

# An Entropy based Feature for Whisper-Island Detection within Audio Streams<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, especially whispered speech, has strong negative impact on speech system performance. It is therefore necessary to detect whisper-islands embedded within neutral speech prior to subsequent processing steps. Detecting whisper-islands in speech audio streams can contribute to improved modeling, speech analysis, and understanding. Speech technology can also benefit by allowing for suppression/obscuring of sensitive data (names, credit card numbers, etc.) in audio archives, call centers, or for spoken document retrieval systems. This study focuses on detecting whisper-island from neutral speech within audio streams using a proposed new entropy-based feature. The new feature focused on effectively detecting vocal effort change points between whisper and neutral speech. Experimental results employing a multi-error score show that the new feature has superior performance over a previous method introduced in [2]. Overall, the detection accuracy of 97% (for male) and 96.7% (for female) indicate effective performance in whisper-island detection, and suggests a viable algorithm to assist speech and language technology when whisper is present.

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

## 1. Introduction

Current speech processing systems are generally designed for normally phonated speech data. However, speech signals can be generally classified into five categories based on the vocal efforts differences: whispered speech, soft speech, neutral speech, loud speech and shouted speech. In [1], test results for a close-set Speaker-ID system showed that speech with vocal effort other than neutral mode results in a significant reduction in speech system performance. From the experiments in [1], we observe that whispered speech has the most dramatic loss for speech processing systems. This is mainly because the fundamental difference in speech production of whispered speech: the absence of all periodic/harmonic excitation, so all speech is unvoiced. Therefore, detecting and identifying whispered islands embedded in the speech signal before further processing is useful to eliminate the negative impact of whispered speech on subsequent speech systems (ASR, Speaker ID, etc.). However, whispered speech has a high probability of conveying confidential or sensitive information [2]. For a spoken document retrieval system or a call center monitoring system, detection and identification of whispered islands in

speech files can help in the retrieval of desired confidential or sensitive information.

Several studies have investigated methods to identify whispered speech or segment non-neutral speech. In [3], 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 4.8s analysis frame length prevents it from being applied to detect the precise boundaries between whispered and neutral speech. In [2], an effective method of detecting vocal effort change points 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 this study, we formulate an algorithm which can both locate and identify whispered speech embedded within a neutral audio speech stream using a new entropy-based feature. The results from experiments show that the new feature has improved performance in detecting the vocal effort change points between whisper and neutral speech. Experiments are also performed on utterances to evaluate the algorithm's performance in whisper-island detection.

The remainder of this paper is organized as follows: Sec.2 introduces the new entropy-based feature used in detecting vocal effort change points between whisper and neutral speech; Sec.3 introduces the audio corpus used for algorithm development and evaluation; Sec.4 presents the whisper-island detection algorithm, followed by segmentation experiments using the new feature, and experiments of whisper-island detection. Finally, we close in Sec.6 with some concluding remarks and possible future studies (Sec. 7).

## 2. A New Entropy-Based Feature

In [2], five acoustic features were evaluated for vocal effort change point (VECP) detection. Measured by the proposed multi-error score [2], the ratio, represented as ER, between the energy in the high band (2800-3000 Hz) versus low band (450-650 Hz) showed the best performance. Furthermore, motivated by the entropy-based feature in endpoint detection for speech recognition in noisy environment proposed in [9], a new 9-D entropy-based feature is developed here. The new proposed 9-D feature consists of first 4-D represented by the 4-D spectral information entropy (SIE) as calculated in [2]. 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

---

1. This project was funded by AFRL under a subcontract to RADC Inc. under FA8750-05-C-0029, and by Univ. of Texas at Dallas under project EMMITT

frame is evenly divided into 4 sub-frames. The SIE of each sub-frame is calculated to form a 4-Dimension SIE feature for each frame. The 5<sup>th</sup>-8<sup>th</sup> dimensions of the feature are the spectral information entropy of the 4 sub-bands evenly divided over the frequency range 300-3000 Hz, 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  $k$  band versus the entire response is written as,

$$p(k) = \frac{|X(k)|^2}{\sum_{j=k_1}^{k_M} |X(j)|^2}, \quad k = k_1, k_2, \dots, k_M \quad (1)$$

Since  $\sum_{k=k_1}^{k_M} p(k) = 1$ ,  $p(k)$  has the property of probability.

The spectral information entropy for the sub-band can then be calculated as,

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

Using 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. Here, the energy ratio described in [3] is modified to be the ratio between spectral information entropy of the high band (2800-3000Hz) versus the low band (450-650Hz) for the final dimension of the new 9-D feature vector. The performance of new feature in detecting vocal effort change point (VECP) between whisper and neutral speech will be illustrated in experimental results in Sec 5.1.

### 3. Corpus Description

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. The corpus named UT-VocalEffort is primary corpus used in experiments for this study. In Vocal Effort, a total of 12 male, native English-speaking subjects participated in data collection. For each subject, the speech was recorded in a series 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 a calibration test tone in all recordings to obtain a ground truth baseline for vocal effort. 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, location of the calibration test tone speaker, and 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 parts corresponding to the five speech modes. In each part, 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 the spontaneous speech of one-minute duration in each vocal mode. Vocal models include: whispered, soft, neutral, loud, and shouted.

In addition to data from UT-VocalEffort, the whispered and neutral speech recorded in the same environment as in UT-VocalEffort, from 10 male and 9 females to be used as test utterances in whisper-islands detection performance evaluation. Each test utterance contains 21 TIMIT sentences 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 speech data produced with close-talking microphone were used in this study.

## 4. System Description

The whisper-island detection system developed consists of two main steps: segmentation and classification. The structure of the system is illustrated in Figure.

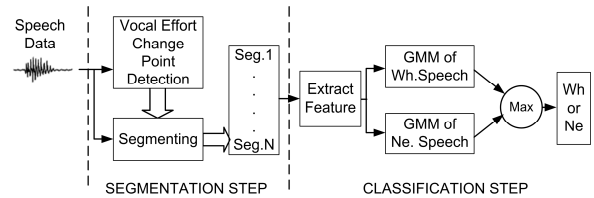


Figure 1: Block diagram of whisper-island detection system.

The VECPs of the input speech data with whisper-islands are detected in the segmentation step (left part of Fig. 1). Based on the detected vocal effort change points, the speech data is divided into segments. In this study,  $T^2$ -BIC algorithm, described in [5] and [2], is deployed here to detect the VECPs between whisper and neutral speech using the new feature introduced in Sec.2.  $T^2$ -BIC algorithm detects acoustic change points based on the input feature data. In [6, 7], 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 [5], 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 [4], a  $T^2$ -BIC algorithm which combined  $T^2$ -statistics and BIC was proposed. It is shown in [4, 8] that,  $T^2$ -BIC segmentation works well for both segments that are less than 5 seconds, and segments greater than 5seconds. In our study, the segment length is between 1-3seconds, thus the  $T^2$ -BIC algorithm could be an effective method to detect the VECPs between whisper and neutral speech if an effective feature for whisper is employed.

In the classification step, a modified GMM-based vocal effort classifier was developed to label the vocal effort of each speech segments from the previous step. GMMs of whispered and neutral speech are respectively trained with all whispered and neutral speech data from UT-VocalEffort. 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 segment vocal effort.

## 5. Experiments and Results

Several experiments were conducted to evaluate the capability of the new feature in detecting VECP between whisper and neutral speech and to test the performance of the proposed system for whisper-island detection. The evaluation of the new feature is presented first.

### 5.1. New Feature Evaluation

#### 5.1.1. Measurement Tool: MES

In [2], a new measurement multi-error score (MES) was proposed to evaluate the acoustic features in detecting VECP between non-neutral and neutral speech. In this study, MES results for the new feature are compared with the best results from [2] in detecting VECP between whisper and neutral speech. For the segmentation case, mismatch can be described by three error scores: miss detection rate, false alarm rate, and average mismatch in milliseconds. Here, the false alarm rate denotes the percentage of false alarms versus all detection results, while the miss detection rate is the percentage of VECPs not detected over the total VECPs in all test sentences. For the vocal effort segmentation case, the false alarm error 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, false alarm error is less important than miss detection error in the evaluation of segmentation. The average mismatch in milliseconds between detected and actual VECPs is an important norm which reflects break point accuracy for the feature and data. As shown in [2], average mismatch in milliseconds was modified as the average mismatch rate by averaging the percentage of the mismatch of total duration of two segments corresponding to the actual VECPs to fuse with false alarm rate and miss detection rate in percentage. MES can be expressed by the following equation [2]:

$$MES = \text{False Alarm Rate} + \text{Miss Detection Rate} * 3 + \text{Mismatch Rate} * 2 \quad (3)$$

An ideal segmentation has zero false alarms, zero miss detections, 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 denotes a fair performance bound in segmentation. If the 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 to be poor/unaccepted.

#### 5.1.2. Experimental Results in MES

The speech data in whisper and neutral vocal efforts from UT-VocalEffort Phase I was used to construct the testing utterances. As described in [1, 2], the speech data were calibrated using a 75 dB-SPL test tone. A frame-energy based approach was used to remove silence at the lead and tail parts of each sentence. In **SAME** speaker scenario, each whispered sentence was inserted between two neutral sentences (from the same speaker) to form a concatenated sentence with two VECPs (Ne+Wh+Ne, where Wh, Ne denote vocal effort of whispered and neutral respectively). However, in the **DIFFERENT** speaker scenario, the two neutral sentences'

speaker is different from the inserted speaker of the whispered sentence (Ne1+Wh2+Ne1). The segmentation results for **SAME** speaker scenario and **DIFFERENT** speaker scenario are compared with results using the energy ratio (ER) in Table 1 and Table 2 respectively:

Table 1: Segmentation results of  $T^2$ -BIC algorithm based on the new entropy-based feature and energy ratio for VECPs between whispered and neutral speech in **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 | MDR  | FAR   | MMR  | MES   |
|--------------|------|-------|------|-------|
| Energy Ratio | 6.67 | 10.30 | 8.81 | 47.92 |
| New Feature  | 2.50 | 7.06  | 7.20 | 28.96 |

Table 2: Segmentation results of  $T^2$ -BIC algorithm based on the new entropy-based feature and energy ratio for VECPs between whispered and neutral speech in **DIFFERENT** speaker scenario.

| Feature Type | MDR  | FAR  | MMR  | MES   |
|--------------|------|------|------|-------|
| Energy Ratio | 6.67 | 6.15 | 8.17 | 42.50 |
| New Feature  | 3.75 | 8.55 | 7.45 | 34.71 |

Based on the results in Table 1&2, our new feature has 18.96 and 7.8 MES level improvement versus ER for the **SAME** speaker and **DIFFERENT** speaker scenarios, respectively. The most attractive property of the new feature is the low miss detection rate, which is more important in VECP detection between whisper and neutral speech. In [2], ER was evaluated as the feature with the lowest average MES in detecting VECP of non-neutral and neutral speech out of five acoustic features. Further experiments are carried out to evaluate the new feature for detecting the VECPs between non-neutral speech and neutral speech. The MES results are shown in Table 3.

Table 3: Average MES results of  $T^2$ -BIC algorithm based on the new entropy-based feature and energy ratio for VECPs between non-neutral speech and neutral speech.

| Scenario \ Feature Type | Energy Ratio | New Feature |
|-------------------------|--------------|-------------|
| SAME Speaker            | 56.49        | 40.43       |
| DIFF. Speaker           | 54.45        | 39.36       |

In Table 3, the MES improvements for the new feature in both scenarios indicate the new feature is not only effective in detecting VECPs between whisper and neutral speech, but it also works well for VECPs in detection between non-neutral and neutral speech within audio streams.

### 5.2. System Evaluation

The test utterances from 10 male and 9 female speakers are used in experiments to evaluate the whisper-island detection system. Since each test utterance has 21 TIMIT sentences read in neutral and whisper alternatively, with the 13<sup>th</sup> and 14<sup>th</sup> sentence blocks neutral, there are ten regions of interest for each test sequence. In total we have 100 whisper-islands to detect for male speech and 90 whisper-islands to detect for female. The GMM-based classifier was trained in two ways for different experimental scenarios. In the first scenario, the whisper and neutral speech from 12 male speakers in UT-VocalEffort Phase I were used to train the GMMs of vocal effort for whisper and neutral respectively. To compensate for

the false alarm detections in the segmentation step, successive segments identified as the same vocal effort are merged to form one segment of constant vocal effort. The experimental results are shown in Table 4.

Table 4: *Whisper-island detection results (Scenario 1).*

| Data Type | Detection Accuracy |
|-----------|--------------------|
| Male      | 97.0% (97/100)     |
| Female    | 93.3% (84/90)      |

Here, the detection accuracy means the percentage of detected whisper-islands out of the total actual whisper-islands.

In the second scenario, the classifier was trained using a GMM-UBM manner [10] with the universal background model (UBM) constructed from a separate set of male speakers from the UT-SCOPE corpus [11]. A vocal effort specific MAP adapted Gaussian Mixture Model (GMM) is obtained from the UBM for each of the two trained vocal efforts. The detection results are shown in Table 5.

Table 5: *Whisper-island detection results (Scenario 2).*

| Data Type | Detection Accuracy |
|-----------|--------------------|
| Male      | 97.0% (97/100)     |
| Female    | 96.7% (87/90)      |

Despite the fact that all training data in both scenarios are male speakers, the results for female speech shown in Table 4&5 illustrate little gender-dependent differences in the whisper model, especially in Scenario 2. The improvement in detection accuracy for the GMM-UBM scenario for female data may indicate that better whisper vocal effort modeling of the UBM adapted model. In addition, there is no false alarm of whisper detection in all experiments from this study. This fact indicates that the algorithm may miss some whisper-islands, but does not mistake neutral speech as whisper.

## 6. Conclusion and Discussion

In this study, an algorithm was developed which detects whisper-islands from a neutral speech audio stream. The algorithm consists of a segmentation step and classification step. A proposed new feature was deployed within a feature-based  $T^2$ -BIC algorithm to detect the vocal effort change points between whisper and neutral speech. The obvious improvements in multi-error score (MES) in the experiments indicate the new feature is better than a previous energy ratio for VECP detection between non-neutral and neutral speech. The overall performance of whisper-island detection algorithm was evaluated, with the best detection performance of 97% (for male) and 96.7% (for female) obtained with a vocal effort classifier trained with a UBM. During all experiments, although several whisper-islands were not detected, no neutral speech segments were ever falsely labeled as whisper.

Although the classifier was trained with speech data from only male speakers, the similar detection accuracy between genders in Table 5 indicates the obtained UBM adapted GMM has little influence based on gender differences in test and training data in modeling whispered vocal effort. Furthermore, although there is no change in detection accuracy for male data, the detection accuracy for female data increased from 93.3% to 96.7% between Table 4 & 5. This increase confirms that the UBM adapted GMM's are better in modeling vocal effort for whisper. Here, the fact that the UBM was trained from neutral speech from male speakers is

motivation for improved performance by training the GMM for vocal effort of whisper with a larger and gender balanced data base.

## 7. Future Work

The algorithm shows good performance in whisper-island detection. However, several directions are possible for improved performance.

First, as discussed in Sec. 6, a larger and gender-balanced corpus of whisper may be used to train a GMM which models the vocal effort of whisper for both male and female whispers. Alternatively, a gender identifier could work as a pre-processing step before vocal effort classification. The gender labeled segments would be classified by the GMMs trained by male whisper and female whisper.

Alternatively, we note that, the GMM-based classifier was trained using MFCC feature only in this study. The new feature which is sensitive to vocal effort changes was merely used to detect the VECPs. Intuitively, GMMs trained with the proposed new feature may model the whispered and neutral vocal effort more accurately. Thus, by using the proposed new feature in the classification step instead of MFCCs may also bring better performance in whisper-island detection. The improved algorithm then can be applied in whisper detection for call-center data or in document retrieval applications.

## 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] Zhang, C. and Hansen, J.H.L., "Effective Segmentation based on Vocal Effort Change Point Detection", ITRW, Aalborg, 2008 June.
- [3] Wenzdt, S.J., Cupples, J., and Floyd, M., "A Study on the Classification of Whispered and Normal Phonated Speech", INTERSPEECH2002, pp649-652, 2002.
- [4] 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.
- [5] 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.
- [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] Shen, J., Hung, J. and Lee, L., "Robust Entropy-based endpoint detection for speech recognition in noisy environments", ICSLP, 1998, pp.232-235, 1998.
- [10] Angkitittrakul, et. al, "Cluster-dependent Modeling and Confidence Measure Processing", ICSLP 2004
- [11] Varadarajan, V. S. and Hansen, J.H.L., "Analysis of Lombard Effect under Different Types and Levels of Noise with Application to In-set Speaker ID Systems", INTERSPEECH 2006-ICSLP, pp.937-940, 2006.
- [12] Reynolds, D.A., et. al., "Speaker Verification Using Adapted Gaussian Mixture Models", IEEE, Digital Signal Processing, pp.19-41, 2000.