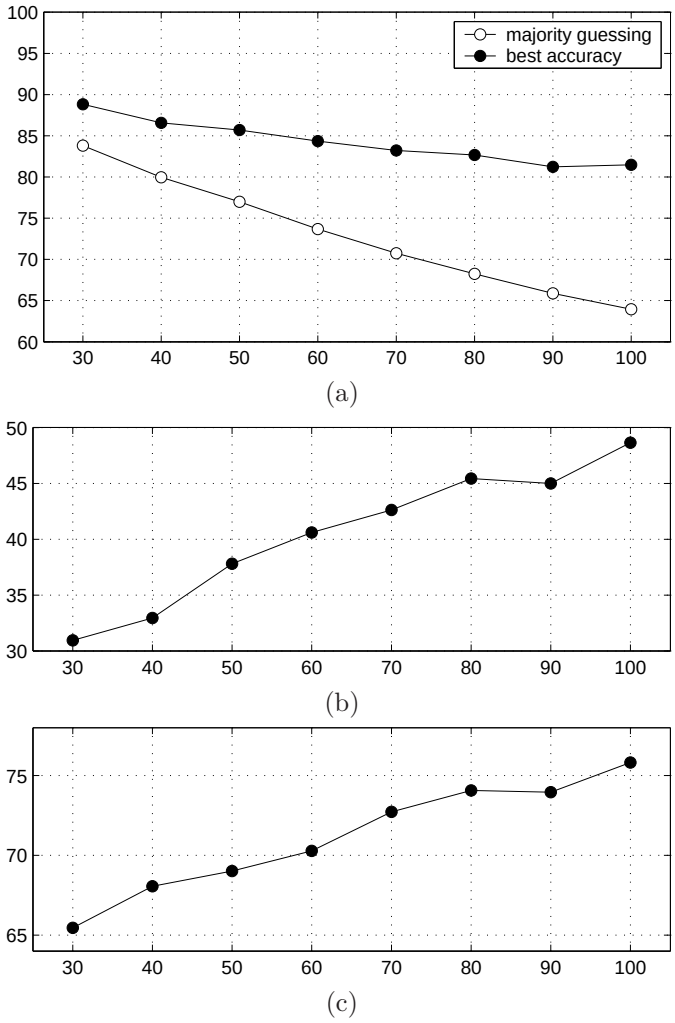


Figure 15.13: EvalSet performance of retrieval system, in %, as a function of interval size (in seconds): (a) classification accuracy; (b) relative improvement over majority class guessing; and (c) F-score.

15.5.13 Retreiving Amused versus Involved Speech

Finally, experiments are conducted in which the task is not the retrieval of intervals containing involved speech, but that of intervals containing speech exhibiting a specific type of involvement. Dialog acts marked as involved in the ICSI Meeting Corpus are annotated additionally with membership to a structured Version2 hotspot entity, one of whose attributes is primary hotspot type. The most frequently assigned hotspot types are amusement, other, and agreement, accounting for 82%, 6%, and 1%, respectively, of involved speech by time.

amusement involvement is present in 21.2-21.8% of 60-second intervals, at an interval step of 15 seconds, in the three data sets, leading to a higher classification accuracy obtainable by always guessing the majority class than when

all involvement types are taken together. The SVM classifier and feature selection over the features **F**, described in the previous section, lead to an EvalSet accuracy of 84.30%, representing a 28.0% relative reduction in classification error over majority-class guessing. This is a slightly lower absolute accuracy than that for intervals containing any type of involvement. Furthermore, the F-score achieved on EvalSet is only 66.8%, compared to 70.3% for all involvement. This result is somewhat surprising, as it was assumed that the association between colocated laughter and amusement would be stronger than that between laughter and involvement, and suggests that other types of involvement may also entail laughter from the participants, in contexts that are similar under the feature representation **F**.

15.6 Detecting Attempts to Amuse

The previous two sections have attempted to characterize emotional valence and emotional activation, the two most popular axes of a continuous, multi-dimensional conceptualization of emotion [200]. It was shown that laughter, when its presence can be detected, appears to play an important role in signalling the emotional states of vocalizing participants. What has not been shown is what leads to those states.

An evident observation is that participants to conversation deliberately seek to change the emotional states of their interlocutors, particularly by deviating from the current topic to tell a joke. There is much room for conjecture whether those telling jokes are also modifying their own state, which may change over the course of the joke's telling, whether interlocutors' laughter is precedent or subsequent or even accompanied by those interlocutors' changes in emotional state, and what the emotional states are of those participants which do not laugh. The experiments in this section sidestep these questions, and attempt merely to predict whether a joke is being told, based on observed and known laughter (and speech) activity.

15.6.1 Dataset Use

As in the preceding section, the data used here is the ICSI Meeting Corpus. Its split into a TrainSet, a DevSet, and a EvalSet is different than in Section 15.5. Instead, the split is as in Chapter 14. The dialog act (DA) annotation, described in Chapter 4 which accompanies the corpus, includes an optional binary label (`j`) per DA indicating a humorous or sarcastic nature. A single joke or sarcastic comment can of course consist of multiple consecutive `j` DAs.

In constrast to Chapter 14, `s` and `q` DAs possessing the `j` label are treated as a separate DA type in the ensuing experiments. In the ICSI corpus as a whole, they account for 0.53–0.73% of speaking time. The breakdown by dataset, complementary to that shown in Table 14.1 in which the `j` label was ignored, is provided in Table 15.26. As can be seen by comparing Tables 14.1 and 15.26, the overwhelming majority of `j` labels are applied to statements rather than questions.

DA Type	TrainSet	DevSet	EvalSet
`fh`, floor holder	2.32	2.29	3.00
`h`, hold	0.21	0.36	0.26
`fg`, floor grabber	0.55	0.58	0.62
`b`, backchannel	2.86	2.65	2.83
`bk`, acknowledgement	1.41	1.41	1.48
`aa`, accept	1.17	1.13	1.10
`s`, statement	84.28	83.63	82.24
`q`, question	6.47	7.41	7.86
`j`, joke/sarcasm	**0.73**	**0.53**	**0.62**

Table 15.26: The 9 DA types of interest in this section, and their prior probability distribution (in %) in the three datasets used in this chapter; attempts at humor (`j`) shown in bold.

15.6.2 Assessment of Performance

To facilitate the construction of receiver operating characteristic (ROC) curves, performance in this section is assessed by computing the false alarm and miss rates. The false alarm rate is given as the number of false positives divided by the total amount of speech not annotated with `j`, as in Equation 11.7. The miss rate (MS) is the ratio of false negatives to the total amount of speech annotated with `j`, as in Equation 11.8. Numerators and denominators in both instances are in seconds, or frames. The measure minimized in this section is the sum of the false alarm and miss rates, ER = FA + MS.

ROC curves are the locii of points (FA, 1.0 − MS), obtained by modifying a selected system parameter which controls the trade-off between misses and false alarms.

15.6.3 Baseline

The baseline system is a hidden Markov model (HMM) Viterbi decoder, with a frame size and frame step of 100 ms. It is identical in structure to the DA recognizer of Chapter 14, except that speech annotated with `j` is modeled as a separate, ninth DA type.

The performance of the baseline is shown in Table 15.27. "T1" refers to the baseline topology of Section 14.6 but with nine DA subtopologies, whose transition probabilities are inferred from TrainSet. "T0" is the same topology, with all licensed transitions having a probability of unity. Both entries in the table "T0" and "T1" refer to topology-only systems, as in Subsection 14.6.5. Low error rates for system T1 relative to that of T0 would indicate that `j` DAs can be predicted from talkspurt duration and from the sequencing of `j` DAs with respect to other DA types. As can be seen, the miss rates achieved by both T0 and T1 are quite high, indicating that `j` talk cannot be predicted in this way.

System	DevSet			EvalSet		
	FA	MS	ER	FA	MS	ER
T0	8.1	90.6	98.7	8.3	92.5	100.7
T1	0.3	96.7	97.0	0.2	94.0	94.2
LEX w/o T	53.6	32.8	86.4	53.7	32.9	86.6
LEX w/ T0	40.2	42.9	83.1	40.5	44.2	84.7
LEX w/ T1	12.7	67.0	79.6	12.8	70.5	83.3

Table 15.27: Detection performance, in %, of the topology with equiprobable transition probabilities (T0), the topology with transition probabilities trained using TrainSet (T1), and the lexical emission model with both T0 and T1 (LEX w/ T0 and LEX w/ T1, respectively). FA is the false alarm rate, MS is the miss rate, and ER = FA + MS.

"Lex" refers to the lexical bigram emission model described in Section 14.12. Lines 3 and 4 in Table 15.27 represent the performance of systems for which the bigram emission probabilities are embedded in the proposed topology, with equiprobable transition probabilities (T0) and transition probabilities inferred from TrainSet (T1), respectively. Although performance is significantly above random guessing, these systems demonstrate that lexical bigram features do not discriminate very successfully between attempts at humor and other DA types (in contrast to their utility for discriminating among non-`j` DAs.

In the rest of this work, the system "LEX w/ T1" in the line 5 of the table will be referred to as simply LEX.

15.6.4 Modeling the Vocal Activity Context

The local neighborhood snapshot of each participant, at every speaking instant t, is modeled separately for speech $\mathcal{S}$ and laughter $\mathcal{L}$. The size of the snapshot (10 seconds, from $t-5$ to $t+5$), the ranking of interlocutors within the snapshot (locally most talkative or laughing in the 10-second snapshot), and the number of interlocutors modeled (top 3) are identical to what was done in Chapter 14. It should be noted that the laughter context was not modeled in Chapter 14; Table 15.28 lists the proportion of vocalizing time that is accounted for by laughter, in each of the three datasets used. It can be seen that laughter is significantly more frequent by time, in all datasets, than is speech implementing attempts to amuse (cf. Table 15.26).

Laughter	TrainSet	DevSet	EvalSet
all, $\mathcal{L}$	10.0	8.9	10.0
voiced, $\mathcal{L}_V$	7.6	6.3	6.6
unvoiced, $\mathcal{L}_U$	2.4	2.6	3.4

Table 15.28: Proportion by vocalizing time, in %, of laughter, for all three datasets.

The results of experiments using the neighborhood snapshot of speech activity and of laughter activity are shown in Table 15.29. Both speech $\mathcal{S}$ and laughter $\mathcal{L}$ context features are seen to be significantly better than lexical features; on DevSet, they yield error rates which are only a quarter of the error rate achieved with the baseline system, due mostly to much lower miss rates.

System	DevSet			EvalSet		
	FA	MS	ER	FA	MS	ER
$\mathcal{L}$	14.0	5.3	19.3	15.6	8.1	23.7
$\mathcal{L}_V$	8.7	16.0	24.7	9.5	9.9	19.4
$\mathcal{L}_U$	12.4	21.2	33.6	13.8	17.4	31.2
$\mathcal{L}_V$ Ⓜ $\mathcal{L}_U$	7.4	15.7	23.1	8.0	13.6	21.7
$\mathcal{L}_V$ Ⓕ $\mathcal{L}_U$	14.2	6.6	20.8	15.7	7.0	22.7
$\mathcal{L}_V$ Ⓒ $\mathcal{L}_U$	14.0	6.3	20.3	15.1	8.3	23.3
$\mathcal{S}$	7.5	47.4	54.9	8.6	62.8	71.4
$\mathcal{L}$ Ⓜ $\mathcal{S}$	9.7	6.6	16.3	11.0	8.4	19.4
$\mathcal{L}$ Ⓕ $\mathcal{S}$	6.0	17.8	23.8	6.8	21.6	28.4
$\mathcal{L}$ Ⓒ $\mathcal{S}$	6.0	16.0	22.0	6.4	17.8	24.2

Table 15.29: Detection performance, in %, of several systems employing topology T1, the laughter $\mathcal{L}$ context and the speech $\mathcal{S}$ context; FA is the false alarm rate, MS is the miss rate, and ER = FA + MS.

The experiments are duplicated using only voiced laughter ($\mathcal{L}_V$) and only unvoiced laughter, respectively, in lines 2 and 3 of the table. It appears that both types of laughter are important, with unvoiced laughter less relevant than voiced laughter. Model-space and feature-space combination of $\mathcal{L}_V$ and $\mathcal{L}_U$ features, denoted Ⓜ and Ⓕ, respectively, in Table 15.29, offer performance which is better than either laughter type alone but not better than all laughter $\mathcal{L}$, indicating that the voicing distinction hurts performance on this task. It should be noted that model-space and feature-space combinations can involve up to 6 interlocutors, since interlocutors are ranked independently for $\mathcal{L}_V$ and $\mathcal{L}_U$ feature computation according to their amount of $\mathcal{L}_V$ and $\mathcal{L}_U$ time. The alternative feature-computation-space combination, using $\mathcal{L}$ to rank interlocutors when extracting $\mathcal{L}_V$ and $\mathcal{L}_U$ features, is shown in line 6 of the table as Ⓒ, and also does not outperform modeling all $\mathcal{L}$ in a single model for DevSet.

The table also shows that speech activity context is much weaker than laughter activity context. However, modeling both speech and laughter offers improvements over modeling laughter alone. The speech context appears to offer complimentary information for predicting attempts at humor.

15.6.5 Performance on Unseen Data

Although laughter appears to be almost as relevant to humor detection in EvalSet as in DevSet, several of the above mentioned trends do not generalize to this set. In particular, voiced laughter does appear to be better than all laughter. As a result, model-space, feature-space, and feature-computation-space combinations perform better than does all laughter, but in this case not better than only voiced laughter. It may be that these differences between DevSet and EvalSet performance are due to the complexity of the feature extraction regions employed here, which were proposed for modeling speech context in Chapter 14 and have not been re-optimized for the current task.

15.6.6 Combining Vocal Activity Context Features with Lexical Features

When all three of the LEX system, the $\mathcal{L}$ system, and the $\mathcal{S}$ system are combined in model space, the results are as shown in Table 15.30. The combination exhibits a relative reduction in error, over the best system $\mathcal{L}$, of 23% for DevSet and 18% for EvalSet. At the minimum ER point, the $\mathcal{S}$ and LEX systems appear to lower the false alarm rate otherwise incurred by the $\mathcal{L}$ system alone.

System	DevSet			EvalSet		
	FA	MS	ER	FA	MS	ER
LEX	12.7	67.0	79.6	12.8	70.5	83.3
$\mathcal{S}$	7.5	47.4	54.9	8.6	62.8	71.4
$\mathcal{L}$	14.0	5.3	19.3	15.6	8.1	23.7
ALL	7.7	7.2	14.8	8.3	11.0	19.4

Table 15.30: Detection performance, in %, of the model-space combination (ALL) of three systems employing topology T1; FA is the false alarm rate, MS is the miss rate, and ER = FA + MS.

15.6.7 Receiver Operating Characteristics

To describe system performance at locations other than the ER minimum, ROC curves are constructed as the convex hull of FA and MS error pairs observed during DevSet tuning. They are shown in Figure 15.14, together with the line of no discrimination and the equal error line for which FA = MS.

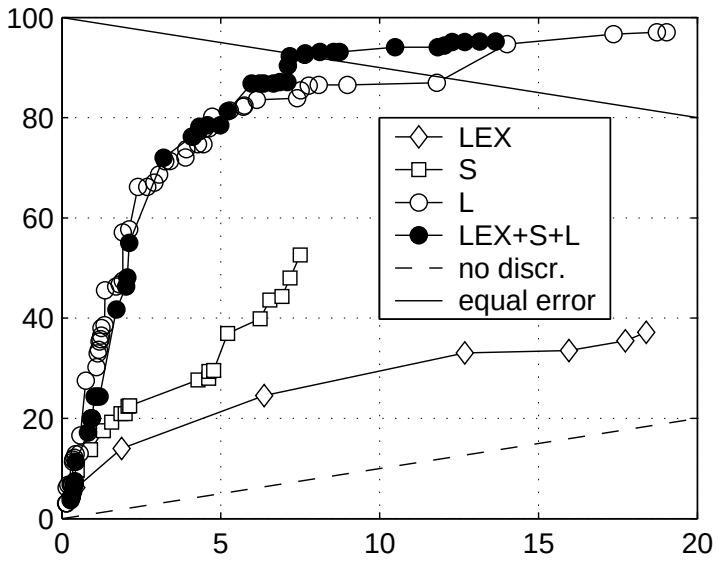


Figure 15.14: Receiver operating characteristic curves for 4 systems employing topology T1, produced using DevSet; also shown is the line of no discrimination and the equal error rate line. False positive and true positive rates shown in % along the x- and y- axes, respectively.

As can be seen, lexical features offer performance above the line of no discrimination over the whole range observed, but performance is much poorer than for any other system explored. Laughter context offers significantly better performance, approximately quadrupling the lexical system true positive rate at the same false positive rates. Speech context

features yield performance which is intermediate between lexical features and laughter context features. The model-space combination of all three systems follows the $\mathcal{L}$-only curve at low false positive rates, but near the equal error rate point achieves miss rates and false alarm rates which are both approximately 5% absolute lower than for the $\mathcal{L}$-only system.

15.6.8 Model Analysis

To analyze what the laughter context models actually learn, a new set of models is inferred without the LDA transform (the latter makes analysis more difficult). Figure 15.15 shows the laughter context emission probability for a DA terminating at time t. The temporal distribution of laughter from the interlocutor who laughs the most in the $[t-5, t+5]$-second window is given in panel (a) for DAs labeled as an attempt at humor, and in panel (b) for all other DA types. Similarly, panels (c) and (d) show the same distributions as (a) and (b), respectively, for the interlocutor who laughs the second most in the $[t-5, t+5]$-second window. As can be seen, attempts at humor are quite different from DAs not so labeled, in terms of how much the two most laughing interlocutors laugh. The peaks in the distributions of (a) and (c) occur just after completion of the humorous DA, with the most laughing interlocutor being more likely to laugh than not. However, panels (a) and (c) also show that interlocutors laugh a significant amount prior to humorous DA completion.

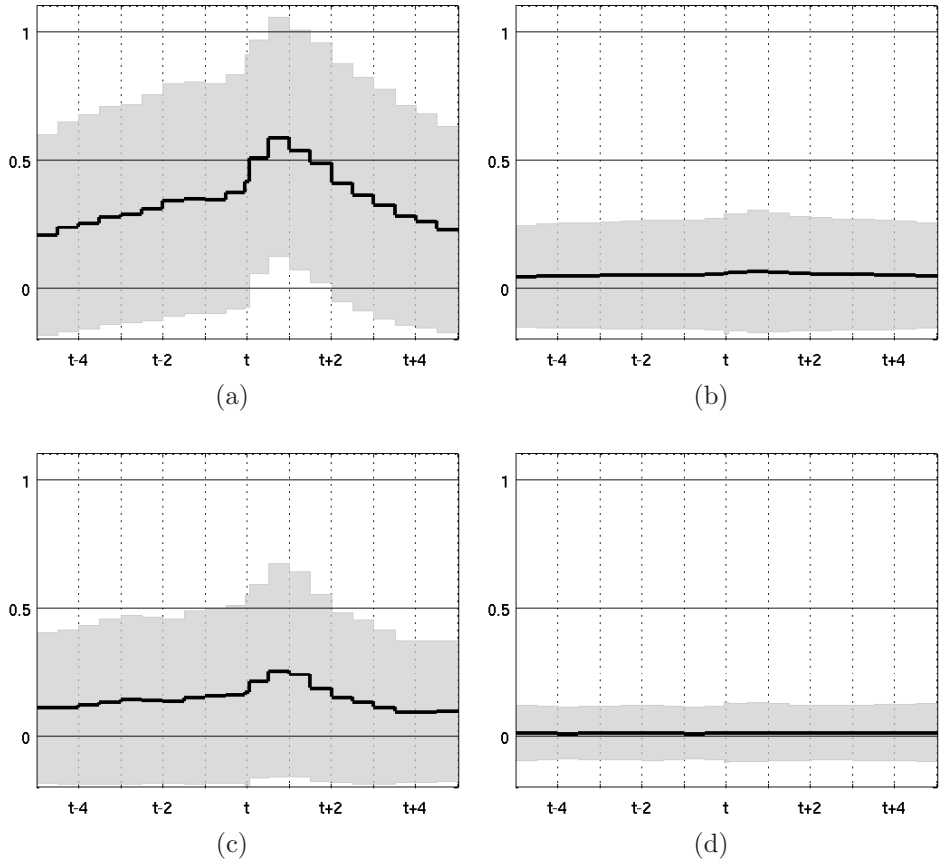


Figure 15.15: Single-Gaussian emission probabilities, in speech states completing DAs, of the raw laughter context produced by the most laughing interlocutor (panels (a) and (b)) and the second-most laughing interlocutor (panels (c) and (d)). Panels (a) and (c) pertain to humorous DAs, while panels (b) and (d) pertain to a model which, for the purposes of analysis, was trained on all other DA types. The x-axis shows time in seconds; probabilities in $[0, 1]$ are shown along the y-axis, with the mean in black and the gray area showing one standard deviation away from the mean.

Figure 15.16 depicts the laughter context emission probability for a participant terminating a DA at time t, for

laughter from him-/her- self only. Somewhat surprisingly, the amount of laughter produced by the participant completing a humourous DA is almost as high as for the most laughing interlocutor, and significantly higher than for the second most laughing interlocutor. This suggests that laughter, like speech, may occur predominantly in dyads. Closer inspection of Figure 15.15(a) and Figure 15.16(a) indicates that the temporal distribution of laughter for the teller of the humorous DA is almost identical to that of the most laughing interlocutor, everywhere in the $[-5, +5]$-second context of that DA's completion except during the interval during which the teller is preoccupied with speaking – from approximately $t - 3$ to t seconds, given a `j` DA terminating at instant t.

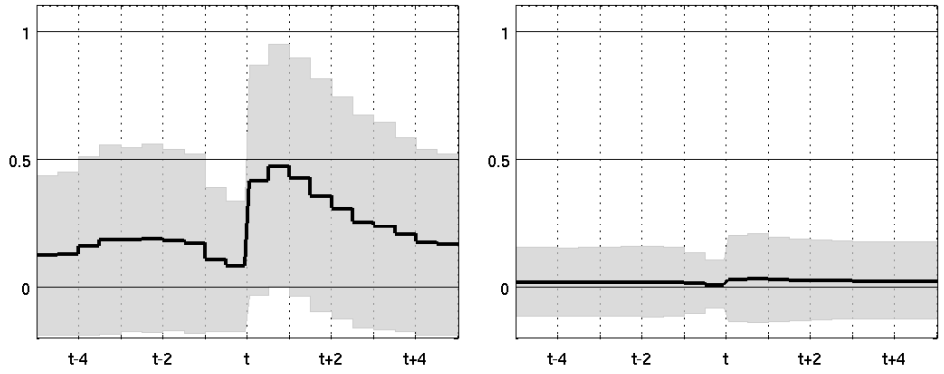


Figure 15.16: Single-Gaussian emission probabilities, in speech states completing DAs, of the raw laughter context produced by the participant completing the DA. Panel (a) pertains to humorous DAs while panel (b) pertains to a model trained on all other DA types. Axes as in Figure 15.15.

15.6.9 Assuming Correct Segmentation

The results presented so far have been frame-level detection errors. This is appropriate in the proposed setting, since the decoder is not exposed to reference DA segmentation and must explicitly segment talkspurts while at the same time classifying them into DA types. However, a frame-level detection error obscures *how many* attempts at humor are detected, as opposed to by how much they are missegmented. To shed light on this issue, the decoder is exposed to reference DA boundaries by disallowing DA-terminal frames from aligning to non-DA-terminal topology states and vice versa. The best Viterbi path can then be scored as done previously, at the frame-level, as well as at the DA level, allowing for comparison between the two metrics.

The results are shown in Table 15.31. The first panel duplicates the results from Table 15.30 for ease of comparison. In the second panel, the original model parameters are retained but DA boundaries are additionally forced-aligned during decoding. ER minima for all three of LEX, $\mathcal{L}$, and $\mathcal{S}$ systems, as well as for their model-space combination, are higher in this condition than when DA boundaries are unknown, due largely to significantly higher miss rates. This suggests that the decoder without forced-alignment of true boundaries oversegments `j` productions and correctly classifies some of the shorter segments as `j`, but that, when inserting DA boundaries is not allowed, it classifies the resulting longer `j` segments as $\neg$`j`. A possible explanation is that DAs labeled as `j` may exhibit intention to amuse during only a fraction of their duration; further analysis is required to assess the extent to which this might be the case.

The third panel in Table 15.31 shows performance when parameters of the three individual systems and of their model-space combination are re-optimized using DevSet for the condition in which DA boundaries are known, and scoring at the DA, rather than the frame, level. Relative to the second panel, DA-level error rates are smaller for DevSet and higher for EvalSet (except for the LEX system).

System	DevSet			EvalSet		
	FA	MS	ER	FA	MS	ER
As in Table 15.30:						
LEX	12.7	67.0	79.6	12.8	70.5	83.3
$\mathcal{L}$	14.0	5.3	19.3	15.6	8.1	23.7
$\mathcal{S}$	7.5	47.4	54.9	8.6	62.8	71.4
ALL	7.7	7.2	14.8	8.3	11.0	19.4
Adding DA boundary information:						
LEX	4.7	79.6	84.3	5.1	77.1	82.2
$\mathcal{L}$	7.3	26.5	33.8	8.7	18.2	26.9
$\mathcal{S}$	6.5	48.6	55.1	5.5	66.5	72.0
ALL	7.0	20.1	27.1	5.8	24.0	29.8
Scoring at the DA-level:						
LEX	7.9	66.7	74.5	8.1	61.9	70.0
$\mathcal{L}$	14.4	12.9	27.3	14.9	19.1	34.0
$\mathcal{S}$	7.3	50.5	57.8	8.1	60.0	68.1
ALL	7.6	16.1	23.7	8.1	20.0	28.1

Table 15.31: Frame-level detection performance in %, for three systems and their model-space combination (ALL), without DA boundary information, with DA boundary information, and with DA boundary information and reoptimized model parameters, the latter scored at the DA-level. FA is the false alarm rate, MS is the miss rate, and ER = FA + MS.

15.6.10 Inviting Laughter with Laughter

The analysis in Section 15.6.8, and in particular Figure 15.16, indicates that those attempting humor themselves laugh. This is especially true immediately following DA completion. In light of this, an alternative system is constructed which excludes laughter from interlocutors and uses only the laughter-context from the participant whose DA productions are being decoded. Detection scores when using this system, with parameters reoptimized for this new task, are shown as $\mathcal{L}'$ in Table 15.32. Performance is lower than when other laughers are considered, but only by 6.3% on unseen EvalSet data, and still considerably lower than when either lexical or speech context features are employed instead.

System	DevSet			EvalSet		
	FA	MS	ER	FA	MS	ER
$\mathcal{L}$	14.0	5.3	19.3	15.6	8.1	23.7
$\mathcal{L}'$	8.7	20.3	28.9	8.5	22.4	31.0

Table 15.32: Detection performance, in %, for the system relying on laughter context from the speaker and the three most laughing interlocutors ($\mathcal{L}$, as in Table 15.29), as well as an alternative system relying on laughter context from the speaker only ($\mathcal{L}'$). FA is the false alarm rate, MS is the miss rate, and ER = FA + MS.

These results indicate that those making attempts at humor communicate their intent by laughing themselves, signaling to interlocutors that it is appropriate for them to take up laughter. Although this finding corroborates qualitative studies in the literature [76], the observed level of performance on the selected corpus is surprising. It suggests that meetings may provide an environment in which producers of `j` DAs deliberately perform additional work (by deploying laughter) to limit the potential ambiguity of their intent, more so than in non-work-oriented conversation.

Furthermore, because producers of `j` DAs are more likely to laugh following DA completion than any but one other interlocutor (cf. Figures 15.15 and 15.16), such DAs appear to be directed at specific other participants rather than the group as a whole. As additional evidence of its dyadic nature, Figure 15.17 depicts the temporal distribution of `j` talk for participants completing a `j` DA, their most `j`-talkative interlocutor, and their second-most `j`-talkative interlocutor, in

panels (a), (b), and (c), respectively, of Figure 15.17.

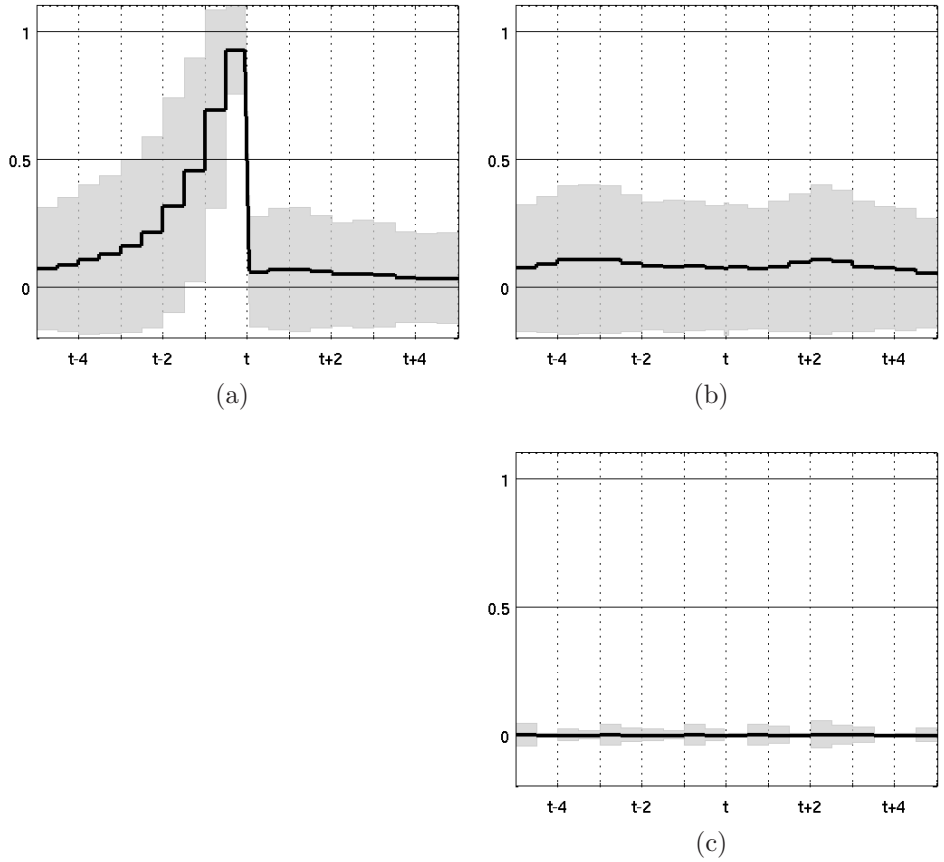


Figure 15.17: Single-Gaussian emission probabilities, for speech states completing j DAs, of the raw j-speech context produced by the participant completing the DA, in panel (a), her or his most-j-speaking interlocutor, in panel (b), and her or his second-most-j-speaking interlocutor, in panel (c). Axes as in Figure 15.15.

As can be seen, inside of the $[-5, +5]$-second context of a terminating j DA, the second-most j-talkative interlocutor is unlikely to be active, i.e. at most two participants produce j talk in any 10 second window centered on a j-terminal DA boundary. Furthermore, the most j-talkative interlocutor has two maxima, at approximately $t - 3$ s and $t + 2.5$ s, indicating that they are likely to be producing j talk before the current DA, following the current DA, or both. Finally, the interval between maxima in the probability of j talk from different participants appears to have a most likely value of approximately 3 seconds. This regularity, which is not explicitly modeled by the explored systems, may improve the detection of attempts at humor in future systems.

15.7 Potential Impact

The impact of this chapter may be in any one or more of four related areas.

First, the annotation effort presented in Section 15.3 has some potential lessons for those planning to annotate similar material. The choice of naive or trained annotators may be relevant to the specific task required of a computational system; as argued here, naive annotators may be the most suitable choice if the final system is a browser in which users enter arbitrary search strings. It appears that the most natural strings for describing emotional phenomena are descriptors of behavior rather than of state. In cases where it is sufficient to determine *whether* an utterance is emotionally inflected,

annotators (and users of browsers) may be expected to agree much more than if also *how* the utterance is inflected is important. Finally, even if a simple three-class ordinal scale is selected, such as that for emotional valence in this chapter, the presence of laughter is likely to strongly influence annotator perception, and annotation scheme design should take that into account.

The influence of the presence of laughter leads to the second potential impact, for the detection, classification, or recognition of emotion or emotional epiphenomena. As shown for the ISL Meeting Corpus, approximately 3% of utterances are of negative valence, and approximately 18% are of positive valence (for EvalSet; for the corpus as a whole, the numbers are 1% and 17%, respectively, as shown in Table 15.4). But 16% of all utterances in EvalSet contain laughter, a proportion that far outweights those utterances which are of positive valence but contain no laughter, as well as those utterances which are of negative valence (of which some contain laughter). This effect further skews an already skewed prior. The classification of valence, at least, a task thought to be much harder than the classification of activation, is likely for the foreseeable future to become an augmented laughter detection task. To move beyond this, emotional valence classification will need to consider utterances containing laughter, or consisting only of laughter where applicable, separately from other utterances, which will require robust laughter detection in the target domain.

Third, as suggested by the experiments and analysis in Section 15.4, prosodic and spectral features appear capable of identifying positive valence utterances, but seemingly only those which contain laughter. If distributions of such features differ between laughter-bearing and laughter-non-bearing utterances, sufficiently to allow for inference of laughter, then it may be necessary to detect and mask out laughter in general studies of speech prosody. When present, laughter appears to have the ability to significantly alter the statistics of prosodic models. For example, the correlates between speech prosody and involvement reported in [225] may in fact be due to the presence of speech laughter in some of the involved dialog acts (and of non-speech laughter, since DAs include inter-word gaps which may bracket laughter); the correlates may actually be between laughter prosody and involvement.

Finally, the three computational tasks explored in this chapter are first of their kind in the domain of meetings or multi-party conversation in general. In the very least they provide baselines which future systems — exploring other sources of information or modeling vocal activity in more appropriate ways — can rely on. Also, since laughter appears to be relevant to all three, they offer an implicit downstream application-level measure of laughter detector performance. This is much needed because efforts in laughter detection (cf. Chapter 13), which currently score performance using metrics such as classification accuracy or F-score, tend to include all laughter in test sets or, worse, only that laughter which is distant from the laugher's speech. As shown in Section 15.5, laughter most proximate to the laugher's speech is most relevant to the detection of the presence of involved speech. The work in this chapter has the potential to focus laughter detection on the most emotionally salient laughter.

15.8 Relevance to Other Chapters

The finding in Section 15.3 regarding the frequency of laughter in the ISL Meeting Corpus, and in Section 15.4 regarding its dominating role in the classification of valence, led to the effort of segmenting laughter in the much larger ICSI Meeting Corpus described in Chapter 12. The experiments described in Section 15.4 were among some of the earlier work in this thesis, and bear relatively little resemblance to other chapters as a result.

Experiments in detecting both involved speech (Section 15.5) and attempts to amuse (Section 15.6) relied on the laughter segmentation produced as described in Chapter 12. The concept of interaction temperature, used in involved speech detection, was developed in Chapter 7, and is a natural extension to the Ising antiferromagnet model described there. In contrast, the infrastructure for detecting joking in Section 15.6 is a natural extension of the dialog act recognizer of Chapter 14.

15.9 Summary

This chapter has explored the role that emotional epiphenomena have on the distribution of vocal activity, in time and across participants. Rather than merely computing the association between vocal interaction and those phenomena, potentially uncovering statistically significant association, attempts were made to determine whether the association is also strong enough to allow for the inference of epiemotional phenomena.

The chapter began with an effort to annotate a large corpus of multi-party conversation. It was found that untrained annotators, when left to their own devices, produced descriptors of *behavior* rather than of *state*. This is important, particularly for browsing conversation corpora, in which entry of an arbitrary search string by an untrained user yields a scenario likely quite similar to the described open label annotation effort. It was also shown that annotators tend to agree that at least 5% of utterances are emotionally inflected (cf. Table 15.1), and that this proportion may reach as high as 25% for the most informal conversations. However, annotators exhibit low agreement on the precise quality of emotional inflection.

Annotators tend to agree, to a level which is on par with other subjective labeling tasks, on the polarity of emotional valence in utterances ($\kappa \in [0.46, 0.67]$, cf. Tables 15.5 and 15.6), when labelers ignoring laughter-only utterances are excluded. Laughter appears to be responsible for the overwhelming proportion of agreement on positive valence utterances (given annotators who were not present in the conversations themselves). The ISL Meeting Corpus, used in this effort, was found by a majority of annotators to consist of 3% of utterances of negative valence, 79–81% of utterances of neutral valence, and 16–18% of utterances of positive valence.

Automatic classification of emotional valence for manually segmented utterances, the first computational task studied in this chapter, achieved an accuracy of 91.32%, where majority-class guessing would have led to 78.67%. This represents a 59%rel reduction in classification error. Over a third of the remaining error is due to an inability to identify *any* negative valence utterances. It was shown that an accuracy of 91.25% is achievable by a single feature, the number of transcribed laughter instances within the utterance. When manual transcription is not used, acoustic features yield an accuracy of only 84.05%. Analysis revealed that the features automatically selected from a very large and comprehensive set of acoustic features were only helping for those utterances which happened to contain laughter; in the absence of laughter, the selected acoustic features, as well as lexical perplexity features, always led to a hypothesis of neutral valence. This is likely due to the fact that laughter is present in the majority of positive valence utterances. The described experiments did not shed light on how to detect positive valence in utterances not containing laughter, how to detect non-positive valence in utterances containing laughter, or how to detect negative valence.

The second computational task studied was the detection of the presence of speech exhibiting marked emotional involvement or activation, as annotated in the ICSI Meeting Corpus in which 24.8% of all 60-second intervals are involvement "hotspots". An accuracy of 85.6% and an F-score of 67.1% were achieved by extracting simple static features from a segmentation of speech laughter, i.e. laughter which is produced simultaneously with speech by the same participant. It was argued that speech laughter may be only unreliably detected in fully automatic settings, and the accuracy achieved by using all non-speech voiced laughter instead was shown to be 81.4%. However, annotating laughter with its proximity to the laughter's speech improves that number to 84.4%, and yields an F-score of 70.3%. The most relevant laughter for this task appears to be that found immediately adjacent to the laugher's speech both before and after it, which is not surprising since speech laughter is likely to be prosodically marked as involved and laughter by itself can be expected to follow and/or precede speech laughter in many cases.

Finally, this chapter explored the detection of attempts to amuse by extending the DA recognition system of Chapter 14. Speech implementing this function accounts for 0.50–0.75% of time in the ICSI Meeting Corpus which was used. It was shown that lexical bigram features, even when true words are known, offers relatively poor performance, namely an error rate (the sum of the miss and false alarm rates) of 83.3%. Features describing the local speech activity neighborhood, on the other hand, yield error rates of 71.4%, while those describing the local laughter activity neighborhood yield error rates of 23.7%. Score-level fusion of all features achieves a 19.4% error. Interestingly, auxillary experiments show that if only the target participant's laughter is used to extract features, an error rate of 31.0% can be achieved. This indicates that those telling jokes in meetings frequently laugh themselves, possibly to indicate to others that laughter is anticipated and to limit confusion and/or embarrassment. Finally, an analysis of trained models seems to indicate that for the majority of attempts to amuse, only one interlocutor laughs within a $[-5, +5]$-second context, and only one also jokes within that context, despite the fact that all the conversations studied involved more than two participants.

While the role of laughter as a cue to a participant's emotional state, and to the efforts by interlocutors to affect it, is not generally surprising, the work in this thesis is the first, for all three tasks, to quantify that role by building computational inference systems. The findings establish that laughter not only can play a role, but may also be the most relevant or even the only emotional cue of import in social conversational settings. The tasks explored also provide a downstream constraint on research in laughter detection, which is currently evaluated using metrics which ignore potential utility to applications.

15.10 Future Directions

Each of the three computational tasks explored in this chapter deserve further attention and closer scrutiny.

In valence detection, an important immediate next step is to repeat experiments following acoustic feature selection separately, for utterances known to contain laughter and for those known not to contain it. As the analysis in Section 15.4 showed, the selected feature set does not yield accuracies which exceed those obtained by always guessing NEUTRAL on utterances containing no laughter. In tandem, since separating out laughter-bearing utterances will exacerbate skewness of class priors, techniques such as sampling and bagging could be applied which are directly aimed at unbalanced class problems. Finally, experiments using the proposed lexical features should be duplicated after removal of all nonverbal productions, since their current performance may be due to the fact that laughter tokens are included. The lexical feature set could also be augmented with more detailed features such as the likelihoods of specific n-grams, to yield a more appropriate benchmark for performance when words are known.

Although laughter was shown to yield good detection results for involved speech hotspots, it appears that approximately 15% of 60-second intervals cannot be classified correctly using laughter in the way proposed. This creates an opportunity for future research, in attempting to describe the remaining intervals. It is possible that the DA recognition system of Chapter 14 should simply be extended to allow for yet other DA types, namely that of speech which is considered involved. This would also simplify the inclusion of lexical n-gram features for this task. However, as for the classification of valence, there is potential that using the DA recognition approach will further imbalance already imbalanced priors. Another possibility is to revisit the features drawn from the speech segmentation. Although neither static nor dynamic features drawn from the speech segmentation seem to yield accuracies significantly in excess of majority-class guessing, the training of a turn-taking model (cf. Chapter 10) for intervals containing involved speech and intervals not containing involved speech may be more successful. Embedding such models as emission probability models, in a non-ergodic HMM framework, may prove better suited to the task (Section 15.5).

The experiments described for the detection of attempts to amuse should be rescored using the framework in Chapter 14, since false alarm rates as computed here penalize insertions by the total amount of non-`j` speech — a large quantity given the small proportion of speech that `j` accounts for by time. Also, the results which were achieved with true segmentation, yielding worse performance than without it, suggest that the laughter context may be indicative of only the final DAs of a joke (which may consist of multiple DAs). The successful detection of attempts at humor, in their entirety, may necessitate duplicate `j` dialog act topologies in sequence, with contrastive emission probability models attached to their individual states.

Chapter 16

Text-Independent Conversation Characterization*

16.1 Introduction

Estimating the likelihood $P(\mathbf{Q} \mid \mathbf{\Theta})$ of a particular conversation's occurrence of speech activity $\mathbf{Q}$, in time and across participants, has so far in this thesis relied on a general model $\mathbf{\Theta}$, whose parameters were inferred from many other conversations. Particularly Chapters 10 and 11 have demonstrated the utility of estimates computed in this way. What has not been shown is how those estimates vary as a function of the training corpus.

As in language modeling, it frequently occurs that, in addition to knowing the sequence of words, one also knows something about the source of that sequence, i.e., its *type*. An example of type, as implied here for text, is *document type* (e.g., newspaper, web page, novel, etc.). An arguably better likelihood estimate in such circumstances is one in which the model used has been trained on other documents of only the *same* type. Similarly, in models of speech activity occurrence, better likelihood estimates may be obtained when models have been trained on other conversations of the same type, rather than on all types. Examples of conversation type may include more or less formal encounters, and may differentiate in the number of central figures, who either produce the majority of contributions or receive them.

That likelihoods should be higher when the training corpus better matches the test conditions is of course not surprising. But the extent to which models reflect the conversational pecularities of their training corpora is highly relevant. A potential application which relies on *type-specific* models $\mathbf{\Theta}_u$, for a type $u \in \mathcal{U}$, is the *inference* of type when the latter is not known for an unseen conversation. Automatic labeling of conversation type can be quite useful; it can, for example, be used to aid in the browsing and indexing of collections of conversation, particularly when such collections are large and labeling each conversation manually would be prohibitively costly. Inference is quite straightforward when $\mathbf{Q}$ is observed,

$$\begin{aligned} u^* &= \arg\max_{u \in \mathcal{U}} P(u \mid \mathbf{Q}) \\ &= \arg\max_{u \in \mathcal{U}} P(u) \cdot P(\mathbf{Q} \mid u) \\ &\doteq \arg\max_{u \in \mathcal{U}} P(u) \cdot P(\mathbf{Q} \mid \mathbf{\Theta}_u) \end{aligned} \tag{16.1}$$

provided that $\mathbf{\Theta}_u$ *successfully captures discriminating aspects* of the various alternatives in $\mathcal{U}$.

The current chapter explores precisely this problem, namely the inference of conversation type given an observed $\mathbf{Q}$. (When $\mathbf{Q}$ is not observed, it must be inferred from the acoustic observable $\mathbf{X}$; in this chapter it is assumed that this earlier inference step has already been performed.) Experimental results indicate that multi-party meetings vary systematically and considerably across project groups, even within a single organization. The chapter explores models which capture holistic group behavior, as well as those which differentiate among the attendees, and shows that combining both model types is beneficial.

The work presented here was published in [145].

*The work in this chapter conducted in collaboration with Mari Ostendorf and Tanja Schultz.

16.2 Related Work

Although it is generally accepted that conversations, and meetings, exhibit variable levels of interaction, there appears to be no literature on automatically classifying different conversation types in their entirety, whether using interaction-related characteristics or otherwise. However, a prototypical system could be expected to be based on the ontology of *speech-exchange systems* proposed in [194], consisting of things such as interactive seminar, debate, formal business meeting, or informal chat. Several variants of these have been characterized in early computational work [125]. In [30], a description of several meeting styles, in terms of observable characteristics, was provided.

Most similar conceptually to the proposed task is the recognition, in time, of group actions in meetings [169], such as "note-taking" or "whiteboard presentation". Although not necessarily stemming from a conduct of conversation in general, these group actions are likely to differ in their interaction patterns. Also related to the current work is the detection of interaction groups in meetings [27].

In the data studied in this chapter, namely the ICSI Meeting Corpus, several observations have been made which have bearing on the proposed approach and task. [203] showed that the meetings of two different groups exhibited large differences in the amount of overlapped speech. When averaged over 5 `Bmr` meetings and 3 `Bro` meetings, 17.0% and 8.8% of words, respectively, were shown to incur some overlap. When backchannels were excluded, the difference persisted; 14.1% and 5.6%, respectively, were produced in overlap. These measurements were not extended to all `Bmr` and all `Bro` meetings, or other meeting types.

16.3 Dataset Use

The experiments in what follows rely on the ICSI Meeting Corpus, described in Section 4.1. For the purposes of this chapter, *conversation type* is equated with *project group* which held the meeting. This is arguably a more interesting partitioning of a corpus, particularly for organizations which hold a lot of meetings (and which record them). The implicit assumption made here is that projects vary not only in participants, who may engage in distinct interaction patterns, but also in the *needs* for group communication. Groups assigned to projects may tend to meet in order to negotiate future actions, to report on their past findings, or merely to maintain affiliation. As such, the interaction patterns observed in any particular meeting can be expected to reflect both the composition of the group and the "type" of conversational (turn-taking) style.

The DataSet in the presented experiments consists of 67 of the 75 meetings in the corpus, representing the longitudinal recording of three project groups: `Bed`, `Bmr`, and `Bro`; the remaining 8 meetings in the corpus were held for other projects, for each of which too few meetings have been recorded to model. The participants attending each meeting type, $u \in \mathcal{U} \equiv \{\texttt{Bed}, \texttt{Bmr}, \texttt{Bro}\}$, are not always the same, but the three populations are broadly distinct. Only three participants attend some meetings of both type `Bmr` and type `Bro`, and one participant attends some meetings of both type `Bed` and type `Bmr`.

Because only 67 meetings are considered, they are not split into distinct training, development, and evaluation sets. Instead, all experiments are conducted in a round-robin fashion. A fold consists of a single meeting, with all other meetings used for training.

Table 16.1 shows several characteristics of DataSet. As can be seen, the prior over the three meeting types is not uniform; in particular, `Bmr` meetings make up 43% of all DataSet meetings by number. It can also be seen that the most likely number of participants to each meeting type is quite similar, either 6 or 7, and that `Bmr` meetings exhibit the largest range in participant number, having both the smallest minimum and the largest maximum. Finally, Table 16.1 indicates that `Bmr` meetings also exhibit the largest variability in participant identity, drawing a sample for each meeting which is similar in size to that of the other two meeting types, but from a larger population $\mathcal{P}$ of potential attendees.

For each meeting, a continuous vocal activity segmentation is formed from the the continuous speech/non-speech segmentation $\mathcal{S}$, without inter-talkspurt gap padding. This continuous segmentation is then discretized at a frame step of 100 ms and a frame size of 200 ms, as described in Subsection 6.2.2, to yield $\mathbf{Q}$.

Meeting Type u	Number of Meetings	Size of Attendee Population $\|\mathcal{P}\|$	Numer of Participants K		
			mod	min	max
`Bed`	15	13	6	4	7
`Bmr`	29	15	7	3	9
`Bro`	23	10	6	4	8

Table 16.1: Overt characteristics of three ICSI meeting types; "mod", "min", and "max" are the most likely, the minimum, and the maximum number of participants, respectively, across meetings of type u.

16.4 Assessment of Performance

Since inference of meeting type for each of the 67 meetings yields one type per meeting, the most natural metric for assessing performance is classification accuracy, over 67 round-robin trials.

16.5 Baseline

The task of classifying meeting type in the ICSI Meeting Corpus has not been previously attempted, and no baselines are readily available from the literature.

To aid in interpreting presented classification accuracy, several naive system baselines are described here. First, stochastic guessing under the assumption of a uniform prior across the three meeting types yields an accuracy of

$$\begin{aligned} A_{sto}^{flat} &= \frac{1}{67}\left(\left\{15\left(\frac{1}{3}\right)\right\}+\left\{29\left(\frac{1}{3}\right)\right\}+\left\{23\left(\frac{1}{3}\right)\right\}\right) \\ &= 34.3\% \,, \end{aligned}$$

where $\{\cdot\}$ represents rounding to the nearest integer. Although the prior over the three classes of meeting type is not uniform, taking it into account does not change the accuracy of guessing,

$$\begin{aligned} A_{sto}^{prior} &= \frac{1}{67}\left(\left\{15\left(\frac{14}{66}\right)\right\}+\left\{29\left(\frac{28}{66}\right)\right\}+\left\{23\left(\frac{22}{66}\right)\right\}\right) \\ &= 34.3\% \,. \end{aligned}$$

Finally, deterministically always choosing the majority class leads to

$$\begin{aligned} A_{det}^{maj} &= \frac{1}{67}(29) \\ &= 43.3\% \,. \end{aligned}$$

16.6 Assuming Participants to Be Identical

Applying Equation 16.1 to each meeting, after training type-specific EDO models for $u \in \mathcal{U} \equiv \{\texttt{Bed}, \texttt{Bmr}, \texttt{Bro}\}$ using all the other meetings in DataSet in round-robin fashion, yields a confusion matrix as shown in Table 16.2. All EDO models were trained using $K_{max} = 4$ (preliminary experiments indicated that this yielded the best match of training corpus size and model complexity in terms of number of parameters; for $K_{max} > 3$, differences appear to be numerically negligible in any case, cf. Table 10.6).

The classification accuracy is 79.1%. As the table shows, 93% of `Bmr` meetings are classified correctly, while only 53% and 78% of `Bed` and `Bro` are, respectively. Recall rates are positively correlated with distribution priors over the three types, suggesting that the 14 `Bed` meetings available for training the `Bed` model in any fold may be too few in number, given average meeting durations in the ICSI corpus.

Actual Type	Estimated		
	Bed	Bmr	Bro
Bed	8	3	4
Bmr	1	27	1
Bro	4	1	18

Table 16.2: Confusion matrix among the three ICSI meeting types studied, for *maximum a posteriori* classification using 3 type-specific EDO models and $P(u)$ given by the number of meetings in DataSet less one.

16.7 Assuming Participants to Be Different

EDO models treat all participants as identical, and therefore can be said to capture mean participant behavior in a particular meeting type. However, meeting types may differ in within-meeting differences observed among their participants, and potentially more so than in between-meeting similarities.

To investigate this, a model which captures and contrasts the behavior of specific participants k in $\mathbf{Q}$ needs to be employed. The most obvious and easy choice is to invoke the assumption that participant states are conditionally independent at instant t, given their joint behavior at instant $t-1$.

16.7.1 Participant Profiles

An explicit compositional model of this type was described in Subsection 6.3.2, and is denoted $\mathbf{\Theta}^{CI} \equiv \{\mathbf{\Theta}^{CI}_k\}$, $1 \leq k \leq K$. Assuming that participants are not identical entails entertaining the possibility that $\mathbf{\Theta}^{CI}_k$ not equal $\mathbf{\Theta}^{CI}_j$, for some $j \neq k$. To formalize this, the notion of participant profiles is introduced.

A *participant profile* g is an element of the set $G \equiv \{g_1, g_2, g_3, \cdots, g_{N_G}\}$. It identifies a single participant's style of vocal interaction. The multi-participant *group profile* $\mathbf{g}$ is then an ordered K-length vector whose elements $\mathbf{g}[k]$ correspond to the participants numbered 1 through K. The space from which a particular $\mathbf{g}$ is drawn is denoted $\mathbb{G}$; in the most general case, it may be assumed that participants are independently assigned profiles in $\mathbf{g}$ from G, making $\|\mathbb{G}\| = N_G^K$.

As argued in Subsection 6.3.4, a compositional conditionally independent model is not $\mathbf{R}$-invariant when participants are not identical. Rotating the rows of $\mathbf{Q}$ using an arbitrary rotation $\mathbf{R}$ requires that the submodels of $\mathbf{\Theta}^{CI} \equiv \{\mathbf{\Theta}^{CI}_{\mathbf{g}[k]}\}$ also be rotated to yield $\mathbf{\Theta}^{CI} \equiv \{\mathbf{\Theta}^{CI}_{(\mathbf{R}\cdot\mathbf{g})[k]}\}$ (it additionally requires the rotation of the conditioning state space when factoring $\mathbf{Q}$ in time). This presents few problems when $\mathbf{g} = \mathbf{g}^*$ is known, and Equation 16.1 may be rewritten as

$$\begin{aligned} u^* &= \arg\max_{u \in \mathcal{U}} P(u \,|\, \mathbf{Q}, \mathbf{g}^*) \\ &= \arg\max_{u \in \mathcal{U}} P(u) \cdot P(\mathbf{g}^* \,|\, u) \cdot P(\mathbf{Q} \,|\, \mathbf{g}^*, u) \\ &\doteq \arg\max_{u \in \mathcal{U}} P(u) \cdot P(\mathbf{g}^* \,|\, u) \cdot P\left(\mathbf{Q} \,|\, \mathbf{\Theta}^{CI}_{u,\mathbf{g}^*}\right) . \end{aligned} \tag{16.2}$$

However, $\mathbf{g}^*$ is, in general, not known. To avoid estimating it separately, estimation of u must marginalize over the alternatives in $\mathbb{G}$,

$$\begin{aligned} u^* &= \arg\max_{u \in \mathcal{U}} \sum_{\mathbf{g} \in \mathbb{G}} P(u, \mathbf{g} \,|\, \mathbf{Q}) \\ &= \arg\max_{u \in \mathcal{U}} \sum_{\mathbf{g} \in \mathbb{G}} P(u) \cdot P(\mathbf{g} \,|\, u) \cdot P(\mathbf{Q} \,|\, \mathbf{g}, u) \\ &\doteq \arg\max_{u \in \mathcal{U}} P(u) \sum_{\mathbf{g} \in \mathbb{G}} P(\mathbf{g} \,|\, u) \cdot P\left(\mathbf{Q} \,|\, \mathbf{\Theta}^{CI}_{u,\mathbf{g}}\right) . \end{aligned} \tag{16.3}$$

16.7.2 Hypermodels over Model Parameters

Equation 16.3 indicates that a $K \times T$ vocal activity record $\mathbf{Q}$ must be evaluated by N_G^K different models. Since T may be quite large, there is scope for improving time complexity by *not* iterating over $\mathbf{Q}$, in time, for each alternative in $\mathbb{G}$.

Instead, this chapter proposes to infer from each unlabeled conversation $\mathbf{Q}$ a model $\mathbf{\Theta}^{CI}(\mathbf{Q})$, and then to score the parameters of that model as features against u-conditioned hypermodels of those parameters. As a further simplification, only those parameters that govern each participant's behavior in the context of *at most one other* vocalizing participant at any $t-1$ are considered. This limits the representation of observed interaction in $\mathbf{Q}$ to what individual participants do independently, and how pairs of participants interact. Finally, the model parameters in this reduced set, comprising a feature vector $\mathbf{F}\left(\mathbf{\Theta}^{CI}(\mathbf{Q})\right)$, are assumed to be mutually independent. This makes it possible to train models using conversations of arbitrary K. More formally,

$$\mathbf{F}\left(\mathbf{\Theta}^{CI}(\mathbf{Q})\right) = \bigcup_{k=1}^{K} \left\{ f_k^{(1)}, \bigcup_{j=1}^{K} \left\{ f_{k,j}^{(2)} \right\} \right\} , \tag{16.4}$$

where $f_k^{(1)}$ are selected $\mathbf{\Theta}^{CD}_{\mathbf{g}[k]}$ parameters which describe the kth participant's behavior in the absence of speech activity from other participants and $f_{k,j}^{(2)}$ are selected $\mathbf{\Theta}^{CD}_{\mathbf{g}[k]}$ parameters which describe the kth participant's behavior in the presence of speech activity from participant j only.

Under these modifications, Equation 16.3 may be rewritten as

$$\begin{aligned} u^* &= \arg\max_{u \in \mathcal{U}} \sum_{\mathbf{g} \in \mathbb{G}} P(u, \mathbf{g} \mid \mathbf{Q}) \\ &= \arg\max_{u \in \mathcal{U}} \sum_{\mathbf{g} \in \mathbb{G}} P(u) \cdot P(\mathbf{g} \mid u) \cdot P(\mathbf{Q} \mid \mathbf{g}, u) \\ &\doteq \arg\max_{u \in \mathcal{U}} P(u) \times \qquad (16.5) \\ &\quad \sum_{\mathbf{g} \in \mathbb{G}} \underbrace{P\left(\mathbf{g} \mid \mathbf{\Theta}_u^M\right)}_{\substack{\text{Membership} \\ \text{Model}}} \cdot \underbrace{P\left(\mathbf{F}\left(\mathbf{\Theta}^{CI}(\mathbf{Q})\right) \mid \mathbf{\Theta}^B_{u,\mathbf{g}}\right)}_{\substack{\text{Behavior} \\ \text{Model}}} . \qquad (16.6) \end{aligned}$$

Implementations of the *membership model* $\mathbf{\Theta}_u^M$ and the *behavior model* $\mathbf{\Theta}^B_{u,\mathbf{g}}$ are proposed following a description of the specific features being modeled.

16.7.3 Model Parameters as Vocal Interaction Features

The features $f_k^{(1)}$ and $f_{k,j}^{(2)}$ used in the experiments of this chapter are of five types. The first feature type is *participant talkativity*, and, unlike the remaining four feature types, is computed from $\mathbf{Q}$ directly rather than $\mathbf{\Theta}^{CI}(\mathbf{Q})$,

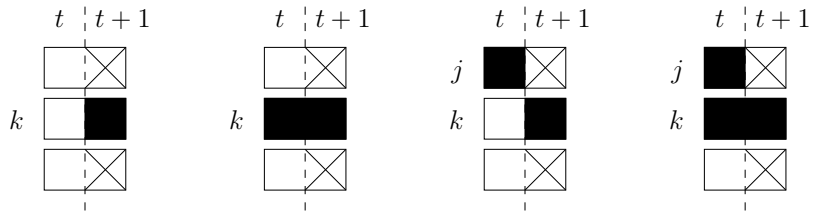
$$f_k^T = \frac{1}{T} \sum_{t=1}^{T} \delta\left(\mathbf{q}_t[k], \blacksquare\right) . \tag{16.7}$$

Each of the remaining four feature types is a subset of the parameters of $\mathbf{\Theta}^{CI}(\mathbf{Q})$, shown in Figure 16.1.

The second feature type (VI) is the probability that participant k initiates vocalization when all other participants are silent,

$$\begin{aligned} f_k^{VI} &= P\left(\mathbf{q}_{t+1}[k] = \blacksquare \mid \mathbf{q}_t[i] = \square, \mathbf{\Theta}^{CI}(\mathbf{Q})\right) \qquad (16.8) \\ &\quad \forall \;\; 1 \leq i \leq K . \end{aligned}$$

The context for this event type, and hence feature type, is shown in Figure 16.1(a). One such feature is computed for each participant k, $1 \leq k \leq K$, yielding K features of type f_k^{VI} for the conversation as a whole.



(a) f_k^{VI} (b) f_k^{VC} (c) $f_{k,j}^{OI}$ (d) $f_{k,j}^{OC}$

Figure 16.1: The context for 4 event types for which the probability of occurrence represents a feature type in $\mathbf{F}$, for a conversation with $K = 3$ participants. Participants' vocal activity tracks, evolving in time from left to right, are shown from top to bottom. ■, □, and ⊠ represent vocalization, non-vocalization, and don't-care frames.

Third, again for a participant k, the probability of continuing vocalization (VC) when all other participants are silent is given by

$$\begin{aligned} f_k^{VC} &= P\left(\mathbf{q}_{t+1}[k] = \blacksquare \,|\, \mathbf{q}_t[k] = \blacksquare,\, \mathbf{q}_t[i] = \square, \boldsymbol{\Theta}^{CI}(\mathbf{Q})\right) \\ &\quad \forall \;\; i{\neq}k\,,\; 1{\leq}i{\leq}K\,. \end{aligned} \tag{16.9}$$

The context for this event type is shown in Figure 16.1(b). As for f_k^{VI} features, one such feature is computed for each participant k, and there are therefore K features of type f_k^{VC} for the conversation as a whole.

Fourth, for participant k, the probability of initiating vocalization when participant j is already talking, thereby initiating overlap (OI), is

$$\begin{aligned} f_{k,j}^{OI} &= P\left(\mathbf{q}_{t+1}[k] = \blacksquare \,|\, \mathbf{q}_t[j] = \blacksquare,\, \mathbf{q}_t[i] = \square, \boldsymbol{\Theta}^{CI}(\mathbf{Q})\right) \\ &\quad \forall \;\; i{\neq}j\,,\; 1{\leq}i{\leq}K \text{ and } j{\neq}k. \end{aligned} \tag{16.10}$$

The context for this event type, and hence feature type, is shown in Figure 16.1(c). $K-1$ such features are computed for each participant k, $1{\leq}k{\leq}K$, $1{\leq}j{\leq}K$, $j{\neq}k$. The total number of $f_{k,j}^{OI}$ features for the conversation as a whole is $K{\cdot}(K-1)$.

The fifth and last of the feature types shown in Figure 16.1(d), for participant k, is the probability of continuing vocalization when the participant of j is also talking, thereby continuing overlap (OC),

$$\begin{aligned} f_{k,j}^{OC} &= P\left(\mathbf{q}_{t+1}[k] = \blacksquare \,|\, \mathbf{q}_t[k] = \mathbf{q}_t[j] = \blacksquare,\right. \\ &\qquad \left.\mathbf{q}_t[i] = \square, \boldsymbol{\Theta}^{CI}(\mathbf{Q})\right) \\ &\quad \forall \;\; i{\neq}j\,,\; i{\neq}k\,,\; 1{\leq}i{\leq}K \text{ and } j{\neq}k\,. \end{aligned} \tag{16.11}$$

The context for this event type is shown in Figure 16.1(d). As for $f_{k,j}^{OI}$ features, $K-1$ such probabilities are computed for each participant rank k. The total number of $f_{k,j}^{OC}$ features is $K \cdot (K-1)$.

Given these five feature types and a number K of participants, the total number of features in the feature vector $\mathbf{F}$ is $3K + 2K\,(K-1)$. The ordering of specific features within the vector corresponds to the extrinsic ordering of participants in $\mathbf{Q}$.

16.7.4 Behavior Model

The role of the behavior model $\boldsymbol{\Theta}_{u,\mathbf{g}}^B$ is to provide estimates of $P\left(\mathbf{F}\left(\boldsymbol{\Theta}^{CI}(\mathbf{Q})\right) \,|\, \boldsymbol{\Theta}_{u,\mathbf{g}}^B\right)$. The size of $\mathbf{F}$ may vary from conversation to conversation, due to heterogenous K; this can be mitigated by assuming that the individual features in $\mathbf{F}$ are mutually independent. A model over these features is merely a composition of models

$$\boldsymbol{\Theta}_{u,\mathbf{g}}^B = \bigcup_{k=1}^{K} \left\{ \theta_{u,\mathbf{g}[k]}^{(1)}, \bigcup_{j=1}^{K} \left\{ \theta_{u,\mathbf{g}[k],\mathbf{g}[j]}^{(2)} \right\} \right\} \tag{16.12}$$

and

$$
\begin{aligned}
P\left(\mathbf{F}\left(\mathbf{\Theta}^{CI}\left(\mathbf{Q}\right)\right) \mid \mathbf{\Theta}^{B}_{u,\mathbf{g}}\right) &= \prod_{k=1}^{K} P\left(f^{T}_{k} \mid \theta^{T}_{u,\mathbf{g}[k]}\right) \\
&\times P\left(f^{VI}_{k} \mid \theta^{VI}_{u,\mathbf{g}[k]}\right) \times P\left(f^{VC}_{k} \mid \theta^{VC}_{u,\mathbf{g}[k]}\right) \\
&\times \prod_{j \neq k}^{K} P\left(f^{OI}_{k,j} \mid \theta^{OI}_{u,\mathbf{g}[k],\mathbf{g}[j]}\right) \times P\left(f^{OC}_{k,j} \mid \theta^{OC}_{u,\mathbf{g}[k],\mathbf{g}[j]}\right) .
\end{aligned} \tag{16.13}
$$

In the above, θ are models of individual one-participant and two-participant features; for simplicity, the models are single Gaussians, characterized by a mean μ and a variance σ^2. It should be noted that since the features are probabilities, a Gaussian model is likely a very suboptimal parametrization.

Training the individual submodels $\theta \equiv \{\mu, \sigma^2\}$ for a participant profile g_k (or a pair of profiles g_k and g_j) is achieved by first computing a global, participant-independent (or participant-pair-independent) model $\theta_0 \equiv \{\mu_0, \sigma_0^2\}$, over all participants and all training conversations. Then, if the participant profile g_k (or the pair g_k and g_j) occurs more than n_0 times in the training material, $\theta \equiv \{\mu, \sigma^2\}$ is assigned its maximum likelihood values for the participant profile (or participant profile pair). Otherwise, it is assigned $\theta = \theta_0$. The threshold n_0 is estimated by maximizing round-robin DataSet classification accuracy; to limit overfitting to the ICSI corpus, a single n_0 is assumed for all feature types. This form of θ estimation is of course an instance of backing off to a global, undifferentiating model.

16.7.5 Defining Participant Profiles for Training

To train the model described in Equation 16.13, the participant profile assignment $\mathbf{g}$ of all participants must be known in the training conversations. Profiles could be manually designed and assigned, or they could be determined using some form of clustering of all training material participants, based on their features. In the current chapter, a simplified form of the latter is accomplished, by equating profile (during training) with *participant talkativity rank*. This greatly simplifies model construction, and offers significant time-complexity savings during search. This is because any conversation of K participants will contain participants of rank 1 through K only; $\mathbf{g}$ is then merely a permutation of these K ranks, and $\mathbb{G}$ is simply the symmetric group on K symbols whose number of elements is $K!$ rather than N_G^K.

16.7.6 Membership Model

Given the above constraint on $\mathbb{G}$ and $\mathbf{g}$, the membership model $\mathbf{\Theta}^{M}_{u}$ in Equation 16.6 plays no role in the experiments of this chapter (in contrast to the subsequent chapter). This is because any conversation of K participants, regardless of type, will consist of participant profiles 1 through K, representing talkativity rank.

There is scope for the membership model to favor specific K, or specific index assignment representing seating arrangement or channel preferences of particular participants in longitudinal recordings, but the experiments in this chapter aim to detect differences in conversational style which is independent of these aspects.

16.7.7 Estimation of Non-Parametric Model Features

Maximum likelihood estimation of those model parameters which are the features of the proposed classifier relies merely on counting bigrams in the observed $\mathbf{Q}$. For example,

$$
f^{OC}_{k,j} = \frac{\sum_{t=1}^{T} \delta\left(\mathbf{q}_{t-1}, \mathbf{S}_{k,j}\right) \delta\left(\mathbf{q}_{t}\left[k\right], \blacksquare\right)}{\sum_{t=1}^{T} \delta\left(\mathbf{q}_{t-1}, \mathbf{S}\right)} \tag{16.14}
$$

where

$$
\mathbf{S}_{k,j}\left[i\right] = \begin{cases} \blacksquare & \text{if } i \in \{k, j\} \\ \square & \text{otherwise} \end{cases} . \tag{16.15}
$$

A classifier which relies only on f_k^T achieves a round-robin DataSet accuracy of 68.7%, with the cutoff $n_0 = 0$ (estimated using the same data). A classifier which relies not only on f_k^T but also on the other four feature types, f_k^{VI}, f_k^{VC}, f_k^{OI}, and f_k^{OC}, yields an accuracy of 64.2%. When combining features, a single cutoff is used to build models for all of them, in order to avoid overfitting to the data; for the 5-feature-type combination, the optimal n_0 was found to be 10. The combined-feature-type accuracy of 64.2% is lower than that achieved by f_k^T alone by 3 exemplars, indicating that the additional feature types, when estimated using maximum likelihood, are on average not helpful for this task.

To explore the impact of individual feature types, experiments were performed using a classifier relying on at most one feature type, with and without f_k^T. The results are shown in Table 16.3. A single cutoff n_0 is used for each experiment, which is duplicated: globally optimized "g-Opt" systems, shown in columns 2 and 3, use $n_0 = 10$ as found to be optimal when all 5 feature types are used; condition-optimized "c-Opt" systems, in columns 4 and 5, generally exhibit higher accuracies because n_0 is re-optimized for that feature type (or feature type pair, with f_k^T).

Feature(s)	g-Opt		c-Opt	
	w/o f_k^T	w/ f_k^T	w/o f_k^T	w/ f_k^T
f_k^{VI}	49.3	59.7	58.2	62.7
f_k^{VC}	58.2	76.1	65.7	80.6
$f_{k,j}^{OI}$	53.7	53.7	61.2	59.7
$f_{k,j}^{OC}$	56.7	65.7	64.2	70.1
$\langle f_{k,j}^{OI} \rangle_j$	43.3	47.8	46.3	62.7
$\langle f_{k,j}^{OC} \rangle_j$	52.2	64.2	52.2	68.7

Table 16.3: Leave-one-out meeting-type classification accuracy on DataSet, in which features are estimated using maximum likelihood. Each feature type is evaluated by itself ("w/o f_k^T") and in combination with f_k^T ("w/ f_k^T").

The table indicates that no feature type exceeds the accuracy of 68.7% which is obtained using f_k^T. However, f_k^T and f_k^{VC}, together, yield an accuracy of 80.6%. Also, $f_{k,j}^{OC}$, when combined with f_k^T, yields 70.1%. The remaining feature types, however, generally hurt performance.

The table also shows the performance of systems based on "summary" feature types $\langle f_{k,j}^{OI} \rangle_j$ and $\langle f_{k,j}^{OC} \rangle_j$, which are averages for each participant profile over their interlocutors j. It can be seen that in the absence of f_k^T features, these summary features appear to be worse than individual feature types which model interaction among specific participant ranks k and j.

It should be noted that tuning n_0 for specific feature types, shown in the "c-Opt" columns as compared to the "g-Opt" columns, yields improved performance in the majority of cases. However, when combining feature types, retaining the optimal "c-Opt" n_0 value for each feature type yields no improvement, relative to retaining the same n_0 for all of them. This indicates that these parameters must be optimized jointly across feature types; this is not performed here, to avoid round-robin overfitting to the data.

16.7.8 Estimation of Parametric Model Features

An alternative means of computing features is not to rely on maximum likelihood estimation using the non-parametric conditionally independent model, but to instead use an approximation. This subsection explores the approximation entailed by assuming the parametric Ising model representation of Chapter 7. There, the $K \times 2 \times 2^K$ parameters of the non-parametric model are replaced by $K \times (K + 1)$ Ising antiferromagnet parameters. Their values are obtained from an observed $\mathbf{Q}$ as described in Section 7.5.

The Ising model parameters $\mathbf{b} = \{b_k\} \in \mathbb{R}^K$ and $\mathbf{W} = \{w_{k,j}\} \in \mathbb{R}^{K \times K}$, once available, yield particularly prosaic

expressions for the feature types under study:

$$f_k^{VI} = \frac{1}{1+e^{-b_k}} \ , \tag{16.16}$$

$$f_k^{VC} = \frac{1}{1+e^{-b_k-w_{k,k}}} \ , \tag{16.17}$$

$$f_{k,j}^{OI} = \frac{1}{1+e^{-b_k-w_{k,j}}} \ , \tag{16.18}$$

$$f_{k,j}^{OC} = \frac{1}{1+e^{-b_k-w_{k,j}-w_{k,k}}} \ . \tag{16.19}$$

The classification accuracy obtained with feature types computed in this way is 76.1%, when combined with f_k^T. The optimal n_0 value is 10, as in the case when the feature values are estimated using maximum likelihood. However, the achieved accuracy is higher by 7.4%abs, or 5 exemplars, than in the maximum likelihood case. This is likely due to the fact that when estimated using the parametric model, parameters are inferred from all of the training material.

Classification results for "g-Opt" and "c-Opt" systems, using individual Ising model feature types with and without f_k^T, are shown in Table 16.4.

Feature(s)	g-Opt		c-Opt	
	w/o f_k^T	w/ f_k^T	w/o f_k^T	w/ f_k^T
f_k^{VI}	50.7	59.7	58.2	70.1
f_k^{VC}	58.2	70.1	58.2	71.6
$\langle f_{k,j}^{OI} \rangle_j$	65.7	65.7	68.7	71.3
$\langle f_{k,j}^{OC} \rangle_j$	64.2	76.1	64.2	79.1
$f_{k,j}^{OI}$	62.7	65.7	68.7	67.2
$f_{k,j}^{OC}$	68.7	74.6	73.1	74.6

Table 16.4: Leave-one-out meeting-type classification accuracy on DATASET using individual feature types, by themselves ("w/o f_k^T") and in combination with f_k^T ("w/ f_k^T").

When participant talkativity is not modeled, the best feature type is that which represents the probability that a participant will continue overlap with a specific other participant ($f_{k,j}^{OC}$), for both the "g-Opt" and the "c-Opt" systems. However, when talkativities f_k^T are modeled alongside each feature type, the best performance is obtained using the probability that each participant continues overlap with *any* other participant ($\langle f_{k,j}^{OC} \rangle$). The best accuracy observed is 79.1%, achieved by the "c-Opt" system relying on talkativity and $\langle f_{k,j}^{OC} \rangle$. Using all 6 feature types from Table 16.4 in addition to talkativity also results in 79.1%.

16.8 Combining Model Scores

The two approaches covered in the two preceeding sections respectively model: (1) participant-independent interaction in conversation, via a model of the probabilities of entry and egress into contrasting degrees of multi-participant vocalization overlap; and (2) participant-dependent speech activity deployment preferences, via models of the probability of speech deployment in specific circumstances, indexed by participant talkativity rank. The performance of these approaches is compared in Table 16.5.

As already observed, talkativity alone reduces classification error, from the majority class guessing baseline, by 25.4%abs or 44.8%rel. The system which assumed identical participants achieves an accuracy of 79.1%, which is also observed for the best system which assumes non-identical participants. This represents a further reduction of error of 10.4%abs, or 33.2%rel.

Table 16.5 also shows, in the last column, the number of parameters which were optimized in the round-robin regime. No such parameters were optimized for the EDO model (although it could be argued that selecting $K_{max} = 4$, from

Approach	**using**	A, %	#
Stochastic guessing	flat prior	34.3	0
	true prior	34.3	0
Deterministic guessing	majority class	43.3	0
Identical Participants	$\mathbf{\Theta}_{EDO}^{CD}$	79.1	0
Non-Identical Participants	f_k^T only	68.7	1
	$\mathbf{\Theta}_{ML}^{CI}$	76.1	1
	$\mathbf{\Theta}_{Ising}^{CI}$	79.1	1
Combined Identical & Non-Identical Participants	f_k^T only	82.1	2
	$\mathbf{\Theta}_{ML}^{CI}$	83.6	2
	$\mathbf{\Theta}_{Ising}^{CI}$	85.1	2

Table 16.5: Meeting type classification accuracies on DATASET, in percent, for the approaches described in this chapter. "#" indicates number of classifier parameters optimized using the same round-robin approach and data as used to produce these results.

Table 10.6, on the entire ICSI Meeting Corpus constitutes tuning). All classifiers which model participants as non-identical had n_0 tuned to maximize round-robin accuracy. A combination of these non-identical-participant systems with the identical-participant EDO-based classifier, required selection of interpolation parameter α. Combined system performance, as a function of α, is shown in Figure 16.2.

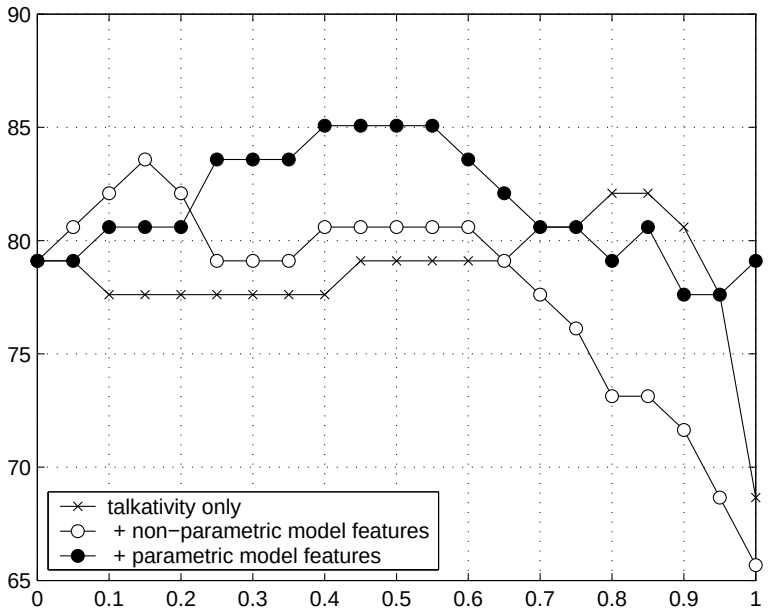


Figure 16.2: DATASET classification accuracy (along the y-axis) following linear interpolation of the EDO model approach, which assumes no differentiation among participant, and three selected models which do. Linear weight α shown along the x-axis, with zero representing only the EDO model approach.

Figure 16.2 indicates that all 3 of the systems in Table 16.5 which do not assume participants to be identical can be usefully combined with the EDO-based model which does. The best performance, of 85.1%, is achieved by equally weighting the EDO model and the parametric non-identical-participant model.

16.9 Potential Impact

The experiments in this chapter demonstrate that, while the EDO model can be used to model the distribution of speech activity in general multiparty conversation, it also successfully discriminates between conversation types, even when they are defined by groups meeting for different projects within the same organization. This makes it deployable as a means of classifying conversations. However, when training EDO models for other purposes, it may be important to train on data as close to the target domain as possible.

The experiments also show that certain type-discriminative information is not captured by the EDO model. In particular, differences between participants are not captured. This reflects the original intent of the EDO model. One feature shown to be important for conversation-type classification is the relative talkativity of participants.

Conversation type classification systems can of course be expected to leverage other information than that modeled here, in particular the number K of participants. The experiments described in this chapter have intentionally ignored that feature, since the goal was to demonstrate discriminative aspects of K-independent representations of interaction.

16.10 Relevance to Other Chapters

This chapter has explored the applicability of the EDO model of Chapter 6 as a means of discriminating among different conversation types, with "type" characterized by project and participant population. It has also applied another form of model from that chapter, namely that which assumes participants' speech activity states to be conditionally independent at instant t, given the joint state at instant $t-1$. Finally, it has exercised the parametric form of the conditionally independent model, proposed in Chapter 7.

Several ideas in the current chapter, particularly the concepts of participant groups and the submodel permutation, are employed in the subsequent chapter (Chapter 17) in an effort to characterize conversation participants, individually, as opposed to characterizing conversations as a whole.

16.11 Summary

This chapter has explored the classification of conversations into types, within a framework which is text-independent and therefore also topic-independent. It was argued that classification can be beneficial, particularly for audio browsing applications. Experiments have shown that, when conversation types are equated with project groups getting together to collaboratively address project-specific needs, text-independent characterization of the distribution of speech activity in time and across participants is possible and useful. In particular, models which capture the time-independent degree of interaction, via probabilities of entry and egress from contrastive levels of overlap, as well as models capturing differences between participants indexed by talkativity rank, both reduce naive baseline error rates by almost 50%. The approaches are complementary to a certain extent, yielding an accuracy of 85% in a subset of the ICSI Meeting Corpus for which guessing the majority class yields an accuracy of 43%.

The experiments also show that estimating participant-pair interaction is possible with the parametric Ising antiferromagnet model, leading to slightly improved performance over uninterpolated maximum likelihood estimates of conditionally independent participant-pair state transitions. While preliminary, due to the paucity of data and the simplicity of classifier stucture, the empirical results of this section offer a promising means of inferring time-indepenent, K-independent characteristics of conversations.

16.12 Future Directions

This chapter has presented a preliminary study which may serve as a baseline for future work on the classification of conversational interaction or style. In the interests of expediency, a number of decisions were made which are likely to have been suboptimal.

Future studies in this area should be conducted using a much larger corpus, if such becomes available. Distinctions among systems, presented here, cannot be usefully qualified for statistical significance; the differences involve small absolute variation, and no unseen data was held out to assess generalization.

A second aspect which should be explored, in the context of those models which assume that participants are non-identical, is the selection of an appropriate distribution for the proposed features. The latter are probabilities, often quite close to zero or unity, and modeling them for the purposes of discrimination with a single-Gaussian generative model is likely to be very sub-optimal.

As for other chapters in this thesis, it is also important to asses the extent to which the proposed modeling approaches are robust to noisy speech activity references, conditions which represent a truly automatic setting.

An approach to conversation classification which promises to be felicitous is not to infer models from each test conversation and then to use model parameters as features, but to simply score test conversations with type-specific models, as was done with the EDO model. To address participant variation, the same conditionally independent models can be used, with one modification. Both training and test conversation $\mathbf{Q}$ need to be independently row-rotated, for example by requiring that $\mathbf{Q}$ is first row-sorted by participant global talkativity rank. The K-specificity of compositional conditionally independent models can be addressed by padding all $\mathbf{Q}$ to a maximum observed K.

Finally, conversation classification systems deployed in specific circumstances are likely to need to rely on participant number, on lexical features — even for topic-independent classification — and on the distribution of other vocalization such as laughter. There may also be significant benefit to modeling time-dependent phenomena, such as conversation phases and their sequencing grammar.

Chapter 17

Text-Independent Participant Characterization*

17.1 Introduction

An important task in automatic conversation understanding is the inference of social structure governing participant behavior; in many conversations, the maintenance or expression of that structure is an implicit goal, and may be more important than the propositional content of what is said.

There are many social dimensions along which participants may differ [16]. Research in social psychology has shown that such differences entail systematic differences in observed turn-taking and floor-control patterns (e.g. [7], [210], [37]), and that participant types are not independent of the types and sizes of conversations in which they appear. It is therefore possible that many participant characteristics may be inferrable from as low-level a representation of conversation as is offered by the speech chronogram.

This chapter aims to quantify, for the first time, the correlation between individual participants' observed probabilities of deploying speech in specific overlap circumstances and both (1) assigned roles in enacted meetings and (2) social status characteristics in naturally occurring meetings. The experiments presented go beyond correlation, in attemtping to classify participants into one of several discrete classes along several dimensions. The proposed tasks include the inference of assigned role, assigned leadership, gender, seniority, and identity. On all tasks except gender classification, the accuracies are significantly higher than chance.

The research in this chapter first appeared in [146] and [152].

17.2 Related Work

Work which inspired the subsequently described experiments included the classification of interaction groups [27], or cliques, and the classification of dynamic role [230]. The latter is somewhat different from what is proposed in this chapter, since dynamic roles as defined in [230] are not time-invariant. More similar, and equally inspirational, was work on dominance classification [190] and influence ranking [191]. Both of these works relied on manual annotation of a time-independent characteristic of each meeting participant. Specific participant detection using chronogram features was explored in [125].

In contrast to these cited works, the experiments presented here attempt to recover the metadata associated with each participant, in the two largest complete corpora of multi-party conversation currently available. These are the AMI Meeting Corpus and the ICSI Meeting Corpus. In the former, individual participants occur rarely, but are *assigned* one of four roles which occur in every meeting. In the latter, some participants occur frequently, and although no roles were assigned (the meetings are naturally occurring), several self-descriptions are available for each participant. Among these is education level which is significantly and strongly associated with age. At the time the experiments of this chapter were

*The work in this chapter was conducted in collaboration with Mari Ostendorf.

conducted [146, 152], none of the tasks they tackle had been attempted on either corpus or on any other data of similar size and variability.

Since the publication of [146, 152], the task of classifying assigned roles in the AMI Meeting Corpus has received a lot of attention. Like the experiments described here, the majority of the relevant works relies on a text-independent representation of each meeting. Using an approach based on social network analysis, employed to advantage for role recognition in radio broadcasts [218], the authors of [59], [68], [58], and [195] reported manual segmentation classification accuracies of 43.6%, 49.5%, 44.4%, and 56.0%, respectively. The accuracy reported here, and in [146], is 53%. It should be noted that these numbers may not be exactly comparable, in particular because [59, 68, 58, 195] achieve the quoted performance using leave-one-out classification, while the current chapter splits the data into three sets with only two thirds of the AMI meetings used for training. The 20 meetings used here for final evaluation could also be easier or harder than the average AMI meeting.

17.3 Dataset Use

Two different corpora of multi-party meetings are used. The first, the scenario subset of the AMI Meeting Corpus [35], consists of meetings involving $K = 4$ participants who play different specialist roles in a product design team. The recommended division of this data has been observed, into: AMITRAINSET of 98 meetings; AMIDEVSET of 20 meetings; and AMIEVALSET, also of 20 meetings. Although each participant takes part in approximately 4 meetings, the 3 sets are disjoint in participants. Only the provided word alignments of these meetings are used, to produce a discretized $\mathbf{Q}$. The corpus is accompanied by metadata which specifies the gender and assigned role of each participant.

The second corpus consists of the `Bed`, `Bmr`, and `Bro` meeting types in the ICSI Meeting Corpus [108]. Each meeting is identified by one of $\{\texttt{Bed}, \texttt{Bmr}, \texttt{Bro}\}$, as well as a numerical identifier d. These have been divided into: ICSITRAINSET, consisting of the 33 meetings for which $d \bmod 4 \in \{1, 2\}$; ICSIDEVSET, consisting of the 18 meetings for which $d \bmod 4 \equiv 3$; and ICSIEVALSET, consisting of the 16 meetings for which $d \bmod 4 \equiv 0$. The three sets are not disjoint in participants, and the number of instrumented participants K varies from meeting to meeting, between 3 and 9. The corpus is accompanied by metadata specifying the gender, age, and education level of each participant. Only the forced alignments of these meetings are used, available in the accompanying MRDA Corpus [202], and are discretized to produce the required $\mathbf{Q}$. Participant identifiers in the corpus are kept across meetings.

In the last part of this chapter, three separate multiparticipant vocal activity segmentations are contrasted. All three segmentations are binary, in that at any point in time t each participant k is considered to be either vocalizing or not vocalizing. The first segmentation, $\mathcal{S}$, consists of all talkspurts and is constructed from the forced-alignment lexical item endpoints found in the ICSI MRDA Corpus [202]; inter-item gaps shorter than 0.3 seconds are bridged. The second segmentation, $\mathcal{S} - \mathcal{B}$, is constructed in the same way, but only non-backchannel lexical items are considered. Finally, the third segmentation $\mathcal{L}$ of laugh bouts is as described in [135]. Each of the three segmentations is discretized [149] using a particular frame step ΔT and frame size T_s.

17.4 Assessment of Performance

The natural measure of performance, when classifying participants, is classification accuracy or error rate. Although this chapter will argue for classifying conversational groups jointly, by simultaneously positing participant classes for each pariticipant present, accuracy is computed not for the group as a whole but for each participant independently. This means that if 3 out of 4 participants are correctly classified, the accuracy is 75% rather than 0%.

17.5 Holistically Classifying Participant Groups

The approach taken in this chapter is to classify the group $\mathbf{g}$ of participants *holistically*, rather than classifying the type $\mathbf{g}[k]$ of each of k participants, $1 \leq k \leq K$, independently. This allows the application of constraints on the group as a whole, by limiting the search space over $\mathbf{g}$ to only those single-participant type combinations which are licensed. For example, when attempting to determine the leader of a conversation, only one of the conversants should be thus classified; classifying participants independently might lead to more than one conversant labeled as the leader, which would then

require subsequent hypothesis recombination. Additionally, treating groups holistically allows for modeling the interactions between specific pairs of participant types.

The method proposed is similar to that in Section 16.7 of the previous chapter, where conversation type was inferred from a set of features $\mathbf{F}$ computed from $\mathbf{Q}$ by marginalizing over possible participant type group assignments,

$$\begin{aligned} u^* &= \arg\max_{u \in \mathcal{U}} \sum_{\mathbf{g} \in \mathbb{G}} P(u, \mathbf{g} \,|\, \mathbf{Q}) \\ &= \arg\max_{u \in \mathcal{U}} \sum_{\mathbf{g} \in \mathbb{G}} P(u) \cdot P(\mathbf{g} \,|\, u) \cdot P(\mathbf{Q} \,|\, \mathbf{g}, u) \\ &\doteq \arg\max_{u \in \mathcal{U}} P(u) \times \\ &\quad \sum_{\mathbf{g} \in \mathbb{G}} \underbrace{P\left(\mathbf{g} \,|\, \boldsymbol{\Theta}^M_u\right)}_{\substack{\text{Membership} \\ \text{Model}}} \cdot \underbrace{P\left(\mathbf{F}\left(\boldsymbol{\Theta}^{CI}(\mathbf{Q})\right) \,|\, \boldsymbol{\Theta}^B_{u,\mathbf{g}}\right)}_{\substack{\text{Behavior} \\ \text{Model}}} . \end{aligned} \tag{17.1}$$

17.5.1 Marginalizing over Conversation Types

The desired dual of Equation 17.1 is of course

$$\begin{aligned} \mathbf{g}^* &= \arg\max_{\mathbf{g} \in \mathbb{G}} \sum_{u \in \mathcal{U}} P(u, \mathbf{g} \,|\, \mathbf{Q}) \\ &= \arg\max_{\mathbf{g} \in \mathbb{G}} \sum_{u \in \mathcal{U}} P(u) \cdot P(\mathbf{g} \,|\, u) \cdot P(\mathbf{Q} \,|\, \mathbf{g}, u) \\ &\doteq \arg\max_{\mathbf{g} \in \mathbb{G}} \sum_{u \in \mathcal{U}} P(u) \times \\ &\quad \underbrace{P\left(\mathbf{g} \,|\, \boldsymbol{\Theta}^M_u\right)}_{\substack{\text{Membership} \\ \text{Model}}} \cdot \underbrace{P\left(\mathbf{F}\left(\boldsymbol{\Theta}^{CI}(\mathbf{Q})\right) \,|\, \boldsymbol{\Theta}^B_{u,\mathbf{g}}\right)}_{\substack{\text{Behavior} \\ \text{Model}}} . \end{aligned} \tag{17.2}$$

It is readily seen that the membership model and the behavior model are identical in form for both inference tasks. The remainder of this section describes the specific features $\mathbf{F}$ and the behavior model $\boldsymbol{\Theta}^B_{u,\mathbf{g}}$ used in this chapter, which are both very similar to those in Section 16.7, as well as alternative search spaces $\mathbb{G}$ and the membership model $\boldsymbol{\Theta}^M_u$.

17.5.2 Features

The same feature types[1] as used in Chapter 16 are also used in the current chapter. They include: f^T_k, the probability that participant k vocalizes at (any) time t; f^{VI}_k, the probability that participant k initiates vocalization at time t when no-one else was speaking at $t-1$; f^{VC}_k, the probability that participant k continues vocalization at time t when no-one else was speaking at $t-1$; $f^{OI}_{k,j}$, the probability that participant k initiates vocalization at time t when participant j was speaking at $t-1$; and $f^{OC}_{k,j}$ the probability that participant k continues vocalization at time t when participant j was speaking at $t-1$. The values of features of these feature types are time time-independent probabilities, and are estimated using the parametric model of Chapter 7. Additionally, single-participant averages of the two-participant features are computed: $\langle f^{OI}_{k,j} \rangle_j$, $\langle f^{OI}_{j,k} \rangle_j$, $\langle f^{OC}_{k,j} \rangle_j$, and $\langle f^{OC}_{j,k} \rangle_j$. The complete feature vector $\mathbf{F}$ for a conversation of K participants then consists of $7K$ one-participant features, and $2(K^2 - K)$ two-participant features.

It should be noted that multiple phenomena contribute to the overlap features. The features $f^{OI}_{k,j}$ are based on counts from interruptions, backchannels, and precise floor handoffs, and potentially other phenomena. The features $f^{OC}_{k,j}$ include counts from interruptions, attempts to hold the floor, and backchannels. Both feature types also contain counts incurred during schism, when the conversation splits into two sub-conversations.

[1] Feature type superscripts indicate talkspurt initiation (I) or continuation (C), for either single-participant vocalization (V) or vocalization overlap (O).

17.5.3 Behavior Model

As was done in Chapter 16, the behavior model $\mathbf{\Theta}^B_{u,\mathbf{g}}$ is factored since the size of the feature vector $\mathbf{F}$ is variable across training and test conversations. Each feature is assumed to be independent of the others, and is described by its own univariate Gaussian model $N\left(\mu, \sigma^2\right)$. The parameters μ and σ are nominally estimated as

$$\hat{\mu} = \frac{C^1}{C^0} \tag{17.3}$$

$$\hat{\sigma}^2 = \frac{C^2}{C^0} - \hat{\mu}^2 \,, \tag{17.4}$$

where C^m is the zeroth, first, or second order cumulant ($m \in \{0, 1, 2\}$). For one-participant feature types characterizing participant $\xi \in G$, and for two-participant feature types characterizing participants ξ and ζ together, these cumulants are computed from training data, consisting of R meetings, via

$$C^m_{u;\xi} = \sum_{r=1}^{R} \delta\left(u_r, u\right) \sum_{k=1}^{K_r} \delta\left(\mathbf{g}_r\left[k\right], \xi\right) \times \left(f_{r,k}\right)^m \,, \tag{17.5}$$

$$C^m_{u;\xi,\zeta} = \sum_{r=1}^{R} \delta\left(u_r, u\right) \sum_{k=1}^{K_r} \delta\left(\mathbf{g}_r\left[k\right], \xi\right) \times \sum_{j=1}^{K_r} \delta\left(\mathbf{g}_r\left[j\right], \zeta\right) \times \left(f_{r,k,j}\right)^m \,, \tag{17.6}$$

respectively. Here, δ is the Kronecker delta, and r enumerates over the R meetings in the training corpus. u_r is the type of the rth meeting, and $f_{r,k}$ (and $f_{r,k,j}$) are the features from the kth (and jth) participant in the rth meeting.

As certain types of participants may be rare in the training data, the above cumulants are interpolated with less specific cumulants. For one-participant feature types, this includes meeting-type-independent cumulants

$$C^m_{*;\xi} = \sum_u C^m_{u;\xi} \,, \tag{17.7}$$

meeting-type-specific but participant-independent cumulants

$$C^m_{u;*} = \sum_\xi C^m_{u;\xi} \,, \tag{17.8}$$

and meeting-type-independent and participant-independent cumulants

$$C^m_{*;*} = \sum_u \sum_\xi C^m_{u;\xi} \,. \tag{17.9}$$

For two-participant feature types, this also includes meeting-type-independent cumulants

$$C^m_{*;\xi,\zeta} = \sum_u C^m_{u;\xi,\zeta} \,; \tag{17.10}$$

meeting-type-specific but participant-independent cumulants

$$C^m_{u;\xi,*} = \sum_\zeta C^m_{u;\xi,\zeta} \tag{17.11}$$

$$C^m_{u;*,\zeta} = \sum_\xi C^m_{u;\xi,\zeta} \tag{17.12}$$

$$C^m_{u;*,*} = \sum_\xi \sum_\zeta C^m_{u;\xi,\zeta} \,; \tag{17.13}$$

and meeting-type-independent and participant-independent cumulants

$$C^m_{*;\xi,*} = \sum_u \sum_\zeta C^m_{u;\xi,\zeta} \tag{17.14}$$

$$C^m_{*;*,\zeta} = \sum_u \sum_\xi C^m_{u;\xi,\zeta} \tag{17.15}$$

$$C^m_{*;*,*} = \sum_u \sum_\xi \sum_\zeta C^m_{u;\xi,\zeta} \; . \tag{17.16}$$

Model estimates are computed by combining these cumulants using

$$\tilde{C}^m_{u;\xi} = C^m_{u;\xi} + \lambda_{*;\xi} \, C^m_{*;\xi} + \lambda_{u;*} \, C^m_{u;*} + \lambda_{*;*} \, C^m_{*;*} \; . \tag{17.17}$$

and

$$\begin{aligned} \tilde{C}^m_{u;\xi,\zeta} &= C^m_{u;\xi,\zeta} + \lambda_{*;\xi,\zeta} \, C^m_{*;\xi,\zeta} + \lambda_{u;\xi,*} \, C^m_{u;\xi,*} \\ &+ \lambda_{*;\xi,*} \, C^m_{*;\xi,*} + \lambda_{u;*,\zeta} \, C^m_{u;*,\zeta} + \lambda_{*;*,\zeta} \, C^m_{*;*,\zeta} \\ &+ \lambda_{u;*,*} \, C^m_{u;*,*} + \lambda_{*;*,*} \, C^m_{*;*,*} \; . \end{aligned} \tag{17.18}$$

In this chapter, the λ interpolation factors are allowed to be different for different feature types, and are optimized by maximizing classification accuracy on development data; however, it should be noted that minimal effort has gone into optimizing these values. This is in contrast to model estimation in Chapter 16, where a single back-off threshold n_0 was used for all feature types, because there — due to the limited number of meetings (as opposed to participants in this chapter) — no data was available for development purposes.

17.5.4 Search Space Types

The search space $\mathbb{G}$ over possible group assignments depends on the types considered for each participant, as well as on group constraints. Generally, given $N_G \equiv |G|$ possible alternative classes per participant, and K participants, the specific group assignment $\mathbf{g}$ is a combination of any K of the N_G types, with or without replacement. There are 3 broad categories of group assignment. First, it may be the case that $N_G = K$, namely that all possible participant classes are unique and are known to be present in the conversation under study. This makes $\mathbf{g}$ a permutatation of the elements of G, and $\mathbb{G}$ the symmetric group on K symbols; the number of candidate group assignments is $K!$. Second, there may be no group constraints whatsoever, such that each participant may be of any class in G. The number of candidate group assignments in this case is N_G^K. Finally, a mixed situation is possible, where some elements of G must be unique in $\mathbf{g}$, but others can be duplicated. All three scenarios are explored in this chapter.

17.5.5 Membership Model

The proposed membership model $\mathbf{\Theta}^M_u$ in this chapter is quite simple,

$$P\left(\mathbf{g} \,|\, \mathbf{\Theta}^M_u\right) = \prod_{k=1}^{K} P\left(\mathbf{g}\,[k] \,|\, \mathbf{\Theta}^M_u\right) \; , \tag{17.19}$$

where $P\left(\mathbf{g}\,[k]\right)$ is the probability that the k-th participant is of type $\mathbf{g}\,[k]$. This simplicity makes it independent of the number K of participants in test conversations, but fails to capture potentially important constraints such as "there should be at least one g_i in each conversation", for some i. The probabilities of specific types are maximum likelihood estimates from the training data.

17.6 Inferring Assigned Role

The inference of assigned role is explored using the AMI corpus. All meetings consist of $K = 4$ participants, and each participant is assigned one of four roles: project manager (PM), marketing expert (ME), user interface designer (UI), or industrial designer (ID).

Since roles $\mathbf{g}[k] \in G \equiv \{\text{PM}, \text{ME}, \text{UI}, \text{ID}\}$ are unique, and present in every meeting, there are 4! possible group assignments $\mathbf{g} \in \mathbb{G}$. A result of this is that the membership model described in Subsection 17.5.5 plays no role, and inference of group assignment is performed using

$$\begin{aligned} \mathbf{g}^* &\doteq \arg\max_{\mathbf{g}\in\mathbb{G}} \sum_{u\in\mathcal{U}} P(u) \times \\ &\qquad P\left(\mathbf{g} \,|\, \mathbf{\Theta}^M_u\right) \cdot P\left(\mathbf{F}\left(\mathbf{\Theta}^{CI}(\mathbf{Q})\right) \,|\, \mathbf{\Theta}^B_{u,\mathbf{g}}\right) \\ &\doteq \arg\max_{\mathbf{g}\in\mathbb{G}} P\left(\mathbf{F}\left(\mathbf{\Theta}^{CI}(\mathbf{Q})\right) \,|\, \mathbf{\Theta}^B_{\mathbf{g}}\right) \end{aligned} \tag{17.20}$$

where in the second line marginalization over meeting type u has been elided, since the AMI data set studied here does not differentiate between meeting types.

amiTrainSet is used to train the behavior model $\mathbf{\Theta}^B_{\mathbf{g}}$, as described in Subsection 17.5.3. The best three features types for this task are then identified using amiDevSet, by analyzing feature types individually, one at a time. Early experiments revealed that performance levels off, and even falls, once more than a handful of the best performing features are used. These were found to be f^{VI}_k, $\langle f^{OI}_{k,j}\rangle_j$, and $f^{OI}_{k,j}$, capturing the probability of initiating a talkspurt in silence, of initiating a talkspurt when someone else is speaking, and of initiating a talkspurt when a participant in a specific other role is speaking, respectively. On amiEvalSet, these feature types lead to single-feature-type 4-way classification rates of 41%, 29%, and 53%, respectively. When all three types are used together ($3K + K^2$ features in total), the rate is 53%. amiEvalSet accuracy when all feature types are used is 46%, indicating that some feature types are detrimental to this task.

Ref	Hypothesized			
	ID	ME	PM	UI
ID	**8**	6	4	2
ME	5	**8**	4	3
PM	3	4	++**12**	− 1
UI	4	2	−− 0	++**14**

Table 17.1: Confusion matrix for role classification on amiEvalSet; reference assignment is found in the rows, hypothesized assignment in columns. Correctly classified roles, along the diagonal, are highlighted in bold. Statistical significance of association at the $p < 0.005$ level per class, using a 2×2 χ^2-test, is shown using "++" and "−−", for above chance and below chance values, respectively; the same is true of "+" and "−", for significance at the $0.005 \leq p < 0.05$ level.

The confusion matrix for classification using the three best feature types is shown in Table 17.1. The matrix shows that association between the reference assignment of PM, as well as of UI, and the hypothesized assignment based on the three feature types mentioned is statistically significant. On the other hand, assignment of ID and ME does not deviate significantly from chance.

17.7 Finding the Assigned Leader

A simpler task than the joint inference of all participant types is the identification of the leader in a conversational group. Using the same data as in the preceding section, the project manager (PM) role is equated with conversation leadership L, while the remaining roles are assigned to the class $\neg L$. In the AMI meetings, which take a product design from start to prototype, the project manager is expected to make the group run smoothly.

Since a leader exists in every AMI meeting, the number of unique group assignments is only 4. Training the behavior model as in the preceding section, using amiTrainSet, and assessing each feature type individually using amiDevSet classification accuracy, identifies the same three feature types as independently the best-performing, namely f_k^{VI}, $\langle f_{k,j}^{OI} \rangle_j$, and $f_{k,j}^{OI}$. Using these three types, the leader $L \equiv$ PM is correctly identified in 12 of the 20 meetings in amiEvalSet. Using them together also yields only a 60% accuracy. However, when all feature types are used, the classification accuracy climbs to 75%. The baseline with random guessing would result in 25%.

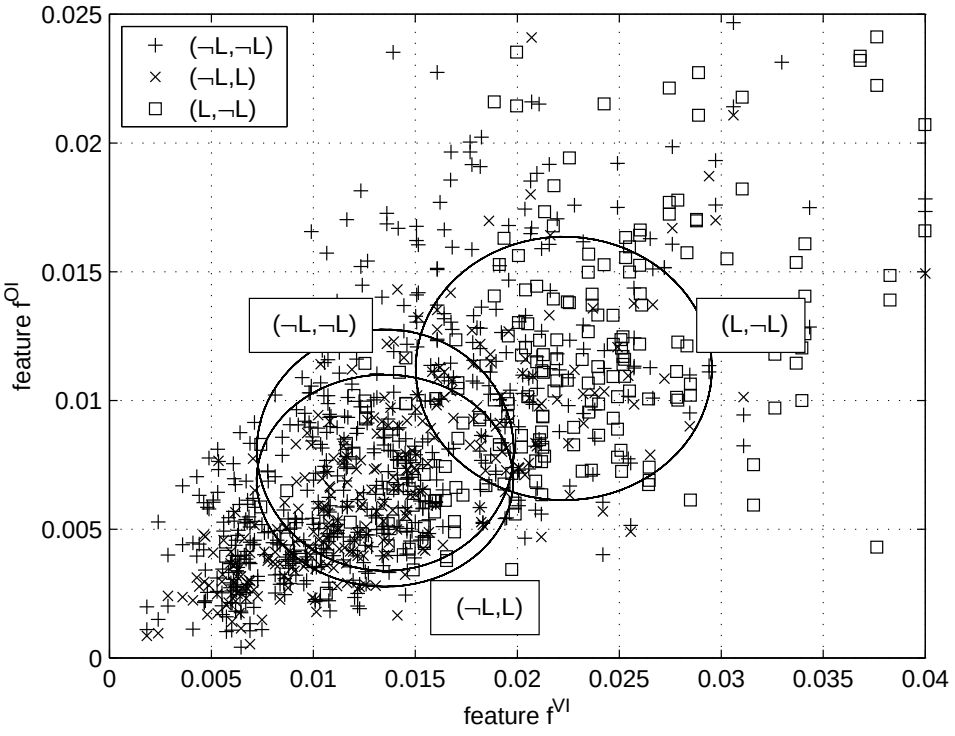


Figure 17.1: Distribution of $(f_k^{VI}, f_{k,j}^{OI})$ pairs for each of $(\neg L, \neg L)$, $(\neg L, L)$, and $(L, \neg L)$. Ellipses are centered on amiTrainSet means and encompass one standard deviation.

Figure 17.1 shows the distribution of two of the selected features, f_k^{VI} and $f_{k,j}^{OI}$, for the data in amiTrainSet; also shown are the first standard deviations of the single-Gaussian diagonal-covariance models induced. It can be seen that f_k^{VI} and $f_{k,j}^{OI}$ are correlated, i.e. that the probability of beginning a talkspurt in silence is positively correlated with the probability of beginning a talkspurt when someone else is speaking. L consistently begins more talkspurts, both in silence and during other people's speech. It is also interesting that $\neg L$ is slightly less likely to initiate a talkspurt when L is already speaking than when another $\neg L$ is. This suggests that $\neg L$ participants consistently observe the L-status of the already speaking party when contemplating talkspurt production. Finally, it is clear that neither the probability of continuing a talkspurt f_k^{VC} (related to talkspurt duration) nor f_k^V (related to overall amount of talk) are by themselves good $L/\neg L$ discriminators.

17.8 Inferring Gender

Gender classification is an example of a task in which the search space $\mathbb{G}$ is a Cartesian product of G; each participant can independently be male or female. This makes the membership model relevant. Experiments in inference of gender were conducted using both the AMI Meeting Corpus and the ICSI Meeting Corpus. In both corpora, gender is encoded in the first letter of each participant's unique identifier. The ratio of male to female occurrences is 2 : 1 in amiTrainSet, and

4 : 1 in ICSITRAINSET. Choosing the majority class leads to gender classification rates of 65% and 81% on AMIEVALSET and ICSIEVALSET, respectively.

Somewhat surprisingly, no single feature type of those studied leads to AMIEVALSET or ICSIEVALSET classification rates higher than those obtained by hypothesizing all participants to be male. On AMIDEVSET, one feature type ($f^{OI}_{k,j}$) yields negligibly better accuracy, but does not generalize to the corresponding evaluation data. Furthermore, the association between reference gender labels and hypothesized gender labels, on both evaluation sets, does not appear to be statistically significant at the $p < 0.05$ level. This finding, that males and females do not differ significantly in their deployment of talkspurts, is likely a consequence of the social structure of the particular groups studied. The fact that AMI roles are acted may also have an effect.

17.9 Inferring Seniority

A second example with non-unique participant classes, and arguably the most interesting of this chapter, is the inference of organizational seniority in the ICSI Meeting Corpus. Following corpus recording, participants indicated, among other things, their level of education. These self-annotations have been clustered for the purposes of this section into $N_G = 3$ mutually exclusive seniority categories[2]. Each participant's seniority is drawn independently from $G = \{\text{GRAD}, \text{PHD}, \text{PROF}\}$; a breakdown for ICSITRAINSET is shown in Table 17.2. Choosing the majority class ($P(\text{PHD}) = 0.444$ on ICSITRAINSET) yields a classification accuracy of 45% on ICSIEVALSET. Education level is closely correlated with age group.

Seniority	Number of		
	spkrs	occur	meets
GRAD	15	81	33
PHD	13	87	29
PROF	3	28	28
all	31	196	33

Table 17.2: Breakdown by seniority $\mathcal{S}$ in ICSITRAINSET by the number of unique participants (spkrs), the number of occurrences (occur), and the number of meetings (meets) in which each seniority occurs.

17.9.1 Classifying Participant Types Independently of Conversation Type

As a first experiment in this section, seniority is classified independently of meeting type,

$$\mathbf{g}^* \doteq \underset{\mathbf{g} \in \mathbb{G}}{\arg\max} \; P\left(\mathbf{g} \,|\, \mathbf{\Theta}^M_u\right) \cdot P\left(\mathbf{F}\left(\mathbf{\Theta}^{CI}(\mathbf{Q})\right) \,|\, \mathbf{\Theta}^B_{\mathbf{g}}\right) \tag{17.21}$$

The three best feature types, determined using ICSIDEVSET, are f^V_k, $f^{OI}_{k,j}$, and $f^{OC}_{k,j}$. These represent the probability of speaking, of beginning a talkspurt when a specific seniority participant is already speaking, and of continuing a talkspurt when a specific seniority participant is speaking, respectively; classification rates are 52%, 59%, and 59%, respectively. When used together, these three feature types produce the confusion matrix shown in Table 17.3 and a rate of 61%, better than when all feature types are used (58%). This represents a 28% relative error reduction over majority-class guessing. As can be seen in the table, association between the reference and hypothesized seniority assignments is statistically significant on unseen data. It is also evident that confusion between GRAD and PROF is lower than between more proximate seniority levels.

Figure 17.2 shows the distribution of $(f^V_k, f^{OC}_{k,j})$ pairs in ICSITRAINSET, together with the first standard deviation, for each combination of the already speaking seniority participant and the seniority participant initiating a new talkspurt (except for (PROF, PROF), since there is at most one PROF in each ICSITRAINSET meeting).

[2] GRAD includes "`Grad`", as well as "`Undergrad`", "`B.A.`", and "`Finished BA in 2001`", due to their small number of exemplars; PHD includes "`PhD`" and "`Postdoc`"; and PROF includes "`Professor`" only.

Ref	Hyp		
	GRAD	PHD	PROF
GRAD	++**11**	26	3
PHD	− 2	++**41**	− 3
PROF	0	−− 6	++**10**

Table 17.3: Confusion matrix for seniority classification on ICSIEVALSET; reference assignment is found in the rows, hypothesized assignment in columns. Highlighting and use of "++", "+", "−", and "−−" as in Table 17.1.

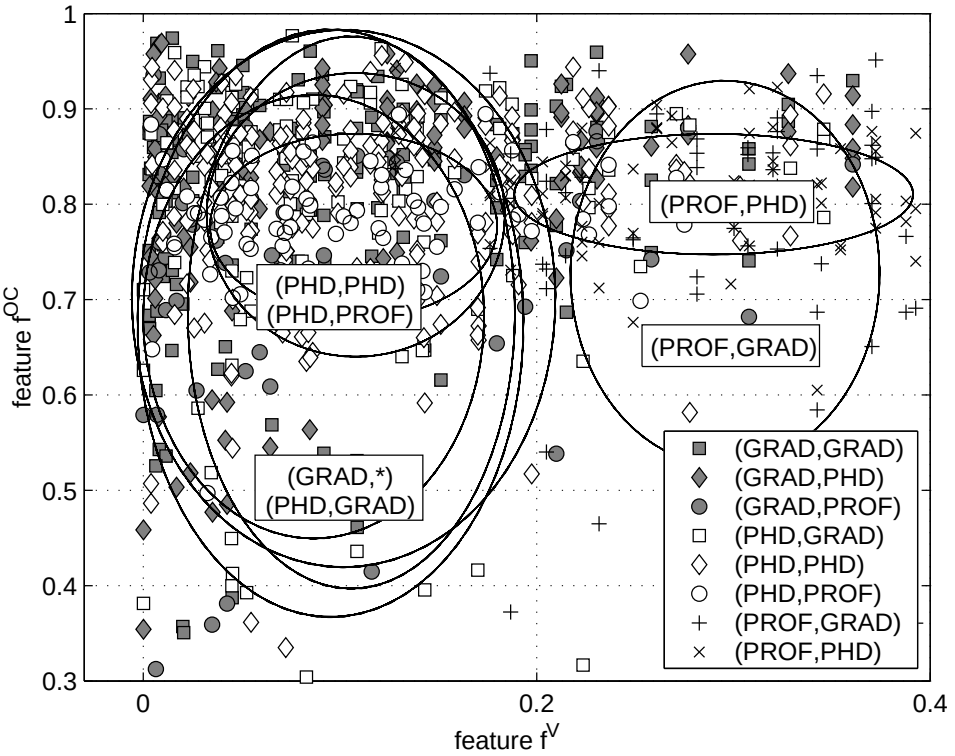


Figure 17.2: Distribution of $(f_k^V, f_{k,j}^{OC})$ feature value pairs for each of the (k, j) participant pairs (GRAD, GRAD), (GRAD, PHD), (GRAD, PROF), (PHD, GRAD), (PHD, PHD), (PHD, PROF), (PROF, GRAD), and (PROF, PHD). Ellipses are centered on ICSITRAINSET means and encompass one standard deviation.

As is clear from the figure, PROF participants in this data talk more than either of the two other seniority types. The figure also demonstrates a difference of behavior during speech overlap. The four ellipses describing GRAD behavior when overlapping with any of the other three classes, as well as PHD behavior when overlapping with GRAD participants, are relatively broad and indicate the absence of strong tendency or preference. However, PHD participants are more likely to continue vocalizing in overlap with other PHD participants, and even more likely to continue through overlap with PROF participants. A similar trend is apparent for PROF participants: the mean likelihood that they continue vocalizing in overlap with GRAD participants lies below $\mu - \sigma$ (bottom 17%) of their model with PHD participants. It appears that the senior researchers in this data are consciously minimizing their overlap with students, who talk less, to make it easier for the latter to speak up.

17.9.2 Conditioning on Conversation Type

The same experiments are repeated, using Equation 17.2, in which separate models are constructed for all three meeting types $T \equiv \{\texttt{Bed}, \texttt{Bmr}, \texttt{Bro}\}$, and compete with one another. The classification accuracy, when behavior model parameters $\left(\mu_u, \sigma_u^2\right)$ are simply their maximum likelihood estimates, is identically 61%, showing no improvement over the experiments in which meeting type was ignored. However, when the parameters are smoothed towards their meeting-type-independent maximum likelihood values $\left(\mu, \sigma^2\right)$, using

$$\hat{\mu}_u = \alpha\mu_u + (1-\alpha)\,\mu\,, \tag{17.22}$$

$$\hat{\sigma}_u^2 = \alpha\sigma_u + (1-\alpha)\,\sigma^2\,, \tag{17.23}$$

an accuracy of 63% on ICSIEVALSET is achieved; the value of $\alpha = 0.7$ was selected using ICSIDEVSET. Furthermore, if instead of estimating the prior on conversation type $P(u)$ from the training data, the meeting type estimates from Chapter 16 are used, the classification rate increases to 67%. A control experiment in which the true type u_{test} of each test meeting is known, i.e. $P(u) = 1$ if $u_{test} = u$ and 0 otherwise, shows that the maximum accuracy achievable under optimal $P(u)$ estimation is 73%.

17.10 Inferring Identity

The last task considered in this chapter is the inference of each participant's identity, based only on the occurrence of speech activity in time and across all participants. Since the AMI meetings contain each of the recorded participants only a handful of times at the most, the ICSI meeting corpus provides a more appropriate testbench for this task.

Because many of the individuals recorded in the ICSI Meeting Corpus occur rarely, the space of possible identities in the experiments presented consists of only those individuals which occur in at least 7 of the ICSITRAINSET meetings; there are 14 such participants. All other participants are mapped to the class OTHER. This renders G a set of 15 alternatives, $G \equiv \{g_1, g_2, \cdots, g_{14}, \text{OTHER}\}$. It should be noted that although the named participants g_1 through g_{14} may occur only once in any meeting's $\mathbf{g}$, OTHER can occur multiple times. Under these constraints, the total number of possible ordered drawings $\mathbf{g}$ of K participants from $\mathbb{G}$ is

$$|\mathbb{G}| = \sum_{M=0}^{K} \frac{K!}{(K-M)!\,M!} \cdot \frac{(|\mathcal{G}|-1)!}{(|\mathcal{G}|-1-M)!}\,. \tag{17.24}$$

The first term in Equation 17.24 represents the number of combinations of M indices in $\mathbf{g}$ at which M non-OTHER participants are found; the second term represents the number of permutations of $|\mathcal{G}|-1$ non-OTHER participants, taken M at a time.

As Equation 17.24 illustrates, the number of possible multiparticipant alternatives $\mathbf{g}$ can be intractably large. A greedy algorithm (shown below) will be applied, which does not exhaustively iterate over $\mathbb{G}$. At every point in the algorithm's execution, $\mathcal{G}'$ is the set of currently unhypothesized specific participants, and $\mathcal{I}$ is the set of indices in $\{1, 2, \cdots, K\}$ currently unoccupied by specific participants.

1. $\mathcal{G}' = \mathcal{G}$. $\mathcal{I} = \{1, 2, \cdots, K\}$. $\mathbf{g}[k] =$ OTHER, for all $1 \leq k \leq K$. Estimate u-conditioned MMs and BMs. Compute $LL = \text{score}(\mathbf{F}, \text{MM}, \text{BM})$. $LL^* = LL$.
2. While $\mathcal{I} \neq \emptyset$,
   (a) $g^* = \emptyset$. Set g to the first element of $\mathcal{G}'$.
   (b) Set i to the first element in $\mathcal{I}$.
   (c) $\mathbf{g}[i] = g$. Estimate u-conditioned MMs and BMs. Compute $LL = \text{score}(\mathbf{F}, \text{MM}, \text{BM})$. If $LL > LL^*$, $\mathbf{g}^* = \mathbf{g}$, $LL^* = LL$, $g^* = g$, and $i^* = i$.
   (d) $\mathbf{g}[i] =$ OTHER. Set i to the next element of $\mathcal{I}$. If $i \neq \emptyset$, return to Step 2c.
   (e) Set g to the next element of $\mathcal{G}'$. If $g \neq \emptyset$, return to Step 2b.
   (f) Remove g^* from $\mathcal{G}'$. Remove i^* from $\mathcal{I}$. Return to Step 2.

The algorithm aims to identify all K participants, one participant per iteration at a time. Step 1 hypothesizes a background participant model at each index $1 \leq i \leq K$. The algorithm then enumerates over the currently still unhypothesized specific participants $\mathcal{G}'$. Each such participant is evaluated as being at each currently still unused index in $\mathcal{I}$. Once the first participant is located at his/her best index, the algorithm proceeds to identify a next participant, by enumerating over all remaining participants and over remaining vacant indices.

The numbers presented in this chapter should be contrasted with majority class guessing. Always guessing the most probable single participant in ICSITRAINSET yields accuracies of 11.9% and 11.8% on ICSIDEVSET and ICSIEVALSET, respectively (always guessing OTHER yields 20.6% and 22.9%, respectively, but effectively fails to identify any specific participants).

Two different experiment suites are presented. In the first, it is assumed that the set of participants is known, and that only their correct permutation must be determined. This makes it possible to assess the behavior model under the assumption of a perfect membership model. The second experiment suite assumes nothing about membership, and effectively exercises both models at the same time.

17.10.1 Known group g

To exercise the behavior model in isolation, the true type u^* of each ICSI meeting is assumed to be known. Also assumed known is the identity of all participants, but not their permutatation. Estimation of the correct permutation in each test meeting is achieved by computing

$$\mathbf{g}^* \doteq \arg\max_{\mathbf{g} \in \mathbb{G}} P\left(\mathbf{F}\left(\mathbf{\Theta}^{CI}(\mathbf{Q})\right) \mid \mathbf{\Theta}^{B}_{u^*,\mathbf{g}}\right) . \tag{17.25}$$

The number of permutations $|\mathbb{G}|$ is nominally $K!$, except that where two or more participants are of class OTHER.

With the behavior model trained as in other sections of this chapter, the ICSIDEVSET accuracy is 60.2%. To complement this finding, the experiment is repeated with different framing policies, each of which is described by a frame step $\Delta T \in \{50, 100, 200, 400, 800, 1600\}$ milliseconds, and a frame size of $T_s = 2\Delta T$. This leads to 6 feature families $\mathbf{F}^{\mathcal{V}}_{\Delta T/T_s}$. The classification results are shown in Table 17.4, when $\mathcal{V}$ is the speech segmentation $\mathcal{S}$, in column 3.

ΔT	T_s	Segmentation Type $\mathcal{V}$		
(ms)	(ms)	$\mathcal{S}$	$\mathcal{S}-\mathcal{B}$	$\mathcal{L}$
50	100	55.9	57.6	36.4
100	200	**60.2**	56.8	**42.4**
200	400	**60.2**	53.4	35.6
400	800	55.1	**58.5**	31.4
800	1600	47.5	47.5	38.1
1600	3200	54.2	56.8	32.2

Table 17.4: Identity classification accuracy using all feature types in each feature family $\mathbf{F}^{\mathcal{V}}_{\Delta T/T_s}$, on ICSIDEVSET, for 6 different framing policies and 3 different binary segmentations. For each test meeting, participant identities are known (but not attributed), and the meeting type is known. Best performing $\mathbf{F}^{\mathcal{V}}_{\Delta T/T_s}$'s for each segmentation type are shown in bold.

As the table shows, performance appears to be relatively stable as a function of the framing policy, with the best accuracy for $\mathcal{S}$ observed for the $\Delta T = 100$ ms frame step used elsewhere in this chapter. The table also provides contrastive performance for systems which rely not on all speech $\mathcal{S}$, but on all speech excluding backchannels $\mathcal{S}_{\neg \mathfrak{b}}$ and only all laughter $\mathcal{L}$. Excluding backchannels appears not to benefit classification rates, but the differences are small. Accuracies obtained with laughter, on the other hand, are almost 20%abs lower than those obtained with speech.

To determine whether any of the 18 feature families shown in Table 17.4 are complimentary, forward feature selection was performed while maximizing ICSIDEVSET accuracy. The top five families were identified to be $\mathbf{F}^{\mathcal{S}}_{0.1/0.2}$, $\mathbf{F}^{\mathcal{S}}_{0.2/0.4}$, $\mathbf{F}^{\mathcal{S}}_{0.8/1.6}$, $\mathbf{F}^{\mathcal{S}-\mathcal{S}_{\neg \mathfrak{b}}}_{0.2/0.4}$, and $\mathbf{F}^{\mathcal{L}}_{0.1/0.2}$. It should be noted that this superset contains all three of the considered vocalization

types and several different framing policies. Accuracies for each of these five families separately and together, for both ICSIDEVSET and ICSIEVALSET, are shown in Table 17.5. Also shown in each row, representing each feature family, is the accuracy obtained with the other four families.

Feature family	ICSIDEVSET		ICSIEVALSET	
	accur	compl	accur	compl
$\mathbf{F}^{\mathcal{S}}_{0.1/0.2}$	60.2	60.2	50.0	52.0
$\mathbf{F}^{\mathcal{S}}_{0.2/0.4}$	60.2	61.9	45.1	51.0
$\mathbf{F}^{\mathcal{S}}_{0.8/1.6}$	47.5	61.0	**54.9**	50.0
$\mathbf{F}^{\mathcal{S}-\mathcal{B}}_{0.2/0.4}$	53.4	59.3	48.0	50.0
$\mathbf{F}^{\mathcal{L}}_{0.1/0.2}$	42.4	63.6	29.4	**54.9**
all 5 $\mathbf{F}^{\mathcal{V}}_{\Delta T/T_S}$'s	**69.5**		53.9	

Table 17.5: Identity classification accuracy (%) using each of the best five feature families by themselves (accur), together (all 5 $\mathbf{F}^{\mathcal{V}}_{\Delta T/T_S}$'s), and leaving each of the five out, one at a time (compl), for both ICSIDEVSET and ICSIEVALSET. For each test meeting, the meeting type and participant identities are known (but not attributed); the best-performing feature families are shown in bold.

Table 17.5 shows that on ICSIDEVSET, while there is significant variation in individual feature family performance, all 5 feature families are nececssary to achieve the combined system accuracy of 69.5%. However, on ICSIEVALSET, $\mathbf{F}^{\mathcal{S}}_{0.8/1.6}$ alone outperforms the combined system. It also appears that $\mathbf{F}^{\mathcal{L}}_{0.1/0.2}$, the only feature family relying on the laughter segmentation $\mathcal{L}$, hurts rather than helps overall system performance.

The combined system achieves an accuracy of 53.9% on completely unseen ICSIEVALSET data, which exceeds guessing the most frequently occurring named (non-OTHER) participant by 42%abs, yielding a reduction of classification error of 48%rel. Closer analysis of the behavior of the combined system indicates that of the 14 named participants, 2 are recognized correctly 100% of the time. The average recall over all 14 is 54%. Seven of these named participants are occasionally hypothesized as OTHER, but the average leakage to the OTHER model per named participant is 13%. Finally, of the 27 errors involving the misclassification of a non-OTHER participant as another non-OTHER participant, 10 are same-seniority and 13 are same-gender; no consistent typology of errors is evident.

17.10.2 Unknown group g

Repeating the above experiments, but this time without knowledge of meeting type or the (unordered) set of participants in advance, requires the computation of all factors of Equation 17.2. The results are shown in Table 17.6.

Feature family	ICSIDEVSET		ICSIEVALSET	
	accur	compl	accur	compl
$\mathbf{F}^{\mathcal{S}}_{0.1/0.2}$	39.0	30.5	27.5	33.3
$\mathbf{F}^{\mathcal{S}}_{0.2/0.4}$	30.5	35.6	20.6	**37.3**
$\mathbf{F}^{\mathcal{S}}_{0.8/1.6}$	16.9	31.4	28.4	32.4
$\mathbf{F}^{\mathcal{S}-\mathcal{B}}_{0.2/0.4}$	29.7	33.1	24.5	**37.3**
$\mathbf{F}^{\mathcal{L}}_{0.1/0.2}$	16.9	**40.7**	21.6	24.5
all 5 $\mathbf{F}^{\mathcal{V}}_{\Delta T/T_S}$'s	29.7		30.4	

Table 17.6: Identity classification accuracy (%) when meeting type and participant identities are not known; symbols as in Table 17.5.

As can be seen in the table, classification accuracies are significantly lower when the K group participants must be

not only permuted into the correct arrangement but also first drawn from a population larger than K. When all 5 feature families are used, classification accuracy is reduced by 44% relative on ICSIEVALSET; it is reduced by 32% relative when only the best performing feature family combinations from among those shown are used. These reductions are likely attributable to both the membership model and the search algorithm, but also to a lack of discrimination in the behavior model which clearly allows many participants to be mistaken for one another. Nevertheless, the presented accuracies are significantly above always guessing the most frequently occurring named or OTHER participants in the training material, indicating that participants exhibit idiosyncratic preferences for vocal activity deployment timing.

It is likely that for this more complicated task, much larger amounts of data are needed to train participant-specific models. Retraining the combined 5 feature family system in Table 17.6 on both ICSITRAINSET and ICSIDEVSET, leads to an increase in ICSIEVALSET accuracy of 3.9%abs, or 13%rel. The relative increase in the amount of training material in this case, in terms of the number of meetings, is 50%rel.

17.11 Potential Impact

That participants to conversation may vary in their preferences for the deployment of speech activity in time is well known and has been extensively studied by the sociolinguistic community. The current chapter has shown that the observed occurrence of speech activity in time and across participants actually discriminates among participant profiles along several dimensions. This has some implications which go beyond the mere classification of participant role or social status characteristics.

First, the models explored are hypermodels over model probabilities of choosing to speak or not to speak. These probabilities are precisely those which guide speech activity detection systems. Since they discriminate among participant types, it is reasonable to assume that improved speech activity detection can be achieved by conditioning those models used in detection on participant characteristics.

Second, that speech activity deployment probabilities discriminate among participant profiles has some potentially important lessons for dialogue system design. It is currently believed that naturally interacting agents should exhibit coherent synthetic personalities. Such personalities, defined at least in part in terms of social characteristics or intended system role in interaction, may appear more compelling and easier to interact with if their speech activity deployment decisions are trained on observed human-human specific-personality material.

Finally, there is strong likelihood that specific participant profiles are also significantly associated with specific types of behavior, such as illocutionary intent or epiemotional behavior. It may turn out that joint inference of conversation type, participant type, and dialog act structure leads to improved performance on all three tasks.

17.12 Relevance to Other Chapters

The experiments in this chapter have relied on the modeling of probabilities obtained using the parametric state-space multiparticipant model proposed in Chapter 7. In many ways, they are the dual of the experiments of Chapter 16, in which the task was to marginalize out participant types while classifying conversation types.

17.13 Summary

This chapter has demonstrated that the fine-grained dynamics of vocal participation to conversation are not only different for different participants but also that they discriminate among participant types. The proposed techniques leverage the observed probabilities of speech/non-speech deployment for all participants individually and for all participant pairs, and attempt to classify the group of participants holistically, allowing only certain configurations of groups to be considered.

Several tasks were considered, summarized in Table 17.7. These included assigned role, assigned leadership, gender, seniority, and identity (the latter not shown). As can be seen in the table, for all tasks except inference of participant gender, the performance of the proposed systems is much higher than guessing using priors. Most relevant to the classification of assigned roles appear to be the probabilities of beginning speech in both silence (f_k^{VI}) and when someone else is already speaking ($f_{k,j}^{OI}$ and $\langle f_{k,j}^{OI} \rangle_j$). The classification of social status characteristics, based on observations from seniority

classification, benefits most from overall talkativity (f_k^T), the probability of beginning speech when a specific other seniority participant is already speaking ($f_{k,j}^{OI}$), and the probability of continuing to speak in overlap with a specific other seniority participant ($f_{k,j}^{OC}$).

Feature Type	AMI			ICSI		
	$\mathcal{R}$	$\mathcal{L}$	$\mathcal{H}$	$\mathcal{H}$	$\mathcal{S}$	$\mathcal{S}\|t^*$
f_k^V	44	—	—	—	*52	*57
f_k^{VI}	*41	*60	—	—	52	56
f_k^{VC}	34	—	—	—	—	62
$\langle f_{j,k}^{OI} \rangle_j$	44	—	—	—	47	56
$\langle f_{k,j}^{OI} \rangle_j$	*29	*60	—	—	49	59
$f_{k,j}^{OI}$	*53	*60	64	—	*59	*59
$\langle f_{j,k}^{OC} \rangle_j$	24	—	—	—	—	57
$\langle f_{k,j}^{OC} \rangle_j$	—	—	—	—	54	59
$f_{k,j}^{OC}$	—	—	—	—	*59	*63
top 3*	53	60	—	—	61	67
all	46	75	43	47	58	57
priors	25	25	65	81	45	45

Table 17.7: Comparative classification performance for 3 experiments on AMIEVALSET and 3 experiments on ICSIEVALSET, per feature type; $\mathcal{R}$, $\mathcal{L}$, $\mathcal{H}$, and $\mathcal{S}$ represent role, leadership, gender, and seniority. Also shown is performance on the best three feature types (selected using development data) and all feature types, as well as that when choosing the majority class ("prior"), informed by training data priors; for $\mathcal{R}$ and $\mathcal{L}$ classification, "prior" performance is equal to random guessing. "—" indicates that a feature type, by itself, did not perform above the corresponding "prior" rate; top-3 feature type selection indicated by "*".

Numerically, the performance of the proposed text-independent systems leaves much room for improvement, but it should be noted that all of these tasks were conducted for the first time. Since the original presentation of these results in [146, 152], however, role recognition in the AMI Meeting Corpus has become a popular area of research. To date, the text-independent approaches relying on reference speech activity segmentation in [59, 68, 58] have not achieved the accuracies reported here, while [195] has obtained a 3%abs improvement.

17.14 Future Directions

The evidence presented suggests that probabilities of speech/non-speech activity are surprisingly discriminative of social characteristics and assigned roles in meetings. This suggests that there is significant potential in participant-specific modeling of observed vocal interaction.

Future work attempting tasks such as those in this chapter should investigate more suitable approaches to model probabilities. Here, probabilities were modeled with single Gaussians, an expedient but clearly suboptimal choice, especially for values close to zero or unity. Models which are equally simple, but offer a better match to empirical distributions, are likely to lead to better discrimination.

Another decision made in the current work was the factorization of the membership model. Although it allowed for K-independence during training and testing, the proposed form does not allow for the enforcement of potentially desirable constraints, such as likelihood of the presence of a specific number of specific participant profiles.

Finally, the presented experiments should be carried out using automatically detected speech activity, rather than only reference speech activity. This would make the techniques deployable in real circumstances, in which manual or human-mediated segmentation is not available. Doing so would also provide auxillary metrics or novel constraints on speech activity detection itself, whose traditional metrics may turn out to be suboptimal for specific applications such as participant characterization. Most felicitous would be if the models used in speech activity detection could be adapted

using only audio and first-pass speech activity hypotheses, and if the adapted model parameters would then serve as features in online or offline classification of participants.

Part IV

CLOSING MATERIAL

Chapter 18

Summary

Spoken conversation is a ubiquitous and crucial social activity. It is overwhelmingly the main means of communication in all settings in which the parties are co-present, and many other social activities are organized around it. Spoken conversation can arguably be said to define us as a species.

It is also uncommonly variable. Depending on ambient acoustics, cultural setting, group size and social status, desired and anticipated outcome, the potentially fluid intentions and internal states of individual participants, and other circumstantial factors, the forms that conversations take are essentially non-enumerable. This is true even when we ignore the language that is used, the words that are spoken, and the way in which those words are spoken.

Despite this variability, conversational form is patently not random. All conversations exhibit a common set of emergent characteristics, which their participants generally appear to intentionally maintain. This set, the systematics of conversation, is sufficiently permissive to license a large degree of observed conversational variability, but not all. Social science has diligently documented ways in which conversational form appears conditioned on various circumstantial factors.

This thesis has provided a computational counterpart to that theoretical and empirical body of work. A crucial element has been the design of a **generic prior probability model of multi-participant conversational dynamics**, which is both time-invariant and independent of the number and identity of participants. The model is important, if only because it entails a data-driven mapping from the space of any and all possible conversations to the one-dimensional probability space of likelihoods. It enables or facilitates, for the first time: (1) quantitative assessment of the variability in the deployment of speech activity between arbitrary conversations (Chapter 10); (2) quantitative comparison of the variability in the deployment of vocal activity between different vocal activity types (Chapter 12); (3) quantitative comparison of the variability in the deployment of speech activity between different conversation types (Chapter 16); and (4) quantitative comparison between alternate multi-participant speech activity states during detection, leading to improved, state-of-the-art speech activity detection systems (Chapter 11). As such, the proposed time-invariant, group-size-independent, and participant-independent modeling approach can be seen to effectively quantify the qualitative models described by conversation analysts for decades.

However, a computational theory of the distribution of vocal activity in time and across participants must not only be able to identify systematic departure from global statistical norms but it must also make it possible to account for those departures. This thesis has achieved that goal by separately relaxing two assumptions of the generic time-invariant, group-size-independent, and participant-independent model.

The first assumption which was relaxed was the time-invariance of the model. Parameters of models describing the distribution of vocal activity in conversation were shown to depend on the instantaneous illocutionary intent and instantanous socio-emotional state of participants, both of which were assumed to vary over the course of a conversation but not to differentiate among participants. This variability was successfully modeled to allow for inference of both intent and state. Attempts to recognize dialog acts differing in illocutionary intent led to the first-ever text-independent dialog act recognition system (Chapter 14). When combined with word-independent prosodic features, the performance of this system was shown to approach that of an oracle lexical system, particularly for turn-boundary phenomena. Based on the largest available corpus of naturally occurring multi-party conversation, the presented findings indicate that departure from time-invariant norms of speech activity distribution in time can be computationally accounted for by what participants

are trying to achieve by speaking at each instant.

In contrast, socio-emotional state was found to be a much stronger function of the occurrence of laughter activity in time and across participants than of the distribution of speech activity. Several different systems, all of which were first-of-their-kind and currently state-of-the-art, validated this finding (Chapter 15). A classifier of emotional valence in manually segmented utterances was revealed to rely almost exclusively on the presence of speaker laughter (Section 15.4). A detector of "hotspots", containing emotionally involved speech, relied more on the distribution of laughter from all participants than on the distribution of speech from all participants (Section 15.5). Finally, the detection of attempts to amuse based on local laughter context considerably outperformed detection based on local speech context and on oracle lexical bigrams (Section 15.6). These experiments, conducted on two of the largest available multi-party conversation corpora, indicate that departure from time-invariant norms of laughter activity distribution in time can be computationally accounted for by what participants feel, how strongly they feel, and how they are trying to make others feel.

The second assumption which was relaxed was the participant-independence of the model. Parameters of models describing the distribution of vocal activity in conversation were shown to depend on the assigned roles and social status characteristics of participants, both of which were assumed to not vary over the course of a conversation but to differentiate among participants. This variability was successfully modeled to allow for inference of both role and status (Chapter 17). It led to the first-ever classifier of assigned role in the AMI Meeting Corpus (Section 17.6), whose error rate has since — despite several published attempts — been beaten only once, by 6%rel with a 50% increase in training material. The same system was also successfully applied to the inference of seniority in the ICSI Meeting Corpus (Section 17.9), and to date remains the only baseline for this task. A similar system, in which no participant labels were relied on, was also applied to improve the classification of conversation type (Chapter 16). These experiments indicate that departure from time-invariant norms of vocal activity distribution across participants can be computationally accounted for by participant characteristics, whether assigned or inherent.

The successful inference of such a broad spectrum of attributes, using a most prosaic representation of spoken conversation, constitutes an achievement of computational understanding. This achievement is remarkable in part because it does not rely on semantics; the representation used — a speaker-attributed, temporal segmentation of vocal activity — entirely ignores words and their boundaries. Instead, the explored attributes are pragmatic, and account for emergent but possibly not voluntary or conscious tactical behavior. In modeling how they shape and contextualize participant contributions to conversation, this thesis helps to begin to close an increasingly glaring gap in speech understanding systems and conversational interfaces.

Chapter 19

Future Research Enabled by This Thesis

This thesis has argued for, and demonstrated, the potential value for speech understanding systems of modeling a text-independent, joint multi-participant representation of conversation. For many of the explored tasks, the specific approach which was chosen should be treated as a baseline in subsequent work. As described in individual chapters, there is much scope for improving performance on the various tasks individually.

To complement this, however, a future recommended course of action is to consolidate many of the proposed techniques into a more parsimonious framework. Although the models deployed to address the needs of specific applications in this thesis have much in common, their differences are cumbersome. In most cases, those differences are due to application-specific assumptions without which the corresponding search spaces would be intractably large. Effort should be invested in identifying and clustering subspaces such that the same models can be used for a variety of tasks. For example, it would be preferable to the current state of affairs if the same model form as used in speech activity detection could be used for participant characterization, by adapting the parameters of a participant-independent first-pass model. Techniques similar to this, but focusing on acoustic modeling, are widely used in speaker recognition, by adapting speaker-independent speech recognition models. (There is of course also scope for improving speech activity estimates via a second pass with the adapted models.) Another example is improvement in dialog act recognition, given estimates of speaker characteristics, and vice versa. For these kinds of joint or cross inference, a single model form would be much more appealing than the array of model forms explored in this work.

An avenue of research which is important in qualifying many of the results presented is the application of the techniques to two-party dialogue. An important design criterion adopted at the outset of this thesis was that the eventual techniques should be applicable to conversations of arbitrary numbers of participants. Many modeling decisions were taken to make this possible, and it may be the case that some of them are sub-optimal. This can be assessed in two-party dialogue, since the complete search space is likely to be directly manageable. An example where significantly better performance can be expected is dialogue act recognition, where dialog act types and dialog act boundary types could be estimated jointly for both participants, in state space.

Finally, the findings of this thesis should be explored within the framework of online (two-party and multi-party) dialogue systems. In their current form, the proposed systems are more suitable to the observation of conversations which have already occurred, and in which the "observer" is not intended to participate. This justifies, for example, the use of models which describe joint multi-participant behavior. For systems in which the observer is a participant, such models — while important to understand what others are doing — is inconvenient for planning one's own participation. Dialogue systems, and particularly those deployed in potentially multi-party environments, can usefully modify some of the techniques in this thesis to better predict what others are about to do, and thereby to better estimate the appropriateness of the actions they themselves are planning.

Bibliography

[1] Acero, A., Crespo, C., de la Torre, C., and Torrecilla, J. Robust HMM-based endpoint detector. In *Proc 3rd European Conference on Speech Communication and Technology (EUROSPEECH)* (Berlin, Germany, September 1993), International Speech Communication Association (ISCA), pp. 1551–1554. http://www.isca-speech.org/archive/eurospeech_1993/e93_1551.html.

[2] Aist, G., Kort, B., Reilly, R., Mostow, J., and Picard, R. Experimentally augmenting an intelligent tutoring system with human-supplied capabilities: Adding human-provided emotional scaffolding to an automated reading tutor than listens. In *Proc 4th International Conference on Multimodal Interfaces (ICMI)* (Pittsburgh PA, USA, October 2002), Institute of Electrical and Electronics Engineers, pp. 483–490. doi:10.1109/ICMI.2002.1167044.

[3] Ang, J., Dhillon, R., Krupski, A., Shriberg, E., and Stolcke, A. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *Proc 7th International Conference on Spoken Language Processing (ICSLP)* (Denver CO, USA, September 2002), J. Hansen and B. Pellom, Eds., International Speech Communication Association (ISCA), pp. 2037–2040. http://www.isca-speech.org/archive/icslp_2002/i02_2037.html.

[4] Ang, J., Liu, Y., and Shriberg, E. Automatic dialog act segmentation and classification in multiparty meetings. In *Proc 30th International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Philadelphia PA, USA, March 2005), Institute of Electrical and Electronics Engineers, pp. 1061–1064. doi:10.1109/ICASSP.2005.1415300.

[5] Argamon-Engelson, S., Koppel, M., and Avneri, G. Style-based text categorization: What newspaper am i reading? In *Proc Workshop on Learning for Text Categorization* (1998), Association for the Advancement of Artificial Intelligence (AAAI), pp. 1–4. http://www.aaai.org/Papers/Workshops/1998/WS-98-05/WS98-05-001.pdf.

[6] Bachorowski, J.-A., Smoski, M., and Owren, M. The acoustic features of human laughter. *Journal of Acoustical Society of America 110*, 3 (September 2001), 1581–1597. doi:10.1121/1.1391244.

[7] Bales, R. *Interaction Process Analysis: A Method for the Study of Small Groups*. Addison-Wesley Press, Inc., Cambridge MA, USA, 1950.

[8] Bales, R., Strodtbeck, F., Mills, T., and Roseborough, M. Channels of communication in small groups. *American Sociological Review 16*, 4 (August 1951), 461–468. http://www.jstor.org/stable/2088276.

[9] Banerjee, S., Rose, C., and Rudnicky, A. The necessity of a meeting recording and playback system, and the benefit of topic-level annotations to meeting browsing. In *Human Computer Interaction - INTERACT; IFIP TC13 International Conference* (Rome, Italy, September 2005), M. Costabile and F. Paternò, Eds., vol. 3585 of *Springer Lecture Notes in Computer Science*, Springer, pp. 643–656. doi:10.1007/11555261_52.

[10] Banse, R., and Scherer, K. Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology 70*, 3 (March 1996), 614–636. doi:10.1037/0022-3514.70.3.614.

[11] Batliner, A., Fischer, K., Huber, R., Spilker, J., and Nöth, E. Desperately seeking emotions: Actors, wizards, and human beings. In *Proc Tutorial and Research Workshop on Speech and Emotion* (Newcastle, Northern Ireland, September 2000), R. Cowie, E. Douglas-Cowie, and M. Schröder, Eds., International Speech Communication Association (ISCA), pp. 195–200.

[12] Batliner, A., Hacker, C., Steidl, S., Nöth, E., and Haas, J. From emotion to interaction: Lessons from real human-machine-dialogues. In *Affective Dialogue Systems, Tutorial and Research Workshop (ADS)*, E. André, L. Dybkiaer, W. Minker, and P. Heisterkamp, Eds., vol. 3068 of *Lecture Notes in Artificial Intelligence*. Springer, Kloster Irsee, Germany, June 2004, pp. 1–12. doi:10.1007/978-3-540-24842-2_1.

[13] Batliner, A., Steidl, S., Hacker, C., Nöth, E., and Niemann, H. Private emotions vs social interaction — Towards new dimensions in research on emotion. In *Proc Workshop on Adapting the Interaction Style to Affective Factors at the 10th International Conference on user Modeling* (Edinburgh, UK, July 2005). *(page numbers not assigned)*.

[14] Batliner, A., Steidl, S., Schuller, B., Seppi, D., Laskowski, K., Vogt, T., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., and Aharonson, V. Combining efforts for improving automatic classification of emotional user states. In *Proc 5th Slovenian and 1st International Language Technologies Conference (IS-LTC)* (Ljubliana, Slovenia, October 2006), T. Erjavec and J. Zganec-Gros, Eds., Informacijska Druzba, pp. 240–245.

[15] Beranek, L. *Acoustics*. American Institute of Physics, Woodbury NY, USA, 1986. ISBN 978-0883184943.

[16] Berger, J., Rosenholtz, S., and Zelditch Jr., M. Status organizing processes. *Annual Review of Sociology 6* (August 1980), 479–508. http://www.jstor.org/stable/2946017.

[17] Birdwhistell, R. *Kinesics and Context: Essays on Body Motion Communication*. University of Pennsylvania Publications in Conduct and Communication. University of Pennsylvania Press, Philadelphia PA, USA, 1970. ISBN 978-0812210125.

[18] Bishop, C. *Neural Networks for Pattern Recognition*. Oxford University Press, New York NY, USA, 1995. ISBN 978-0198538646.

[19] Boakye, K. *Audio segmentation for meetings speech processing*. PhD thesis, University of California, Berkeley CA, US, December 2008.

[20] Boakye, K., and Stolcke, A. Improved speech activity detection using cross-channel features for recognition of multiparty meetings. In *Proc 9th International Conference on Spoken Language Processing (INTERSPEECH)* (Pittsburgh PA, USA, September 2006), International Speech Communication Association (ISCA), pp. 1962–1965. http://www.isca-speech.org/archive/interspeech_2006/i06_1824.html.

[21] Boersma, P. Praat, a system for doing phonetics by computer. *Glot International 5*, 9/10 (2001), 341–345.

[22] Bohus, D., and Horvitz, E. Models for multiparty engagement in open-world dialog. In *Proc 10th Annual SIGDial Meeting on Discourse and Dialogue* (London, UK, September 2009), Association for Computational Linguistics (ACL), pp. 225–234. http://www.aclweb.org/anthology/W09-3933.pdf.

[23] Bohus, D., and Horvitz, E. Computational models for multiparty turn-taking. Tech. Rep. Microsoft Technical Report MSR-TR-2010-115, Microsoft Research, Redmond WA, USA, 2010.

[24] Brady, P. A technique for investigating on-off patterns of speech. *Bell Systems Technical Journal 44*, 1 (January 1965), 1–22. http://bstj.bell-labs.com/BSTJ/images/Vol44/bstj44-1-1.pdf.

[25] Brady, P. A statistical analysis of on-off patterns in 16 conversations. *Bell Systems Technical Journal 47*, 1 (January 1968), 73–91. http://bstj.bell-labs.com/BSTJ/images/Vol47/bstj47-1-73.pdf.

[26] Brady, P. A model for generating on-off speech patterns in two-way converation. *Bell Systems Technical Journal 48*, 9 (September 1969), 2445–2472. http://bstj.bell-labs.com/BSTJ/images/Vol48/bstj48-7-2445.pdf.

[27] Brdiczka, O., Maisonnasse, J., and Reignier, P. Automatic detection of interaction groups. In *Proc 7th International Conference on Multimodal Interfaces (ICMI)* (Trento, Italy, October 2005), Association for Computing Machinery, pp. 32–36. doi:10.1145/1088463.1088473.

[28] Burger, S. Transliteration spontansprachlicher Daten, Lexikon der Transliterationskonventionen in Verbmobil II. Tech. Rep. Verbmobil Technisches Dokument Nr. 56, Institut für Phonetik und Sprachverarbeitung, Ludwig-Maximilians-Universität, München, Germany, 1997.

[29] Burger, S., Laskowski, K., and Wölfel, M. A comparative cross-domain study of the occurrence of laughter in meeting and seminar corpora. In *Proc. 6th International Conference on Language Resources and Evaluation (LREC)* (Marrakech, Morocco, May 2008), European Language Resources Association (ELRA), p. (page numbers not assigned).

[30] Burger, S., MacLaren, V., and Yu, H. The ISL Meeting Corpus: The impact of meeting type on speech style. In *Proc 7th International Conference on Spoken Language Processing (ICSLP)* (Denver CO, USA, September 2002), International Speech Communication Association (ISCA), pp. 301–304. http://www.isca-speech.org/archive/icslp_2002/i02_0301.html.

[31] Burger, S., and Sloane, Z. The ISL Meeting Corpus: Categorical features of communicative group interactions. In *Proc Rich Transcription Meeting Recognition Evaluation Workshop* (Montreal, Canada, May 2004), National Institute of Standards and Technology (NIST). *(page numbers not assigned).*

[32] Burgoon, J., Stern, L., and Dillman, L. *Interpersonal Adaptation: Dyadic Interaction Patterns.* Cambridge University Press, New York NY, USA, 1995. ISBN 978-0521451208.

[33] Cappella, J., and Streibel, M. Computer analysis of talk-silence sequences: The FIASSCO System. *Behavior Research Methods & Instrumentation 11*, 3 (1979), 384–392.

[34] Carletta, J. Assessing agreement on classification tasks: The Kappa Statistic. *Computational Linguistics 22*, 2 (June 1996), 249–254.

[35] Carletta, J. Unleashing the killer corpus: Experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation Journal 41*, 2 (2007), 181–190. doi:10.1007/s10579-007-9040-x.

[36] Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, M., Post, W., Reidsma, D., and Wellner, P. The AMI Meeting Corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction: 2nd International Workshop (MLMI)* (Edinburgh, UK, July 2005), S. Renals and S. Bengio, Eds., vol. 3869 of *Lecture Notes in Computer Science*, Springer, pp. 28–39. doi:10.1007/11677482_3.

[37] Carletta, J., Garrod, S., and Fraser-Krauss, H. Communication and placement of authority in workplace groups — The consequences for innovation. *Small Group Research 29*, 5 (October 1998), 531–559. doi:10.1177/1046496498295001.

[38] Carletta, J., Isard, A., Isard, S., Kowtko, J., and Doherty-Sneddon, G. HCRC dialogue structure coding manual. Tech. Rep. HCRC/TR-82, University of Edinburgh, Edinburgh, UK, 1996.

[39] Carletta, J., Isard, A., Isard, S., Kowtko, J., Newlands, A., Doherty-Sneddon, G., and Anderson, A. The reliability of a dialogue structure coding scheme. *Computational Linguistics 23*, 1 (March 1997), 13–31.

[40] Cassotta, L. *The stability and modification of the vocal behavior of individuals in stress and nonstress interviews.* PhD thesis, New York University, New York NY, USA, June 1966.

[41] Cavnar, W., and Trenkle, J. N-Gram-based text categorization. In *Proc 3rd Annual Symposium on Document Analysis and Information Retrieval (SDAIR)* (Las Vegas NV, USA, April 1994), pp. 161–175.

[42] Çetin, O., and Shriberg, E. Overlap in meetings: ASR effects and analysis by dialog factors, speakers, and collection site. In *Machine Learning for Multimodal Interaction, 3rd International Workshop (MLMI)* (Bethesda MD, USA, May 2006), S. Renals, S. Bengio, and J. Fiscus, Eds., vol. 4299 of *Lecture Notes in Computer Science*, Springer, pp. 212–224. doi:10.1007/11965152_19.

[43] Chapple, E. Quantitative analysis of the interaction of individuals. *Proceedings of the National Academy of Sciences of the United States of America 25*, 2 (February 1939), 58–67. http://www.jstor.org/stable/87221.

[44] Chapple, E. Measuring human relations: An introduction to the study of the interaction of individuals. *Genetic Psychology Monographs 22*, 1 (February 1940), 3–147.

[45] Chapple, E. The Interaction Chronograph: Its evolution and present application. *Personnel 25*, 4 (January 1949), 295–307.

[46] Chapple, E., Chapple, M., Wood, L., Miklowitz, A., Kline, N., and Saunders, J. Interaction chronograph method for analysis of differences between schizophrenics and controls. *Archives of General Psychiatry 3*, 2 (August 1960), 160–167.

[47] Chapple, E., and Donald Jr., G. An evaluation of department store salespeaple by the interaction chronograph. *The Journal of Marketing 12*, 2 (October 1947), 173–185. http://www.jstor.org/stable/1245356.

[48] Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement 20*, 1 (April 1960), 37–46.

[49] Cuendet, S., Hakkani-Tur, D., Shriberg, E., Fung, J., and Favre, B. Cross-genre feature comparisons for spoken sentence segmentation. *International Journal of Semantic Computing 1*, 3 (September 2007), 335–346. doi:10.1142/S1793351X07000202.

[50] Dabbs Jr., J., and Ruback, R. Vocal patterns in male and female groups. *Personality and Social Psychology Bulletin 10*, 4 (December 1984), 518–525. doi:10.1177/0146167284104004.

[51] Dabbs Jr., J., and Ruback, R. Dimensions of group process: Amount and structure of vocal interaction. *Advances in Experimental Social Psychology 20* (1987), 123–169. doi:10.1016/S0065-2601(08)60413-X.

[52] Dabbs Jr., J., Ruback, R., and Evans, M. "Grouptalk": Sound and silence in group conversation. In *Nonverbal Behavior and Communication*, A. Siegman and S. Feldstein, Eds., 2nd ed. Lawrence Erlbaum Associates, Inc., Hillsdale NJ, USA, 1987, pp. 501–520. ISBN 978-0805800180.

[53] Dellaert, F., Polzin, T., and Waibel, A. Recognizing emotions in speech. In *Proc 4th International Conference on Spoken Language Processing (ICSLP)* (Philadelphia PA, USA, October 1996), vol. 3, International Speech Communication Association (ISCA), pp. 1970–1973. http://www.isca-speech.org/archive/icslp_1996/i96_1970.html.

[54] Dhillon, R., Bhagat, S., Carvey, H., and Shriberg, E. Meeting recorder project: Dialog act labeling guide. Tech. Rep. TR-04-002, International Computer Science Institute, Berkeley CA, USA, February 2004.

[55] Dielmann, A., and Renals, S. Recognition of dialogue acts in multiparty meetings using a switching dbn. *IEEE Transactions on Audio, Speech, and Language Processing 16*, 7 (September 2008), 1303–1314. doi:10.1109/TASL.2008.922463.

[56] Dines, J., Vepa, J., and Hain, T. The segmentation of multi-channel meeting recordings for automatic speech recognition. In *Proc 9th International Conference on Spoken Language Processing (INTERSPEECH)* (Pittsburgh PA, USA, September 2006), International Speech Communication Association (ISCA), pp. 1213–1216. http://www.isca-speech.org/archive/interspeech_2006/i06_1548.html.

[57] Douglas-Cowie, E., Campbell, N., Cowie, R., and Roach, P. Emotional speech: Towards a new generation of databases. *Speech Communication 40*, 1–2 (April 2003), 33–60.

[58] Favre, S., Dielmann, A., and Vinciarelli, A. Automatic role recognition in multiparty recordings using social networks and probabilistic sequential models. In *Proc 17th International Conference on Multimedia (MM)* (Beijing, China, October 2009), Association for Computing Machinery, pp. 585–588. doi:10.1145/1631272.1631362.

[59] Favre, S., Salamin, H., Dines, J., and Vinciarelli, A. Role recognition in multiparty recordings using social affiliation networks and discrete distributions. In *Proc of the 10th International Conference on Multimodal Interfaces (ICMI)* (Chania, Greece, October 2008), Association for Computing Machinery, pp. 29–36. doi:10.1145/1452392.1452401.

[60] Fay, N., Garrod, S., and Carletta, J. Group discussion as interactive dialogue or serial monologue: The influence of group size. *Psychological Science 11*, 6 (November 2000), 487–492. http://www.jstor.org/stable/40063620.

[61] Feldstein, S. Impression formation in dyads: The temporal dimension. In *Interaction Rhythms: Periodicity in Communicative Behavior*, M. Davis, Ed. Human Sciences Press, Inc., New York NY, USA, 1982. ISBN 978-0898850031.

[62] Feldstein, S., Alberti, L., and BenDebba, M. Self-attributed personality characteristics and the pacing of conversational interaction. In *Of Speech and Time: Temporal Speech Patterns in Interpersonal Contexts*, A. Siegman and S. Feldstein, Eds. Lawrence Erlbaum Associates, Inc., Hillsdale NJ, USA, 1979, pp. 73–87. ISBN 978-0470268315.

[63] Filippelli, M., Pellegrino, R., Iandelli, I., Misuri, G., Rodarte, J. R., Duranti, R., Brrusasco, V., and Scano, G. Respiratory dynamics during laughter. *Journal of Applied Physiology 90*, 4 (April 2001), 1441–1446. PMID 11247945.

[64] Fügen, C., Ikbal, S., Kraft, F., Kumatani, K., Laskowski, K., McDonough, J., Ostendorf, M., Stüker, S., and Wölfel, M. The ISL RT-06S speech-to-text system. In *Machine Learning for Multimodal Interaction, 3rd International Workshop (MLMI)* (Bethesda MD, USA, May 2006), S. Renals, S. Bengio, and J. Fiscus, Eds., vol. 4299 of *Lecture Notes in Computer Science*, Springer, pp. 407–418. doi:10.1007/11965152_36.

[65] Galley, M., McKeown, K., Fosler-Lussier, E., and Jing, H. Discourse segmentation of multi-party conversation. In *Proc 41st Annual Meeting of the Association for Computational Linguistics* (Sapporo, Japan, July 2003), Association for Computational Linguistics (ACL), pp. 562–569. doi:10.3115/1075096.1075167.

[66] Galley, M., McKeown, K., Hirschberg, J., and Shriberg, E. Identifying agreement and disagreement in conversational speech: Use of Baysesian networks to model pragmatic dependencies. In *Proc 42nd Annual Meeting of the Association for Computational Linguistics* (Barcelona, Spain, 2004), Association for Computational Linguistics (ACL). doi:10.3115/1218955.1219040.

[67] Garfinkel, H. *Studies in Ethnomethodology*, 8th reprinting of 1st ed. Polity Press, Englewood Cliffs NJ, USA, 1984. ISBN 978-0745600055.

[68] Garp, N., Favre, S., Salamin, H., Hakkani-Tür, D., and Vinciarelli, A. Role recognition for meeting participants: An approach based on lexical information and social network analysis. In *Proceedings of the 16th International Conference on Multimedia (MM)* (Vancouver BC, Canada, October 2008), Association for Computing Machinery, pp. 693–696. doi:10.1145/1459359.1459462.

[69] Germesin, S., and Wilson, T. Agreement detection in multiparty conversation. In *Proc 11th International Conference on Multimodal Interfaces and 6th Workshop on Machine Learning for Multimodal Interaction* (Cambridge MA, USA, November 2009), Association for Computing Machinery, pp. 7–14. doi:10.1145/1647314.1647319.

[70] Glenn, P. *Laugh and the world laughs with you; shared laughter sequencing in conversation.* PhD thesis, University of Texas, Austin TX, USA, 1987. *(unpublished).*

[71] Glenn, P. Initiating shared laughter in multi-party conversations. *Western Journal of Speech Communication 53*, 2 (Spring 1989), 127–149. doi:10.1080/10570318909374296.

[72] Glenn, P. Some techniques for extending conversational shared laughter. In *Proc of the 75th Annual Meeting of the Speech Communication Association* (San Francisco CA, USA, November 1989). *(page numbers not assigned).*

[73] Glenn, P. Shared laughter, intimacy, and play. *(page numbers not assigned).*

[74] Glenn, P. Current speaker initiation in two-party shared laughter. *Research on Language and Social Interaction 25*, 1–4 (1991), 139–162. doi:10.1080/08351819109389360.

[75] Glenn, P. *Situated Order: Studies in the Social Organization of Talk and Embodies Activities*. University Press of America, Lanham MD, USA, 1995, ch. Laughing at and laughing with: Negotiating participant alignments through conversational laughter, pp. 43–56. ISBN 978-0819196255.

[76] Glenn, P. *Laughter in Interaction*, vol. 18 of *Studies in Interactional Sociolinguistics*. Cambridge University Press, Cambridge, UK, 2003. ISBN 978-0521772068.

[77] Glenn, P., Hoffman, E., and Hopper, R. Woman, laughter, man: Gender and the sequential organization of laughter. American Association of Applied Linguistics. *(page numbers not assigned)*.

[78] Goetsch, G., and McFarland, D. Models of the distribution of acts in small discussion groups. *Social Psychology Quarterly 43*, 2 (June 1980), 173–183. http://www.jstor.org/stable/3033620.

[79] Guerrero, L., DeVito, J., and Hecht, M., Eds. *The Nonverbal Communication Reader: Classic and Contemporary Readings*, 2nd ed. Waveland Press, Inc., Prospect Heights IL, USA, 1999. ISBN 978-1577660408.

[80] Gustafson, J., Heldner, M., and Edlund, J. Potential benefits of human-like dialogue behavior in the call routing domain. In *Perception in Multimodal Dialogue Systems, Proc 4th Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems (PIT)* (Kloster Irsee, Germany, June 2008), E. André, L. Dybkær, W. Minker, H. Neumann, R. Pieraccini, and M. Weber, Eds., vol. 5078 of *Lecture Notes in Computer Science*, Institute of Electrical and Electronics Engineers, Springer, pp. 240–251. doi:10.1007/978-3-540-69369-7_27.

[81] Guze, S., and Mensh, I. An analysis of some features of the interview with the interaction chronograph. *Journal of Abnormal Psychology 58*, 2 (March 1959), 269–271. doi:10.1037/h0046036.

[82] Haakana, M. *Laughing matters; a conversation analytical study of laughter in doctor-patient interaction*. PhD thesis, University of Helsinki, Department of Finnish Language, Helsinki, Finland, 1999. *(unpublished)*.

[83] Hain, T., Burget, L., Dines, J., Garau, G., Karafiat, M., Lincoln, M., McCowan, I., Moore, D., Wan, V., Ordelman, R., and Renals, S. The 2005 AMI system for the transcription of speech in meetings. In *Machine Learning for Multimodal Interaction: 2nd International Workshop (MLMI)* (Edinburgh, UK, July 2005), S. Renals and S. Bengio, Eds., vol. 3869 of *Lecture Notes in Computer Science*, Springer, pp. 450–462. doi:10.1007/11677482_38.

[84] Hain, T., Burget, L., Dines, J., Garau, G., Karafiat, M., Lincoln, M., Vepa, J., and Wan, V. The AMI Meeting Transcription System: Progress and performance. In *Machine Learning for Multimodal Interaction, 3rd International Workshop (MLMI)* (Bethesda MD, USA, May 2006), S. Renals, S. Bengio, and J. Fiscus, Eds., vol. 4299 of *Lecture Notes in Computer Science*, Springer, pp. 419–431. doi:10.1007/11965152_37.

[85] Hakkani-Tür, D. Towards automatic argument diagramming of multiparty meetings. In *Proc 34th International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Taipei, Taiwan, April 2009), Institute of Electrical and Electronics Engineers, pp. 4753–4756. doi:10.1109/ICASSP.2009.4960693.

[86] Hall, E. *The Hidden Dimension*. Knopf Doubleday & Company, Inc., New York NY, USA, 1966. ISBN 978-0385084765.

[87] Hare, A. A study of interaction and consensus in different sized groups. *American Sociological Review 17*, 3 (June 1952), 161–267. http://www.jstor.org/stable/2088071.

[88] Hargreaves, W. A model for speech unit duration. *Language and Speech 3*, 3 (July/September 1960), 164–173. doi:10.1177/002383096000300305.

[89] Harper, R., Wiens, A., and Matarazzo, J. *Nonverbal Communication: The State of the Art*. Wiley Series on Personality Processes. John Wiley & Sons, Inc., New York NY, USA, 1978. ISBN 978-0471026723.

[90] Hebb, D. *The Organization of Behavior: A Neuropsychological Theory*. John Wiley & Sons, New York NY, USA, 1949.

[91] Heldner, M., Edlund, J., Laskowski, K., and Pelcé, A. Prosodic features in the vicinity of silences and overlaps. In *Proc. 10th Nordic Conference on Prosody* (Helsinki, Finland, August 2008), pp. 95–105.

[92] Hertz, J., Krogh, A., and Palmer, R. *Introduction to the Theory of Neural Computation*. Sante Fe Institute Studies in the Sciences of Complexity. Perseus Books Publishing, Reading MA, USA, 1991. ISBN 978-0201515602.

[93] Heylen, D., Nijholt, A., and Reidsma, D. Determining what people feel and think when interacting with humans and machines: Notes on corpus collection and annotation. In *Proc 1st California Conference on Recent Advances in Engineering Mechanics* (Fullerton CA, USA, January 2006), California State University, pp. 1–6.

[94] Hillard, D., Ostendorf, M., and Shriberg, E. Detection of agreement vs. disagreement in meetings: Training with unlabeled data. In *Proc Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volumne, Short Papers* (Edmonton, Canada, May 2003), Association for Computational Linguistics (ACL), pp. 34–36. doi:10.3115/1073483.1073495.

[95] Hopfield, J. Neural networks and physical systems with emergent collective computational abilities. *Proc National Academy of Science of the United States of America 79*, 8 (April 1982), 2554–2558. doi:10.1073/pnas.79.8.2554.

[96] Hopper, R., and Glenn, P. Repetition and play in conversation. In *Repetition in Discourse: Interdisciplinary Perspectives*, B. Johnstone, Ed., vol. 2 of *Advances in Discourse Processes (Volume XLVIII)*. Praeger, Norwood NJ, USA, 1994, pp. 29–40. ISBN 978-0893918316.

[97] Horton Jr., A. The occurrence and effect of lockout in telephone connections involving two echo suppressors. *Bell Systems Technical Journal 17*, 4 (April 1938), 258–280. http://bstj.bell-labs.com/BSTJ/images/Vol17/bstj17-2-258.pdf.

[98] Huang, Z., and Harper, M. Speech activity detection on multichannels of meeting recordings. In *Machine Learning for Multimodal Interaction: 2nd International Workshop (MLMI)* (Edinburgh, UK, July 2005), S. Renals and S. Bengio, Eds., vol. 3869 of *Lecture Notes in Computer Science*, Springer, pp. 415–427. doi:10.1007/11677482_35.

[99] Hutchby, I., and Wooffitt, R. *Conversation Analysis*, 2nd ed. Polity Press, Cambridge, UK, 2008. ISBN 978-0745638652.

[100] Iizuka, H., and Ikegami, T. Adaptive coupling and intersubjectivity in simulated turn-taking behavior. In *Proc 7th European Conference on Advances in Artificial Life (ECAL)* (Dortmund, Germany, September 2003), W. Banzhaf, T. Christaller, P. Dittrich, J. Kim, and J. Ziegler, Eds., vol. 2801 of *Lecture Notes in Computer Science*, Springer, pp. 336–345. doi:10.1007/978-3-540-39432-7_36.

[101] Ising, E. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei 31*, 1 (February 1925), 253–258. doi:10.1007/BF02980577.

[102] Jaffe, J. Computer assessment of dyadic interaction rules from chronographic data. In *Proc Research in Psychotherapy Conference* (Chicago IL, USA, May 1968), J. M. Shlien, Ed., vol. 3, Americal Psychological Association, pp. 260–276.

[103] Jaffe, J., Cassotta, L., and Feldstein, S. Markovian model of time patterns of speech. *Science (New Series) 144*, 3620 (May 1964), 884–886. doi:10.1126/science.144.3620.884.

[104] Jaffe, J., and Feldstein, S. *Rhythms of Dialogue*. Personality and Psychopathology. Academic Press, New York NY, USA, 1970. ISBN 978-0123798503.

[105] Jaffe, J., Feldstein, S., and Cassotta, L. Computation of information measures in diagnostic interviews. In *Data Acquisition and Processing in Biology and Medicine, Proceedings of the Rochester Conference*, K. Enslein, Ed., vol. 3. Pergamon Press, Inc., New York NY, USA, 1964, pp. 143–150.

[106] Jaffe, J., Feldstein, S., and Cassotta, L. Markovian models of dialogic time patterns. *Nature 216* (October 1967), 93–94. doi:10.1038/216093a0.

[107] James, J. A preliminary study of the size determinant in small group interaction. *Americal Sociological Review 16*, 4 (August 1951), 474–477. http://www.jstor.org/stable/2088278.

[108] Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., and Wooters, C. The ICSI Meeting Corpus. In *Proc 28th International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Hong Kong, China, April 2003), vol. 1, Institute of Electrical and Electronics Engineers, pp. 364–367. doi:10.1109/ICASSP.2003.1198793.

[109] Jefferson, G. Side sequences. In *Studies in Social Interaction*, D. Sudnow, Ed. Free Press, New York NY, USA, 1972. ISBN 978-0029323601.

[110] Jefferson, G. A technique for inviting laughter and its subsequent acceptance declination. In *Everyday language: Studies in ethnomethodology*, G. Psathas, Ed. Irvington, New York NY, USA, 1979, pp. 79–96. ISBN 978-0470266700.

[111] Jefferson, G. An exercise in the transcription and analysis of laughter. In *Handbook of Discourse Analysis; Volume 3: Discourse and Dialogue*, T. van Dijk, Ed. Academic Press, London, UK, 1985, pp. 25–34. ISBN 978-0127120010.

[112] Jefferson, G. On the organization of laughter in talk about troubles. In *Structures of social action: Studies in conversation analysis*, J. Atkinson and J. Heritage, Eds. Cambridge University Press, Cambridge, UK, 1985, pp. 346–369. ISBN 978-0521318624.

[113] Jefferson, G. A note on laughter in 'male-female' interaction. *Discourse Studies 6*, 1 (February 2004), 117–133. doi:10.1177/1461445604039445.

[114] Jefferson, G. Some features of the serial construction of laughter. University of Massechusetts, *(no date)*.

[115] Jefferson, G., Sacks, H., and Schegloff, E. Preliminary notes on the sequential organization of laughter. Tech. rep., *Pragmatics Microfiche*, Department of Linguistics, Cambridge University, Cambridge, UK, 1977.

[116] Jelinek, F. *Statistical Methods for Speech Recognition*. MIT Press, Cambridge MA, USA, 1999. ISBN 978-0262100663.

[117] Jin, Q., Laskowski, K., Schultz, T., and Waibel, A. Speaker segmentation and clustering in meetings. In *Proc. NIST RT-04 Spring Meeting Recognition Evaluation Workshop at the 29th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (Montreal, Canada, May 2004), National Institute of Standards and Technology (NIST).

[118] Kangasharju, H., and Nikko, T. Emotions in organizations: Joint laughter in workplace meetings. In *Journal of Business Communication* (January 2009), vol. 46, pp. 100–119. doi:10.1177/0021943608325750.

[119] Kathol, A., and Tur, G. Extracting question/answer pairs in multi-party meetings. In *Proc. 33rd International Conference Acoustics, Speech and Signal Processing (ICASSP)* (Las Vegas NV, USA, March 2008), Institute of Electrical and Electronics Engineers (IEEE), pp. 5053–5056. doi:10.1109/ICASSP.2008.4518794.

[120] Kennedy, L., and Ellis, D. Laughter detection in meetings. In *Proc Rich Transcription Meeting Recognition Evaluation Workshop* (Montreal, Canada, May 2004), National Institute of Standards and Technology (NIST), pp. 118–121.

[121] Knapp, M., and Hall, J. *Nonverbal Communication in Human Interaction*, 3rd ed. Holt Rinehart & Winston, Fort Worth TX, USA, 1992. ISBN 978-0030625831.

[122] Knox, M., and Mirghafori, N. Automatic laughter detection using neural networks. In *Proc 8th Annual Conference of the International Speech Communication Association (INTERSPEECH)* (Antwerpen, Belgium, August 2007), International Speech Communication Association (ISCA), pp. 2973–2976. http://www.isca-speech.org/archive/interspeech_2007/i07_2973.html.

[123] Knox, M., Morgan, N., and Mirghafori, N. Getting the last laugh: Automatic laughter segmentation in meetings. In *Proc 9th Annual Conference of the International Speech Communication Association (INTERSPEECH)* (Brisbane, Australia, September 2008), International Speech Communication Association (ISCA), pp. 797–800. http://www.isca-speech.org/archive/interspeech_2008/i08_0797.html.

[124] Kolář, J., Shriberg, E., and Liu, Y. Using prosody for automatic sentence segmentation of multi-party meetings. In *Text, Speech and Dialogue, 9th International Conference (TSD)* (Brno, Czech Republic, September 2006), P. Sojka, I. Kopecek, and K. Pala, Eds., vol. 4188 of *Lecture Notes in Computer Science*, Springer, pp. 629–636. doi:10.1007/11846406_79.

[125] Kominek, J., and Kazman, R. Accessing multimedia through concept clustering. Association for Computing Machinery (ACM), pp. 19–26. http://doi.acm.org/10.1145/258549.258567.

[126] Langkilde, I., and Knight, K. Generation that exploits corpus-based statistical knowledge. In *Proc 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL)* (Montréal, Canada, August 1998), Association for Computational Linguistics (ACL), Morgan Kaufmann Publishers, pp. 704–710. doi:10.3115/980451.980963.

[127] Laskowski, K. Modeling vocal interaction for text-independent detection of involvement hotspots in multi-party meetings. In *Proc. 2nd Workshop on Spoken Language Technology (SLT)* (Goa, India, December 2008), Institute of Electrical and Electronics Engineers (IEEE), pp. 81–84.

[128] Laskowski, K. Contrasting emotion-bearing laughter types in multiparticipant vocal activity detection for meetings. In *Proc 34th International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Taipei, Taiwan, April 2009), Institute of Electrical and Electronics Engineers, pp. 4765–4768. doi:10.1109/ICASSP.2009.4960696.

[129] Laskowski, K. Detecting attempts at humor in multiparty meetings. In *Proc. 3rd International Conference on Semantic Computing (ICSC)* (Berkeley CA, USA, September 2009), Institute of Electrical and Electronics Engineers (IEEE), pp. 9–16.

[130] Laskowski, K. Finding emotionally involved speech using implicitly proximity-annotated laughter. In *Proc. 35th International Conference on Acoustics, Speech and Signal Processing* (Dallas TX, USA, March 2009), Institute of Electrical and Electronics Engineers (IEEE), pp. 5226–5229.

[131] Laskowski, K. A frame-synchronous prosodic decoder for text-independent dialog act recognition. In *Proc. 5th International Conference on Speech Prosody* (Chicago IL, USA, May 2010), International Speech Communication Association, p. (page numbers not assigned).

[132] Laskowski, K. Modeling norms of turn-taking in multi-party conversation. In *Proc 48th Annual Meeting of the Association for Computational Linguistics (ACL)* (Uppsala, Sweden, July 2010), Association for Computational Linguistics (ACL), pp. 999–1008. http://www.aclweb.org/anthology/P10-1102.

[133] Laskowski, K., and Burger, S. Development of an annotation scheme for emotionally relevent behavior in multiparty meeting speech. *(poster presented at)* the 2nd International Workshop on Machine Learning for Multimodal Interaction (MLMI), July 2005.

[134] Laskowski, K., and Burger, S. Annotation and analysis of emotionally relevant behavior in the ISL Meeting Corpus. In *Proc 5th International Conference on Language Resources and Evaluation (LREC)* (Genoa, Italy, May 2006), European Language Resources Association (ELRA).

[135] Laskowski, K., and Burger, S. Analysis of the occurrence of laughter in meetings. In *Proc 8th Annual Conference of the International Speech Communication Association (INTERSPEECH)* (Antwerpen, Belgium, August 2007), International Speech Communication Association (ISCA), pp. 1258–1261. http://www.isca-speech.org/archive/interspeech_2007/i07_1258.html.

[136] Laskowski, K., and Burger, S. Annotation guide for laughter in multi-party conversation. Tech. Rep. CMU-LTI-07-013, Carnegie Mellon University, Pittsburgh PA, USA, December 2007.

[137] Laskowski, K., and Burger, S. On the correlation between perceptual and contextual aspects of laughter in meetings. In *Proc Workshop on the Phonetics of Laughter at the 16th International Congress of Phonetic Sciences (ICPhS)* (Saarbrücken, Germany, August 2007), pp. 55–60.

[138] Laskowski, K., Edlund, J., and Heldner, M. An instantaneous vector representation of delta pitch for speaker-change prediction in conversational dialogue systems. In *Proc 33rd International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Las Vegas NV, USA, March 2008), Institute of Electrical and Electronics Engineers, pp. 5041–5044. doi:10.1109/ICASSP.2008.4518791.

[139] Laskowski, K., Edlund, J., and Heldner, M. Learning prosodic sequences using the fundamental frequency variation spectrum. In *Proc. 4th International Conference on Speech Prosody* (Campinas, Brazil, May 2008), International Speech Communication Association (ISCA), pp. 151–154.

[140] Laskowski, K., Edlund, J., and Heldner, M. A single-port non-parametric model of turn-taking in multi-party conversation. In *Proc 36th International Conferences on Acoustics, Speech and Signal Processing (submitted)* (Praha, Czech Republic, 2011), Institute of Electrical and Electronics Engineers (IEEE).

[141] Laskowski, K., Fügen, C., and Schultz, T. Simultaneous multispeaker segmentation for automatic meeting recognition. In *Proc 15th European Signal Processing Conference (EUSIPCO)* (Poznań, Poland, September 2007), European Association for Signal Processing, pp. 1294–1298.

[142] Laskowski, K., Heldner, M., and Edlund, J. Exploring the prosody of floor mechanisms in English using the fundamental frequency variation spectrum. In *Proc. 17th European Signal Processing Conference (EUSIPCO)* (Glasgow, UK, August 2009), European Association for Signal Processing (EURASIP0, pp. 2539–2543.

[143] Laskowski, K., Heldner, M., and Edlund, J. A general-purpose 32 ms prosodic vector for hidden Markov modeling. In *Proc. 10th Annual Conference of the International Speech Communication Association (INTERSPEECH)* (Brighton, UK, September 2009), International Speech Communication Association (ISCA), pp. 2783–2786.

[144] Laskowski, K., Jin, Q., and Schultz, T. Crosscorrelation-based multispeaker speech activity detection. In *Proc 8th International Conference on Spoken Language Processing (INTERSPEECH)* (Jeju Island, South Korea, October 2004), International Speech Communication Association (ISCA), pp. 973–976. http://www.isca-speech.org/archive/interspeech_2004/i04_0973.html.

[145] Laskowski, K., Ostendorf, M., and Schultz, T. Modeling vocal interaction for text-independent classification of conversation type. In *Proc SIGdial* (Antwerpen, Belgium, September 2007), Association for Computational Linguistics (ACL), pp. 194–201.

[146] Laskowski, K., Ostendorf, M., and Schultz, T. Modeling vocal interaction for text-independent participant characterization in multi-party meetings. In *Proc 9th SIGDial Workshop on Discourse and Dialogue* (Columbus OH, USA, June 2008), Association for Computational Linguistics (ACL), pp. 148–155. http://www.aclweb.org/anthology/W/W08/W08-0124.pdf.

[147] Laskowski, K., and Schultz, T. Unsupervised learning of overlapped speech model parameters for multichannel speech activity detection in meetings. In *Proc 31st International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Toulouse, France, May 2006), vol. 1, Institute of Electrical and Electronics Engineers, pp. 993–996. doi:10.1109/ICASSP.2006.1660190.

[148] Laskowski, K., and Schultz, T. A geometric interpretation of non-target-normalized maximum cross-channel correlation for vocal activity detection in meetings. In *Proc Human Language Technologies Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT); Companion Volume, Short Papers* (Rochester NY, USA, April 2007), C. Sidner, T. Schultz, M. Stone, and C. Zhai, Eds., Association for Computational Linguistics (ACL), pp. 89–92. http://www.aclweb.org/anthology/N/N07/N07-2023.pdf.

[149] Laskowski, K., and Schultz, T. Modeling vocal interaction for segmentation in meeting recognition. In *Machine Learning for Multimodal Interaction: 4th International Workshop (MLMI)* (Brno, Czech Republic, June 2007), A. Popescu-Belis, S. Renals, and H. Bourlard, Eds., vol. 4892 of *Lecture Notes in Computer Science*, Springer, pp. 259–270. doi:10.1007/978-3-540-78155-4_23.

[150] Laskowski, K., and Schultz, T. A supervised factorial acoustic model for simultaneous multiparticipant vocal activity detection in close-talk microphone recordings of meetings. Tech. Rep. CMU-LTI-07-017, Carnegie Mellon University, Pittsburgh PA, USA, December 2007.

[151] Laskowski, K., and Schultz, T. Detection of laughter-in-interaction in multichannel close-talk microphone recordings of meetings. In *Machine Learning for Multimodal Interaction: 5th International Workshop (MLMI)* (Utrecht, The Netherlands, September 2008), A. Popescu-Belis and R. Stiefelhagen, Eds., vol. 5237 of *Lecture Notes in Computer Science*, Springer, pp. 149–160. doi:10.1007/978-3-540-85853-9_14.

[152] Laskowski, K., and Schultz, T. Recovering participant identities in meetings from a probabilistic description of vocal interaction. In *Proc 9th Annual Conference of the International Speech Communication Association (INTERSPEECH)* (Brisbane, Australia, September 2008), International Speech Communication Association (ISCA), pp. 82–85. http://www.isca-speech.org/archive/interspeech_2008/i08_0082.html.

[153] Laskowski, K., and Shriberg, E. Modeling other talkers for improved dialog act recognition in meetings. In *Proc 10th Annual Conference of the International Speech Communication Association* (Brighton, UK, September 2009), International Speech Communication Association (ISCA), pp. 2783–2786. http://www.isca-speech.org/archive/interspeech_2009/i09_2783.html.

[154] Laskowski, K., and Shriberg, E. Comparing the contributions of context and prosody in text-independent dialog act recognition. In *Proc. 35th International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Dallas TX, USA, March 2010), Institute of Electrical and Electronics Engineers (IEEE), pp. 5374–5377.

[155] Laskowski, K., Wölfel, M., Heldner, M., and Edlund, J. Computing the fundamental frequency variation spectrum in conversational spoken dialogue systems. In *Proc. 155th Meeting of the Acoustical Society of America* (Paris, France, June 2008), Acoustical Society of America, pp. 3305–3310.

[156] Lazarus, R. *Emotion and Adaptation*. Oxford University Press, New York NY, USA, 1991. ISBN 978-0195092660.

[157] Lee, C., Narayanan, S., and Pieraccini, R. Recognition of negative emotion from the speech signal. In *Proc Workshop on Automatic Speech Recognition and Understanding (ASRU)* (Madonna di Campiglio, Italy, December 2001), Institute of Electrical and Electronics Engineers, pp. 240–243. doi:10.1109/ASRU.2001.1034632.

[158] Levinson, S. *Pragmatics*, 16th printing of 1st ed. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge, UK, 1983. ISBN 978-0521294140.

[159] Licht, M. The staff-resident interaction chronograph: Observational assessment of staff performance. *Journal of Psychopathology and Behavioral Assessment 1*, 3 (October 1979), 185–197. doi:10.1007/BF01321875.

[160] Litman, D., and Forbes-Riley, K. Annotating student emotional states in spoken tutoring dialogues. In *Proc 5th SIGdial Workshop on Discourse and Dialogue* (Cambridge MA, USA, April 2004), Association for Computational Linguistics (ACL), pp. 144–153. http://www.aclweb.org/anthology/W/W04/W04-2326.pdf.

[161] Litman, D., and Forbes-Riley, K. Predicting student emotions in computer-human tutoring dialogues. In *Proc 42nd Annual Meeting of the Association for Computational Linguistics* (Barcelona, Spain, July 2004), Association for Computational Linguistics (ACL), pp. 351–358. doi:10.3115/1218955.1219000.

[162] Litman, D., and Forbes-Riley, K. Predicting student emotions in computer-human tutoring dialogues. In *Proc 42nd Annual Meeting of the Association for Computational Linguistics* (Barcelona, Spain, July 2004), Association for Computational Linguistics (ACL), pp. 351–358. doi:10.3115/1218955.1219000.

[163] Little, W. The existence of persistent states in the brain. *Mathematical Biosciences 19* (February 1974), 101–120. doi:10.1016/0025-5564(74)90031-5.

[164] Little, W., and Shaw, G. A statistical theory of short and long term memory. *Behavioral Biology 14* (June 1975), 115–133. doi:10.1016/S0091-6773(75)90122-4.

[165] Little, W., and Shaw, G. Analytic study of the memory storage capacity of a neural network. *Mathematical Biosciences 39* (June 1978), 281–290. doi:10.1016/0025-5564(78)90058-5.

[166] Lustig, M. Computer analysis of talk-silence patterns in triads. *Communication Quarterly 28*, 4 (Fall 1980), 3–12. doi:10.1080/01463378009369377.

[167] Ma, S.-K. *Statistical Mechanics*. World Scientific Publishing Company, Singapore, 1985. ISBN 978-9971966072.

[168] Matarazzo, J., Wiens, A., Matarazzo, R., and Saslow, G. Speech and silence behavior in clinical psychotherapy and its laboratory correlates. In *Proc Research in Psychotherapy Conference* (Chicago IL, USA, May 1968), J. Shlien, Ed., vol. 3, American Psychological Association, p. 347.

[169] McCowan, I., Bengio, S., Gatica-Perez, D., Lathout, G., Barnard, M., and Zhang, D. Automatic analysis of multimodal group actions in meetings. *IEEE Transactions on Pattern Analysis and Machine Intelligence 27*, 3 (March 2005), 305–317. doi:10.1109/TPAMI.2005.49.

[170] McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., Post, W., Reidsma, D., and Wellner, P. The AMI Meeting Corpus. In *Proc Symposium on Annotating and Measuring Meeting Behavior at the 5th International Conference on Methods and Techniques in Behavioral Research* (Wageningen, The Netherlands, September 2005). *(page numbers not assigned).*

[171] McGrath, J. *Groups: Interaction and Performance*. Prentice-Hall, Inc., Englewood Cliffs NJ, USA, 1984. ISBN 978-0133657005.

[172] Metze, F., Jin, Q., Fügen, C., Laskowski, K., Pan, Y., and Schultz, T. Issues in meeting transcription — The ISL Meeting Transcription System. In *Proc. 8th International Conference on Spoken Language Technology* (Jeju Island, South Korea, October 2004), International Speech Communication Association (ISCA), pp. 1709–1712.

[173] Neiberg, D., Elenius, K., and Laskowski, K. Emotion recognition in spontaneous speech using GMMs. In *Proc 9th International Conference on Spoken Language Processing (INTERSPEECH)* (Pittsburgh PA, USA, September 2006), International Speech Communication Association (ISCA), pp. 809–812. http://www.isca-speech.org/archive/interspeech_2006/i06_1581.html.

[174] Norwine, A., and Murphy, O. Characteristic time intervals in telephonic conversations. *Bell Systems Technical Journal 17*, 4 (April 1938), 281–291. http://bstj.bell-labs.com/BSTJ/images/Vol17/bstj17-2-281.pdf.

[175] Nwokah, E., Hsu, H.-C., Davies, P., and Fogel, A. The integration of laughter and speech in vocal communication: a dynamic systems perspective. *Journal of Speech, Language & Hearing Research 42* (August 1999), 880–894. http://jslhr.asha.org/cgi/reprint/42/4/880.pdf.

[176] op den Akker, H., and Schultz, C. *Exploring features and classifiers for dialogue act segmentation*, vol. 5237 of *Lecture Notes in Computer Science*. Springer, Utrecht, The Netherlands, September 2008, pp. 196–207. doi:10.1007/978-3-540-85853-9_18.

[177] Owren, M., and Bachorowski, J.-A. Reconsidering the evolution of nonlinguistic communication: The case of laughter. *Journal of Nonverbal Behavior 27*, 3 (Fall 2003), 183–199. doi:10.1023/A:1025394015198.

[178] Padilha, E., and Carletta, J. A simulation of small group discussion. In *Proc 6th Workshop on the Semantics and Pragmatics of Dialogue (EDILOG)* (Edinburgh, UK, September 2002), pp. 117–124.

[179] Papp, G., and GauthierDickey, C. Characterizing multiparty voice communication for multiplayer games. In *Proc International Conference on Measurement and Modeling of Computer Systems* (Annapolis MD, USA, 2008), Association for Computing Machinery, pp. 465–466. doi:10.1145/1375457.1375523.

[180] Patterson, M. *Nonverbal Behavior: A Functional Perspective*. Springer Series in Social Psychology. Springer, New York NY, USA, 1983. ISBN 978-0387908465.

[181] Pfau, T., Ellis, D., and Stolcke, A. Multispeaker speech activity detection for the ICSI Meeting Recorder. In *Proc Workshop on Automatic Speech Recognition and Understanding (ASRU)* (Madonna di Campiglio, Italy, December 2001), Institute of Electrical and Electronics Engineers, pp. 107–110. doi:10.1109/ASRU.2001.1034599.

[182] Picard, R. *Affective Computing*. MIT Press, Cambridge MA, USA, 1997. ISBN 978-0262161701.

[183] Polzin, T., and Waibel, A. Pronunciation variations in emotional speech. In *Proc Tutorial and Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition (MPV)* (Rolduc, The Netherlands, May 1998), H. Strik, J. M. Kessens, and M. Wester, Eds., European Speech Communication Association (ESCA), pp. 103–108.

[184] Press, W., Flannery, P., Teukolsky, S., and Vetterling, W. *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed. Cambridge University Press, 1992. ISBN 978-0521431088.

[185] Provine, R. Contagious laughter: Laughter is a sufficient stimulus for laughs and smiles. *Bulletin of the Psychonomic Society 30*, 1 (1992), 1–4.

[186] Raaijmakers, S., Truong, K., and Wilson, T. Multimodal subjectivity analysis of multiparty conversation. In *Proc Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Honolulu HI, USA, October 2008), Association for Computational Linguistics (ACL), pp. 466–474. http://www.aclweb.org/anthology/D/D08/D08-1049.pdf.

[187] Rabiner, L. A tutorial on hidden markov models and selected applications in speech recognition. *Proc Institute of Electrical and Electronics Engineers 77*, 2 (February 1989), 257–286. doi:10.1109/5.18626.

[188] Raux, A., and Eskenazi, M. Finite state turn taking model for spoken dialog systems. In *Proc Human Language Technologies Conference of the North American Chapter of the Association for Computational Linguistics* (Boulder CO, USA, June 2009), Association for Computational Linguistics (ACL), pp. 629–637. http://www.aclweb.org/anthology/N/N09/N09-1071.pdf.

[189] Reidsma, D., Heylen, D., and Ordelman, R. Annotating emotions in meetings. In *Proc 5th International Conference on Language Resources and Evaluation (2006)* (Genoa, Italy, May 2006), European Language Resources Association (ELRA).

[190] Rienks, R., and Heylen, D. Dominance detection in meetings using easily obtainable features. In *Machine Learning for Multimodal Interaction, 2nd International Workshop (MLMI)* (Edinburgh, UK, July 2005), S. Renals and S. Bengio, Eds., vol. 3869 of *Lecture Notes in Computer Science*, Springer, pp. 76–86. doi:10.1007/11677482_7.

[191] Rienks, R., Zhang, D., Gatica-Perez, D., and Post, W. Detection and application of influence rankings in small-group meetings. In *Proc 8th International Conference on Multimodal Interfaces (ICMI)* (Banff, Canada, November 2006), Association for Computing Machinery, pp. 257–264. doi:10.1145/1180995.1181047.

[192] Ries, K. *Accessing Spoken Interaction through Dialogue Processing*. PhD thesis, University of Karlsruhe, Karlsruhe, Germany, December 2001.

[193] Russell, J., Bachorowski, J.-A., and Fernandez-Dols, J.-M. Facial and vocal expressions of emotion. *Annual Review of Psychology 54* (February 2003), 329–349. doi:10.1146/annurev.psych.54.101601.145102.

[194] Sacks, H., Schegloff, E., and Jefferson, G. A simplest semantics for the organization of turn-taking for conversation. *Language 50*, 4 (December 1974), 696–735. http://www.jstor.org/stable/412243.

[195] Salamin, H., Favre, S., and Vinciarelli, A. Automatic role recognition in multiparty recordings: Using social affiliation networks for feature extraction. *IEEE Transactions on Multimedia 11*, 7 (November 2009), 1373–1380. doi:10.1109/TMM.2009.2030740.

[196] Saslow, G., Matarazzo, J., and Guze, S. The stability of interaction chronograph patterns in psychiatric interviews. *Journal of Consulting Psychology 19*, 6 (December 1955), 417–430. doi:10.1037/h0047036.

[197] Schegloff, E. Overlapping talk and the organization of turn-taking for conversation. *Language in Society 29*, 1 (March 2000), 1–63. http://www.jstor.org/stable/4168983.

[198] Schegloff, E. *Sequence Organization in Interaction: A Primer in Conversation Analysis*, vol. 1. Cambridge University Press, Cambridge, UK, 2007. ISBN 978-0521532792.

[199] Schiel, F., Burger, S., Geumann, A., and Weilhammer, K. The partitur format at BAS. In *Proc 1st International Conference on Language Resources and Evaluation (LREC)* (Granada, Spain, May 1998), vol. 2, European Language Resources Association (ELRA), pp. 1295–1301.

[200] Schlosberg, H. Three dimensions of emotions. *Psychological Review 61*, 2 (March 1954), 81–88. PMID 13155714.

[201] Sellen, A. Speech patterns in video-mediated conversations. In *Proc SIGCHI Conference on Human Factors in Computing Systems* (Monterey CA, USA, 1992), Association for Computing Machinery, pp. 49–59. doi:10.1145/142750.142756.

[202] Shriberg, E., Dhillon, R., Bhagat, S., Ang, J., and Carvey, H. The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. In *Proc 5th SIGdial Workshop on Discourse and Dialogue* (Cambridge MA, USA, April 2004), M. Strube and C. Sidner, Eds., Association for Computational Linguistics (ACL), pp. 97–100. http://www.aclweb.org/anthology/W/W04/W04-2319.pdf.

[203] Shriberg, E., Stolcke, A., and Baron, D. Observations on overlap: Findings and implicatins for automatic processing of multi-party conversation. In *Proc 7th European Conference on Speech Communication and Technology (EUROSPEECH)* (Aalborg, Denmark, September 2001), P. Dalsgaard, B. Lindberg, H. Benner, and Z. Tan, Eds., vol. 2, International Speech Communication Association (ISCA), pp. 1359–1362. http://www.isca-speech.org/archive/eurospeech_2001/e01_1359.html.

[204] Siegman, A., and Feldstein, S., Eds. *Multichannel Integrations of Nonverbal Behavior*. Lawrence Erlbaum Associates, Inc., Hillsdale NJ, USA, 1985. ISBN 978-0898595666.

[205] Somasundaran, S., Ruppenhofer, J., and Wiebe, J. Detecting arguing and sentiment in meetings. In *Proc 8th SIGdial Workshop on Discourse and Dialogue* (Antwerpen, Belgium, September 2007), Association for Computational Linguistics (ACL), pp. 26–34.

[206] Stasser, G., and Taylor, L. Speaking turns in face-to-face discussions. *Journal of Personality and Social Psychology 60*, 5 (May 1991), 675–684.

[207] Stolcke, A., Anguera, X., Boakye, K., Çetin, O., Grézl, F., Janin, A., Mandal, A., Peskin, B., Wooters, C., and Zheng, J. Further progress in meeting recognition: The ICSI-SRI Spring 2005 speech-to-text evaluation system. In *Machine Learning for Multimodal Interaction: 2nd International Workshop (MLMI)* (Edinburgh, UK, July 2005), S. Renals and S. Bengio, Eds., vol. 3869 of *Lecture Notes in Computer Science*, Springer, pp. 463–475. doi:10.1007/11677482_39.

[208] Stolcke, A., and Shriberg, E. Automatic linguistic segmentation of conversational speech. In *Proc 4th International Conference on Spoken Language Processing (ICSLP)* (Philadelphia PA, USA, Oct 1996), vol. 2, International Speech Communication Association (ISCA), pp. 1005–1008. doi:10.1109/ICSLP.1996.607773.

[209] Stolcke, A., Shriberg, E., Bates, R., Ostendorf, M., Hakkani, D., Plauche, M., Tür, G., and Liu, Y. Automatic detection of sentence boundaries and disfluencies based on recognized words. In *Proc 5th International Conference on Spoken Language Processing (ICSLP)* (Syndey, Australia, November 1998), International Speech Communication Association (ISCA), pp. 2247–2250. http://www.isca-speech.org/archive/icslp_1998/i98_0059.html.

[210] Tannen, D. *Gender & Discourse*. Oxford University Press, 1996. ISBN 978-0195101249.

[211] Tardelli, J., Gatewood, P., Kreamer, E., and Follette, P. L. The benefits of multi-speaker conferencing and the design of conference bridge control algorithms. In *Proc 18th International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Minneapolis MN, USA, April 1993), vol. 2, Institute of Electrical and Electronics Engineers, pp. 435–438. doi:10.1109/ICASSP.1993.319333.

[212] Thomas, D., Loomis, A., and Arrington, R. *Observational Studies of Social Behavior*, vol. I: Social Behavior Patterns. Institute of Human Relations, Yale University, New Haven CT, USA, 1933.

[213] Truong, K., and van Leeuwen, D. Automatic detection of laughter. In *Proc 9th European Conference on Speech Communication and Technology (INTERSPEECH)* (Lisboa, Portugal, September 2005), International Speech Communication Association (ISCA), pp. 485–488. http://www.isca-speech.org/archive/interspeech_2005/i05_0485.html.

[214] Truong, K., and van Leeuwen, D. Automatic discrimination between laughter and speech. *Speech Communication 49*, 2 (February 2007), 144–158. doi:10.1016/j.specom.2007.01.001.

[215] Truong, K., and van Leeuwen, D. Evaluating automatic laughter segmentation in meetings using acoustic and acoustics-phonetic features. In *Proc Proc Workshop on the Phonetics of Laughter at the 16th International Congress of Phonetic Sciences (ICPhS)* (Saarbrücken, Germany, August 2007), pp. 49–53.

[216] van Leeuwen, D., and Huijbregts, M. The AMI Speaker Diarization System for NIST RT06s meeting data. In *Machine Learning for Multimodal Interaction, 3rd International Workshop (MLMI)* (Bethesda MD, USA, May 2006), S. Renals, S. Bengio, and J. Fiscus, Eds., vol. 4299 of *Lecture Notes in Computer Science*, Springer, pp. 371–384. doi:10.1007/11965152_33.

[217] Venkataraman, A., Liu, Y., Shriberg, E., and Stolcke, A. Does active learning help automatic dialog act tagging in meeting data? In *Proc 9th European Conference on Speech Communication and Technology (INTERSPEECH)* (Lisboa, Portugal, September 2005), International Speech Communication Association (ISCA), pp. 2777–2780. http://www.isca-speech.org/archive/interspeech_2005/i05_2777.html.

[218] Vinciarelli, A. Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling. *IEEE Transactions on Multimedia 9*, 6 (October 2007), 1215–1226. doi:10.1109/TMM.2007.902882.

[219] Wahlster, W., Ed. *Verbmobil: Foundations of Speech-to-Speech Translation*. Artificial Intelligence. Springer, Berlin, Germany, 2000. ISBN 978-3540677836.

[220] Wasserman, L. *All of Statistics: A Concise Course in Statistical Inference*. Springer Texts in Statistics. Springer, New York NY, USA, 2004. ISBN 978-0387402727.

[221] Weilhammer, K., Reichel, U., and Schiel, F. Multi-tier annotations in the Verbmobil Corpus. In *Proc 3rd International Conference on Language Resources and Evaluation (LREC)* (Las Palmas, Canary Islands, May 2002), European Language Resources Association (ELRA), pp. 912–917.

[222] Weinstein, R., Feldstein, S., and Jaffe, J. The interaction of vocal context and lexical predictability. *Language and Speech 8*, 1 (January 1965), 56–67. doi:10.1177/002383096500800105.

[223] Wrede, B., Bhagat, S., Dhillon, R., and Shriberg, E. Meeting Recorder project: Hot spot labeling guide. Tech. Rep. TR-05-004, International Computer Science Institute, Berkeley CA, USA, May 2005.

[224] Wrede, B., and Shriberg, E. The relationship between dialogue acts and hot spots in meetings. In *Proc Workshop on Automatic Speech Recognition and Understanding (ASRU)* (St. Thomas, US Virgin Islands, November 2003), Institute of Electrical and Electronics Engineers, pp. 180–185. doi:10.1109/ASRU.2003.1318425.

[225] Wrede, B., and Shriberg, E. Spotting "hot spots" in meetings: Human judgements and prosodic cues. In *Proc 8th European Conference on Speech Communication and Technology (EUROSPEECH)* (Geneva, Switzerland, September 2003), International Speech Communication Association (ISCA). http://www.isca-speech.org/archive/eurospeech_2003/e03_2805.html.

[226] Wrede, B., and Shriberg, E. Reliability analysis for hot spot annotations in the MRDA Corpus. (*internal document*), International Computer Science Institute, Berkeley CA, USA, April 2005.

[227] Wrigley, S., Brown, G., Wan, V., and Renals, S. Speech and crosstalk detection in multichannel audio. *IEEE Transactions on Speech and Audio Processing 13*, 1 (January 2005), 84–91. doi:10.1109/TSA.2004.838531.

[228] Xie, S., Liu, Y., and Lin, H. Evaluating the effectiveness of features and sampling in extractive meeting summarization. In *Proc. 2nd Workshop on Spoken Language Technology* (Goa, India, December 2008), Institute of Electrical and Electronics Engineeers (IEEE), pp. 157–160. doi:10.1109/SLT.2008.4777864.

[229] Yang, F., Tur, G., and Shriberg, E. Exploiting dialogue act tagging and prosodic information for action item identification. In *Proc. 33rd International Conference Acoustics, Speech and Signal Processing (ICASSP)* (Las Vegas NV, USA, March 2008), Institute of Electrical and Electronics Engineers (IEEE), pp. 4941–4944. doi:10.1109/ICASSP.2008.4518766.

[230] Zancanaro, M., Lepri, B., and Pianesi, F. Automatic detection of group functional roles in face to face interactions. In *Proc 8th International Conference on Multimodal Interfaces (ICMI)* (Banff, Canada, November 2006), Association for Computing Machinery, pp. 28–34. doi:10.1145/1180995.1181003.

[231] Zimmermann, M. Joint segmentation and classification of dialog acts using conditional random fields. In *Proc 10th Annual Conference of the International Speech Communication Association (INTERSPEECH)* (Brighton, UK, September 2009), International Speech Communication Association (ISCA), pp. 864–867. http://www.isca-speech.org/archive/interspeech_2009/i09_0864.html.

[232] Zimmermann, M., Liu, Y., Shriberg, E., and Stolcke, A. A* based joint segmentation and classification of dialog acts in multiparty meetings. In *Proc Workshop on Automatic Speech Recognition and Understanding (ASRU)* (Cancun, Mexico, November 2005), Institute of Electrical and Electronics Engineers, pp. 215–219. doi:10.1109/ASRU.2005.1566537.

[233] Zimmermann, M., Liu, Y., Shriberg, E., and Stolcke, A. Toward joint segmentation and classification of dialog acts in multiparty meetings. In *Machine Learning for Multimodal Interaction: 2nd International Workshop (MLMI)* (Edinburgh, UK, July 2005), S. Renals and S. Bengio, Eds., vol. 3869 of *Lecture Notes in Computer Science*, Springer, pp. 187–193. doi:10.1007/11677482_16.