

# Advancements in Whisper-Island Detection within Normally Phonated Audio Streams

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

In this study, several improvements are proposed for improved whisper-island detection within normally phonated audio streams. Based on our previous study, an improved feature, which is more sensitive to vocal effort change points between whisper and neutral speech, is developed and utilized in vocal effort change point (VECP) detection and vocal effort classification. Evaluation is based on the proposed multi-error score, where the improved feature showed better performance in VECPs detection with the lowest MES of 19.08. Furthermore, a more accurate whisper-island detection was obtained using the improved algorithm. Finally, the experimental detection rate results of 95.33% reflects better whisper-island detection performance for the improved algorithm versus that of the original baseline algorithm.

**Index Terms:** whisper, feature, segmentation, detection, vocal effort,  $T^2$ -BIC

## 1. Introduction

Whisper is one mode of natural speech communication with reduced perceptibility and significant loss in understanding. Whispered speech can occur with normal physiological blocking of vocal fold vibrations[5]. Furthermore, as a paralinguistic phenomenon, whispered speech can be used in different contexts. One may wish to communicate clearly, but be in a situation where the loudness of normal speech is prohibited, such as in a library or in a formal setting. On the other hand, one may be whispering to avoid being overheard, in which case some loss of understanding in context may be desirable[1]. However, current speech processing systems are generally designed for normally phonated speech. In [3], based on experimental results for a close-set Speaker-ID system, we observed that whispered speech has the most dramatic loss in performance for speech processing systems. This is mainly because of the fundamental difference in speech production mechanism of whisper and neutral speech[12][3]. Therefore, whispered speech requires further study not only on its acoustic characteristics but also for speech applications. Since whispered speech can be effectively used for quiet and private communications, processing techniques for whispered speech become more important for speech communication systems.

Recently, several studies have emerged in the field of whispered speech processing. The investigation of whispered speech is interesting from a theoretical point of view in speech production and perception, and for practical needs in whispered speech and speaker recognition. In [1], acoustic properties and a recognition method for whispered speech were discussed. The isolated whispered and phonated vowels produced by male adults were acoustically compared in [2]. In [5], the study was narrowed on the analysis of consonants in whispered speech in Serbian language. However, little systematic effort has been reported on how to detect and locate whispered speech within a normally phonated audio stream. To utilize techniques that cope with whispered speech, the locations of whispered speech need to be identified within audio streams. Furthermore, because of the high probability for whispered speech conveying confidential or sensitive information, the detection and identification of whispered speech in audio files can help a spoken documents retrieval system or a call center monitoring system. Several preliminary studies have investigated methods to identify whispered speech or segment non-neutral speech. In [7], a technique for automatically classifying normally phonated speech and whispered speech was proposed. Although the highest correct classification rate of the technique is 95% (57/60), the 4.8 seconds analysis frame length prevents it from being applied to detect the precise boundaries between whispered and neutral speech. In [6], an effective method of detecting vocal effort change points (VECPs) between non-neutral speech and neutral speech was presented. However, the vocal effort of the speech segment between two consecutive vocal effort change points cannot be identified using that algorithm. In [13], an entropy-based feature was developed to effectively detect VECPs between whisper and neutral speech. The algorithm of whisper-island detection based on that feature was also proposed in [13]. In this study, improvements of the entropy-based feature and algorithm in [13] are proposed. The experimental results showed that the improved feature and algorithm have better performance in VECPs detection between whisper and neutral speech and in whisper-island detection as well. To evaluate the performance of the algorithm, two corpora having speech data under different vocal efforts and normally phonated speech embedded with whispered speech were developed. All analysis and experiments were performed on the data from the corpora.

The remainder of this study is organized as follows. First, the corpora developed are introduced in Sec. 2, followed by the baseline system of whisper-island detection in Sec. 3. Next, details of the improved feature and improved algorithm are il-

---

This project was funded by AFRL under a subcontract to RADC Inc. under FA8750-05-C-0029. Recently approved for public release; distribution unlimited.

illustrated in Sec. 4. Evaluations are presented in Sec. 5. Finally, discussion and conclusions are presented.

## 2. Corpus Description

In order to consider whispered speech processing, two corpora were developed with different focuses. Corpus UT-VocalEffort(UT-VE) I consists of speech under five vocal efforts: whispered, soft, neutral, loud and shouted, while corpus UT-VocalEffort(UT-VE) II focuses on neutral speech embedded with whispered speech “islands”. Both corpora were collected in an ASHA certified, single walled sound booth using a multi-track FOSTEX 8-channel synchronized digital recorder with gain adjustments for individual channels.

### 2.1. UT-VocalEffort I

For UT-VE I, a total of 12 male, native English-spoken subjects participated in the data collection. For each subject, the speech was recorded for a range of 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 the sound pressure level(SPL) of the test tone measured  $75dB$  using a QUEST sound level meter(SLM). 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  $75cm$ (also shown in Fig. 1(a)). The data collection procedure was divided into 3 phases for each subject. Phase I consists of 2 sessions which has 5 tokens corresponding to 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 a neutral mode. Phase III includes spontaneous speech of one-minute duration in each vocal mode.

### 2.2. UT-VocalEffort II

In addition to UT-VE I, a much larger corpus named UT-VE II was developed in the same environment as in UT-VE I. Whisper and neutral speech from 37 male and 75 female subjects were collected. Unlike corpus UT-VE I that focuses on five vocal efforts, corpus UT-VE II focuses on neutral speech embedded with whispered speech. The corpus consists of a spontaneous part and reading part for each subject. In the spontaneous part, the collection environment was assumed to be a cyber cafe scenario. Three subjects were positioned as shown in Fig.1(b). Subject 1 and 2 engage in conversation (seated across from each other, laptop in front of subject 1). Before recording, a list of information, such as names, addresses, phone numbers or credit card numbers, was given to subject 1. Next, some key information was randomly chosen for whisper from the list by subject 1. Furthermore, subject 1 was told that subject 3 is trying to pick up as much key information as possible, thus subject 1 was persuaded to produce the speech as low as he/she can when the key information was mentioned in conversation. In so doing, when subject 1 introduces information from the list to subject 2, subject 1 was lead to produce whispered speech for key information in neutrally phoned conversation. In the reading part of UT-VE II, only subject 1 was enrolled and required to read materials in either neutral or whispered modes. Three types of materials were used in the reading part. The first type of material are sen-

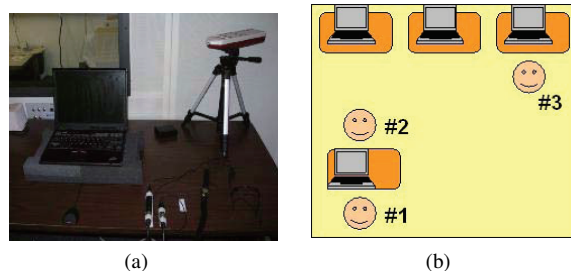


Figure 1: (a): Table setting of UT-VE-I; (b): Setting of Data Collection for UT-VE-II

tences selected from the TIMIT database. 41 TIMIT sentences were produced alternatively in neutral and whisper mode, with the 14<sup>th</sup> and 15<sup>th</sup> sentences both read in neutral mode. The second type of material are two paragraphs from a newspaper. For each paragraph, four whisper-islands were produced, with each island consisting of 1-2 sentences. The third type of material is the same paragraph as those of the second type. However, for each paragraph, five phrases were read in whisper mode, with each phrase 2-3 words long. In this study, the speech data produced with close-talking microphone in UT-VE I&II were used for analysis and tests.

## 3. Baseline Algorithm

A baseline routine for whisper-island detection consists of two main steps: segmentation and classification. The structure of the routine can be illustrated in Fig. 2. The VECPs of

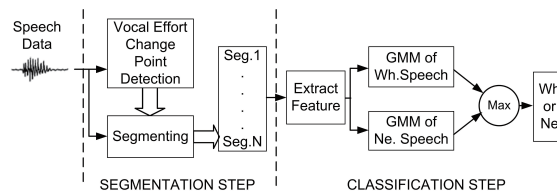


Figure 2: Flow Diagram of Whisper-Island Detection

the input speech data with whisper-islands are detected in the segmentation step (left part of Fig. 2). Based on the detected VECPs, the speech data is divided into segments. In this study,  $T^2$ -BIC algorithm, described in [6][13] and [8], is deployed here to detect the VECPs between whisper and neutral speech.  $T^2$ -BIC detects acoustic change points based on the input feature data. In [9], the Bayesian Information Criterion (BIC) has been shown to be an effective metric in segmentation for segments greater than 5 seconds in duration. However, BIC needs secondary statistics (i.e., the covariance), which is a problem due to estimation error with insufficient data for segments less than 5 seconds. In [8], the  $T^2$ -statistic was incorporated to detect break points of speaker change. For segments even less than 2 seconds, the covariance is assumed to be an identity matrix to calculate  $T^2$ -mean distance. In [10], a  $T^2$ -BIC algorithm which combines  $T^2$ -statistics and BIC was proposed. It is shown in [10], [11] that,  $T^2$ -BIC segmentation works well for both segments that are less than 5 seconds, and segments greater than 5 seconds. In our study, the segment length is between 1-3 seconds, thus the  $T^2$ -BIC algorithm could be an effective method to detect the VECPs between whisper and neutral speech if a proper feature is deployed.

In the classification step, a modified GMM-based vocal ef-

fort classifier was developed to label the vocal effort of each speech segment obtained from the previous step. GMMs of whisper and neutral speech are respectively trained with all whisper and neutral speech data from UT-VE I. The scores obtained by comparing the detected segment with two vocal effort models were sorted, and the model with the highest score is identified as the model which best fits the vocal effort of segment.

## 4. Algorithm Advancement

### 4.1. Feature Advancements

In [13], motivated by an entropy-based feature proposed for endpoint detection for speech recognition in noisy environments [4], a 9-D entropy-based feature was proposed for whisper speech. For each frame, the spectrum obtained from an FFT can be viewed as a vector of coefficients in an orthonormal basis. Hence, the probability density function (PDF) can be estimated by a normalization over all frequency components. The spectral information entropy (SIE) can be obtained from this estimated PDF. In [13], the SIE calculation was considered over the spectrum from 300 to 3000 Hz. Also, each frame was evenly divided into 4 sub-frames across the time domain. The SIE of each sub-frame is calculated to form first a 4-Dimension SIE feature for each frame. The 5th-8th dimensions of the feature are the SIE of the 4 sub-bands evenly divided over the frequency range 300-3000Hz, respectively. For each sub-band, the spectral information entropy can be obtained as follows. Assuming  $X(k)$  is the power spectrum of speech frame  $x(n)$ ,  $k$  varies from  $k_1$  to  $k_M$  in a sub-band; then that portion of the frequency content in the  $k$  band versus the entire response is written as,

$$p(k) = \frac{|X(k)|^2}{\sum_{j=k_1}^{k_M} |X(j)|^2}, \quad k = k_1, \dots, k_M. \quad (1)$$

Since  $\sum_{k=k_1}^{k_M} p(k) = 1$ ,  $p(k)$  can be viewed as an estimated probability. Next, the spectral information entropy (SIE) for the sub-band can then be calculated as,

$$H = - \sum_{k=k_1}^{k_M} p(k) \cdot \log p(k). \quad (2)$$

Based on the power spectrum of each frame, the above calculation is performed for each of 4 sub-bands, so that the 4-D SIE over the frequency domain is obtained. Next, the energy ratio described in [7] is modified to be the ratio between SIE of the high band (2800-3000Hz) versus the low band (450-650Hz) for the final dimension of the 9-D feature vector.

In this study, the SIE calculation for the 4 sub-frames of each frame was performed over 300-8000Hz instead of 300-3000Hz. The power spectrum of each frame was evenly divided into two sub-bands over 300-8000Hz. Thus, the 9-D feature becomes into a 7-D feature including 1-D of SIE ratio between the high band (2800-3000Hz) versus the low band (450-650Hz), 2-D of SIE for 2 evenly-divided sub-band over 300-8000Hz and 4-D of SIE over 300-8000Hz for 4 sub-frames of each frame, which were respectively illustrated by Fig. 3, Fig. 4 and Fig. 5.

With simplified feature calculation and smaller dimension, the improved 7-D feature shows more sensitivity to vocal effort change between whisper and neutral speech. Performance of

the improved feature in detecting VECPs between whispered and neutral speech will be illustrated by experimental results later.

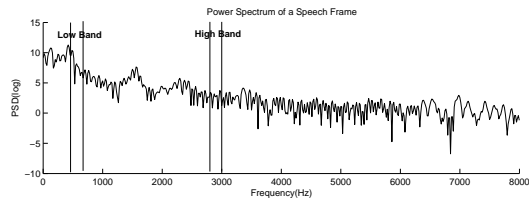


Figure 3: Entropy Ratio between High and Low Band Frequencies

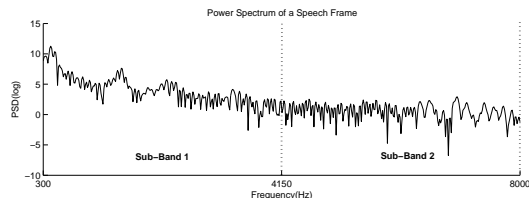


Figure 4: Two Bands over Frequency Domain

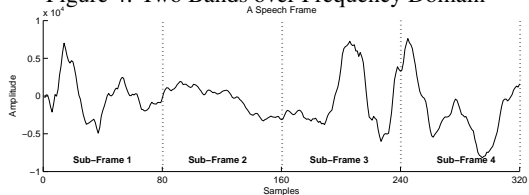


Figure 5: Four Sub-Frames over Time Domain

### 4.2. $T^2$ -BIC Algorithm Improvements

In [10] and [8], the  $T^2$ -BIC algorithm calculated  $T^2$  statistics to detected possible acoustic change points, and verified actual acoustic change points by obtaining BIC scores on possible VECPs. In this study, the  $T^2$  statistics was used to detect the possible VECPs as well. Next, BIC algorithm was used to scan the 100 frames range which centered around the possible VECP in order to refine the position of the detected VECP or extend the scan range for the highest BIC score. The modified  $T^2$ -BIC algorithm can obtain not only more accurate VECPs positions, but also at a much lower miss detection rate. The performance of VECP detection can be measured by the MES score introduced in Sec. 5.1.

### 4.3. Classifier Improvements

In [3][13], the GMM based vocal effort classifier was trained and tested with MFCC features. However in [6], MFCC was shown to be a feature which is not sensitive to vocal effort changes. In the discussion of [13], a proposed 9-D feature was planning to be used for training and testing in vocal effort classification. In this study, experimental results of VECP detection, which will be shown in Sec. 5, indicates the improved 7-D feature is more suitable to detect VECPs between whisper and neutral speech. Motivated by this fact, the GMM based vocal effort classifier was trained and tested with the improved 7-D feature described in Sec. 4.1. The advances of classifier trained with improved 7-D feature will be illustrated by experimental results shown in next section.

## 5. Evaluation Results

### 5.1. Short Introduction of Multi-Error Score

In [6] and [13], the Multi-Error Score (MES) was developed and introduced to evaluate the performance of acoustic features in

detecting VECPs. The MES consists of 3 types of error of segmentation mismatch: miss detection rate, false alarm rate and average mismatch in milliseconds normalized by dual-segment duration. Fig. 6 illustrates these three types of error .

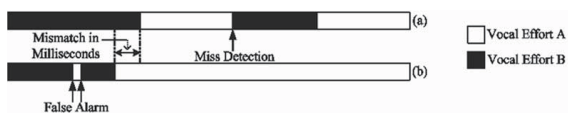


Figure 6: Three Types of Segmentation Error

The calculation of MES can be illustrated by the following equation:

$$MES = False\ Alarm\ Rate + 2 \times Mismatch\ Rate + 3 \times Miss\ Detection\ Rate \quad (3)$$

The mismatch rate is obtained by calculating the percentage of the mismatch in milliseconds of the total duration of two segments corresponding to the actual break points. More details of MES can be found in [6] and [13]. Miss detection rate and mismatch rate are more costly errors for whisper island detection, so these errors are scaled by 3 and 2 respectively.

## 5.2. Experimental Results in MES

An audio stream with 41 sentences produced alternatively in neutral and whispered mode by each subject from UT-VE II were manually labeled for VECPs in transcript files. As described in Sec. 3, the 14<sup>th</sup> and 15<sup>th</sup> sentences in the audio stream were both produced in neutral mode. The audio files from 60 subjects were enrolled in experiments. The transcript files of these audio streams were used to compare with VECF detection results obtained from  $T^2$ -BIC algorithm using different features, so that the MES can be calculated. The lower MES denotes better performance in VECF detection. An MES score of 0 occurs if all 3 error rates are 0%, increases to 60 if all error rates are 10%(acceptable segmentation) and to 120 if all error rates climb to 20%(unacceptable). The experimental results in MES are shown in Table 1 for 9-D feature from [13], and 7-D feature with original  $T^2$ -BIC, and improved  $T^2$ -BIC in this study. The reduction in MES from 80 to 19 is quite remarkable, noting that even a slight increase in FAR can be addressed in a post-processing stage.

Table 1: Evaluation for Vocal Effort Change Points Detection

Feature Scenario	MDR(%)	FAR(%)	MMR(%)	MES
9-D& $T^2$ -BIC	21.56	6.89	4.30	80.18
7-D& $T^2$ -BIC	10.58	5.25	3.04	43.07
9-D&New $T^2$ -BIC	5.29	11.42	3.48	34.28
7-D&New $T^2$ -BIC	2.17	7.58	2.48	<b>19.08</b>

## 5.3. Experimental Results of System

With lower MES in the VECFs detection experiments, the 7-D feature shows the best performance in sensing vocal effort changes between whisper and neutral speech versus the 9-D feature in [13]. The overall performance of the improved system was experimented and compared with the performance of a baseline system using the previous 9-D feature in segmentation and MFCC for classification. The same audio streams used in last subsection were employed here. Thus, with 20 whisper-islands for each audio stream, there are 1201 whisper-islands in total for detection. The GMM based vocal effort classifier was trained with whisper and neutral speech from 12 male subjects from UT-VE I using MFCC features, while for the improved

system the classifier was trained with the same audio streams using the proposed 7-D feature. Experimental results of both systems are illustrated in Table 2. There is significant advancement in whisper-island detection performance.

Table 2: Evaluation for Overall Whisper Island Detection

System	Detection Accuracy(%)
Baseline System	62.28
Improved System	<b>95.33</b>

## 6. Discussion and Conclusion

Whisper island detection is a challenging research problem which has received little attention in the research community. There are profound differences in speech production under whisper(e.g. no voicing) for all speech which renders speech system technology ineffective(ASR, speaker ID, coding, etc). Effective whisper island detection is the first step needed to ensure the engagement of effective subsequent speech processing steps to address whisper. Here an improved 7-D feature was developed, along with novel advancements for  $T^2$ -BIC segmentation that have resulted in the best MES score to date on VECF detection. To date, this represents the best solution available for whisper island detection. Finally, GMMs trained with the same 7-D feature set also showed outstanding(95.3% vs. 62.3%) performance for whisper detection. This advancement is an important step towards addressing mixed vocal effort(neutral/whisper) in conversational speech for applications such as speech recognition or speaker ID in telecommunication.

## 7. References

- [1] Ito T, Takeda K, Itakura F., "Analysis and recognition of whispered speech", *Speech Commun.*, 2005; 45:139-152.
- [2] Kallail, K., "An Acoustic Comparison of Isolated Whispered and Phonated Vowel Samples Produced by Adult Male Subjects", *Journal of Phonetics*, 12, pp. 175-186, 1984.
- [3] Zhang, C, Hansen, J.H.L., "Analysis and Classification of Speech Mode: Whispered through Shouted", *INTERSPEECH 2007-ICSLP*, pp.2289-2292, 2007.
- [4] J. Shen, J. Huang, L. Lee, "Robust Entropy-based endpoint detection for speech recognition in noisy environments", *ICSLP*, 1998, pp.232-235, 1998.
- [5] Jovicic, S., Saric, Z., "Acoustic Analysis of Consonants in Whispered Speech", *Journal of Voice*, Vol. 22, No. 3, pp.263-274.
- [6] C. Zhang, J.H.L. Hansen, "Effective Segmentation based on Vocal Effort Change Point Detection", *ITRW*, Aalborg, 2008 June.
- [7] Wenndt, S.J., Cupples, J., Floyd, M., "A Study on the Classification of Whispered and Normal Phonated Speech", *INTER-SPEECH2002*, pp649-652, 2002.
- [8] Huang, R., Hansen, J.H.L., "Advances in Unsupervised Audio Segmentation for the Broadcast News and NGSW Corpora", *ICASSP 2004*, pp.741-744, 2004.
- [9] Johnson, S., "Speaker Tracking", Master Thesis, Engineering Department, Cambridge University, UK, 1997.
- [10] Zhou, B., Hansen, J.H.L., "Efficient Audio Stream Segmentation via the Combined T2 Statistic and Bayesian Information Criterion", *IEEE Trans. Speech and Audio Processing*, Vol. 13, No.4, July 2005.
- [11] Huang, R., 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.
- [12] Ceballos, L., "Analysis and Modeling of Speech for Laryngeal Pathology Assessment", Ph. D. Thesis, Department of Biomedical Engineering, Duke University, Durham, North Carolina, 1995.
- [13] Zhang, C., Hansen, J.H.L., "An Entropy based Feature for Whisper-Island Detection within Audio Streams", *INTER-SPEECH 2008-ICSLP*, 2008.