# GENRE EFFECTS ON AUTOMATIC SENTENCE SEGMENTATION OF SPEECH: A COMPARISON OF BROADCAST NEWS AND BROADCAST CONVERSATIONS

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

[1]Dept. of Cybernetics, Faculty of Applied Sciences, University of West Bohemia, Pilsen, Czech Rep.
[2]Department of Computer Science, University of Texas at Dallas, Richardson, TX, USA
[3]SRI International, Menlo Park, CA, USA [4]ICSI, Berkeley, CA, USA

## ABSTRACT

We investigate genre effects on the task of automatic sentence segmentation, focusing on two important domains – broadcast news (BN) and broadcast conversation (BC). We employ an HMM model based on textual and prosodic information and analyze differences in segmentation accuracy and feature usage between the two genres using both manual and automatic speech transcripts. Experiments are evaluated using Czech broadcast corpora annotated for sentence-like units (SUs). Prosodic features capture information about pause, duration, pitch, and energy patterns. Textual knowledge sources include words, part-of-speech, and automatically induced classes. We also analyze effects of using additional textual data that is not annotated for SUs. Feature analysis reveals significant differences in both textual and prosodic feature usage patterns between the two genres. The analysis is important for building automatic understanding systems when limited matched-genre data are available, or for designing eventual genre-independent systems.

***Index Terms***— Spoken language understanding, sentence segmentation, broadcast news, broadcast conversations, prosody

## 1. INTRODUCTION

Automatic sentence segmentation is important for enriching speech recognition output and for aiding downstream language processing. Several past approaches to this task have used lexical features, prosodic features, or a combination of such features, for example, [1, 2, 3, 4, 5]. Studies on sentence segmentation have been conducted in different domains, including broadcast news, conversational telephone speech, lectures, and meetings. Broadcast news is a widely-studied domain for sentence segmentation, partly because of the benchmark test of speech recognition in this area. In contrast, there is much less work on broadcast conversations, which have only recently received attention [6]. An understanding of how genres differ on the task of sentence segmentation is important for building automatic understanding systems when limited matched-genre data are available, or for designing eventual genre-independent systems.

In this paper, we focus on a comparison between the better-studied broadcast news (BN) genre and the lesser-studied broadcast conversation (BC) genre, in terms of various knowledge sources. An additional novel aspect is that we examine data from the Czech language. Czech is different from English in many aspects that make it interesting for this task. Czech belongs to the family of Slavic languages, which are highly inflectional and derivational, and thus have an extremely large number of distinct word forms. In addition, colloquial Czech has a different morphology than standard Czech – prefixes and endings are often changed in the former. Another difference is a relatively free word order in Czech. There are also differences in prosody. For example, while sentence-final pitch falls/rises are present in both languages, intrasentential pitch movements (e.g., at prosodic phrase boundaries) are typically less steep in Czech than in English. Furthermore, preboundary lengthening is less emphatic in Czech because vowel length has a lexical function, thereby limiting the scope of prosodically-motivated lengthening.

## 2. DATA AND EXPERIMENTAL SETUP

We use two Czech corpora – one in the BN domain (containing TV and radio news – mostly prepared speech) and the other in the BC domain (containing radio debates – mostly spontaneous speech). The two corpora were annotated based on LDC's Metadata Extraction (MDE) standard [7], as described in [8]. The annotation included labeling of sentence-like unit (SU) boundaries which were used in this work. Note that the SUs are defined based on a set of strict segmentation rules designed to achieve good annotation consistency even on conversational speech. The data in each corpus were split into a training set, a development set, and a test set. For BN, the data sets comprised 174.8k words for training, 28.2k for development, and 31.2k words for testing. For BC, the data included 159.1k words for training, 24.1k words for development, and 24.6k words for testing. All experiments were evaluated using both human-generated reference transcripts (REF) and automatic speech recognition (ASR) transcripts. The ASR output was obtained from the UWB LVCSR system tailored for real-time recognition of highly inflective languages [9]. The overall word error rates were 12.4% for BN and 29.3% for BC.

## 3. METHOD

For a given word sequence, our task is to determine the location of sentence boundaries using textual and acoustic information. The model we use in this study is a hidden Markov model (HMM) [1]. This approach has been widely used for sentence segmentation and generally achieves comparable performance compared to other approaches. The HMM model describes the joint distribution of words $W$, prosodic features $P$, and SU boundaries $S$, $P(W,P,S)$. The model assumes that prosodic features depend only on the events (sentence boundary or not), and not on the words. The observation likelihood comes from the prosodic classifier. The transition probability is based on an $n$-gram language model, which is trained by explicitly including the SU boundary as a token in the vocabulary. During testing, the model performs forward-backward decoding to

This work was supported by the Ministry of Education of the Czech Republic under projects No. 2C06020 and ME909.

find the SU boundaries (hidden states) given the word sequence and corresponding prosodic features (observations). The following two sections describe the two components in the HMM – observation probabilities and transition probabilities.

### 3.1. Prosodic features and models

Our prosodic features are designed to reflect breaks in temporal, intonational, or energy contours. The features are extracted from an alignment of the speech signal with word-level and phone-level time alignment information from an automatic speech recognizer. Note that this approach computes features directly from the signal, without the need for any human labeling of prosodic events [1].

The prosodic features can be grouped into four broad feature classes: *pause*, *pitch*, *duration*, and *energy*. The features are associated with particular interword boundaries. In order to capture local prosodic dynamics, we also use features associated with boundaries after the previous and after the following word. In addition to the purely prosodic features, the automatic prosodic classifiers also have access to a limited number of "other" features, capturing phenomena such as speaker change.

For prosody modeling, we used decision tree classifiers. Since SU boundaries are much less frequent than non-SU boundaries, we had to cope with the problem of data skew. To overcome this problem and to decrease classifier variance, we use a combination of ensemble sampling with bagging [10].

We performed feature selection to identify a small set of prosodic features in two steps [11]. First, for each of the broad prosodic feature categories, we selected those features each of which has a feature usage statistics higher than a predefined threshold. Then using these features, we performed leave-one-out feature selection and removed a feature if its deletion did not yield any performance loss. This feature reduction algorithm ended up selecting 11 features for BN and 17 features for BC. Furthermore, we investigated whether there is a gain from using the richer set of prosodic features in comparison with using pause information alone. The alternative pause-only feature set contains only those features that capture pause duration after the previous, current, and following word, plus the speaker change feature.

### 3.2. Textual features and $n$-gram LM

We use various information to improve the word-based LM trained from the MDE training corpus, including using class-based LMs and additional corpora. All of the LMs are trigram LMs with modified Kneser-Ney smoothing.

#### *3.2.1. Word-based LMs*

For word-based LMs, we use two corpora. The first (MDE-Word) is the training set of the corresponding MDE corpus (either BN or BC). The second (Aux-Word) corresponds to an auxiliary corpus of Czech broadcast transcripts, which contains about 107M words. Note that the latter data were not annotated for SU boundaries in terms of the MDE guidelines, but only contained standard punctuation.

#### *3.2.2. Automatically induced classes (AICs)*

Data sparseness is a common problem for word-based LMs. One solution to this problem is to group words with similar properties into classes. We used a well-known clustering algorithm that minimizes the perplexity of the induced class-based $n$-gram with respect to the provided word bigram counts [12]. The SU boundary token was excluded from merging, however, its statistics still affected the clustering. The optimal number of word clusters was empirically estimated on development data as 300 for BN and 275 for BC. We also experimented with removing frequent words from the clustering, but this did not yield improvement.

#### *3.2.3. POS tags*

The AICs reflect word usage in our datasets, but do not form clusters with a clearly interpretable linguistic meaning. In contrast, part-of-speech (POS) tags describe grammatical features of words. Unlike English, highly inflected languages (such as Czech) often use structured morphological tagsets. In addition to labeling words with a POS category, these structured tagsets use tags comprising of "subtags" providing information about morphological categories. For Czech, the most popular tagset is the positional tag system from the Prague Dependency Treebank (PDT) [13]. In this tagset, every tag is represented as a string of 15 subtags which approximately fit the formal Czech morphology categories. While the English *Penn Treebank Tagset* contains just 36 POS tags plus 12 tags for punctuation, there are more than 1,500 different Czech tags in the PDT tagset. For the experiments herein, we used automatic tags obtained from the Morče tagger [14].

In addition to the purely POS-based models, we also tested models that combine tags with frequent words (POSmix). The idea behind this approach is to preserve information about certain frequent words that correlate strongly with sentence boundaries or nonboundaries. This approach can be viewed as a form of backoff: we back off from words to tags for rare words, but keep word identities for frequent words. Optimizing the model on the development data, we ended up with 1,600 most frequent word forms being kept for the BN corpus, and 2,000 word forms being kept for the BC corpus.

## 4. RESULTS AND DISCUSSION

We measure SU segmentation performance using $F$-measure, which is the harmonic mean of Precision ($P$) and Recall ($R$):

$$F = \frac{2PR}{P+R} \tag{1}$$

In the discussions below, we also use a comparative metric $\delta_F$ reflecting relative error reduction in terms of $1 - F$:

$$\delta_F = \frac{(1-F_1)-(1-F_2)}{1-F_1} = \frac{F_2-F_1}{1-F_1} \tag{2}$$

where $F_2 > F_1$ are $F$-measures of two compared systems.

### 4.1. Results based on textual information only

Fig. 1 displays $F$-measures achieved by different textual feature groups (i.e., MDE-Word, AIC, POS, POSmix, Aux-Word) and their combination in the four test sets: BN REF, BN ASR, BC REF, and BC ASR. We use lines to connect points corresponding to the same test set in order to increase readability.

As shown, BN and BC differ considerably in performance when using word-based LMs. The MDE-Word model performs better on BC than on BN, even though BC is more spontaneous and seems more difficult. One explanation for this difference is that conversational speech contains a number of cue words (such as discourse markers) that signal sentence boundaries. Since Czech does not have a fixed word order, such cue words are more important for LMs than in English.

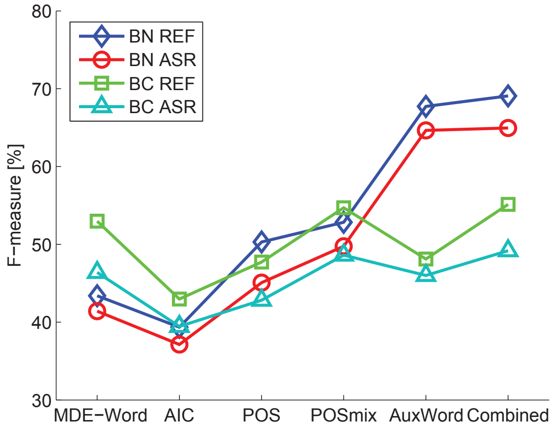

**Fig. 1**. SU segmentation F-measures for BN and BC for different textual feature groups in REFerence and ASR conditions

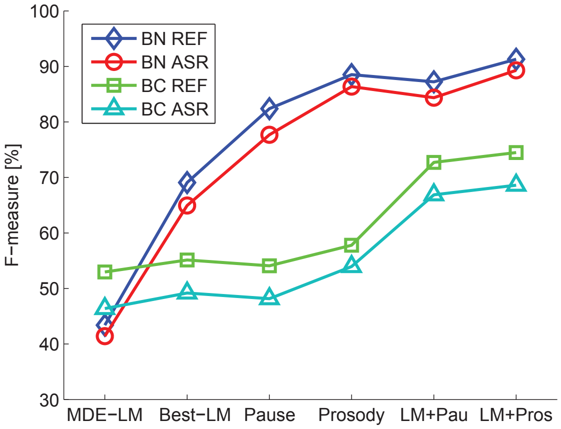

**Fig. 2**. SU segmentation F-measures for BN and BC using different knowledge sources in REFerence and ASR conditions

Aux-Word performs well on BN but not as well on BC. In fact, for BN the best result using a single model is achieved by using Aux-Word. This suggests that the auxiliary training text more closely matches the BN than the BC data. Although Aux-Word also contains transcripts of discussions, these transcripts are not strictly verbatim; transcribers of this database often left out filler words and disfluencies, and "standardized" colloquial word forms. Aux-Word is helpful for SU segmentation on BC only when combined with MDE-Word. The relative error reduction from adding Aux-Word is 1.0% for BC REF and 1.1% for BC ASR. Both improvements are statistically significant at $p < 0.01$ using the Sign test. For comparison, $\delta_F$ is 34.4% for BN REF and 25.7% for BN ASR. The above-presented results suggest that for building the LM for BC, in-domain speech transcripts are needed to achieve good results. On the other hand, the LM for BN achieves good performance even if only trained on the auxiliary textual data.

Of the two POS-based feature sets, POSmix performed better than POS in all four test sets, indicating that keeping frequent words in the POS model is helpful. The POSmix model mitigates the data sparseness problem by grouping infrequent words into POS-based classes, but it also preserves important details about frequent words. POSmix achieves the best performance among the single models for BC, but on BN it is not as effective as Aux-Word.

The least useful feature set for all the conditions was AIC. It yields the poorest performance when used on its own, and does not provide any gain when combined with other models. This is in contrast with our previous results on English [15] where adding AIC information significantly improved results. Overall, the best results for both BN and BC were achieved by a combination of Aux-Word, MDE-Word, and POSmix. The best $F$-measures are significantly higher for BN than BC.

### 4.2. Results using prosody only and its combination with LMs

Results using only prosodic information and its combination with LMs are displayed in Fig. 2. For combination, we use the best LMs from Section 4.1.

The prosody-only models perform much better for BN than for BC. For both BN and BC, the best prosody models outperformed the best LMs, however, the prosody model dominance was much more visible in the BN corpus. The richer prosodic feature sets outperformed pause-only feature sets in all test conditions. The rich prosody prevalence was higher in the BN corpus. $\delta_F$ for this difference was 34.6% for BN REF, 38.9% for BN ASR, 8.1% for BC REF, and 11.2% for BC ASR. Using the Sign test, all differences were significant at $p < 10^{-6}$ or better. For the combined results using prosody and LMs, the best results were achieved by the LM combined with the rich prosodic model. The best results were $F = 91.3\%$ for BN REF, $F = 89.3\%$ for BN ASR, $F = 74.5\%$ for BC REF, and $F = 68.6\%$ for BC ASR.

The result comparison for BN indicates that very good results on this data may be achieved when only the prosody model is used. The prosody-only model is better than the LM+Pause model and only slightly worse than the LM+Prosody model. On the other hand, the BC result comparison shows that both LM+Pause and LM+Prosody perform much better than the prosody-only model. The same comparison also indicates that, unlike BN, LM+Prosody is only slightly better than LM+Pause in BC – corresponding $\delta_F$ values are 6.5% for BC REF, and 5.2% for BC ASR.

### 4.3. Prosodic feature usage

To better understand the prosodic model, we look at the results broken down by feature usage. The usage metric reflects the number of times a feature is queried in a decision tree, weighted by the number of samples it affects at each node. The total feature usage within a tree sums to 1. The feature usage distributions are displayed in Fig. 3. These results are based on averaging results over multiple trees generated in bagging.

The graph shows that duration and pause feature groups were most important for both BN and BC; however, the feature group usage distributions differ between them. The difference is most prominently displayed in pause features, which were more frequently queried in BN, indicating that pause information is a relatively better cue in prepared speech. By contrast, duration features were more heavily used in BC. Another difference between the two distributions is in the proportion of pitch and energy features. BN prefers pitch features, while energy features are used more in BC.

Regarding individual prosodic features, there are also some differences between BN and BC. From the duration group, normalized

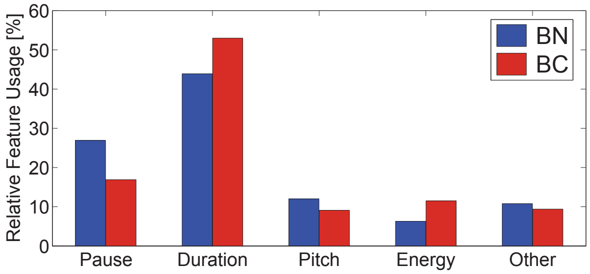

**Fig. 3**. Prosodic feature group usage in Czech BN and BC

duration of the last vowel was the most important feature for BNs (9.4% of overall usage), while the raw word duration feature was dominant for BC (16.5%). The pitch group in BN heavily uses a feature reflecting the ratio between the last $F_0$ value and the speaker's $F_0$ baseline (12.0%), suggesting that radio anchors tend to mark statement boundaries with significant pitch falls. This pitch feature was also important for BC, but to a lesser extent (4.1%). In both corpora, the most used energy feature was normalized maximal RMS value from the word following the boundary in question (6.3% in BN, 5.8% in BC). This feature captures the phenomenon that speakers typically start sentences at a higher level of vocal effort and fall off in loudness toward the end.

We also compared the feature usage statistics with those for English corpora. Because to our best knowledge, there are no published usage statistics for English BC, we only could make the comparison for BN. We used the statistics published in [16]. The most used feature (pause duration at the boundary in question) was the same for both languages. A comparison of other most used features demonstrated that features capturing final lengthening were more important for English, while features capturing final pitch fall were more important for Czech. This finding is in agreement with the fact that in comparison to English, Czech offers less opportunity for final lengthening because length also serves a lexical function in Czech.

## 5. CONCLUSION

We explored the task of automatic sentence segmentation in Czech broadcast news and broadcast conversations, using an HMM model based on textual and prosodic information. The experiments with language models showed that a large database of broadcast transcripts is important for training models for BN, while in-domain speech transcripts are essential for achieving good results on BC. We also found that POS information is used more efficiently when we keep word identities for frequent words and back off to POS for infrequent words. In general, language models combining several textual knowledge sources worked better than models using just a single information source.

Regarding prosodic information, we found that prosodic features benefit SU segmentation more for BN than for BC. Prosodic features beyond pause were also relatively more helpful for BN than for BC. Overall, the best performance was achieved by combining the prosodic and LM information for all the test conditions. As expected, much better overall performance was achieved for BN than for BC.

Feature analysis revealed that BN and BC differ in prosodic feature usage patterns. Furthermore, a cross-lingual comparison showed that pause information is most important for sentence segmentation of both English and Czech BN, but the second most used features differ. Features capturing preboundary lengthening are more important for English, while Czech prefers the features capturing final pitch fall. Overall, this finding should help guide strategies for automatic sentence segmentation across genres in future work.

## 6. REFERENCES

[1] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, vol. 32, no. 1-2, pp. 127–154, 2000.

[2] J. Huang and G. Zweig, "Maximum entropy model for punctuation annotation from speech," in *Proc. of ICSLP 2002*, Denver, USA, 2002, pp. 917–920.

[3] D. Wang and S. Narayanan, "A multi-pass linear fold algorithm for sentence boundary detection using prosodic cues," in *Proc. of ICASSP 2004*, Montreal, Canada, 2004.

[4] M. Tomalin and P. C. Woodland, "Discriminatevely trained Gaussian mixture models for sentence boundary detection," in *Proc. ICASSP*, Toulouse, France, 2006, pp. 549–552.

[5] J. Kolář, Y. Liu, and E. Shriberg, "Speaker adaptation of language models for automatic dialog act segmentation of meetings," in *Proc. Interspeech 2007*, Antwerp, Belgium, 2007.

[6] S. Cuendet, E. Shriberg, B. Favre, J. Fung, and D. Hakkani-Tür, "An analysis of sentence segmentation features for broadcast news, broadcast conversations, and meetings," in *Proc. SIGIR 2007 – SSCS*, Amsterdam, The Netherlands, 2007.

[7] S. Strassel, "Simple metadata annotation specification V6.2," http://www.ldc.upenn.edu/Projects/MDE/Guidelines/SimpleMDE_V6.2.pdf, 2004.

[8] J. Kolář and J. Švec, "Structural metadata annotation of speech corpora: Comparing broadcast news and broadcast conversations," in *Proc. LREC'08*, Marrakech, Morocco, 2008.

[9] A. Pražák, L. Müller, J. V. Psutka, and J. Psutka, "Live TV subtitling: Fast 2-pass LVCSR system for online subtitling," in *Proc. SIGMAP 2007*, Barcelona, Spain, 2007.

[10] Y. Liu, N. Chawla, M. Harper, E. Shriberg, and A. Stolcke, "A study in machine learning from imbalanced data for sentence boundary detection in speech," *Computer Speech and Language*, vol. 20, pp. 468–494, 2006.

[11] J. Kolář, "Automatic segmentation of speech into sentence-like units," Ph.D. dissertation, University of West Bohemia, Pilsen, Czech Republic, 2008.

[12] P. Brown, V. D. Pietra, P. de Souza, J. Lai, and R. Mercer, "Class-based n-gram models of natural language," *Computational Linguistics*, vol. 18, no. 4, pp. 467–479, 1992.

[13] J. Hajič et al., "Prague Dependency Treebank 2.0," Linguistic Data Consortium, CD-ROM LDC2006T01, 2006.

[14] D. Spoustová, J. Hajič, J. Votrubec, P. Krbec, and P. Květoň, "The best of two worlds: Cooperation of statistical and rule-based taggers for Czech," in *Proc. of the ACL Workshop on Balto-Slavonic NLP*, Prague, Czech Republic, 2007.

[15] J. Kolář, "A comparison of language models for dialog act segmentation of meeting transcripts," in *Proc. TSD'2008*, Brno, Czech Republic, 2008.

[16] Y. Liu, "Structural event detection for rich transcription of speech," Ph.D. dissertation, Purdue University, 2004.