A Tale of Two Tasks: Detecting Children's Off-Task Speech in a Reading Tutor

Wei Chen [1], *Jack Mostow*[2]

[1] Language Technologies Institute, School of Computer Science, Carnegie Mellon University, USA
[2] Project LISTEN, School of Computer Science, Carnegie Mellon University, USA

weichen@cs.cmu.edu, mostow@cs.cmu.edu

Abstract

How can an automated tutor detect children's off-task utterances? To answer this question, we trained SVM classifiers on a corpus of 495 children's 36,492 computer-assisted oral reading utterances. On a test set of 620 utterances by 10 held-out readers, the classifier correctly detected 88% of off-task utterances and misclassified 17% of on-task utterances as off-task. As a test of generality, we applied the same classifier to 20 children's 410 responses to vocabulary questions. The classifier detected 84% of off-task utterances but misclassified 57% of on-task utterances. Acoustic and lexical features helped detected off-task speech in both tasks.

Index Terms: off-task speech detection, acoustic feature, lexical feature, children speech

1. Introduction

Off-task speech is speech that strays away from an intended task. It occurs in many dialog applications, such as intelligent tutors [1], virtual games [2]¸ health communication systems [3] and human-robot cooperation [4]. On the one hand, an automated agent capable of detecting off-task speech could track users' attention and maintain natural dialogs by bringing a user back on task [5, 6]. Also, knowledge of where off-task speech events are likely to have occurred can help in analyzing automatic speech recognition (ASR) errors. On the other hand, off-task speech detection faces challenges of informal conversational style and potentially unbounded scope [3] that hamper accurate speech recognition. Despite the opportunities and challenges presented by off-task speech, we are not aware of previous research explicitly focused on it.

The goal of our research is to address this gap. As a step towards this goal, we start by detecting and analyzing off-task speech in children's interactions with an automated reading tutor. Section 2 relates this goal to prior work; Sections 3 and 4 describe our data and features, respectively; Section 5 presents results; Section 6 discusses their generality; and Section 7 concludes.

2. Relation to prior work

Off-task speech is similar to but distinct from some previously studied phenomena.

Off-task speech resembles out-of-domain (OOD) speech [7] in referring to information outside the scope of the system. For example, a user may ask a travel system *When is the next train from London to Aldeburgh* (which has no train service)*?* Even though this query is out of scope, the utterance asks for travel information and hence is on-task. 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. Hence existing OOD methods focus mostly on word cues for topic modeling, while off-task speech detection uses both acoustic and lexical features.

Off-task speech includes speech addressed to a third party [8], but is not restricted to it. For example, off-task speech addressed to the system during oral reading includes questions such as *Can I stop here?* and comments about classmates.

Off-task speech resembles spontaneous speech [9] in speaking style, but spontaneous speech is not necessarily off-task. For instance, children's on-task responses to vocabulary questions often involve disorganized spontaneous speech.

3. Two tasks

We now describe our data sets and annotation scheme for children's oral reading and responses to vocabulary questions.

3.1. Oral reading and automatic annotation

The oral reading corpus consists of utterances collected by Project LISTEN's Reading Tutor [10] during children reading story sentences out loud. The training data contains 36,492 utterances spoken by 495 children totaling 43 hours of audio recordings. The test data contains 659 utterances spoken by 10 randomly chosen children who do not appear in the training data, with total audio length of 1 hour 3 minutes.

For oral reading, off-task means 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.

To train and test classifiers, we needed to label the data. Rather than hand-labeling so much training data as on- or off-task, we use the following "deviation length" heuristic to find off-task speech in already transcribed oral reading. 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. Figure 1 shows a typical alignment result.

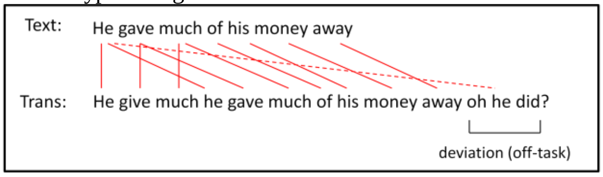

Text:
He gave much of his money away
Trans:
He give much he gave much of his money away oh he did?
deviation (off-task)

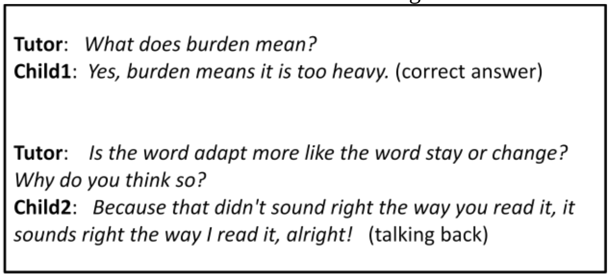
Figure 1: *Alignment with disfluency and misreading.*

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. To recognize misreading (e.g., *gave* as *give*), we measure the length of the deviation as the min of orthographic and phonetic edit distances between non-matching words, and relative deviation length as deviation length / transcription length. If this ratio exceeds 0.5, we labeled the utterance as off-task. If it falls between 0.36 and 0.5, we labeled it as partially off-task and excluded it from the current study. We tuned the thresholds on a separate development set of 467 utterances to maximize Kappa agreement (0.93) compared to hand labels by one rater. Our automated method labeled 4,236 (12%) utterances as off-task, 29,196 as on-task, and the other 3,060 as partial off-task.

Two annotators independently labeled the 659 test utterances for the oral reading task with inter-rater agreement of Kappa = 0.96. They labeled 51 utterances as off-task, 569 as on-task, and 31 as partial off-task, which we excluded from analysis, along with the 8 where the raters disagreed.

3.2. Vocabulary task

The vocabulary task prompted children to explain the meaning of a word or to compare semantic relatedness between words. As Figure 2 shows, the vocabulary questions elicited utterances less constrained than oral reading.



Tutor: *What does burden mean?*
Child1: *Yes, burden means it is too heavy.* (correct answer)

Tutor: *Is the word adapt more like the word stay or change? Why do you think so?*
Child2: *Because that didn't sound right the way you read it, it sounds right the way I read it, alright!* (talking back)

Figure 2: *Example responses to vocabulary questions.*

An undergraduate summer intern annotated the vocabulary responses with finer grained categories such as correct answer (as in Figure 2), wrong answer (e.g., *a burden means you're a thief*), no response, playing (e.g., *Haha! Oh oh!*), and talking back (as in Figure 2). We categorized correct and wrong answers as on-task, and the rest as off-task. The first author later independently annotated the utterances using only on- and off-task labels. The Kappa score was 0.83, with most disagreement occurring on the wrong answer category. After filtering out the 23 utterances with annotator disagreement, we obtained 410 utterances, 139 (34%) of them labeled off-task. We used these data as a second test set to investigate the generality of models trained on off-task oral reading.

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. Only 12% of oral reading utterances are off-task, versus 34% for the vocabulary task.

4. Features used

To characterize the content of off-task speech, i.e., what was said, we used lexical features of ASR output. To characterize speaking style, i.e., how it was said, we used acoustic features computed directly from the speech signal, without ASR.

4.1. Acoustic features and feature selection

Table 1 summarizes five groups of low level acoustic descriptors extracted using Praat [11] scripts. Before extracting features from the acoustic descriptors, we used a Praat script to segment each audio recording into voiced, unvoiced, and silence regions. For each region and entire utterance, we calculated statistics to summarize frame based acoustic descriptor values, including mean, minimum, maximum, quartiles, and the first four moments.

Training a classifier with 1,250 acoustic features on only 4,236 off-task instances is doomed to overfit. Therefore we applied the Adaboost learning algorithm to choose the 50 features most informative for oral reading data. We used the Adaboost algorithm to overcome the imbalanced training data problem (4,236 off-task vs. 29,196 on-task) by boosting the weight on errors made on the minority class. Table 2 lists the top 10 features ranked by descending absolute weight.

Table 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, pulses

Table 2. *The top 10 acoustic features.*

Feature name	Weight
Pitch – 3rd quartile	0.079
Loudness (voiced) – 3rd moment	0.072
Degree of voice breaks	0.068
Mean harmonics-to-noise ratio	-0.061
Duration	-0.053
Shimmer	0.049
4th MFCC (voiced) – variance	0.044
Number of pulse periods	-0.039
Spectrum (200-1000Hz) – mean skewness	0.038
LTAS (0-8000Hz) – 4th moment	0.034

4.2. Lexical features

Lexical features characterize the content of an utterance. We extracted lexical features from the Sphinx-3 speech recognizer's output and confidence scores, using a 32 Gaussian mixture acoustic model with vocal tract length normalization, trained on 43 hours of children's oral reading data (the same training corpus described in Section 3.1).

We designed language models for speech recognition to cover both the task domain and frequent off-task language, so they vary by task. Our approach was to interpolate a task language model with a trigram language model built from an off-task corpus. For oral reading, the task language model consisted simply of the trigrams in the sentence being read. Since we do not have 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 [12] and WordNet [13], along with words we expected in children's word explanations, such as *means* and *something*.

We trained a general off-task trigram language model from transcriptions of the 4,236 off-task utterances in the oral reading training data, which comprise 18,040 tokens of 2,012 distinct word types. To avoid overfitting, our off-task language model includes just the 200 most frequent of these word types, which we call *off-task words*. They covered 74% of the off-task tokens and occurred in 86% of the off-task utterances in the training data. The 10 most frequent words were *I* (860 tokens), *you* (552), *it* (397), *to* (354), *the* (350), *what* (331), *on* (300), *go* (283), *this* (271), and *that* (262). For these 10 words, the Kullback–Leibler divergence (measured in bits) $D_{KL}(P_{offtask}||P_{ontask}) = 0.71$, where the functions $P_{offtask}$ and P_{ontask} represent the probability distributions of the 10 words in off-task and on-task utterances, respectively.

We computed three lexical features of each ASR hypothesis: (1) percentage of off-task words, (2) percentage of off-task words with ASR confidence scores higher than a threshold, and (3) percentage of on-task words with confidence scores lower than a threshold. We used the threshold to decide whether to classify a hypothesis word as recognized correctly. To minimize classification error, we tuned this threshold on the oral reading training data. Percentage of off-task words contributes the most to detection rate. This single feature detects 55% of the off-task speech. On the other hand, percentage of off-task words with confidence scores higher than a threshold contributes the most to classifying on-task speech (96%).

5. Evaluation

To study the predictive power of the proposed features, we trained SVM classifiers and tested them on oral reading and vocabulary tasks, using LIBSVM-3.0 [14] with its radial basis function kernel and default settings except for the data weighting parameters. We used receiver operating characteristic (ROC) curves to summarize performance of the trained classifiers for 21 threshold values ranging from -2 to 2.

There were 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 [15]. The direct impact of the data imbalance on our study was that using the natural distribution of the data to train an SVM classifier yielded a degenerate solution that classified 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 used 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 obtained a classifier that weighted off-task data 6 times as much as on-task data. It detected 88% of off-task utterances and falsely classified 17% of on-task utterances. Figure 3 shows the ROC curves for oral reading test data with off-task training utterances weighted 1, 2, or 6 times as much as on-task utterances. When weight was 1 (i.e., using the original data distribution), only one of the decision thresholds tested gave non-degenerate results; all other points clustered at [0,0], and [1,1]. The shape of the ROC curves did not change much for weights higher than 2.

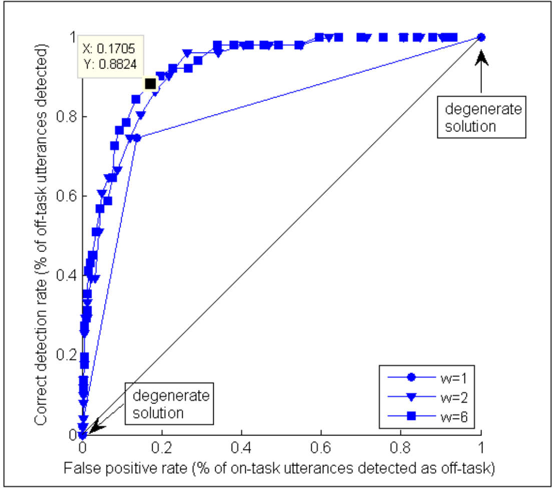


Figure 3: *ROC curves on oral reading test data of classifiers trained with different data weights.*

Figure 4 shows how the oral reading off-task detector performed on test data for the oral reading and vocabulary tasks. Transferring this detector to the vocabulary task reduced detection only slightly, from 88% to 84%, but increased the false positive rate (percentage of on-task speech misclassified as off-task) from 17% to 57%. Figure 4 also shows performance of classifiers trained on oral reading using only acoustic features or only lexical features, discussed shortly.

To evaluate the effect of training on the 410 vocabulary utterances using the same features, we used leave-one-out cross validation. Compared to classifiers trained on oral reading, cross-validated performance on the vocabulary task showed similar ROC curves for training on all features or lexical features only. That is, training on a small amount of vocabulary data did just as well as training on oral reading data. Using only acoustic features, training on vocabulary data vs. oral reading detected almost the same percentage of off-task vocabulary utterances (81% vs. 82%), but with a lower false positive rate (57% vs. 70%).

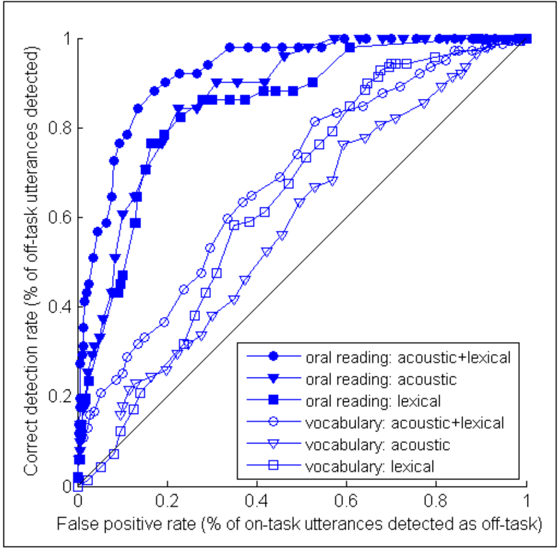


Figure 4: *ROC curves of classification results on oral reading and vocabulary tasks.*

6. Discussion

The classifier trained on oral reading detected a similar percentage of off-task speech on the vocabulary task, but performed much less accurately on the on-task speech. Why?

The similar detection rates for off-task speech in oral reading (84%) and vocabulary responses (82%) using only acoustic features suggest that their acoustic characteristics are similar. In contrast, the marked difference in false positive rates for oral reading (22%) and vocabulary (70%) suggests a difference in speaking style. Oral reading is easier than the vocabulary task because the word sequence only needs to be read rather than composed. The spontaneous speaking style of vocabulary responses shares characteristics of off-task speech that make it difficult to distinguish from on-task speech.

Some types of off-task utterances not observed in the training data occurred in vocabulary responses: singing, humming, and null responses (i.e., audio recordings containing background noise but no user vocalization). Yet the acoustic features generalized to these unseen types. 100% of the 8 utterances with humming or singing and 98% of the 65 null responses were correctly classified as off-task using only acoustic features. This finding needs to be replicated on other tasks, but provides some reason to hope for the existence of task-independent acoustic features for detecting some types of off-task speech.

The strength of a language model depends on its ability to predict which words the speaker will utter, and in what order. Thus our language model of oral reading is very strong because both the text words and their order are known, and even disfluent reading deviates from them only somewhat. The predictability of the on-task utterances directly affects accuracy of ASR hypotheses and subsequently the quality of the lexical features extracted from the hypotheses.

Using lexical features only, the false positive rate for the vocabulary task was 56% compared to 19% for oral reading. 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. Overall our lexicon for the vocabulary task covered only 40% of the on-task tokens. With few children's vocabulary responses to train a language model on, we used a unigram language model. Due to the weak language model, ASR performance was poor.

The words used in off-task speech are less predictable than in oral reading, but surprisingly predictable compared to vocabulary explanations, and the quality of our lexical features depends on the predictability of its words. The 200 most frequent off-task words extracted from the training data covered 77% of the word tokens in the off-task test data for oral reading, and 65% of the off-task vocabulary responses. The off-task detection rate using only lexical features was comparable for the two tasks: 78% for oral reading and 79% for vocabulary responses. For the vocabulary task, off-task words that has always occurred in correctly detected off-task utterances include *can't*, *gonna*, *ha*, *mister*, and *spell*.

Table 3 summarizes ASR performance for on- and off-task speech in the oral reading and vocabulary tasks in terms of word error rate (# insertions, deletions, and substitutions / # transcribed words) and recognition accuracy (# correctly recognized words / # transcribed words). We observed high insertion error rate due to background noise and speech in the classroom. We have adjusted penalty for inserting silences and filler words to reduce WER. However it caused classification accuracy to drop too. We therefore kept high insertion and used confidence thresholding to compute the features.

Table 3. *ASR performance for the two tasks.*

	WER		Recognition accuracy	
	Reading	Vocab	Reading	Vocab
On-task	26%	93%	86%	41%
Off-task	92%	99%	22%	26%
Overall	32%	96%	79%	33%

7. Conclusions

This paper introduces the problem of detecting off-task speech, which is closely related to but different from the previously studied problems of out-of-domain detection, addressee identification, and spontaneous speech detection. We studied off-task speech detection in two tutorial activities for children: oral reading and vocabulary tasks. To automate the annotation of transcribed oral reading as off- or on-task, we used a deviation length heuristic calculated by aligning transcripts against story sentences. To characterize the difference between on- and off-task speech, we used acoustic features to capture speaking style, and lexical features to capture content. To investigate the generality of acoustic and lexical features across tasks, we trained an SVM classifier on the task of oral reading and compared its performance on both oral reading and vocabulary tasks. To deal with naturally imbalanced training data, we used an existing data weighting method in the SVM trainer to boost the weight of the minority class. For off-task speech, both the acoustic and lexical features generalized well across the two tasks, yielding comparable detection rates of around 80%. Acoustic features generalized even to untrained types of off-task speech, namely null responses, humming, and singing. We attribute the higher false positive rate on vocabulary responses to the overlapping speaking style between its off- and on-task speech and the potentially wide choice of words in the task language.

Future work includes applying off-task speech detection to other tasks, both within and beyond tutorial activities and children's speech. We need better features and task language models to improve accuracy, but the ultimate evaluation of off-task speech detection is how it affects a dialog system.

8. Acknowledgements

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B070458. The opinions expressed are those of the authors and do not necessarily represent the views of the Institute and the U.S. Department of Education. We thank Corinne Durrette and Donna Gates for annotating our data.

9. References

[1] Chen, W., Mostow, J., and Aist, G., "Exploiting Predictable Response Training to Improve Automatic Recognition of Children's Spoken Questions," in ITS2010, Pittsburgh, PA, 2010.

[2] Kluwer, T., Adolphs, P., Xu, F., Uszkoreit, H., and Cheng, X., "Talking NPCs in a Virtual Game World," in the ACL 2010 System Demonstrations, 2010.

[3] Bickmore, T. and Giorgino, T., "Some Novel Aspects of Health Communication from a Dialogue Systems Perspective," in AAAI Fall Symposium on Dialogue Systems for Health Communication, Washington DC, 2004.

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

[5] Traum, D. R. and Gendve, U. d., "Book Reviews: Spoken Natural Language Dialogue Systems: A Practical Approach by Ronnie W. Smith and D. Richard Hipp," *Computational Linguistics,* vol. 22, 1996.

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

[7] Lane, I., Kawahara, T., Matsui, T., and Nakamura, S., "Out-of-Domain Utterance Detection Using Classification Confidences of Multiple Topics," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 15, pp. 150 – 161, 2007.

[8] Jovanovic, N. and Akker, R. o. k., "Towards automatic addressee identification in multi-party dialogues," in the 5th SIGdial Workshop on Discourse and Dialogue, Boston, MA, 2004.

[9] Dufour, R., Jousse, V., Est`eve, Y., B´echet, F., and Linar`es, G., "Spontaneous Speech Characterization and Detection in Large Audio Database," in SPECOM, St. Petersburg, 2009.

[10] Mostow, J. and Beck, J., "When the Rubber Meets the Road: Lessons from the In-School Adventures of an Automated Reading Tutor that Listens," *Scale-Up in Education,* vol. 2, pp. 183-200, 2007.

[11] Boersma, P. and Weenink, D. (2010) *Praat: doing phonetics by computer [Computer program], Version 5.1.44, retrieved from http://www.praat.org/, 2010.*

[12] *Wordsmyth*. http://www.wordsmyth.net/

[13] Fellbaum, C., *WordNet: An Electronic Lexical Database*: MIT Press, Cambridge, MA, 1998.

[14] Chang, C.-C. and Lin, C.-J. (2001) *LIBSVM : a library for support vector machines*. www.csie.ntu.edu.tw/~cjlin/libsvm

[15] Chawla, N. V., Japkowicz, N., and Kolcz, A., "Editorial: special issue on learning from imbalanced data sets," *ACM SIGKDD Explorations Newsletter,* vol. 6, pp. 1-6, 2004.