A PROSODY-ONLY DECISION-TREE MODEL FOR DISFLUENCY DETECTION

Elizabeth Shriberg[1] *Rebecca Bates*[2] *Andreas Stolcke*[1]

[1]Speech Technology and Research Laboratory, SRI International, Menlo Park, California
{ees,stolcke}@speech.sri.com; http://www.speech.sri.com
[2]Dept. of Electrical Engineering, Boston University, Boston, Massachusetts
becky@raven.bu.edu; http://raven.bu.edu

ABSTRACT

Speech disfluencies (filled pauses, repetitions, repairs, and false starts) are pervasive in spontaneous speech. The ability to detect and correct disfluencies automatically is important for effective natural language understanding, as well as to improve speech models in general. Previous approaches to disfluency detection have relied heavily on lexical information, which makes them less applicable when word recognition is unreliable. We have developed a disfluency detection method using decision tree classifiers that use only local and automatically extracted prosodic features. Because the model doesn't rely on lexical information, it is widely applicable even when word recognition is unreliable. The model performed significantly better than chance at detecting four disfluency types. It also outperformed a language model in the detection of false starts, given the correct transcription. Combining the prosody model with a specialized language model improved accuracy over either model alone for the detection of false starts. Results suggest that a prosody-only model can aid the automatic detection of disfluencies in spontaneous speech.

1. INTRODUCTION

1.1. Why detect disfluencies?

Disfluencies (filled pauses, repetitions, repairs, and false starts) are prevalent in natural, spontaneous speech. The ability to detect and correct disfluencies is clearly important for natural language understanding, since most NLU systems are trained to interpret fluent utterances. Recent studies suggest that disfluency detection is also relevant at other levels of speech processing. For example, work on statistical language modeling has shown that perplexity is reduced if disfluencies are removed from the N-gram context [12]. Additional analyses suggest that speakers hesitate before less-predictable words; thus, transition probabilities should be dynamically adjusted in the vicinity of hesitations [9]. Automatic detection of disfluencies could also benefit higher-level modeling, for example, the automatic segmentation of speech into sentences [11], and the modeling of discourse or topic structure [13].

1.2. Why use prosody?

Various approaches to automatic disfluency detection have been proposed in past work [8, 1, 7, 4]. These studies have focused on task-oriented dialog and have used a combination of lexical and prosodic features. Results have shown a heavy reliance on lexical information, although prosodic information was also useful when constrained by the lexical information.

Such approaches are limited, however, if lexical information is unreliable. In past work, a correct transcription was assumed—a reasonable approach given that error rates for the corpora used were quite low (typically under 5% word error rate). For more natural speech corpora, even state-of-the-art systems are much less accurate (e.g., about 40% WER for Switchboard as of 1996 [5]). Thus, while prosody played a lesser role in studies based on correct transcriptions, it could be an important knowledge source for detecting disfluencies when word hypotheses are less reliable.

The goal of the present work was to determine whether a prosody-only model could provide information helpful for automatic disfluency detection. In the context of this work, "prosody-only" will refer to information that does not rely on word or phone identity. The long-term goal is to develop a model using locally and automatically extracted prosodic features, and to combine this independent model with standard models in speech recognition to increase detection accuracy over that of the separate models alone.

2. METHOD

2.1. Data

Speech data consisted of more than 1000 conversations from the Switchboard corpus of human-human telephone dialogs on prescribed topics [3]. The data set represents 364 different speakers (45% male, 55% female). Data were divided into randomly selected independent training (500,000 words) and test (60,000 words) sets, with no speaker overlap. We prepared a speech database that combined information from various sources, and at various levels of resolution, including

- Word transcripts
- Hand-labeled disfluency annotations and sentence segmentations prepared by the Linguistic Data Consortium (LDC) [6]
- Phone-level time marks produced by forced alignment of the word transcripts using the SRI Decipher(TM) speech recognizer, as used in the 1996 LVCSR evaluations [5]
- Raw acoustic measurements for the prosodic features described below, such as fundamental frequency (F0) and signal-to-noise ratio (SNR) values.

To limit computation and to facilitate integration with a language model, our datapoints consisted of each inter-word boundary, as determined by the forced alignments. We note that in principle, however, the word information was unnecessary, since we could have evaluated all features at, for example, each frame.

The disfluency types examined are shown below. The "extra" words in each disfluency are indicated in boldface font. In our experiments, the goal was to automatically detect the inter-word boundary at the right edge of the boldface material (represented by "*"):

Filled pause	he **uh** * liked it
Repetition	**he** * he liked it
Repair	**he** * she liked it
False start	**it was** * he liked it

For each boundary in the database,[1] a feature vector was recorded. The feature vector contained information extracted from regions spanning from 70 frames (700 milliseconds) before the boundary to 70 frames following the boundary. The boundary itself could also contain a silent pause. The acoustic feature types examined included duration, F0, distance from a pause, and SNR features. In addition, "gender" features included the gender of the speaker as well as that of the listener.

Duration features included the duration of pauses, vowels, and continuously voiced segments preceding the boundary. Although vowel durations and pause durations were based on information from the forced alignments, they did not rely on phone or word identity, and we expect that pause alignments would approximate those for hypothesized recognition output. Durations of voiced regions were obtained independent of alignment information, using the Entropic Systems Waves/ESPS probability-of-voicing measure. F0 values were extracted using the Waves pitch tracker. F0 was measured from the speech preceding the boundary and from the speech following the boundary; the difference in F0 across the boundary, as well as the F0 derivative before the boundary were also computed. Distance features used pauses in alignments as landmarks, and computed the distance from the landmark to the boundary. SNR features were intended to capture energy of the speaker (rather than the background). Using SRI's telephone-bandwidth front end, each waveform was searched to find the noise floor, and the instantaneous SNR was computed at each time and frequency region. Various normalization methods were also used. Duration, F0, and SNR features were used both in "raw" form and "globally" normalized (using information from an entire conversation side). Certain F0 features were also normalized "locally" (using information available within 70 frames of the boundary).

2.2. Decision Trees and Language Models

For our prosodic model, we chose decision trees because they can be inspected to determine the role of features and feature combinations in classification. We used CART-style decision trees [2], a widely used data modeling algorithm convenient for replication of results. The decision trees (DTs) take a collection of acoustic features X as input and predict disfluency events D by asking questions of the features. The DT outputs posterior probability estimates $P(D|X)$.

In one case, we also compared the DT with a classifier based on a statistical language model (LM). The LM yields a joint probability $P(W, D)$, where W is the word sequence and D are the disfluency events. From this we can obtain another posterior probability estimate $P(D|W) = P(D,W)/P(W)$. The LM used was a disfluency N-gram model of the type used in [12], and was trained on 1.4 million words of Switchboard transcripts, hand-annotated for disfluencies by LDC [6].

Finally, we want to combine the DT classifier based on acoustic information with the LM classifier based on word information for a combined estimate. This can be done as follows:

$$\begin{aligned} P(D|W,X) &= \frac{P(D|X)P(W|D,X)}{P(W|X)} \\ &\approx \frac{P(D|X)P(W|D)}{P(W|X)} \qquad (1) \\ &= \frac{P(D|X)P(D|W)P(W)}{P(W|X)P(D)} \\ &\propto \frac{P(D|X)P(D|W)}{P(D)} \qquad (2) \end{aligned}$$

[1] We note that we were unable to include data for boundaries adjacent to a word fragment (a word cut off by the speaker before completion) because fragments, which are not in the lexicon, had been removed from the acoustically-segmented data in preparation for acoustic training. Therefore, results apply only to disfluencies in which no word fragments were involved. (Based on hand-labeled data for a subset of Switchboard [10], we estimate this set to comprise about 80% of all disfluencies).

Approximation (1) holds if words and acoustic features are chosen to be largely independent of one another, given D, i.e., $P(W|D,X) \approx P(W|D)$. This condition was met in our case since none of the features we included in the trees depended on word or phone identity. The proportionality in (2) is obtained by dropping all terms that are independent of D, i.e., those which can be ignored when comparing the posteriors for different values of D.

As explained below, we also downsampled all of our training and test data sets to equate the prior probabilities for different values of D. This allows us to drop $P(D)$ from Equation (2). Furthermore, because both DT and LM are approximations we insert a language model weight λ to empirically balance the dynamic ranges of the two models:

$$P(D|W,X) \propto P(D|X)P(D|W)^{\lambda}$$

This weight serves a function similar to that of the LM weight used in combining acoustic and language models for automatic speech recognition.

3. RESULTS AND DISCUSSION

Independent experiments were run for each of the four disfluency types (filled pauses, repetitions, repairs, and false starts). In each case, we used a CART-style tree with binary classes; the task for the tree was to classify each inter-word boundary as either "disfluent" or as "other" (fluent, or other type of disfluency). Performance metrics included

- **accuracy:** correct classifications / all datapoints
- **recall:** disfluencies detected / disfluencies
- **false alarm rate:** others called disfluencies / others

Prior class probabilities and therefore, chance performance vary widely for the different classification tasks. To enable comparable analyses across disfluency types and test sets we decided to downsample the data to assure an equal number of cases in each class; therefore chance performance on all three metrics was 50% in all experiments. Downsampling also yields more informative trees as the DT algorithm otherwise tends to devote very few resources to classes with low prior probabilities (assigning differential costs to classification errors would be another way to prevent this).

The accuracy measure summarizes both the false positive and the false negative errors, giving them equal weight. Since we have no reason to assign different costs to these errors in the present work, accuracy is a reasonable overall error statistic for comparison purposes.

3.1. Detection of filled pauses

The goal of the decision tree in this experiment was to discriminate boundaries following a filled pause ("uh" or "um") from all other boundaries. In the case of filled pauses, unlike other disfluencies, successful detection of the disfluency event is equivalent to correct recognition of "uh" and "um", both of which can be modeled as words. Table 1 shows results for filled pause detection using the prosody-only tree model. For comparison purposes, results using the SRI recognizer are also provided; this rate reflects recognizer performance after comparable downsampling (which effectively increases the recognition rate relative to that for the full set of data).

Table 1. Classification Rate for Filled Pauses (%)

	Leaves	Accuracy	Recall	False Al.
Tree	47	89.7	92.3	12.9
Recognizer		77.8	56.7	1.1

As shown by the accuracy measure, the prosody-only model is superior to the recognizer at discriminating filled pauses from

other words. It should be noted, however, that the prosody-only result is not strictly comparable with the recognition result, since the former is optimistic because of the knowledge of correct word boundaries. For the practical reasons noted earlier, we have not run the appropriate "fair" test, which would involve creating a database with one feature vector per *hypothesized* word boundary.

Nevertheless, we find the present results promising. We predict that even with word boundaries from recognition output, the DT model will work well because (1) most filled pauses (FPs) are followed by silent pauses, and recognition for silence is typically quite good, (2) our features do not rely on word or phone identity, as explained earlier, and (3) there is already some noise in the location of word boundaries in the forced alignment procedure used. The tree revealed that the primary features involved were duration, distance from pause, and F0 features. The leaf count is large compared to the number of features used, and the features were queried in complex sequences, suggesting that different speakers use these features differently in producing filled pauses. Assuming equal costs for false alarms and false rejections, the prosodic model has better recall than the recognizer, and the recognizer has a lower false alarm rate. Thus, a future goal is to improve recognition performance by integrating the prosodic model with standard acoustic models and with a filled-pause language model.

3.2. Detection of repetitions

Like filled pauses, repetitions should be (largely) detectable from a correct word transcription. However, given unreliable recognition, prosody could provide a helpful knowledge source for repetition detection if results for the prosody-only model are better than chance. Table 2 shows results for repetition detection.

Table 2. Classification Rate for Repetitions (%)

	Leaves	Accuracy	Recall	False Al.
Tree	31	77.5	83.5	28.5

As indicated, the accuracy of the tree model is significantly above chance. Performance is lower for repetitions than for filled pauses (see Table 1); this may occur because repetitions comprise a greater range of possible words than do filled pauses. The main features used were duration, distance from pause, and F0. The leaf count for repetitions was lower than that for filled pauses, suggesting that speakers may be more similar to each other in their prosodic production of repetitions than in their prosodic productions of filled pauses. Future work will aim to integrate prosody with acoustic and language models, to reduce the rate of false alarms.

3.3. Detection of repairs

Compared with filled pauses and repetitions, repairs are more difficult to detect based on words alone, and therefore prosody could play an important role if performance is better than chance. Table 3 shows results for the tree model in repair detection.

Table 3. Classification Rate for Repairs (%)

	Leaves	Accuracy	Recall	False Al.
Tree	11	75.5	77.0	25.9

Again, the tree model is able to classify the data at better than chance. In addition, the repair-detection tree has a low leaf count, suggesting that repairs are cued by similar features (in this case duration and distance from pause). In future work we plan to compare these rates to rates using a language model alone, and to attempt to reduce false alarms by combining prosodic and language models.

3.4. Detection of false starts

False starts are the most difficult type of disfluency to detect using lexical information, since there is no relationship between the abandoned and the following material. In addition, false starts occur at high rates in natural discourse. Thus, it would be helpful if prosody could provide a cue to their detection. Table 4 shows results for our false-start detection task using the prosody-only model, as well as results using a false-start language model (based on reference transcriptions), and results after combining the two models using an optimal language model weight. Figure 1 contains ROC plots showing the tradeoff between recall and false alarm rate for all three classfiers.

Table 4. Classification Rate for False Starts (%)

	Leaves	Accuracy	Recall	False Al.
Tree	4	74.0	74.0	26.0
False-start LM		60.7	23.7	2.3
Combined		77.9	72.8	17.1

The tree model, as shown, outperforms the language model in classification accuracy. The superiority of the prosodic model is actually underrepresented here, since the language model is likely to be less helpful given actual recognition output, whereas the prosodic model would remain unchanged. In addition, the prosodic model for false starts is extremely parsimonious; it contains only four leaves (all using normalized duration information)—a striking result given the large number of different speakers represented.

The language model, however, is nevertheless helpful for reducing the high false alarm rate associated with the prosodic model. Comparison of results for the combined DT/LM model to those for the DT alone reveals that for a given recall rate, the combined model achieves substantially fewer false alarms than the DT model, thereby improving the accuracy overall. This behavior is shown in Figure 1: the combined classifer performs as well or better than either the prosodic or the LM classifier alone over the entire operating range. Based on these results, we expect similar improvements from combining prosodic and LM classifiers for the other three disfluency types.

3.5. Feature types used

To obtain a general summary of the relative contribution of different types of features in our trees, we examined feature usage across the four disfluency detection tasks. To compare results across tasks, we looked at a measure of "relative feature usage." We counted the number of times the decision tree asked a question of a given feature, over all test samples, divided by the total number of questions asked. Note that relative feature usage is a crude way to assess the importance of a feature for a task. (A better way to gauge feature importance would be to remove a feature from the model and retrain the decision tree, for each feature in turn. This approach, however, is much more time consuming than could be justified for the present work.)

Features were grouped by type into five classes (duration,

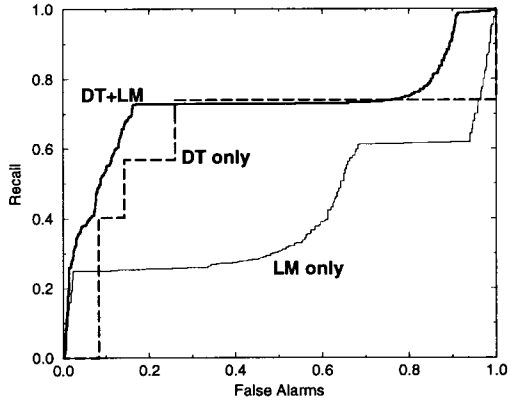

Figure 1. ROC plots for the prosodic DT, LM, and combined DT+LM classifiers.

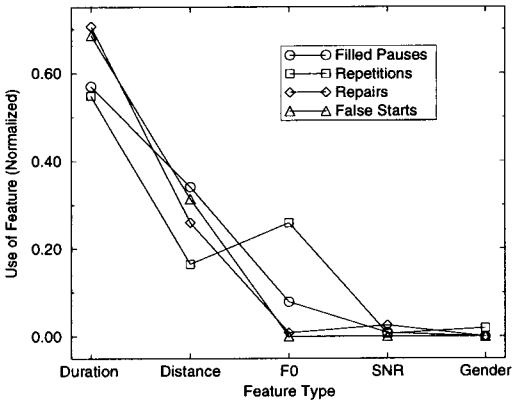

Figure 2. Feature Types Used by Task

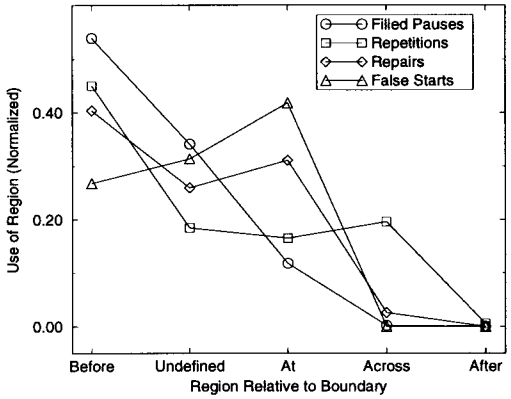

Figure 3. Feature Extraction Regions Used by Task

distance from pause, F0, SNR, and gender), collapsing over extraction region and over normalization method. Results are shown in Figure 2. As shown, the main features used across tasks were duration, distance from pause, and F0. Duration comprised both the duration of the last voiced region preceding the boundary, and the duration of the boundary itself. Both were heavily used in the trees. Independent analyses revealed that overall classification results improved when durations were globally normalized (using the mean and variance from all tokens from a particular speaker over a conversation); this is likely to be due in particular to the normalization of pause durations at the boundary.

As visible in Figure 2, F0 features were used for repetitions and to a small extent for filled pauses, but not used for the other two disfluency types. A possible explanation is that typically at boundaries containing a pause (including pauses at grammatical boundaries), F0 differences are large and quite variable. For hesitation phenomena, however (including repetitions and filled pauses, but not the other two disfluency types), intonation is "suspended" but resumed near its previous value after the pause, because the content of the utterance is not changed by the hesitation. Thus, at boundaries containing a pause, F0 across the boundary is typically lower for hesitations than for fluent tokens. Preliminary inspection of the trees is consistent with this hypothesis, but further work is warranted.

3.6. Feature use by extraction region

In addition we compared usage of the various features across tasks when features were grouped not by type, but rather by the regions from which they were extracted. Results are shown in Figure 3. As shown, classification across tasks was based almost exclusively on features extracted before or at the boundary. This suggests that disfluency detection may be possible before speech resumes. For speech applications, results suggest that endpointing could be improved by dynamically adjusting the threshold based on information preceding a pause. If preceding information suggests that a speaker is hesitating, the threshold could be increased to prevent premature cutoff; conversely, it could be decreased to speed processing if no such indication were present.

4. CONCLUSION

This work revealed that a prosody-only model performed significantly better than chance at detecting four disfluency types. The prosody model outperformed a state-of-the-art recognizer in detecting filled pauses. It also outperformed a language model in the detection of false starts, given the correct transcription. Combining the prosody model with a specialized language model improved accuracy over either model alone for the detection of false starts. The main features used in classification were duration, distance from a pause, and F0; in general, the relative usage of these features was also similar for the four different disfluency types. Classification was based almost exclusively on features extracted before or at the boundary, suggesting that on-the-fly disfluency detection could be used to improve endpointing. Results suggest that prosody is a valuable knowledge source for the automatic detection of disfluencies in spontaneous speech.

ACKNOWLEDGMENTS

This research was supported by DARPA and NSF, under NSF Grants IRI-9314967 and IRI-8905249. The views herein are those of the authors and should not be interpreted as representing the policies of DARPA or NSF. Special thanks go to Mari Ostendorf for valuable comments, and to Mitch Weintraub for feature extraction algorithms and computing resources.

REFERENCES

[1] J. Bear, J. Dowding, and E. Shriberg. Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog. In *Proc. ACL*, pp. 56–63, University of Delaware, Newark, Delaware, 1992.

[2] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, 1984.

[3] J. J. Godfrey, E. C. Holliman, and J. McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *Proc. ICASSP*, vol. I, pp. 517–520, San Francisco, 1992.

[4] P. A. Heeman and J. Allen. Detecting and correcting speech repairs. In *Proc. ACL*, pp. 295–302, New Mexico State University, Las Cruces, NM, 1994.

[5] *LVCSR Hub 5 Workshop*, Linthicum Heights, MD, 1996.

[6] M. Meteer et al. Dysfluency annotation stylebook for the Switchboard corpus. Linguistic Data Consortium, 1995. Revised June 1995 by Ann Taylor.

[7] C. H. Nakatani and J. Hirschberg. A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustical Society of America*, 95(3):1603–1616, 1994.

[8] D. O'Shaughnessy. Correcting complex false starts in spontaneous speech. In *Proc. ICASSP*, vol. I, pp. 349–352, Adelaide, Australia, 1994.

[9] E. Shriberg and A. Stolcke. Word predictability after hesitations: A corpus-based study. In *Proc. ICSLP*, vol. 3, pp. 1868–1871, Philadelphia, 1996.

[10] E. E. Shriberg. *Preliminaries to a Theory of Speech Disfluencies*. Ph.D. thesis, Department of Psychology, University of California, Berkeley, CA, 1994.

[11] A. Stolcke and E. Shriberg. Automatic linguistic segmentation of conversational speech. In *Proc. ICSLP*, vol. 2, pp. 1005–1008, Philadelphia, 1996.

[12] A. Stolcke and E. Shriberg. Statistical language modeling for speech disfluencies. In *Proc. ICASSP*, vol. 1, pp. 405–408, Atlanta, 1996.

[13] M. Swerts, A. Wichmann, and R.-J. Beun. Filled pauses as markers of discourse structure. In *Proc. ICSLP*, vol. 2, pp. 1033–1036, Philadelphia, 1996.