An Analysis of Sentence Segmentation Features for Broadcast News, Broadcast Conversations, and Meetings

Sebastien Cuendet[1]
cuendet@icsi.berkeley.edu

Elizabeth Shriberg[1,2]
ees@speech.sri.com

Benoit Favre[1]
favre@icsi.berkeley.edu

James Fung[1]
jgf@icsi.berkeley.edu

Dilek Hakkani-Tur[1]
dilek@icsi.berkeley.edu

[1]ICSI
1947 Center Street
Berkeley, CA 94704, USA

[2]SRI International
333 Ravenswood Ave
Menlo Park, CA 94025, USA

ABSTRACT

Information retrieval techniques for speech are based on those developed for text, and thus expect structured data as input. An essential task is to add sentence boundary information to the otherwise unannotated stream of words output by automatic speech recognition systems. We analyze sentence segmentation performance as a function of feature types and transcription (manual versus automatic) for news speech, meetings, and a new corpus of broadcast conversations. Results show that: (1) overall, features for broadcast news transfer well to meetings and broadcast conversations; (2) pitch and energy features perform similarly across corpora, whereas other features (duration, pause, turn-based, and lexical) show differences; (3) the effect of speech recognition errors is remarkably stable over features types and corpora, with the exception of lexical features for meetings, and (4) broadcast conversations, a new type of data for speech technology, behave more like news speech than like meetings for this task. Implications for modeling of different speaking styles in speech segmentation are discussed.

General Terms

Prosodic Modeling, Sentence Segmentation

Keywords

Spoken Language Processing, Sentence Segmentation, Broadcast Conversations, Spontaneous Speech, Prosody, Word Boundary Classification, Boosting.

1. INTRODUCTION

We investigate the role of identically-defined lexical and prosodic features when applied to the same task across three different speaking styles—broadcast news (BN), broadcast conversations (BC), and face-to-face multi-party meetings (MRDA). We focus on the task of automatic sentence segmentation, or finding boundaries of sentence units in the otherwise unannotated (devoid of punctuation, capitalization, or formatting) stream of words output by a speech recognizer.

Sentence segmentation is of particular importance for speech understanding applications, because techniques aimed at semantic processing of speech input—such as machine translation, question answering, information extraction—are typically developed for text-based applications. They thus assume the presence of overt sentence boundaries in their input [9, 7, 3]. In addition, many speech processing tasks show improved performance when sentence boundaries are provided. For instance, speech summarization performance improves when sentence boundary information is provided, as observed in [2]. Similarly, named entity extraction and part-of-speech tagging in speech is improved using sentence boundary cues in [4], and the use of sentence boundaries for machine translation is shown to be beneficial for machine translation in [8] . Sentence boundary annotation is also important for aiding human readability of the output of automatic speech recognition systems [5], and could be used for determining semantically and prosodically coherent boundaries for playback of speech to users in tasks involving audio search.

While sentence segmentation of broadcast news, and to some extent of meetings, has been studied in previous work, little is known about broadcast conversations. Indeed, data for this task has only recently become available for work in speech technology. Studying the properties of broadcast conversations and comparing them with those of meetings and broadcast news is of interest both theoretically, and also practically, especially because there is currently less data available for broadcast conversations than for the other two types studied here. For example, if two speaking styles share characteristics, one can perform adaptation from one to another to improve the performance of the sentence segmentation, as proved previously for meetings by using conversational telephone speech [1].

The goal of this study is to analyze how different sets of features, including lexical features, prosodic features, and

their combination, perform on the task of automatic sentence segmentation for different speaking styles. More specifically we ask the following questions:

1. How do different feature types perform for the different speaking styles?
2. What is the effect of speech recognition errors on performance, and how does this effect depend on the feature types or on the speaking style?
3. For this task, are broadcast conversations more like broadcasts or more like conversations?

Results have implications not only for the task of sentence boundary detection, but more generally for prosodic modeling for natural language understanding across genres.

The next section describes the data set, features, and approach to sentence segmentation. Section 3 reports on experiments with prosodic and lexical features, and provides further analysis and a discussion of usage of various feature types (or groups) and comparison across speaking styles. A summary and conclusions are provided in Section 4.

2. METHOD

2.1 Data and annotations

To study the differences between the meetings, BN and BC speech for the task of sentence segmentation, we use the ICSI Meetings (MRDA) [12], the TDT4 English Broadcast News [15], and the GALE Y1Q4 Broadcast Conversations corpora.

The ICSI Meeting Corpus is a collection of 75 meetings, including simultaneous multi-channel audio recordings, word-level orthographic transcriptions. The meetings range in length from 17 to 103 minutes, but generally run just under an hour each, summing to 72 hours. We use a 73 meeting subset of this corpus that was also used in the previous research [12] with the same split into training, held-out and test sets. TDT4 Corpus was collected by LDC and includes multilingual raw material, news wires and other electronic text, web audio, broadcast radio and television. We use a subset of TDT4 English broadcast radio and television data in this study. The GALE Y1Q4 Broadcast Conversations Corpus, also collected by LDC, is a set of 47 in-studio talk shows with two or more participants, including formal one-on-one interviews and debates with more participants. Three shows last for half an hour and the rest of the shows run an hour each, for a total of about 45 hours.

In the experiments to follow, classification models are trained on a set of data, tuned on a held-out set, and tested on an unseen test set, within each genre. The corpora are available with the words transcribed by humans (reference) and with the words output by the speech recognizer (STT). For the reference conditions, word start and end times are obtained by using a flexible alignment procedure [14]. Reference boundaries in speech recognizer output and flexible alignments are obtained by aligning these with manual transcriptions with annotations. Statistics on these data sets are shown in Table 1 for the STT conditions.

Note that the three different speaking styles differ significantly in mean sentence length, with sentences in meetings being only about half the length on average as those in broadcast news. Meetings (and conversational speech in general) tend to contain syntactically simpler sentences and significant pronominalization. News speech is typically read from a transcript, and more closely resembles written text. It contains for example appositions, center embeddings, and proper noun compounds, among other characteristics, that contribute to longer sentences. Discourse phenomena also obviously differ across corpora, with meetings containing more turn exchanges, incomplete sentences, and higher rates of short backchannels (such as "yeah" and "uhhuh") than speech in news broadcasts and in the broadcast conversations.

	MRDA	TDT4	BC
Training set size	456,486	800,000	270,856
Test set size	87,576	82,644	40,598
Held-out set size	98,433	81,788	37,817
Vocabulary size	11,894	21,004	12,502
Mean sentence length	7.7	14.7	12.6

Table 1: Data set statistics. Values are given in number of words, based on the output of the speech recognizer (STT).

Sentence boundary locations are based on reference transcriptions for all three corpora. Sentences boundaries are annotated in BN transcripts directly. For the meeting data, boundaries are obtained by mapping dialog act boundaries to sentence boundaries. The meetings data are labeled according to 5 classes of dialog acts: backchannels, floor-grabbers and floor-holders, questions, statements, and incompletes. In order to be able to compare the three corpora, all dialog act classes are mapped to the sentence boundary class. The BC Corpus frequently lacked sentence boundary annotations, but included line breaks and capitalization as well as dialog act tag annotations. In order to use this data, we implemented heuristic rules based on human analysis to produce punctuation annotations. For BC data, we similarly mapped the 4 types of dialog acts defined (statements, questions, backchannels, and incompletes) to sentence boundaries. Note that we have chosen to map incomplete sentence boundaries to the boundary class, even though they are not "full" boundaries. This is because the rate of incompletes, while not negligible, was too low to allow for adequate training of a third class in the BC data given the size of the currently available data. We thus chose to group it with the boundary class, even though incompletes also share some characteristics with non-boundaries. (Namely, material to the left of an incomplete resembles non-boundaries, whereas material to the right resembles boundaries).

2.2 Automatic speech recognition

Automatic speech recognition results for the ICSI Meetings data, the TDT4 data and the BC data were obtained using the state-of-the-art SRI conversational speech recognition system [17], BN system [16], and BC system [14], respectively. The meetings recognizer was trained using no acoustic data or transcripts from the analyzed meetings corpus. The word error rate for the recognizer output of the complete meetings corpus is 38.2%. Recognition scores for the TDT4 corpus is not easily definable as only closed captions are available that frequently do not match well with the actual words of the broadcast news shows. The estimated word error rate lies between 17% and 19%. The word error rate for the recognizer output of the BC data is 16.8%.

2.3 Features

Sentence segmentation can be seen as a binary classification problem, in which every word boundary has to be labeled as a sentence boundary or as a non-sentence boundary[1]. We define a large set of lexical and prosodic features, computed automatically based on the output of a speech recognizer.

Lexical features.

Previous work on sentence segmentation in broadcast news speech and in telephone conversations has used lexical and prosodic information [13, 6]. Additional work has studied the contribution of syntactic information [10]. Lexical features are usually represented as N-grams of words. In this work, lexical information is represented by 5 N-gram features for each word boundary: 3 unigrams, 2 bigrams and 1 trigram. Naming the word preceding the word boundary of interest as the *current* word, and the preceding and following words as the *previous* and *next* word respectively, the 5 lexical features are as follows:

- unigrams: {previous}, {current}, {next},
- bigrams: {current, next},
- trigram: {previous, current, next}.

Prosodic Features.

Prosodic information is represented using mainly continuous values. We use 68 prosodic features, defined for and extracted from the regions around each inter-word boundary. Features include pause duration at the boundary, normalized phone durations of the word preceding the boundary, and a variety of speaker-normalized pitch features and energy features preceding, following, and across the boundary. Features are based in part on those described in [13]. The extraction region around the boundary comprises either the words or time windows on either side of the boundary. Measures include the maximum, minimum, and mean of pitch and energy values from these word-based and time-based regions. Pitch features are normalized by speaker, using a method to estimate a speaker's baseline pitch as described in [13]. Duration features, which measure the duration of the last vowel and the last rhyme in the word before the word boundary of interest, are normalized by statistics on the relevant phones in the training data. We also include "turn" features based on speaker changes.

2.4 Boosting Classifiers

For classification of word boundaries, we use the AdaBoost algorithm [11]. Boosting aims to combine weak base classifiers to come up with a strong classifier. The learning algorithm is iterative. In each iteration, a different distribution or weighting over the training examples is used to give more emphasis to examples that are often misclassified by the preceding weak classifiers. For this approach, we use the BoosTexter tool described in [11]. BoosTexter handles both discrete and continuous features, which allows for a convenient incorporation of the prosodic features described above (no binning is needed). The weak learners are one-level decision trees (stumps).

[1]More detailed models may distinguish questions from statements, or complete from incomplete sentences.

2.5 Metrics

The quality of a sentence segmentation is usually computed with F-measure and NIST error. The F-measure is the harmonic mean of the recall and precision measures of the sentence boundaries hypothesized by the classifier to the ones assigned by human labelers. The NIST error rate is the ratio of the number of wrong hypotheses made by the classifier to the number of reference sentence boundaries. In this work, we report only the F-Measure performances.

2.6 Chance performance computation

What is of interest in the following experiments is the performance gain obtained by the classifier towards the baseline performance that one would achieve without any knowledge about the data but the prior of the classes. The easiest way of doing so is to compute the prior probability $p_t(s)$ of having a sentence boundary on the training set, and classify each word boundary in the test set as a sentence boundary with probability $p_t(s)$. Concretely, the chance score is evaluated by computing the probability of each error and correct class (true positives, false positives, and false negatives) and the ensuing value for the F-Measure computation. The final chance performance only depends on the prior probabilities of having a sentence boundary on the training set and on the test set. Therefore, the chance performance can differ slightly on the reference and STT experiments, due to word insertion and deletion errors introduced by automatic speech recognition.

3. RESULTS AND DISCUSSION

This section discusses results for the three different corpora, in two subsections. The main section, Section 3.1, presents results for feature groups (e.g., pitch, energy, pause) and for combinations of the groups. Section 3.2 examines feature subgroups within the pitch and energy features (for example, pitch reset versus pitch range features), to gain further understanding of which features contribute most to performance.

3.1 Performance by feature group

Performance results for experiments using one or more feature types are summarized for reference in Table 2. To convey trends, they are plotted in Figure 1; lines connect points from the same data set for readability. The feature conditions on the X-axis are arranged in approximate order of performance for the best-performing condition, i.e. for MRDA using reference transcriptions.

Although chance performance is higher for MRDA than for the broadcast corpora, consistent with the shorter average sentence length in meetings, all corpora have low chance performance. While chance performance changes slightly for reference versus automatic transcriptions, they are close within a corpus. As a consequence, one can compare the F-Measure results almost directly across conditions. To simplify the discussion, we define δ the relative error reduction. Since the F-Measure is a harmonic mean of two error types, one can compute the relative error reduction for a model with F-Measure F and the associated chance performance c as:

$$\delta = \frac{(1-c)-(1-F)}{1-c} = \frac{F-c}{1-c} \quad (1)$$

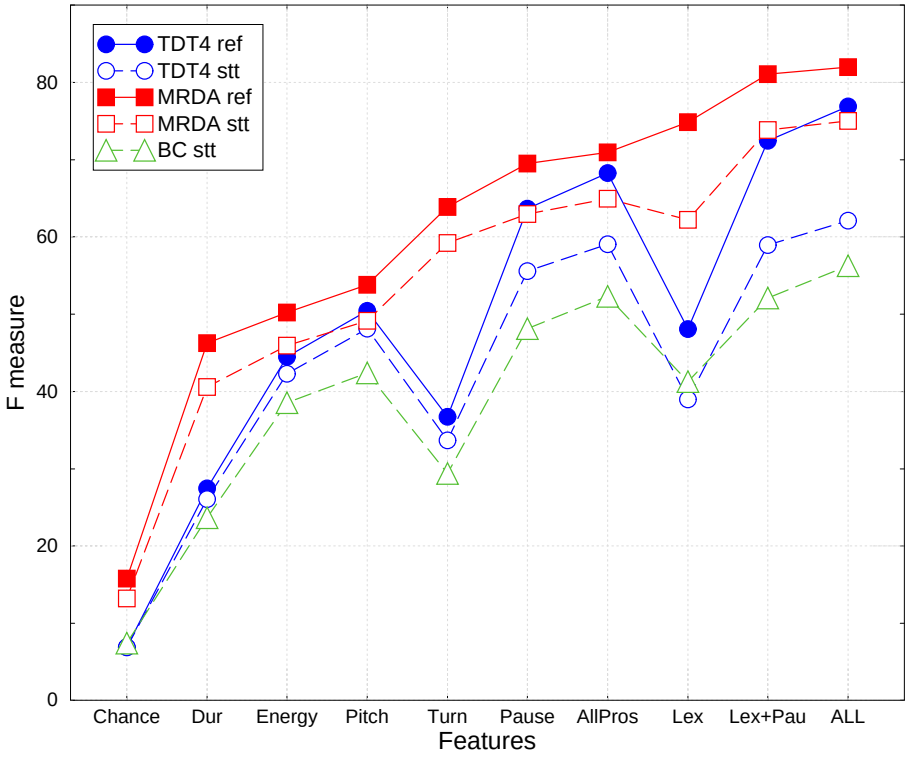

Figure 1: F-measure Results by Condition and Features Included in Model. Dur = duration features, AllPros = all prosodic feature types, Lex = lexical features, Lex+Pau = lexical feature plus pause features, ALL = lexical plus all prosodic features.

Features	TDT4 ref	TDT4 STT	MRDA ref	MRDA STT	BC STT
Chance	6.9	6.9	15.8	13.2	7.4
Duration	27.5	26.0	46.3	40.6	23.6
Energy	44.5	42.3	50.3	46.0	38.5
Pitch	50.5	48.1	53.8	49.2	42.4
Turn	36.7	33.7	63.9	59.2	29.4
Pause	63.7	55.6	69.5	63.0	48.1
All pros	68.3	59.1	71.0	65.0	52.3
Lex only	48.1	39.0	74.9	62.2	41.3
Lex+pause	72.4	59.0	81.1	73.9	52.1
ALL(lex+pros)	76.9	62.1	82.0	75.0	56.3

Table 2: Overall results: F-Measure scores for each group of features showed in the first column, for all combinations of corpus/conditions.

Since chance error rates are near 100%, the relative reduction in error after normalizing for chance performance is nearly the same value as the F-Measure itself. That is, an F-measure of 70 corresponds to a relative error reduction δ of about 70% for the data sets considered.

When comparing performance within a corpus (TDT4 or MRDA) for reference versus automatic transcripts, results show a remarkably consistent performance drop associated with ASR errors. This implies that all feature types are affected to about the same degree by ASR errors. The interesting, clear exception is the performance of lexical features for the meeting data, which degrades more in the face of ASR errors than do other conditions. For example, relative to the all-prosody features condition, lexical features in this corpus give better performance for reference transcripts, but worse performance for automatic transcripts. In contrast, TDT4 shows about the expected drop for lexical features from ASR features. One possible explanation is that in MRDA, there is a high rate of backchannel sentences (such as "uh-huh") which comprise a rather small set of words, sometimes with fairly low energy, that are more prone to recognition errors or that cause class errors when misrecognized. The same argument could be made for other frequent words in MRDA that are strong cues to sentence starts, such as "I" and various fillers and discourse markers. Further analysis, in which selected dialog acts such as backchannels are removed from the train and test data, could shed light on these hypotheses.

If we consider that the BC data set is much smaller than the other two sets, and thus the training material for the

classifier smaller, all three corpora are quite similar in performance in both energy and pitch features (although we will see in the next section that within these feature classes, there are some corpus differences). The corpora also share the trend that duration features are less useful than pitch or energy features, and that pause features are the most useful individual feature type. Interestingly, duration features alone are more useful in MRDA than in either of the broadcast corpora. A listening analysis using class probabilities of errors from the model revealed a possible explanation. In broadcast speech, speakers do lengthen phones before sentence ends, but they also lengthen phones considerably in other locations, including during the production of frequent prominences, and at the more frequent sub-sentential syntactic boundaries found in news speech. Both characteristics appear to lead to considerable false alarms on duration features in the broadcast corpora.

Another noticeable difference across corpora is visible for turn features. Here again, the meeting data differs from the broadcast data. This result reflects both the higher rate of turn changes in the meeting data, especially for short utterances such as backchannels, and the way that the data is processed. As already mentioned in Section 2, the turn is computed differently in the meetings than in the two other corpora. In the broadcast data, the turn is only estimated by an external diarization system that may introduce errors, whereas in the meetings the turn information is the true one since each speaker have their own channel. Furthermore, while turns in both broadcast and meetings data are broken by 0.5 second pauses, the meeting pauses are derived from the reference or STT transcript while the broadcast data pauses come from the less-sophisticated speech/non-speech preprocessor of the diarization system.

A final observation from Figure 1 concerns the patterns for the BC data. This is a newer corpus in the speech community and little is understood about whether it is more like broadcast news speech or more like conversational speech. The results here, both for chance performance and for performance across feature types, clearly indicate that in terms of sentence boundary cues, broadcast conversations are more like broadcast news, and less like conversations. The overall lower results for BC data are as noted earlier, likely explained simply by the smaller set of training data available. The one exception visible from Figure 1 in this trend is in the condition using lexical features only. We would expect the BC result here to be lower than that for TDT4 STT, given the overall lower performances for BC than TDT4. But instead we see a higher-than-expected result for BC in this condition, similar in trend to the pattern seen for MRDA STT for lexical features. We hypothesize that BC shares with conversational data the added utility of lexical features from either backchannels (that start and end sentences) or from words like fillers, discourse markers, and first person pronouns (that tend to start sentences). Further analyzes of the BC data suggest that while backchannels may not play a large role in broadcast conversations, the second class of words, i.e. those that tend to start sentences, are fairly frequent and thus probably aid the lexical model.

3.2 Performance by feature subgroup

A further feature analysis step is to look more closely at the two feature types that capture frame-level prosodic values, namely pitch features and energy features. These are also the two feature types that are normalized for the particular speaker (or speaker estimated via diarization) in our experiments. Our feature sets for each of these two feature types consisted of three subgroups.

Subgroup 1 uses an estimate of "baseline" pitch or energy, intended to capture the minimum value for the particular talker. Features in this subgroup reflect pitch or energy values in the word or short time window preceding the boundary in question, and compares those values to the estimated baseline value for the talker. The idea is to capture how high or low the pre-boundary speech is, relative to that speaker's range, with lower values correlating with sentence ends. We refer to these features as *range* features, since they capture the speaker's local value within their range. Note that because these features look at information prior to the boundary itself, they could be used for online processing, i.e. to predict boundary types before the following word is uttered.

Subgroup 2 looks *across* the boundary, i.e. at words or time windows both before and after the boundary in question. These features compare various pitch and energy statistics (for example maximum, mean, minimum) of the preceding and following speech, using various normalizations. The idea is to capture "resets" typical of sentence ends, in which the pitch or energy has become low before the end of a sentence, and is then reset to a higher value at the onset of the next sentence. Such features are defined only when both the preceding and following speech is present (within a time threshold for any pause at the boundary). We refer to these as *reset* features.

Subgroup 3, like subgroup 1, looks only at words or time windows at one side of the boundary. The idea is to capture the size of pitch or energy excursions by using the *slope* of regions close to the boundary. The slope is taken from linear fits of pitch and energy contours after various preprocessing techniques to remove outliers. Sentence ends in conventional linguistic studies of prosody are associated with a large "final fall", which these features are intended to capture. They may also however capture excursions related to prominent syllables at non-boundaries.

Results for the three feature types, for both energy and pitch, are shown in Figure 2. For ease of readability, lines connect points for the same condition. We look only at the three STT conditions, since reference results for TDT4 and MRDA show a similar pattern to their respective STT results, and using STT results allows us to compare all three corpora. To compare relative usage of the different feature subgroups directly, we look at relative error reduction results (see previous section), although as noted there the absolute F-measure results will look similar.

A first point to note about Figure 2, which can be construed by comparing to results in Figure 1, is that in all conditions, subgroups perform less well on their own than the all-energy and all-pitch groups. This indicates that the subgroups contribute some degree of complementary information. Second, across all conditions and across the two feature types, it is clear that the *reset* features perform better than features based on local *range* or *slope*. The interpretation is that it is better to use information from words or windows on both sides of a boundary in question than to look only at one side or the other. Third, pitch features are relatively more useful than energy features for all three corpora, but the largest differential is for the TDT4 data.

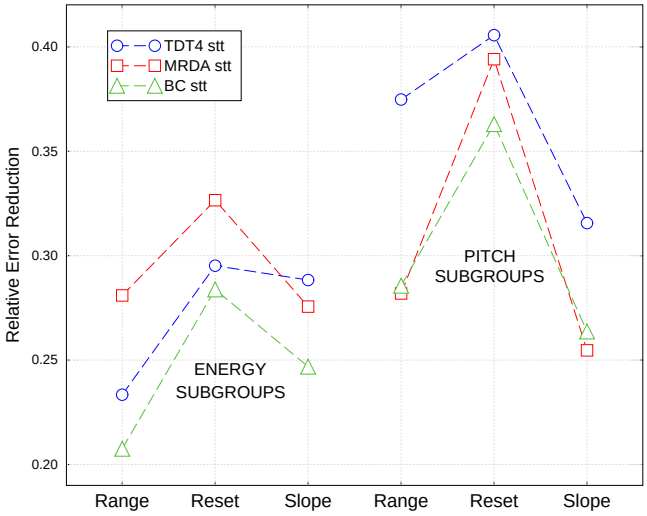

Figure 2: Relative error reduction for the three subgroups of features *range*, *reset* and *slope*, for both energy and pitch.

Note however that the strongest subgroup, i.e. pitch reset, shows nearly identical relative error reduction performance for both TDT4 and MRDA; BC is not far behind given the much smaller amount of data available for training.

Finally, these results show one example in which the BC data compares more closely with meeting data than with broadcast news data. TDT4, more so than the two conversational corpora, can make use of additional subtypes such as slope for energy features, and range for pitch features. Thus, although BC looks more like TDT4 than like MRDA when examining overall feature usage (see Figure 1), it shares with MRDA that certain feature subtypes, such as those based on looking at only one side of a boundary, are much less robust than the reset features that look across boundaries. This suggests that the conversational data may have greater range variation and less defined excursions than read news speech.

4. SUMMARY AND CONCLUSIONS

We have studied the performance of sentence segmentation across two spontaneous speaking styles—broadcast conversations and meetings— and a more formal one—broadcast news. The average length of sentences and comparison of the lexical and prosodic feature types performance showed that in terms of sentence boundary cues, broadcast conversations are more like broadcasts and less like meetings. However, the performance of the lexical features suggests that BC shares with meetings the added utility of lexical features from word like fillers, discourse markers, and first person pronouns. Other similarities between meetings and BC were also observed, such as the benefit of prosodic features that looking at characteristics of both sides (rather than only one side) of an inter-word boundary.

The three speaking styles showed similarities in the role of individual features. Pitch and energy features, as overall groups, perform surprisingly similarly in absolute terms for all three corpora. Also, for all corpora pause features are the most useful individual type of features, and duration features are less useful than energy and pitch. However, while the rank of the feature types was the same, the duration features were comparatively more useful in meetings than in the two other corpora, most likely because of the tendency of broadcast speakers to lengthen phones not only near sentence boundaries, but also in other locations.

A closer look at pitch and energy features in terms of feature subgroups revealed that subgroups provide complementary information, but some subgroups are clearly better than others. For all three corpora, there was greatest benefit from features that compare speech before and after inter-word boundaries. Broadcast news differed from the conversational corpora in being able to also take good advantage of features that look only at one side of the boundary, likely reflecting the more careful and regular prosodic patterns associated with read (as opposed to spontaneous) speech.

Comparisons of the reference and speech-to-text conditions showed, interestingly, that nearly all feature types are affected to about the same degree by ASR errors. The exception was lexical features in the case of the meetings, which degrade more than expected from ASR errors. Possible explanations for this are that sentence segmentation performance in meetings relies more heavily on certain one-word utterance like backchannels, as well as on a small class of highly predictive sentence onset words such as "I", fillers, and discourse markers.

In future work we plan to explore methods for improving performance on BC data, including adaptation and addition of similar data from other corpora. We also plan to study the impact of removing specific classes of dialog acts from the meetings, to determine the behavior of lexical features for this corpus, as just described above, is related to specific dialog acts or to some other phenomenon. Finally, we hope that further work along the lines of the studies described herein, can add to our longer term understanding of the relationship between speaking style and various techniques and features for natural language processing.

5. ACKNOWLEDGMENTS

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under contract No. HR0011-06-C-0023. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of DARPA. The authors would like to thank Matthias Zimmermann, Yang Liu, and Mathew Magimai Doss for their help and suggestions.

6. REFERENCES

[1] S. Cuendet, D. Hakkani-Tür, and G. Tur. Model adaptation for sentence unit segmentation from speech. In *Proceedings of SLT*, Aruba, 2006.

[2] S. Furui, T. Kikuchi, Y. Shinnaka, and C. Hori. Speech-to-text and speech-to-speech summarization of spontaneous speech. *Speech and Audio Processing, IEEE Transactions on*, 12(4):401–408, 2004.

[3] D. Hakkani-Tür and G. Tur. Statistical sentence extraction for information distillation. In *Proceedings of ICASSP*, Honolulu, HI, 2007.

[4] D. Hillard, Z. Huang, H. Ji, R. Grishman, D. Hakkani-Tur, M. Harper, M. Ostendorf, and W. Wang. Impact of automatic comma prediction on

pos/name tagging of speech. In *Spoken Language Technologies (SLT)*, 2006.

[5] D. Jones, W. Shen, E. Shriberg, A. Stolcke, T. Kamm, and D. Reynolds. Two experiments comparing reading with listening for human processing of conversational telephone speech. In *Proceedings of EUROSPEECH*, pages 1145–1148, 2005.

[6] Y. Liu, E. Shriberg, A. Stolcke, B. Peskin, J. Ang, D. Hillard, M. Ostendorf, M. Tomalin, P. Woodland, and M. Harper. Structural metadata research in the EARS program. In *Proceedings of ICASSP*, 2005.

[7] J. Makhoul, A. Baron, I. Bulyko, L. Nguyen, L. Ramshaw, D. Stallard, R. Schwartz, and B. Xiang. The effects of speech recognition and punctuation on information extraction performance. In *In Proc. of Interspeech*, Lisbon, 2005.

[8] E. Matusov, D. Hillard, M. Magimai-Doss, D. Hakkani-Tür, M. Ostendorf, and H. Ney. Improving speech translation with automatic boundary prediction. In *Proceedings of ICSLP*, Antwerp, Belgium, 2007.

[9] J. Mrozinski, E. W. D. Whittaker, P. Chatain, and S. Furui. Automatic sentence segmentation of speech for automatic summarization. In *Proc. ICASSP*, Philadelphia, PA, 2005.

[10] B. Roark, Y. Liu, M. Harper, R. Stewart, M. Lease, M. Snover, I. Shafran, B. Dorr, J. Hale, A. Krasnyanskaya, and L. Yung. Reranking for sentence boundary detection in conversational speech. In *Proceedings of ICASSP*, Toulouse, France, 2006.

[11] R. E. Schapire and Y. Singer. Boostexter: A boosting-based system for text categorization. *Machine Learning*, 39(2/3):135–168, 2000.

[12] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of SigDial Workshop*, Boston, MA, 2004.

[13] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tur. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 2000.

[14] A. Stolcke, B. Chen, H. Franco, V. R. R. Gadde, M. Graciarena, M.-Y. Hwang, K. Kirchhoff, A. Mandal, N. Morgan, X. Lin, T. Ng, M. Ostendorf, K. Sonmez, A. Venkataraman, D. Vergyri, W. Wang, J. Zheng, and Q. Zhu. Recent innovations in speech-to-text transcription at SRI-ICSI-UW. In *IEEE Trans. Audio, Speech and Language Processing*, volume 14, pages 1729 – 1744, 2006.

[15] S. Strassel and M. Glenn. Creating the annotated TDT-4 Y2003 evaluation corpus. In *TDT 2003 Evaluation Workshop, NIST*, 2003.

[16] A. Venkataraman, A. Stolcke, W. Wang, D. Vergyri, V. Gadde, and J. Zheng. SRIs 2004 broadcast news speech to text system. In *EARS Rich Transcription 2004 workshop*, Palisades, 2004.

[17] Q. Zhu, A. Stolcke, B. Chen, and N. Morgan. Using MLP features in SRIs conversational speech recognition system. In *Proceedings of INTERSPEECH*, pages 2141 – 2144, Lisbon, Portugal, 2005.