# Word Boundary Estimation for Continuous Speech Using Higher Order Statistical Features

Vijayakrishna Naganoor, Akshay Kumar Jagadish, and Krishnan Chemmangat*
*Department of Electical and Electronics Engineering,
National Institute of Technology Karnataka, India.

***Abstract*—Detection of the start and the end time of words in a continuous speech plays a crucial role in enhancing the accuracy of Automatic Speech Recognition (ASR). Hence, addressing the problem of efficiently demarcating word boundaries is of prime importance. In this paper, we introduce two new acoustic features based on higher order statistics called Density of Voicing (DoV) and Variability of Voicing (VoV) obtained from the bispectral distribution, which when used along with the popular prosodic cues helps in drastically reducing the recognition error rate involved. An ensemble of three different models has been designed to minimize the false alarms, during word boundary detection, by maximizing the uncorrelatedness in prediction from each model. Finally, the majority-voting rule was used to decide if the given frame encompasses a word boundary. The contribution of the work lies in: (i) Converting word boundary detection into a supervised learning problem (ii) Introduction of two new higher order statistical features (iii) Using ensemble methods to find the best model for prediction. Experimental results for NTIMIT Database shows the efficacy of the proposed method and its robustness to noise added during telephonic transmission.**

## I. Introduction

Determination of the word boundaries for continuous speech is a challenging task and finds immense significance in the field of Automatic Speech Recognition (ASR) as it helps to reduce the ASR problem into a more simpler single word transcription problem. Word boundary detection can be utilised to extract the Out of Vocabulary (OoV) words such as proper nouns [1] as well as for the rich transcription of speech.

Prosodic features have proven to be superior to word-level information as the speakers use prosody to impose structure on both spontaneous and read speech. Extensive work has been done for German and Indian Language using the prosodic cues [2]. Studies were conducted for observing the behaviour of the pitch pattern across the speech utterances. It was observed that the pitch frequency $F_0$ fell gradually from the beginning to the end of the utterance. The fact that $F_0$ rose from the first syllable to the last syllable in a word and fell to the first syllable in the next word was utilized. Log of the average energy [3] is another feature which has been used along with short-term energy [4] to localize the word boundary.

However, these popular 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. The problem becomes further challenging when noise is introduced in the audio files. Basic features become less reliable in the presence of different kinds of sound artifacts and noise, especially when it is non-stationary. In this paper, we propose the usage of the rudimentary acoustic features and higher-order statistical (HOS) features like kurtosis, skewness combined with two new simple yet powerful features, derived from HOS, to improve the robustness of the system. Majority-voting method was used to decide the outcome of each frame from the ensemble of three models namely, Support Vector Machines (SVM), Artificial Neural Network (ANN) and Random Forest Classifier (RFC).

The paper has been organized in the following manner, Section II and III, briefly discusses the features that were explored in the past ( [1], [5], [6] ) and the additional features that has been proposed in this work. In Section IV, experimental setup is explained with the information about the corpus, classifier setup, evaluation and the implementation of the algorithm. The results are presented in section V. Section VI concludes the work with brief description on future possibilities.

## II. Rudimentary Acoustic Features

The following are the basic acoustic cues extracted from the prosodic information and used in word boundary detection:

### A. *Short-time pitch frequency [4]*

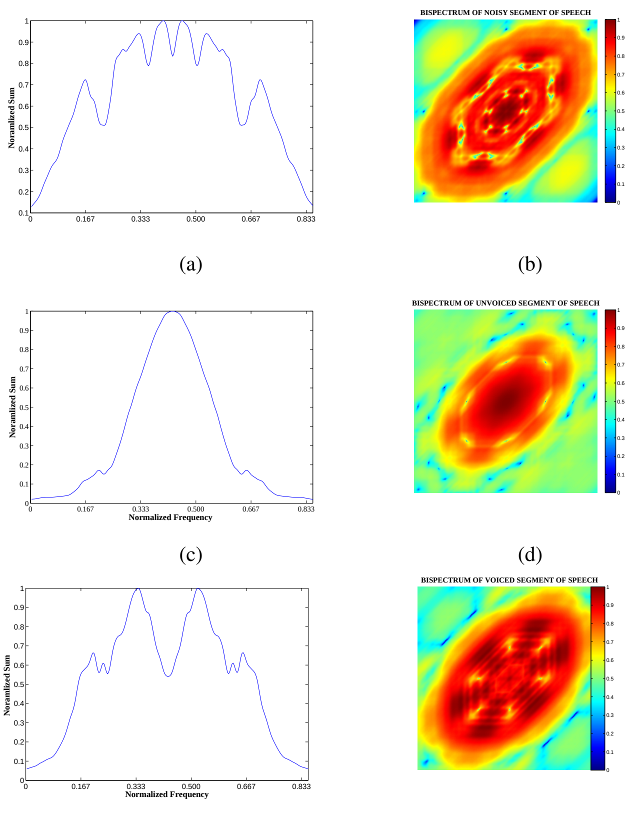
Pitch can be defined only for the voiced portion of the speech. It takes on very low values, close to zero, at segments or frames corresponding to the unvoiced region or to those that contain only noise. In general, the frame with word boundary is surrounded by the unvoiced frames and hence, it is in the region where pitch defined is zero.

### B. *Zero Line Crossing [4]*

Zero Line Crossing gives the number of times the the signal crosses the zero mark within the particular frame. This quantity is comparatively lower for frames in and around the voiced portion of speech and has higher values for the segments which correspond to the word to word transition.

### C. *Log Energy*

It has been observed that majority of the frames belonging to the transition between words have lower energy. This could be used as a cue to decide whether the frame is within the a word or closer to a word transition. It is derived from the root-mean squared energy of each frame.



(a) (b)

(c) (d)

(e) (f)

Fig. 1. **Bispectrum:** *Left Coloumn:* (a) Noisy (c) Unvoiced (e) Voiced Spectrum, **Normalized Cumalative Sum:** *Right Coloumn:* (b) Noisy (d) Unvoiced (f) Voiced Cumulated Sum

### D. *Probability of Voicing*

It gives the probability of the given frame belonging to the voiced part of the speech, instead of having a hard bound to decide the same. It is computed using Bayesian Rules on the observation metric having two components namely, maximum unnormalized template-frame correlation and minimum of the crossframe correlation.

## III. Proposed Higher-order Statistical Features

Higher-order statistics (HOS) refers to functions which consider the behaviour of third or higher order power of the given data like kurtosis, skewness,etc as opposed to more conventional techniques of lower-order statistical features like mean and variance. The general motivation behind using this higher order spectra is to derive the much needed information with regard to the deviation from Gaussianness (normality).

### A. *Skewness [7]*

It can be used to quantify the symmetry in the signal. Skewness is calculated as follows:

$$skew(X_f) = E\left(\left[\frac{X_f - \mu_f}{\sigma_f}\right]^3\right) \tag{1}$$

where $E(.)$ is the expectation operator, $X_f$ is the time series speech signal, $\mu_f$ and $\sigma_f$ are the mean and variance of the given speech frame, *f*. The effectiveness of skewness in voiced segment detection [6] and noise classification motivates its usage in the given problem.

### B. *Kurtosis*

Kurtosis [8] measures both the "peakedness" of the distribution and the heaviness of its tail and is defined as :

$$Kurt(X) = \frac{E[(X_f - \mu_f)^4]}{E[(X_f - \mu_f)^2]^2} \tag{2}$$

Kurtosis quantifies whether the shape of the data distribution matches the Gaussian distribution.

### C. *Bispectral Features*

These features are obtained by taking the 2-dimensional (2-D) Fourier transform of third order cumulant. Cumulant of a random variable (here, the time series speech signal, $X_f$) is defined as $K_f$ and is the natural logarithm of the moment-generating function:

$$K_f(t) = \log E[\exp(tX_f)] \tag{3}$$

The bispectral spectrum can be given as :

$$B_f(\omega_1, \omega_2) = \sum_{\tau=-\infty}^{\infty} K_f(\tau_1, \tau_2) e^{-j\omega_1\tau_1} e^{-j\omega_2\tau_2} \tag{4}$$

where $K_f(\tau_1, \tau_2)$ can more intuitively be seen as:

$$K_f(\tau_1, \tau_2) = E(x(\tau_1)x(\tau_2)) - E(g(\tau_1)g(\tau_2)) \tag{5}$$

In Fig.1, the bispectral distribution for noisy, voiced and unvoiced frames, obtained by averaging over multiple frames from different training examples, has been shown. By taking sum along each of the columns of the bispectrum, which is hermitian symmetric in 2-D, for normalized frequency range of [0,1]. It can be observed that the resulting distribution is symmetric about the center (0.5) and hence, only one portion has been considered for further analysis. This distribution derived for each frame, $X_f$ is referred to as $B_f$ in the rest of the paper.

On comparing the distributions and its statistical metrics for noisy, voiced and unvoiced segments. The observation made were as follows:

- *Voiced Region*: Have very few peaks in its cumulative distribution. Further, most of them are concentrated about the center of the spectrum. It can be concluded that the voiced segment has its peaks close to the origin and are relatively few in number.
- *Unvoiced Region*: Have lesser number of peaks when compared to voiced segment, and are located very close to the center with a well distinguished peak which is clearly dominant over the other ones. On proper thresholding one can conclude that unvoiced segment have distinguished peak(s) close to the point of symmetricity.

- *Noisy Spectrum*: Follows a random "noisy" distribution have many peak(s) with no clear distinguished peak. The peaks are also in close proximity to one-another.

Further, to capitalize on the characteristic behaviours mentioned above, two new features were extracted from the bispectral spectrum called Density of Voicing (DoV) and Variability of Voicing (VoV). The idea behind using both the features is to differentiate a noise-like or silent region (most likely to have word boundary) from unvoiced parts of speech (usually results as false alarms). The process of extraction of the features and their significance has been elaborated below:

*1) Density of Voicing (DoV):* Considering the idiosyncratic behaviour of each segment, a feature has been derived which takes into account the relative distances between the peaks in the cumulative plot. The feature was computed after excerpting one of the symmetric segment $B_f$. Mean subtraction was performed to remove any dc offset and thresholding to remove any noise-like structures and prevent occurrence of false peaks. Following which the feature was computed by averaging the relative distances between the consecutive peaks. *DoV* essentially gives density of the peaks that is how closely they are placed with respect to one another. It can be given by,

$$DoV(X) = \frac{1}{N_p} \sum_{i \in N_p} (P_f(i) - P_f(i-1)), \forall\, i = 1, 2, ..N_p \quad (6)$$

where $N_p$ is the total number of peaks in $Y_f$ which is the distribution (shown in Fig. 1), of one segment obtained from the bispectrum, for frame $f$ after subjecting it to mean-subtraction and thresholding ($\epsilon$) i.e $Y_f = (B_f - \mu_x) > \epsilon$ and $P_f(i)$ gives peak distance at $i$ given by:

$$P_f(i) = loc(Y_f^{ind} == 1) \quad (7)$$

where *loc(.)* is a operator designed to give the location of non-zero instances and

$$Y_f^{ind} = \begin{cases} 1, & \text{if } Y_f > 0 \\ 0, & \text{otherwise} \end{cases}$$

*2) Variability of Voicing (VoV):* The number of peaks present in cumulative distribution is fundamentally different for each of the voiced, unvoiced and silence or noisy segment of speech. It is relatively small for the unvoiced part in comparison to noisy segment. By taking into account, the variance of the distribution resulted in the desired feature called Variability of Voicing (VoV) which can be given by:

$$VoV(X) = N_p * Var(Y_f) \quad (8)$$

## IV. Experimental Set-Up

### A. About the corpus

**TIMIT** is a corpus of phonemically and lexically transcribed speech of American English speakers of different sexes and dialects. The NTIMIT (Network TIMIT) dataset has been used for conducting experiments and comparing results. NTIMIT [9] is a telephone bandwidth version of TIMIT. It was collected by transmitting the TIMIT database over the telephone network. The database consists of 630 speakers, 438 male and 138 female speakers. Speakers are categorized to one of the eight dialect regions and approximately match to the speech dialects in the American Language.

### B. Feature Extraction

The features [10] were extracted from each audio file of the NTMIT database after segmenting it into smaller segments of 320 samples each ( Which roughly corresponds to 20ms) with an overlap stride of 160 samples (50%) overlap.

### C. Classifier Setup

The central theme of the work is to convert the problem of word boundary detection into a generalized supervised learning problem. Considering dynamics of speech signals, the word boundary predicted can be assumed to be correct if the estimated location is within a few frames of the actual boundary location. Thus, providing more landmarks to train the classifiers. A total of $K$ frames in the neighbourhood of the annotated frame were chosen and labeled to the class indicative of the word boundary, based on the hypothesis that behaviour of features is similar or identical its neighbourhood. Here, we have chosen that the estimated word boundary location is considered to be correct if it is within 10 frames of the actual boundary location. Hence, $K \in [1, 5]$, where $K$= 5 indicates that 5 frames (or instances) before and after the word boundary annotated window belong to class of word boundary.

Further, to make the classification task more robust, an ensemble of learning algorithms was built, in which multiple models are learned and then, combined to improve accuracy during prediction. In this work, majority voting (soft and hard thresholding) to combine an ensemble of three classifier systems namely SVM, ANN and Random Forests.

### D. Evaluation

The evaluation metric chosen to report the result is F-score because the training data upon feature extraction is essentially skewed (atleast with the ratio 1:3, when $K$=5). This imbalance was rectified by fine tuning the hyper-parameters of the training model such that each class is weighted proportionate to ratio of the classes. Random undersampling [11] of the majority class was performed as a preprocessing step to increase the sensitivity of the classifier.

## V. Results

The parameter setting for each of the model were obtained by performing Grid Search to find the best or optimal F-measure. The resultant parameters are as follows:

- **SVM:** *Kernal*= Radial-basis function, *C*= 10, *Gamma*= 0.0001, *Probability Measure Used*= True.
- **Random Forests:** *Number of Estimators*= 10000, *Bootstrap*= True, *Minimum Sample Leafs*= 9, *Minimum Samples Split*= 3, *Criterion*= Gain in Information (gini) and *Maximum Features*= 1.
- **ANN:** *Number of Hidden Layer*= 1, *Number of Hidden Neurons*= 100, *Learning Rate*= 0.005, *Maximum Iterations*= 200, *Activation Function*= Rectified Linear Unit.

TABLE I
F-Scores for Different Prediction Frame Widths ($K$) For Different Methodologies

| Frame-Width | Ensemble Methodology | |
|---|---|---|
| K | Traditional | Modified |
| 1 | **0.221** | 0.211 |
| 2 | **0.246** | 0.239 |
| 3 | **0.348** | 0.334 |
| 4 | **0.402** | 0.400 |
| 5 | **0.467** | 0.466 |

The F-score obtained using the aforementioned three learning models for detection of word boundary were 0.460, 0.462 and 0.457 respectively, keeping the Frame width (K) as 5.

In order to further enhance the performance, ensemble of these three learning models was tried, such that each model can balance out the weaknesses of the other two. These classifiers were combined using voting classifiers - majority voting (hard voting) and weighted average probabilities (soft voting), since the classifiers considered are conceptually different. It was observed that hard voting results in higher F-score of 0.467 in comparison to soft voting which results in 0.431. Further, this test also validated that ensemble of learning models is better than each learning model taken in isolation.

Given the better performance of hard voting, a modified hard voting classifier which predicts the occurrence of word boundary if just one of the three model predict was tested and compared with traditional majority vote classifier in Table I. The conclusion drawn was that the performance of traditional majority voting based classifier is better than the modified voting classifier especially, for lower values of frame width.

In Table II, the F-score of word boundary detection for different frame widths ($K$) has been shown and as expected it was seen that F-score is better for greater value of K due to the higher margin for error.

Following the determination of the best ensemble model, the contribution of each feature for prediction of word boundary was computed using [12]. It was seen that the HOS features alone make a contribution of 46.74 % HOS, and when considered together with DoV and VoV, it contributes more than 75%. As shown in Fig. II, these results were backed up by improvement in F-scores indicating that the newly introduced features play a pivotal role during detection of word boundaries.

TABLE II
F-Scores for Different Prediction Frame Widths ($K$) for Different Set of Features

| Frame-Width | F-Measure | |
|---|---|---|
| K | Acoustic only | Acoustic+HOS+DoV+VoV |
| 1 | 0.159 | **0.221** |
| 2 | 0.215 | **0.246** |
| 3 | 0.271 | **0.348** |
| 4 | 0.336 | **0.402** |
| 5 | 0.383 | **0.467** |

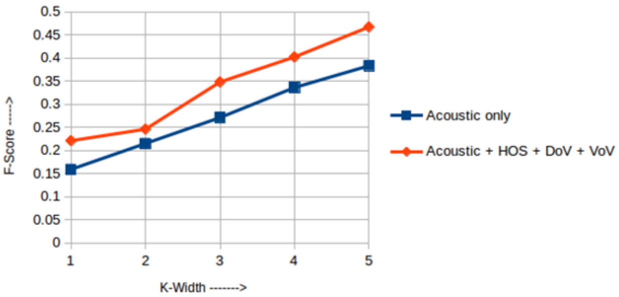


Fig. 2. **Effect of Proposed Features:** The Improvement in F-Measure using Variability of Voicing and Density of Voicing

## VI. Conclusion

In conjunction with other works in speech recognition, efficient estimation of word boundaries is an essential precursor for ASR, and will eventually lead to an overall powerful speech recognition system. In this paper, the problem of world boundary detection was converted to a simple classification problem and two new features called DoV and VoV has been proposed which when coupled with acoustic features and other higher order statistical features improves the detection rate of word boundary.

## References

[1] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell System Technical Journal*, vol. 54, no. 2, pp. 297–315, 1975.

[2] G. R. Rao and J. Srichland, "Word boundary detection using pitch variations," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 2. IEEE, 1996, pp. 813–816.

[3] A. Agarwal, A. Jain, N. Prakash, and S. Agrawal, "Word boundary detection in continuous speech based on suprasegmental features for hindi language," in *Signal Processing Systems (ICSPS), 2010 2nd International Conference on*, vol. 2. IEEE, 2010, pp. V2–591.

[4] A. Tsiartas, P. K. Ghosh, P. Georgiou, and S. Narayanan, "Robust word boundary detection in spontaneous speech using acoustic and lexical cues," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 4785–4788.

[5] L. F. Lamel, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilpon, "An improved endpoint detector for isolated word recognition," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 29, no. 4, pp. 777–785, 1981.

[6] B. Reaves, "Comments on" an improved endpoint detector for isolated word recognition," *Signal Processing, IEEE Transactions on*, vol. 39, no. 2, pp. 526–527, 1991.

[7] J. Kaur and M. Jureka, "Speech detection using high order statistic (skewness)."

[8] R. Kompe, J. Siekmann, and J. Carbonell, *Prosody in speech understanding systems*. Springer-Verlag New York, Inc., 1997.

[9] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "Ntimit: A phonetically balanced, continuous speech, telephone bandwidth speech database," in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*. IEEE, 1990, pp. 109–112.

[10] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, pp. 321–357, 2002.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.