Detecting Off-Task Speech

Wei Chen

CMU-LTI-12-003

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213
www.lti.cs.cmu.edu

Thesis Committee:

Jack Mostow, Ph.D., Research Professor, Carnegie Mellon University
Alan Black, Ph.D., Associate Professor, Carnegie Mellon University
Florian Metze, Ph.D., Assistant Research Professor, Carnegie Mellon University
Gregory Aist, Ph.D., Assistant Professor, Iowa State University
Diane Litman, Ph.D., Professor, University of Pittsburgh

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
In Language and Information Technologies

Abstract:

Off-task speech is speech that strays away from an intended task. It occurs in many dialog applications, such as intelligent tutors, virtual games, health communication systems and human-robot cooperation. Off-task speech input to computers presents both challenges and opportunities for such dialog systems. On the one hand, off-task speech contains informal conversational style and potentially unbounded scope that hamper accurate speech recognition. On the other hand, an automated agent capable of detecting off-task speech could track users' attention and thereby maintain the intended conversation by bringing a user back on task; also, knowledge of where off-task speech events are likely to occur can help the analysis of automatic speech recognition (ASR) errors. Related work has been done in confidence measures for dialog systems and detecting out-of-domain utterances. However, there is a lack of systematic study on the type of off-task speech being detected and generality of features capturing off-task speech. In addition, we know of no published research on detecting off-task speech in children's interactions with an automated agent. The goal of this research is to fill in these blanks to provide a systematic study of off-task speech, with an emphasis on child-machine interactions.

To characterize off-task speech quantitatively, we used acoustic features to capture its speaking style; we used lexical features to capture its linguistic content; and we used contextual features to capture the relation of off-task speech to nearby utterances. Using these features, we trained an off-task speech detector that yielded 87% detection rate at a cost of 10% false positives on children's oral reading. Furthermore, we studied the generality of these types of features by detecting off-task speech in data from four tutorial tasks ranging from oral reading to prompted free-form responses. In addition, we examined how the features help detect adults'

off-task speech in data from the CMU Let's Go bus information system. We show that lexical features detect more task-related off-task speech such as complaints about the system, whereas acoustic features detect more unintelligible speech and non-speech events such as mumbling and humming. Moreover, acoustic features tend to be more robust than lexical features when switching domains. Finally, we demonstrate how off-task speech detection can improve the performance on application-relevant metrics such as predicting fluency test scores in oral reading and understanding utterances in the CMU Let's Go bus information system.

Acknowledgements

This thesis could not have been completed without help from many people.

First and foremost, I would like to thank my advisor Prof. Jack Mostow, who provided a tremendous amount of support to me throughout the four and half years of my PhD study. It was really fortunate for me to have an advisor like Jack, who has an excellent grasp of both the big picture and fine details of research ideas. He has been and will continue to be a role model to me in my academic career. Prof. Gregory Aist, with his sharp thinking, guided me through my PhD study. Prof. Alan Black, Prof. Florian Metze, and Prof. Diane Litman helped me make this dissertation better by raising many important and thought-provoking questions. Prof. Scott Fahlman, who was my advisor during the first year of my Masters program, has continuously provided me with mental support and insightful discussions on my thesis and research career.

Most of my time during my PhD study was spent at Project LISTEN at Carnegie Mellon University. In Project LISTEN, I had the great opportunity to work with some really creative and supportive colleagues: Juliet Bey, Donna Gates, Octavio Juarez-Espinosa, Martin Kantorzyk, Lynnetta Miller, Joe Valeri, and Anders Weinstein.

I enjoyed sharing ideas and research tools with my fellow students in Project LISTEN, especially Liu Liu, Minh Duong, Yanbo Xu, Weisi Duan, Yuanpeng Li, José González-Brenes, Morten Højfeldt Rasmussen, Sunayana Sitaram, Hyeju Jang, Mdahaduzzaman Munna, and Lucas Tan.

Finally, I give my deepest thanks to my loving and lovely family: my husband Shanqing, my ten-month-old son David, my parents and parent-in-laws (especially my mom Lu Xin and mom-in-law Peihua Qiu, who came from China to the US to give me and Shanqing help raising David while both of us were in the final stages of our PhD studies). Without the tremendous support and sacrifice on their part, none of this would have been possible.

Contents

List of Figures

List of Tables

1. Introduction

Off-task speech is speech that deviates from an intended task. As a prevalent speech event, off-task speech has been acknowledged in many dialog systems, such as health communication systems (Bickmore & Giorgino, 2004), human-robot cooperation (Dowding, Alena, Clancey, Sierhuis, & Graham, 2006), tutoring systems (Chen, Mostow, & Aist, 2010), and virtual games (Kluwer, Adolphs, Xu, Uszkoreit, & Cheng, 2010). Off-task speech is an important and challenging issue in dialog. Traum and Gendve (1996) pointed out that a successful dialog participant should "respond to what is said," including off-task speech, and that an example response to off-task speech is to bring the subject back to topic. Bickmore and Giorgino (2004) described that the "potentially unbounded scope" and "informal conversational" style of off-task speech make it difficult for speech recognizers to perform accurately. Moreover, off-task speech often contains out-of-vocabulary (OOV) words and different speaking styles (e.g., spontaneous rather than planned speech), both of which are known sources of speech recognition errors (Chase, 1997a; Scharenborg, 2007). For intelligent tutoring systems, the capability to detect and monitor off-task behavior can contribute to better understanding and tracking the attention and progress of students (R.S. Baker, 2007).

The goal of this dissertation is to conduct a systematic study of off-task speech and of the roles played by acoustic and lexical features in the automatic detection of off-task speech in a variety of tasks. The principal tasks that we study are children's verbal interactions with a reading tutor during reading comprehension activities, including oral reading, self-questioning, defining words, and thinking aloud. In addition, we test if the conclusions drawn from children's speech generalize to the Let's Go corpus, which consists of adults' spoken interactions by

telephone with a local bus information system. Both children's reading comprehension and adults' phone calls for bus information provide naturally occurring off-task speech rather than elicited or acted off-task speech.

The rest of this chapter is organized as the following. We introduce a taxonomy of off-task speech in Section 1.1. We then summarize the thesis statement in Section 1.2. Finally we describe research platforms in Section 1.3.

1.1 A taxonomy of off-task speech

For some kinds of off-task speech the definition is task-dependent, and for other kinds the definition is general across many tasks. From empirical study of off-task utterances collected during children's interactions with a reading tutor, we develop a taxonomy of off-task speech. Despite its limited scope and size, the corpus still reveals some of the variation in "off-task-ness." The taxonomy shows that only some of the off-task speech depends on the task; other types of off-task speech commonly occur in many tasks and domains. The examples shown below come from real users. The number after each category of off-task speech denotes the frequency of off-task utterances in that category. We counted these frequencies from an annotated corpus of 410 utterances[1] totaling 1,729 words spoken by 20 children ages 7-11 during vocabulary activities with the Reading Tutor (Jack Mostow & Beck, 2007) in 2010. Two annotators independently labeled 96 utterances with three possible tags: null responses, general off-task speech, and task-related off-task speech, and the inter-rater reliability computed as the intraclass correlation (Shrout & Fleiss, 1979) is 0.66. We assigned each utterance to a unique category, so the percentages add to 100%.

- **Null responses (non-speech events) [36%]**:

[1] We use the term "utterance" loosely. Each utterance corresponds to a recording, which may or may not contain speech.

- **Silence [24%]**: a type of off-task behavior (Gobel, 2008).
- **Noises from non-vocal sources made by the user [7%]** (e.g., clicking and tapping sound).
- **Singing and humming [5%]**

- **General off-task speech** (i.e., off-task speech that is independent of the task domain and can occur in many tasks; includes many instances where the intended hearer is not the system (see Section 1.3.1.7 for related work on addressee identification)) **[44%]**:

  Examples:

  Going to bathroom:

  Mrs Matheson I have to go bathroom bad.

  Chatting with others:

  Come on come on Brianna we have to go.

  Talking to oneself:

  I'm tired.

  Complaining about the computer, microphone, and other system parts:

  This computer's all screwed up.

  Going back:

  No I want to go back good.

  Checking current status of progress:

  Sixty minutes today.

  …

- **Task-related off-task speech** (i.e., thinking about the task but not directly addressing the task; not counted as off-task in some education literature (e.g., Gobel, 2008)) **[20%]**

Examples:

Questioning in oral reading:

[Sentence text: *They snuggle together to block the wind.*]

Child: *How do they get that fat?*

Repeating prompt:

Hey it says "what are you wondering about!"

Do not know answer

I forgot this, ummm...

…

Notice that we tagged all silences as off-task, although silence could, and may often mean that the user is thinking. Strictly speaking silence does not necessarily indicate off-task behavior. However, a tutor can behave similarly to both silence and off-task speech. For example, it can wait patiently both when a user is thinking silently and when a user is talking to his classmate. For another example, the tutor can provide hints both when a user is thinking silently and when a user is expressing frustration.

1.2 Goals of the thesis

The goal of this dissertation is to test hypotheses to address the following research questions:

a. How to detect children's off-task speech?

a.1. What features characterize off-task speech?

a.2. How to detect off-task speech when training labels are difficult to acquire?

b. Do the features generalize across tasks?

c. Does off-task speech detection help?

1.3 Research platforms

The principal research platform for detecting off-task speech is a reading tutor that listens as children read aloud and respond to vocabulary and reading comprehension prompts. We train classifiers and study features of off-task speech using data collected by this tutor. As a test of generality, we use a corpus collected by the Let's Go bus system to evaluate specific features of off-task speech and the detection approach in general adult speech for a different task.

1.3.1 Project LISTEN's Reading Tutor

Project LISTEN's Reading Tutor (Jack Mostow & Beck, 2007) is a computer program that listens to children read aloud, provides help when necessary (J. Mostow & Aist, 1999), and teaches them vocabulary and reading comprehension strategies (J. Mostow et al., 2009; Jack Mostow et al., 2010). All these activities involve children's verbal interactions with the tutor. The Reading Tutor displays stories on a computer screen, adding one sentence at a time. A child reads the displayed sentence out loud. The Reading Tutor tracks the child's position in the sentence. The core technology that enables tracking of children's oral reading is the SphinxII speech recognizer, using a finite state grammar automatically generated from the story sentence which penalizes repeating and skipping of words.

The vocabulary activities in the Reading Tutor engage a child in explaining the meaning of a word and comparing semantic similarities between pairs of words (J. Mostow et al., 2010). These vocabulary activities are introduced outside a story.

In contrast, activities for reading comprehension strategies entirely sit in the context of stories in the Reading Tutor specifically augmented to teach reading comprehension strategies.

The main reading comprehension strategies taught include self-questioning, summarizing, visualizing, and vocabulary activities. For example, the self-questioning activity follows a five-step model (Duke & Pearson, 2002) that gradually transfers responsibility of strategy use from the tutor to the child: explain the strategy, demonstrate the strategy, scaffold the use of the strategy, prompt the child to ask questions, and finally let the child ask questions freely. Figure 1.1 shows a screenshot of the prompting step in the self-questioning activity. In addition to the reading comprehension strategies, there is also "think-aloud," which is an outcome measurement activity.

Depending on the nature of the task, children's off-task speech can be a serious problem for the Reading Tutor. For some tasks that prompt free-form spoken responses, the percentage of off-task speech can be higher than 30%. Therefore, detection children's off-task speech in the Reading Tutor is an important issue for automatic understanding of children's responses.

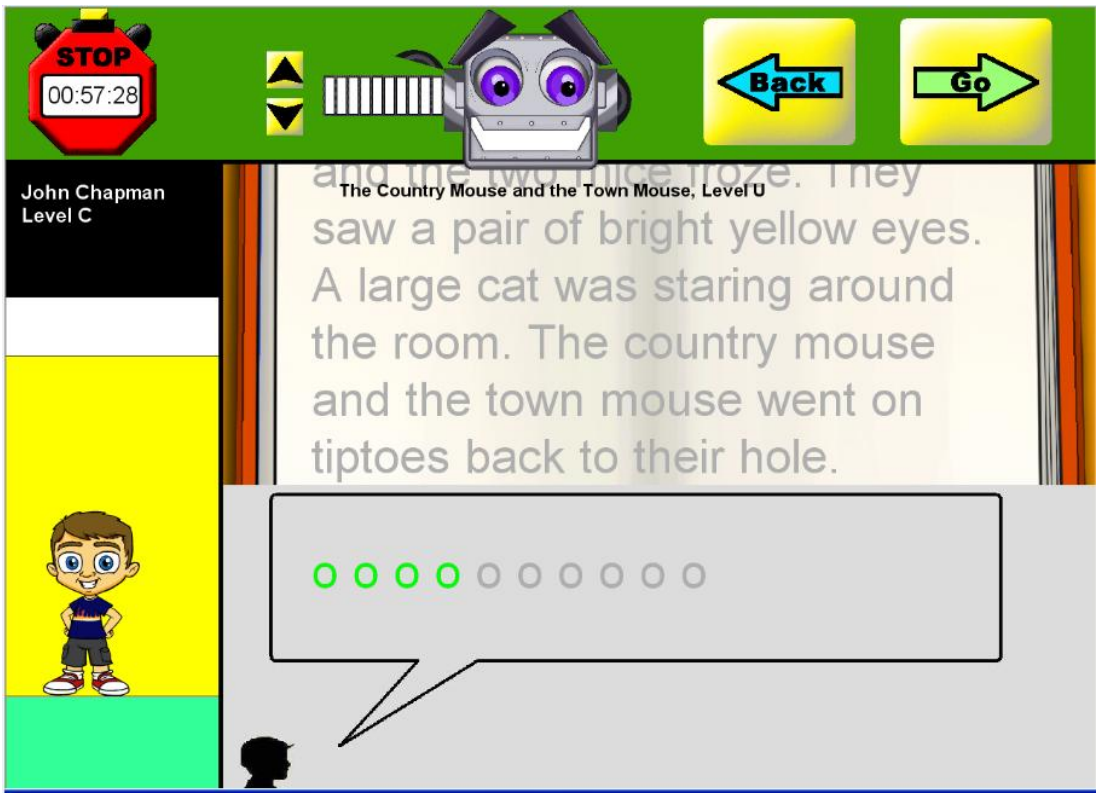

Figure 1.1 The Reading Tutor prompts a child to ask a question about the story: "What are you wondering about now?" The spoken response is recorded and stored into a back-end database. Green circles reflect the amount of speech detected.

1.3.2 The CMU Let's Go Bus information system

The CMU Let's Go Bus information system (henceforth called the Let's Go system) is a spoken dialog system that provides Pittsburgh local bus information (e.g., bus schedule and route information) by telephones (Raux, Langner, Black, & Eskenazi, 2005). When a person calls the system, it automatically identifies the information that the caller needs and presents the information to the caller in speech. The system consists of five parts: a speech recognizer to transcribe the caller's speech, a dialog manager to identify information requested and decide the next action of the system, a back-end database to retrieve bus information, a language generator to construct the system response, and a speech synthesizer to turn generated language into speech.

During each call, the system records speech from the caller and logs the ASR result, parser output, and dialog context (e.g., the order of system prompts and user utterances). Many calls recorded in 2003 have been made publicly available as a downloadable corpus (Raux, Langner, Black, & Eskenazi, 2003). We use this resource to train and test detection of off-task speech in adult speech.

2. Problems related to off-task speech detection

This chapter compares off-task speech detection with off-task behavior detection, dialog act classification, out-of-domain utterance detection, out-of-vocabulary detection, confidence measures, emotion detection, speaking style classification, and addressee detection.

2.1 Off-task behavior detection

Off-task behavior is an important phenomenon under study in educational psychology (e.g., Abramowitz, O'Leary, & Rosen, 1987) and intelligent tutoring (e.g., R. S. Baker, Corbett, Koedinger, & Wagner, 2004). In particular, Baker, D'Mello et al. (2010) correlated off-task behavior with boredom and frustration; Baker, Corbett et al. (2004) found off-task behavior an indicator of poor learning; Beal, Mitra et al. (2007) suggested that detecting off-task behavior may help improve effectiveness of an intelligent tutoring system.

Off-task speech is a spoken form of off-task behavior. Besides off-task speech, off-task behaviors (especially during interaction with a computer) also include random mouse movement, typing unrelated to the task, as well as any non-speech behavior that does not involve the software, such as surfing the web (R. S. Baker, et al., 2004).

Detection of off-task behavior appeared in the literature only recently. Baker et al. (2004) introduced a key approach to off-task behavior detection, which used rich information about students and tutoring sessions to detect off-task behaviors. Recently, Cetintas et al. (2010) found that mouse movement features add significant improvement to a detector using only time and student performance features.

2.2 Dialog act classification

Early approaches to dialog act classification relied on theories of planning and beliefs (Cohen & Perrault, 1979; Perrault & Allen, 1980). As large corpora with hand annotations became available, many machine learning methods have been applied to the problem (e.g., Stolcke et al., 2000). Recently, having noticed that the choice of features matters more than the choice of machine learning methods, Webb (2010) developed methods to extract cue phrases for dialog act classification automatically from training corpora.

The bulk of research on dialog act classification concentrates on human dialogs, rather than human-computer conversations (Webb, 2010). There are many notable differences between the two types of interactions. For example, Booth (1989) stated that human dialogs use truly natural languages (i.e., error-tolerant, sometimes fragmental and ungrammatical, involving rich background knowledge and common sense), whereas people choose abstract and abbreviated language to communicate with a machine.

Off-task speech can be a type of dialog act. Even though many dialog act classification programs neglect off-task speech (e.g., Stolcke, et al., 2000), a comprehensive taxonomy of dialog acts often contains off-task speech as one category (Alexandersson et al., 1998; Orkin & Roy, 2010). Therefore some dialog act classification provides methods to detect off-task speech. However, since there are many ways to design a taxonomy of dialog acts, not every taxonomy has one class for off-task speech, nor does every taxonomy provide a clean-cut criterion for distinguishing on- and off-task speech. That is, some taxonomies mix on- and off-task speech in their classes. For example, the "explain" dialog act in the MAPTASK corpus (Anderson, Bader, Bard, Boyle, & Doherty, 1991) can include both a person explaining the current Terrain (on-task, e.g., "I've got farmed land here") and a person explaining his feeling about the task after the task has been completed (off-task, e.g., "It's probably nothing like it, but heh ... and um ..." "I don't

think I banged in ... I think it was sounded alright"). Moreover, off-task speech does not have to occur in dialogs. For example, children's oral reading with the Reading Tutor is not a dialog in a conventional sense, so off-task speech in this task is not a conventional dialog act.

2.3 Out-of-domain utterance detection

Out-of-domain (OOD) utterance detection sits in the context of dialog systems. It deals with the problem of users requesting information that does not exist in a back-end database. Lane et al. introduced OOD utterance detection in 2004 (Ian Lane, Kawahara, Matsui, & Nakamura, 2004), where they applied topic classification and a linear discriminant model to verify in-domain utterances. They later extended the approach by incorporating features of dialog context (I. Lane, Kawahara, Matsui, & Nakamura, 2007).

Unlike off-task speech detection that detects disengagement in speech, OOD detection finds utterances that request information which does not belong to any of the pre-defined topics (I. Lane, et al., 2007). Some OOD utterances are on-task. For example, a user may ask a travel system "When is the next train from London to Aldeburgh?", where Aldeburgh has no train service. Even though this query is out of scope (thus OOD), the utterance asks for travel information and hence is on-task. Furthermore, existing work on OOD has not explicitly addressed off-task speech phenomena such as a user talking to himself, speaking to a third party, uttering nonsense, and even humming. OOD methods focus mostly on word or n-gram occurrences for topic modeling, while off-task speech detection uses both acoustic and lexical features.

2.4 Out-of-vocabulary detection

Out-of-vocabulary (OOV) detection finds out spoken words not included in the ASR vocabulary, which cannot be correctly processed by the ASR. Since off-task speech is often unexpected, words in off-task speech are usually hard to predict as well. Although OOV words occur in off-task speech, they are not robust indicators of off-task speech. For example, misreading can cause OOV words but it is not off-task for oral reading.

The methods we use to detect off-task speech are partially motivated by methods used for OOV detection. There are two major approaches used for OOV detection. The first one uses confidence measures, which is computed either from a feature vector that aggregates a large number of features extracted from the ASR process and dialog contexts (e.g., White, Droppo, Acero, & Odell, 2007) or by the posterior probability of a word given the acoustic observation and a time interval for uttering the word (Kemp & Schaaf, 1997; Wessel, Schlüter, Macherey, & Ney, 2001). The second approach explicitly models OOV words by filler or garbage models (Hazen & Bazzi, 2001; Schaaf, 2001), which use a phone loop to absorb unknown words.

2.5 Confidence measures

Confidence measures are used for finding ASR errors, detecting OOV words, and detecting OOD utterances (San-Segundo, Pellom, Hacioglu, & Ward, 2001). As we argued previously, OOD is not necessarily off-task. Therefore confidence measures do not provide a sufficient tool for detecting and exploiting off-task speech. That said, confidence measures can detect *some* off-task speech when they use a good task language model. But they will still fail at accepting on-task utterances that contain errors (e.g., misreading and speech repairs).

Research in confidence measures originated from detecting OOV words in large vocabulary speech recognition (e.g., Asadi, Schwartz, & Makhoul, 1990). Since then, three common approaches appeared in the literature (Jiang, 2005). The first one computes confidence

scores as word posterior probabilities (e.g., Kemp & Schaaf, 1997). The second approach combines features generated from lattices, language models, posterior probabilities, and many other resources to help compute confidence scores (e.g., Chase, 1997b). The third approach formulates confidence measures as statistical hypothesis tests, especially for utterance verification (e.g., Rose, Juang, & Lee, 1995).

2.6 Emotion detection and speaking style classification

Off-task behavior has been found to correlate with emotion (Sabourin, Rowe, Mott, & Lester, 2011). The rich literature on detecting emotion in speech provides foundations for methods to detect off-task speech. Emotion detection in speech has a long history. Early research on emotion detection started in two separate aspects: words (Austin, 1962) and prosodic cues (Lieberman & Michaels, 1962; Williams & Stevens, 1972). More recently, the two types of emotional cues (i.e., words spoken and speaking style expressed in acoustic signals) have been combined to detect emotions in human-machine interactions (Cowie et al., 2001). To find out which features are salient for detecting emotions in speech, Batliner et al. (2010) extracted more than 4000 features characterizing both linguistic and acoustic aspects of emotions. They found that the performances of linguistic and acoustic features are comparable, and combining them together brings improvements.

Speaking style classification focuses only on the differences in speech acoustics between different speaking styles such as spontaneous speech, infant-directed speech and dictation (Shinozaki, Ostendorf, & Atlas, 2009). Among these speaking styles, we are most interested in the differences between spontaneous off-task speech and planned on-task speech.

2.7 Addressee identification

Off-task speech includes speech addressed to entities other than the system. Therefore addressee identification (e.g., Jovanovic & Akker, 2004) can help identify some off-task speech, but may not distinguish off-task speech addressed to the system (e.g., complaints about the system, asking to return to a previous step) from on-task speech also addressed to the system.

Traum (2003) described the earliest approach to identify addressees, where he hand-coded decision rules to capture utterance and conversational context features. One example rule says that "the current addressee is the previous speaker." Since then, other people added gaze features to utterance and conversation contextual features in a statistical classifier such as the Naïve Bayes (e.g., Jovanovic & Akker, 2006). Recent trend in addressee identification has seen more attention in non-verbal cues such as gaze and head movement (e.g., Huang, Baba, & Nakano, 2011).

2.8 Features used in related problems

There is no way to detect off-task speech without a proper representation of it. Features provide ingredients for the representation. We now review relevant features used in related problems. These features serve as partial motivation for our features to detect off-task speech.

2.8.1.1 Acoustic features

Generally speaking, there are two types of acoustic features: word level acoustic scores from the ASR (San-Segundo, et al., 2001) and features computed directly from acoustic signals without going through the ASR (e.g., Batliner, et al., 2010). This dissertation cares about speaking

styles of off-task speech. Therefore, we use signal level acoustic features similar to those used in emotion detection and speaking style classification.

The feature extraction follows a two step process. First, extract frame level (typically 10 ms) acoustic descriptors from the signal. Second, compute statistics of the acoustic descriptors on the time span of an entire speech segment. A speech segment is usually defined based on units of meaning, such as words, phrases, and sentences. Batliner, Stefan et al. (2010) summarized seven types of acoustic descriptors for emotion detection: duration, energy, pitch, spectrum, cepstrum, voice quality, and wavelets. Based on these descriptors, they compute the following statistics to make features: extremes, means, percentiles, higher statistical moments (e.g., variance, skewness, kurtosis), combinations of these primitive functions (e.g., mean of max), distributional (e.g., number of segments) and regressional (e.g., linear regression coefficients) functions. Through feature selection, the most relevant descriptors found were energy and duration; the most relevant statistical functions included the mean and combinations of primitive functions.

2.8.1.2 Lexical features

We refrain from calling the features extracted from transcribed utterances "linguistic features" in order to avoid conceptual overlap between speech acoustics and linguistics. Instead, we call word-related features "lexical features."

OOD, OOV, emotion, and addressee detection all use lexical features, especially bag-of-words and n-grams (i.e., the occurrence of each word or ngram in a dictionary is a feature with a value of either 0 or 1), because these features directly encode what is spoken. OOD utterance detection uses bag of words and n-grams in latent semantic analysis or a topic model to classify an utterance into one of the pre-defined topics or a group of topics. OOV word detection uses

word context to help identify names likely to be OOV (e.g., Parada, Dredze, Filimonov, & Jelinek, 2010). Emotion detection looks for subjective words such as "great" (for excitement) and "stupid" (for frustration or anger) to help identify the emotion a speech segment expresses. Addressee detection uses person names to find out who the speaker is addressing. We can loosely group the above features into two classes – local indicators (i.e., those used in OOD and emotion detection to indicate a class directly) and context features (i.e., those used in OOV and addressee detection to detect the occurrence of a class).

2.8.1.3 Other features

Besides acoustic and lexical features, dialog context, the ASR lattice, language model scores, and gaze have also proved useful in solving related problems. Many of these features are useful for off-task speech detection as well. In particular, the dialog context feature assumes that consecutive utterances are likely to be about the same topic, which is helpful for detecting OOD utterances (I. R. Lane & Kawahara, 2005). Lattice features compute the cost for aligning a word level lattice against a phone level lattice, which helps identify OOV regions (White, Zweig, Burget, Schwarz, & Hermansky, 2008). Language model scores, frequencies of words occurring in an N-best list, and phone perplexity are all valid features for detecting OOV words or out-of-model events (Bansal & Ravishankar, 1998). Finally, gaze, a modality different from speech, can be a crucial indicator for identifying the addressee of an utterance (Jovanovic & Akker, 2004).

Many of these features, such as lattice features, language model scores, and N-best list features, focus on the ASR process rather than characteristics of off-task speech itself, even though in some cases they can be powerful for detecting off-task speech. Features based on the behavior of the decoder may change as ASR technology changes, whereas off-task speech

features that are properties of human behavior are more likely to remain similar over time. Therefore rather than enumerating and duplicating all the relevant features including those characterizing the ASR process more than the speech itself, we use only acoustic and lexical features, as well as the changes in the feature values during a session.

2.9 Summary

Spoken interactions are natural to humans. As long as a user is not strictly trained on how to speak to a system (which breaks the naturalness of the communication), off-task speech is quite likely to occur. Without prior knowledge about off-task speech, a system will fail both at recognizing off-task speech and at how to respond to the speech. Therefore, knowing the features of off-task speech and being able to detect off-task speech become necessary. Although many researchers have done work similar to off-task speech detection, the phenomena and the context of the phenomena they study are not the same as the problem discussed and addressed in this dissertation. In the following chapters, we will discuss features of off-task speech, how to use them to detect off-task speech, how well they detect different types of off-task speech, whether and how they generalize across tasks and domains, and the role of off-task speech detection in other applications beyond detecting off-task speech itself.

3. Methods to detect off-task speech

This chapter addresses the following research question:

How to detect off-task speech? What features characterize off-task speech?

We use machine learning methods to train a classifier to identify off-task speech. Classifier training requires a large number of examples consisting of utterances and labels indicating whether the utterances are off-task. Therefore a sub-question we ask is:

How to detect off-task speech when training labels are difficult to acquire?

Our goal for analyzing off-task speech is two-fold: first, to be able to detect off-task speech; and second, to study features of off-task speech through their roles in detection. Therefore, the focus of this chapter is not to search for an algorithm that gives the best classification accuracy, but to provide a reasonable detection method that incorporates the features we would like to study and to propose an evaluation metric that we can use to measure the effect of features.

3.1 Features

We represent off-task speech using its features. The most interesting features are those that distinguish off-task speech from on-task speech. The features we study can be loosely divided into two types: local features such as acoustic and lexical features extracted from an utterance, and context features extracted from the history of the speaker's speech. Many of the features we describe in this section are motivated by the related problems discussed in above.

3.1.1 Acoustic features and feature selection

Off-task speech is usually spontaneous and casual, whereas on-task speech tends to be more deliberate and formal. Shinozaki et al. (2009) found that spontaneity in speech correlates with

pitch and energy variations. That is, speech acoustics contain some information to characterize off-task speech. Thus we extract features from acoustic signals of an utterance. We do not know beforehand which features are the most promising. Therefore we generate many acoustic features and select the ones that have the strongest predictive power.

Since acoustic signals evolve on a time axis, most acoustic features are statistics summarizing signal samples over a period of time. The most common approach to extract acoustic features is to first extract acoustic Low Level Descriptors (LLD) such as pitch and intensity values on small time intervals (e.g., 10ms fames), and then calculate statistical function values of the LLDs. Figure 2.1 summarizes the feature extraction process. We tuned the number of acoustic features (45) to maximize classification accuracy on a development data set of 200 hand labeled utterances in oral reading. The 200 utterances were spoken by 4 children, 2 of whom appeared in our training data. We did not use the test data to tune any parameter.

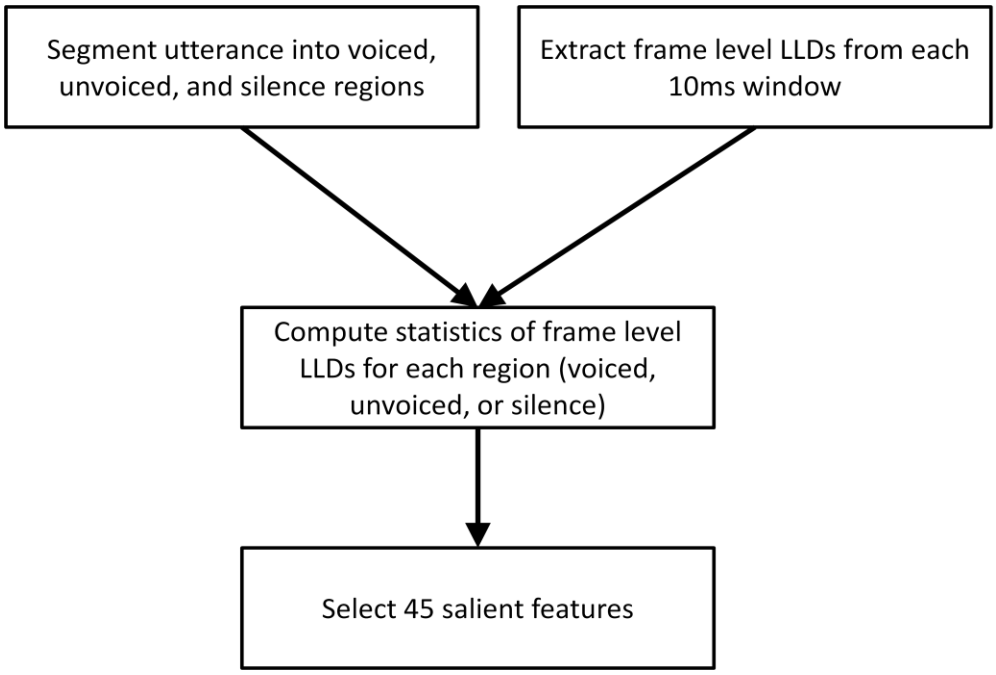

Figure 3.1 Process to exact acoustic features.

Table 1 summarizes five groups of low level acoustic descriptors we extract using Praat (Boersma & Weenink, 2010) scripts. Before extracting features from the acoustic descriptors, we use a Praat script to segment each audio recording into voiced, unvoiced, and silence regions. For each region and entire utterance, we calculate statistics to summarize frame based acoustic descriptor values, including mean, minimum, maximum, first, second and third quartiles, variance, the third central moment (to measure the lopsidedness of the distribution), and the fourth central moment (to measure the peakedness of the distribution). For formants, we also compute F1-to-F2, F1-to-F3, and F2-to-F3 ratios to capture descriptions of vowels (Miller, 1989). In total we extract 1,250 acoustic features.

Table 3.1. Five groups of low level acoustic descriptors.

Category	Members
Energy	Intensity, perceptual loudness
Spectrum	Pitch, first four formant frequencies with bandwidths, long-term average spectrum (LTAS)
Cepstrum	12 mel frequency cepstral coefficients (MFCC)
Voice quality	Jitter, shimmer, harmonics-to-noise ratio, degree of voice breaks
Miscellaneous	Zero crossing rate, number of pulses, utterance duration, percentage of unvoiced regions

Next we select features from the pool of 1,250 features we computed in the previous step. The goal for selecting features is two-fold. First, in order to interpret the features, we need a

sparser representation which contains only the features that have strong predictive power on detecting off-task speech. Second, the number of training examples of off-task speech is 4,236, which is on the same order of magnitude as the number of features. Therefore training a classifier with 1,250 acoustic features on only 4,236 off-task instances is doomed to overfit. Notice that regularization can avoid some overfitting by searching for a sparse solution during training (Lee, Lee, Abbeel, & Ng, 2006). However, regularization is tied to specific machine learning methods (i.e., each method has its own version of regularization). In fact there are many ways to select or rank features, and the goal of this thesis is not to search for the optimal learning algorithm on our data set(s). In fact the optimal learning algorithm on our data set(s) may not be the best one on other data sets. Therefore we chose a feature selection + classifier learning procedure commonly used in related problems such as emotion classification (Batliner, et al., 2010).

We select features that have the strongest predictive power. Therefore we apply the AdaBoost learning algorithm (Viola & Jones, 2001) to choose the top features. AdaBoost selects and combines many weak classifiers to obtain a strong classifier. Since it chooses weak classifiers incrementally based on their contributions to minimizing classification errors, AdaBoost naturally provides a feature selection scheme that selects the most predictive features when each weak classifier uses only one or a small number of features. At each step, AdaBoost chooses one weak classifier that maximizes the classification accuracy. Therefore this feature selection process can be treated as a special case of the popular sequential forward selection method. Besides its objective function, we use the AdaBoost algorithm to overcome the imbalanced training data problem (i.e., there is much more on-task speech than off-task speech in real data) by boosting the weight on errors made on the minority class. Furthermore, AdaBoost

automatically adjusts weights of training examples based on the errors made by the current model. For example, if the current model made an error on a training example, that example will be assigned a higher weight in the next training iteration. Ideally, the classifier under this scenario should not overlook classification errors on off-task speech, but boost them. The original AdaBoost algorithm (Freund & Schapire, 1996) is sensitive to noise. To accommodate noise in training data, we adopt the AdaBoost algorithm used in the Viola-Jones object detector (Viola & Jones, 2001), which is known to be more robust to noise and outliers (Vezhnevets & Vezhnevets, 2005). For the purpose of feature selection, the weak classifiers we use are single node decision stumps. Figure 3.2 illustrates an example decision stump.

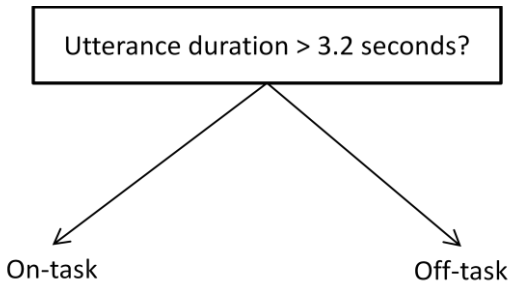

Figure 3.2 Example decision stump used in AdaBoost.

Besides selecting features, the AdaBoost algorithm produces a classifier that gives 69% detection rate with 24% false positives, where detection rate and false positives are calculated as:

$$\text{detection rate} = \frac{\text{number of offtask utterances classified as offtask}}{\text{number of offtask utterances}}$$

$$\text{false positive rate} = \frac{\text{number of } on\text{task utterances classified as offtask}}{\text{number of } on\text{task utterances}}$$

Table 2 lists the top 10 features ranked by descending absolute weights. The signs on the weights reflect the direction that the features correlate with off-task speech. That is, a positive weight indicates that a larger feature value suggests off-task speech; a negative weight indicates that a smaller feature value suggests off-task speech. On average off-task utterances sound higher, softer, shimmerier, and shorter.

Table 2. The top 10 acoustic features ranked by absolute weight value.

Feature name	**Weight**
Pitch – 3rd quartile	0.041
LTAS (0-8000Hz, voiced) – variance	-0.020
Spectrum (200-1000Hz, voiced) – median	0.017
Loudness (voiced) – 3rd moment	-0.013
Pitch – median	0.012
Shimmer	0.010
Loudness (unvoiced) – minimum	-0.0099
Voice break rate	0.0096
3rd MFCC – 4th moment	-0.0091
Utterance duration	-0.0088

3.1.2 Lexical features

Lexical features characterize the content of an utterance. We extract lexical features from the Sphinx-3 speech recognizer's output and confidence scores (CMU, 2010), using a 32 Gaussian mixture acoustic model trained on 43 hours of children's oral reading data recorded by the Reading Tutor.

We design language models for speech recognition to cover both the task domain and frequent off-task language. Our approach is to interpolate a task language model with a trigram language model built from an off-task corpus. Notice that the purpose of including a language model for off-task speech is not to improve ASR, but to output words that occur in off-task speech. Because the task language model and vocabulary vary by task, the overall language model we use to recognize both on- and off-task speech varies by task as well. For example, the task language model for children's oral reading consists simply of the trigrams in the sentence being read. To reflect generality of off-task speech, the language model for off-task speech stays the same across tasks. Figure 3.3 shows the general structure of the overall language model for recognizing both on- and off-task speech. Intuitively speaking the interpolation weight for off-task language model should follow the percentage of off-task speech (around 10% for the task of oral reading). However, words in off-task speech tend to be shorter than in on-task speech, and ASR is likely to output shorter words when there is background noise. Therefore we have to assign a much smaller weight to the off-task language model. The weight 0.001 was tuned to minimize WER on the development set of 200 utterances in children's oral reading.

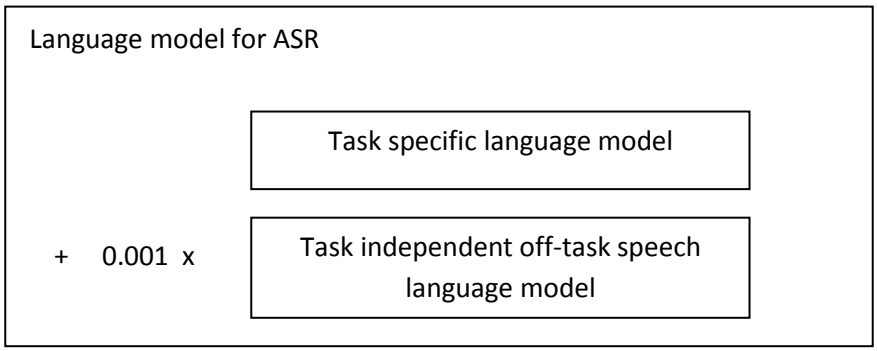

Figure 3.3 Language model components for recognizing both on-task and off-task speech.

Word distributions in off- and on-task speech are different. Our training corpus for off-task speech consists of transcriptions of 4,236 off-task utterances in oral reading training data, which comprise 18,040 tokens and 2,012 distinct word types. The 200 most frequent of these word types cover 74% of the off-task tokens and occurred in 86% of the off-task utterances in the training data. The 10 most frequent of these words are *I* (860 tokens), *you* (552), *it* (397), *to* (354), *the* (350), *what* (331), *on* (300), *go* (283), *this* (271), and *that* (262). Figure 3.4 shows differences of word frequency in on- and off-task speech. Notice that the most frequent word *the* in written English only ranks number 5 in children's off-task speech, which is roughly one third of its frequency in on-task speech.

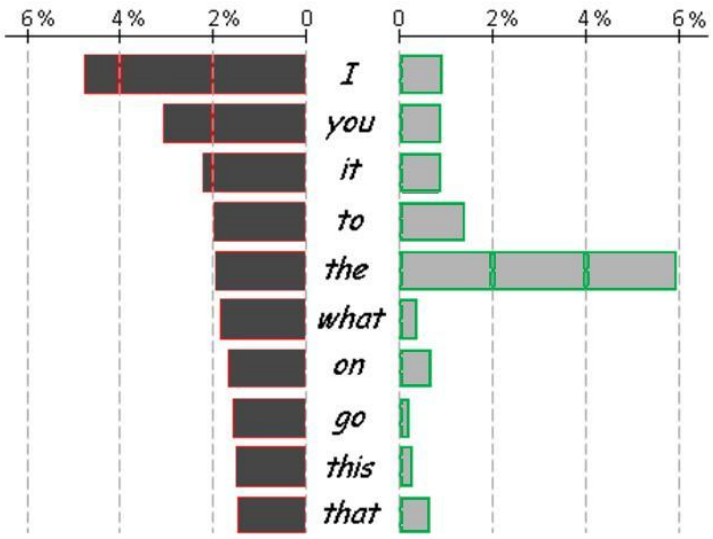

Figure 3.4 Distributions of the 10 most frequent words in off-task speech. Dark bars to the left denote word relative frequencies in off-task speech; light bars to the right denote word relative frequencies in on-task speech. Word distributions are calculated from children's transcribed oral reading data.

Since we recognize both on-task and off-task speech using a combined language model, we decide whether an utterance is on-task by comparing the number of words commonly occur in off-task speech with words less frequent in off-task speech. Since some words occur frequently in both types of speech (e.g., *the*), we focus only on words that occur more frequently in off-task speech than on-task speech. We rank these words by their differences in frequency in off- and on-task speech. 1,339 word types in our off-task corpus occur more frequently than in on-task speech. Table 3.2 shows the 10 words with the largest positive difference in frequencies in off- vs. on-task speech.

Table 3.2 10 words with the largest positive difference in frequencies in off- and on-task speech.

Off-task word	Frequency difference in off- vs. on-task speech (%)
I	3.81
you	2.22
what	1.42
go	1.34
it	1.28
this	1.20
oh	1.08
I'm	1.07
me	1.01
on	0.94

We train a trigram language model from the transcriptions of off-task speech. To avoid overfitting, we choose only the 350 words with the largest positive frequency difference in off- and on-task speech. These word types cover 66% of word tokens and occur in 85% of the utterances in our corpus of off-task speech. We call these 350 words more likely to occur in off-

task speech "off-task words," and all the other words in the language model "on-task words."

Due to recognition errors in ASR output, we use ASR confidence scores provided by the Sphinx3 speech recognizer[2] to compute two of our four features. In particular, we compute four lexical features from the ASR hypothesis of each utterance, shown in Table 3.3. We use these features to capture the distributional difference of off-task words in on- and off-task utterances.

Table 3.3 Lexical features for off-task speech detection, ordered alphabetically.

Lexical features
(1) Percentage of off-task words;
(2) Percentage of off-task words with ASR confidence scores higher than a threshold;
(3) Percentage of on-task words with confidence scores lower than a threshold;
(4) Percentage of silences: number of silences divided by the number of words and silences.

Features (2) and (3) use the confidence threshold to decide whether to classify a hypothesis word as recognized correctly. To minimize this classification error, we tune this threshold on the oral reading training data. The confidence score in Sphinx3 ranges from a large negative number to a large positive number. The default confidence threshold used by Sphinx3 is -2000. The threshold we tuned on children's oral reading data is 0. More than 94% of the correctly recognized words have confidence scores reaching or exceeding the threshold 0. Figure 3.5 shows the box plots of confidence scores on words misrecognized and correctly

[2] The ASR confidence score in the Sphinx3 speech recognizer corresponds to the posterior probability of a word hypothesis with starting and ending time.

recognized.

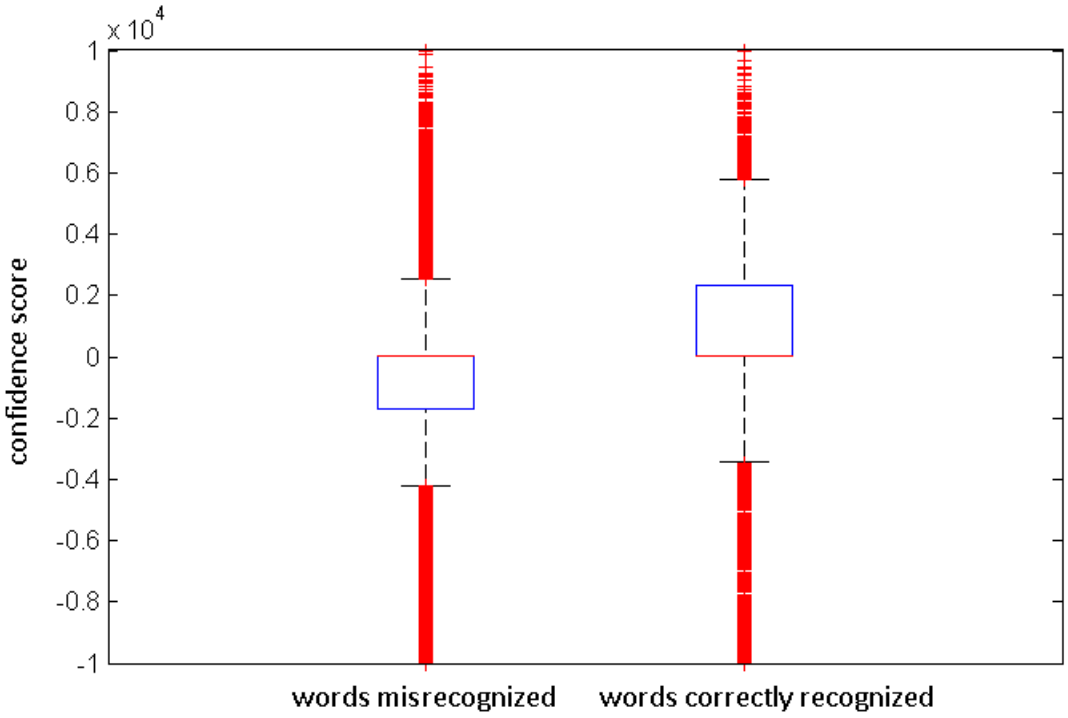

Figure 3.5. Box plots of confidence scores on words mis-recognized and correctly recognized. For words misrecognized, confidence score 0 is both the median and the 3rd quartile. For words correctly recognized, confidence score 0 is both the median and the lowest quartile.

3.1.3 Acoustic and lexical features in context

Off-task speech sits in a context of other utterances spoken by the same user, unless the user speaks only one utterance in a session (which is rarely the case). Therefore we extract features to represent the contextual factors that correlate with off-task speech.

Prior work on dialog context feature assumes that consecutive utterances are likely to be about the same topic, which has been shown to be helpful for detecting OOD utterances (I. R. Lane & Kawahara, 2005). We test this intuition on our data by asking two questions about the context of off-task speech. First, do off-task speech and on-task speech occur in chunks?

Second, is an off-task utterance likely to co-occur with other off-task utterances? We answer the first question by counting the transitions between on- and off-task utterances in our training corpus of 36,492 utterances from children's oral reading. Table 3.4 shows probabilities of the transitions.

Table 3.4 Probabilities of transitions between on-task and off-task utterances.

	To on-task	**To off-task**
From on-task	0.93	0.07
From off-task	0.51	0.49

Table 3.4 suggests that in most cases a child speaks only off-task for one time and then immediately returns back to the task again. Table 3.5 shows an example context of off-task speech.

Table 3.5 Example of what happens before, during, and after an off-task utterance in children's oral reading.

Context	**Utterance**
Before off-task	*Mice froze.* *They saw a pair of bright yellow eyes.*
During off-task	*I tell you I can hear myself.*
Back to on-task	*They they saw a pair of bright yellow eyes* *A large cat was staring staring around the room*

Based on this observation, we propose the actual working feature, which computes the difference between each local feature values of the current utterance and its running average feature value so far in the session. The goal was to find utterances that are much longer, much shorter, much louder, much quieter, much more recognizable, or much less recognizable than the utterances so far. Because the first utterance has no running average to compare with, and because 81% of the first utterances in our training data are on-task, we define its difference to be 0. To inform the classifier about the location of the utterance in a session, we have another feature with value 1 for the beginning utterance and value 2 for all the successive utterances in the session. Since we compute differences between the running average and current feature value for each feature, we double the number of acoustic and lexical features. After adding this location indicator, we have 45 (acoustic features) x 2 + 4 (lexical features listed in Table 3.3) x 2 + 1 (location indicator) = 99 utterance features in total.

3.2 Training and testing an off-task speech detector for children's oral reading

In this section, we build an off-task speech detector using the 99 features described above. The data we use to train and test the detector come from children's oral reading in the Reading Tutor.

3.2.1 Automatic labeling of training data using heuristics based on deviation from text

The oral reading corpus consists of utterances collected by Project LISTEN's Reading Tutor (Jack Mostow & Beck, 2007) during children reading out loud. The training data contains 36,492 utterances spoken by 495 children ages 7-10 totaling 43 hours of audio recordings.

Our definition for off-task speech in oral reading is that the utterance was not an attempt

to read the sentence text. Notice that on-task utterances include not only correct readings, but also disfluent or even incorrect readings. For example, if the screen displays "As soon as the boar perceived the tailor it ran at him", the utterance "I'm never gonna finish this story" is off-task, whereas "As soon as … as soon as the bear …" (i.e., repetition disfluency and miscues in oral reading) is on-task.

To train classifiers, we needed to label the data. Rather than hand-labeling so much training data as on- or off-task, we used a heuristic based on "deviation length" to find off-task speech in already transcribed oral reading. Computation of deviation length is developed previously in Project LISTEN by Evandro Gouvea in order to analyze the Reading Tutor's ASR accuracy separately on oral reading. The heuristic works as follows. First we align each word in a transcribed utterance against the sentence text using a dynamic programming algorithm similar to edit distance, but with lower penalties for repetitions than for insertions. Figure 3.6 shows a real example of alignment.

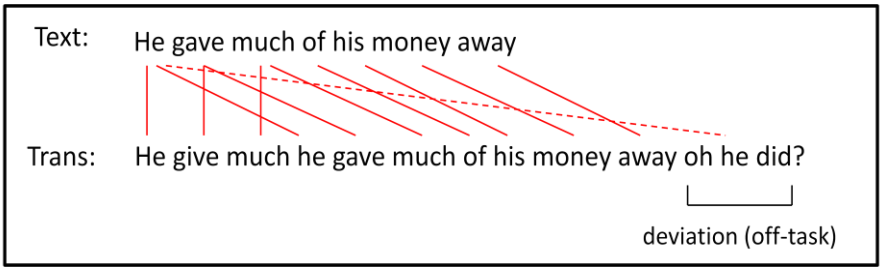

Figure 3.6 MultiMatch with disfluency and misreading

We define a sequence of *n* transcribed words as a deviation if none of them match the text words they are aligned against, except for isolated words (e.g., "he" in "oh he did?"), which we assume match by accident. It is easy to see that deviation length provides useful information for labeling off-task speech. Intuitively, if the absolute deviation length or the normalized deviation length

(i.e., deviation length / transcription length) reaches some threshold (e.g., =1 for 100% off-task speech), we can infer that the utterance contains off-task speech. Clearly, whether this heuristic works depends critically on the robustness of the alignment, which subsequently influences the computation of deviation length.

Robustness of the alignment means the following. First, the alignment algorithm has to account for re-read and skipped words. Second, it has to correctly match transcription words to the *intended* sentence words. The first requirement rules out conventional word alignment algorithms used for ASR scoring, such as the NIST alignment (Fiscus, Ajot, Radde, & Laprun, 2006). Instead we use an alignment algorithm that is designed explicitly for dealing with jumping around the words. This alignment algorithm is called MultiMatch. MultiMatch is mentioned by Tam et al. (Rasmussen, Mostow, Tan, Lindberg, & Li, 2011; 2003), but we know of no published description of the algorithm. To summarize, MultiMatch searches for the word alignment that minimizes word mismatch cost and jump cost subject to the constraint that the alignment satisfies a many-to-one mapping from transcription words to sentence words. Word mismatch cost takes into account edit distances computed from both spelling and pronunciations of words. Jump cost penalizes alignments of words not in the text order. MultiMatch, by design, meets the robustness requirement on matching re-read and skipped words. However, its current implementation does not meet the second robustness requirement on matching transcription words to the intended sentence words, even if the words were misread.

The second robustness requirement means that the alignment algorithm should be able to pair both correctly read and misread transcription words with corresponding sentence words. This is critical for computing deviation length, especially for utterances with few words. For example, when given the word “tipped” to read, one child says “okay” and another child says

"tiptoed." In the previous implementation of the MultiMatch algorithm, both are considered mismatches, so the deviation length for both cases is 1. However, the first case should be labeled as off-task, whereas the second one is more likely to be misread.

What causes "okay" to be treated the same as "tiptoed" in the previous MultiMatch is that the algorithm considers mismatch as a 0-1 hard decision. That is, two words mismatch as long as they are not spelled identically, even though the alignment algorithm does take into account phonetic and spelling distances in minimizing the word match cost.

Based on the above observations, we use the following adjustment to improve off-task labeling using transcription-sentence alignment. Instead of the 0-1 matching decision, we use word matching cost as a soft matching score. Word mismatch cost is defined as the smaller value of the spelling and pronunciation costs defined by the edit distance. In particular, to recognize misreading (e.g., "gave" as "give"), we measure the length of the deviation as the minimum of orthographic edit distance (normalized by the number of letters in the target word) and phonetic edit distance[3] (normalized by the number of phonemes in the target word) (Fogarty, Dabbish, Steck, & Mostow, 2001), summed over non-matching words. We then measure the relative deviation length as deviation length / transcription length. If this ratio exceeds 0.5, we label the utterance as off-task. If it falls between 0.36 and 0.5, we label it as partially off-task and exclude it from the training data. Algorithm 1 describes the off-task tagging procedure using this modified deviation length based on summation of word matching costs. We tuned the thresholds on a separate development set of 467 utterances from 7 children labeled by Becky Kennedy to maximize Kappa agreement (= 0.93) between Kennedy's annotation and our automatic labeling heuristic. One of the 7 children in the development set appeared in our

3 In the Colorado miscue database that contains oral reading miscues annotated by Richard Olson et al., the average and median phonetic distances between a target word and miscue, normalized by the length of the target word, are 0.64 and 0.6, respectively.

training data. Our automated method labels 4,236 utterances (12%) in training data as off-task, 29,196 as on-task (80%), and the other 3,060 (8%) as partially off-task.

Algorithm 1: Labeling off-task speech by modified deviation length

Input: Transcription T, Sentence S, Threshold thr

Output: **TRUE** if off-task, else **FALSE**

```
1.  MultiMatch(T, S). Store word matching costs for each alignment.
2.  max_deviation_length←0,  correct_count←0,  deviation_length←0
3.  for t←1 to LENGTH(T)
4.      // get mismatch cost for target word at position t and the transcription word aligned to the
        target word.
        cost ← GET-MISMATCH-COST(t, GET-ALIGNED-WORD[t])
5.      if cost = 0
6.          then  correct_count←correct_count+1  // extend match
7.              if correct_count > 1   // not isolated match
8.                  then if max_deviation_length < deviation_length
9.                          then max_deviation_length←deviation_length
10.                     deviation_length←0              // Start new mismatch
11.     deviation_length←deviation_length+cost
12. end for
13. if max_deviation_length < deviation_length   // the last deviation
14.     then max_deviation_length←deviation_length
15. if max_deviation_length / length(T) > thr
16.     then return TRUE
```

17. **else return FALSE**

3.2.2 Test data

The test data contains 651 oral reading utterances spoken by 10 randomly chosen children who do not appear in the training data, with total audio length of 1 hour 3 minutes. Two annotators independently labeled the 651 test utterances for the oral reading task with inter-rater agreement of Kappa = 0.96. They labeled 51 utterances (8%) as off-task, 569 (87%) as on-task, and 31 (5%) as partial off-task, which we exclude from analysis, along with the 8 where the raters disagreed.

The percentage of off-task utterances in the test data (8%) is lower than in the training data (12%). There are two reasons accounting for this difference. First, a small portion of the difference can be explained by labeling method. We labeled the training data automatically using the deviation heuristic, whereas we labeled the test data by hand. The deviation heuristic labeled 9% (56) of the test utterances as off-task, which is slightly higher than 8% (51), labeled by hand. Second, percentage of off-task speech varies by individual. Figure 3.7 illustrates percentage of off-task utterances varying by individual students. The figure shows only data from 69 students who spoke off-task, among 200 students randomly picked from the oral reading training data.

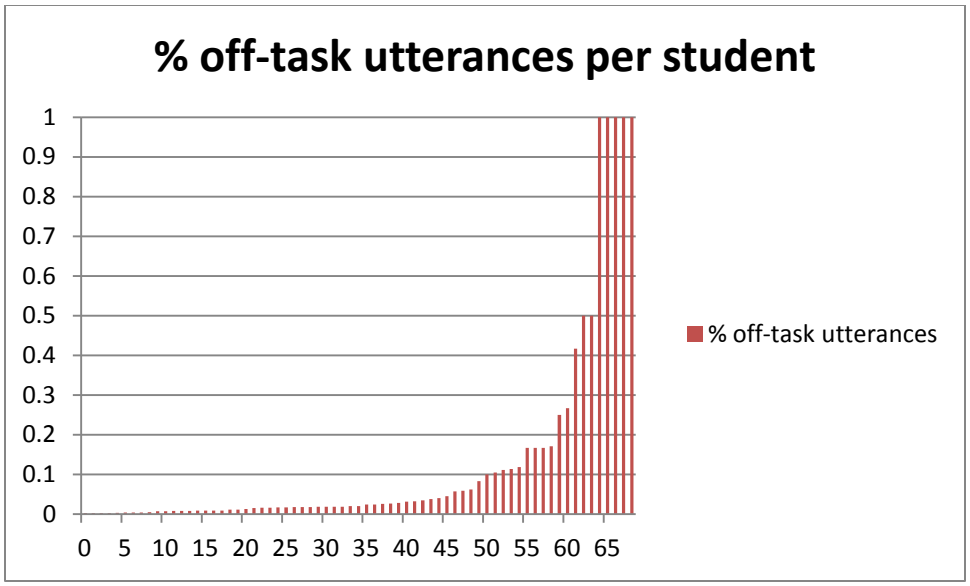

Figure 3.7 Percentage of off-task utterances per student

Since the test set comes from only 10 students, their percentage of off-task speech might have differed from the training set of students just by chance.

3.2.3 Evaluating the detector

The goal of this section is merely to demonstrate the overall classification accuracy of our approach and propose evaluation criteria for measuring classification accuracy. We will leave detailed analysis of classification results and predictive power of individual groups of features to Chapter 4.

We train an SVM classifier and test it on oral reading data, using LIBSVM-3.0 (Chang & Lin, 2001) with its radial basis function kernel and default settings except for the data weighting parameters. The SVM classifier outputs a score for each utterance. Whether an utterance is classified as off-task depends on a decision threshold. To avoid this parameter in evaluation, we use receiver operating characteristic (ROC) curves to summarize performance of the trained classifiers for 21 threshold values ranging from -2 to 2. Larger area under the ROC curve (AUC) indicates better classification accuracy.

There are many fewer off-task utterances than on-task utterances in our data (4,236 vs. 29,196 in oral reading training data and 51 vs. 569 in oral reading test data). A learning algorithm that aims to maximize overall classification accuracy is likely to fail on the minority class (Chawla, Japkowicz, & Kolcz, 2004). The direct impact of the data imbalance on our study is that using the natural distribution of the data to train an SVM classifier yields a degenerate solution that classifies every utterance as on-task. Such a result is useless, despite its overall classification accuracy of 92% on oral reading test data. To solve this problem, we use LIBSVM's data weighting parameter to assign different relative weights to off- and on-task utterances. By adjusting these weights to maximize the area under the ROC curve, we obtain a classifier that weights off-task data 8 times as much as on-task data. It detects 88% of off-task utterances and falsely classifies 11% of on-task utterances, i.e., approximately equal accuracy on both classes. Figure 3.8 shows the ROC curves for oral reading test data with off-task training utterances weighted 1, 2, or 8 times as much as on-task utterances. When weight is 1 (i.e., using the original data distribution), only 5 of the decision thresholds tested gave non-degenerate results; all other points cluster around [0,0], and [1,1]. The shape of the ROC curves does not change much for weights higher than 2.

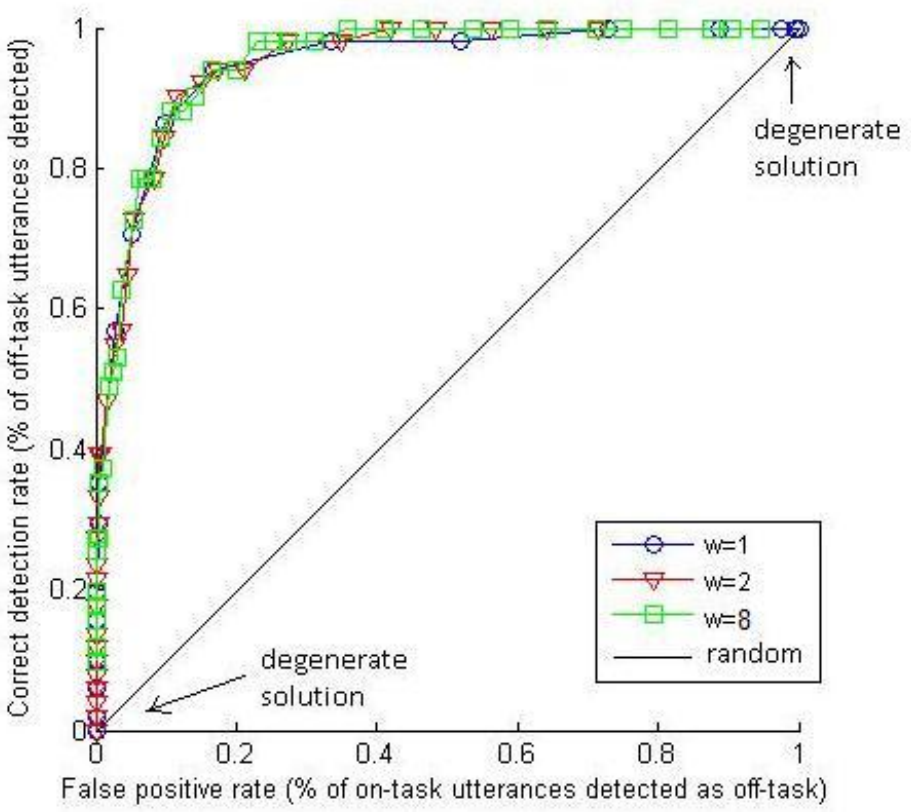

Figure 3.8 ROC curves on oral reading test data of classifiers trained with different data weights.

Notice that the diagonal line shows the ROC curve for a baseline classifier which randomly assigns labels to utterances. The ROC curve is not sensitive to data distribution. Consider a baseline classifier that randomly assigns labels according to data distribution in training data. For example, our training data for oral reading contains 8% off-task utterances. According to this distribution, the baseline classifier randomly classifies 8% of the utterances as off-task, and the rest as on-task, which means that the expected percentage of utterances classified as off-task is 8% for both true off-task and true on-task utterances. Therefore the expected true positive rate for the baseline classifier is (0.08 * # true off-task) / # true off-task = 0.08, and

the expected false positive rate is (0.08 * # true on-task) / # true on-task = 0.08. Hence the diagonal line in the ROC diagram captures the performance of the baseline classifier.

To further prove that our detector outperforms the baseline classifier that randomly assigns labels based on percentage of off-task speech, we run the baseline classifier on the test data for children's oral reading. The purple line with triangle markers in Figure 3.9 shows the ROC curve of the baseline classifier, which is not too far away from its expected curve (i.e., the black diagonal line).

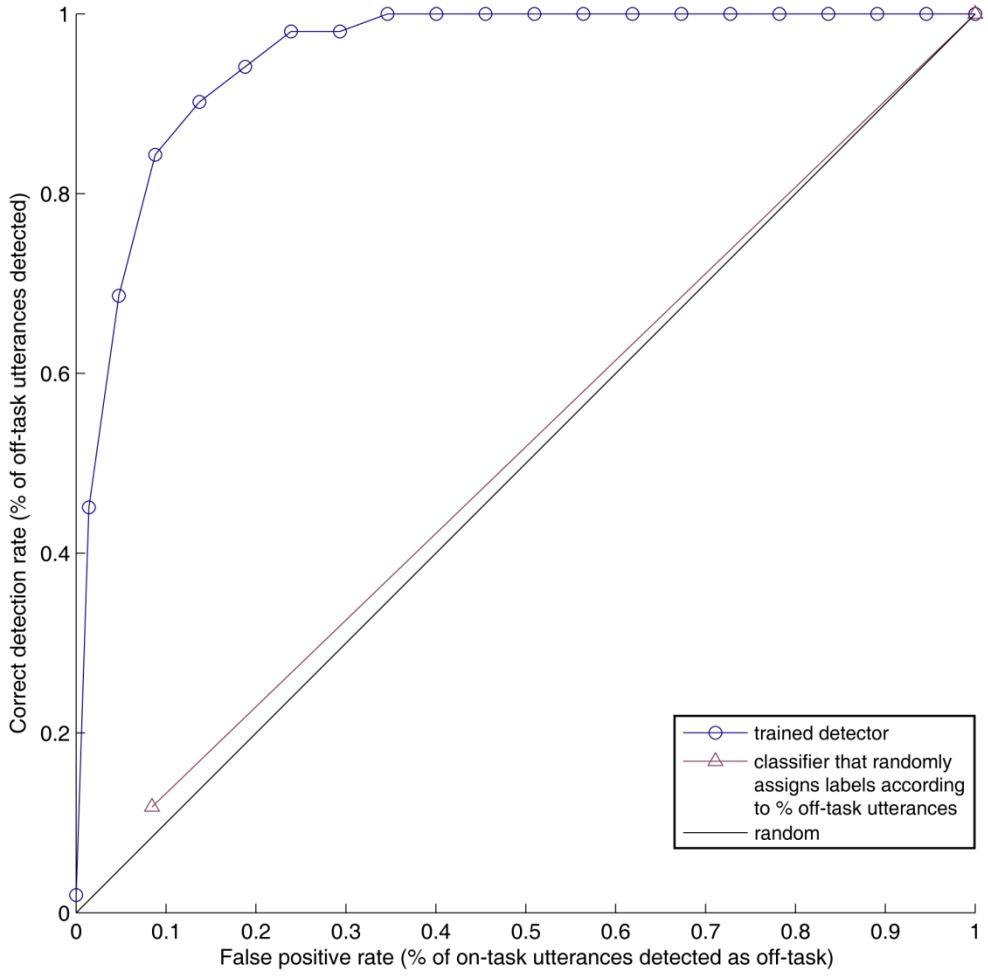

Figure 3.9 Comparison between the trained detector and a baseline classifier that randomly assigns labels based on percentage of off-task speech

3.3 Detecting off-task speech segments

During people's verbal interaction with an automated agent, their off-task speech may occur in two forms: 1) as a whole recorded utterance or 2) as a portion of a recorded utterance. A single utterance can contain both on- and off-task speech. Therefore it is natural to refine our analysis from utterance level to segment level. However, due to lack of labeled data, it is challenging to do this kind of more finer-grained analysis. This section shows how we detect off-task speech within an utterance using automatically derived labels and even when segment level labels are completely missing. Figure 3.10 shows real examples of oral reading utterances containing both on- and off-task speech.

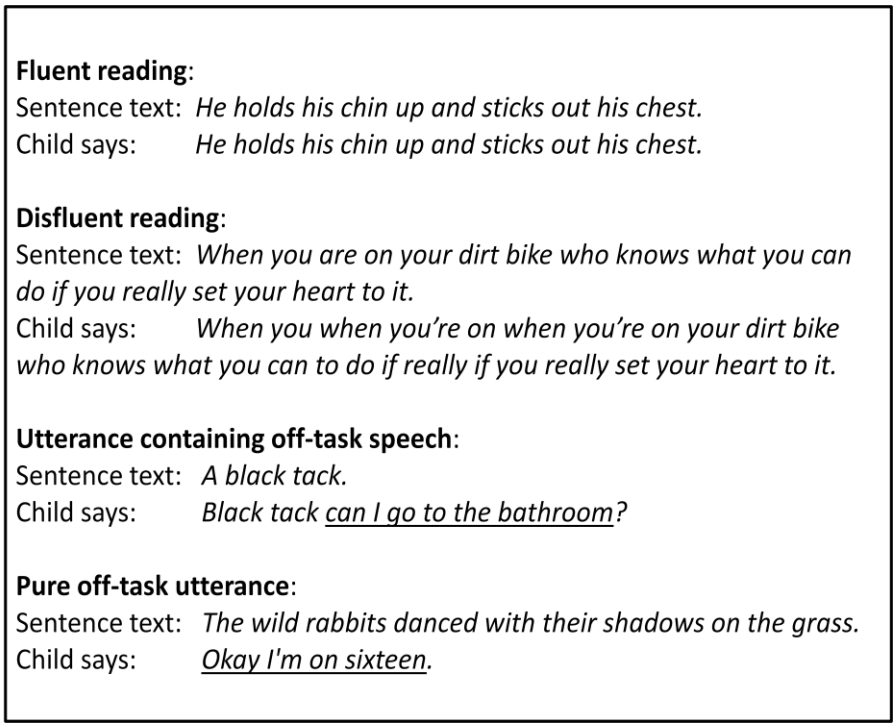

Figure 3.10 Real examples of on-task and off-task speech in oral reading. Underlined text denotes transcription of off-task speech.

To detect off-task speech for a given utterance, two questions need to be addressed. First, does an utterance contain off-task speech? Second, if yes, where in the utterance is the off-task

speech? The approaches discussed earlier in this section all assume that the training examples are labeled on the utterance level, and the test instance will be classified as either completely off-task or completely on-task. In this section, we demonstrate how to develop an off-task speech detector to detect off-task speech on finer-grained units than utterances.

3.3.1 Segmenting utterances

The natural units we get directly from audio recordings are utterances. In this section, we work on finer-grained audio segments. First we need to decide the size of segments. There is a trade-off between resolution and content of segments. If the segment size is too small, we may not get enough information. If we use a large size, we miss accurate transitions from one type of speech to another. We choose to train and classify speech regions (i.e., voiced and unvoiced speech areas) of utterances because they roughly mark separated speech areas. We use a Praat script (Boersma & Weenink, 2010) to extract the speech regions for each utterance. In the end we extract 56,551 speech segments from 33,432 training utterances. We label an audio segment as off-task if over half of its time interval is occupied by words in off-task regions. Some neighboring segments in the same utterance have the same category label. There are 5,381 utterances (15%) containing both on- and off-task speech. On average each utterance has only 0.26 transitions between on- and off-task segments.

3.3.2 Training with automatically generated segment level labels

We use an automated process to identify off-task speech regions in training utterances. The input to the algorithm is the audio recording of an utterance and its transcription provided by human transcribers. For each utterance, the extraction process outputs time intervals with labels indicating whether an audio segment is off-task. The process works as follows. First, we run Viterbi alignment on an utterance and its transcription using the CMU Sphinx3 force alignment

tool and the CMU dictionary for word pronunciations. This step marks the audio stream with time points indicating the start and end time of each word in the transcript. Not all the training utterances are successfully aligned, due to 49 transcribed words not found in the CMU dictionary. Next we use MultiMatch to align transcript words to sentence text. This step labels words that deviate from the reading sentence as off-task. The difference between MultiMatch and conventional NIST alignment is its ability to account for misreading and disfluency, as described previously. Algorithm 2 shows the details of the automatic labelling process. Once we get the off-task words and time marks for each word, we can label the off-task regions using time marks.

Algorithm 2: Labeling off-task speech segments in an utterance

Input: Transcription T, Sentence S, audio file au

Output: Offtask_Word_Segs, Offtask_Time_Segs

```
1.  MultiMatch(T, S). // Store word matching costs for each alignment.
2.  correct_count←0,  deviation_length←0, start ←0
3.   for t←1 to LENGTH(T)
4.        cost ← GET-MATCHING-COST(t, GET-ALIGNED-WORD[t])
5.        if cost = 0
6.          then correct_count←correct_count+1
7.                if correct_count > 1
8.                  then if deviation_length > 0
9.                        then PUSH(Offtask_Word_Segs, MAKE-PAIR(start, start+deviation_length)),
                                start←start+deviation_length, deviation_length←0
10.       else correct_count ←0
11.       deviation_length←deviation_length+cost
12.  end for
```

```
13.  if deviation_length > 0     // the last deviation
14.       then PUSH(Offtask_Word_Segs, MAKE-PAIR(start, start+deviation_length))
15. Word_Time_Marks ← FORCE-ALIGN(au, T)
16. foreach p in Offtask_Word_Segs
17.    do PUSH(Offtask_Time_Segs,
                    MAKE-PAIR(Word_Time_Marks[p.Start].Begin_Time,
                              Word_Time_Marks[p.End-1].End_Time))
```

Similar to utterance level off-task speech detection, we use an SVM classifier trained directly on automatically generated segment level labels to classify audio segments. Utterance level labels can be easily derived from the segment labels: the utterance contains off-task speech if there exists an off-task segment in the utterance; otherwise the utterance is pure on-task.

3.3.3 Training without automatically generated segment level labels

Due to environmental noise in the audio recordings and limitations of our speech recognizer, using forced alignment for deriving segment level labels introduces additional noise to the training data. However, there is a way to bypass the segment level labels. With multiple-instance learning (MIL), we can learn classifiers to detect off-task speech on both utterance and segment levels even when segment level labels are missing during training. Later we will show that the MIL algorithms generate detection rates comparable to those achieved by a conventional support vector machine trained on segment labels.

In MIL, training examples are bags of instances. Each bag receives a label. A positive bag label indicates that the bag contains at least one positive instance. Otherwise the bag is assigned a negative label. The goal of MIL is to learn the labels of bags and instances simultaneously. For the case of off-task speech detection, utterances correspond to bags of

instances, and each instance is an audio segment that is either off-task or on-task. Thus MIL can be a suitable tool to solve the problem of multiple level off-task speech detection.

MIL was first proposed by Keeler et al. (1990) to solve hand-written digit recognition in images without knowing where the digits are located. The problem was formally defined and named by Dietterich et al. (1997). MIL algorithms have many variations. To compare with conventional SVM, we use the multiple-instance SVM algorithm implemented in the MILL toolkit (Yang, 2008) on our data. Multiple-instance SVM (Andrews, Tsochantaridis, & Hofmann, 2002) assumes that the target instance-level concept can be separated out by a hyper-plane in a high dimensional feature space. What makes multiple-instance SVM different from a conventional SVM is its constraint over the relation of instance labels and their bag labels. To compare classifiers trained with and without labels on speech segments, we use multiple-instance SVM to detect off-task speech under two conditions: when only the utterance labels are available and when the segment labels are available too.

The problem of detecting off-task speech segments in utterances resembles problems in audio segmentation. Common audio segmentation problems focus on detection of significant transitions in audio signals, such as transitions from music to speech, from vocal to non-vocal and silence, from verse to chorus, and transitions between other homogeneous audio components. The main methods used include the identification of significant changes in properties of the audio signal by measuring signal differences (e.g., Foote, 2000) and applying machine learning algorithms to classify windows of audio segments (e.g., Lu, Li, & Zhang, 2001). Our problem in off-task speech detection is different from existing audio segmentation applications in that the transitions we aim to detect are within a single type of signal, that is, speech, and within the same speaker. In addition, each of our audio recordings contains a single utterance, which is only 5

seconds long on average. This is much shorter than the materials used in common audio segmentation problems, such as a song, a movie, or a conversation.

3.3.4 Evaluation of the detector

The utterances in our test data are the same as in the test data for evaluating utterance level off-task speech detection. The test data contains 51 pure off-task utterances, 31 mixed utterances that contain both off-task and on-task speech, and 569 pure on-task utterances. We extract 1,918 speech segments from the utterances. For evaluation, we further mark time intervals of off-task speech in the 31 mixed utterances. We assign labels to segments accordingly. If a segment spans both types of speech, we assign to it the label of the type that takes more than half of the time interval. Segments in pure off-task utterances are assigned off-task labels. Segments in pure on-task utterances are assigned on-task labels. In all, 430 of the audio segments are labeled as off-task, and the other 1,488 on-task.

To deal with the imbalanced data problem, we set the weight on off-task speech to be four times higher than that on the on-task speech. We use the same weights for both the conventional SVM and MIL-SVM.

The features we use are the same as utterance level features. But because we work directly on speech segments, we do not segment an utterance further into speech and silence regions, as we did for utterance level feature extraction. We select acoustic features using the same AdaBoost algorithm. Table 3.6 shows the 10 acoustic features with largest absolute weights. Extraction of lexical features and context features on speech segments follows exactly the same process as utterance level off-task speech detection. On average off-task speech sound softer and shorter. The signs on the weights are consistent with those for the features selected using utterance-level data.

Table 3.6 The top 10 acoustic features with largest absolute weights assigned by the AdaBoost algorithm for segment level off-task speech detection. The weight values are larger than for whole-utterance features because whole utterances have more acoustic features, due to including features extracted from both speech and silence regions.

Feature name	Weight
Intensity – variance	-0.029
10^{th} MFCC – variance	0.026
Harmonics to noise ratio – first quartile	-0.020
Segment duration	-0.019
5^{th} MFCC – variance	0.013
Pitch – variance	-0.012
Intensity – first quartile	-0.0115
4^{th} MFCC – max	-0.0113
1^{st} MFCC – 4^{th} moment	-0.010
Pitch – 4^{th} moment	-0.009

Figure 3.11 compares the ROC curves of the SVM trained on automatically generated segment level labels, the MIL-SVM trained on automatically generated segment level labels, and MIL-SVM trained using only the utterance level labels.

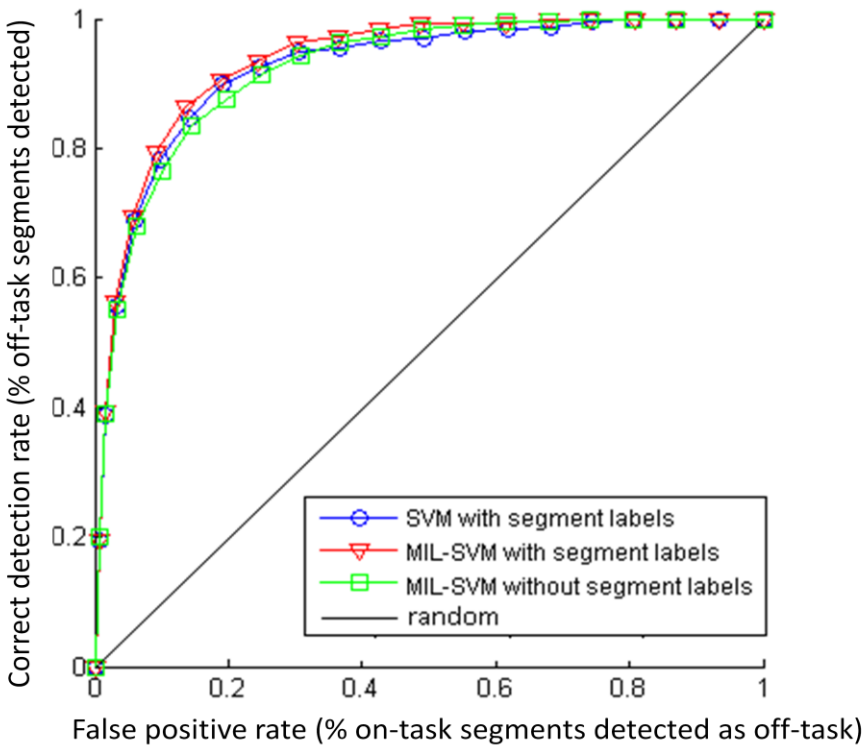

Figure 3.11 ROC curves of three classification methods for segment level off-task speech detection.

By modelling constraints between segment and utterance labels, MIL-SVM slightly out-performs conventional SVM. Overall, MIL-SVM without segment labels works almost as well as conventional SVM trained with segment labels. Segment level off-task speech detection is less accurate than utterance level detection, due to additional errors in labels of audio segments caused by Viterbi alignment. At 10% false positives, we correctly detect 87% of off-task utterances, and detect only 80% of off-task segments.

3.4 Summary

This chapter described how to detect off-task speech in children's oral reading, and how to evaluate an off-task speech detector. These aspects serve as the foundation for analysis of features and detector generality in future chapters. The feature set includes 45 acoustic features

selected from 1250 acoustic features, 4 lexical features motivated by the difference between word distributions in off-task and on-task speech, one context feature specifying the position of the utterance in an interaction session, and 49 other context features tracing the differences between acoustic and lexical feature values and the running averages. We use the SVM to train and classify speech. The SVM provides a max-margin approach to discriminate off-task speech from on-task speech. We provide methods to detect off-task speech on two levels: utterance level and segment level. The majority of this thesis is built on utterance level detection. Although segment level detection is not as accurate as utterance level detection, it provides an opportunity to observe features of off-task speech at a finer grained scale.

Another contribution of this chapter is that instead of hand-labeling hundreds of thousands of utterances that may require significant labor and time to obtain, we describe an automatic approach. Both the utterance and segment level labels are obtained using automatic methods. These automatic labeling methods, especially for deriving utterance level labels, are defined in terms of target text. However, as Chapter 4 will show, the data that the labeling methods generate help train models that generalize to other tasks.

In the following chapters, we will use the detection methods to investigate the predictive power of features through their effects on detection accuracy.

4. Generalizing the off-task speech detector to other Reading Tutor tasks

To test the generality of the off-task speech detector trained on oral reading, we apply it to three other tutorial tasks in the Reading Tutor: self-questioning, vocabulary activities, and think-aloud.

4.1 Tutorial activities in the Reading Tutor

Apart from oral reading, we have Reading Tutor data from six different tutorial tasks for vocabulary learning and reading comprehension, namely self-questioning, vocabulary activities, think-aloud, summarizing, visualizing, and activating background knowledge. All these tasks involve prompted spoken responses from children. Table 4.1 summarizes the characteristics of speech in each of the Reading Tutor tasks.

Table 4.1 Characteristics of speech and its lexical content in different Reading Tutor tasks.

Task	**Type of speech**	**Predicted task lexicon**	**Predicted task language**
Activating background knowledge	Prompted speech.	Target concept words (e.g., "goat"), words that commonly co-occur with target concept words.	Story ngram + ngrams in documents about target concept.
Oral reading	Read speech.	Sentence words to read.	Sentence to read + disfluency.
Self-questioning	Prompted or scaffolded speech.	"I", "wonder," question words (e.g., "what," "why"), story words.	I wonder what/why/who/where/when/how /if …
Summarizing	Prompted speech.	Story words.	Story ngram.
Think-aloud	Prompted speech.	Story words.	Story ngram.
Visualizing	Prompted speech.	"I," "see," story words.	Story ngram.

Vocabulary activities	Prompted speech.	Words in dictionary definition.	<target word> means …

We investigated how to detect off-task speech in four of the tasks: oral reading, self-questioning, defining words, and think-aloud. These four tasks cover all the speech types in the Reading Tutor tasks: read, scaffolded, and prompted speech. They also illustrate different approaches for modeling lexical content of the speech. To model summarizing, we use ngrams from story text, which is the same as think-aloud. To model visualizing, self-questioning, and think-aloud, we use story ngrams as well as opening phrases such as "I see," "I wonder" or "I'm wondering." To model activating background knowledge and vocabulary activities, we use target words and external resources relevant to the target words. Oral reading, self-questioning, think-aloud, and vocabulary activities sample diverse types of speech and approaches to modeling.

To detect off-task speech, we need to compute acoustic and lexical features. The feature extraction process for acoustic features is completely task-independent. Once the ASR output is ready, feature extraction process for lexical features is also task-independent. However, to obtain ASR output, we need to apply different language models to model on-task speech. The rest of this chapter will discuss the language models for on-task speech in each task.

4.2 Self-questioning

Self-questioning is a reading comprehension activity, where the Reading Tutor prompts children to ask questions about the text they are reading. Our data for the self-questioning activity comes from 34 children ages 7-10 responding to questioning prompts while reading 10 stories. Two of

the texts are fictional, and the other 8 are informational. Figure 4.1 shows example dialog fragments in the self-questioning activity.

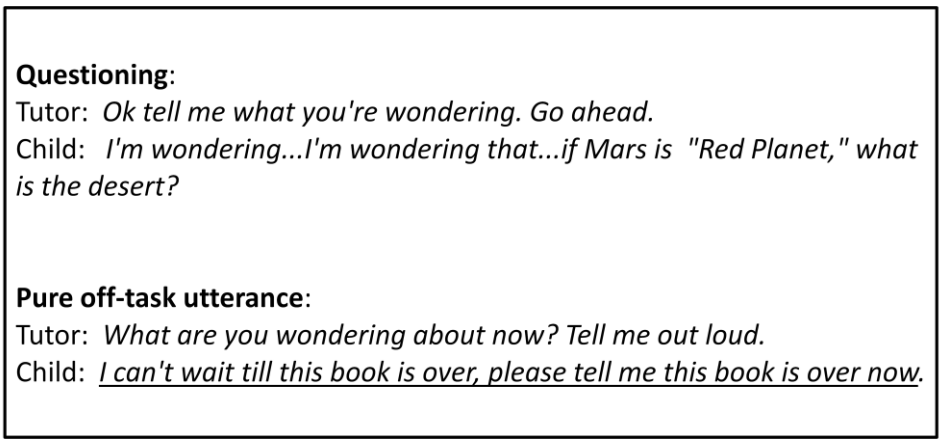

Figure 4.1 Prompts and responses in self-questioning

In contrast to oral reading, we do not have enough data from the self-questioning activity for training. Therefore we use it only as a test task to evaluate the generality of the off-task speech detector trained from oral reading utterances.

Among the 250 self-questioning responses, we manually labelled 157 on-task utterances, 69 utterances containing off-task speech, and 24 null responses (i.e., non-speech events such as singing, humming, babbling, and silence). Two annotators independently labelled the data. The inter-rater reliability Kappa = 0.784.

4.2.1 Language model

To build a language model for on-task speech, we generate questions to predict children's responses to the self-questioning prompt (Chen, Mostow, & Aist, 2011). We use off-the-shelf natural language processing tools to annotate text and generate questions from the annotations. Unlike syntactic approaches that transform statements to make questions (Gates, 2008), we generate questions by filling in question templates. Since the prompts (shown in Figure 4.1) include the meta phrase "I wonder," we expect children to follow similar phrasing too. So all the

question templates begin with the phrase “I wonder” or “I’m wondering.” The remainder of the question templates depend on the information requested.

Template 1:

```
I wonder | I’m wondering
how|why|if|when <THING> <VERB-PHRASE>.
```

Example:

I wonder how wind makes electricity.

Template 2:

```
I wonder | I’m wondering
who|what <VERB-PHRASE>.
```

Example:

I wonder what lives on Mars.

We generate items to fill in <THING> and <VERB-PHRASE> by running the ASSERT semantic role labeler (Pradhan, Ward, & Martin, 2008) on the story text. We extract text marked by the tag [ARG0] (verb argument in front of the verb) to fill in <THING>. We combine text marked by [TARGET] (verb) and [ARG1] (verb argument after the verb) to fill in <VERB-PHRASE>. For example, from the story sentence annotated using ASSERT: [ARG0 *wind*] [TARGET *makes*] [ARG1 *electricity*] [ARGM-MNR *all by itself*], we get the question “I wonder how *wind makes electricity.*”

To predict children’s speech, we need a language model to set constraints on vocabulary and word order. We do not have sufficient training data to train the language model on

children's spoken questions. Therefore, we use the automatically generated questions as a synthetic corpus to build the language model. In particular, we construct a probabilistic finite state grammar (PFSG) that incorporates the generated questions, with equal probabilities for the transitions from each state due to absence of data.

The coverage of the PFSG is limited. We dealt with this problem along three dimensions. First, to improve coverage of the language model, we added the Dolch list (Dolch, 1936) of 220 words common in children's books. We expected children's questions to be about story text, so we added all the story words. We used a morphology generator to add all inflections of each verb. We use the resulting vocabulary for the interpolated language model that we describe now. Second, to make the language model for children's questions more robust, we interpolate the PFSG with part of speech (POS) bigrams. We train the bigrams from a POS corpus generated from 673 children's stories from Project LISTEN's Reading Tutor. The stories contain 158,079 words. We use the Stanford POS tagger (Toutanova, Klein, Manning, & Singer, 2003) to find the POS of the words in the stories. We then train a bigram model using the SRILM toolkit (Stolcke, 2002). To incorporate this model in the PFSG, we add a state for each POS tag. We add a transition from states immediately preceding <VERB-PHRASE> to the VB (verb) state and assign it a heuristic probability .0001, and transitions between POS states their POS-bigram probabilities. We tag each word with its most frequent POS. Thus this model approximates Pr(drink the milk) as .0001 * Pr(DT | VB) * Pr(NN | DT). Third, to cover responses that are not questions (i.e., off-task speech), we interpolate the language model with the trigram language model trained from the off-task portion of the oral reading corpus.

The coverage of our language model directly affects ASR accuracy. We measure the coverage using the out-of-vocabulary (OOV) rate computed as the percentage of transcribed

word tokens not included in the language model. The overall OOV rate (i.e., percentage of word tokens not covered by the language model vocabulary) for our language model is 7%. Fully 32.7% of the word tokens in the spoken responses do not occur in any of the generated questions.

4.2.2 Off-task speech detection result

After obtaining lexical features using the ASR output gained with the language model described in previous section, we combine them with acoustic and contextual features, and we apply the classifier trained on oral reading data to detect off-task speech in self-questioning. Figure 4.2 compares the ROC curves for the same classifier on both oral reading and self-questioning.

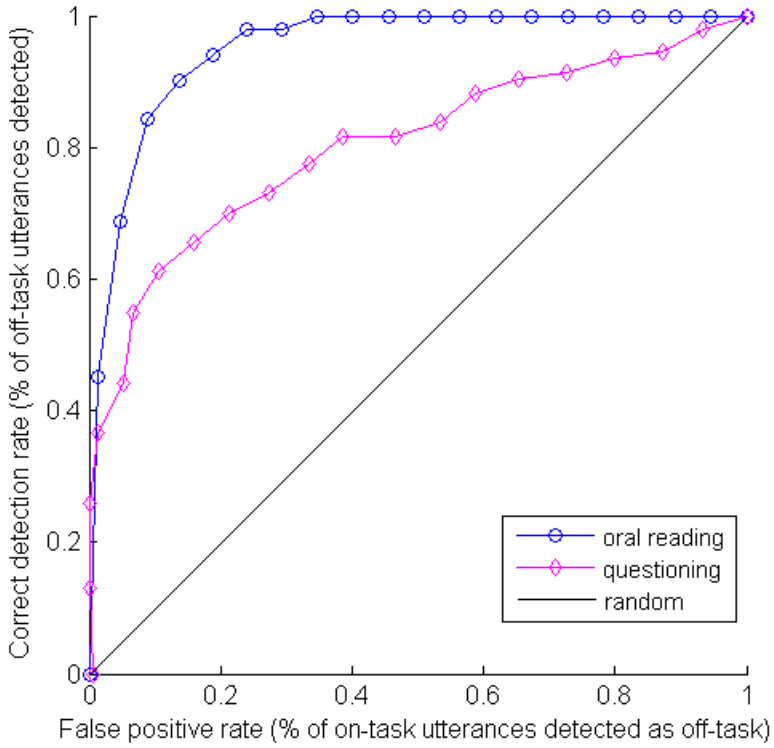

Figure 4.2 Comparison of ROC curves for off-task speech detection on oral reading and self-questioning using a classifier trained on oral reading.

Due to differences in the nature of speech between the two tasks, accuracy of the classifier on self-questioning is much lower than on oral reading. One reason is that self-questioning consists of scaffolded and prompted free-form responses, which allows more freedom than read speech. Another reason is that responding to self-questioning prompts involves higher cognitive load than oral reading, so the speech tends to sound more spontaneous, i.e., has more speech repairs, pauses, and casual expressions. This spontaneity makes the speech harder to distinguish from off-task speech, which also shares some aspects of spontaneous speech.

4.3 Think-aloud

Think-aloud is a research indicator of children's reading comprehension processes. During oral reading, now and then the Reading Tutor asks what the child is thinking about. The child then responds by thinking aloud. Figure 4.3 demonstrates example dialogs in think-aloud.

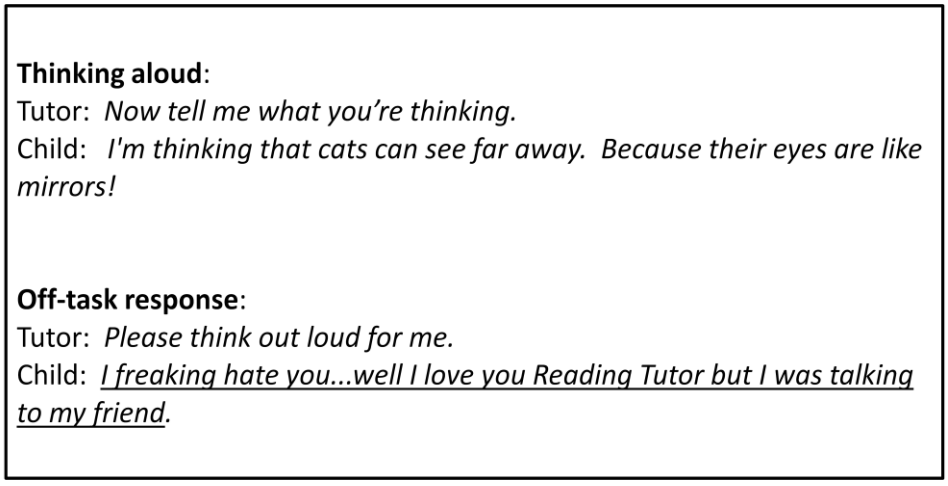

Figure 4.3 Prompts and responses in think-aloud.

Our think-aloud data comes from children ages 9-10 responding to think-aloud prompts while reading two informational texts, "Seeing Eye to Eye" and "People and Goats." Among the 284 think-aloud utterances, 168 utterances were labeled as on-task, 47 as off-task, and 69

utterances contained no spoken responses. Two annotators independently labeled the data, with inter-rater reliability Kappa = 0.916.

4.3.1 Language model

Compared to the other Reading Tutor tasks, our language model for think-aloud is relatively simple. We expect children to follow scenarios in the story. Therefore it is likely for them to echo things described in the story. For example, in the first example response shown in Figure 4.3, the child talks about eyes of cats and that they resemble mirrors. We can find the corresponding story sentences talking about exactly the same thing:

"… Some of them have really big eyes! That helps them catch more light. Others, like cats, have an extra eye part. It's like a mirror. The eye catches the light coming in. And then catches it again when it bounces off the extra part …"

Thus we simply use trigrams in the stories to model children's responses, although a more complicated approach could assign heavier probability mass on recently read sentences.

The language model used for ASR interpolates story trigrams with off-task trigrams trained from oral reading. When applying this language model, the OOV rate on all the utterances, including off-task, is 7.1%.

4.3.2 Off-task speech detection result

We follow the feature extraction procedure described in section 3.1 to compute features from acoustic signals, ASR output, and utterance context. We then use the classifier trained on oral reading data to detect off-task speech in think-aloud. Figure 4.4 compares the ROC curves for the same classifier on both oral reading and think-aloud. For similar reasons as self-questioning,

the classifier trained on oral reading is less accurate on utterances in think-aloud than on oral reading.

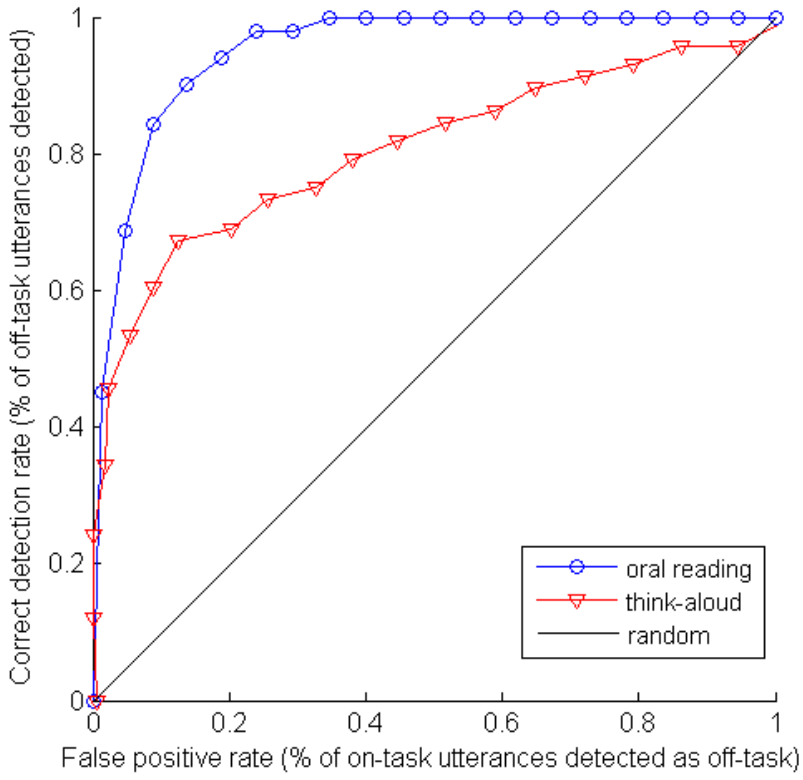

Figure 4.4 Comparison of ROC curves for off-task speech detection on oral reading and think-aloud using a classifier trained on oral reading.

4.4 Vocabulary activities

The vocabulary task prompts children to explain the meaning of a word or to compare which of two words is closer in meaning to a target word. As Figure 4.5 illustrates, the vocabulary prompts elicited utterances less constrained than oral reading.

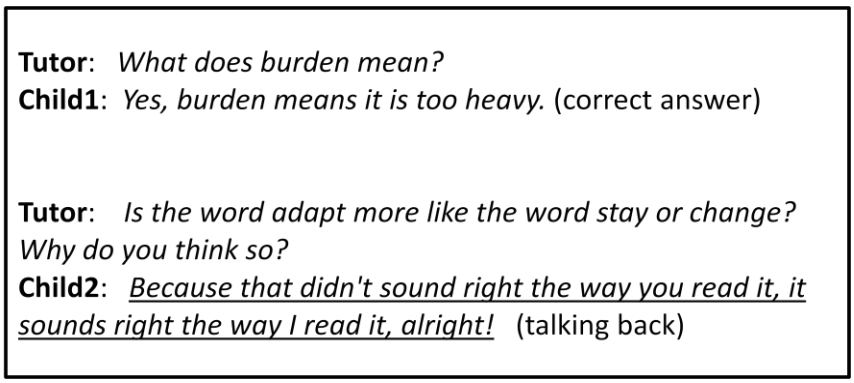

Figure 4.5 Prompts and responses in vocabulary activities.

An undergraduate summer intern, Corinne Durrette, annotated the vocabulary responses with finer grained categories such as correct answer (e.g., the first response in Figure 4.5), wrong answer (e.g., *a burden means you're a thief*), no response, playing (e.g., *Haha! Oh oh!*), and talking back (e.g., the second response in Figure 4.5) (Durrette, 2010). We categorized correct and wrong answers as on-task, and the rest as off-task. The utterances were later independently annotated by the author using only on- and off-task labels. The Kappa score was 0.83, with most disagreement occurring on the wrong answer category. We obtained 410 utterances, 139 (34%) of them labeled off-task.

4.4.1 Language model

Unlike the other Reading Tutor tasks, vocabulary activities do not occur in the context of a story. To model children's responses to vocabulary prompts, we used children's dictionary definitions of the target words. Since the dictionary definitions do not provide enough data to train a language model for the vocabulary task, we used a unigram language model consisting of the words in the definitions and synonyms of the target vocabulary word in Wordsmyth Children's

Dictionary (Wordsmyth, 2010) and WordNet (Fellbaum, 1998), along with words we expected in children's word explanations, such as *means* and *something*.

Unlike the highly predictable on-task utterances in oral reading, explanations of word meaning vary both lexically and syntactically. For example, many children used the word "heavy" to explain "burden." Although "heavy" often characterizes burden, it did not appear in either of the two definitions we used. Wrong but on-task answers are even harder to predict because they may not even include any content words semantically related to the target vocabulary word. Thus the OOV rate is 22%, higher than the other Reading Tutor tasks. To further break down OOV rate by the two vocabulary tasks, defining words has an OOV rate of 22%, whereas comparing which of two words is closer in meaning to a target word has a OOV rate of 19%.

4.4.2 Off-task speech detection result

Figure 4.4 compares the ROC curves for the same classifier on both oral reading and vocabulary activities.

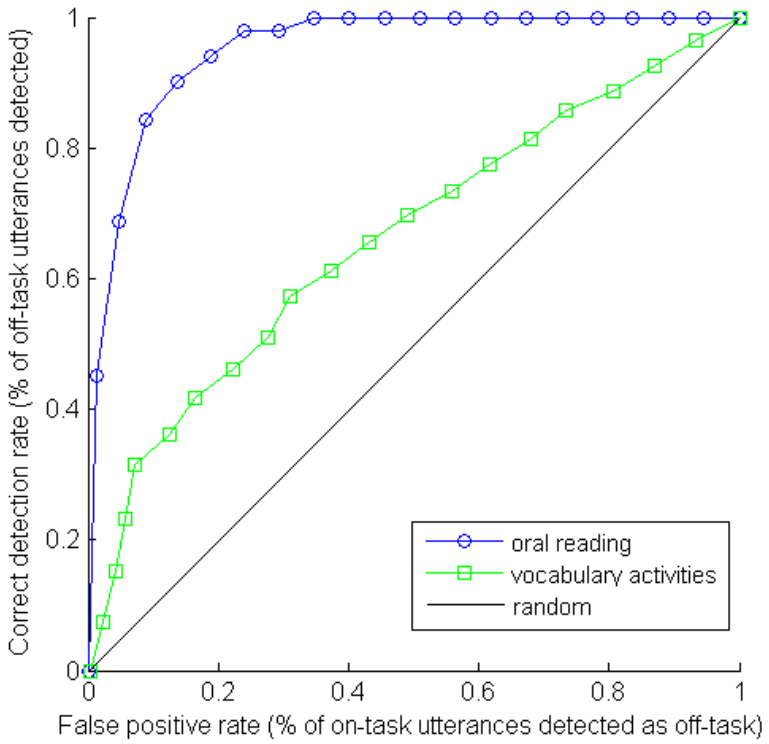

Figure 4.6 Comparison of ROC curves for off-task speech detection on oral reading and vocabulary activities using a classifier trained on oral reading.

Compared to oral reading, detection accuracy is much lower for vocabulary activities, which is again due to the differences in nature of speech prompted in the two tasks. Oral reading and vocabulary tasks are dissimilar in terms of task difficulty for the user. Oral reading requires mostly recognition of words (although expressive reading requires some comprehension). In contrast, explaining the meaning of a word requires both understanding the word and translating that mental representation into speech. This difference in cognitive load is reflected in the percentage of off-task utterances in our data. 34% of the utterances in vocabulary tasks are off-task versus only 12% in oral reading.

4.5 Task difficulty and its relation to the performance of the detector

Transferring the off-task speech detector trained on oral reading data directly to the other three Reading Tutor tasks resulted in decreased but better-than-chance classification accuracy. Figure 4.7 summarizes classification accuracy on the four tasks using the same off-task speech detector. Besides the difference caused by training-testing discrepancy between oral reading and the other test tasks, the three test tasks themselves also exhibit differences in off-task speech detection accuracy. These differences are associated with two types of task difficulties: (1) cognitive load involved during completion of the task, and (2) the difficulty of predicting children's responses to the task prompts.

Higher cognitive load can result in hesitations, incomplete sentences, frequent speech repairs, and even frustration that may lead to off-task speech (Bortfeld, Leon, Bloom, Schober, & Brennan, 2001 ; Merlo & Mansur, 2004; Oviatt, 1995; Shriberg & Lickley, 1993). Cognitive load is partially determined by how much help the child receives. The instruction in the Reading Tutor follows the Duke-Pearson (2002) five-step model. Duke and Pearson (2002) introduced a 5-step instructional model to teach reading comprehension, which gradually transfers task responsibility from tutor to student. The 5 steps are: describing a reading comprehension strategy, demonstrating the use of the strategy, scaffolding the use of the strategy through collaborative use, prompting the use of the strategy, and letting students use the strategy independently. Among the three non-reading activities, self-questioning is the only one that implements scaffolding. The other two tasks merely prompt the child for a response. Therefore, even though self-questioning can involve more complicated cognitive processes than think-aloud, the scaffolding instruction simplifies the task.

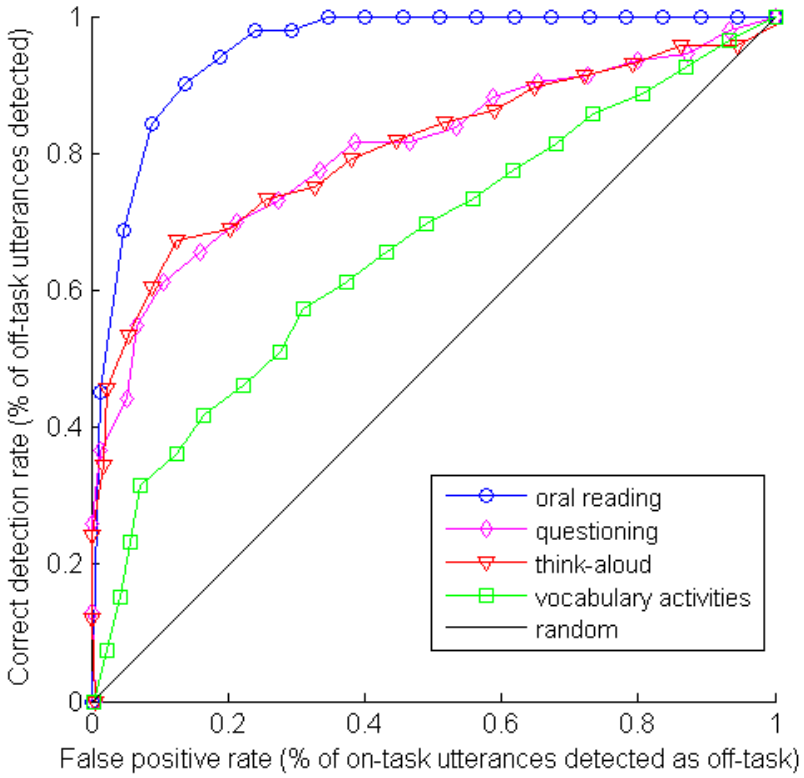

Figure 4.7 Comparison of ROC curves for off-task speech detection of oral reading, questioning, vocabulary activities, and think-aloud using a classifier trained on oral reading.

Classification accuracy of the lexical features varies by at least two factors. First, the property of the task itself bounds the performance of the lexical features. Figure 4.8 compares ROC curves of models trained from lexical features generated from ASR output with an upper bound of the lexical features generated from human transcripts. The relative performance of lexical features on the four Reading Tutor tasks is the same for both cases (i.e., features generated from ASR output and features generated from human transcripts), namely ROC(oral reading) > ROC(questioning) > ROC(think-aloud) > ROC(vocabulary activities). The difference in ROC curves for the four tasks using lexical features generated from human transcript reveals that the relative performance of the lexical features depends on nature of the responses prompted in the

task. For example, in vocabulary activities where children explain the meaning of a word, we observed some high frequency words and phrases that occur in off-task speech as well, such as "you are like ....", whereas in oral reading, this type of casual conversational phrases almost never occur in on-task speech.

Second, the ASR performance affects the performance of the lexical features. As Figure 4.8 illustrates, the ROC curves of models trained from lexical features generated from ASR outputs are worse than those generated from human transcripts, due to inaccurate ASR. The ASR performance depends on both the acoustic model and language model. Since we used the same acoustic models for all four Reading Tutor tasks, the difference comes only from the language model. We use the perplexity to quantify the fit of a language model to a task. Lower perplexity indicates better fit of the language model. The language model perplexity for on-task utterances in oral reading, self-questioning, think-aloud, and vocabulary activities are measured at 5, 582, 405, and 684, respectively. Notice that for oral reading, each sentence has a different language model, so 5 is the perplexity averaged across all the oral reading utterances. Each on-task language model has a small vocabulary size, which on average is 7. We can see that the fit of the language model can roughly predict the performance of the lexical features, except that the language model for think-aloud fits the on-task utterances better than the language model for self-questioning, but the classification accuracy for think-aloud is lower than self-questioning.

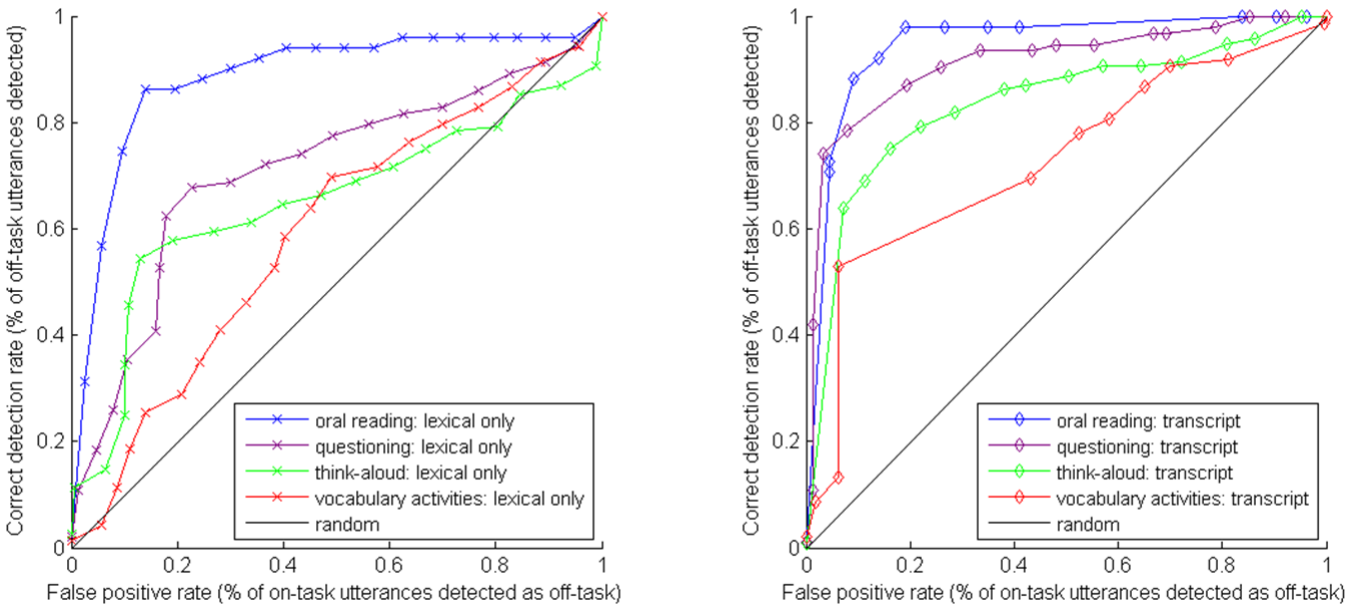

Figure 4.8 Chart on the left: ROC curves of models trained from lexical features generated from ASR outputs. Chart on the right: ROC curves of models trained from lexical features generated from human transcripts for the four Reading Tutor tasks.

To further examine the importance of ASR accuracy, we relate the ASR accuracy on on- and off-task speech to the detector accuracy, that is, when ASR is perfect for off-task speech, how much improvement we can see, and when ASR is perfect for on-task speech, how much improvement we can see. The goal of this analysis is to find out if there is any difference in the importance of ASR accuracy between on- vs. off-task speech. Figure 4.9 illustrates how predictive lexical features generated from different sources are. For each Reading Tutor task, we compare lexical features generated under four conditions: when realistic (using ASR output), when ASR is perfect for off-task speech, when ASR is perfect for on-task speech, and when ASR is perfect for both types of speech. For think-aloud and vocabulary activities, perfect ASR for off-task speech generates significantly more accurate classifiers than perfect ASR for on-task speech ($p=0.01$ for

think-aloud, p=8e-6 for vocabulary activities). As the figure illustrates, for detecting off-task speech, ASR needs to be accurate not only on on-task speech, but on off-task speech as well.

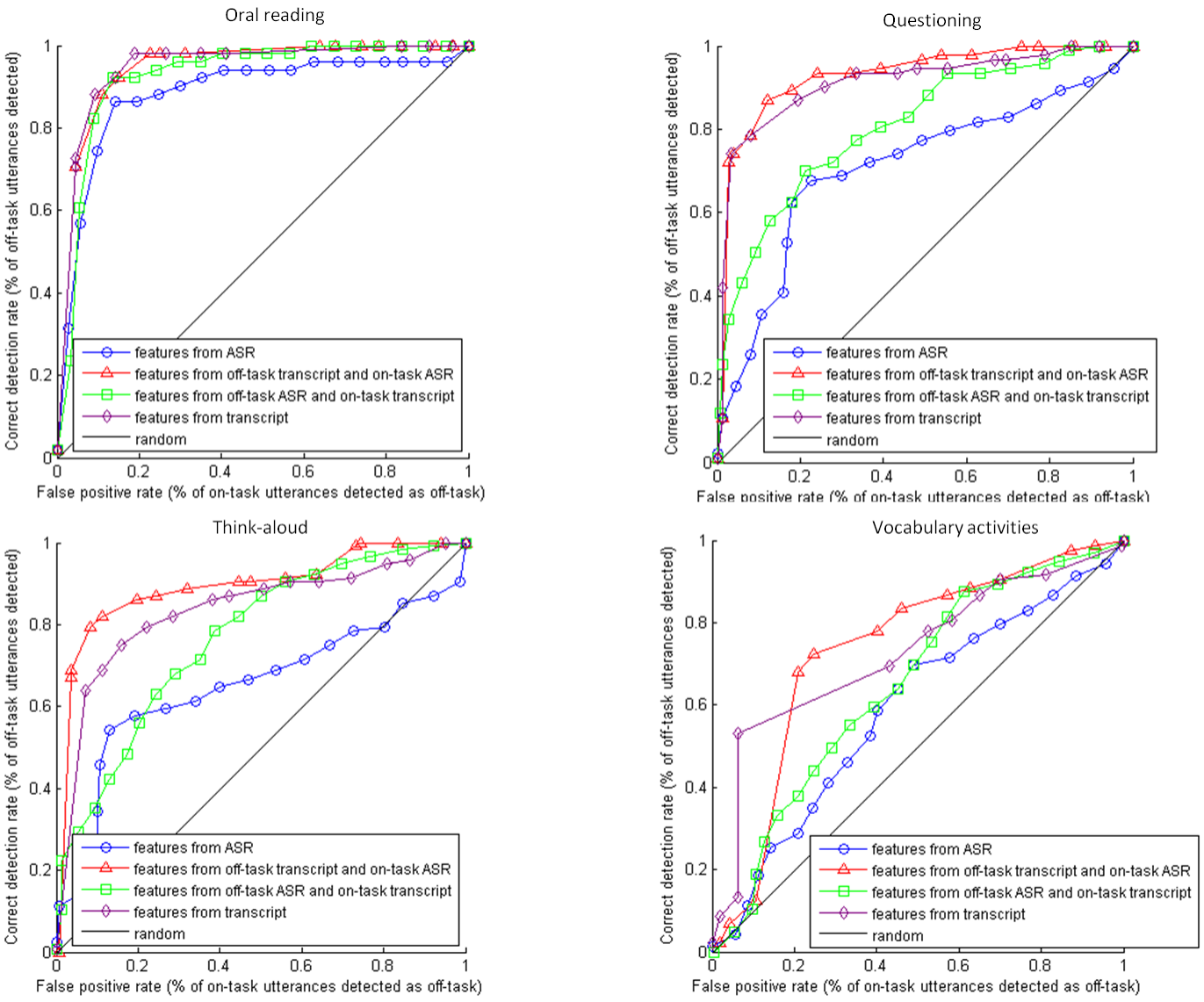

Figure 4.9 Comparison of the importance of ASR accuracy for lexical features. Lexical features are generated from four sources: pure ASR output, transcript for off-task speech and ASR output for on-task speech, transcript for on-task speech and ASR output for off-task speech, and pure transcript.

Figure 4.10, Figure 4.11, and Figure 4.12 summarize ASR performance for the four tasks, broke down by on- and off-task speech. The OOV rate for vocabulary activities is much higher than the other tasks, which leads to worse ASR performance than the other tasks as well. Even though the OOV rate for off-task speech is only moderately higher than for on-task speech (and even lower than on-task speech for vocabulary activities), the recognition error for off-task speech is much higher than for on-task speech. ASR struggles with spontaneous speech. The short words frequent in off-task speech such as *I*, *you*, and *'cause* also increase errors in ASR.

Table 4.2 ASR performance for the four Reading Tutor tasks.

		Oral reading	**Self-questioning**	**Think-aloud**	**Vocabulary activities**
OOV	**Overall**	3%	7%	7%	20%
	On-task	3%	7%	5%	22%
	Off-task	6%	8%	7%	15%
WER	**Overall**	32%	87%	88%	96%
	On-task	26%	82%	82%	93%
	Off-task	91%	94%	97%	99%
Recognition accuracy	**Overall**	78%	31%	26%	33%
	On-task	85%	32%	27%	41%
	Off-task	20%	27%	22%	26%

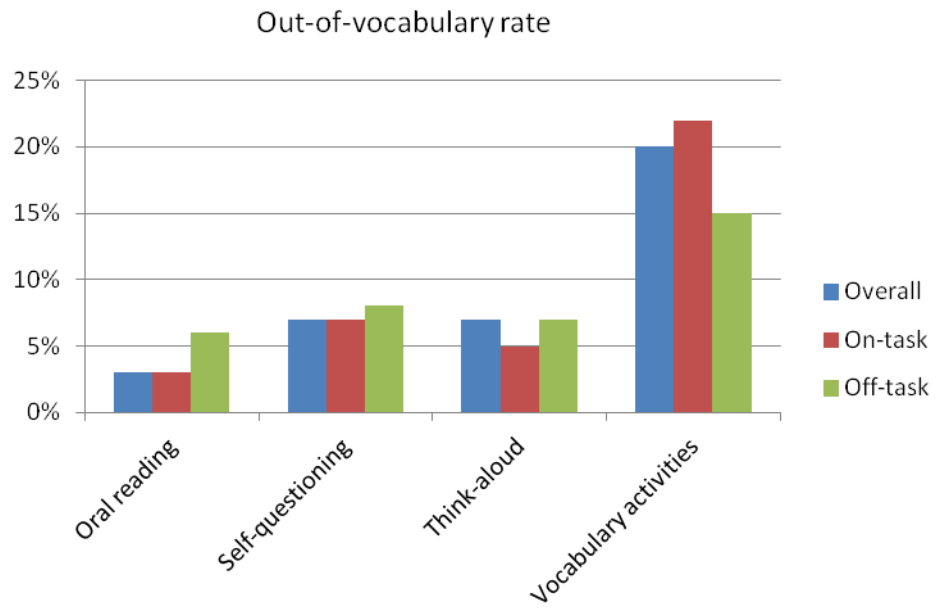

Figure 4.10 OOV rate in oral reading, self-questioning, think-aloud, and vocabulary activities.

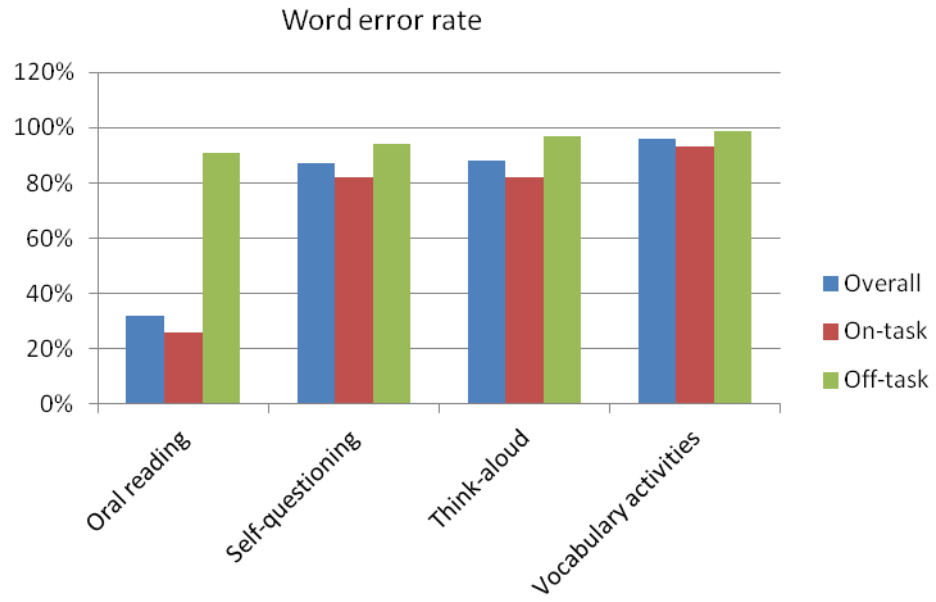

Figure 4.11 WER in oral reading, self-questioning, think-aloud, and vocabulary activities.

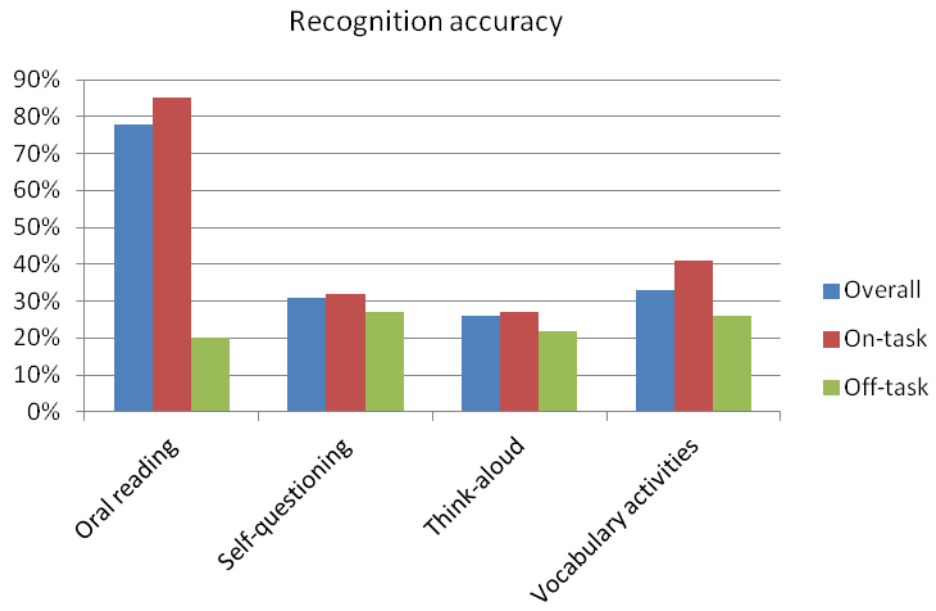

Figure 4.12 Recognition accuracy in oral reading, self-questioning, think-aloud, and vocabulary activities.

5. The role of features in characterizing off-task speech

This chapter analyzes the role and generality of acoustic, lexical, and contextual features in detection of off-task speech across different tasks.

5.1 Acoustic features

Acoustic features characterize the speaking style of off-task speech, i.e., how it was said.

5.1.1 Speaking style of off-task speech

We extracted 5 groups of acoustic features: energy, spectrum, cepstrum, voice quality, and others. Some of the features are more intuitive than others, such as energy features, which we perceive as loudness. Others are less intuitive, even though perceivable, such as features extracted from the spectrogram. To describe the speaking style of off-task speech, we choose four types of intuitive acoustic features in the top 45 features selected by AdaBoost learning algorithm on oral reading data. Table 5.1 lists the four types of features occurring in the selected 45 features, their weights, and the rank of their weights. Compared to oral reading, average pitch in off-task speech was higher by 11%, and average loudness lower by 16%. In addition, off-task utterances averaged 10% shorter than oral reading, and had a lower harmonics-to-noise ratio. In Section 5.4 we will discuss how well acoustic features predict each types of off-task speech.

Table 5.1 Four types of intuitive acoustic features.

Feature type	Feature name	Feature weight	Feature rank
Pitch	Pitch – 3rd quartile	0.041	1

	Pitch – median	0.012	5
	Pitch – 3rd moment	-0.0035	28
	Pitch – minimum	-0.0025	35
Energy	Loudness (voiced) – 3rd moment	-0.013	4
	Loudness (unvoiced) – minimum	-0.0099	7
	Loudness (silence) – 1st quartile	-0.0048	22
	Loudness – 3rd quartile	0.0034	29
Duration	Utterance duration	-0.0088	10
Voice quality	Shimmer	0.010	6
	Voice break rate	0.0096	8
	Harmonics to noise ratio – mean	-0.0080	12
	Harmonics to noise ratio – 4th moment	-0.0060	17

	Harmonics to noise ratio – 3rd quartile	-0.0022	38

To further investigate the role of pitch, energy, duration, and voice quality features on off-task speech detection, we train and test off-task speech detectors on oral reading using each single type of feature and compare their classification performance against the full fledged detector. Figure 5.1 shows this comparison. Although the top one feature is a pitch feature, pitch features in general did not show an advantage over other types of features. In particular, voice quality features perform almost as well as pitch features.

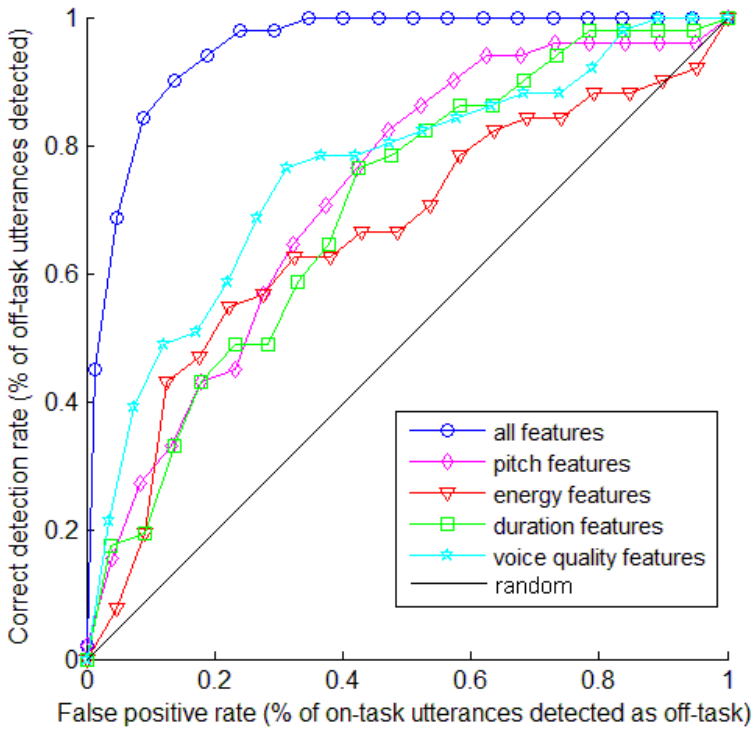

Figure 5.1 ROC curves of off-task speech detectors using different types of acoustic features on oral reading.

To test the generality of each of the four types of acoustic features, we run the classifiers trained using only one type of features on all four Reading Tutor tasks described previously. Figure 5.2 shows the classification performance of pitch, energy, duration, and voice quality features on each Reading Tutor task. Duration is the most consistent feature on all four tasks. Pitch features perform well on oral reading and think-aloud, but not on vocabulary activities. Energy features are good at characterizing off-task speech in think-aloud and questioning. Overall the four types of acoustic features perform consistently well on think-aloud utterance, but not on utterances in vocabulary activities.

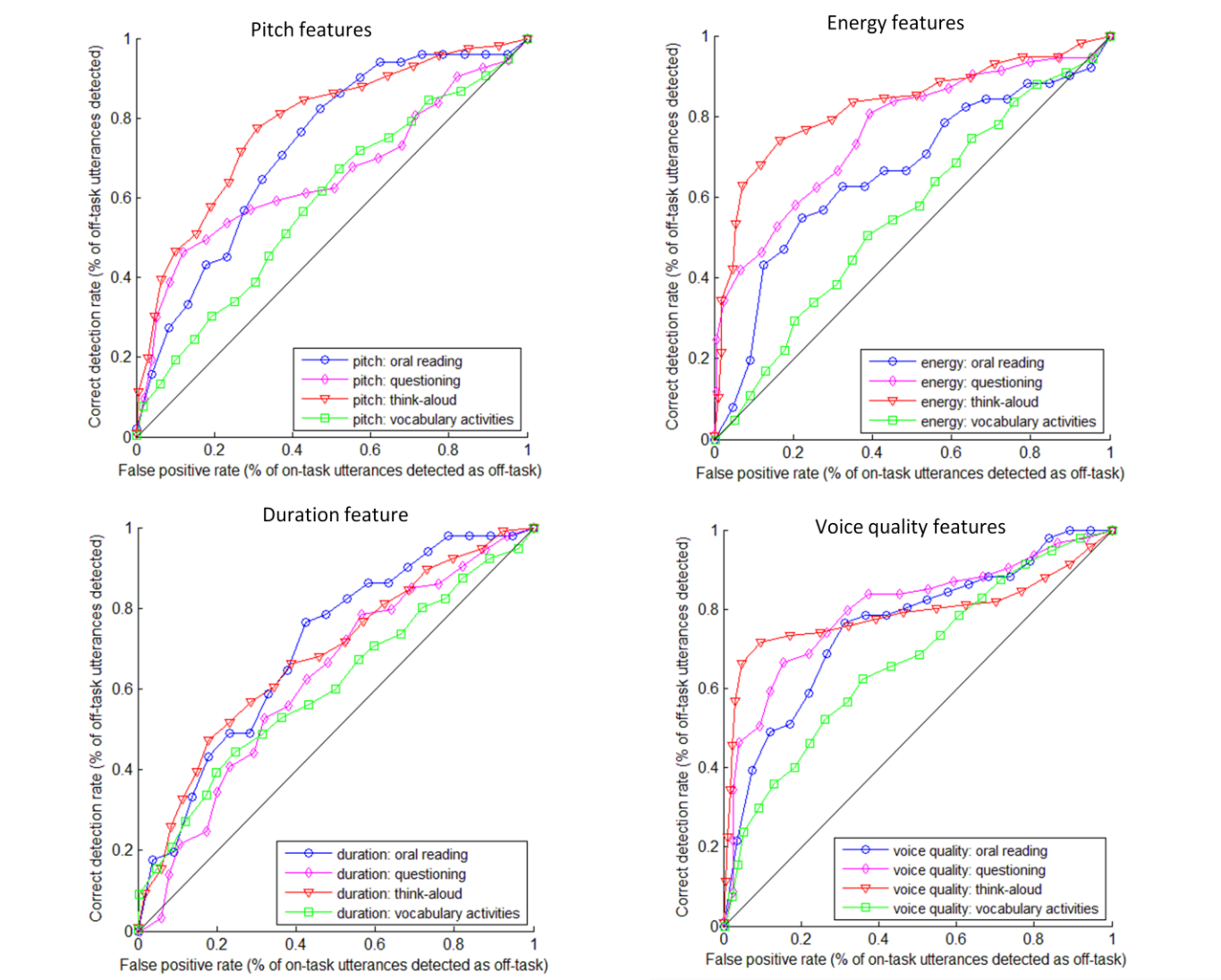

Figure 5.2 Classification performance of pitch, energy, duration, and voice quality features on four Reading Tutor tasks.

5.1.2 Generalization of acoustic features

How well do acoustic features work in general for detecting off-task speech? To answer this question, we train an off-task speech detector on oral reading using only acoustic features and compare its classification performance with the full-fledged detector with acoustic, lexical, and contextual features. To observe the generality on tasks other than oral reading, we also test the detector with only acoustic features on other Reading Tutor tasks. Figure 5.3 demonstrates the

comparison. Surprisingly, the detector with only acoustic features even significantly out-performed the full-fledged detector on think-aloud (we performed a chi-squire test on the outputs of the two classifiers, where $p < 0.01$), which indicates that (1) there is overfitting during training; (2) acoustic features of off-task speech in think-aloud have stronger predictive power than lexical features.

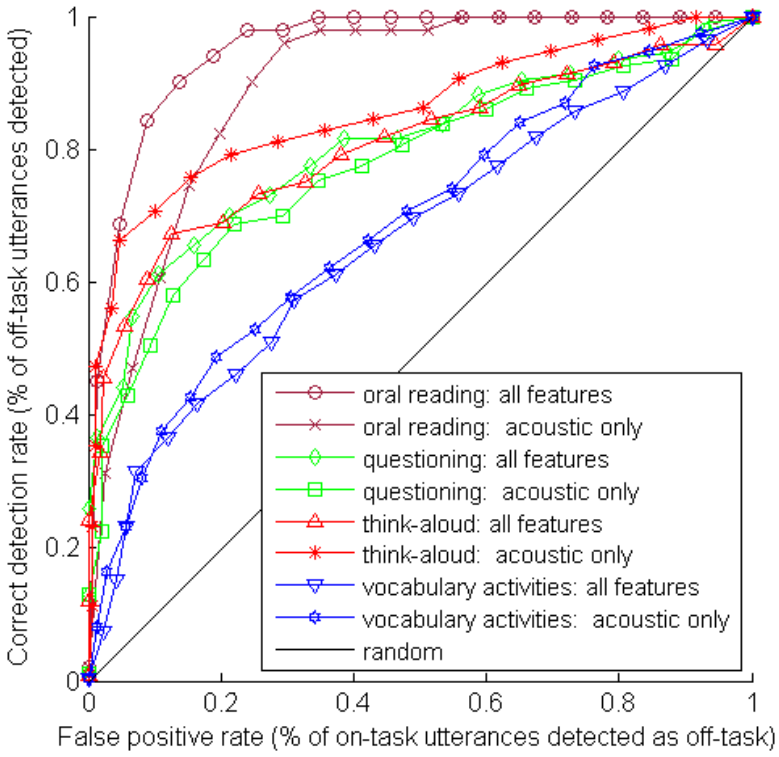

Figure 5.3 Comparison of acoustic feature performance on all four Reading Tutor tasks.

5.2 Lexical features

Lexical features characterize the content of an utterance, i.e., what was said.

5.2.1 Individual predictive power of lexical features

We used four lexical features: percentage of off-task words, percentage of off-task words where the ASR is confident, percentage of on-task words that the ASR is not sure of, and ratio of the number of silences to the number of words. Figure 5.4 breaks down the discriminative power of lexical features on oral reading and compares it to the complete detector. The most informative feature is the percentage of off-task words. Features with confidence thresholding are less strong, suggesting that the confidence scores are noisy.

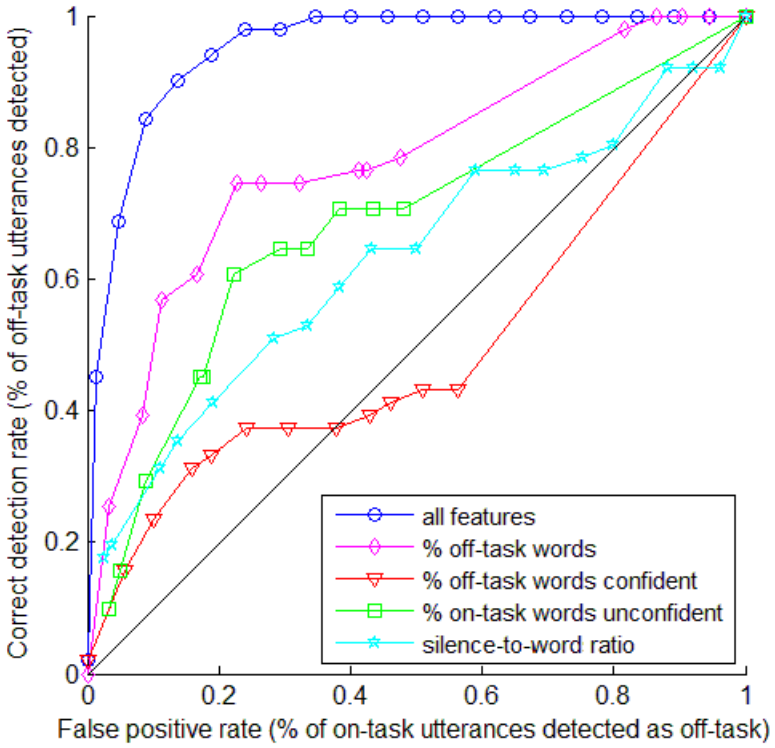

Figure 5.4 Performance of each lexical feature on oral reading.

Figure 5.6 shows the generality of each lexical feature across the four Reading Tutor tasks. Percentage of off-task words exceeding a confidence threshold is the worst feature on all four tasks, sometimes worse than chance. However, combining percentage of off-task words

confident with percentage of off-task words sometimes improves classifier accuracy marginally, as illustrated in Figure 5.7. As shown by the box plots of confidence scores in Figure 3.5, at least ¼ of the words, both recognized correctly and incorrectly, were assigned a single confidence score = 0. To find out how informative confidence scores are, in addition to the classification results illustrated in Figure 5.6 and Figure 5.7, we performed the Mann-Whitney U test on two data sets from children's oral reading. The first data set consists of the mean confidence scores for each off-task utterances. The second data set consists of the mean confidence scores for each on-task utterances. The mean confidence score for an utterance is the arithmetic average of the confidence scores for the words in the utterance. In addition to the mean, we performed the Mann-Whitney U test on median and variance of confidence scores as well. The statistical test showed significant difference in median, mean, and variance of confidence scores for off- vs. on-task utterances ($p<0.01$ for median, mean, and variance). Confidence scores in off- and on-task speech share the same median (= 0). Off-task utterances have a lower average confidence score (-4012) than on-task utterances (-1933), and the variance in confidence scores for off-task utterances is larger (1.5e8) than that of on-task utterances (1.2e8). The three statistics on confidence scores can classify off- vs on-task utterance, as illustrated by Figure 5.5. Because off-task speech and on-task speech share the same median confidence score, confidence thresholding can be tricky and may lose some distributional information in confidence scores.

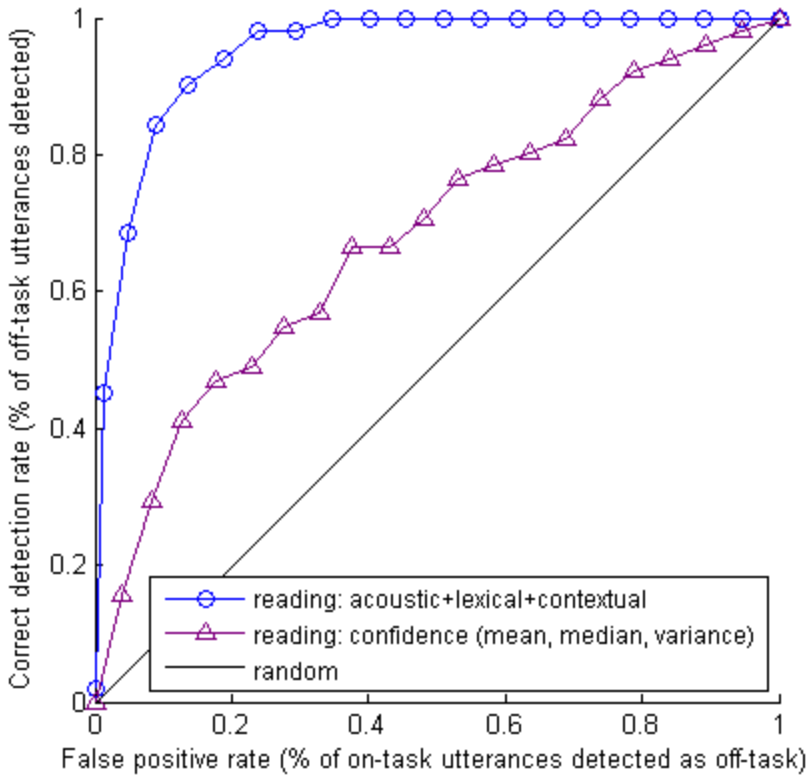

Figure 5.5 Confidence scores demonstrate a positive role in detecting off-task speech in children's oral reading.

Since the detectors are trained on oral reading utterances, the features in general work better on oral reading utterances as well. However, the silence-to-word ratio performs best on think-aloud, compared to other lexical features. Although expressed in terms of the number of words rather than the percentage of the signal, this feature is quite similar to acoustic features in that it also characterizes how the speech sounds, i.e., how often the speaker pauses.

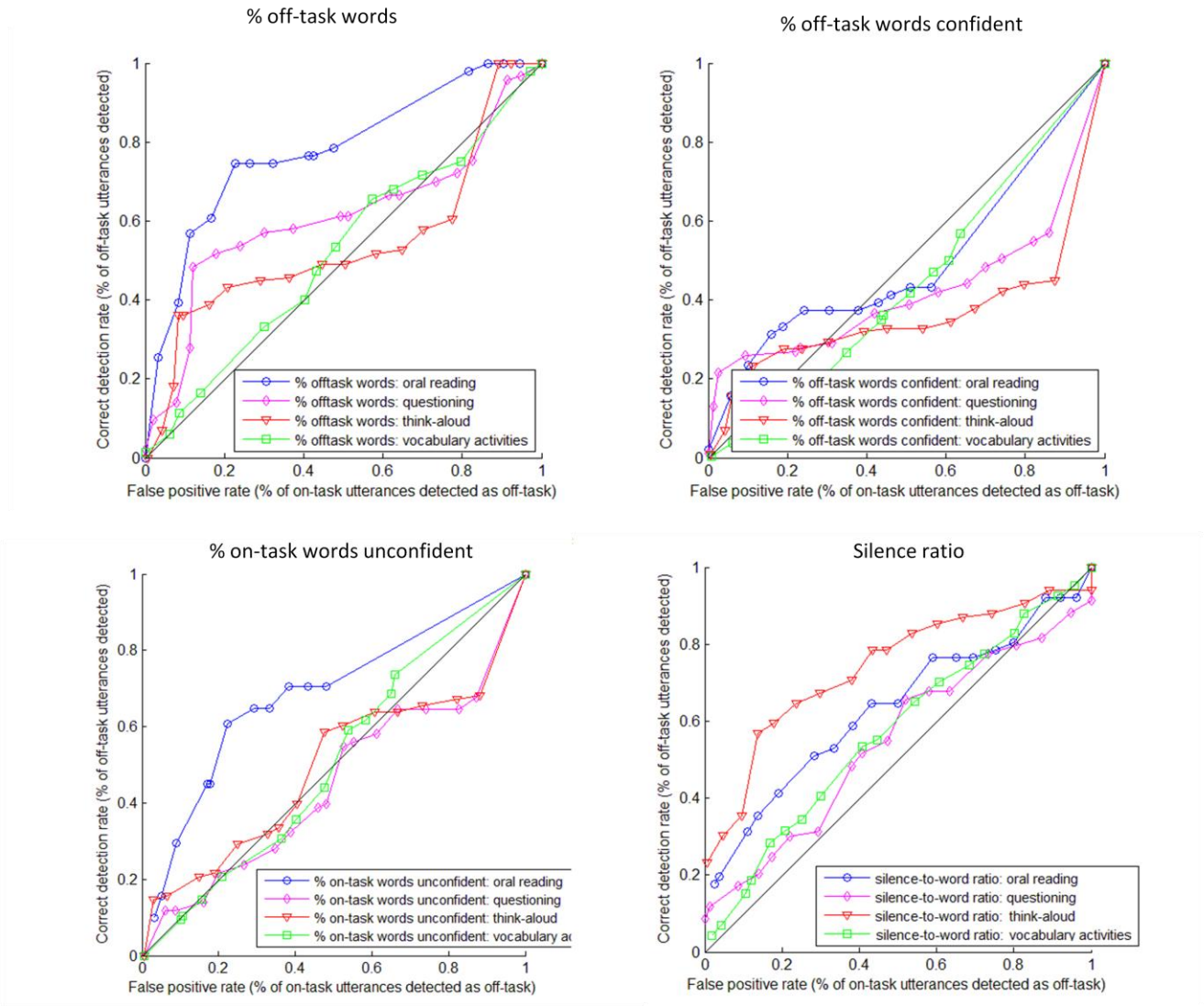

Figure 5.6 Generality of each lexical feature among the four Reading Tutor tasks.

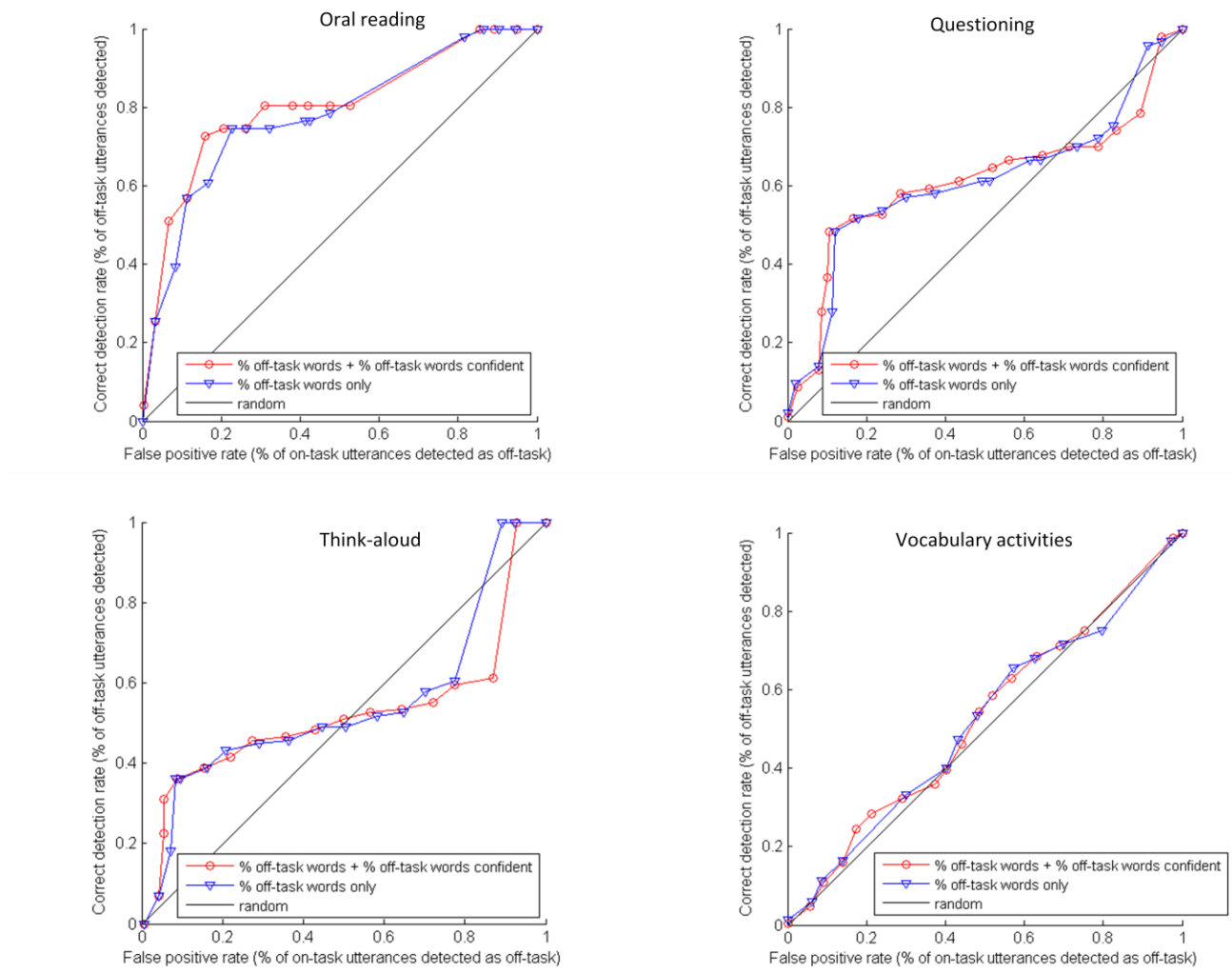

Figure 5.7 The effect of adding percentage of off-task words confident to percentage of off-task words.

5.2.2 Generalization of lexical features

To investigate the role of lexical features as a whole, we train an off-task speech detector using only lexical features and compare its performance on all four Reading Tutor tasks. Figure 5.8 summarizes the comparisons. Unlike the detector trained using acoustic features, the detector trained using only lexical features consistently performs worse than the complete detector. This disadvantage is especially pronounced on think-aloud utterances. Compared to self-questioning, which has similar overall detection accuracy as think-aloud, the lexical features of think-aloud

perform much worse. As Figure 4.11 and Figure 4.12 showed, self-questioning has a lower WER and higher percentage of words correctly recognized than think-aloud. Hence the lexical features appear to be more helpful when there is a language model that fits the task better.

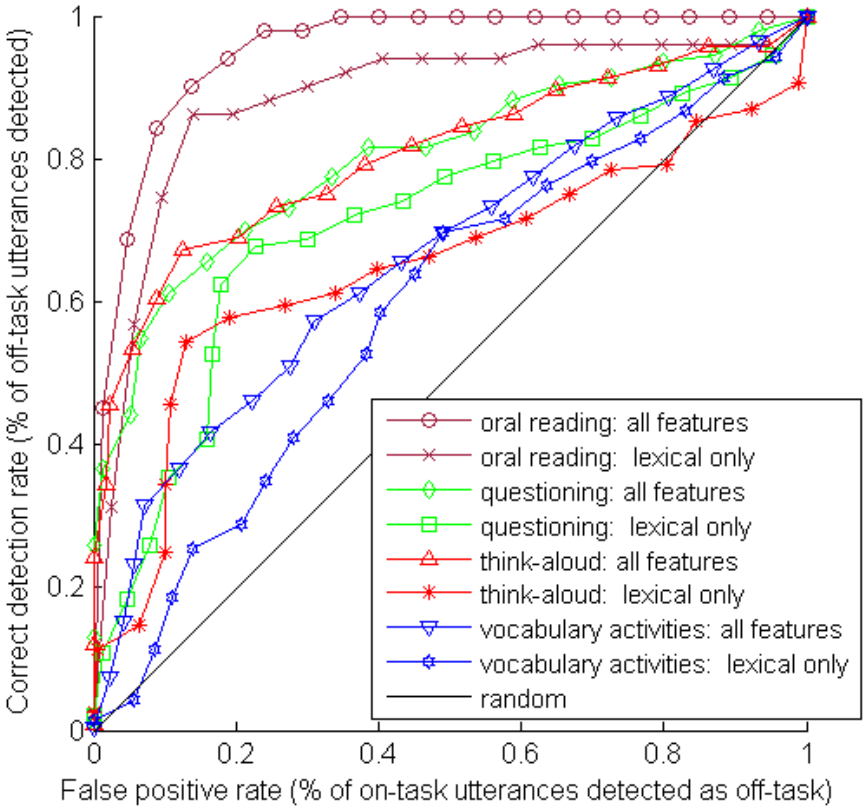

Figure 5.8 Comparison of classification performance between detector trained using only lexical features and the full detector.

5.3 Contextual features

Our assumption in using contextual features is that off-task speech will "stand out" in the majority of on-task speech. If this assumption is true, the contextual features ought to improve accuracy for detecting off-task speech.

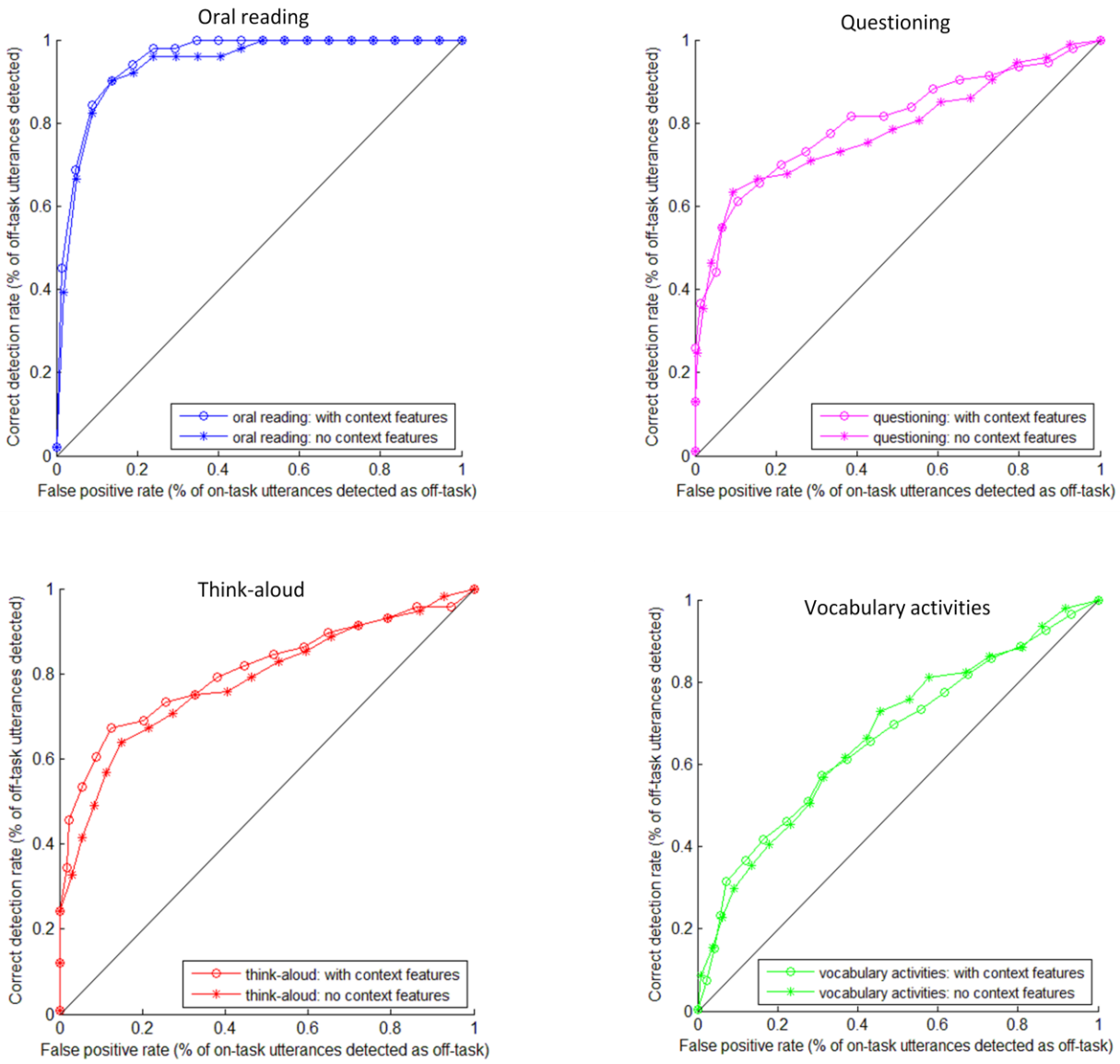

Figure 5.9 Comparison of detectors with and without contextual features.

As Figure 5.9 shows, in oral reading, questioning, and think-aloud, contextual features indeed improve classification accuracy. The improvement in oral reading and think-aloud are significant ($p < 0.05$ by a Chi-square test). For vocabulary activities, contextual features do not improve classifier performance. Vocabulary activities are the most difficult tasks for detecting off-task speech, as shown previously. The acoustic and especially lexical features we computed

may be less informative than for the other tasks. Therefore the calculation of the running average may not be accurate either, which could have hindered the performance of contextual features.

5.4 Roles of features in different types of off-task speech

Off-task speech contains various types. According to the taxonomy described in Introduction, we can categorize it roughly into null responses (silence, singing and other non-spoken events), general off-task speech, and task-related off-task speech. Figure 5.10 shows the break-down of off-task speech in each of the four Reading Tutor tasks. Due to previous processing operations which had filtered out null responses from our database, oral reading does not contain null responses. This does not mean that oral reading has no null responses.

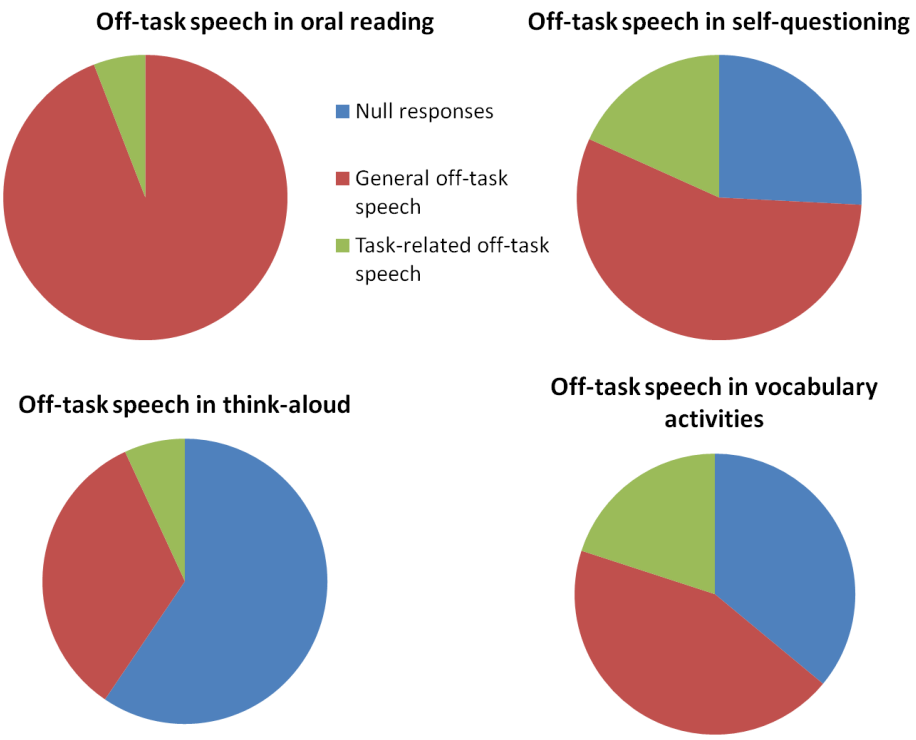

Figure 5.10 Break-down of off-task speech in four Reading Tutor tasks.

To evaluate the roles of acoustic and lexical features in classifying different types of off-task speech, we break down classification accuracy by types of feature and types of off-task speech. We compute classification accuracy of a class X as $\frac{number\ of\ X\ correctly\ classified}{number\ of\ true\ X}$. We use the classifier trained from off-task speech weighted 8 times higher than on-task speech in oral reading. As Figure 5.11 illustrates, for the most cases, lexical features work better at detecting general and task-related off-task speech. Acoustic features detect more null responses. However, when ASR is inaccurate, such as for vocabulary activities, acoustic features detect more general and task-related off-task speech.

Figure 5.11 shows the percentage of off-task utterances detected in each type of off-task speech. For each reported percentage, the false positive rate is fixed at 10%. The percentages should not be over interpreted as "recall." The detector does not distinguish among the three types of off-task speech. The chart is intended only at analyzing the percentage of each type of off-task speech detected. It does not show how accurately the detector can distinguish each type of off-task speech from other types of off-task speech, nor does it capture how accurately the detector can distinguish each type of off-task speech from on-task speech.

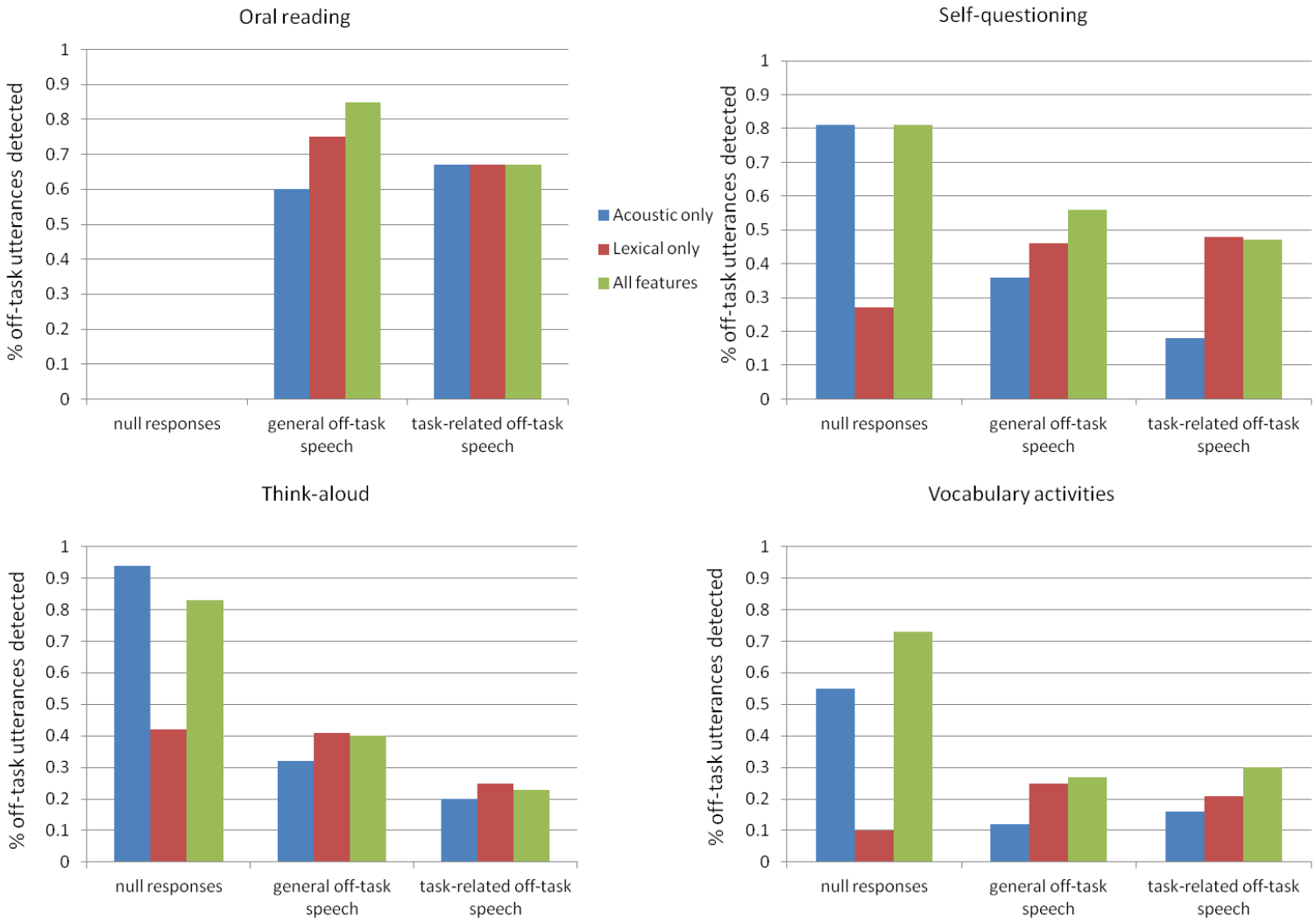

Figure 5.11 Predictive power of acoustic and lexical features on various types of off-task speech.

5.5 Summary

This chapter analyzed the roles of features in detecting off-task speech by four Reading Tutor tasks: oral reading, self-questioning, think-aloud, and vocabulary activities. Acoustic features capture the speaking style of off-task speech. To describe the speaking style, we broke down classifier performance by four groups of acoustic features: pitch, energy, duration, and voice quality features. None of the four groups of features demonstrated significant advantage over the other groups of features. Off-task speech tends to have higher pitch than oral reading. Its

loudness changes less dramatically, and it tends to be shorter, and less intelligible (by voice quality features) than oral reading. Lexical features capture the linguistic content of off-task speech. The most informative lexical feature is the percentage of off-task words. Performance of lexical features strongly relies on the ASR accuracy. When ASR is inaccurate, lexical features are less robust than acoustic features. Contextual features capture the stand-alone characteristic of off-task speech. Adding contextual features helps improve accuracy for off-task speech detection.

Acoustic and lexical features are each more suitable for detecting different types of off-task speech. Lexical features are more powerful for detecting general and task-related off-task speech, whereas acoustic features detect more null responses, including singing, humming, mumbling, and silences.

6. Example applications of off-task speech detection

All applications involving verbal input may use off-task speech detection to track attention of users or filter out undesired task-irrelevant utterances. This chapter uses two examples to demonstrate how off-task speech detection can help improve other applications. The two example applications include one Reading Tutor application and one non-Reading-Tutor application that involves adult telephone speech.

6.1 Improving prediction of fluency test scores for oral reading

During the summer of 2006, Virnik and Mostow (2006) re-designed a teacher tool (Alpern, Minardo, O'Toole, Quinn, & Ritzie, 2001) to help teachers gain knowledge about their students. As part of the tool, they developed a method that uses data logged in the Reading Tutor to estimate children's reading rate by the end of a school year. Human testers provided the fluency test scores for each student. Fluency test scores in oral reading are defined by *word correct per minute* (WCPM), which is the average number of words read one-minute. Mostow and Virnik normalized the fluency test score by multiplying it by the average of word lengths in the reading level of the child, and dividing the score by 60 seconds to obtain a letter-per-second (LPS) rate. Higher grade levels use more difficult texts, which contain longer words on average. To predict this normalized fluency score, they calculated the average of LPS reading rate using the data collected during the school year, with some heuristics (Beck, Jia, & Mostow, 2004) to filter out data likely to distort the prediction. Their goal was to develop an algorithm for measuring a reading rate in LPS that correlated most strongly with the normalized fluency score. The evaluation criterion is the correlation coefficient R between the computed LPS and the

normalized fluency score, given a linear regression model that minimizes the squared error of the prediction. The data come from 60 children ages 7-10 in grades 2-4 who used the Reading Tutor in the 2005-2006 school year.

Virnik and Mostow (2006) mentioned that off-task speech may explain some of the variation in the data that distorted fluency predictions. To study whether this was the case, we apply off-task speech detection to filter out off-task utterances from the data and compare the resulting fluency prediction with the highest R = 0.891 achieved by Mostow and Virnik, which we use as the baseline.

To get the baseline result of R (0.891), Virnik and Mostow computed LPS reading rate as the following. First they collected information about each word read, including the child ID, the date, the start and end time of the reading, as well as if the child clicked the word for help, and if the word was accepted by the ASR. To exclude effects of recency, Mostow and Virnik selected only the first encounter of each word on a given day. To reduce the number of outliers that may distort fluency estimation, they winsorized per-word reading time to 5 seconds. They also excluded words rejected by the ASR and words on which the child clicked for help. Notice that the baseline R = 0.891 is a strong result, considering the correlation between two hand-coded fluency tests is often lower than 0.9 (Good & Jefferson, 1998).

To improve the correlation between estimated LPS and normalized fluency test scores, we apply automatic off-task speech detection to the oral reading utterances and filter out words in utterances classified as off-task. Notice that words read are found by ASR, not by humans. Therefore the ASR may have misrecognized some of the off-task speech as oral reading. To achieve a high detection rate, we use the off-task speech detector trained from off-task utterances

weighted 12 times higher than oral reading utterances. On the test data described in section 3.2.2, the detector correctly detected 94% of the off-task utterances and falsely classified 15% of oral reading utterances as off-task. On the fluency prediction data, this detector identified 63,393 off-task utterances to exclude from the total of 171,854 utterances. The detector classified more than 1/3 of the utterances as off-task, many of which can be false positives. However, to filter out off-task speech as completely as possible, we weigh detection rate heavier than false positives.

The correlation R increased from 0.891 to 0.894 after filtering out off-task speech. However, this improvement is only marginal ($p = 0.4$), according to a test of significance on two dependent correlations (Blalock, 1972).

Notice that filtering out off-task utterances also reduces the size of data, which may reduce predictor accuracy. In fact randomly removing the same number of utterances as we detected (63,393) decreased R from 0.894 to 0.886.

The baseline method did not consider speech unaligned with any target words, which may exclude some off-task speech. Besides, the baseline method may have filtered out additional off-task speech by excluding words rejected. Therefore applying additional off-task speech filtering did not achieve significant improvement in fluency prediction. In particular, without the constraint that the word has to be accepted as read correctly, applying off-task speech filtering would have increased R from 0.864 to 0.875, which is a statistically significant difference ($p < 0.01$). Notice that adding the constraint to include only words accepted by the ASR as read correctly provides stronger correlation than filtering off-task speech (0.891 vs. 0.875). However, this difference is not statistically significant ($p = 0.14$), which suggests that

applying off-task speech detection and filtering out words rejected by the ASR can improve fluency prediction by similar degrees.

6.2 Improving utterance understanding in the Let's Go corpus

An essential component of the CMU Let's Go bus information system is natural language understanding (NLU). The responsibility of NLU is to help identify the intent of an utterance, namely the *dialog act*, from the ASR output. The NLU component in the CMU Let's Go bus information system uses the Phoenix parser (Ward & Issar, 1994) to identify key concepts such as the name of a bus stop. These key concepts determine dialog acts. One challenge for NLU is that its input (i.e., the ASR hypotheses) can contain errors, and the error rate is especially high in realistic environments. Besides, Raux et al. (2005) found that certain speaking styles such as frustration are associated with low understanding rates. Both these problems may have some connection with off-task speech.

This section reports the effect of off-task speech detection on understanding user utterances in the Let's Go system. Instead of applying the off-task speech detector in real time to the system, we run the detector off-line on utterances in the Let's Go corpus and see if detecting off-task speech would have improved the understanding rate on the utterances (calculated as the fraction of user utterances correctly understood over the total number of user utterances).

6.2.1 Detecting off-task speech in the Let's Go corpus

The Let's Go corpus contains 1,464 recorded calls to the Let's Go bus information system. It includes user utterances, hand transcriptions of the utterances, ASR outputs for the utterances, files summarizing system and user turns in the dialogs, and log files from the system. To detect off-task speech, we obtained 1,170 dialogs that contained at least one recorded user utterance,

the transcription, and the ASR output for the utterance. To train an off-task speech detector on this data, we randomly chose 1,000 dialog sessions as training data, and the other 170 as test data. The training data contains 6,606 utterances with transcriptions and ASR output, of which we labeled 795 (12%) as off-task. The test data contains 1,067 utterances with transcriptions and ASR output, of which we labeled 107 (10%) as off-task. Off-task speech in the Let's Go corpus includes complaints about the system (e.g., "I'm trying it's not working"), talking to oneself or other people (e.g., "it says what can I do for you"), and null responses (sighing, mumbling, silence, etc.).

We computed the same 45 acoustic features and 4 lexical features for each utterance as for the Reading Tutor tasks. We extracted the set of off-task words from the transcriptions in training data. Since there were only 322 words which are more frequent (percentage-wise) in off-task speech than in on-task speech, we included all the 322 words in our list of off-task words. We ranked the off-task words by their frequency difference in off- vs. on-task speech. The ones likelier to occur in off-task speech are ranked higher. Although only 141 off-task words actually appeared in the ASR output in the Let's Go corpus, the 20 off-task words ranked highest all got recognized by the speech recognizer in the Let's Go system. Table 6.1 shows the 10 words with largest difference in frequencies in off- and on-task speech. Words underlined also appeared in the top 10 off-task words in children's oral reading.

Table 6.1 10 words with largest difference in frequencies in off- and on-task speech in the training data of the Let's Go corpus.

Off-task word	Frequency difference in off- vs. on-task speech (%)
you	2.44
it	1.71
are	1.20
still	1.06
don't	0.94
me	0.93
not	0.91
it's	0.90
okay	0.89
there	0.81

Figure 6.1 shows the test result for detecting off-task speech on the Let's Go corpus. In particular, we break down the test result by feature types. Although we adapted the set of off-task words to the Let's Go corpus, lexical features are still less informative compared to acoustic features. Notice that we did not select acoustic features for detecting off-task speech on the Let's

Go corpus. Instead, we directly applied features selected based on children's oral reading. This result suggests that the difference between off- and on-task speech in speaking style may be more robust than lexical content.

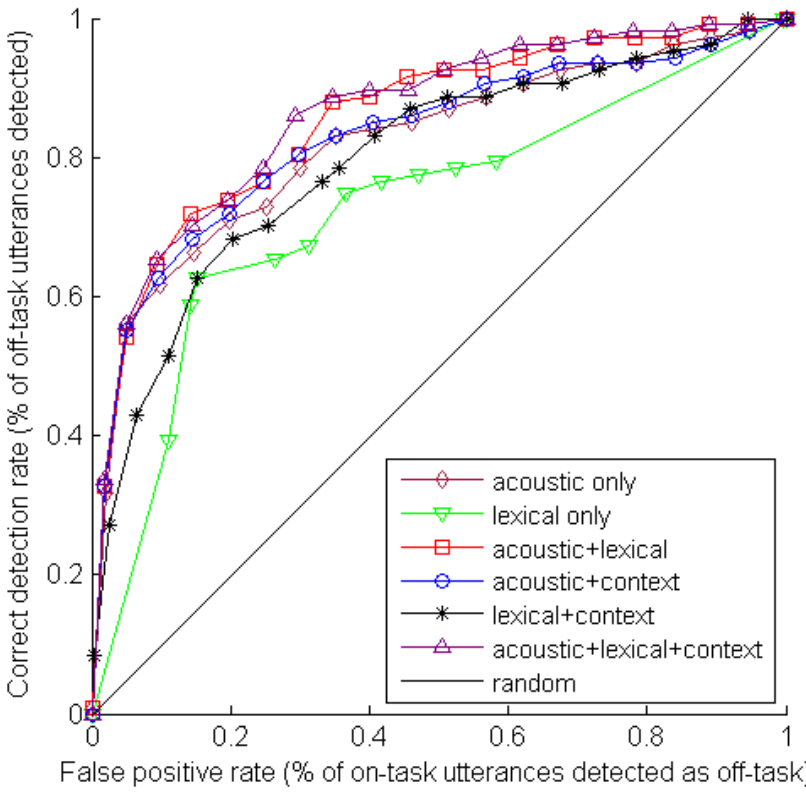

Figure 6.1 Predictive power of features on the Let's Go corpus.

Because feature extraction for telephony speech can be different from feature extraction for speech collected by a microphone, the salient features for telephony speech may be different from those extracted from microphone speech. In order to find out salient acoustic features on the Let's Go data, we re-ranked acoustic features on the training data from the Let's Go corpus and selected 45 features with largest weights assigned by the AdaBoost algorithm. Table 6.2 shows the top 10 acoustic features with largest weights. The features occurred in both lists of 45

salient features selected from children's oral reading and the Let's Go corpus include pitch (minimum, median, and 3rd moment), loudness (minimum), HNR (1st quartile), and shimmer.

Table 6.2 The top 10 acoustic features with largest weights assigned by running the AdaBoost algorithm on the Let's Go data.

Feature name	**Weight**
Intensity – variance	0.028
F2 – variance	0.016
Shimmer	0.014
HNR – 1st quartile	0.013
2nd MFCC (voiced) – median	0.013
10th MFCC –3rd quartile	0.012
3rd MFCC (unvoiced) – 1st quartile	-0.011
Jitter	0.011
MFCC 208	0.010
Ltas – 3rd moment	0.010

Figure 6.2 shows the ROC curves of performance of two sets of acoustic features tested on the Let's Go data. One set of acoustic features were selected from children's oral reading, and the other set of acoustic features were selected from the Let's Go training data. Re-ranking the acoustic features on the Let's Go data improves the detection accuracy. However, the difference is not statistically significant, according to a Chi-square test ($p = 0.08 > 0.05$).

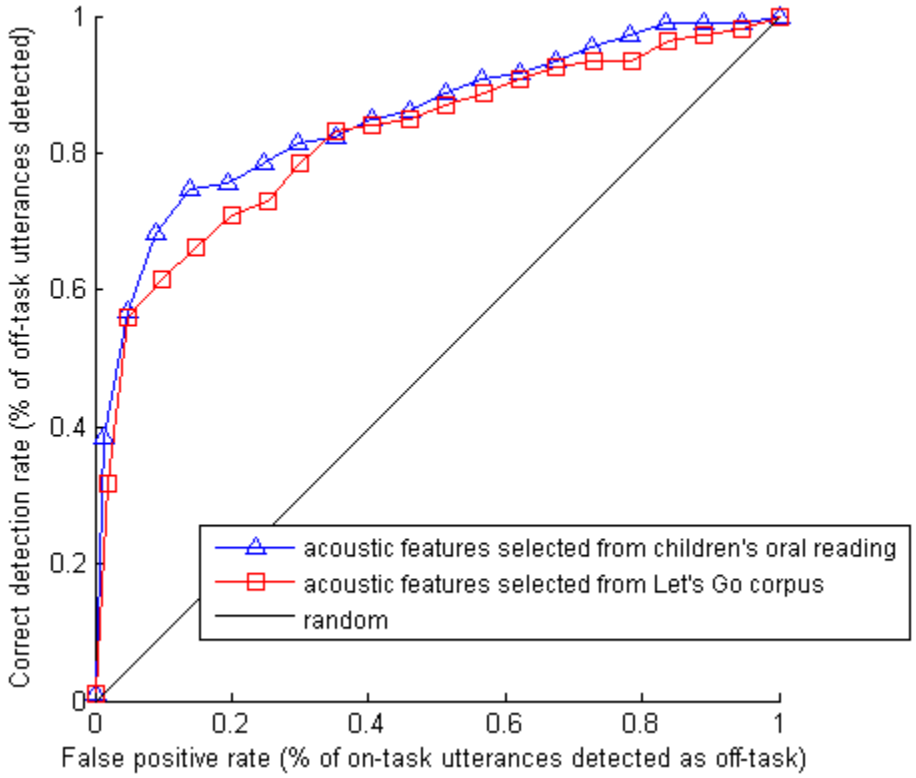

Figure 6.2 ROC curves of acoustic features selected from children's oral reading and the Let's Go corpus, respectively.

6.2.2 Improving understanding rate on user utterances in the Let's Go corpus

The Let's Go bus information system does not explicitly deal with off-task speech, although the Sphinx-II speech recognizer and the Phoenix parser reject some off-task speech. Off-task speech

can cause the speech recognizer to make errors, and the ASR errors can later mislead the parser and cause it to map the off-task utterance to some on-task dialog acts. We define *utterance-level understanding errors* (or *understanding error* for short) as mismatches between the true intent the user aims to convey through speech and what the system decodes the intent to be. The parser outputs the system's final understanding of the intent. However, an understanding error does not necessarily result from a parsing error: an understanding error can result from errors either of the ASR and/or of the parser. Therefore understanding errors need to be distinguished from parsing errors. Figure 6.3 shows an example understanding error made on off-task speech. The error first originated in the ASR output, where the user's complaint about being unable to understand the dialog system was mis-transcribed as some street names, which were in turn recognized by the parser as place information. Consequently, this understanding error caused the system to generate a wrong prompt in the following step.

System: Are you sure you want to quit?

User: *I don't understand it*

ASR: TALBOT AND EIGHTH AND ANN

Parse:
```
PlaceInformation
[SinglePlace] ( [stop_name]
                ( [part1]  ( [place_name] ( TALBOT ) ) ) )
[SinglePlace] ( [stop_name]
                ( [part1] ( [place_name] ( EIGHTH ) )
                  [place_type] ( AND )
                  [part2] ( [place_name] ( ANN ) ) ) )
```

System: I think you said **TALBOT AND ANN** . Do you want to leave from there, or to go there.

Figure 6.3 Example understanding error made on off-task utterance. The error (shaded) originated in the ASR output, and resulted in a wrong system prompt in the following step.

The parser was designed to be robust against errors made by ASR (Raux, et al., 2003). Figure 6.4 shows an example of the parser successfully rejecting conspicuously incorrect ASR output, which in turn rejects the off-task speech correctly.

System:	I think you said **ELEVENTH AND BRADDOCK** . Do you want to leave from there, or to go there.
User:	*it gave me the*
ASR:	BEFORE THAT
Parse:	[no_parse]

Figure 6.4 Parser rejecting error made by ASR.

To reduce errors caused by unexpected off-task speech, we apply off-task speech detection on each user utterance. If an utterance is purely off-task, the parser should not assign any on-task dialog acts to it.

On the test data, our off-task speech detector correctly detected 62 (58%) out of 107 off-task utterances. On these correctly detected off-task utterances, the detector corrected 32 understanding errors. On the other hand, the same off-task speech detector mis-classified 47 (4.9%) out of 960 on-task utterances as off-task. However, in the absence of the off-task speech detector, the system already mis-understood 29 (62%) of these mis-classified on-task utterances. Hence the off-task speech detector increased the number of understanding errors on the on-task speech by only 18. By subtracting these 18 errors from the 32 corrections, the off-task speech detector would have reduced the number of understanding errors on the 1,067 test utterances by 14, thereby reducing the understanding error rate by 1.4% absolute. The Let's Go corpus that is

available for download from the Let's Go project website (http://www.speech.cs.cmu.edu/letsgo/letsgodata.html) does not provide hand labels on whether an utterance is understood correctly by the system. To analyze the absolute percentage of improvement in the understanding rate by the off-task speech detector, we labeled only the utterances classified as off-task, whether correctly or incorrectly. We cannot directly evaluate the relative percentage improvement due to the lack of hand labels on correctness of utterance understandings in the entire Let's Go corpus. Raux, et al. (2005) reported an understanding rate of 45% on a different corpus which was not available at the time this thesis was written. Assuming similar understanding rate on the two corpora, the relative improvement after detecting off-task speech would be 3%. In fact, the more recent 2005 Let's Go data may be easier for understanding because most of the users speak English as their first language, whereas the 2003 corpus (the downloadable version) contained utterances from the system developers, many of whom are non-native English speakers.

The understanding rate is often counted at the turn level rather than at the utterance level (Raux, et al., 2005). However, because we detected off-task speech at the utterance level, we counted the understanding rate at the utterance level. improving utterance-level understanding is also likely to improve understanding at the turn level. A user turn can have multiple utterances, as illustrated in Figure 6.5. But the majority of the user turns (> 97%) contained only one utterance. Besides, we counted only utterances processed by the parser, as shown in Figure 6.5.

Prompt: I need you to tell me which stop you're leaving from. If you don't know the exact name of the stop, try the name of the closest intersection, like FORBES AT MURRAY

{
Speech: *More*
ASR: WHAT

Speech: *Morewood_Avenue*
ASR: [empty]

Speech: *okay*
ASR: CMU
}

Parse: PlaceInformation[SinglePlace] ([stop_name] ([monument] (CMU)))

Figure 6.5 A user turn (in curly brackets) containing multiple utterances.

6.3 Summary

This chapter demonstrated how off-task speech detection can help other applications. The two example applications we showed in this chapter differ in many aspects. Table 6.3 summarizes some major differences between the two applications. All these differences may affect the quality of off-task speech detection. For example, the frequency band of land-line telephone speech is limited to 250 Hz to 4000 Hz, and thus the acoustic features extracted from utterances carried through a phone line would perform differently than from a close-talking microphone. Despite these differences, off-task speech detection is able to improve the performance of both applications.

Table 6.3 Major differences between the two applications.

Application	User age group	Transmission device	Speech type
Predicting fluency test scores	Children ages 7-10	Close-talking microphone	Oral reading
Understanding user utterances in the Let's Go bus information system	Adults	Telephone	Prompted speech in a dialog

Notice that even without an explicit off-task speech detector, both applications carry some mechanism to filter out off-task speech. In predicting fluency test scores, Mostow and Virnik used only words read correctly (except that WCPM counted time for both correct and incorrect reading); in understanding user utterances, the parser in the Let's Go system already rejects some off-task speech. Even with many off-task utterances filtered out by the existing techniques, applying an off-task speech detector still achieved additional gains.

7. Conclusions, limitations, and future work

The goal of this research is to provide a systematic study of off-task speech and the roles of its acoustic and lexical features in automatic detection of off-task speech in a variety of tasks. The thesis contributes to the field of speech science in that it presents three types of features that distinguish off-task from on-task speech. The thesis contributes to the field of speech technology in that it proposes a general framework of how to detect off-task speech, and that it suggests the effectiveness of acoustic and lexical features under different situations. We now summarize how we addressed the research hypotheses stated at the beginning of this dissertation and point out future directions for continuing the research on off-task speech detection.

7.1 How we addressed hypotheses

This thesis addressed the following research questions:

a. How to detect children's off-task speech? What features characterize off-task speech?

Chapter 3 described methods to detect whether an utterance contains off-task speech and where the off-task speech may have occurred in the utterance. We characterized off-task speech using 45 acoustic and 4 lexical features, as well as contextual features to complement the "stand-alone" characteristics of off-task speech (i.e., off-task speech is sparsely interspersed in speech) in the Reading Tutor. We used SVM to train an off-task speech detector on 36,492 oral reading utterances, which yielded 88% detection with 11% false positives on a test set of 651 utterances spoken by 10 randomly chosen children who did not appear in the training data.

Chapter 4 described how to detect off-task speech in other Reading Tutor tasks: self-questioning, think-aloud, and vocabulary activities. Due to insufficient training data, we transferred the off-task speech detector trained on oral reading data to the other Reading Tutor tasks. Task difficulty directly affected accuracy of off-task speech detection. In particular, some difficult tasks such as the vocabulary activities made children's responses difficult to predict, which lead to high ASR error rate and a blurry boundary in speaking styles between on- and off-task speech.

Table 7.1 shows whether our detector significantly out-performs a baseline classifier that randomly assigns labels according to percentage of off-task utterances observed in the training data. Notice that the Chi-square test is sensitive to the distribution of off- and on-task speech. Strictly speaking, the Chi-square test is not the perfect statistical test to account for both true positives and false positives on imbalanced data. However, it is the best that we can find by far.

Table 7.1 Comparison between our trained detector and a baseline classifier that randomly assigns labels according to percentage of off-task utterances observed in training data. Detectors underlined significantly out-perform the baseline ($p<0.05$), according to a Chi-square test.

Comparison to a classifier that randomly assigns labels according to % off-task utterances	**P value from a Chi-square test**
Detector for oral reading	0.04
Detector for self-questioning	0.6
Detector for think-aloud	0.09
Detector for vocabulary activities	0.01

a.1. How to detect off-task speech when training labels are difficult to acquire?

On children's oral reading data, we applied a heuristic based on length of deviation from the text to label utterances as off- or on-task. We tuned this labeling heuristic on 467 development utterances, which reached high agreement with hand labels (Kappa = 0.93). To label speech segments within utterances, we applied two alternative methods: (1) force-align transcriptions with utterances to label speech segments in training data; (2) use multiple-instance learning to derive segment labels automatically. On a test data of 651 utterances segmented into 1,918 speech segments, the two methods provided similar detection accuracies on the speech segments at 80% with 10% false positives.

b. What are the roles of acoustic and lexical features in detecting off-task speech? How do they generalize?

Chapter 5 discussed roles of features of off-task speech by studying the generality of predictive power of each type of feature across four Reading Tutor tasks.

Acoustic features characterized the speaking style of off-task speech. Compared to oral reading, average pitch in off-task speech was higher by 11%, and average loudness lower by 16%. In addition, off-task utterances tended to be shorter than oral reading by 10%, and had a lower harmonics-to-noise ratio. The above acoustic properties tend to cause off-task speech to be less intelligible than oral reading.

Lexical features characterized the linguistic content of off-task speech. Noticing the difference in distribution of vocabulary in on- and off-task speech, we characterized lexical content of off-task speech by percentage of off-task words and on-task words. Percentage of off-task words turned out to be the most powerful lexical feature across all four Reading Tutor tasks.

For most cases, lexical features performed better at detecting general (e.g., "I need to go to the bathroom") and task-related off-task speech (e.g., asking text related questions during oral reading). Acoustic features detected more null responses. However, when ASR was inaccurate, such as for vocabulary activities, acoustic features out-performed lexical features in detecting all types of off-task speech.

Table 7.2 Comparison-related conclusions and their statistical significance.

Conclusion	**Evidence task**	**P value from a Chi-square test**
When ASR is bad, acoustic features alone out-perform acoustic+lexical+contextual	Think-aloud	0.005
	Vocabulary activities	0.07
Acoustic features better at identifying null responses than lexical features	Self-questioning	0.04
	Think-aloud	3e-5
	Vocabulary activities	1
	Let's Go	1e-4
Lexical features better at identifying general and task-related off-task speech than acoustic features	Self-questioning	0.1
	Think-aloud	0.3
	Vocabulary activities	0.5
	Let's Go	0.08
Contextual features help improve classification accuracy	Oral reading	0.04
	Self-questioning	0.65
	Think-aloud	0.04
	Vocabulary activities	0.9
	Let's Go	0.9

We used the ROC curve to evaluate classifier accuracies. The ROC curve evaluates model quality regardless of the class distribution. In practice, for detecting rare events, some parts of the ROC may be of more importance than the other parts, depending on which is more costly: misses or false alarms. For detecting off-task speech, the left part of the ROC curve (i.e., the part that has lower false positive rate) can be more important because frequently misclassifying on-task utterance as off-task can be annoying to users. Besides, off-task speech is

rare compared to on-task speech, therefore over-detecting off-task utterances will lead to high number of utterances misclassified as off-task.

c. How can other applications benefit from off-task speech detection?

Chapter 6 showed that detecting off-task speech can help filter out undesirable events that may affect prediction or classification. For example, filtering out off-task utterances in oral reading can improve accuracy of automatic prediction of fluency test scores. Detecting off-task speech can also improved the understanding rate of utterances in the Let's Go dialog system. The two applications are distinct in many aspects, such as the age group of user (children vs. adults), speech transmission device (close-talking microphone vs. telephone), and the main speech type (read vs. prompted speech), suggesting generality of off-task speech detection across different problem domains.

The motivation of detecting off-task speech originated from difficulties in automatic understanding of children's free-form spoken responses to Project LISTEN's Reading Tutor. Therefore, besides improvement in predicting oral reading fluency scoring, another major contribution of off-task speech detection to Project LISTEN lies in understanding children's free-form spoken responses. In particular, we demonstrated that off-task speech detection can significantly improve classification of children's responses to self-questioning prompts (Chen, et al., 2011), compared to a baseline utterance classifier using only information from ASR output.

7.2 Limitations and future work

We have not yet deployed the off-task speech detector in a system to work in real time. To make the detector run in real time, we may need to modify some features depending on what "real time" means for a particular task. For a dialog system such as Let's Go, "real time" can mean

utterance-by-utterance detection. The running time of the detector includes time for ASR, feature extraction, and classification. ASR and feature extraction can take seconds to complete. So in order to run the detector in real time, there still needs to be some mechanism for speeding up the detection. Tracking oral reading would require slight modification of the current detector to work on short segments of speech to mimic real time performance with imperceptible delay. We have described how to detect off-task segments. By restricting information to only current and previous speech segments (i.e., neglecting future speech segments in an utterance), the technique can be adapted to a real-time off-task speech detector.

This thesis studied only the acoustic and lexical features, in a single utterance and in neighboring contexts. The features we explored are not exhaustive. Other features that fall into these three types can be included in the future. Acoustic, lexical, and contextual features are probably the most universally accessible features for any spoken input. As more knowledge is available, we may include other types of features. For example, behavioral features such as mouse clicks and keyboard strokes may provide additional cues: if a child keeps on clicking the "Next" button and mumbles something, he is likely to be speaking off-task. In addition, the current features are not speaker specific. Normalizing the features by speaker may improve detector accuracy.

We were able to avoid hand labels for training the off-task speech detector, due to a special characteristic of oral reading that enables us to apply heuristics based on deviation length. There may not always be a heuristic for labeling the data automatically. In such cases, we can minimize the cost of hand labels by carefully selecting the most valuable training instances to label. One tool that can be used is *active learning* (Bonwell, 1991), which automatically selects the most valuable training instances. The Amazon's Mechanical Turk

(https://www.mturk.com/mturk/) may also be used to acquire relatively cheap hand labels, but the quality of the labels needs to be carefully controlled.

7.2.1 Fields that can make use of this work

The features and techniques we described for detecting off-task speech are potentially useful for many applications. First, any intelligent interface that involves verbal interaction and detection of engagement may apply the speech features described here. Second, the lexical features may be applicable to some natural language processing applications as well, such as using the percentage of off-task words to detect out-of-topic threads in an online forum processor. Third, since off-task speech is a speech form of off-task behavior, any intelligent tutoring system that uses spoken input may apply the features discussed in this thesis to help detect off-task behaviors. Finally, off-task speech detection can improve a dialog system by helping the system better understand users' spoken responses.

8. References

Abramowitz, A. J., O'Leary, S. G., & Rosen, L. A. (1987). Reducing off-task behavior in the classroom: A comparison of encouragement and reprimands. *Journal of Abnormal Child Psychology, 15*, 153-163.

Alexandersson, J., Buschbeck-Wolf, B., Fujinami, T., Kipp, M., Koch, S., Maier, E., Reithinger, N., Schmitz, B., & Siegel, M. (1998). Dialogue Acts in VERBMOBIL-2 (second edition) *Vm report 226*. Saarbrucken and Stuttgart: Universities of Berlin.

Alpern, M., Minardo, K., O'Toole, M., Quinn, A., & Ritzie, S. (2001). *Project LISTEN: Design Recommendations and Teacher Tool Prototype (Unpublished Group Project for Masters' Lab in Human-Computer Interaction)*. Carnegie Mellon University. Pittsburgh.

Anderson, A., Bader, M., Bard, E., Boyle, E., & Doherty, G. (1991). The HCRC Map Task Corpus. *Language and Speech, 34*, 356-366.

Andrews, S., Tsochantaridis, I., & Hofmann, T. (2002). *Support Vector Machines for Multiple-Instance Learning*. The Neural Information Processing Systems.

Asadi, A., Schwartz, R., & Makhoul, J. (1990). *Automatic detection of new words in a large vocabulary continuous speech recognition system*. The IEEE International Conference on Acoustics, Speech, and Signal Processing.

Austin, J. L. (1962). *How to do things with words*. Cambridge, MA: Harvard U. P. .

Baker, R. S. (2007). *Modeling and understanding students' off-task behavior in intelligent tutoring systems*. The Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, San Jose, California, USA.

Baker, R. S., Corbett, A. T., & Koedinger, K. R. (2004). *Detecting Student Misuse of Intelligent Tutoring Systems*. The the 7th International Conference on Intelligent Tutoring Systems.

Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004). *Off-Task Behavior in the Cognitive Tutor Classroom: When Students "Game the System"*. The ACM CHI 2004: Computer-Human Interaction.

Baker, R. S., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to Be Frustrated than Bored: The Incidence, Persistence, and Impact of Learners' Cognitive-Affective States during Interactions with Three

Different Computer-Based Learning Enviroments. *International Journal of Human-Computer Studies, 64*(4), 223-241.
Bansal, D., & Ravishankar, M. K. (1998). *New Features for confidence Annotation*. The International Conference on Spoken Language Processing.
Batliner, A., Steidl, S., Schuller, B., Seppi, D., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Aharonson, V., Kessous, L., & Amir, N. (2010). Whodunnit – Searching for the most important feature types signaling emotion – related user states in speech. *Computer Speech and Language*.
Beal, C. R., Mitra, S., & Cohen, P. R. (2007). *Modeling learning patterns of students with a tutoring system using Hidden Markov Models*. The Artificial intelligence in education (AIED): Building technology rich learning contexts that work.
Beck, J. E., Jia, P., & Mostow, J. (2004). Automatically assessing oral reading fluency in a computer tutor that listens. *Technology, Instruction, Cognition and Learning, 2*(1-2), 61-81.
Bickmore, T., & Giorgino, T. (2004). *Some Novel Aspects of Health Communication from a Dialogue Systems Perspective*. The AAAI Fall Symposium on Dialogue Systems for Health Communication, Washington DC.
Blalock, H. (1972). *Social Statistics*: NY: McGraw-Hill.
Boersma, P., & Weenink, D. (2010). Praat: doing phonetics by computer [Computer program], Version 5.1.44, retrieved from http://www.praat.org/, 2010.
Bonwell, C. E., J. (1991). *Active Learning: Creating Excitement in the Classroom AEHE-ERIC Higher Education Report No. 1.* : Washington, D.C.: Jossey-Bass.
Booth, P. A. (1989). *An Introduction to Human-Computer Interaction*. East Sussex, U.K.: Lawrence Erlbaum Associates Ltd.
Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. E. (2001). Disfluency rates in spontaneous speech: Effects of age, relationship, topic, role, and gender. *Language and Speech, 44*, 123-147.
Cetintas, S., Si, L., Xin, Y. P., & Hord, C. (2010). Automatic Detection of Off-Task Behavior in Intelligent Tutoring Systems with Machine Learning Techniques. *IEEE Transactions on Learning Technologies, 3*(3).
Chang, C.-C., & Lin, C.-J. (2001). LIBSVM : a library for support vector machines, from http://www.csie.ntu.edu.tw/~cjlin/libsvm
Chase, L. (1997a). *Error-Responsive Feedback Mechanisms for Speech Recognizers.* Ph.D. , Carnegie Mellon University.

Chase, L. (1997b). *Word and acoustic confidence annotation for large vocabulary speech recognition*. The European Conference on Speech Communication and Technology.

Chawla, N. V., Japkowicz, N., & Kolcz, A. (2004). Editorial: special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter, 6*(1), 1-6.

Chen, W., Mostow, J., & Aist, G. (2010). *Exploiting Predictable Response Training to Improve Automatic Recognition of Children's Spoken Questions*. The 10th International Conference on Intelligent Tutoring Systems, Pittsburgh, PA.

Chen, W., Mostow, J., & Aist, G. (2011). *Using Automatic Question Generation to Evaluate Questions Generated by Children*. The 2011 AAAI Fall Symposium on Question Generation, Arlington, VA.

CMU. (2010). CMU Sphinx: Open Source Toolkit For Speech Recognition, from http://cmusphinx.sourceforge.net/

Cohen, P. R., & Perrault, C. R. (1979). Elements of a Plan-Based Theory of Speech Acts. *Cognitive Science, 3*(3), 177-212.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. G. (2001). Emotion recognition in human-computer interaction *Signal Processing Magazine, IEEE 18*(1), 32-80.

Dietterich, T. G., Lathrop, R. H., & Lozano-Perez, T. (1997). Solving the Multiple-Instance Problem with Axis-Parallel Rectangles. *Artificial Intelligence Journal, 89*.

Dolch, E. (1936). A basic sight vocabulary. *Elementary School Journal, 36*, 456-460.

Dowding, J., Alena, R., Clancey, W. J., Sierhuis, M., & Graham, J. (2006). *Are You Talking To Me? Dialogue Systems Supporting Mixed Teams of Humans and Robots*. The AAAI Fall Symposium on Aurally Informed Performance: Integrating Machine Listening and Auditory Presentation in Robotic Systems, Arlington, Virginia.

Duke, N. K., & Pearson, P. D. (2002). Effective Practices for Developing Reading Comprehension. In A. E. Farstrup & S. J. Samuels (Eds.), *What Research Has To Say about Reading Instruction* (Third ed., pp. 205--242). Newark, DE: International Reading Association.

Durrette, C. (2010). *PSLC poster: What do children say to an intelligent tutoring system?* Carnegie Mellon University.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*: MIT Press, Cambridge, MA.

Fiscus, J. G., Ajot, J., Radde, N., & Laprun, C. (2006). *Multiple Dimension Levenshtein Edit Distance Calculations for Evaluating Automatic Speech Recognition Systems During Simultaneous Speech*. The International Conference on language Resources and Evaluation (LERC).

Fogarty, J., Dabbish, L., Steck, D., & Mostow, J. (2001). *Mining a database of reading mistakes: For what should an automated Reading Tutor listen?* The Artificial Intelligence in Education: AI-ED in the Wired and Wireless Future, San Antonio, Texas.

Foote, J. (2000). *Automatic Audio Segmentation using a Measure of Audio Novelty.* the Proceedings of IEEE International Conference on Multimedia and Expo.

Freund, Y., & Schapire, R. E. (1996). *Game theory, On-line Prediction and Boosting*. The Proceedings of the Ninth Annual Conference on Computational Learning Theory.

Gates, D. (2008). *Generating Look-Back Strategy Questions from Expository Texts*. The Workshop on the Question Generation Shared Task and Evaluation Challenge, NSF, Arlington, VA. http://www.cs.memphis.edu/~vrus/questiongeneration//1-Gates-QG08.pdf

Gobel, P. (2008). Student off-task behavior and motivation in the CALL classroom. *International Journal of Pedagogies and Learning, 4*(4), 4-18.

Good, R. H. I., & Jefferson, G. (Eds.). (1998). *Contemporary perspectives on curriculumbased measurement validity*. New York: Guilford.

Hazen, T. J., & Bazzi, I. (2001). *A Comparison and Combination of Methods for OOV Word Detection and Word Confidence Scoring*. The Proceedings of the International Conference on Acoustics, Speech and Signal Processing.

Huang, H.-H., Baba, N., & Nakano, Y. (2011). *Making virtual conversational agent aware of the addressee of users' utterances in multi-user conversation using nonverbal information*. The International Conference on Multimodal Interaction.

Jiang, H. (2005). Confidence measures for speech recognition: A survey. *Speech Communications, 45*, 455-470.

Jovanovic, N., & Akker, R. o. d. (2004). *Towards automatic addressee identification in multi-party dialogues*. The the 5th SIGdial Workshop on Discourse and Dialogue, Boston, MA.

Jovanovic, N., & Akker, R. o. d. (2006). *Addressee identification in face-to-face meetings*. The European Chapter of the Association for Computational Linguistics.

Keeler, J. D., Rumelhart, D. E., & Leow, W.-K. (1990). *Integrated segmentation and recognition of hand-printed numerals*. The NIPS-3: Proceedings of the 1990 conference on Advanced in neural information processing systems 3, San Francisco, CA, USA.

Kemp, T., & Schaaf, T. (1997). *Estimating confidence using word lattices*. The European Conference on Speech Communication and Technology.

Kluwer, T., Adolphs, P., Xu, F., Uszkoreit, H., & Cheng, X. (2010). *Talking NPCs in a Virtual Game World*. The Annual Meeting of the Association for Computational Linguistics: 2010 System Demonstrations.

Lane, I., Kawahara, T., Matsui, T., & Nakamura, S. (2004). *Out-of-domain detection based on cinfidence measures from multiple topic classification*. The IEEE International Conference on Acoustics, Speech, and Signal Processing.

Lane, I., Kawahara, T., Matsui, T., & Nakamura, S. (2007). Out-of-Domain Utterance Detection Using Classification Confidences of Multiple Topics. *IEEE Transactions on Audio, Speech, and Language Processing, 15*(1), 150 – 161.

Lane, I. R., & Kawahara, T. (2005). *Incorporating dialog context and topic clustering in out-of-domain detection*. The IEEE International Conference on Acoustics, Speech, and Signal Processing.

Lee, S.-I., Lee, H., Abbeel, P., & Ng, A. Y. (2006). *Efficient L1 Regularized Logistic Regression*. The Twenty-First National Conference on Artificial Intelligence (AAAI-06).

Lieberman, P., & Michaels, S. B. (1962). Some aspects of fundamental frequency and envelope amplitude as related to emotional context of speech. *J. Acoust Soc. Am., 34*, 922-927.

Lu, L., Li, S. Z., & Zhang, H.-J. (2001). *Content-based audio segmentation using support vector machines* the IEEE International Conference on Multimedia and Expo.

Merlo, S., & Mansur, L. L. (2004). Descriptive discourse: topic familiarity and disfluencies. *Journal of Communication Disorders, 37*, 489-503.

Miller, J. D. (1989). Auditory-perceptual interpretation of the vowel. *J. Acoust. Soc. Am., 85*(5), 2114-2134.

Mostow, J., & Aist, G. (1999). Giving Help and Praise in a Reading Tutor with Imperfect Listening – Because Automated Speech Recognition Means Never Being Able to Say You're Certain. *CALICO Journal Special issue (M. Holland, Ed.), Tutors that Listen: Speech recognition for Language Learning, 16*(3), 407-424.

Mostow, J., Aist, G., Corbett, A., Beck, J., Duke, N., McKeown, M., P., S., Trotochaud, C., Chen, W., Duong, M., Liu, L., Mills-Tettey, A., Bey, J., Gates, D., Juarez, O., Kantorzyk, M., Miller, L., Valeri, J., & Weinstein, A. (2009). *Automated generation of reading instruction in fluency, vocabulary, and comprehension*. The Institute of Education Sciences Research Conference.

Mostow, J., Aist, G., Gates, D., Liu, L., Valeri, J., Bey, J., Weinstein, A., Kantorzyk, M., & Duan, W. (2010). *A funny thing happened on the way to the system: a morphology surprise while developing a vocabulary tutor*. The Institute of Education Sciences Research Conference.

Mostow, J., & Beck, J. (2007). When the Rubber Meets the Road: Lessons from the In-School Adventures of an Automated Reading Tutor that Listens. *Scale-Up in Education, 2*, 183-200.

Mostow, J., Chen, W., Corbett, A., Valeri, J., Bey, J., Weinstein, A., Juarez, O., & Kantorzyk, M. (2010). *Challenges in automated instruction of reading comprehension strategies*. The Institute of Education Sciences Research Conference.

Orkin, J., & Roy, D. (2010). *Semi-Automated Dialogue Act Classification for Situated Social Agents in Games*. The Agents for Games & Simulations Workshop at the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS), Toronto, Canada.

Oviatt, S. (1995). Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language, 9*, 19-35.

Parada, C., Dredze, M., Filimonov, D., & Jelinek, F. (2010). *Contextual Information Improves OOV Detection in Speech*. The North American Chapter of the Association for Computational Linguistics (NAACL).

Perrault, C. R., & Allen, J. F. (1980). A Plan-Based Analysis of Indirect Speech Acts. *Computational Linguistics, 6*, 167-182.

Pradhan, S., Ward, W., & Martin, J. H. (2008). Towards Robust Semantic Role Labeling. *Computational Linguistics Special Issue on Semantic Role Labeling, 34*(2), 289-310.

Rasmussen, M. H., Mostow, J., Tan, Z.-H., Lindberg, B., & Li, Y. (2011). *Evaluating Tracking Accuracy of an Automatic Reading Tutor*. The SLaTE: ISCA (International Speech Communication Association) Special Interest Group (SIG) Workshop on Speech and Language Technology in Education, Venice, Italy.

Raux, A., Langner, B., Black, A., & Eskenazi, M. (2003). *LET'S GO: Improving Spoken Dialog Systems for the Elderly and Non-natives*. The European

Conference on Speech Communication and Technology, Geneva, Switzerland.
Raux, A., Langner, B., Black, A., & Eskenazi, M. (2005). *Let's Go Public! Taking a Spoken Dialog System to the Real World*. The Annual Conference of the International Speech Communication Association (Interspeech), Lisbon, Portugal.
Rose, R. C., Juang, B. H., & Lee, C. H. (1995). *A training procedure for verifying string hypothesis in continuous speech recognition*. The IEEE International Conference on Acoustics, Speech, and Signal Processing.
Sabourin, J., Rowe, J., Mott, B., & Lester, J. (2011). *When Off-Task in On-Task: The Affective Role of Off-Task Behavior in Narrative-Centered Learning Environments*. The 15th International Conference on Artificial Intelligence in Education (AIED-2011), Auckland, New Zealand.
San-Segundo, R., Pellom, B., Hacioglu, K., & Ward, W. (2001). *Confidence Measures for Spoken Dialogue Systems.* the IEEE International Conference on Acoustics, Speech, and Signal Processing.
Schaaf, T. (2001). *Detection of OOV Words Using Generalized Word Models and a Semantic Class Language Model*. The European Conference on Speech Communication and Technology.
Scharenborg, O. (2007). Reaching over the Gap: A Review of Efforts to Link Human and Automatic Speech Recognition Research. *Speech Communications, 49*, 336–347.
Shinozaki, T., Ostendorf, M., & Atlas, L. (2009). Characteristics of Speaking Style and Implications for Speech Recognition. *J. Acoust Soc. Am.,, 126*(3), 1500-1510.
Shriberg, E., & Lickley, R. (1993). Intonation of clause-internal filled pauses. *Phonetica, 50*, 172-179.
Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psych. Bull., 86*(2), 420-428.
Stolcke, A. (2002). *SRILM - An Extensible Language Modeling Toolkit*. The Intl. Conf. Spoken Language Processing, Denver, Colorado.
Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Meteer, M., & Ess-Dykema, C. V. (2000). Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics, 26*(3), 339-371.
Tam, Y.-C., Mostow, J., Beck, J., & Banerjee, S. (2003). *Training a Confidence Measure for a Reading Tutor that Listens*. The Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003), Geneva, Switzerland.

Toutanova, K., Klein, D., Manning, C., & Singer, Y. (2003). *Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network*. The Human Language Technologies: The 4th Annual Conference of the North American Chapter of the Association for Computational Linguistics

Traum, D. (2003). Issues in multi-party dialogs. In F. Dignum (Ed.), *Advances in agent communication*: Springer-Verlag LNCS.

Traum, D. R., & Gendve, U. d. (1996). Book Reviews: Spoken Natural Language Dialogue Systems: A Practical Approach by Ronnie W. Smith and D. Richard Hipp. *Computational Linguistics, 22*(3).

Vezhnevets, A., & Vezhnevets, V. (2005). *„Modest AdaBoost'' - Teaching AdaBoost to Generalize Better*. The Graphicon-2005, Novosibirsk Akademgorodok, Russia.

Viola, P., & Jones, M. (2001). *Robust Real-time Object Detection*. The Proceedings of 2nd Int'l Workshop on Statistical and Computational Theories of Vision -- Modeling, Learning, Computing and Sampling, Vancouver, Canada.

Virnik, M., & Mostow, J. (2006). *What do teachers want to know about their students who use the Reading Tutor?* . Carnegie Mellon University. Pittsburgh.

Ward, W., & Issar, S. (1994). *Recent improvements in the CMU spoken language understanding system.* The ARPA Human Language Technology Workshop Plainsboro, NJ.

Webb, N. (2010). *Cue-Based Dialog Act Classification.* PhD, University of Sheffield, Sheffield, UK.

Wessel, F., Schlüter, R., Macherey, K., & Ney, H. (2001). Confidence Measures for Large Vocabulary Continuous Speech Recognition. *IEEE Trans. on Speech and Audio Processing*.

White, C., Droppo, J., Acero, A., & Odell, J. (2007). *Maximum Entropy Confidence Estimation For Speech Recognition*. The IEEE International Conference on Acoustics, Speech, and Signal Processing.

White, C., Zweig, G., Burget, L., Schwarz, P., & Hermansky, H. (2008). *Confidence Estimation, OOV Detection and Language in Using Phone-toWord Transduction and Phone-level Alignments*. The IEEE International Conference on Acoustics, Speech, and Signal Processing.

Williams, C. E., & Stevens, K. N. (1972). Emotion and speech: Some acoustical correlates. *J. Acoust Soc. Am., 52*, 1238-1250.

Wordsmyth. (2010). Wordsmyth, from http://www.wordsmyth.net/

Yang, J. (2008). MILL: A Multiple Instance Learning Library, from http://www.cs.cmu.edu/~juny/MILL