# ROBUST WORD BOUNDARY DETECTION IN SPONTANEOUS SPEECH USING ACOUSTIC AND LEXICAL CUES

Andreas Tsiartas, Prasanta Kumar Ghosh, Panayiotis Georgiou and Shrikanth Narayanan

Speech Analysis and Interpretation Laboratory
Department of Electrical Engineering,
University of Southern California,
Los Angeles, CA 90089

tsiartas@usc.edu, prasantg@usc.edu, georgiou@sipi.usc.edu, shri@sipi.usc.edu

## ABSTRACT

We consider the problem of word boundary detection in spontaneous speech utterances. Acoustic features have been well explored in the literature in the context of word boundary detection; however, in spontaneous speech of Switchboard-I corpus, we found that the accuracy of word boundary detection using acoustic features is poor (F-score $\sim$ 0.63). We propose a new feature - that captures lexical cues in the context of the word boundary detection problem. We show that including proposed lexical feature along with the usual acoustic features, the accuracy of the word boundary detection improves considerably (F-score $\sim$ 0.81). We also demonstrate the robustness of our proposed feature in presence of different noise levels for additive white and pink noise.

***Index Terms***— word boundary, sentence segmentation, OOV detection

## 1. INTRODUCTION

Automatic word boundary detection, a topic that has been investigated for several decades, is still an active area of research due to its impact in diverse applications and, the challenging nature of the problem. Initial applications have included detection of the exact word boundaries to assess speech recognition performance and to make recognizers faster. Other, applications of word boundary detection include detecting regions of out of vocabulary (OOV) words and detecting exact boundaries for unknown named entities in speech. Word boundary information can also be helpful for rich transcription of speech such as in detecting emphatic (prominent) words [1].

In the past, researchers have tried to address this problem using just the acoustic information from speech. It has been shown that to some extent word boundaries can be successfully estimated from acoustic information. For example, Junqua et al [2] used energy, based on a frequency sub-band analysis to detect word boundaries and showed that the frequency band 250-3500 Hz provides useful clues for word boundary detection, even in the presence of noise. Rajendran et al [3] showed that pitch patterns can provide useful information for word boundary detection. Lin et al [4] also used acoustic information and showed that a multi-band energy approach along with background noise estimation can improve word boundary detection in the presence of some noise types. While acoustic cues carry useful word boundary information, they suffer from certain limitations.

Importantly, acoustic cues often fail to give clues about the word boundaries, particularly when the beginning of a word gets co-articulated with the end of the previous word in many lexical contexts. This is especially common in spontaneous speech, where the word boundaries are not acoustically distinct (even though they are lexically defined) unlike isolated words, which makes any acoustic cues based on boundary detection algorithms incapable of handling such cases. It is, in these cases, where lexical cues could become an important factor for estimating word boundaries. For instance, Harrington et al [5] used information from phoneme strings derived from speech transcriptions to estimate word boundaries. In a different context, Cettolo et al [6] used lexical and acoustic information to recognize semantic word boundaries. Automatic Speech Recognition(ASR), in particular, uses both acoustic and lexical features to obtain the best hypothesis of word sequences of a given utterance. However, even the best recognition hypothesis of an ASR often turns out to be noisy or erroneous in terms of estimation of actual word boundaries; this situation is especially notable in the the presence of OOV items in speech.

In our work, we focus on capturing lexical information from the ASR framework (not necessarily best hypothesis) and incorporating them with improved acoustic features to obtain a robust word boundary estimate. The rationale here is that although the lexical hypotheses may be erroneous, they provide rough, albeit potentially imperfect, word segmentation information which can be advantageously used in con-

[†] Work supported by DARPA.

978-1-4244-2354-5/09/$25.00 ©2009 IEEE

junction with additional acoustic information for improved word boundary detection. Furthermore, in contrast to many of the earlier efforts on word boundary detection that have focused on isolated words [2, 4], we concentrate on detecting word boundaries in spontaneous speech (Switchboard Corpus).

This paper is structured as follows. In the next section, we present a brief background of the proposed features and the underlying formulation used in this work. In section 3, we describe the experimental setup and the evaluation methodology used in our approach. In section 4, we present the results of this work compared to prior efforts by work [4]. A discussion of the results of the proposed and a baseline alternative method follows in section 5. Finally, we summarize this work and propose some future directions.

## 2. DESCRIPTION OF THE FEATURES

### 2.1. Acoustic Features

Below we describe the acoustic features that we used for word boundary detection in this paper.

#### *2.1.1. Short-time energy:*

Since speakers frequently reduce their loudness level while making transitions from one word to the next, it is expected that the signal variation might give a cue to determine whether a frame falls within a word or close to the word boundary. We compute the fullband averaged log rms energy value as a feature at every frame(m), which is defined as

$$E[m] = \frac{1}{3} \sum_{i=m-1}^{m+1} \tilde{E}[i] \qquad (1)$$

$$\text{where } \tilde{E}[i] = ln \left\{ \sqrt{\frac{1}{N} \sum_{iN_{\text{sh}}-N/2}^{iN_{\text{sh}}+N/2} x^2[n]} \right\}$$

where $x[n]$ is the speech signal and $N_{\text{sh}}$ is frame shift in number of samples and $N$ is the length of a analysis frames in number of samples.

#### *2.1.2. Short-time zero crossing rate (ZCR):*

When there is reduced speech activity in a word to word transition, we expect the ZCR at that word boundary to be different from that within the word. ZCR is computed by finding the number of times the signal crosses level zero within an analysis frame.

#### *2.1.3. Short-time Pitch Frequency:*

The speech signal is pseudo-periodic only during the voiced portions and can be tracked by computing the pitch value in every analysis frame. If there is an unvoiced sound or no coarticulation near a word boundary, the pitch gives a meaningful cue for detecting such regions.

### 2.2. Lexical Features

Let us first consider a speech segment of a sample utterance from the Switchboard corpus. In Fig. 1 we see the speech signal of the utterance “Yeah and it was usually” spoken by a female speaker. The actual word boundaries are at .53, .83, .92 ,.98 and 1.13 sec. As it is clear from the plot, the boundary between ’was’ and ’usually’ is acoustically distinct from the amplitude variation while the other word boundaries are not. In the case between words ‘was’ and ‘usually’, the short-time energy shows a clip and also the pitch shows a discontinuity; the zero-crossing rate however, does not show any significant change. For the other word boundaries in this example, these acoustic features do not show any significant cues. The lack of consistent acoustic cues of word boundary in this example is typical in spontaneous speech. This makes the performance of solely acoustic-feature based word boundary detection poor. Below we explain how lexical features obtained from the ASR lattice can be used for improving word boundary detection.

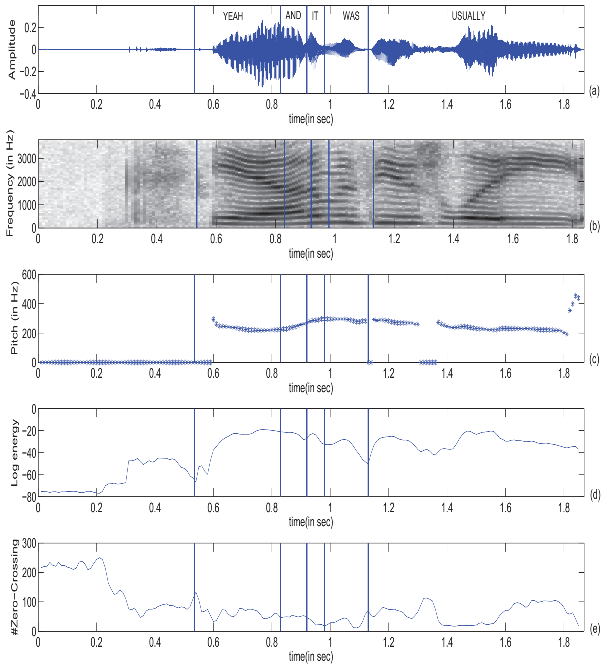

**Fig. 1**. An example speech utterance to illustrate the limitations of the acoustic features in detecting word boundaries: (a) time signal (b) spectrogram (c) short-time pitch (d) short-time energy (e) short-time zero crossing rate. (Analysis window length is 20ms and shift is 10ms)

Let us consider the word lattice at the $l$th frame as shown in Fig. 2, where the nodes indicate the possible word boundaries and arcs show different possible words $w_i$ that end at word boundaries and $w_j$ that start from the respective word boundary. Using a language model, one can determine the most probable paths and list them as hypotheses in a descending order of probability, which is computed over the path from start to end node in the entire lattice.

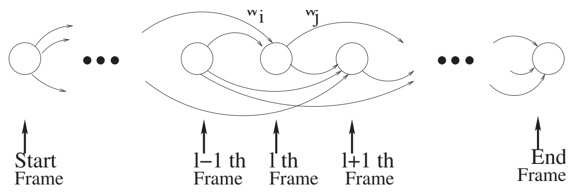

**Fig. 2**. A typical word lattice in ASR.

Let $N_l$ be the total number of paths in the lattice passing through frame $l$. Let us say $p_k$ is the probability of the $k$th path in the lattice from start frame to end frame. We define lexical information based *boundary confidence coefficient* (BCC) at frame $l$ as follows :

$$BCC(l) = \frac{1}{N_{\text{th}}} \sum_{k=1}^{N_l} g(p_k) \tag{2}$$

where

$$g(p_k) = \begin{cases} 1 & \text{if } p_k > p_{\text{th}} \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

where $p_{\text{th}}$ is the optimally chosen threshold of probability. $N_{\text{th}}$ is the number of paths in the lattice with probability $>\ p_{\text{th}}$. This means $BCC(l)$ takes values between 0 and 1. $g(.)$ can be chosen as any function other than this threshold based step function, e.g. a sigmoidal function. In this paper, instead of choosing a fixed $p_{\text{th}}$ for all utterances we vary $p_{\text{th}}$ so that $g(.)$ is 1 for top $N_{\text{th}}$ hypotheses. And we choose $N_{\text{th}}$ in our experiment.

## 3. EXPERIMENTAL SETUP

### 3.1. Corpus

To show the usefulness of our proposed feature, we perform word boundary detection on spontaneous speech utterances obtained from the SwitchBoard corpus (phase-1). We randomly selected 7000 utterances from SwitchBoard, of which 5000 utterances were used for training purposes and 2000 for testing purposes. The recognition system that we use in this setup is Sphinx-3. In order to train Sphinx-3 system, WSJ and TIMIT acoustic data were used. The features that Sphinx-3 was trained on, are 12 MFCCs and energy along with first and second derivatives. We use word boundaries tagged by Misissipi State University [1] as the reference to train and evaluate the system.

### 3.2. Features extraction

Initially, a phoneme recognition is performed on all utterances using Sphinx. We rescore the phoneme lattice using general purpose language model (LM) and we extracted the BCC feature. The LM that we use has a perplexity score 148 against the test set. This LM was created using KN-discounting. In total, the LM is composed of 16K unigrams, 1.6M bigrams and 1.7M trigrams; it is built using DARPA Transtac English side of persian-english transcripts and internet data [7]. We use averaged log Energy, the number of zero crossings and pitch per frame as additional acoustic features (denoted by notation PEZ in Fig. 3 and 4).

### 3.3. Classifier setup

We use the features described above to train a Hidden Markov Model (HMM) based classifier for word boundary detection. We use the HTK toolkit to train the HMMs. We train the HMMs on clean speech only. In this experimental setup, we have two symbols that the HMM classifier can recognize. It can recognize a boundary and a non-boundary symbol. In this experiment, we assume that our features are Gaussian distributed in a multi-dimensional space. We use features extracted from 5000 utterances to estimate the HMM parameters for each symbol. Finally, features from 2000 utterances were decoded using HTK and the most likely sequence of boundary and nonboundary regions was extracted.

### 3.4. Evaluation

In order to evaluate the proposed features, we conducted experiments with various combinations of the above features. Also, we compare the performance of features based on acoustic information only and features based on both the lexical and acoustic information. Additionally to evaluate the robustness of our approach, we performed experiments with speech in additive white and pink noise with different noise levels, from 5dB up to 20dB. The noise samples are obtained from NOISEX-92 database[2]. To evaluate the performance, we first compute the precision and recall [8] of the word boundary detection and finally we report the F-score [8]. An estimated word boundary location is considered to be correct if it is within 10 frames of the actual boundary location. We chose the following parameters for our experiment $N = 320$, $N_{\text{sh}} = 160$ and $N_{\text{th}} = 50$.

## 4. RESULTS

Fig. 3 and 4 show the F-score of word boundary detection for various noisy conditions for different feature combinations with additive white and pink noise respectively. It can be seen

[1] http://www.ece.msstate.edu/research/isip/projects/switchboard/

[2] http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html

that the BCC feature combined with acoustic features consistently gives higher performance compared to acoustic features or BCC feature only. As expected, the performance degrades with lower SNR in all cases when we add noise to speech. In addition, we observe that the accuracy of word boundary estimates using lexical features only is consistently higher than using acoustic features only. We also observe that the rate in drop in performance of word boundary detection using acoustic features only is less than that of using the BCC feature. For colored noise the performance of the word boundary detection is found to be worse compared to that of white noise for a fixed SNR value.

This observations make it clear that in colored noise, the acoustic and/or lexical cues are not as significant as that for additive white noise case for word boundary detection task. It should be noted that our recognizer was trained on clean speech; thus testing on various noisy conditions reveals that the BCC feature obtained in the white noise case is more representative of the actual word boundaries compared to the additive pink noise. It is worth mentioning that the trend of F-score across different SNR's remains unchanged even if a stricter measure of a correct match is used by reducing the 10-frame actual vs estimate window, despite the fact that the absolute performance degrades.

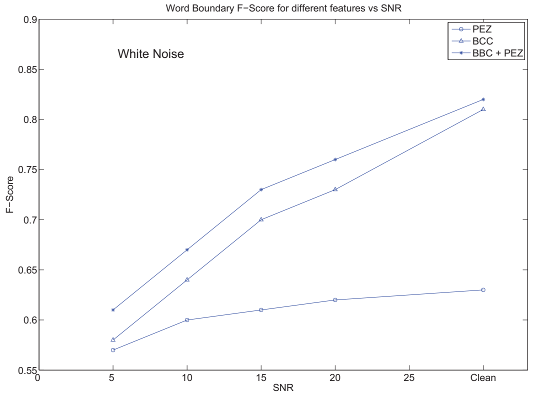

**Fig. 3**. F-score of the word boundary detection for various features at different Noise level (White Noise)

## 5. SUMMARY

In this paper, we proposed a novel lexically derived feature called the boundary confidence coefficient (BCC) in addition to acoustic features to improve automatic word boundary detection in spontaneous speech. The robustness of our feature was demonstrated through the experimental evaluation at various noise levels of white and pink noise upto 5dB. We chose 2000 utterances randomly from SwitchBoard Corpus as test set; we found that they contained 4% OOVs. In that respect, the results obtained using lexical cues are significant. We also observe that the F-score for colored noise is poorer than that of white noise at the same SNR level. At lower SNRs, the acoustic features do not significantly contribute to the word boundary detection; BCC feature also suffers due to low SNR since the ASR lattice becomes more noisy due to poor acoustics of noisy speech. However, adding these two features still improves the word boundary detection accuracy.

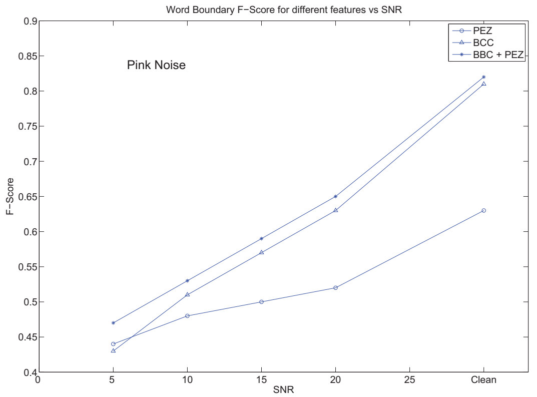

**Fig. 4**. F-score of the word boundary detection for various features at different Noise level (Pink Noise)

## 6. REFERENCES

[1] Daniel M. Brenier, Jason M. and Daniel Jurafsky, "The detection of emphatic words using acoustic and lexical features," *INTERSPEECH*, 2005.

[2] Junqua J.-C. Mak B. and Reaves B., "A robust algorithm for word boundary detection in the presence of noise," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 406–412, February 1994.

[3] S. Rajendran and B. Yegnanarayana, "Word boundary hypothesization for continuous speech in hindi based on f0 patterns," *Speech Communication*, vol. 18, pp. 21–46, January 1996.

[4] Jiann-Yow Lin Chin-Teng Lin and Gin-Der Wu, "A robust word boundary detection algorithm for variable noise-level environment in cars," *IEEE Transactions on intelligent transportation systems*, vol. 3, pp. 89–101, March 2002.

[5] J. Harrington and M. Cooper, "Word boundary detection in broad class and phoneme strings," *Computer Speech and Language*, pp. 367–382, 1989.

[6] M. Cettolo and D. Falavigna, "Automatic detection of semantic boundaries based on acoustic and lexical knowledge," *5th International Conference on Spoken Language Processing*, 1998.

[7] Abhinav Sethy, Panayiotis Georgiou, and Shrikanth Narayanan, "Text data acquisition for domain-specific language models," *In Proceedings of EMNLP*, Sydney, Australia, 2006.

[8] B. Ribeiro-Neto R. Baeza-Yates, "Modern information retrieval," *New York: ACM Press, Addison-Wesley*, 1999.