

# Effective Segmentation based on Vocal Effort Change Point Detection<sup>1</sup>

Chi Zhang and John H.L. Hansen

Center of Robust Speech Systems (CRSS)  
Erik Jonsson School of Engineering & Computer Science  
University of Texas at Dallas, Richardson, Texas 75083, USA  
{cxz055000, john.hansen}@utdallas.edu <http://crss.utdallas.edu>

## Abstract

Non-neutral speech data has a strong negative impact on speech processing systems such as Automatic Speech Recognition (ASR) or speaker ID systems [1]. It is therefore necessary to detect and segment non-neutral speech data before further processing steps. Alternatively, the detection and segmentation of non-neutral speech segments from an input speech stream can be used in speech analysis and understanding, or in speech file retrieval systems to detect speech files containing whispered speech representing sensitive information, or shouted speech denoting strong emotion. This study addresses the segmentation problem for vocal effort change by deploying an improved feature based  $T^2$ -BIC algorithm. Several features are considered as input to the  $T^2$ -BIC algorithm in this study. A new fused evaluation criterion, Multi-Error Score (MES), is proposed to explore which feature conveys the most information on vocal effort. Results show that the lowest mean MES (56.49) occurs for the energy ratio feature for segmentation of different vocal effort speech segments based on vocal effort change point detection. Finally, recommendations are made for integrating this framework to advance knowledge processing for subsequent speech systems.

**Index Terms:** Segmentation, vocal effort, change point detection, whispered speech, shouted speech,  $T^2$ -BIC

## 1. Introduction

The purpose of speech segmentation is to partition an input speech stream into a number of segments based on the characteristic differences between these segments. The most characteristic differences of segments are generally based on content (speech, music or noise, etc), source (speaker change), environment (noisy or silent), or channel. As a preliminary stage, speech segmentation is necessary for speech content analysis, speech information retrieval [9], speech clustering and collection of training data for Automatic Speech Recognition (ASR) systems or speaker ID systems. While much work has been conducted in segmentation, most studies have considered one homogeneous vocal effort of speech data as neutral speech.

From [1] we know that the speech signal can be generally classified into five categories: whispered, soft, neutral, loud and shouted, based on differences in vocal effort. As shown in [1], different vocal effort speeches have discriminative acoustic characteristics such as sound intensity level, sentence/silence duration, frame energy distribution and spectral tilt. It is also known that speech data other than

neutral speech will negatively affect the performance of speech processing systems. For example, speaker ID systems degrade rapidly when tested with mismatched vocal effort speech [1]. Thus, it is necessary to detect non-neutral speech segment from the speech stream, before it is processed further. A corresponding compensation can be performed to reduce the negative impact of the non-neutral speech to maintain performance. In most cases, the vocal effort of speech can indicate the importance of the information conveyed by the speech signal or the emotion the speaker has while speaking. For example, whispered speech has high probability of conveying confidential or sensitive information; shouted speech generally denotes angry emotion a speaker might have for interactive systems. Hence for a speech file retrieval system, the vocal effort based segmented speech files should not only allow retrieval of the speech file containing non-neutral speech, but also help find speech conveying confidential information or speaker emotion. To segment the speech stream, the acoustic change points between segments need to be detected. Some prior studies have explored audio segmentation based on the detection of acoustic change points. In [2], new features have been deployed in the audio segmentation for unsupervised multi-speaker change detection. In [3], an efficient audio stream segmentation algorithm has been developed via the combined  $T^2$  statistic and Bayesian information criterion. However, no research has focused on the speech segmentation of vocal effort change.

In this paper, we report on an effective unsupervised speech segmentation algorithm based on the vocal effort change point detection. The  $T^2$ -BIC algorithm described in [2] has been incorporated to segment an audio stream for vocal effort change. Several acoustics features, within the  $T^2$ -BIC algorithm, are considered for performance in vocal effort change point detection. Finally, a new multi-error score is proposed to evaluate the performance of each feature.

This paper is organized as follows; Sec. 2 develops a new evaluation criterion: multi-error score, Sec. 3 introduces the  $T^2$ -BIC algorithm used for audio segmentation. Sec. 4 addresses the acoustic features employed within the  $T^2$ -BIC algorithm. Next, Sec. 5 describes the speech corpus developed of five speech vocal effort modes, followed by details on experiments of segmentation for vocal effort change points performed on the corpus. Finally, we close in Sec. 7 with some concluding remarks.

## 2. Multi-Error Score

The goal of reliable segmentation in audio streams requires that we assess the mismatch between hand/human

---

<sup>1</sup> This work was funded by RADC under contract A40104, and by University of Texas at Dallas under Project EMMITT

segmentation and automatic segmentation. Traditional methods emphasize frame precision scores, but these do not take into account the desired continuity conditions needed for speech recognition. Mismatch can be described by three error scores: miss detection rate, false alarm rate and average mismatch in milliseconds. The following figure illustrates these three types of error:

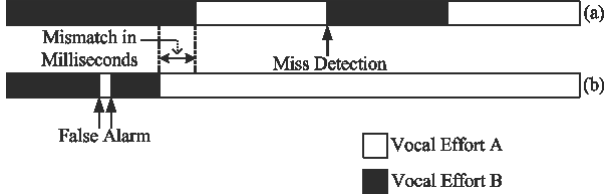


Figure 1: An example of break points detection of an audio file. (a) Actual break points; (b) Experimental break points

For the vocal effort segmentation case, the false alarm error rate can be compensated by merging two very close segments of common vocal effort, or by merging two adjacent segments classified as the same vocal effort in the later vocal effort classification step. Hence, the false alarm error is less important than the miss detection error in the evaluation of segmentation. Furthermore, the average mismatch between experimental and actual break points is an important norm which reflects break point accuracy for the feature and data. We propose a new combined evaluation criterion that fuses these three error scores into an overall performance measure. To fuse the average mismatch in milliseconds with false alarm rate and miss detection rate in percentage, we obtain the average mismatch rate by averaging the percentage of the mismatch of total duration of two segments corresponding to the actual break points. The multi-error score can be expressed with the following equation:

$$\text{Multi-Error Score} = \text{False Alarm Rate} + \text{Miss Detection Rate} * 3 + \text{Mismatch Rate} * 2 \quad (1)$$

The ideal segmentation has zero false alarms, zero miss detection and zero mismatches in millisecond, and thus the multi-error score will be zero in the ideal case. If in the case that the mismatch rate, false alarm rate and miss detection rate are all no greater than 10%, the resulting multi-error score is no more than 60, which can denote a fair performance in segmentation. If mismatch rate is still no greater than 10%, but the miss detection rate and mismatch rate are greater than 15%, the multi-error score is greater than 80 which means the segmentation is collectively considered as performing bad.

### 3. $T^2$ -BIC Algorithm

Bayesian Information Criterion (BIC) has been proved as an effective metric based method in segmentation for the segments greater than 5 seconds in [6] and [7]. However the BIC method needs the second statistics (i.e., the covariance), which is problem due to estimation error with insufficient data for segments less than 5 seconds. In [2], the  $T^2$ -statistic has been incorporated to detect the break points of speaker change. For the segments even less than 2 seconds, the covariance is assumed to be an identity matrix to calculate the  $T^2$ -mean distance. In [3], a  $T^2$ -BIC algorithm which combined  $T^2$ -statistics and BIC has been proposed. It is shown in [3] and [8] that,  $T^2$ -BIC segmentation works well for both segments that are less than 5 seconds and segments more than 5 seconds. In this study, the segments length is between

1-3seconds, thus the  $T^2$ -BIC algorithm has been used to detect the change points with different vocal efforts.

## 4. Features Used in Experiments

$T^2$ -BIC algorithm detects acoustic change points based on the input feature data. In this study, the following features are considered to explore the most effective features in segmentation for vocal effort change point detection:

- 1) MFCC: one of the most widely used features, 13-Dimension MFCCs are used here.
- 2) ZEPS: a 4-Dimension feature contains Zero Cross Rate, Energy, Pitch and Energy slope.
- 3) Energy Ratio (ER): the ratio between the energy in high band (2800-3000 Hz) and the energy in the low band (450-650 Hz), (fully described in [4]).
- 4) Spectral Information Entropy (SIE) [5]: For each frame, the spectrum obtained from FFT can be viewed as a vector of coefficients in an orthonormal basis. Hence, the probability density function (pdf) can be estimated by the normalization over all frequency components. The SIE can be obtained from this estimated pdf. Here, only the spectrum between 300 Hz and 3000 Hz is considered. Each frame is evenly divided into 4 sub-frames. The SIE of each sub-frame is calculated to form a 4-Dimension SIE feature of each frame.
- 5) Spectral Tilt (ST): As shown in [1], spectral tilt is a distinguish feature between vocal efforts. Here, a 1-Dimension spectral tilt feature is used as input to  $T^2$ -BIC.

## 5. Corpus Development

All speech data used for this study are recorded in an ASHA certified, single walled sound booth using a multi-track FOSTEX 8-channel synchronized digital recorder with gain adjustments for individual channels. A total of 12 male, native English-speaking subjects participated in the data collection. For each subject, the speech was recorded in several tokens using three microphones- a throat microphone, a close-talking and a far field microphone. A 1 kHz sinusoid signal generated by an NTI analog audio generator was played by an ALTEC speaker as the calibration test tone in all recordings. At the beginning of each token, the volume of the test tone was carefully adjusted to ensure a 75 dB sound pressure level (SPL) of the test tone using a QUEST sound level meter. The test tone was recorded with all three microphones. The position of the subject, the location of the calibration test tone speaker, and the location of the sound level meter were all positioned in an equidistant triangle separated by 75cm.

The data collection procedure was divided into 3 phases for each subject. Phase I consists of 2 sessions which has 5 tokens corresponding to the five speech modes. In each token, 5 sentences from the TIMIT database were read in one of five speech modes and recorded. Phase II consists of 20 sentences which were read in the neutral mode. Phase III includes the spontaneous speech of one-minute duration in each vocal mode. Vocal models include: whispered, soft, neutral, loud, shouted.

## 6. Experiments and Results

Speech data in Phase I from the 12 subjects, produced with the close-talking microphone was used in the following

experiments. In this study, the carefully measured and recorded 75 dB-SPL test tone is used to normalize the speech data under all five vocal efforts. A frame-energy based approach has been used to remove the silence at the lead and tail parts of each sentence. All the experiments are categorized into two scenarios: same speaker scenario and different speaker scenario. The results of the experiments are discussed below.

### 6.1. Same Speaker Scenario

To test the performance of  $T^2$ -BIC algorithm using 5 candidate features for vocal effort change point detection, each sentence from four types of non-neutral speeches has been inserted between two neutral sentences to form a concatenated sentence with two vocal effort change points (Ne+Wh+Ne; Ne+So; Ne+Lo+Ne; Ne+Sh+Ne, where Wh, So, Ne, Lo, Sh denote the vocal effort whispered, soft, neutral, loud and shouted respectively). In the same speaker scenario, all three sentences for concatenation are from the same speaker. The segmentation results are shown as follows:

Table 1-5: Segmentation results of  $T^2$ -BIC algorithm based on five feature sets for different vocal effort change points for same speaker scenario. MDR denotes miss detection rate, FAR denotes false alarm rate, MMR denotes mismatch rate and MES denotes multi-error score.

Feature Type: 13-D MFCC		Average MES:64.38		
Vocal effort	MDR	FAR	MMR	MES
Whispered	9.58	7.11	7.29	50.43
Soft	20.00	7.20	9.91	87.02
Loud	16.67	5.88	8.03	71.95
Shouted	10.42	5.02	5.91	48.10

Feature Type: 4-D ZEFS		Average MES:69.82		
Vocal effort	MDR	FAR	MMR	MES
Whispered	9.58	4.14	5.56	44.00
Soft	23.75	5.41	9.41	95.48
Loud	20.00	2.89	8.98	80.85
Shouted	13.75	2.64	7.53	58.95

Feature Type: 4-D SIE		Average MES: 78.03		
Vocal effort	MDR	FAR	MMR	MES
Whispered	7.91	4.11	7.15	42.17
Soft	27.08	9.09	10.79	111.93
Loud	22.50	7.58	11.42	97.92
Shouted	12.08	5.17	9.34	60.11

Feature Type: 1-D ER		Average MES:56.49		
Vocal effort	MDR	FAR	MMR	MES
Whispered	6.67	10.30	8.81	47.92
Soft	10.83	8.94	11.04	63.53
Loud	10.41	6.30	10.82	59.29
Shouted	8.33	8.87	10.68	55.23

Feature Type: 1-D ST		Average MES:89.92		
Vocal effort	MDR	FAR	MMR	MES
Whispered	10.00	3.79	9.12	52.03

Soft	25.83	5.41	12.77	108.46
Loud	25.42	4.71	11.68	104.33
Shouted	22.50	4.47	11.43	94.84

The segmentation results in each table of Table 1-5 show that the MES of soft and loud speech is much higher than those of whispered and shouted speech. These facts correspond to the results in [1], where vocal efforts of soft speech and loud speech are most close to that of neutral speech. Comparing the experiment results for these five features, it is obvious that the ER feature has the lowest mean MES (56.49) for 4 non-neutral vocal efforts, which suggests that the ER feature is more sensitive to the vocal effort change than the other 4 features.

### 6.2. Different Speaker Scenario

In this scenario, each sentence from all four non-neutral speeches has been inserted between two neutral sentences to form a concatenated sentence with two vocal effort change points (same as Sec. 6.1). However, two segments besides each vocal effort change points are from different speakers (Ne1+Wh2+Ne1, Ne1+So2+Ne1, Ne1+Lo2+Ne1, Ne1+Sh2+Ne1, where 1 and 2 denotes different speakers). The segmentation results are shown as follows:

Table 6-10: Segmentation results of  $T^2$ -BIC algorithm based on five feature sets for different vocal effort change points for different speaker scenario. MDR denotes miss detection rate, FAR denotes false alarm rate, MMR denotes mismatch rate and MES denotes multi-error score.

Feature Type: 13-D MFCC		Average MES: 46.44		
Vocal effort	MDR	FAR	MMR	MES
Whispered	5.42	6.46	7.12	36.96
Soft	11.25	5.42	8.12	55.41
Loud	12.08	4.74	5.99	52.96
Shouted	7.50	5.66	6.18	40.52

Feature Type: 4-D ZEFS		Average MES: 64.14		
Vocal effort	MDR	FAR	MMR	MES
Whispered	8.75	4.50	6.11	42.97
Soft	20.41	2.83	7.29	78.64
Loud	19.58	3.86	8.90	80.40
Shouted	11.25	3.89	7.96	53.56

Feature Type: 4-D SIE		Average MES: 69.25		
Vocal effort	MDR	FAR	MMR	MES
Whispered	6.25	7.87	7.26	41.16
Soft	19.58	5.60	10.10	84.57
Loud	22.08	4.85	9.91	90.93
Shouted	12.50	4.80	9.02	60.36

Feature Type: 1-D ER		Average MES: 54.45		
Vocal effort	MDR	FAR	MMR	MES
Whispered	6.67	6.15	8.17	42.50
Soft	11.25	8.33	11.13	64.34
Loud	10.00	8.13	10.80	59.73
Shouted	7.50	6.97	10.89	51.25

Table 10

Feature Type: 1-D ST			Average MES: 79.90	
Vocal effort	MDR	FAR	MMR	MES
Whispered	10.83	6.25	8.83	56.42
Soft	20.41	4.26	11.37	88.26
Loud	21.67	5.36	11.29	92.96
Shouted	19.17	3.31	10.57	81.97

Similar to that seen in Sec. 6.1, the MES of soft and loud speech is much higher than those of whispered and shouted speech. Compared with segmentation results in corresponding tables in the Sec. 6.1, the MES are generally smaller, which indicates improvement in segmentation. These improvements illustrate the fact that these features are not only carrying information of vocal effort but also conveying more or less speaker dependant information. The algorithm detects not only the vocal effort changes but also the speaker changes based on these features. Therefore, although MFCCs case has the smallest average MES, the MES score difference for ER feature between Table 4 and 9 is obviously the smallest compared with the MES change between the other 4 pair result tables, implying that ER feature is more reliable on vocal effort change detection. The experiment results in both scenarios indicate that the ER feature not only is more sensitive to vocal effort change but also containing more vocal effort information than the other 4 features.

## 7. Conclusion and future work

This study has considered a segmentation application for partitioning the speech audio into different vocal effort speech segments. In this study, a feature based  $T^2$ -BIC algorithm has been developed to fulfill the segmentation task. Five candidate acoustic features were considered to determine the best performance in vocal effort change point detection. A new evaluation criterion, named Multi-Error Score (MES) was proposed. The smaller the value of MES, the better the performance in segmentation. The experiment results in MES show that the ratio between the energy in high band (2800-3000 Hz) and the energy in the low band (450-650 Hz) can be an accurate and reliable feature for vocal effort segmentation.

The experimental results are consistent with the facts found in [1] that, whispered and shouted speech have the most dramatic vocal effort difference compared to neutral speech. Furthermore, that whispered speech usually conveys sensitive information and the shouted speech mostly denotes strong emotion. These facts motivated us to focus on segmentation and classification of speech with whispered and shouted vocal efforts versus only looking at speech with neutral vocal effort, which means the detection of whispered/shouted islands embedded in neutral speech. These advances can help improve speech recognition/speaker ID for continuous audio stream where variable vocal effort is presented.

## 8. References

[1] Zhang, C. and Hansen, J.H.L., "Analysis and Classification of Speech Mode: Whispered through Shouted", INTERSPEECH 2007-ICSLP, pp.2289-2292, 2007.

[2] Huang, R. and Hansen, J.H.L., "Advances in Unsupervised Audio Segmentation for the Broadcast News and NGSW Corpora", ICASSP 2004, pp.741-744, 2004.

[3] Zhou, B. and Hansen, J.H.L., "Efficient Audio Stream Segmentation via the Combined  $T^2$  Statistic and Bayesian Information Criterion", IEEE Trans. Speech and Audio Processing, Vol. 13, No.4, July 2005.

[4] Wenndt, S.J., "A study on the Classification of whispered and Normally Phonated Speech", INTERSPEECH2006, pp.649-652, 2002.

[5] Shen, J., Hung, J. and Lee, L., "Robust Entropy-based endpoint detection for speech recognition in noisy environments", ICSLP, 1998, pp.232-235, 1998.

[6] Chen, S. and Gopalakrishnan, P., "Speaker, Environment and Channel Change Detection and Clustering via The Bayesian Information Criterion", Proc. Broadcast News Transcr. & Under., Workshop, 1998.

[7] Johnson, S., "Speaker Tracking", Master Thesis, Engineering Department, Cambridge University, UK, 1997.

[8] Huang, R. and Hansen, J.H.L., "Advances in Unsupervised Audio Classification and Segmentation for the Broadcast News and NGSW Corpora", IEEE Trans., Audio, Speech, and Language Processing, pp.907-919, 2006

[9] Hansen, J.H.L., Rongqing Huang, etc, "SpeechFind: Advances in Spoken Document Retrieval for a National Gallery of the Spoken Word", IEEE Trans., Speech and Audio Processing, vol.13, pp.712-730, 2005