Using Prosody for Automatic Sentence Segmentation of Multi-Party Meetings

Jáchym Kolář[1,2], Elizabeth Shriberg[1,3], and Yang Liu[1,4]

[1] International Computer Science Institute, Berkeley, CA, USA
[2] Department of Cybernetics, University of West Bohemia in Pilsen, Czech Republic
[3] SRI International, Menlo Park, CA, USA
[4] University of Texas at Dallas, TX, USA
{jachym,ees,yangl}@icsi.berkeley.edu

Abstract. We explore the use of prosodic features beyond pauses, including duration, pitch, and energy features, for automatic sentence segmentation of ICSI meeting data. We examine two different approaches to boundary classification: score-level combination of independent language and prosodic models using HMMs, and feature-level combination of models using a boosting-based method (BoosTexter). We report classification results for reference word transcripts as well as for transcripts from a state-of-the-art automatic speech recognizer (ASR). We also compare results using the lexical model plus a pause-only prosody model, versus results using additional prosodic features. Results show that (1) information from pauses is important, including pause duration both at the boundary and at the previous and following word boundaries; (2) adding duration, pitch, and energy features yields significant improvement over pause alone; (3) the integrated boosting-based model performs better than the HMM for ASR conditions; (4) training the boosting-based model on recognized words yields further improvement.

1 Introduction

Standard automatic speech recognition systems output only a raw stream of words, leaving out important structural information such as punctuation. Punctuation, in particular that associated with sentence boundaries, is crucial to human readability. Sentence boundaries also benefit various natural language processing techniques (e.g., machine translation, information extraction and retrieval, text summarization) which are typically trained on formatted input such as text. Previous efforts in sentence segmentation have studied the role of both lexical and prosodic features, in data from news broadcasts (mostly read speech) and from spontaneous telephone conversations (two-party conversations) [1–9]. Work on multi-party meetings has been more recent, and has generally examined the use of prosody for sentence segmentation using only pause information, for example [10–12]. In this paper we explore the use of prosodic features beyond pauses, including duration, pitch, and energy features, for automatic sentence segmentation of a large set of data from the publicly available ICSI meeting corpus [13].

2 Method

2.1 Speech Data and Experimental Setup

The ICSI meeting corpus [13] contains approximately 72 hours of multichannel conversational speech data. For the sentence segmentation experiments herein, we used 73 out of the total 75 available meetings (two meetings were excluded because of their very different character from the rest of the data). The 73 meetings were split into a training set (51 meetings, 539k words), a development set (11 meetings, 110k words), and a test set (11 meetings, 102k words). The test set contains unseen speakers, as well as speakers appearing in the training data as it is typical for the real world applications.

A crucial step when performing sentence segmentation of spontaneous speech is to define the notion of a "sentence", since spontaneous utterances do not consist of sentences as defined in written text. Although the original manual transcripts of the ICSI corpus do contain punctuation, the punctuation is highly inconsistent. Transcribers were instructed to focus on transcribing words as quickly as possible; there was not a focus on consistency or conventions for marking punctuation. As a result, different transcribers used different approaches to punctuation annotation. We used instead punctuation marks from a project on annotation of dialog acts in the same corpus [14, 15]. In this annotation pass, labelers carefully annotated both dialog acts and their boundaries, using a set of segmentation conventions for the latter.

For training and testing our models we have used both forced alignment of reference transcripts and ASR output. Recognition results were obtained using the state-of-the-art SRI CTS system [16], which was trained using no acoustic data or transcripts from the analyzed meeting corpus. To represent a completely automatic system, we used automatic speech/nonspeech segmentation. Word error rates for this difficult data are still quite high, the used ASR system performed at 38.2% (on the whole corpus). To generate the "reference" sentence boundaries for the ASR words, we aligned the reference setup to the ASR hypotheses with the constraint that two aligned words may not occur further apart than a fixed time. The possible sentence boundaries for ASR output were then merged from corresponding aligned words from the reference. Since the ASR hypotheses tend to miss short backchannels that are usually followed by a sentence boundary, sentence boundaries are less frequent (in our data 13.9%) than in reference conditions (15.9%).

2.2 Prosodic Features

We developed a database of 270 prosodic features describing pause, pitch, duration, and energy information in the vicinity of each word boundary, inspired by [2, 17]. Features were extracted directly from the automatically aligned speech signal, so that no hand-labeling of prosody (such as ToBI) was necessary in model training. A number of features were highly correlated, differing only in the normalization approach. To reduce the feature space, we combined similar features

into groups, and then selected the features from each group that were most frequently used in a first set of decision trees. We then used the resulting smaller set of 40 features to train the models reported in this paper.

Pause features consisted of the pause duration after the current, previous, and following words. Duration features included the duration of vowels, final rhymes, and whole words, aiming mainly to reflect the phenomenon of preboundary lengthening. We used raw durations as well as duration features normalized using phoneme duration statistics from the whole database. Pitch features included features describing minimal, maximal and mean values, f_0 slopes, and differences and ratios of values across word boundaries. These features were extracted both from raw f_0 value and from a f_0 contour stylized by a piece-wise linear function. Energy features were represented by maximal, minimal, and mean frame-level RMS values, both raw and per-channel normalized. Statistics showing which prosodic feature were used by our models are provided at the end of Section 3.

2.3 Classifiers

We report results obtained using two different approaches: (1) a combination of independent language and prosodic models in an HMM framework, and (2) a boosting-based algorithm (BoosTexter) that uses one integral model containing both lexical and prosodic features.

HMM Approach The approach for sentence boundary detection from speech that has received the most attention in recent years is a hidden Markov model (HMM) [1, 2, 5]. This approach provides a convenient way for combining lexical and prosodic features and is computationally efficient. In the HMM, the word/event pairs correspond to states, and the words as well as other (in our case prosodic) features correspond to observations. That is, the words appear both in the states and in the observations, with the transition probabilities given by the N-gram language model. Transition probabilities are estimated using standard N-gram techniques from text data, in which sentence boundaries are marked by a special tag (which is for training purposes treated in the same way as other word tokens). The HMM observation likelihoods are estimated by converting posteriors obtained by a prosodic classifier into likelihoods, under the assumption that prosodic features depend only on the events, and not on the words.

After several simplifications [2, 5], to combine prosodic and lexical scores we use the relation

$$P(e_i|F,W) \propto P(e_i|W) \left(\frac{P(e_i|f_i)}{P(e_i)} \right)^{\lambda} \tag{1}$$

where e_i and f_i stand for i-th event (type of boundary, in our case "sentence boundary" or "no boundary") and vector of prosodic features, respectively, and W and F for the sequences of words and prosodic features, respectively. λ is

an exponential scaling factor estimated using held-out data, which allows us to weight the relative contributions from the two models.

In past work, the most popular classifiers for estimating posteriors $P(e_i|f_i)$ have been decision trees, since they handle features with undefined values, combine continuous and categorical features, and are easily human-readable and interpretable. When training a sentence boundary classifier, we have to deal with the problem of imbalanced data [18], since sentence boundaries occur only at approximately 16% of all word boundaries in our corpus. The skewed distribution of training data may cause decision trees to miss out on inherently valuable features that are dwarfed by data priors. One solution to this problem is to train classifiers on data downsampled to equal class priors. To take advantage of all available data, we apply ensemble sampling instead of simple downsampling. Ensemble sampling is performed by randomly splitting the majority class into int(N) nonoverlapping subsets, where N is the ratio between the number of samples in the majority and minority classes. Each subset is joined with all minority class samples to form int(N) balanced sets to train classifiers. It is also advantageous to employ bagging [19], which decreases classifier variance by averaging results obtained by multiple classifiers. The classifiers are trained from different datasets sampled with replacement from the original training set. A combination of these two methods (applying bagging on ensemble samples) makes up *ensemble bagging*, which we used in our experiments.

Boosting-based Approach The HMM approach has also a couple of disadvantages. The combination of prosodic and lexical models makes strong independence assumptions, which are not typically met in actual language data. Moreover, the HMM training maximizes the joint probability of data and hidden events, but a criterion more closely related to classification error is the posterior probability of the correct hidden variable assignment given the observations. To overcome these drawbacks, models based on maximum entropy [4, 6] or conditional random fields [9] have been proposed in past work. These models provide a more principled way to combine prosodic and overlapping lexical features. In this work, we explore a different approach, by integrating prosodic and lexical features into one model based on boosting. We provide some detail here on the approach, since it has not been described for this task in previous work.

The principle of boosting is to combine many weak learning algorithms to produce an accurate classifier. The algorithm generates weak classification rules by calling the weak learners repeatedly in series of rounds. Each weak classifier is built based on the outputs of previous classifiers, focusing on the samples that were formerly classified incorrectly. This general method can be basically combined with any classifier. We used an algorithm called BoosTexter, that combines weak classifiers having a basic form of one-level decision trees using confidence-rated predictions [20]. The test at the root of each tree can check for the presence of a word N-gram, or for a value of a continuous feature. This allows a straightforward combination of lexical and prosodic features.

For training the classifier, we have used exactly the same set of prosodic features as was used for prosodic classification using decision trees. We have added the following N-gram features (w_0 denotes the word before the classified boundary, w_1 the word after the boundary, and so on) – unigrams: w_0, w_1, bigrams: w_0w_1, $w_{-1}w_0$, trigrams: $w_{-1}w_0w_1$, $w_0w_1w_2$, and a binary feature indicating whether the word before the boundary is identical with the following word.

3 Experimental Results and Discussion

Our experimental results testing both approaches in reference and ASR conditions are summarized in Table 1. As an evaluation metric we use a "boundary error rate" described by

$$E = \frac{I + M}{N_W} \quad [\%] \tag{2}$$

where I denotes the number of false sentence boundary insertions, M the number of misses, and N_W the number of words in the test set. For each system, we report error rates using lexical features alone, pause durations alone (including both pause duration at the boundary, and at the previous and following word boundaries), all prosodic features alone, lexical features plus pause durations, and lexical features plus all prosodic features. We also present chance performance, or the performance achieved by classifying every word boundary as the class having the highest prior probability (which is "no sentence boundary" in our case). Note that chance performance differs for reference versus ASR conditions as described in Section 2.1. To enable a comparison of relative gain from each set of features for both conditions, we also report the relative error reduction with respect to chance error.

Results for the reference condition (REF) indicate that the language model alone performs better than the prosodic model alone, and either model is outperformed by a combination model. The gain from additional prosodic features (beyond pause) is larger when lexical information is not accessible. However, when combining with the lexical model, there is a significant gain from adding prosodic features beyond pauses, both for the HMM (0.12% absolute, 2.0% relative, $p < 0.01$) and for BoosTexter (0.10% absolute, 1.7% relative, $p < 0.01$). Results using both approaches are very similar for the reference condition.

Results for the ASR condition show that recognition errors cause more degradation for lexical than for prosodic features. Note that while degradation is expected for lexical features, it is not the case that prosodic features should be completely robust to word error: some prosodic features depend on phone or word boundary information for extraction or normalization. The level of degradation for individual models is also visible from values of the exponential weight λ (optimized on development data), which is used for combining the prosodic and language models in the HMM. For the reference condition, the optimal value was 0.8 (giving a slightly higher weight to the lexical model), while for the ASR condition it was 1.1 (actually giving a higher weight to the prosodic model).

Table 1. Sentence boundary detection error rates (defined as count of false alarms and misses divided by the total number of words [%]) for different models (HMM trained on reference, BoosTexter trained on reference /TrREF/ and BoosTexter trained on recognized words /TrASR/) and test conditions (REFerence and ASR), numbers within parentheses correspond to relative reductions over chance error rate

		HMM approach	BoosTexter-TrREF	BoosTexter-TrASR
REF	chance	15.92 (0.0)	15.92 (0.0)	N/A
	LM	7.47 (53.1)	7.73 (51.4)	N/A
	pause	8.96 (43.7)	8.78 (44.8)	N/A
	all prosody	8.06 (49.4)	8.28 (48.0)	N/A
	LM+pause	5.89 (63.0)	5.88 (63.1)	N/A
	LM+all prosody	5.77 (63.8)	5.78 (63.7)	N/A
ASR	chance	13.85 (0.0)	13.85 (0.0)	13.85 (0.0)
	LM	9.43 (31.9)	9.59 (30.8)	9.48 (31.6)
	pause	8.97 (35.2)	8.85 (36.1)	8.82 (36.3)
	all prosody	8.30 (40.1)	8.31 (40.0)	8.35 (39.7)
	LM+pause	7.03 (49.2)	6.84 (50.6)	6.81 (50.8)
	LM+all prosody	7.00 (49.4)	6.67 (51.8)	6.58 (52.5)

An interesting observation is that the BoosTexter model using both prosodic and textual features proved to be more robust to recognition errors than the HMM. Note that for the language model and prosody alone, boosting does not help. However, when using both prosodic and lexical cues, it yields some gain over the HMM. This fact supports the hypothesis that it is advantageous to more tightly integrate prosodic and textual features. We tried training BoosTexter on recognized rather than reference words, to see how training on data more matched to the test data (i.e., containing many word errors) would affect performance. As shown, this approach outperforms the approach of training on reference in the case of boosting, but not when using the HMM. Finally, there was a significant gain from adding prosodic features beyond pauses. For the best BoosTexter model the gain was 0.23% absolute and 3.4% relative, significant at $p < 0.001$.

To explore which prosodic features were useful in this task, we analyzed prosodic decision trees from the HMM approach, because they are much easier to interpret than the resulting BoosTexter model. We used the measure "feature usage" [2], which counts how many times (by token) each feature is queried in a decision tree. Results were averaged over all trees generated during ensemble bagging. The statistics for each group of features as well as the best features from each group are listed in Table 2. The statistics show that the most frequently used features were pause duration after the current word, raw word duration, pause after the following word, and normalized duration of the last rhyme in the word. Sums of usages of features from the four basic groups show that those most frequently queried in decision trees were duration features, followed by pause, pitch, and energy features.

Table 2. Prosodic feature usage (percentage of total feature usage) for groups of prosodic features

group	total usage	two most frequently used features from each group
pause	24.8	pause after current word(16.1), pause after previous word(5.7)
duration	48.9	word duration(9.3), last rhyme normalized duration(5.5)
pitch	21.4	first slope of following word(3.6), min f_0 of last voic. region(3.2)
energy	4.9	mean RMS following word(2.3), min of voiced norm RMS (1.5)

4 Summary and Conclusions

We explored the use of prosody including pauses, duration, pitch, and energy features, for automatic sentence segmentation of a large set of data from the ICSI meeting corpus. We have examined two different approaches to the boundary classification: HMM and a boosting-based classifier BoosTexter. Results indicate that (1) information from pauses is important, including pause duration both at the boundary, and at the previous and following word boundaries; (2) adding duration, pitch, and energy features yields a further significant improvement; (3) an integrated boosting-based model performs better than an HMM for ASR conditions; (4) training the boosting-based model on recognized words yields additional improvement.

From this work, we conclude that prosody can make an important contribution to meeting understanding, via helping to find boundaries of sentences or dialog acts. Features beyond pauses are worth exploring in future work, as are modeling techniques that can tightly integrate prosodic and lexical features. Finally, for corpora in which state-of-the-art ASR performance is still rather poor, it may be useful to train models on recognized rather than reference words.

5 Acknowledgments

The authors thank Dilek Hakkani-Tur and Gokhan Tur for help with the boosting software, and Ozgur Cetin for generating ASR output. This work was supported by the European Union 6th FWP ISR Integrated Project AMI (FP6-506811), the DARPA CALO project (NBCHD-030010), NSF project IIS-0121396, DARPA Contract No. HR0011-06-C-0023, and by the Academy of Sciences of the Czech Republic (project No. 1QS101470516). The views expressed are those of the authors, and not the funding agencies.

References

1. Stolcke, A., Shriberg, E., Bates, R., Ostendorf, M., Hakkani, D., Plauche, M., Tur, G., Lu, Y.: Automatic Detection of Sentence Boundaries and Disfluencies Based on Recognized Words. In: Proc. ICSLP 98, pp. 2247–2250, Sydney (1998)

2. Shriberg, E., Stolcke, A., Hakkani-Tur, D., Tur, G.: Prosody-based Automatic Segmentation of Speech into Sentences and Topics. In: Speech Communication, vol. 32, no. 1–2, p. 127–154 (2000)
3. Warnke, V., Kompe, R., Niemann, H., Nöth, E.: Integrated Dialog Act Segmentation and Classification Using Prosodic Features and Language Models. In: Proc. EUROSPEECH 97, pp. 207–210, Rhodes, Greece (1997)
4. Huang, J., Zweig, G.: Maximum Entropy Model for Punctuation Annotation from Speech. In: Proc. ICSLP 2002, pp. 917–920, Denver (2002)
5. Kim, J.H., Woodland, P.: A Combined Punctuation Generation and Speech Recognition System and Its Performance Enhancement Using Prosody. In: Speech Communication, vol. 41, no. 4, pp. 563–577 (2003)
6. Liu, Y., Stolcke, A., Harper, M., Shriberg, E.: Comparing and Combining Generative and Posterior Probability Models: Some Advances in Sentence Boundary Detection in Speech. In: Proc. EMNLP, Barcelona, Spain (2004)
7. Liu, Y., Shriberg, E., Stolcke, A., Hillard, D., Ostendorf, M., Peskin, B., Harper, M.: The ICSI-SRI-UW Metadata Extraction System. In: ICSLP 2004, Jeju, Korea (2004)
8. Kolář, J., Švec, J., Psutka, J.: Automatic Punctuation Annotation in Czech Broadcast News Speech. In: Proc. SPECOM 2004, St. Petersburg, Russia (2004)
9. Liu, Y., Stolcke, A., Shriberg, E., Harper, M.: Using Conditional Random Fields for Sentence Boundary Detection in Speech. In: Proc. ACL pp. 451–458, Ann Arbor (2005)
10. Ang, J., Liu, Y., Shriberg, E.: Automatic Dialog Act Segmentation and Classification in Multiparty Meetings. In: Proc. IEEE ICASSP-2005, pp. 1061–1064, Philadelphia (2005)
11. Ji, G., Bilmes, J.: Dialog Act Tagging Using Graphical Models. In: Proc. IEEE ICASSP-2005, pp. 33–36, Philadelphia (2005)
12. Zimmermann, M., Stolcke, A., Shriberg, E.: Joint Segmentation and Classification of Dialog Acts in Multiparty Meetings. In Proc.: IEEE ICASSP-2006, Toulouse, France (2006)
13. Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., Wooters, Ch.: The ICSI Meeting Corpus. In: Proc. IEEE ICASSP-2003, pp. 364–367, Hong Kong (2003)
14. Dhillon, R. et al.: Meeting Recorder Project: Dialog Act Labeling Guide. ICSI Technical Report TR-04-02, International Computer Science Institute, Berkeley (2004)
15. Shriberg, E. et al.: The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. In: Proc. SIGDIAL, Cambridge, MA, USA (2004)
16. Zhu, Q., Stolcke, A., Chen, B., Morgan, N.: Using MLP Features in SRI's Conversational Speech Recognition System. In: Proc. INTERSPEECH 2005, pp. 2141–2144, Lisboa (2005)
17. Buckow, J., Warnke, V., Huber, R., Batliner, A., Nöth, E., Niemann, H.: Fast and Robust Features for Prosodic Classification. In: Proc. TSD'99 Marienbad, pp. 193–198, Springer Verlag, Berlin (1999)
18. Liu, Y., Shriberg, E., Stolcke, A., Harper, M.: Using Machine Learning to Cope with Imbalanced Classes in Natural Speech: Evidence from Sentence Boundary and Disfluency Detection. In: Proc ICSLP 2004, Jeju, Korea (2004)
19. Breiman, L.: Bagging Predictors. In: Machine Learning 24(2), pp. 123–140 (1996)
20. Schapire, R.E., Singer, Y.: BoosTexter: A Boosting-based System for Text Categorization. In: Machine Learning, 39(2/3), pp. 135–168 (2000)