Frame-Level Vocal Effort Likelihood Space Modeling for Improved Whisper-Island Detection

Chi Zhang and *John H.L. Hansen*

Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering & Computer Science
University of Texas at Dallas, Richardson, TX 75080, USA
{cxz055000;john.hansen}@utdallas.edu

Abstract

In this study, a frame-based vocal effort likelihood space modeling framework for improved whisper-island detection within normally phonated audio streams is proposed. The proposed method is based on first training a traditional Gaussian mixture model for whisper and neutral speech, which is then employed to extract a newly proposed discriminative feature set entitled Vocal Effort Likelihood (VEL), for whisper-island detection. The VEL feature set is integrated within a BIC/T^2-BIC segmentation scheme for vocal effort change point(VECP) detection. With the dimension-reduced VEL 2-D feature set, the proposed framework has reduced computational costs versus prior method [1]. Experimental results using the UT-VocalEffort II corpus for whisper-island detection using the proposed framework are presented and compared with a previous algorithm introduced in [1]. The proposed algorithm is shown to improve performance in VECP detection with the lowest Multi-Error Score(MES) of 6.33. Furthermore, very accurate whisper-island detection was obtained using proposed algorithm, which is useful for sustained performance in speech systems (ASR, Speaker-ID, etc.)which might experience whisper speech. Finally, experimental performance achieves a 100% detection rate for the proposed algorithm, which represents the best whisper-island detection performance with lowest computational costs available in the literature to date.

Index Terms: Vocal Effort Likelihood, Vocal Effort, Whisper-Island Detection, GMM Classifier

1. Introduction

Whisper speech is one mode of natural speech communication which results in reduced perceptibility and a significant reduction in intelligibility. In general, with the absence of vocal fold vibrations, whispered speech may be intentional, or caused by a change in the vocal fold structure, or muscle control due to disease of the vocal system, such as functional aphonia [2], laryngeal cancer [3]. Furthermore, as a paralinguistic phenomenon, whispered speech can be used in environments where loud speech is prohibited, or in cases where the speaker would prefer to keep speech content private from being heard by remote listeners in public settings [4]. Current speech processing systems are generally designed for normally phonated speech, and are therefore severely impacted due to the fundamental change in speech production of whispered speech: the absence of all periodic/harmonic excitation. Whispered speech, within the range of vocal effort from whisper to shouted, has the most dramatic loss in terms of vocal effort for speech processing systems [5]. Therefore, detecting and identifying whispered islands embedded in the speech signal before further processing is useful in order to eliminate the negative impact of whispered speech on subsequent speech systems (ASR, Speaker ID, etc.). Furthermore, whispered speech has a high probability of conveying confidential or sensitive information. 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 algorithms have been developed for identifying whisper-islands within normally phonated audio streams, using different types of features extracted from the time waveform, spectral analysis of the speech signal or linear predictive residual [1, 6, 7]. In [1], an algorithm using a 4-D entropy-based feature set: WhID was proposed and shown to achieve good performance in whisper-island detection for both vocal effort change point detection and vocal effort classification.

In this study, a new framework deploys the proposed discriminative feature set entitled "Vocal Effort Likelihood(VEL)" to detect the vocal effort change point between whisper and neutral speech, and therefore improves the whisper-island detection with lower computational costs. The remainder of this paper is organized as follows. First, for the readers' benefit, the formulation of our previous 4-D WhID feature, and description of the UT-VocalEffort II speech corpus [1] are briefly presented in Sec. 2 and Sec. 3 respectively. The proposed framework and feature set for vocal effort likelihood whisper-island detection are presented in Sec. 4. Evaluations of the proposed algorithm are presented in Sec. 5 and compared to our previous method. Finally, conclusions and discussion are presented.

2. Formulation of Previously Developed WhID Feature

In [1], a 4-D feature set WhID, which is sensitive to vocal effort changes between whisper and neutral was formulated and used for feature extraction from speech and modeling of whisper/neutral speech and whisper-island detection. The 4-D WhID feature set for each 20ms speech frame is formulated as follows:

$$\begin{bmatrix} \text{1-D spectral information entropy ratio(ER);} \\ \text{2-D spectral information entropy(SIE);} \\ \text{1-D spectral tilt(ST).} \end{bmatrix} \quad (1)$$

This project was funded by AFRL under a subcontract to RADC Inc. under FA8750-09-C-0067 and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. Hansen.

Copyright © 2011 ISCA

ER and SIE calculation can be illustrated in Fig. 1 and Fig. 2 respectively. The spectral information entropy(SIE) is obtained

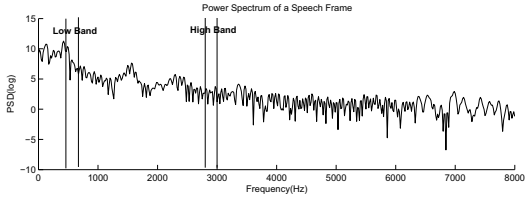

Figure 1: Entropy Ratio is Derived between High and Low Frequency Bands

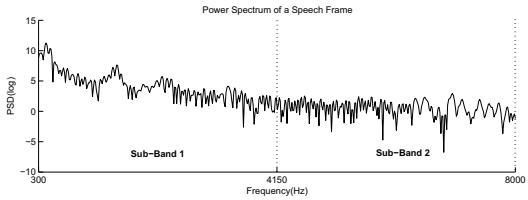

Figure 2: Two Sub-Bands over Frequency Domain for SIE Calculation

as follows. Assume $X(k)$ is the power spectrum of the speech frame $x(n)$, and k varies from k_1 to k_M in a sub-band; then that portion of the frequency content in the k band versus the entire response is written as,

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

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

$$H = -\sum_{k=k_1}^{k_M} p(k) \cdot \log p(k). \tag{3}$$

Furthermore, since the spectral tilt of whispered speech is statistically different from the spectral tilt of neutral speech [5], the spectral tilt can be used as a discriminative feature in differentiating whispered and neutral speech as the 4th dimension of the feature set WhID.

3. Corpus Description

In this study, two corpora were developed with different foci. Corpus UT-VocalEffort(UT-VE) I consists of speech under five vocal efforts: whispered, soft, neutral, loud and shouted, while corpus UT-VocalEffort(UT-VE) II focuses on neutral speech embedded with whispered speech "islands". Both corpora were collected in an ASHA certified, single walled sound booth using a multi-track FOSTEX 8-channel synchronized digital recorder with gain adjustments for individual channels. The details of UT-VE I were presented in [1].

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 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.

In the read part of UT-VE II, each subject was 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 last sentence pair read in the 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.

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

4. Proposed Algorithm

The framework for the proposed algorithm is illustrated in Fig. 4. To easily compare the difference between the proposed WhID-VEL-GMM framework and previous WhID-GMM framework [1], the previous framework is presented in Fig. 3.

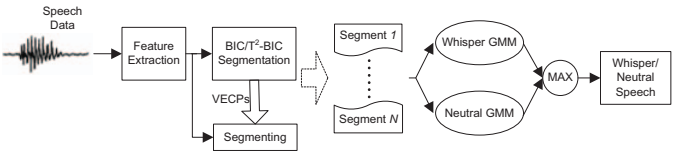

Figure 3: WhID-GMM: Flow Diagram of the Previous Framework for Whisper-Island Detection [1]

Fig.3 shows that, in the previously developed framework, the acoustic feature was extracted from each speech frame and submitted to the BIC/T^2-BIC segmentation algorithm to detect vocal effort change points(VECP) between whisper and neutral speech. Next, the speech feature sequence was segmented according to the detected VECPs and compared with GMMs of whisper and neutral speech to classify the vocal effort of each segment. As illustrated in Fig. 4, instead of directly using the feature WhID for VECP detection as in the previous framework, the proposed algorithm uses the frame-based vocal effort likelihood(VEL) scores as the discriminative feature to detect the VECP between whisper and neutral speech. Although a proper acoustic feature may be sensitive to the vocal effort change between whisper and neutral speech, the changes in content of the speech may introduce variations, which are not dependent on the vocal effort change within the feature. By comparing the speech feature of the current frame with the gender balance, phoneme balance, mono-vocal-effort GMM, which models the vocal effort, the output score can be viewed as the likelihood of the current frame being the vocal effort modeled by the GMM. In this case, the fluctuation of the feature set, which consists of the likelihood scores of speech frames from the whisper and neutral GMMs, is highly dependent on the vocal effort of the speech frames. Therefore, the feature set VEL may be viewed as a discriminative feature space between vocal efforts. In this study, to detect the whisper-islands, the vocal effort likelihood

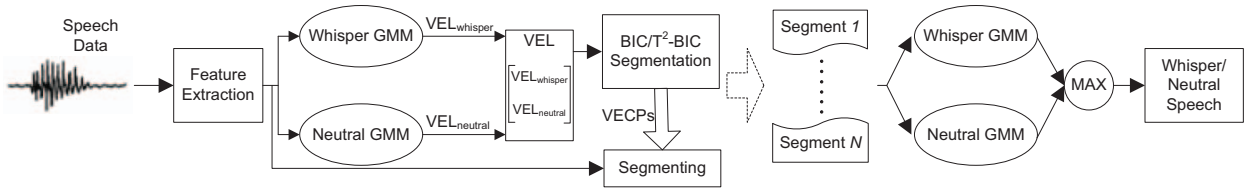

Figure 4: WhID-VEL-GMM: Flow Diagram of the Proposed Framework for Whisper-Island Detection

is calculated for whisper and neutral speech.

Furthermore, the feature set VEL, compared with the acoustical feature used to train the GMMs, has a reduced feature dimension. Although in our algorithm, the BIC algorithm has been improved as BIC/T^2-BIC to reduce the computational cost, the covariance matrix calculation and the updating in the BIC/T^2-BIC algorithm can still be costly in computations if we use a high dimension feature. By deploying the 2-D feature set VEL, compared with the 4-D feature WhID and 13-D feature MFCC [1], the computational costs will be reduced using the BIC/T^2-BIC segmentation scheme.

4.1. Vocal Effort Likelihood Feature Set Formulation

The speech data employs a 20ms frame length with 50% overlap between consecutive frames. For each frame, a specific acoustical feature (e.g., either 4-D feature set WhID or 13-D MFCC) is extracted. For the jth frame, the feature y_j is compared with the GMMs of whisper and neutral speech respectively to estimate the vocal effort likelihood(VEL) of each vocal effort given the feature vector y_n:

$$\begin{aligned} \mathrm{VEL}_{\mathrm{whisper}} &= \mathrm{p}(\mathrm{C}_{\mathrm{whisper}}|\mathrm{y_j}) \\ \mathrm{VEL}_{\mathrm{neutral}} &= \mathrm{p}(\mathrm{C}_{\mathrm{neutral}}|\mathrm{y_j}), \end{aligned} \tag{4}$$

where $\mathrm{C}_{\mathrm{whisper}}$ and $\mathrm{C}_{\mathrm{neutral}}$ represent the GMMs trained with the corresponding acoustical feature for whisper and neutral speech, respectively; $\mathrm{p}(\mathrm{C}_{\mathrm{whisper}}|\mathrm{y_j})$ and $\mathrm{p}(\mathrm{C}_{\mathrm{neutral}}|\mathrm{y_j})$ represent the scores obtained from the comparison between the current frame and GMMs of whisper and neutral speech, respectively. Next, the Vocal Effort Likelihood(VEL) feature vector can be formed as,

$$\begin{bmatrix} \mathrm{VEL}_{\mathrm{whisper}} \\ \mathrm{VEL}_{\mathrm{neutral}} \end{bmatrix} \tag{5}$$

Since GMMs trained with an appropriate acoustical feature can be viewed as a viable representation of the acoustic space of whisper and neutral vocal effort [1], $\mathrm{p}(\mathrm{C}_{\mathrm{whisper}}|\mathrm{y_j})$ and $\mathrm{p}(\mathrm{C}_{\mathrm{neutral}}|\mathrm{y_j})$ will represent the probability of the current frame being whisper or neutral speech. For example, for a given frame which is known to be whisper, $\mathrm{p}(\mathrm{C}_{\mathrm{whisper}}|\mathrm{y_j})$ should be much larger than $\mathrm{p}(\mathrm{C}_{\mathrm{neutral}}|\mathrm{y_j})$. Furthermore, the variances between phonemes, which may influence the acoustical feature, are normalized in the VEL set by comparing the feature of each frame with the GMMs.

4.2. BIC/T^2-BIC Algorithm for VECP Detection

Next, the BIC/T^2-BIC scheme derived based on the T^2-BIC algorithm(Zhou and Hansen [8]), is an unsupervised model-free scheme that detects acoustic change points based on the input feature data. One assumption for applying this algorithm is that the feature employed by the BIC/T^2-BIC algorithm is considered to be discriminative between vocal efforts of whisper and neutral speech.

In [8], the T^2 value is calculated for frame $b \in (1, N)$ to find the candidate boundary frame $\hat{b}$. Next, the BIC value calculation is performed only on frame $\hat{b}$ to verify the decision of the boundary. In this study, for more accuracy and reliable detection, BIC processing is performed within the range $[(\tilde{b} - 50), (\tilde{b} + 50)]$ after the T^2 statistic algorithm is used to detect the possible VECP $\tilde{b}$,

$$\hat{b} = \arg\max_{(\tilde{b}-50)<b<(\tilde{b}+50);BIC(b)>0} BIC(b). \tag{6}$$

Furthermore, T^2-Statistics are integrated within the BIC algorithm in this manner for processing longer audio streams, while the traditional BIC algorithm is used to process short duration blocks. Since most experimental data used in this study represent read TIMIT sentences with different vocal effort levels, which are 2-3s in duration, the BIC algorithm is used for process window, L_w less than 5s, and T^2-BIC is used when L_w is larger than 5s. The implementation of the overall BIC/T^2-BIC segmentation algorithm is detailed in [1].

4.3. Vocal Effort Classification

With the detection of VECPs, a GMM-based vocal effort classifier is deployed to label the vocal effort of each speech segment obtained from the detected VECPs. GMMs of whisper and neutral speech used to extract the VEL feature set can be reused for vocal effort classification. 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.

5. Evaluation Results

5.1. Brief Overview of Multi-Error Score

Since it is a challenging task to collectively evaluate the performance of a segmentation task, we previously developed the Multi-Error Score [1, 9, 10]. The MES consists of 3 error types for segmentation mismatch: miss detection rate, false alarm rate and average mismatch in milliseconds normalized by adjacent-segment duration. Fig. 5 illustrates these three types of error. The calculation of MES can be illustrated by the Eq. 7.

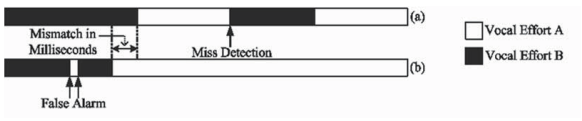

Figure 5: Three Types of Segmentation Error

$$\begin{aligned} \mathbf{MES} &= 1 \times \text{False Alarm Rate(FAR)} \\ &+ 2 \times \text{Mismatch Rate(MMR)} \\ &+ 3 \times \text{Miss Detection Rate(MDR)} \end{aligned} \tag{7}$$

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. The mismatch rate is obtained by calculating the percentage of the mismatch in milliseconds versus the total duration of the two adjacent segments corresponding to the actual breakpoints. Miss detection rate and mismatch rate are more costly errors for whisper island detection, so these errors are scaled by 3 and 2 respectively. MES is bounded by 0, for all 3 error rates at 0%, and 600 for all 3 error rates at 100%. A score of 90 occurs when all 3 error rates are 15%. More details concerning the MES can be found in [1].

5.2. Experimental Results in MES

An audio stream with 41 sentences produced alternatively in neutral and whispered mode by each subject from UT-VE II were manually labeled for VECPs in transcript files. The audio files from 59 subjects were employed for these experiments. The transcript files of these audio streams were used to compare with VECP detection results obtained from the different experimental scenarios, so that the MES can be calculated. The lower MES denotes better performance in VECP detection. A round robin process was used for GMM training to maximize training data size while ensuring open test conditions. Therefore, while testing with each of the 59 subjects, speech data from the other 58 out of 59 subjects was used in GMM training (i.e., open test speaker and open test speech). To illustrate the improvements brought by utilizing the proposed framework, the experiments of the proposed WhID-VEL-GMM framework using MFCC and WhID are presented. To compare with the WhID-GMM algorithm in [1], experimental results are presented in Table 1 and Table 2.

The experimental results in MES are shown in Table 1 for each scenario. It can be seen that when using the proposed framework, the MES has the smallest MES values for both the MFCC and WhID features than using the previous framework. In the bottom row of Table 1, the high MES proves that MFCC may not be an appropriate feature for BIC/T^2-BIC in VECP detection. However, by using the proposed framework, with MFCC, the MES achieves lowest 6.33 with 0.00% MDR as well.

Table 1: *Multi-Error Score Results of Experimental Scenarios*

Framework	*Feature*	*MDR(%)*	*FAR(%)*	*MMR(%)*	**MES**
WhID-VEL-GMM	WhID	0.00	6.75	1.49	9.74
MFCC-VEL-GMM	MFCC	0.00	3.51	1.41	6.33
WhID-GMM	WhID	0.00	8.13	1.69	11.51
MFCC-GMM	MFCC	1.13	27.44	2.63	36.09

5.3. Experimental Results of System

With an extremely low MES(especially 0.00% MDR) in VECP detection, the proposed framework shows excellent performance as does the previous framework in sensing vocal effort changes between whisper and neutral speech. Overall system performance is compared based on the detection rate of whisper-islands within neutral audio speech streams in Table 2. The same audio streams used in the last subsection were employed here. With 20 whisper-islands for each audio stream, there are 1182 potential whisper-islands in total for detection. Clearly, whisper-island detection performance is extremely effective for the WhID-VEL-GMM and MFCC-VEL-GMM systems.

Table 2: *Evaluation for Overall Whisper Island Detection*

Framework	*Feature*	*Detected*	*Detection Accuracy(%)*
WhID-VEL-GMM	WhID	1180/1182	99.83
MFCC-VEL-GMM	MFCC	1182/1182	100.00
WhID-GMM	WhID	1182/1182	100.00
MFCC-GMM	MFCC	572/1182	48.39

6. Conclusion and Discussion

Effective whisper island detection is the first step necessary for engaging robust of effective subsequent speech processing steps to address whisper. With reliable and accurate detection, the computational cost should also be considered carefully, especially for tasks dealing with large amounts of audio data. In this study, the proposed VEL-GMM framework works with MFCC and WhID feature sets to obtain reliable, accurate whisper-island detection with less computation, compared with our previous algorithm [1]. The proposed framework employing MFCC, which provides the lowest MES of 6.33 and 100% detection rate achieves the best whisper-island detection performance to date, and WhID feature is also outstanding(MES of 9.74 and 99.83% detection rate). The experimental results also show that the proposed vocal effect likelihood can be used as a discriminative feature utilized in VECP detection.

Based on these conclusions, it is intuitive to consider that the proposed framework may be used for bi-model segmentation and detection tasks, such as speaker segmentation for conversational speech, gender segmentation and speech/music segmentation. With properly trained GMMs, the VEL feature set can be formulated to be discriminative between the data types for which the GMMs respectively stand. Based on the discriminative VEL, the BIC/T^2-BIC algorithm can detect the change points between the data types within the audio streams. Therefore, the VEL-GMM framework can be employed in other speech tasks as well.

7. References

[1] C. Zhang and J.H.L. Hansen, "An unsupervised effective algorithm for whisper-island detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. PP, no. 99, p. 1, 2010.

[2] J. Koufman, "The spectrum of vocal dysfunction," *The Otolaryngologic Clinics of North America. Voice Disorders*, vol. 24(5), pp. 985–988, 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. on Biomedical engineering*, vol. 43(4), pp. 373–383, Apirl 1996.

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

[5] C. Zhang and J.H.L. Hansen, "Analysis and classification of speech mode: Whispered through shouted," *INTERSPEECH 07*, pp. 2289–2292, 2007.

[6] S. Wenndt, J. Cupples, and M. Floyd, "A study on the classification of whispered and normal phonated speech," *INTERSPEECH02*, pp. 649–652, 2002.

[7] C. Zhang and J.H.L. Hansen, "Advancements in whisper-island detection using the linear predictive residual," *ICASSP2010*, pp. 5170–5173, 2010.

[8] B. Zhou and J.H.L. Hansen, "Efficient audio stream segmentation via the combined T2 statistic and Bayesian information criterion," *IEEE Trans. Speech and Audio Processing*, vol. 13(4), pp. 467–474, July 2005.

[9] C. Zhang and J.H.L. Hansen, "An entropy based feature for whisper-island detection within audio streams," *INTERSPEECH2008-ICSLP*, pp. 2510–2513, 2008.

[10] ——, "Advancements in whisper-island detection within normally phonated audio streams," *INTERSPEECH09*, pp. 860–863, 2009.