

ADVANCEMENTS IN WHISPER-ISLAND DETECTION USING THE LINEAR PREDICTIVE RESIDUAL

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, Texas 75083, USA

{cxz055000, john.hansen}@utdallas.edu, <http://crss.utdallas.edu>

ABSTRACT

In this study, we consider the use of a new entropy-based feature extracted from linear predictive residual for whisper-island detection within normally phonated audio streams. The proposed feature, which is sensitive to vocal effort changes between whisper and neutral speech, is integrated within a BIC/ T^2 -BIC segmentation for vocal effort change point(VECP) detection and utilized for vocal effort classification. Evaluation is based on the proposed multi-error score(MES), where the improved feature is shown to improve performance in VECP detection with the lowest MES of 20.70. Furthermore, more accurate whisper-island detection was obtained using the proposed feature and algorithm. Finally, the experimental detection rate results of 97.37% represents the best whisper-island detection performance available in the literature to date.

Index Terms— whisper, feature, segmentation, detection, vocal effort, T^2 -BIC

1. INTRODUCTION

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

subsequent speech systems (ASR, Speaker ID, etc.). However, whispered speech has a high probability of conveying confidential or sensitive information. 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.

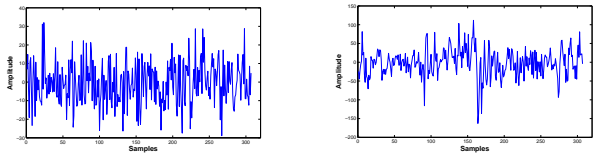
Several algorithms and features have been developed for identifying whisper-islands within normally phonated audio streams, using features extracted from the time waveform or spectral analysis of the speech signal[2][3]. In this study, the Linear Predictive residual(LPR) is considered for whisper-island detection. LPR has been used in many speech processing applications, such as Speaker-ID[4], VAD[5]. However, in this study, we formulate an algorithm which can both locate and identify whispered speech islands embedded within a neutral audio stream using an new entropy-based feature extracted from LPR. The new feature is integrated into a BIC/ T^2 -BIC segmentation algorithm for detecting vocal effort change points between whispered and neutral speech. The measurement strategy named Multi-Error Score(MES) proposed in [3], is used to evaluate performance of vocal effort change detection. In the final stage, a GMM based classifier trained with speech data using the new proposed feature is developed to address the problem of whisper-island detection. The remainder of this paper is organized as follows. First, details of the new proposed feature is addressed in Sec. 2. Next, the corpora developed for this study is introduced in Sec. 3. In Sec. 4, the baseline routine and BIC/ T^2 -BIC algorithm for whisper-island detection are presented. Evaluations using two whisper/Vocal Effort corpora are presented in Sec. 5. Finally, discussion and conclusions of this study are presented.

2. FEATURE FORMULATION

If we consider the speech production mechanism, the speech signal can be represented as a result of the excitation of the vocal tract. Under the framework of LP analysis, the vocal

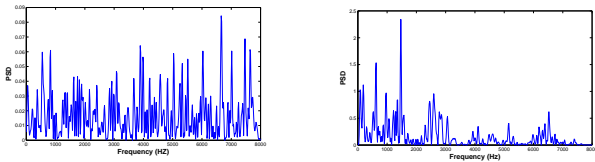
This project was funded by AFRL under a subcontract to RADC Inc. under FA8750-05-C-0029. Recently approved for public release; distribution unlimited.

tract is associated with an overall shaping filter and the excitation to the residual signal. LP analysis consists of estimation of the LP coefficients by minimizing the prediction error. Under traditional LP analysis, theoretically, the linear prediction error(LP-error) is uncorrelated with the speech signal and related to the excitation. With knowledge that the fundamental difference between whisper and neutral speech is the specific structure of the excitation, specifically the absence of vocal fold vibration in whisper, the LP-error can be potentially used in differentiating whisper and neutral speech. Fig. 1.a & 1.b show the LP residual signal for a 20ms whisper and a 20ms neutral speech respectively. The power spectrum of the LP-error signal for the frame shown in Fig. 1.a & 1.b are respectively shown in Fig. 2.a & 2.b(Vowel segment, 16 kHz sampling rate, 13th order LP analysis).



a.LP-error for whisper speech b.LP-error for neutral speech

Fig. 1. The LP-error for a 20 ms frame .



a.PSD for whisper LP-error b.PSD for neutral LP-error

Fig. 2. The PSD for the LP-error of a 20 ms frame .

It is easy to observe that the LP-error of neutral speech is more harmonic than that of whisper, which indicates the excitation differences of whisper and neutral speech. Furthermore, the spectral tilt of the LP-error for whisper and neutral speech are different as well, with whisper being more flat than for neutral. Considering previous whisper research[1][2][3], a 4-D feature vector for each 20ms LP-error frame can be formulated as follows:

$$\begin{bmatrix} \text{1-D spectral information entropy(ER)[3];} \\ \text{2-D spectral information entropy(SIE)[3];} \\ \text{-(1-D spectral tilt(ST)[1]).} \end{bmatrix} \quad (1)$$

ER and SIE calculation can be illustrated in Fig. 3 and Fig. 4 respectively. The spectral information entropy(SIE) is obtained as follows. Assuming $X(k)$ is the power spectrum of speech frame $x(n)$, k varies from k_1 to k_M in a sub-band; then that portion of the frequency content in the k band versus the entire response is written as,

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

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

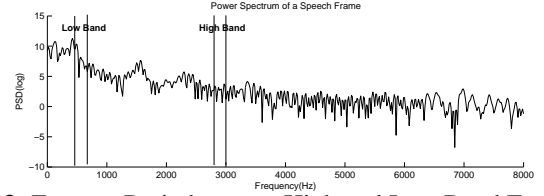


Fig. 3. Entropy Ratio between High and Low Band Frequencies

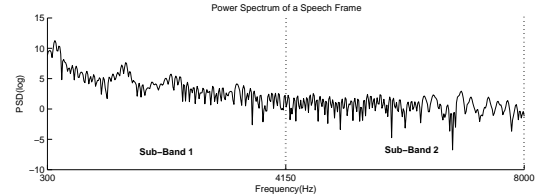


Fig. 4. Two Bands over Frequency Domain

the sub-band can then be calculated as,

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

Since the ER and SIE are all non-zero values, the spectral tilt value will generally be below zero, and so to effectively perform segmentation with T^2 -BIC [6], the sign-inverted spectral tilt value was formulated as the 4th-D of the feature vector.

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.

3.1. UT-VocalEffort I

For UT-VE I, a total of 12 male, native English-speaking subjects participated in data collection. For each subject, speech was recorded for a range of tokens using three microphones: a throat microphone, a close-talking and a far field microphone. A 1 kHz sinusoid signal generated by an NTI analog audio generator was played by an ALTEC speaker as the calibration test tone in all recordings. At the beginning of each token, the volume of the test tone was carefully adjusted to ensure the sound pressure level(SPL) of the test tone measured 75dB using a QUEST sound level meter(SLM). The test tone was recorded with all three microphones. The position of the subject, the location of the calibration test tone speaker, and the location of the sound level meter were all positioned in an equidistant triangle separated by 75cm. The data collection

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

3.2. UT-VocalEffort II

In addition to UT-VE I, a much larger corpus named UT-VE II was developed in the same environment as in UT-VE I. Whisper and neutral speech from 37 male and 75 female subjects were collected. Unlike corpus UT-VE I that focuses on five vocal efforts, corpus UT-VE II focuses on neutral speech embedded with whispered speech. In UT-VE II, the subject was required to read materials in either neutral or whispered modes. 41 TIMIT sentences were produced alternatively in neutral and whisper mode, with the 14th and 15th sentences were both read in neutral mode. In this study, the speech data produced with close-talking microphone in UT-VE I&II were used for analysis and tests.

4. SYSTEM DESCRIPTION

4.1. Baseline System

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. 5. The po-

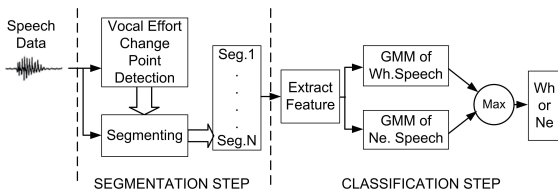


Fig. 5. Flow Diagram of Whisper-Island Detection.(Wh: whiser island, Ne: neutral audio block)

tential 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. 5). Based on the sequence of potential detected VECPs, the speech stream is divided into segments. In this study, 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[6], 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. In this study, the improved BIC/ T^2 -BIC algorithm is considered as a potential effective method to detect the VECPs between whisper and neutral speech if an effective feature sensitive to vocal effort change is employed.

In the classification step, 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.

4.2. BIC/ T^2 -BIC Algorithm

In [6], 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 $[(\hat{b} - 50), (\hat{b} + 50)]$ after the T^2 statistic algorithm is used to detect the possible VECP \tilde{b} ,

$$\hat{b} = \arg \max_{(\hat{b}-50) < b < (\hat{b}+50); BIC(b) > 0} BIC(b). \quad (4)$$

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 2-3 Sec. in duration, the BIC algorithm is used for process window, L_w less than 5 Sec., and T^2 -BIC is used when L_w is larger than 5 Sec. The implementation of the overall proposed segmentation algorithm for vocal effort change point(VECP) detection is described in Fig.6.

5. EVALUATION RESULTS

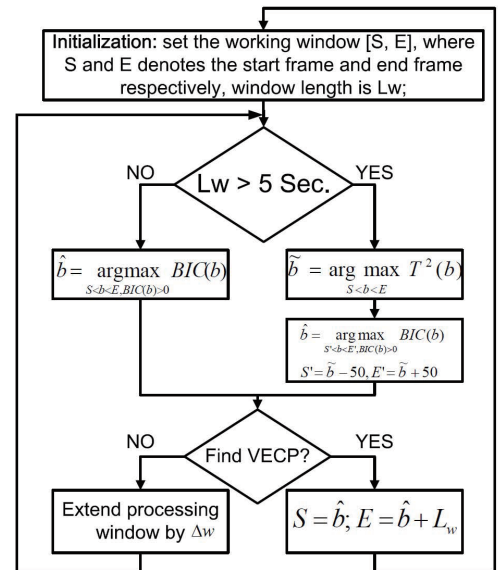


Fig. 6. Segmentation based on BIC/ T^2 -BIC processing algorithm for VECP detection

5.1. Short Introduction of Multi-Error Score

In [3], the Multi-Error Score(MES) was developed and introduced to evaluate performance of acoustic features for detection of VECPs. The MES consists of 3 error types for segmentation mismatch: miss detection rate, false alarm rate and average mismatch in milliseconds normalized by dual-segment duration. Fig. 7 illustrates these three types of error.

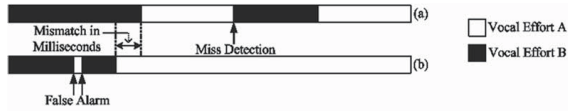


Fig. 7. Three Types of Segmentation Error

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

$$MES = False\ Alarm\ Rate + 2 \times Mismatch\ Rate + 3 \times Miss\ Detection\ Rate \quad (5)$$

The mismatch rate is obtained by calculating the percentage of the mismatch in milliseconds versus the total duration of the two segments corresponding to the actual breakpoints. More details concerning the MES can be found in [3]. 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

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 enrolled in experiments. The transcript files of these audio streams were used to compare with VECP detection results obtained from the BIC/T²-BIC algorithm using different features, so that the MES can be calculated. The lower MES denotes better performance in VECP detection. The experimental results in MES are shown in Table 1 for the 7-D feature set extracted from speech frame[3], and proposed 4-D feature extracted from LPR frame with BIC/T²-BIC algorithm. The reduction in MES from 43.07 to 20.70 is quite remarkable. Noting that even an increase in FAR can be addressed by merging successive segments of identical vocal effort in the classification step.

Table 1. Evaluation for Vocal Effort Change Points Detection

Feature Scenario	MDR(%)	FAR(%)	MMR(%)	MES
7-D feature	10.58	5.25	3.04	43.07
4-D feature	0.43	14.55	2.43	20.70

5.3. Experimental Results of System

With an extremely low MES in VECP detection, the proposed 4-D feature based on the LP residual shows better perfor-

mance in sensing vocal effort changes between whisper and neutral speech versus the previous 7-D feature set. Overall system performance using the proposed 4-D feature was compared with the previous system using the 7-D feature set[3]. The same audio streams used in last subsection were employed here. Thus, with 20 whisper-islands for each audio stream, there are 1182 potential whisper-islands in total for detection. The GMM based vocal effort classifier was trained with whisper and neutral speech from 12 male subjects from UT-VE I. Experimental results of both systems are illustrated in Table 2. There is significant improvement in whisper-island detection performance for the new 4-D feature set.

Table 2. Evaluation for Overall Whisper Island Detection

System	Detection Accuracy(%)
System on 7-D feature	95.33
System on 4-D feature	97.37

6. DISCUSSION AND CONCLUSION

Effective whisper island detection is the first step necessary for engagement of effective subsequent speech processing steps to address whisper. In this study, the LP residual was used in representing the vocal effort difference between whisper and neutral speech, a critical problem since whisper speech contains no vocal fold excitation whatsoever. The proposed 4-D feature set extracted from LPR was developed, along with a novel BIC/T²-BIC segmentation algorithm that resulted in the best MES score to date on VECP detection. To date, this represents the best solution available in the literature for whisper-island detection. Finally, GMMs trained with the same 4-D feature set also showed outstanding(97.37% vs. 95.33%) performance for whisper detection. This advancement is an important step towards addressing mixed vocal effort(neutral/whisper) in conversational speech for applications such as speech recognition or speaker ID in telecommunications. For further work, it is possible to merge the proposed feature with other features to reduce false alarms in VECP detection between whisper and neutral speech.

7. REFERENCES

- [1] C. Zhang and J.H.L. Hansen, "Analysis and classification of speech mode: Whispered through shouted," *INTERSPEECH07*, 2007.
- [2] Cupples J. Wennndt, S.J. and M. Floyd, "A study on the classification of whispered and normal phonated speech," *INTERSPEECH02*, 2002.
- [3] C. Zhang and J.H.L. Hansen, "Advancements in whisper-island detection within normally phonated audio streams," *INTERSPEECH09*, 2009.
- [4] B.Gas J.L.Zarader M. Chetouani., M.Faundez-Zanuy, "Investigation on lp-residual representations for speaker identification," *Pattern Recognition*, vol. 42, pp. 487-494, 2009.
- [5] R. Goubran E. Nemer and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the lpc residual domain," *Trans. on Speech & Audio*, pp. 9(3):217-231, 2001.
- [6] 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, July 2005.