Whisper-Island Detection Based on Unsupervised Segmentation With Entropy-Based Speech Feature Processing

Chi Zhang and John H. L. Hansen, *Fellow, IEEE*

***Abstract*—Whisper island detection is a challenging research problem which has received little attention in the research community. Effective whisper-island detection is the first step necessary to ensure engagement of effective subsequent speech processing steps to address mismatch between whisper and neutral speech production. In this paper, we propose an effective approach for detecting whisper-islands embedded within normally phonated speech via BIC/T^2-BIC using a proposed 4-D feature set. Performance is assessed using our proposed multi-error score (MES), which shows that the new proposed algorithm achieves the lowest MES (11.51) to date and along with a perfect 100% correct whisper/neutral vocal effort labeling. The results show that we can correctly and precisely detect vocal effort change points (VECP) between whisper-islands and neutral speech as well as label the vocal effort of the whisper-island. The proposed feature is sensitive to the vocal effort change between whisper and neutral speech and is gender independent. The result suggests that the proposed algorithm is effective and precise for the whisper-island detection.**

***Index Terms*—Bayesian information criterion (BIC), classification, detection, segmentation, T^2-BIC, vocal effort, whisper.**

I. Introduction

WHISPER speech is one mode of natural speech communication with results in reduced perceptibility and a significant reduction in intelligibility. In general, whispered speech can occur in a variety of settings with the physiological speech production change from neutral of a complete absence of vocal fold vibration [1]. On the other hand, in some voice pathology cases, whispered speech may be caused by a change in the vocal fold structure or physiology or muscle control due to disease of the vocal system, such as functional aphonia [2], laryngeal cancer [3], functional voice disorders [4], or alteration of the vocal folds as a result of medical operations [1]. Furthermore, as a paralinguistic phenomenon, whispered speech can be used in different circumstances. 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 a formal setting. On the other hand, one may be whispering to avoid being overheard, in which case some loss of intelligibility of context by a remote listener in the same speaker environment may be desirable [5]. However, current speech processing systems are generally designed for normally phonated speech data. In [6], using experimental results from a close-set Speaker-ID system evaluation, we observed that for five vocal efforts ranging from whispered to shouted, whispered speech possesses the most dramatic loss in performance for speech processing systems. This is mainly because of the fundamental difference in the speech production mechanism of whispered speech versus neutral speech [6], [7]. Therefore, whispered speech has attracted a series of studies not only for a more comprehensive understanding of the acoustic characteristics, but also for applications in speech and language technology. In the medical domain, knowledge of whispered speech is also necessary as a means for recovering laryngeal surgery patients [8], or in the evaluation of voice disorders such as for aphonic patients [9]. Since whispered speech can be effectively used for quiet and private communications, and cellphone use is significantly increasing, speech processing techniques that effectively address whispered speech are more relevant for emerging and robust speech communication systems.

In recent years, several research studies considered advancements in the field of whispered speech signal processing. The investigation of whispered speech is interesting from a theoretical point of view in speech production and perception, and for practical reasons in whispered speech recognition as well as whispered speaker recognition. In [5], the acoustic properties and a speech recognition method for whispered speech were considered. Another study considered the difference in isolated whispered and normally phonated vowels produced by male adults from an acoustical perspective [10]. Most studies consider whisper for English, but some have considered the analysis of consonants in whispered speech in Serbian [1], vowels in Swedish [11], and Mandarin Chinese [12]. Speaker gender identification from whispered speech has also been investigated in [13], [14], and [11]. In [5] and [6], speech recognition [5] and speaker ID [6], [15], [16] for whispered speech was studied. While many studies have considered the analysis of speech production under whisper and assessing whisper speech impact on speech technology, little if any systematic effort has been reported on how to effectively detect and locate whispered speech within input audio streams. This research is crucial, since for subjects with healthy vocal systems, it is expected that whisper mode will occur in mixed

Manuscript received November 01, 2009; revised July 12, 2010; accepted July 13, 2010. Date of publication August 19, 2010; date of current version March 30, 2011. This work was supported in part by the AFRL under a subcontract to RADC, Inc., under FA8750-09-C-0067 and in part by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. Hansen. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gary Geunbae Lee.

The authors are with the Center for Robust Speech Systems, Electrical Engineering Department, University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: john.hansen@utdallas.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2010.2066967

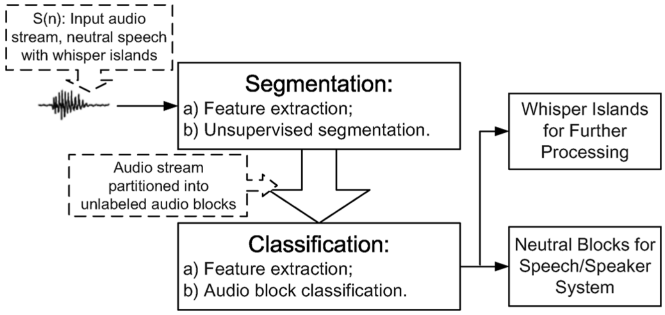
S(n): Input audio stream, neutral speech with whisper islands
Segmentation:
a) Feature extraction;
b) Unsupervised segmentation.
Audio stream partitioned into unlabeled audio blocks
Classification:
a) Feature extraction;
b) Audio block classification.
Whisper Islands for Further Processing
Neutral Blocks for Speech/Speaker System

Fig. 1. High level flow diagram of whisper-island detection.

neutral/whisper combinations depending on the information content within the audio stream (e.g., it is generally the case that normal speakers cannot sustain whisper mode for extensively long periods of time—10 mins-hrs). To utilize speech processing techniques that address whispered speech, the locations of whispered speech must first be identified within the audio stream. Furthermore, because of the high probability for whispered speech to convey confidential or sensitive information, the detection and identification of whispered speech in audio files can help spoken document retrieval systems or call center dialog systems (i.e., credit card numbers, emergency response systems, etc.). Several preliminary studies have investigated methods to identify whispered speech or segment non-neutral speech. In [17], 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-s analysis frame length prevents it from being applied to detect precise boundaries between whispered and neutral speech. In [18], Zhang and Hansen proposed an effective method of detecting vocal effort change points between non-neutral speech and neutral speech. However, vocal effort of the speech segment between two consecutive vocal effort change points cannot be assessed using that algorithm. Considering previous research, in this paper, we formulate an algorithm which can both locate and identify whispered speech islands embedded within a neutral audio stream using a new entropy-based feature. The new feature is integrated within a modified BIC/T^2-BIC unsupervised segmentation algorithm for detecting vocal effort change points between whispered and neutral speech (see Fig. 1 for the Segmentation phase). A new measurement strategy termed the multi-error score (MES) is proposed to evaluate performance of vocal effort change detection. In the final stage (see Fig. 1 for the Classification phase), a Gaussian mixture model (GMM)-based classifier trained with speech data using the new proposed feature is developed to address the problem of whisper-island detection.

To evaluate performance of the algorithm, two corpora which contain speech data under different vocal efforts, as well as normally phonated speech embedded with whispered speech islands are employed. All analysis and experiments presented in this study are performed based on data from these corpora.

The remainder of this paper is organized as follows. First, a short review of whispered speech is presented. Next, the corpora developed for this study is introduced in Section III. In Section IV, the baseline routine for whisper-island detection is presented. Next, details of the proposed new feature and algorithm are addressed in Section V. Evaluations using two whisper/Vocal Effort corpora are presented in Section VI. Finally, Section VII includes discussion and conclusions of this study.

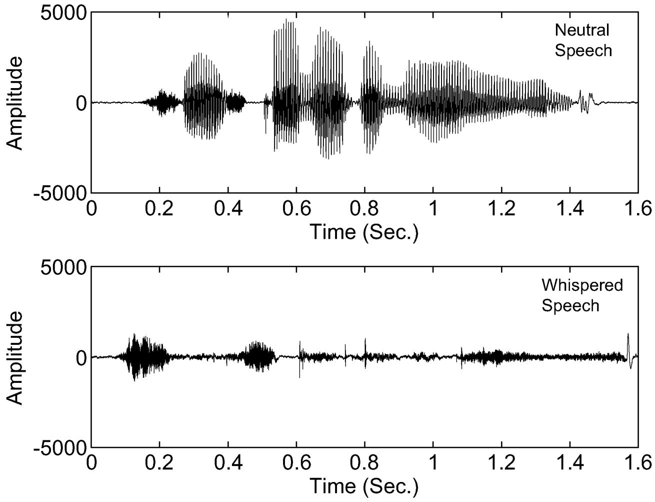

Fig. 2. Waveforms of neutral speech and whispered speech signals for the phrase "she is thinner than I am".

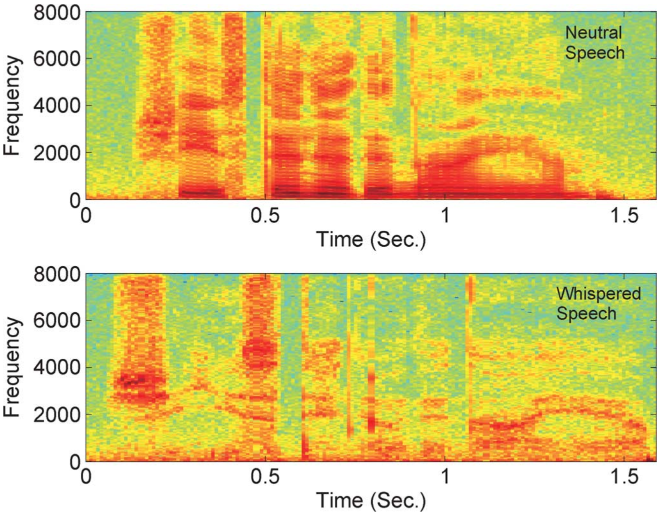

Fig. 3. Spectrograms of neutral speech signal and whispered speech signal.

II. Review of Whispered Speech

In Figs. 2 and 3, waveforms and spectrograms of the speech signal for the TIMIT sentence "She is thinner than I am." for both neutral and whispered speech are displayed. In addition to the fact that the intensity of whispered speech is significantly lower than that of neutral speech, the lack of voiced periodic components in whispered speech can also be observed. These observations indicate the fundamental and significant difference between whispered and neutral speech production. Without voicing, in theory it is not possible to distinguish between unvoiced and voiced stops and fricatives (e.g., /s/ versus /z/, /sh/ versus /zh/, /t/ versus /d/, /p/ versus /b/, etc.).

In fact, the main difference between whispered speech and neutral speech is the complete absence of vocal fold vibration for whispered speech. In neutral speech, voiced phonemes are produced by modulation of the flow air from the lungs through vibration of the vocal folds. However, for whispered speech, with exhalation of air used as the sound source, the shape of the pharynx is adjusted such that the vocal folds do not vibrate [5], [7], [19], [20]. Furthermore, recent studies report that a three-dimensional vocal tract shape measurement from magnetic resonance imaging (MRI) show a narrowing of the tract in the false vocal fold regions and weak acoustic coupling with the subglottal system [1], [21]. In [12], the laryngeal sphincter mechanism was found to be a principle contributing physiological maneuver in the production of whisper. Larynx rising is more evident in whispered tense vowels than lax vowels, and the tongue root is retracted in the tense vowels as well. In [22], the role of lip kinematics in the production of whispered plosives was considered, which support the theory that whispered speech and voiced speech rely on distinct motor control processes.

Due to these physiological differences in the production mechanism, the acoustic characteristics of whispered speech are different from those of neutral speech. In [9], [23], and [24], acoustic analysis of vowels in whispered and neutral speech showed that formant frequencies for vowels in whispered speech shift to higher frequencies compared to neutral speech. In [1] and [6], the fact that the duration of whispered speech is longer than that of neutral speech was also noted. It was shown in both [5] and [6] that the spectral tilt of whispered speech is much flatter than that of neutral speech. Earlier studies focused on a range of speech under stress speech styles, including angry, slow, fast, Lombard, as well as soft and loud. Furthermore, those research studies considered pitch structure, glottal spectral traits, duration, intensity, and formant structure over 200 features and revealed significant changes between soft, neutral, and loud production of speech [25], [26]. However, soft speech still has periodic excitation, and is different than whisper speech. Based on the fundamental acoustic differences between whispered and neutral speech, the detection of whispered speech within a neutral audio stream is much more difficult than simply detecting the speech audio which has lower overall energy. One key reason is that most technology for voice communications employ an automatic gain control (AGC) (e.g., all telephone, TV, radio transmission, etc.), and depending on the adaptation rate it may not be possible to use only gain to decide between neutral/whisper speech islands. Previous research in voice activity detection (VAD) based on frame energy illustrated the challenge in using a single feature to detect speech versus silence, let alone the challenge in separating neutral versus whisper speech segments. It is therefore necessary to develop an effective and robust algorithm to detect real whispered speech within an audio stream instead of simply relying on the low energy pieces caused by distance or gain changes while recording.

III. Corpus

To address the general problem of whispered speech detection as well as contribute to the field of whispered speech processing, two corpora are developed with different foci. The Corpus UT-VocalEffort (UT-VE) I consists of speech under five distinct vocal efforts: whispered, soft, neutral, loud, and shouted; while corpus UT-VocalEffort (UT-VE) II is focused on neutral speech streams embedded with key 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. Calibration test tones (1 kHz at 75 dB-SPL), are employed for all recordings to ensure ground-truth in absolute dB sound levels for all speech.

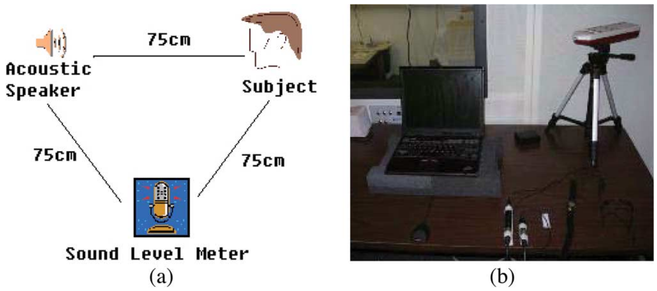

Fig. 4. (a) Table setting of data collection for UT-VE-I. (b) Setting of data collection for UT-VE-I in ASHA certified sound room.

A. UT-VocalEffort I

For UT-VE I, a total of 12 male, native English-speaking subjects participated in the data collection. All speakers were native American English speakers with no history of speech or hearing limitations/disorders. For each subject, speech was recorded for a series of tokens using three positioned microphones: a P-Microphone (physiological microphone) [27], a SHURE Beta-54 close-talking microphone and a SHURE MX391/S far field microphone. A 1-kHz sinusoid signal generated by an NTI analog audio generator was played through an ALTEC speaker at the same physical location as the calibration test tone and included in all recordings. At the beginning of each token, the volume of the test tone was carefully adjusted so the dBA sound pressure level (SPL) of the test tone measures 75 dB using a QUEST sound level meter (SLM). The test tone was recorded for 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 75 cm. The table recording setting is illustrated in Fig. 4(a), along with an image of the ASHA certified $13' \times 13'$ sound room in Fig. 4(b).

The data collection procedure was divided into three phases for each subject. Phase I consists of two sessions with five tokens corresponding to the five speech modes. In each token, five sentences from the TIMIT database were spoken in one of five speech modes and recorded. Each subject was prompted to read the particular sentences from a laptop display positioned in front of the subject. Phase II consists of 20 sentences which were all read sequentially in the neutral speech mode. Phase III includes spontaneous speech of one-minute duration in each of five vocal modes (e.g., whisper, soft, neutral, loud, and shouted). Human transcriber were used to verify speech and vocal effort content for all recordings for UT-VEI.

B. UT-VocalEffort II

In addition to the UT-VE I corpus, a much larger corpus named UT-VE II was constructed in the same acoustic environment as UT-VE I. Here, whispered and neutral speech from 37

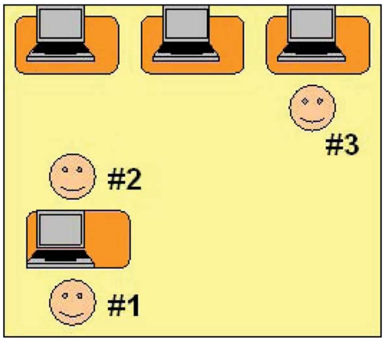

Fig. 5. Setting of data collection for UT-VE-II.

male and 75 female subjects were collected. Unlike the UT-VE I corpus which focused on five vocal efforts, corpus UT-VE II is focused on neutral speech embedded with whispered speech islands. The corpus consists of spontaneous natural exchanges with small blocks of whispered speech consisting of key information parts. For the spontaneous part, the collection environment was explained to the subject to be a cyber cafe scenario. Three subjects were positioned in the ASHA 13 ft × 13 ft sound room as shown in Fig. 5. Subject 1 and 2 engage in a conversation (seated across from each other, where a laptop is placed in front of Subject 1). Here, Subject 1 is the volunteer producing neutral/whispered speech, Subject 2 is the data collector and second party listener for Subject 1, and Subject 3 is a cyber cafe participant attempting to listen-in on the conversation between Subject 1 and 2 while using their computer. In order to achieve completely natural human-to-human conversation, the data collector (Subject 2) was instructed to keep their conversation engaged (e.g., between Subject 1 and 2). In order to satisfy IRB requirements, we did not want to record personal information regarding Subject 1 (e.g., their name, credit card or phone numbers, names of family/friends, etc.). To solve this challenge, random names were generated by selecting different first and last names from the Dallas telephone directory, along with company names and addresses pieced together from directory listings (e.g., "Acme Trucking" and "Dallas Furniture Mart" becomes "Acme Furniture" and "Dallas Trucking"). Each person or business name was printed on sheets of paper, with key parts for whisper production highlighted as sensitive information. The Subject 1 was instructed to be certain that when using the names/information in conversation, any highlighted parts needed to be kept confidential between Subject 1 and 2, so Subject 3 should not hear this information. The list of information, including names, addresses, phone numbers, or credit card numbers, was given to Subject 1. Key information was randomly chosen to be spoken in whisper mode 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, and thus Subject 1 was persuaded to produce the speech as low as he/she can but to convey the key information to Subject 2 in conversation. By doing this, when Subject 1 introduces the information from the list to Subject 2, Subject 1 would be able to work into the audio rather than be required to produce whispered speech for key information in the neutral phonated conversation. In the read part of UT-VE II, only Subject 1 was enrolled and required to read material in either neutral or whispered modes. Three types of read materials were used in the read part. The first type consists of sentences selected from the TIMIT database. Here, 41 TIMIT sentences were produced alternatively in neutral and whispered mode, with the 14th and 15th sentences both read in neutral mode. The second material type consists of two paragraphs selected from a local newspaper. For each paragraph, four whisper-islands were produced, with each island consisting of 1–2 sentences. The third type of material consists of the same paragraphs as those of the second type. However, for each paragraph, five phrases were read in whispered mode, with each phrase 2–3 words in duration.

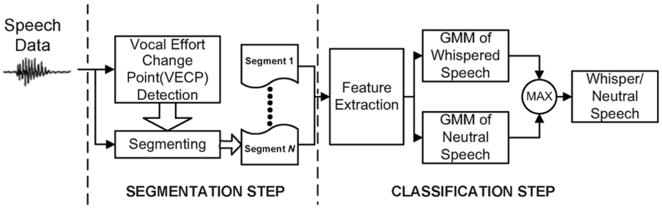

Fig. 6. Flow diagram of whisper-island detection.

In the present study, the speech data produced with the close-talk SHURE Beta-54 microphone in UT-VE I&II were used for analysis and experiments.

IV. Baseline Routine

The baseline routine for whisper-island detection consists of two main algorithmic steps: segmentation and classification. The structure of the routine is illustrated in Fig. 6.

The potential vocal effort change points (VECPs) of the input speech data embedded with whisper-islands are first detected in the segmentation step (left part of Fig. 6). Based on the sequence of potential detected VECPs, the speech stream is divided into segments. In this paper, an improved T^2-BIC algorithm is incorporated to detect the potential VECPs between whisper and neutral speech. The T^2-BIC algorithm, developed by Zhou and Hansen [28] and also described in [18], [29], [30], is an unsupervised model-free scheme that detects acoustic change points based on the input feature data. A range of potential input features for the T^2-BIC algorithm can be used to detect input acoustic changes within audio stream. For example, in [30], the speaker change points within the input audio streams were detected based on T^2-BIC algorithm using a range of features. In this paper, the T^2-BIC algorithm is considered as a potential effective method to detect the VECPs between whisper and neutral speech if an effective feature for vocal effort change is employed. Further details will be presented in Section V-B2.

In the classification step (right part of Fig. 6), a GMM-based vocal effort classifier is 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 whisper and neutral speech data. The scores obtained by comparing the detected segment with two vocal effort models are sorted, and the model with the highest score is identified as the model which best fits the vocal effort of the current segment.

V. Proposed Algorithm

A. New Feature Description

Since a feature-based algorithm is deployed in this study to detect the VECPs between whisper and neutral speech, a proper feature which is sufficiently sensitive to reflect the vocal effort change between whisper and neutral speech is necessary. Although several previous studies have considered the difference in acoustic properties for segment detection of neutral speech, such as formant frequency for vowels [9], [23], [24], duration [1], [6], and spectral tilt [5], [6], little if any effort has considered an effective feature for use in whisper detection, but simultaneously maintains performance for neutral speech. In [17], the energy ratio, which represents the ratio between the energy in a high-frequency band (2800–3000 Hz) to the energy in a low-frequency band (450–650 Hz) was deployed to classify whispered and normally phonated speech. Although the highest correct classification rate of the technique is 95% (57/60), the extensive 4.8-s analysis frame length prevents this technique from being applied for detection of the precise boundaries between whisper and neutral speech in this study. In [31], an entropy-based feature, which was originally developed in [32], [33] for speech boundary detection in realistic noisy environments, was proposed to fulfil the initial/final segmentation for Chinese whispered speech. A threshold for the average spectral information entropy must be properly chosen such that the initial and final points of the segment of whispered Chinese speech could be determined. Based on these previous studies, a new feature is proposed here based on the spectral information entropy (SIE), which can be effectively integrated within the T^2-BIC segmentation algorithm to detect VECPs between whisper and neutral speech.

For each 20-ms analysis frame from the input speech signal, the spectrum obtained from a fast Fourier transform (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 SIE can then be obtained from this estimated pdf. The SIE for a specific frequency band is obtained using the following procedure.

Step 1) Assuming $X(k)$ is the power spectrum of the input speech frame $x(n)$, where k varies from k_1 to k_M in a specific frequency band; then that portion of the frequency content in the kth band versus the entire response is written as

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

Step 2) Since $\sum_{k=k_1}^{k_M} p(k) = 1$, $p(k)$ can be viewed as an estimated probability that describes the energy distribution within this frequency band. Thus, the SIE for the frequency band $(\mathrm{k} = k_1, \ldots, k_M)$ can be calculated as

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

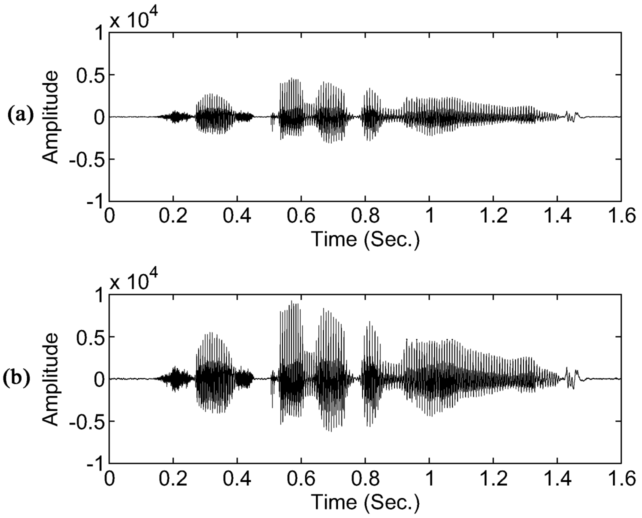

Fig. 7. (a) Waveforms of a speech signal (top). (b) Waveform of the same speech signal amplified by two (bottom).

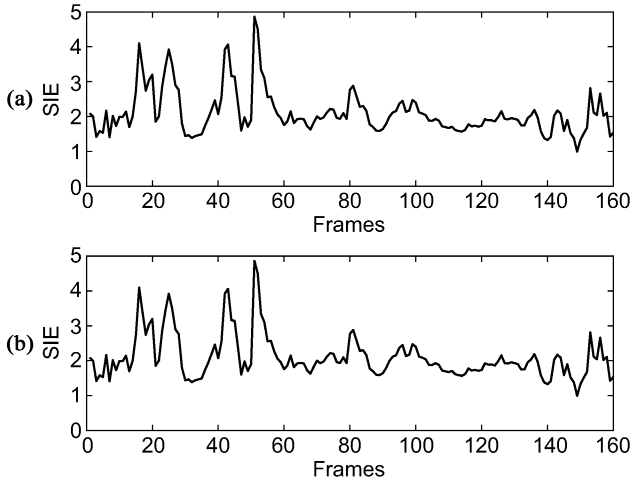

Fig. 8. SIE sequence of the speech signals shown in Fig. 7.

It should be noted that, the SIE is not influenced by the amplitude of actual speech signal waveform, so either an active/disabled automatic gain control (AGC) for any communication device will not impact segmentation performance. Fig. 7 shows a waveform of an original speech signal and the same speech signal which has been amplified by 2.0, respectively. Fig. 8 shows the corresponding spectral information entropy (SIE) sequence using 20-ms analysis frames for the speech signal shown in Fig. 7. It is obvious that the spectral information entropy response is unchanged although the original speech waveform has been amplified; thus, the effect of an automatic gain control system on whisper speech will not affect the value of the spectral information entropy. The spectral information entropy represents the distribution of energy over the frequency domain rather than the total amount of energy over the entire frequency domain. The proposed new feature consists of a 4-D parameter set of spectral information entropy obtained through distinct methods to form the feature vector.

In [17] and [18], the energy ratio between the energy in a high-frequency band (2800–3000 Hz) versus the energy in a low frequency band (450–650 Hz) was shown to statistically represent the acoustic difference between whisper and neutral speech.

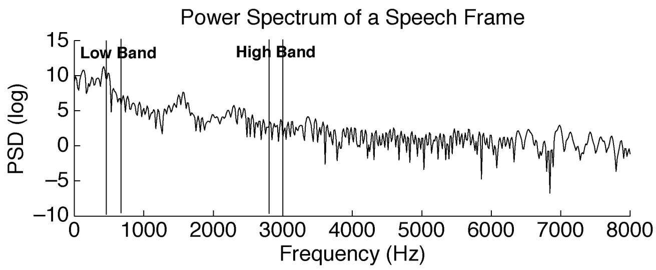

Fig. 9. Entropy ratio between high-frequency band and low-frequency band.

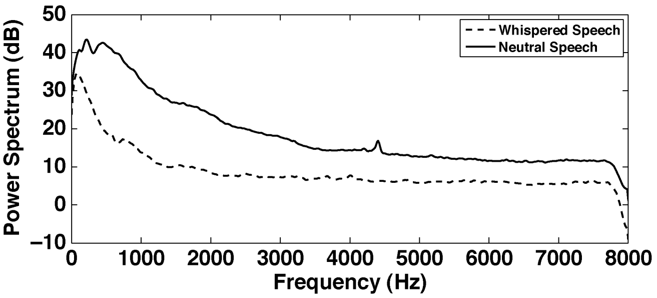

Fig. 10. Power spectrum amplitude for whispered speech and neutral speech.

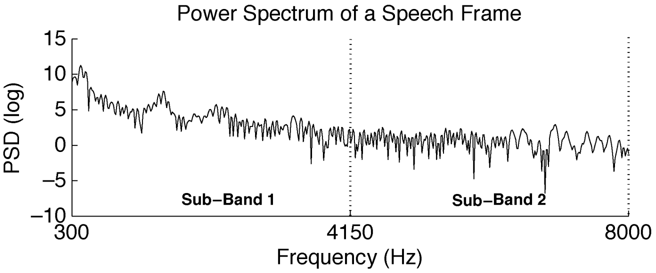

Fig. 11. Two sub-band over frequency domain for whisper analysis.

This fact reflects the observation that high (2800–3000 Hz) and low-frequency bands (450–650 Hz) can be viewed as significant bands in representing the difference between whisper and neutral speech. The difference of the energy distribution within these two bands may also be used to differentiate whisper from neutral speech. Using only energy, it is necessary to sample a larger amount of speech data (4.5 s) in order to reduce the variability across the phoneme space. Instead of roughly using the summation of energy within these two bands, the more detailed SIE within each of these two bands can be calculated to obtain the entropy band ratio. Thus, the entropy ratio between a high-frequency band (2800–3000 Hz) and low-frequency band (450–650 Hz) is used as the 1st-D parameter of the proposed segmentation feature, which is illustrated in Fig. 9.

Earlier studies [32], [33] showed that speech under stressed/speaking styles have spectral slopes for the glottal spectrum of −13.7, −12.1, −9.5 dB/Octave for soft, neutral, and loud speech [33]. In general, when producing speech under loud, neutral, and soft, the spectral slope starts off flatter (−9.5 dB/Octave) and moves to a steep slope (−13.7 dB/Octave) for soft because of the loss of high-frequency energy content. However, for whispered speech, studies have shown a much flatter spectral tilt than neutral speech does [5], [6], due primarily to the absence of voiced/periodic excitation. While the change in spectral slope for soft to loud and shouted is due primarily to the change in high-frequency energy content, under whisper speech condition, the spectral slope becomes flat because there is a significant loss in the low-frequency energy content as well due to the absence of voiced excitation. To illustrate this change, Fig. 10 shows an average power spectrum of whisper and neutral speech from 60 speakers from the UT-VE II corpus using 20-ms analysis frames, which have all been averaged to obtain the two responses. It is easily to see the separation of energy distributed in the high-frequency portion between whisper and neutral speech is smaller than the separation seen for low frequencies. The difference of the energy distribution in the low and high frequencies result in the much flatter spectral tilt for whisper versus that for neutral speech above 300 Hz. Thus, the SIE calculation is performed over the selected high-frequency band and low-frequency band. By careful selection, the frequency band from 300 Hz to 8000 Hz was evenly divided as low (300 Hz–4150 Hz) and high (4150 Hz–8000 Hz) frequency bands (see Fig. 11).

Furthermore, since the spectral tilt of whispered speech is statistically different from the spectral tilt of neutral speech [6], the spectral tilt can be used as a discriminative feature in differentiating the whispered speech and neutral speech. Finally, the 4-D Whisper Island Detection (WhID) feature set proposed in this study can be described as

$$WhID = \begin{bmatrix} 1-\text{D spectral information entropy ratio(ER)} \\ 2-\text{D spectral information entropy(SIE)} \\ 1-\text{D spectral tilt(ST).} \end{bmatrix} \tag{3}$$

B. Segmentation Algorithm

1) Measurement Tool: Multi-Error Score: The goal of reliable segmentation in audio streams requires that we assess the mismatch between hand/human segmentation and automatic segmentation. Traditional methods emphasis frame precision scores, but these do not take into account the desired continuity conditions needed for speech recognition. Toggling action between the two classes (e.g., a flip-flopping effect resulting in short duration blocks) at or near the true boundaries cause problems for the classification stage, and therefore it is more desirable to have longer contiguous blocks versus short blocks that flip-flop often between classes. Mismatch can be described by three error scores: miss detection rate (MDR), false alarm rate (FAR), and average mismatch in milliseconds (MMR: normalized by combined segment durations). Fig. 12 illustrates these three types of error. For the case of vocal effort segmentation, 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 a later vocal effort classification step. Hence, false alarm errors are less important than miss detection errors in the overall 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 the total

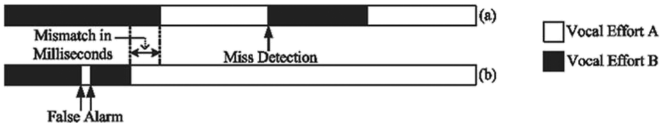

Fig. 12. Three types of segmentation error for whisper islands in audio streams.

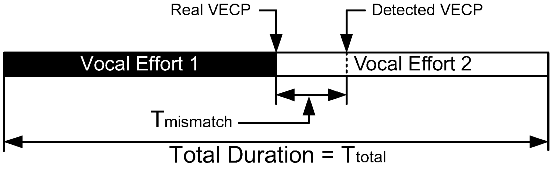

Fig. 13. Illustration of how boundary mismatch (in millseconds) is converted to overall boundary mismatch rate (in %): Mismatch$\% = (T_{\rm mismatch}/T_{\rm total}) \times 100\%$.

duration of two consecutive segments corresponding to the actual break points, which can be illustrated by Fig. 13 with VE1 and VE2 denoting two different vocal efforts respectively (e.g., potentially neutral and whisper). The multi-error score (MES) can be expressed using the following equation:

$$\begin{aligned}\text{MES} =&\ \text{False Alarm Rate(FAR)}\\ &+ 2 \times \text{Mismatch Rate(MMR)}\\ &+ 3 \times \text{Miss Detection Rate(MDR).}\end{aligned} \tag{4}$$

where

$$\begin{aligned}\text{FAR} &= \frac{\text{False Alarm VECP \#}}{\text{VECP \# in Detection Result}} \times 100\%\\ \text{MDR} &= \frac{\text{Not Detected VECP \#}}{\text{Total Real VECP \#}} \times 100\%\\ \text{MMR} &= \frac{\text{Mismatch}}{\text{Total Duration}} \times 100\%.\end{aligned} \tag{5}$$

The costs associated with FAR, MMR, and MDR are set to 1, 2, and 3 respectively (other cost values can be selected based on user goals). Next, we consider the meaning of the resulting MES scores, including an upper and lower bound on performance. The ideal whisper/neutral segmentation has a zero false alarm rate, zero miss detection rate, and 0 mismatch in milliseconds, and thus the multi-error score (MES) will be zero in the ideal/lower-bound case. If in the case that the mismatch rate, false alarm rate, and miss detection rate are all no greater than 10%, the resulting MES is no greater than 60, which we denote as "fair" performance in segmentation. If the false alarm rate is at 10%, but the miss detection rate and mismatch rate are at 15%, the MES becomes 85, which suggests that segmentation is collectively considered to be performing "poorly." If the error rates are all at their maximum value (100%), then the MES will be 600. Therefore, the goal of the segmentation phase is to achieve an MES as small as possible.

2) BIC/T^2-BIC Algorithm: The ultimate goal of segmentation is to produce a sequence of discrete segment blocks with consistent acoustic characteristics within each block (e.g., the goal is to limit any "flip-flop" action between classes, resulting in many short duration segments). The operating characteristics of the specific choice of segmentation structure will depend on the overall goals of the whole system. In this paper, vocal effort in terms of whispered versus neutral speech are chosen as the primary speech/speaker characteristics.

Segmentation can be reformulated as a model selection task between two nested competing models. Bayesian information criterion (BIC), a penalized maximum-likelihood model selection criterion, is employed for model selection [34]. The segmentation decision is derived by comparing BIC values. As a statistical data processing method, BIC requires no prior knowledge concerning acoustic conditions and no prior model training is needed. Instead of choosing a hard threshold for the segmentation decision, BIC statistically finds the difference in the acoustic features of the input frames to determine a point which can separate the data within the processing window into two models.

The problem of model selection is to choose one among a set of candidate models M_i, $i = 1, 2, \ldots, m$, as well as the corresponding model parameters θ_i to represent a given data set $D = (D_1, D_2, \ldots, D_N)$. These candidate models may be nested or non-nested. The BIC value of model M_i for the given data is defined as

$$BIC(M_i) = \log P(D_1, D_2, \ldots, D_N|M_i) - \frac{1}{2}d_i \log N \tag{6}$$

where d_i is the number of independent parameters in the model parameter set, and $P(D_1, D_2, \ldots, D_N|M_i)$ is the maximized data likelihood for the given model. In BIC, the term $(1/2)d_i \log N$ is subtracted from the log-likelihood as a penalty for model complexity, where BIC favors the model which maximizes the BIC values [35].

For a segmentation task, let us denote $X = x_i \in \mathbf{R}^d$, $i = 1, 2, \ldots, N$ as the sequence of frame-based feature vectors extracted from an input audio stream in which there is at most one segment boundary. Frame $b \in (1, N)$ will be tested as a potential whisper/neutral/silence boundary. If we suppose that each acoustically homogeneous speech block can be modeled as one multivariate Gaussian process $X \sim N(\mu, \Sigma)$, the segmentation issue can be viewed as a model selection problem between the following two nested models [34]:

$$\begin{aligned}&M_1 : X = x_1, x_2, \ldots, x_N \sim N(\mu, \Sigma)\\ &M_2 : x_1, x_2, \ldots, x_b \sim N(\mu_1, \Sigma_1);\\ &\qquad x_{b+1}, x_{b+2}, \ldots, x_N \sim N(\mu_2, \Sigma_2).\end{aligned}$$

Here, model M_1 assumes that all samples are distributed as a single Gaussian, which implies there is no boundary within the input stream of data. Model M_2 assumes the first b frame samples are drawn from one Gaussian while the last $N - b$ frame samples are drawn from a separate Gaussian. With these assumptions, if BIC favors M_1 then the data is assumed to be homogeneous, otherwise a boundary frame should be found within this block of data at frame location "b".

For a normal distribution $N(\mu, \Sigma)$, the likelihood of the observation data $x_1, x_2, \ldots, x_N$ is maximized when $\mu = \hat{\mu}$ and $\Sigma = \hat{\Sigma}$, where

$$\hat{\mu} = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{7}$$

and

$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})'. \tag{8}$$

Using (6), the BIC values for models M_1 and M_2 can obtained as

$$\begin{aligned} B\bar{I}C(M_1) = &-\frac{d}{2}N\log 2\pi - \frac{N}{2}\log|\hat{\Sigma}| \\ &-\frac{N}{2} - \frac{1}{2}\lambda\left(d + \frac{1}{2}d(d+1)\right)\log N \end{aligned} \tag{9}$$

and

$$\begin{aligned} B\bar{I}C(M_2) = &-\frac{d}{2}N\log 2\pi - \frac{b}{2}\log|\hat{\Sigma}_1| - \frac{N-b}{2}\log|\hat{\Sigma}_2| \\ &-\frac{N}{2} - \lambda\left(d + \frac{1}{2}d(d+1)\right)\log N \end{aligned} \tag{10}$$

respectively, where $\hat{\Sigma}$, $\hat{\Sigma}_1$, and $\hat{\Sigma}_2$ are ML covariance estimations from corresponding sample data, λ is the penalty factor to compensate for small sample cases, and d is the overall feature dimension size. Next, the BIC value difference ΔBIC can be viewed as a function of the boundary frame point b

$$\begin{aligned} \Delta BIC(b) &= B\bar{I}C(M_2) - B\bar{I}C(M_1) \\ &= \frac{1}{2}\left(N\log|\hat{\Sigma}| - b\log|\hat{\Sigma}_1| - (N-b)\log|\hat{\Sigma}_2|\right) \\ &\quad - \frac{1}{2}\lambda\left(d + \frac{1}{2}d(d+1)\right)\log N. \end{aligned} \tag{11}$$

According to the BIC rule, segmenting this audio stream into two parts at frame b will be favored if $\Delta BIC(b) > 0$. The final segmentation point decision can be achieve via MLE as

$$\hat{b} = \arg\max_{1<b<N;\Delta BIC(b)>0} \Delta BIC(b). \tag{12}$$

Obviously, the BIC-based segmentation algorithm has quadratic complexity. The computational cost is extensive since the determinants of two full covariance matrices for every possible frame break point b in a window must be evaluated. Therefore, Zhou and Hansen [28] proposed the T^2-statistic to detect possible boundary points faster and more efficiently. Hotelling's T^2-statistic is a multivariate analog of the well-known t-distribution [36] was employed. In terms of segmentation, the problem can now be stated as follows: for a given audio stream $X = x_i \in \mathbf{R}^d$, $i = 1, 2, \ldots, N$, determine if the two sets of samples, one containing the frames $[1, b]$ and the second containing $[b+1, N]$, are homogeneous. If the covariance of the audio stream is assumed to be common and unknown, the two sets of samples are homogeneous if and only if they are drawn from the same underlying normal distribution. With this statement, the segmentation problem can be viewed as verifying the hypothesis $H_0 : \mu_1 = \mu_2$ against the alternative $H_1 : \mu_1 \neq \mu_2$, where μ_1 and μ_2 are the means of the two sets of frame samples respectively. From [36], the likelihood ratio test is derived as the following T^2-Statistic [28]:

$$T^2(b) = \frac{b(N-b)}{N}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2) \tag{13}$$

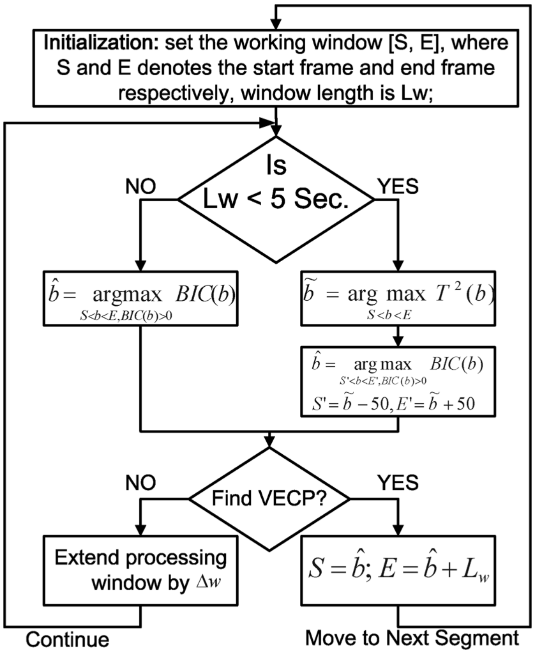
Initialization: set the working window [S, E], where S and E denotes the start frame and end frame respectively, window length is Lw;
NO
Is Lw < 5 Sec.
YES
$\hat{b}$= argmax BIC(b)
S<b<E,BIC(b)>0
$\tilde{b}$ = arg max $T^2(b)$
S<b<E
$\hat{b}$ = arg max BIC(b)
S'<b<E',BIC(b)>0
S' = $\tilde{b}$ − 50, E' = $\tilde{b}$ + 50
NO
Find VECP?
YES
Extend processing window by Δw
$S = \hat{b}$; $E = \hat{b} + L_w$
Continue
Move to Next Segment

Fig. 14. Segmentation based on BIC/T^2-BIC processing algorithm for VECP detection.

where Σ is the common covariance matrix. As stated in [36], the T^2 value is distributed as T^2 with $N-2$ degrees of freedom. Instead of setting a threshold, the T^2 value defined in (14) is used as a distance measure for the sets of samples. Obviously, the smaller the value of T^2, the more similar the two sample distribution sets. Thus,

$$\tilde{b} = \arg\max_{S<b<E} T^2(b) \tag{14}$$

can be viewed as the possible VECP within the current processing window.

In [28], the T^2 value was calculated for frame $b \in (1, N)$ to find the candidate boundary frames in the region near $\hat{b}$. Next, BIC value calculations are performed only on the frames in the neighborhood of $\hat{b}$ to find the best frame breakpoint and verify the decision of the boundary according (11) and (12). In this study, for increased accuracy and reliable detection, the BIC calculation was performed within the range $[(\tilde{b} - 50), (\tilde{b} + 50)]$ after the T^2 statistic algorithm was used to detect the possible VECP $\tilde{b}$. With this, we have

$$\hat{b} = \arg\max_{(\tilde{b}-50)<b<(\tilde{b}+50);BIC(b)>0} BIC(b) \tag{15}$$

which represents the VECP. In this paper, the T^2-Statistic was integrated within the BIC algorithm in this manner for processing shorter audio streams, while the traditional BIC algorithm was used to process long duration blocks. Since most experimental data used in this study represent read TIMIT sentences with different vocal effort levels, which are 2–3 s in duration, the BIC algorithm was used for a process window L_w larger than 5 s, and T^2-BIC was used when L_w was less than 5 s.

The implementation of the overall proposed segmentation algorithm for vocal effort change point (VECP) detection is described in Fig. 14.

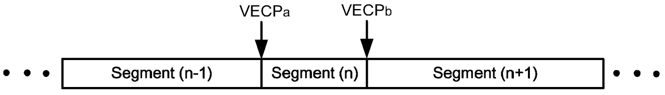

Fig. 15. Example of compensation of false-alarmed VECP.

C. False Alarm Compensation

It is expected that false alarm detection errors of VECPs will occur, and these are troublesome if the segments are too short, and not sufficiently long in duration to be correctly classified by a GMM model classifier. In this paper, a compensation algorithm based on spectral tilt (ST) is used to address the false alarm errors from VECP detection. In [6], it was shown that the average spectral tilt of whispered speech is much flatter than that of neutral speech. As a statistical property, although spectral tilt may not be used for precisely classifying vocal effort for each speech frame, it can be used to eliminate the false alarmed VECP errors in this study. Fig. 15 shows an example of a short segment (n) formed by two detected VECP_a and VECP_b. The corresponding spectral tilts ST_{n-1}, ST_n, and ST_{n+1} of segment (n), previous segment $(n-1)$, and next segment $(n+1)$ are calculated and compared. If the value of ST_n is close to ST_{n-1}, then the VECP_a can be eliminated as a false alarm VECP. Also, if ST_n is close to ST_{n+1}, then VECP_b can also be eliminated as a false alarm VECP. By doing this, some of the false alarm VECPs can be removed from the detected VECPs, such that the audio stream can be partitioned more precisely according to true vocal effort changes. Later on in the classification step, the remaining false alarm VECPs can be addressed by merging two very close segments of common vocal effort, or by merging two adjacent segments classified as the same vocal effort.

D. System Description

By using the proposed entropy-based feature, the BIC/T^2-BIC segmentation algorithm is able to detect VECPs between whisper and neutral speech within the input audio stream. The input audio stream is divided into a sequence of segment blocks according to the VECPs detected. Although there may be some false alarms in VECP detection, the subsequent classification step can compensate for false alarm errors by merging adjacent segments which are correctly classified as the same vocal effort. In the classification step, GMMs which are trained with whisper and neutral speech using the proposed entropy-based feature instead of MFCCs, are used to classify and label the vocal effort segments. The flow diagram of the entire system is illustrated in Fig. 16. In [5], the analysis results showed that the average spectrum of consonants for whisper are similar as the that of the consonants for neutral speech. Furthermore, for both whisper and neutral speech, the spectrum of consonants are flatter than that of the vowels. Thus, a preprocessing step was deployed to remove the frames with spectral tilt flatter than 0.8 of the average spectral tilt of input audio stream. Next, entropy-based feature vectors are extracted from the preprocessed audio speech frames and used by the BIC/T^2-BIC segmentation algorithm to detect the VECPs between whisper and neutral speech within the audio stream. Feature vectors of segments obtained according to the detected VECPs are compared with GMMs for whisper and neutral speech, and the model with the higher score is labeled as that vocal effort for the segment. As a final stage, successive segments labeled as homogeneous vocal effort classes are merged into an entire single segment to compensate for the false alarm error in VECP detection.

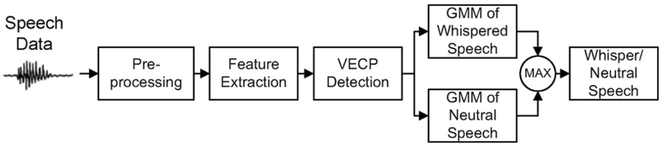

Fig. 16. Flow diagram of entire system.

TABLE I
EVALUATION FOR VOCAL EFFORT CHANGE POINTS DETECTION

Feature Type	*MDR(%)*	*FAR(%)*	*MMR(%)*	**MES**
MFCC Feature	1.13	27.44	2.63	36.09
WhID Feature	0.00	8.13	1.69	**11.51**

VI. EVALUATION

A. Experimental Setup

The experimental results reported in this study are all evaluated on audio streams from UT-VE I & II (from Section III) recorded using a SHURE Bets-54 close-talking microphone. All streams were sampled at 16 kHz with 16-bit sample mono audio. Again, ground-truth knowledge of absolute sound pressure level is known based on integrated 75 dB-SPL test tones at 1 kHz during all recording sessions.

B. Experimental Results of Segmentation

To test performance of the proposed entropy-based feature in VECP detection using the BIC/T^2-BIC algorithm, an experiment was carried out, with experimental results evaluated using the proposed multi-error score. For each subject, the speech audio from UT-VE II, which consists of 41 TIMIT sentences read by the subject alternatively in whisper and neutral mode, was used in the experiments. The vocal effort and onset and offset time of whisper/neutral segment within each speech audio were manually labeled and saved as a text file corresponding to the speech audio. The audio files from 59 subjects were employed in the present experiment to detect the VECPs. The transcript files of these audio streams were used to compare with VECP detection results obtained from the BIC/T^2-BIC algorithm using the proposed feature, so that the MES can be calculated. In addition to these features, the classic 13-D MFCC feature(without energy feature) was also used within our algorithm for experiments as a reference. The MES score and detailed error scores using MFCC and the proposed WhID feature are summarized in Table I, where the penalty factor for BIC/T^2-BIC segmentation in (11) is set to $\lambda = 1.2$.

The reduction of MES from 36.09 to 11.51, as well as reduction in MDR, FAR, and MMR, is quite remarkable. From the experimental result, the zero value of MDR denotes that all the VECPs within the audio stream were detected, with 1.69% of mismatch rate compared to the real VECPs. Among the total VECPs in the detection result, 8.13% of change points are false

TABLE II
TRAINING SCENARIOS FOR GMM-BASED CLASSIFIER

Scenario	*Testing Subjects*	*Training Subjects*	*Feature*
A	each of 59	rest 58 of 59	13-D MFCC
B	each of 59	rest 58 of 59	4-D WhID
C	20 Male	39 Female	4-D WhID
D	39 Female	20 Male	4-D WhID

TABLE III
EVALUATION FOR OVERALL WHISPER ISLAND DETECTION

Scenario	*Detected Number*	*Detection Rate(%)*
A	572	48.39%
B	1182	100%
C	400	100%
D	782	100%

alarms which will be compensated in the subsequent classification step.

C. Experimental Results of the Whisper Island Detection System

Based on the VECPs detected in the segmentation step, the audio stream is partitioned into several segments which are then to be classified by a GMM-based vocal effort classifier. There are four training scenarios for 64 GMMs of whisper and neutral vocal efforts for classifier. Table II shows the details of all four experimental scenarios. The round-robin technique was deployed in experimental Scenarios A&B to obtain the average performance of the vocal effort classification. The same audio streams used in the previous subsection are employed here. The subjects, used include a combination of male and female (Scenarios A&B), as well as gender dependent (Scenarios C&D). From the input audio streams, there are generally 20 whisper-islands for each audio stream, resulting in a maximum of 1182 whisper-islands (782 for female and 400 for male) in total for detection. The detection results for each subject's speech was compared with the human transcript file having manually labeled vocal effort to calculate the detection rate of whisper-island. Table III shows the performance for the overall system in terms of detection rate of whisper-islands.

From the Table III, we can see that the classifier in scenarios B, C, and D, trained using the 4-D proposed WhID feature, correctly detected and labeled all 1182 whispered islands. It should be noted that the false alarmed VECPs were compensated in this step.

VII. DISCUSSION

In whisper-island detection, one critical point for detection accuracy is the ability to find the boundary points between the whisper-island and the normally phonated speech. Therefore, although keeping the other two error rates low is necessary (FAR, MMR), obtaining a 0% Miss Detection Rate is far more important. The reason for this is that once a segment of whisper has been incorrectly incorporated into an adjacent neutral speech block, it is not possible to ever recover that island for alternative speech processing. The algorithm development here for the proposed feature is focused on achieving as close to zero MDR as possible, while also trying for low FAR and MMR as well, which makes the success of the subsequent classifying task possible. The zero miss detection of VECPs shows that the proposed 4-D WhID feature set is sensitive to the vocal change between whisper and neutral speech.

The 100% percent detection rate in Table III for scenarios B, C, D also show that the model trained with the proposed feature can clearly describe the vocal effort difference between whisper and neutral speech. Comparing the detection rate for scenarios C and D, although scenario C and D was trained with all female speech and male speech respectively, the two scenarios have both excellent detection rate performance with 100%. This fact indicates that the proposed feature is gender independent, which can be useful in GMM-based classifier training regardless of the gender ratio in the training data.

Although the proposed algorithm works well with the proposed entropy-based feature in both VECP detection and vocal effort classification, the proposed feature is based on spectral entropy of speech frame, which is depending on the energy distribution in frequency domain. The robustness of the proposed feature and algorithm was not examined under noisy environments. Due to the low energy property of the whispered speech, the noise interference with a moderate signal-to-noise ratio (SNR) for neutral speech segments may cause a relative very low SNR for whispered speech, even bury the whispered speech segments. Thus, before examining the robustness of our algorithm and feature, a balanced and normalized metric is needed to evaluate the noise interference within both neutral and whispered speech segments. With such metric, the performance of our algorithm and feature can be objectively evaluated with noise present.

Finally, it should be noted that whisper island detection is a challenging research task in general, but becomes even more difficult in changing acoustic conditions (e.g., background noise), or when speech capture properties are unknown (e.g., differing microphone pickup, handset, or acoustic microphone placement—near versus far-field microphone). Another related issue which is of interest is the variability of whisper speech across subjects in general, and the issue of how long a subject can sustain a true whisper speaking style. Previous work by Zhang and Hansen [6] illustrated the diversity in vocal speech traits, and data collection for that study pointed to the challenge for subjects to be able to produce true whisper, versus combinations of whisper and soft spoken speech over time. Of course, if the goal is to sustain speech system performance for speaker recognition, speech recognition, or other speech/speaker classification task, then from the users perspective, all that is necessary is to identity those islands which limit speech system performance the most. Alternative processing strategies may be needed for soft versus whisper spoken speech, because of the differences in their production traits.

VIII. 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 virtually ineffective (ASR, speaker ID, coding, etc.). Effective whisper island detection is the first step needed to ensure that subsequent engagement of effective speech processing steps could be employed to address whisper.

In this paper, we proposed a 4-D WhID feature combined with a BIC/T^2-BIC algorithm which has shown to have the lowest proposed multi-error score to date with zero MDR, meaning that all vocal effort change points between whisper and neutral within audio streams have been correctly detected. Next, the segments partitioned according to the detected VECPs were correctly labeled based on vocal effort by the trained GMM-based classifier. It was also shown that the false alarm VECPs produced in the previous segmentation step were all eliminated by merging the adjoint segments with identical vocal effort classes.

The zero miss detection rate of VECPs and 100% detection rates for classifiers trained with both male and female speech indicate that the proposed WhID feature set is not only sensitive to the vocal effort change between whisper and neutral speech but also is gender independent. Furthermore, the BIC/T^2-BIC segmentation algorithm works effectively with the proposed feature in VECP detection. Finally, the proposed algorithm has been applied successively and precisely for detection of whisper-islands embedded in the neutral speech audio streams. Having achieved this stage in the field of whisper speech processing, it is now possible to develop future strategies to improve speech system performance for whisper-based speech in speech/language technology.

References

[1] S. Jovicic and Z. Saric, "Acoustic analysis of consonants in whispered speech," *J. Voice*, vol. 22, pp. 263–274, 1997.

[2] J. Koufman, "The spectrum of vocal dysfunction," *Otolaryngol. Clinics North Amer. Voice Disorders*, vol. 24, no. 5, Oct. 1991.

[3] L. Gavidia-Ceballos and J. H. L. Hansen, "Direct speech feature estimation using an iterative EM algorithm for vocal fold pathology detection," *IEEE Trans. Biomed. Eng.*, vol. 43, no. 4, pp. 373–383, Apr. 1996.

[4] J. H. L. Hansen, L. Gavidia-Ceballos, and J. Kaiser, "A nonlinear operator-based speech feature analysis method with application to vocal fold pathology assessment," *IEEE Trans. Biomed. Eng.*, vol. 45, no. 3, pp. 300–313, Mar. 1998.

[5] T. Ito, K. Takeda, and F. Itakura, "Analysis and recognition of whispered speech," *Speech Commun.*, vol. 45, pp. 139–152, 2005.

[6] C. Zhang and J. H. L. Hansen, "Analysis and classification of speech mode: Whispered through shouted," in *Proc. Interspeech'07*, Antwerp, Belgium, 2007, pp. 2289–2292.

[7] L. Gavidia-Ceballos, "Analysis and modeling of speech for laryngeal pathology assessment," Ph. D. dissertation, Dept. of Biomed. Eng., Duke Univ., Durham, NC, 1995.

[8] N. Solomon, G. McCall, M. Tosset, and W. Gary, "Laryngeal configuration and constriction during two types of whispering," *J. Speech Hear Res.*, vol. 32, no. 1989.

[9] K. Kallail and F. Emanuel, "Formant-frequency differences between isolated whispered and phonated vowel samples produced by adult female subjects," *J. Speech Hearing Res.*, vol. 27, pp. 245–251, 1984.

[10] K. Kallail, "An acoustic comparison of isolated whispered and phonated vowel samples produced by adult male subjects," *J. Phon.*, vol. 12, pp. 175–186, 1984.

[11] I. Eklund and H. Traumuller, "Comparative study of male and female whispered and phonated versions of the long vowels of Swedish," *Phonetica*, vol. 54, pp. 1–21, 1996.

[12] M. Gao, "Ones in whispered Chinese: Articulatory features and perceptual cues," M.A. thesis, Dept. of Linguist., Univ. of Victoria, Victoria, BC, Canada, 2002.

[13] M. Schwartz and H. Rine, "Identification of speaker sex from isolated whispered vowels," *J. Acoust. Soc. Amer.*, vol. 44, pp. 1736–1737, 1968.

[14] K. Lass, K. Hughes, M. Bowyer, L. Waters, and V. Bourne, "Speaker sex identification from voiced, whispered and filtered isolated vowels," *J. Acoust. Soc. Amer.*, vol. 59, pp. 675–678, 1976.

[15] X. Fan and J. H. L. Hansen, "Speaker identification for whispered speech based on frequency warping and score competition," in *Proc. Interspeech'08*, Brisbane, Australia, 2008, pp. 1313–1316.

[16] X. Fan and J. H. L. Hansen, "Speaker identification with whispered speech based on modified lfcc parameters and feature mapping," in *Proc. ICASSP'09*, Taipei, Taiwan, 2009, pp. 4553–4556.

[17] S. Wenndt, J. Cupples, and M. Floyd, "A study on the classification of whispered and normal phonated speech," in *Proc. Interspeech'02*, Denver, CO, 2002, pp. 649–652.

[18] C. Zhang and J. H. L. Hansen, "Effective segmentation based on vocal effort change point detection," in *Proc. ITRW*, Aalborg, Denmark, Jun. 2008.

[19] W. Meyer-Eppler, "Realisation of prosodic features in whispered speech," *J. Acoust. Soc. Amer.*, vol. 29, pp. 104–106, 1957.

[20] I. Thomas, "Perceived pitch of whispered vowels," *J. Acoust. Soc. Amer.*, vol. 46, pp. 468–470, 1969.

[21] M. Matsuda and H. Kasuya, "Acoustic nature of the whisper," in *Proc. Eurospeech*, 1999, vol. 1, pp. 137–140.

[22] M. Higashikawa, J. Green, C. Moore, and F. Minifie, "Lip kinematics for /p/ and /b/ production during whispered and voiced speech," *Folia Phoniar Logop*, vol. 55, pp. 1–9, 2003.

[23] H. Konno, J. Toyama, M. Shimbo, and K. Murata, "The effect of formant frequency and spectral tilt of unvoiced vowels on their perceived pitch and phonemic quality," *IEICE Tech. Rep.*, vol. SP95-140, pp. 39–45, Mar. 1996.

[24] S. Jovicic, "Formant feature differences between whispered and voiced sustained vowels," *Acta Acust. United With Acust.*, vol. 84, no. 4, pp. 739–743, Jul. 1998.

[25] J. H. L. Hansen, "Analysis and compensation of stressed and noisy speech with application to robust automatic recognition," Ph.D. dissertation, Georgia Inst. of Technology, Atlanta, 1988.

[26] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Commun., Special Iss. Speech Under Stress*, vol. 20, no. 2, Nov. 1996.

[27] A. Patil and J. H. L. Hansen, "Detection of speech under physical stress: Model development, sensor selection, and feature fusion," in *Proc. Interspeech'08-ICSLP*, 2008.

[28] B. Zhou and J. H. L. Hansen, "Efficient audio stream segmentation via the combined t2 statistic and Bayesian information criterion," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp. 467–474, Jul. 2005.

[29] C. Zhang and J. H. L. Hansen, "An entropy based feature for whisper-island detection within audio streams," in *Proc. Interspeech'08*, Brisbane, Australia, 2008.

[30] R. Huang and J. H. L. Hansen, "Advances in unsupervised audio classification and segmentation for the broadcast news and NGSW corpora," *IEEE Trans., Audio, Speech, Language Process.*, vol. 14, no. 3, pp. 907–919, May 2006.

[31] X. Li, H. Ding, and B. Xu, "Entropy-based initial/final segmentation for chinese whispered speech," *Acta Acustica-BEIJING*, vol. 30, pt. 1, pp. 69–75, 2005.

[32] J. Shen, J. Huang, and L. Lee, "Robust entropy-based endpoint detection for speech recognition in noisy environments," in *Proc. ICSLP*, 1998, pp. 232–235.

[33] K. Weaver, K. Waheed, and F. Salem, "An entropy-based robust speech boundary detection algorithm for realistic noisy environments," in *Proc. Int. Joint Conf. Neural Netw.*, Jul. 2003, vol. 1, no. 20–24, pp. 680–685.

[34] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proc. Broadcast News Transcr. Under. Workshop*, 1998.

[35] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, pp. 461–464, 1978.

[36] T. Anderson, *An Introduction to Multivariate Statistical Analysis*. New York: Wiley, 1958.

Chi Zhang was born in Nanjing, China. He received B.S.E.E. degree from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2004 and the M.S. degree in electrical engineering from Aalborg University, Aalborg, Denmark, in 2006. He is currently pursuing the Ph.D. degree at the University of Texas at Dallas, Richardson, under supervision of Dr. John. H. L. Hansen.

In August 2006, he joined the Center for Robust Speech Systems, Eric Jonsson School of Engineering and Computer Science, University of Texas at Dallas, as a Research Assistant. His research interests are focused in the field of speech processing, including analysis, modeling, and classification of speech under different vocal effects and whispered speech detection.

John H. L. Hansen (S'81–M'82–SM'93–F'07) received the B.S.E.E. degree from the College of Engineering, Rutgers University, New Brunswick, NJ, in 1982 and the M.S. and Ph.D. degrees in electrical engineering from the Georgia Institute of Technology, Atlanta, in 1983 and 1988, respectively.

He joined the Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas (UTD), Richardson, in the fall of 2005, where he is a Professor and Department Head of Electrical Engineering, and holds the Distinguished University Chair in Telecommunications Engineering. He also holds a joint appointment as a Professor in the School of Brain and Behavioral Sciences (Speech and Hearing). At UTD, he established the Center for Robust Speech Systems (CRSS) which is part of the Human Language Technology Research Institute. Previously, he served as Department Chairman and Professor in the Department of Speech, Language, and Hearing Sciences (SLHS), and Professor in the Department of Electrical and Computer Engineering, at the University of Colorado, Boulder (1998–2005), where he cofounded the Center for Spoken Language Research. In 1988, he established the Robust Speech Processing Laboratory (RSPL) and continues to direct research activities in CRSS at UTD. His research interests span the areas of digital speech processing, analysis and modeling of speech and speaker traits, speech enhancement, feature estimation in noise, robust speech recognition with emphasis on spoken document retrieval, and in-vehicle interactive systems for hands-free human–computer interaction. He has supervised 51 (22 Ph.D., 29 M.S.) thesis candidates, is author/coauthor of 370 journal and conference papers in the field of speech processing and communications, coauthor of the textbook *Discrete-Time Processing of Speech Signals*, (IEEE Press, 2000), coeditor of *DSP for In-Vehicle and Mobile Systems* (Springer, 2004), *Advances for In-Vehicle and Mobile Systems: Challenges for International Standards* (Springer, 2006), *In-Vehicle Corpus and Signal Processing for Driver Behavior Modeling* (Springer, 2008), and lead author of the report "The Impact Of Speech Under 'Stress' on Military Speech Technology," (NATO RTO-TR-10, 2000).

Prof. Hansen was recipient of the 2005 University of Colorado Teacher Recognition Award as voted by the student body. He also organized and served as General Chair for ICSLP/Interspeech-2002: International Conference on Spoken Language Processing, September 16–20, 2002, and served as Co-Organizer and Technical Program Chair for the IEEE ICASSP-2010 Conference in Dallas, TX, in March 2010. In 2007, he was named IEEE Fellow for contributions in "Robust Speech Recognition in Stress and Noise," and is currently serving as a Member of the IEEE Signal Processing Society Speech Technical Committee and Educational Technical Committee. Previously, he has served as Technical Advisor to U.S. Delegate for NATO (IST/TG-01), IEEE Signal Processing Society Distinguished Lecturer (2005–006), Associate Editor for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (1992–1999), Associate Editor for the IEEE SIGNAL PROCESSING LETTERS (1998–2000), Editorial Board Member for the *IEEE Signal Processing Magazine* (2001–2003). He has also served as a Guest Editor of the October 1994 special issue on Robust Speech Recognition for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. He has served on the Speech Communications Technical Committee for the Acoustical Society of America (2000–2003), and is serving as a member of the ISCA (International Speech Communications Association) Advisory Council.