AUTOMATIC PUNCTUATION AND DISFLUENCY DETECTION IN MULTI-PARTY MEETINGS USING PROSODIC AND LEXICAL CUES

Don Baron[1,3] *Elizabeth Shriberg*[1,2] *Andreas Stolcke*[1,2]

[1]International Computer Science Institute, Berkeley, CA
[2]Speech Technology and Research Laboratory, SRI International, Menlo Park, CA
[3]EECS Department, University of California, Berkeley, CA
dbaron@icsi.berkeley.edu {ees,stolcke}@speech.sri.com

ABSTRACT

We investigate automatic approaches to finding "hidden" spontaneous speech events, such as sentence boundaries and disfluencies, in multi-party meetings. Hidden events are characterized prosodically by a large array of automatically extracted energy, duration, and pitch features, and are modeled by decision tree classifiers; lexical cues are modeled by N-gram language models. Both sources of information are combined in a hidden Markov model framework. Results show that combined classifiers achieve higher accuracy than either single knowledge source alone. We also study classifiers that use only the preceding context for predicting events, simulating online processing. We find that prosodic features are more robust than are language model features to this constraint. Finally, we examine the effect of automatic word recognition errors, in both training and testing, on classification accuracy. We find that lexical models degrade much more severely than do prosodic models in this case, again showing the relative robustness of prosodic information for hidden-event detection in natural conversation.

1. INTRODUCTION

Speech researchers have recently taken a greater interest in the automatic processing of natural multi-person meetings. Meetings constitute a ubiquitous form of human communication, and present unique research challenges [1, 2]. While better word recognition is an important goal in much of this work, interest is also shifting toward higher-level tasks, such as information extraction and summarization. For such tasks to succeed, information in text but not currently in speech recognition output, such as punctuation and disfluencies, must be available.

In past work we have demonstrated that prosodic information, especially when combined with lexical cues, can be effective in detecting "hidden events", such as unmarked sentence boundaries and disfluencies, in broadcast speech and in two-person telephone conversations [3]. In this paper we extend and compare those results to the new domain of multi-party meetings. Such meetings are presently recognized with high word error rates. Prosodic cues are less dependent on word identity, and should therefore be more robust to recognition errors than are lexical cues. Also, anticipating the emergence of artificial conversational agents that will ultimately participate in meetings [4], we investigate how well a system can detect hidden events when it must work online, in real-time, and therefore has access only to the past history at any given time.

Table 1. Data used in study, and word recognition error rates (WER) obtained: Even Deeper Understanding (Bed), Meeting Recorder (Bmr) and Robustness (Bro) meetings. Speech duration excludes long silent regions, but counts overlapped speech multiple times. "Spurts" are stretches of speech separated by at least 0.5 second of silence.

	Bed	Bmr	Bro	Total
Meetings	7	13	12	32
Speech duration	7.0h	13.7h	11.2h	31.9h
Transcribed words	67,546	145,150	94,261	306,957
Speech spurts	8,254	15,414	11,821	35,989
Native speakers				
Channels	20	61	41	122
WER	48.0%	43.9%	46.1%	45.2%
Nonnative speakers				
Channels	18	20	24	62
WER	62.5%	76.4%	79.2%	72.2%

2. METHOD

2.1. Data and Annotations

We processed and analyzed data from multi-party meetings collected as part of the ICSI Meeting Recorder Project [2]. We drew data from three types of Berkeley group meetings: "Meeting Recorder" (Bmr) , "Robustness" (Bro), and "Even Deeper Understanding" (Bed), with between 3 and 8 speakers each. Table 1 summarizes the amount of data in each of these meeting types. We split our corpus into a training and a non-overlapping test portion. The test portion consisted of 1 Bed, 2 Bmr, and 2 Bro meetings, chosen so as to make the total amount of data (number of words) in the test set about 18% of the total, keeping roughly equal proportions of data by meeting type across train and test data. Although the speech in training and test sets is disjoint, several speakers do appear in both sets. However, we consider this not atypical of real world applications, where meetings will involve a mix of recurring and unknown participants.

2.2. Automatic speech recognition and time alignment

After meetings were recorded, they were processed by an automatic segmentation routine [5] to detect regions of speech activity. These regions were passed to human labelers who made segmentation corrections as necessary, and then created word transcripts.

Additional labelers then added and corrected various annotations involving punctuation, disfluencies, and incomplete sentences.

The Meeting Recorder Project collects signals from both close-talking and far-field microphones. In this study we used only data from close-talking microphones, since automatic recognition from far-field microphones is currently far too errorful. Even for studies involving forced alignment of correct words, we prefer to use the high-quality signals, in order to obtain the best possible time alignments and prosodic features.

Our prosodic features rely on phone-level time alignments, which we obtained in two different ways. In experiments based on automatic speech recognition (ASR), we obtained word hypotheses and time alignments using a simplified version of the SRI Hub-5 large-vocabulary conversational speech recognizer [6]. The recognizer uses a single decoding pass, and performs channel-based vocal-tract length and cepstral normalization. It also performs unsupervised speaker adaptation using a phone-loop model. Both acoustic and language models were unchanged from the Hub-5 system, which is trained mainly on Switchboard data.

We found that fully automatic meeting segmentation currently incurs at least a 10% degradation in recognition accuracy [5], due to missed speech regions and false recognitions of nonspeech. Since we were interested in the effects of *word* recognition (as opposed to segmentation) errors on our prosodic feature extraction and modeling, we chose to perform recognition on the output of automatic segmentation with hand-corrected boundaries. ASR word error rates for each of the meetings types are also shown in Table 1; they are about 45% for native speakers and 72% for nonnative speakers. A second set of time alignments was derived from forced alignment of the reference transcripts, using the same recognition engine and acoustic models. Though not perfect, these alignments are much more accurate than those derived from ASR output, and serve as a baseline for our event-detection experiments.

2.3. Prosodic features

This section describes some of the features used in the classification tasks. Because of space limitations, only the most cursory explanations are given here; the reader is referred to [3] for a more detailed discussion of these features, although additional features are used in the present work.

The features can be divided into four main groups: pause and duration features, pitch (F0) features, energy features, and other contextual features. Pause features were computed based on alignments, and are fairly robust to recognition errors. Phone durations were obtained from ASR or forced alignments, and were normalized by phone duration statistics obtained from the Switchboard corpus. F0 features were computed by creating linear fits from median filtered raw F0 values, which were extracted using the ESPS pitch tracker `get_f0` [7]. Line fits were used to more succinctly describe general prosodic trends, since slopes can easily be determined from this data. Minimum, mean, and maximum processed F0 values were computed for any given word, and were normalized by baseline F0 values determined by a log-normal tied mixture model [8]. Features were also computed from only the last F0 value, or using a windowed range starting from the last frame of a word and stretching back N frames ($N = 10, 20, 50, 80, 100$). Windowed F0 values provide robustness against short word durations or noisy time boundaries. Using the `get_f0` RMS values, we computed minimum, maximum, and mean energy features over a word. These features were then normalized by statistics computed for the channel, in order to account for variability in microphone gain, or inherent speaker loudness. Finally, a number of non-prosodic contextual features, including speaker name, meeting type (Bmr, Bed, Bro), speaker gender, whether the speaker is a native speaker, and whether or not the speech was in a region of speaker overlap—were included as potential features for the prosodic model, as certain events and the prosodic features themselves can correlate with these contextual features.

2.4. Prosodic and language models

As in earlier work, we used CART-style decision trees [9] as classifiers to predict classes and their posterior probabilities from input features. This is a greedy algorithm, however, and in order to avoid globally suboptimal feature combinations, we used a feature selection algorithm to search for an optimal subset of input features [3]. Trees were built for both raw class distributions and distributions that had class probabilities equated by downsampling. Downsampling to a uniform class distribution allows more sensitivity to the minority classes and avoids skewing posterior probabilities towards decisions that correspond to the majority class. Also, as different meeting types may exhibit different majority/minority class distributions, downsampling can be seen as a type of normalization across meeting types. Finally, downsampling to equal priors allows for direct integration with the LM-based classifier since its posterior probability estimates are proportional to class likelihoods [3].

To model lexical information about hidden events, i.e., their cooccurrence with discourse markers, filled pauses, and high frequency words (such as the pronoun "I" that often starts sentences) we employed N-gram language models (LMs). We trained such models on annotated transcripts, where each event is represented by a tag, and is otherwise treated the same way as a word token. In testing, when only the (transcribed or automatically recognized) regular (non-event) "words" are available, the LM is evaluated as a "hidden event N-gram". That is, the event tags are treated as states in a hidden Markov model (HMM), and their probabilities (conditioned on the words) are computed via the forward-backward algorithm [3]. In all our experiments we used trigram LMs, which performed no worse than higher-order models given the amount of available data. The LM training data was always identical to that for the corresponding decision tree experiments.

Finally, we tested combinations of LMs and prosodic classifiers. Here too we used the LM as an HMM, but in this case we also computed likelihoods for the event states using the prosodic decision trees, and factored them into the computation of event posteriors. This method uses decision trees trained on downsampled data, and is an effective and efficient way of combining both types of models [3]. For comparison, we ran prosody-only experiments by using the same downsampled trees combined with a unigram LM to adjust for event priors. We note that in this case, we could have also used non-downsampled trees (thus incorporating the event priors in training) without any LM in testing. Nondownsampled trees generally yield better results in this task, but were too time-consuming to train; hence we report downsampled results only, which underestimate the prosodic model performance.

2.5. True versus recognized words

As mentioned earlier, one of our interests was the effect of word recognition errors on event detection, for both prosodic and language models. We therefore performed each experiment in three ways: using true (transcribed) words in training and in testing; using recognized words in training and in testing; and using true

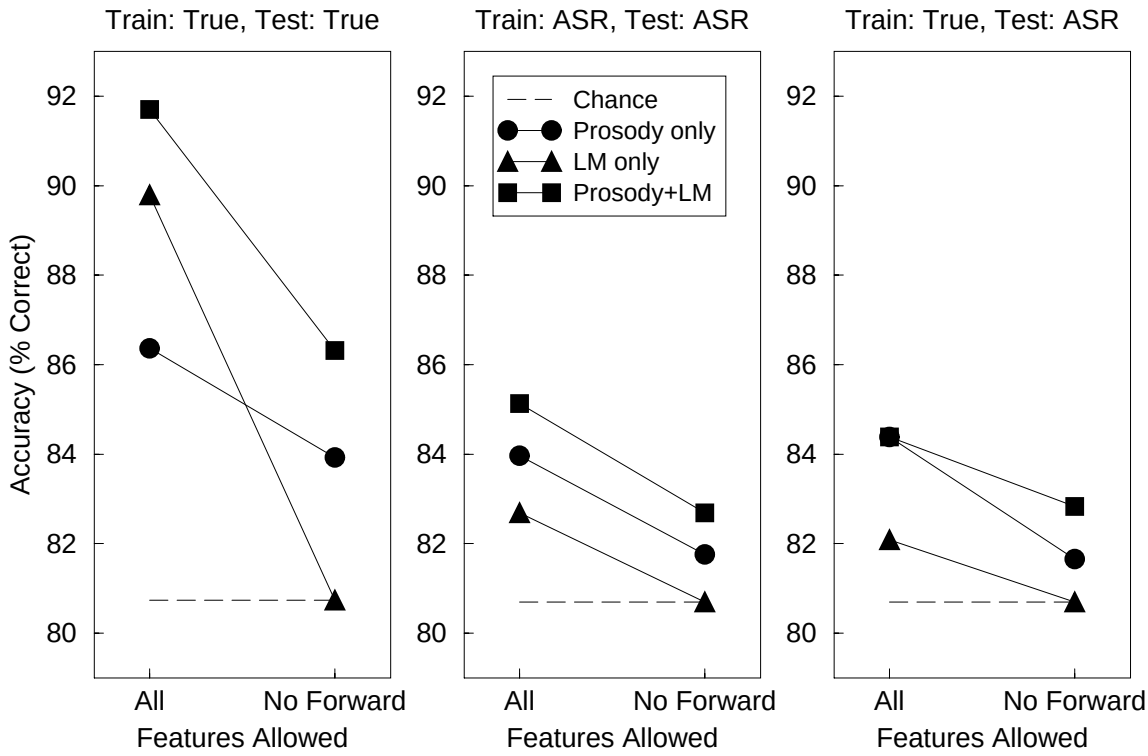

Fig. 1. Event detection accuracy (in %) using different models and different train/test conditions. "True" = true words (forced alignment); "ASR" = 1-best recognizer output; "LM" = language model.

words in training and recognized words in testing. We expect degradation of performance in the latter two conditions relative to the first, but it is not clear a priori whether it is better to train on true or recognized words. Training on true words could lead to "cleaner" models, while training on recognized words would allow the model to learn the error patterns and behave accordingly in testing. Training on true words is also less costly than training on recognized words.

Accuracy scoring for automatic speech recognition presents some interesting problems, since hypothesized and true boundaries are not in direct correspondence. We addressed this problem by first aligning hypothesized and reference words using a distance metric based on phonetic similarity, a method that can deal with fairly high word error rates. Event labels for sentence boundaries and disfluencies were then transferred to corresponding locations in the hypothesized word string, and served as reference labels for event scoring in ASR output. The same alignment procedure was applied to the training data to obtain event-labeled transcripts for training models from recognition output.

2.6. Online event classification

Both decision trees and language models can utilize cues from both before and after events, assuming processing occurs in batch mode, e.g., after the meeting has been fully recorded. However, there are situations where *online* processing is desired, for example, as part of a computer system following an ongoing meeting in real-time, possibly interacting with the participants [4]. In this situation the system would have to classify events instantaneously, using only information *preceding* the location of interest. To simulate this situation, we tested models using only features derived from the left context of a given location. For prosodic models this means limiting the available features to those computable from the region preceding the word boundary. For language models, online processing is simulated by predicting event probabilities using preceding words only (using forward probabilities only, rather than forward-backward computation).

3. EXPERIMENTS AND RESULTS

We report results for a three-way classification task in which each word boundary is to be labeled as either a sentence boundary, a disfluency interruption point (including endings of incomplete sentences), or a fluent sentence-internal word transition. In our test set, about 9% of word boundaries were sentence breaks, 10% were disfluencies or incomplete sentences, and the remaining 81% were fluent boundaries. The latter number also represents the "chance" accuracy of a classifier that always outputs the majority class.

Figure 1 shows the accuracy of three kinds of classifiers (prosody only, LM only, and combined) and with different combinations of true and recognized words in training and test. In each panel, the results using all features are on the left, and simulated online proceeding results (excluding forward-looking features) on the right. All classifiers perform well above chance, with one exception: the LM without future context (in all conditions), which cannot overcome the strong prior for the majority class in this condition. However, all classifiers suffer from the lack of future features. For example, using true words, the error rate (1 – accuracy) of the prosodic tree increases by 18% relative, and for the combined classifier by 65%. Overall, we can conclude that classifiers based on lexical information degrade more than does the prosody-only classifier when following context is removed.

We also observe prosodic classifiers to be more robust with respect to word recognition errors. All classifiers degrade severely when tested on ASR output, however the prosodic classifier less so than the word-based ones. For example, when training and testing using ASR output, the error rate of the prosodic classifier with all features increases by 17% relative, whereas the increase is 70% for the word-only classifier and 79% for the combined classifier. Nevertheless, across all conditions, the combined classifiers outperform those based on prosody or words alone, confirming our past results on other types of data.

Finally, there is an interesting difference in the way that using ASR output in training affects the different classifiers. The LM-only classifier performs better on ASR output when it is also trained on ASR output. We can surmise that the LM learns common word error patterns and how they relate to the overall event distribution. The prosodic tree, on the other hand, performs better when trained on correct words, possibly due to excessive noise in feature values resulting from incorrect word alignments.

Inspection of the prosodic model's feature usage in the different experiment conditions reveals interesting effects due to both true versus recognized words, and to the allowing of forward versus no forward features. We report feature usage as the percentage of decisions that have queried the feature type; thus, features used higher up in the tree have higher usage values. By feature type, we refer to subsets of prosodic features (duration, pitch, etc.), as space does not permit further detail. In the case of training and testing on true words, the all-features model is quite simple, with 67% of its feature usage from raw and normalized vowel and trivowel durations, and 32.5% of its usage from pause information (predominantly from the pause following the boundary in question). We note that the predominance of duration features for this task is consistent with our previous results on event detection in Switchboard [3]. This is not surprising, since we have found similarities in speaking styles in Switchboard and meetings elsewhere [10].

When the following context is not allowed however, feature usage changes. Raw and normalized vowel and trivowel durations still account for most of the decisions (52.7%), but the rest of the usage is from pitch range and slope features (17.7%), normalized energy features (12.5%), *previous* pause duration (9.9%), and presence of speaker overlap (3.4%).The effect of removing forward features is similar for the other two conditions, allowing pitch and energy features to compensate for the omission of the following pause feature. Finally, allowing no forward context and using recognized words causes the appearance of the speaker name feature to be used for about 15% of the time, presumably because the prosodic model has exhausted robust features and thus begins to utilize variations in speaker priors for the various events.

4. CONCLUSIONS

We have investigated the use of prosodic and word-based classifiers for locating sentence boundaries and disfluencies in multiparty meetings. We find that combining these two information sources yields the best results, and that word recognition errors lead to significant performance degradation, even though prosodic classifiers are less affected than word-based ones. Prosodic classifiers also degrade less than do language models when restricted to using only past information, as required for online processing.

5. ACKNOWLEDGMENTS

We thank our colleagues on the ICSI Meeting Project for providing the infrastructure for this research, as well as many stimulating discussions. Sonali Bhagat, Ashley Krupski, Rajdip Dhillon, and Kai Filion annotated the data for disfluencies and corrected punctuation. This work was funded by an ICSI DARPA Communicator project (via U. Washington), supplemented by an award from IBM. Additional support came from DARPA-TRVS, NASA, and NSF-STIMULATE (IRI-9619921) projects at SRI. The views herein are those of the authors and do not reflect the policies of the funding agencies.

6. REFERENCES

[1] A. Waibel, M. Bett, M. Finke, and R. Stiefelhagen, "Meeting Browser: Tracking and summarizing meetings", *in Proceedings DARPA Broadcast News Transcription and Understanding Workshop*, pp. 281–286, Lansdowne, VA, Feb. 1998. Morgan Kaufmann.

[2] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke, "The Meeting Project at ICSI", in J. Allan, editor, *Proc. HLT 2001*, pp. 246–252, San Diego, Mar. 2001. Morgan Kaufman.

[3] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür, "Prosody-based automatic segmentation of speech into sentences and topics", *Speech Communication*, vol. 32, pp. 127–154, Sep. 2000, Special Issue on Accessing Information in Spoken Audio.

[4] Y. Matsusaka, S. Fujie, and T. Kobayashi, "Modeling of conversational strategy for the robot participating in the group conversation", in P. Dalsgaard, B. Lindberg, H. Benner, and Z. Tan, editors, *Proc. EUROSPEECH*, vol. 3, pp. 2173–2176, Aalborg, Denmark, Sep. 2001.

[5] T. Pfau, D. P. W. Ellis, and A. Stolcke, "Multispeaker speech activity detection for the ICSI Meeting Recorder", *in Proceedings IEEE Automatic Speech Recognition and Understanding Workshop*, Madonna di Campiglio, Italy, Dec. 2001.

[6] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. R. Rao Gadde, M. Plauché, C. Richey, E. Shriberg, K. Sönmez, F. Weng, and J. Zheng, "The SRI March 2000 Hub-5 conversational speech transcription system", *in Proceedings NIST Speech Transcription Workshop*, College Park, MD, May 2000.

[7] Entropic Research Laboratory, Washington, D.C., *ESPS Version 5.0 Programs Manual*, Aug. 1993.

[8] K. Sönmez, E. Shriberg, L. Heck, and M. Weintraub, "Modeling dynamic prosodic variation for speaker verification", in R. H. Mannell and J. Robert-Ribes, editors, *Proc. ICSLP*, vol. 7, pp. 3189–3192, Sydney, Dec. 1998. Australian Speech Science and Technology Association.

[9] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth and Brooks, Pacific Grove, CA, 1984.

[10] E. Shriberg, A. Stolcke, and D. Baron, "Observations on overlap: Findings and implications for automatic processing of multi-party conversation", in P. Dalsgaard, B. Lindberg, H. Benner, and Z. Tan, editors, *Proc. EUROSPEECH*, vol. 2, pp. 1359–1362, Aalborg, Denmark, Sep. 2001.