Word Boundary Detection for Continuous speech using Higher Order Statistical features

Vijayakrishna Naganoor, Akshay Kumar J, Krishnan Chenmmangat

Department of Electrical Engineering,
National Institute Of Technology Karnataka.

NITK Surathkal
Mangalore, India

Outline of the Presentation

- Problem formulation
- Advantages of addressing the problem efficiently
- Previous works and their merits
- New ideas introduced to tackle the problem statement
  I. Using statistical features to solve the problem
  II. Features derived from Bispectral spectrum

- Experimental setup
- Improvement because of new features
- Improvement in classification due to ensemble model
- Conclusion and future work

- Problem formulation
- Advantages of addressing the problem efficiently
- Previous works and their merits
- New ideas introduced to tackle the problem statement
    I. Using statistical features to solve the problem
    II. Features derived from Bispectral spectrum

- Experimental setup
- Improvement because of new features
- Improvement in classification due to ensemble model
- Conclusion and future work

Problem formulation

- The problem of estimating the word boundaries corresponds to finding the start and the end time of the words in a sentence.
- 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.
- Here,in this paper we have considered that the estimated word boundary location to be correct if it is within 100 milliseconds from the actual boundary location.
- The problem becomes further challenging when noise is introduced in the audio files.

Outline of the Presentation

- Problem formulation
- Advantages of addressing the problem efficiently
- Previous works and their merits
- New ideas introduced to tackle the problem statement
  I. Using statistical features to solve the problem
  II. Features derived from Bispectral spectrum

- Experimental setup
- Improvement because of new features
- Improvement in classification due to ensemble model
- Conclusion and future work

Why accurate Word Boundary Detection is important ?

- Accurate Boundary detection helps to reduce the Automatic Speech Recognition 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
- It is also helpful for rich transcription of speech

Outline of the Presentation

- Problem formulation
- Advantages of addressing the problem efficiently
- Previous works and their merits
- New ideas introduced to tackle the problem statement
  I. Using statistical features to solve the problem
  II. Features derived from Bispectral spectrum

- Experimental setup
- Improvement because of new features
- Improvement in classification due to ensemble model
- Conclusion and future work

Rudimentary Acoustic features previously used

- ## Short-time pitch frequency
  Pitch can be defined only for the voiced portion of the speech.The frame with word boundary is surrounded by the unvoiced frames and hence, it is in the region where pitch defined is zero.

- ## Zero Line Crossing
  Zero Line Crossing gives the number of times the the signal crosses the zero mark within the particular frame.

- ## Log Energy
  It is derived from the root-mean squared energy of each frame

- ## Probability of Voicing
  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.

Outline of the Presentation

- Problem formulation
- Advantages of addressing the problem efficiently
- Previous works and their merits
- New ideas introduced to tackle the problem statement
  I. Using statistical features to solve the problem
  II. Features derived from Bispectral spectrum
- Experimental setup
- Improvement because of new features
- Improvement in classification due to ensemble model
- Conclusion and future work

Proposed Higher Order Statistical Features used

- **Skewness**
  It is a measure that is used to quantify the symmetry in the signal.
  This has been effectively used in voiced segment detection and noise classification.
- **Kurtosis**
  Kurtosis is the ”peakedness” of the distribution and as well as the heaviness of its tail.
  It quantifies whether the shape of the data distribution matches the Gaussian distribution.
- **Bispectral Features**
  These are obtained by taking the 2-dimensional fourier transform of the third order cumulant.

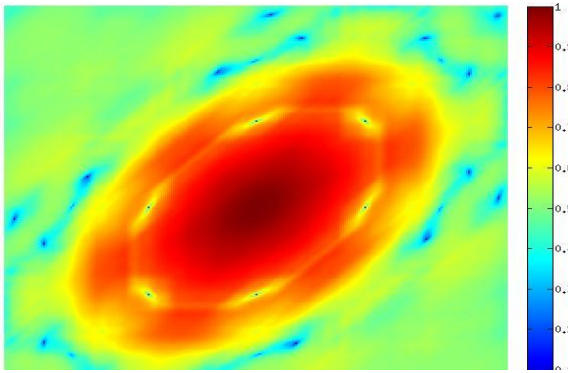
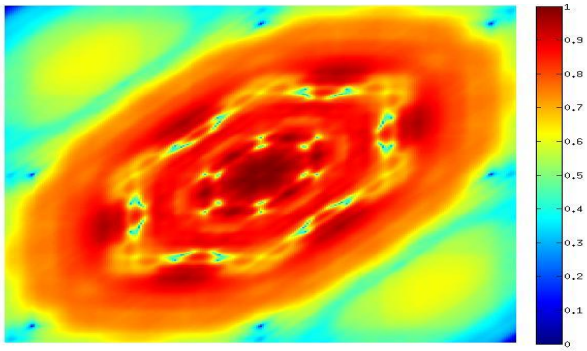
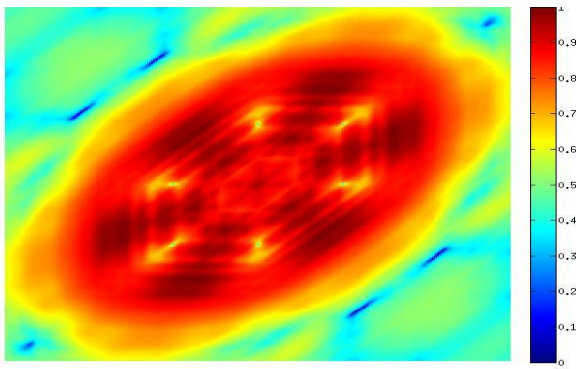
If third order cumulant is defined as K_f-

$$K_f(t) = \log E[\exp(tX_f)]$$

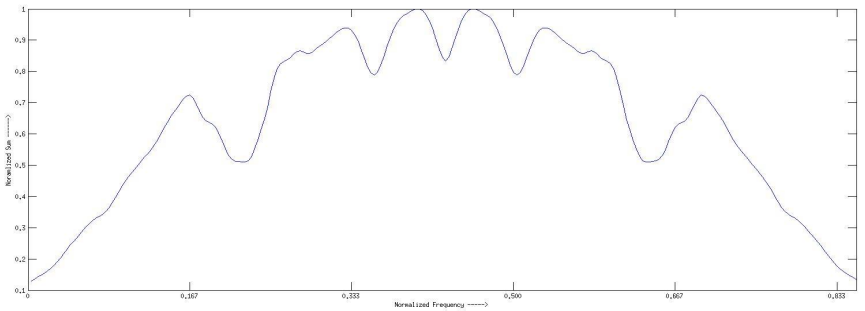
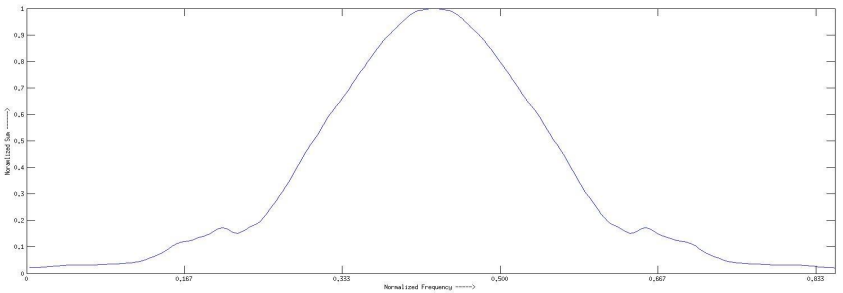
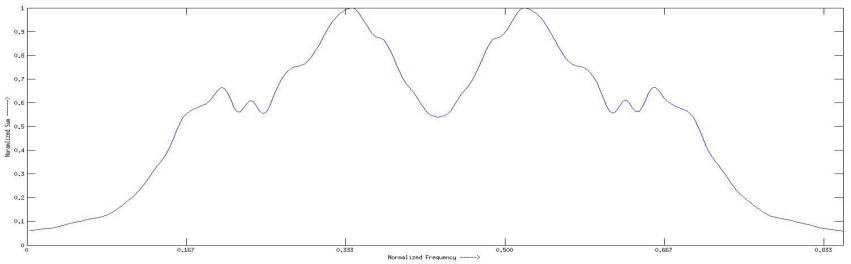
$$B_f(\omega_1, \omega_2) = \sum_{\tau=-\infty}^{\infty} K_f(\tau_1, \tau_2) e^{-j\omega_1\tau_1} e^{-j\omega_2\tau_2}$$

Two features named ‘DoV’ and ‘VoV’ have been derived from the the bispectral distribution.

Analysis using Bispectral Distribution

Noisy Frame	
Unvoiced Frame	
Voiced Frame	

More analysis using Bispectral distribution

Noisy	
Unvoiced	
Voiced	

- Problem formulation
- Advantages of addressing the problem efficiently
- Previous works and their merits
- New ideas introduced to tackle the problem statement
  I. Using statistical features to solve the problem
  II. Features derived from Bispectral spectrum
- Experimental setup
- Improvement because of new features
- Improvement in classification due to ensemble model
- Conclusion and future work

Features extracted using Bispectral Features

- ## Density of Voicing (DoV)

  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.

- ## 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 the product of the number of peaks and the variance of Y_f where, Yf is the distribution of one segment obtained from the bispectrum, for frame f after subjecting it to mean subtraction.

- Problem formulation
- Advantages of addressing the problem efficiently
- Previous works and their merits
- New ideas introduced to tackle the problem statement
    I. Using statistical features to solve the problem
    II. Features derived from Bispectral spectrum

- Experimental setup
- Improvement because of new features
- Improvement in classification due to ensemble model
- Conclusion and future work

Experimental set-up

- **Database**
  The NTIMIT (Network TIMIT) dataset has been used for conducting experiments and comparing results. NTIMIT is a telephone bandwidth version of TIMIT. It was collected by transmitting the TIMIT database over the telephone line.

- **Feature Evaluation**
  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.

- Problem formulation
- Advantages of addressing the problem efficiently
- Previous works and their merits
- New ideas introduced to tackle the problem statement
    I. Using statistical features to solve the problem
    II. Features derived from Bispectral spectrum

- Experimental setup
- Improvement because of new features
- Improvement in classification due to ensemble model
- Conclusion and future work

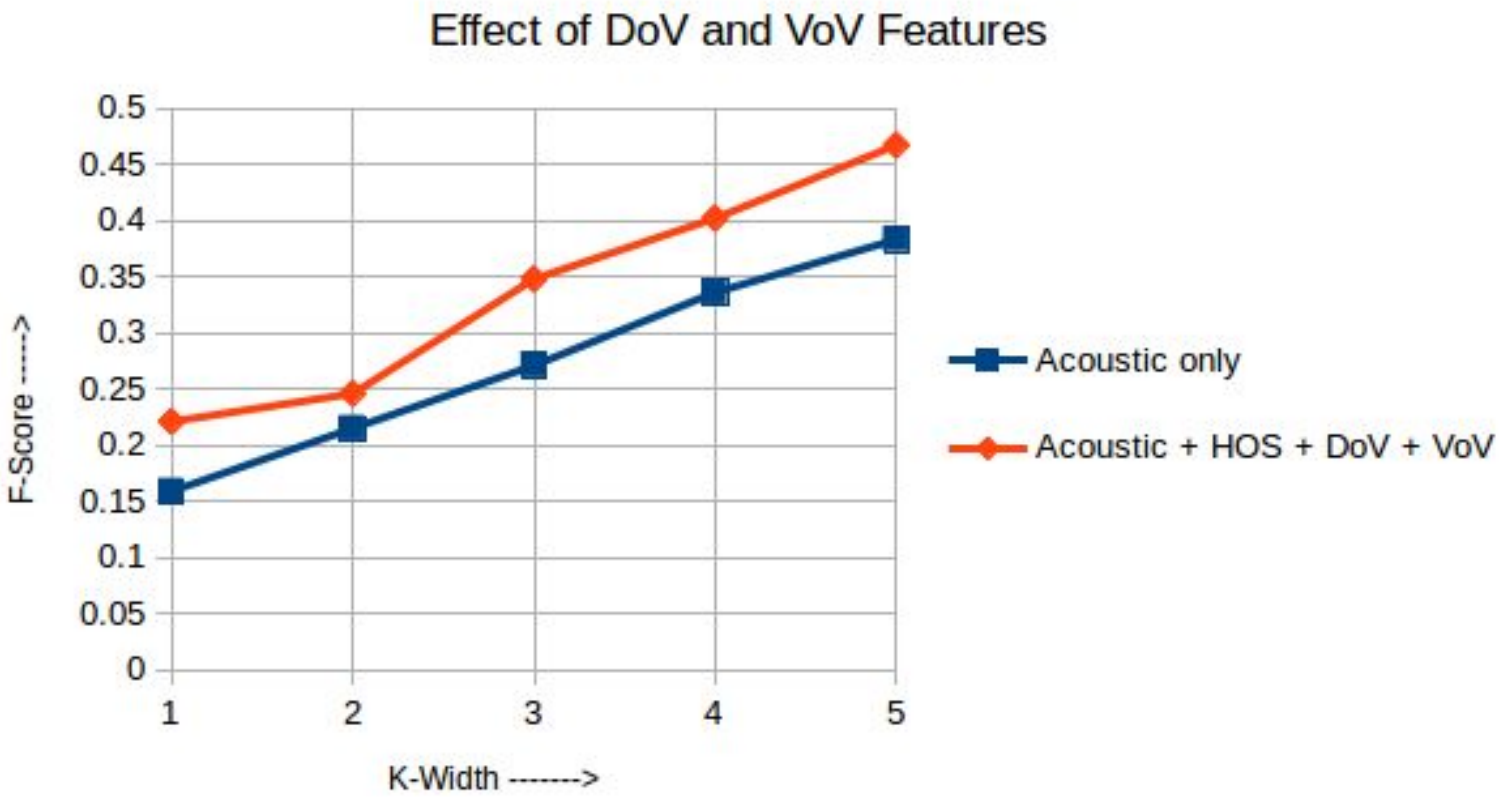
Improvement in the results

F-Score has been to report the results as the ratio of the number of the frames containing word boundaries to non-word boundary is skewed

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

Ensemble Model using Majority Voting

It can be observed that the ensemble model performs consistently better in comparison to the each model taken in isolation



- Problem formulation
- Advantages of addressing the problem efficiently
- Previous works and their merits
- New ideas introduced to tackle the problem statement
    I. Using statistical features to solve the problem
    II. Features derived from Bispectral spectrum

- Experimental setup
- Improvement because of new features
- Improvement in classification due to ensemble model
- Conclusion and future work

Improvement in the results

F-Score has been to report the results as the ratio of the number of the frames containing word boundaries to non-word boundary is skewed

Frame Width	Individual classifiers			Ensemble Methodology	
K	SVM	NN	RFC	Traditional	Modified
1	0.171	0.201	0.201	0.221	0.211
2	0.199	0.225	0.215	0.246	0.239
3	0.248	0.331	0.329	0.348	0.334
4	0.366	0.382	0.374	0.402	0.400
5	0.417	0.451	0.441	0.467	0.466

- Problem formulation
- Advantages of addressing the problem efficiently
- Previous works and their merits
- New ideas introduced to tackle the problem statement
    I. Using statistical features to solve the problem
    II. Features derived from Bispectral spectrum

- Experimental setup
- Improvement because of new features
- Improvement in classification due to ensemble model
- Conclusion and future work

Conclusion and future work

Our contribution

- We have converted word boundary detection into a supervised learning problem
- Introduction of two new higher order statistical features
- Using ensemble methods to find the best model for prediction.

Future Work

- Along with the acoustic features which we have
- Considering syntactic cues such as will further enhance the performance