MICROPHONE ARRAY PROCESSING FOR DISTANCE SPEECH CAPTURE: A PROBE STUDY ON WHISPER SPEECH DETECTION

Chi Zhang, Tao Yu *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;txy073000;john.hansen}@utdallas.edu

ABSTRACT

In this study, we develop a probe system for whisper-island detection for distance speech capture using a microphone array technique. The developed corpus consists of distance speech in neutral vocal effort and embedded with whisper speech which are produced at different distances for this study. The microphone array beamforming technique is used to enhance the distance speech before being processed by the whisper-island detection system. The enhanced distance speech provides better vocal effort change point detection results, which is indicated by either 0.0% miss detection rate or lower Multi-Error Score(MES) compared to unprocessed distance speech. The final whisper-island detection results produce a higher detection rate and lower false alarm rate for enhanced speech, which illustrates the improvement of whisper-island detection for distance speech attained from microphone array processing.

***Index Terms*—** Distant speech, Microphone Array, Whisper-Island Detection, BIC/T^2-BIC, GMM Classifier

1. INTRODUCTION

Distance based speech acquisition via a microphone array is a viable approach for speech recognition, speech enhancement, and speaker identification. Although microphone array beamforming techniques have been widely and extensively studied, the optimality in the sense of minimal detection error is seldom explored. In this study, the performance of vocal effort detection, for example the whisper-island detection, using a microphone array solution is the subject for investigation.

Whisper speech is one mode of natural speech communication which results in reduced perceptibility and a significant reduction in intelligibility. 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 for speech processing systems[4]. Therefore, detecting and identifying whispered islands embedded in the speech signal(specially distant 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.

In this study, the microphone array enhancement technique is merged with whisper-island detection algorithm, which is introduced in [7], to detect whisper-island within distance based neutral speech. The 4-D entropy-based Whisper IDentification(WhID) feature[7] was modified to adapt the spectrum changes caused by channel differences between distant speech and close speech. The remainder of this paper is organized as follows. First, the framework for distant whisper-island detection are presented in Sec. 2. Next, in Sec. 3 the corpus used in this study is briefly described. Experiments and evaluations of the framework are presented in Sec. 4. Finally, discussion and conclusions are presented.

2. FRAMEWORK FOR WHISPER-ISLAND DETECTION WITHIN DISTANCE SPEECH

The high-level framework of whisper-island detection within distance speech is illustrated in Fig. 1.

The distance speech was collected by a 6-channel microphone array, then processed and enhanced by beamforming technique to form a enhanced speech signal. Then the enhanced distance speech signal was fed into the whisper-island detection system introduced in [7] to detect the whispered speech within the distance speech.

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.

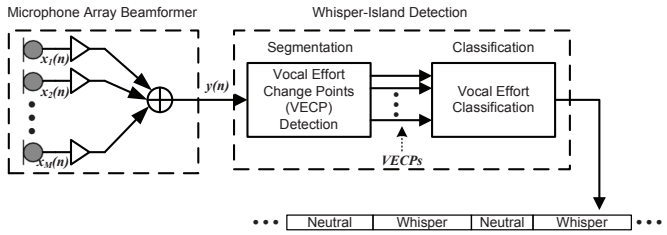

Fig. 1. High-Level Flow Diagram of Whisper-Island Detection of Distance Speech.

2.1. Microphone Array front-end

Microphone array speech signal processing provides an effective method of spatial filtering. Analogous to a temporal filter which processes data collected over a temporal aperture, a spatial filter processes data received over a spatial aperture, and filters signals and interference originating from separate spatial locations.

Consider the case where the desired speech impinges on a array of M microphones. Taking the Short-Time Fourier Transform (STFT) of the time domain signal, the signal model in each time-frame and frequency-bin can be written as,

$$\boldsymbol{x}(t,k) = \boldsymbol{a}(t,k)s(t,k) + \boldsymbol{n}(t,k), \tag{1}$$

where $\boldsymbol{x} \in C^{M\times 1}$ is the array observation data vector, $s \in C$ is the desired speech, $\boldsymbol{a} \in C^{M\times 1}$ is the unknown (maybe time-varying) array steering vector, $\boldsymbol{n} \in C^{M\times 1}$ is the noise (background noise plus interference) vector, and t and k are the time-frame index and frequency-bin index, respectively. In general, we can process each frequency-bin independently; thus, the notation of the frequency-bin index k is be omitted for briefness.

Assuming that each vector components of the model in Eq.(1) are mutually uncorrelated, the autocorrelation matrix for the observed data vector can be expressed as,

$$\begin{aligned} R_{xx} = E\{\boldsymbol{x}(t)\boldsymbol{x}(t)^H\} &= R_{ss} + R_{nn}, \\ &= \sigma_s^2 \boldsymbol{a}\boldsymbol{a}^H + R_{nn}, \end{aligned} \tag{2}$$

where R_s and R_{nn} are the autocorrelation matrices for the desired speech and noise, respectively; and σ_s^2 is the power of the desired speech.

For a single frequency-bin, the optimal beamformer is a linear processor (filter) consisting of a set of complex weights [8]. The output of the beamformer is an estimate of the desired signal and is given by,

$$y(t) = \hat{s}(t) = \boldsymbol{w}^H \boldsymbol{x}(t). \tag{3}$$

The weights are chosen according to some optimization criterion, such as the Minimum Mean Square Error (MMSE), the Minimum Variance Distortionless Response (MVDR), or the Maximum Signal-to-Noise Ratio (Max-SNR). Generally, the optimal weights have the same structure [8], as:

$$\boldsymbol{w}_o = \mu R_{nn}^{-1} \boldsymbol{a}, \tag{4}$$

where μ is a scale factor decided by the optimization criterion. In this study, the distortionless enhancement is employed and henceforth μ is obtained using MVDR criterion as

$$\mu = \frac{1}{\boldsymbol{a}^H R_{nn}^{-1} \boldsymbol{a}}. \tag{5}$$

2.2. Whisper-island Detection Algorithm

The algorithm for whisper-island detection proposed in [7] consists of two main algorithmic steps: segmentation and classification. The structure of the algorithm is illustrated in Fig. 2. The potential vocal effort change points(VECPs)

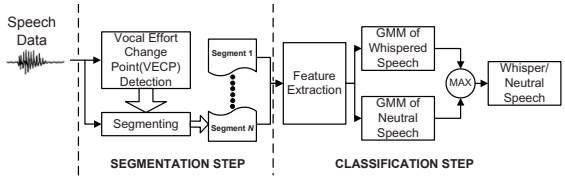

Fig. 2. Flow Diagram of Whisper-Island Detection Framework I.

of the input speech data embedded with whisper-islands are first detected in the segmentation step (left part of Fig. 2). Based on the sequence of potential detected VECPs, the speech stream is divided into segments. In this study, an improved T^2-BIC[9] algorithm, BIC/T^2-BIC[6] is incorporated to detect the potential VECPs between whisper and neutral speech. The T^2-BIC algorithm, developed by Zhou and Hansen[9], is an unsupervised model-free scheme that detects acoustic change points based on the input feature data. One pre-request assumption for applying this algorithm is that the feature employed by the BIC/T^2-BIC algorithm is considered to be sensitive for vocal effort changes between whisper and neutral speech. 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.

2.2.1. Modification of WhID Feature

In [7], the proposed 4-D feature set WhID was formulated as follow for each 20ms speech frame:

$$\begin{bmatrix} \text{1-D spectral information entropy ratio(ER);} \\ \text{2-D spectral information entropy(SIE);} \\ \text{1-D spectral tilt(ST).} \end{bmatrix} \tag{6}$$

In [7], the ER feature was calculated between frequency bands 450-650Hz and 2800-3000Hz. SIE feature was calculated within frequency band of 300-4150Hz and 4150-8000Hz. However in this study, due to the change of channel property of distance speech and the sampling frequency of corpus(8kHz instead of 16kHz), although the WhID feature set remains the same 4-D structure, some modification had been made to WhID feature set: the ER feature was calculated between 450-900Hz and 1000-1450Hz; the SIE feature was calculated within 350-1000Hz and 1800-2500Hz. The ST calculation has remained the same as in [7].

3. CORPUS DESCRIPTION

In this study, a corpus consisting of neutral and whispered speech produced at different distances was developed using a close-talking microphone and a 6-channel microphone array. The corpus was collected in a rectangular conference room using a 8-channel TASCAM US-1641 USB audio interface with two SHURE PG185 condenser lavalier microphone and a 6-channel microphone array. One data collector and one subject wearing condenser lavalier microphones participated and would sit face-to-face over a long table in each collection session. Each session consists of (i) conversation part and (ii) reading part. In the conversation part, the subject answered 9 questions asked by the data collector, but chose 3 of 9 questions to be answered in whisper, and the rest of them answered in neutral speech. The reading part required the subject to read 40 TIMIT sentences alternatively in whisper and neutral mode and then read 4 paragraphs in neutral mode with several sentences and phrases in those paragraphs in whisper. The whole session of conversation and reading will repeated at 1 meter, 3 meter and 5 meter positions along the long side of the table. The setting of the data collection can be illustrated in Fig. 3

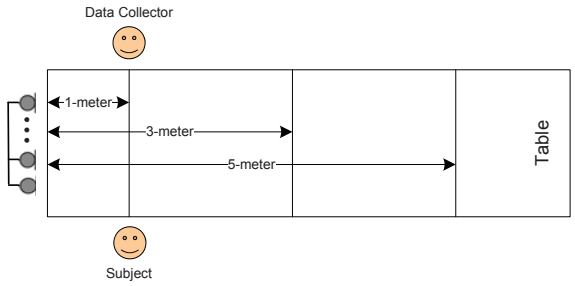

Fig. 3. Table Setting of Data Collection.

All conversations and readings of the collection were recorded by a 6-channel microphone array located at the center of the short side edge of the table and the two lavalier microphones worn by data collector and subject. A digital video recording was also performed. The ultimate aim of this corpus is to enroll 30 subjects(15 males and 15 females) in data collection.

4. EVALUATION RESULTS

4.1. Brief Overview of Multi-Error Score

In [6], 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. 4 illustrates these three types of error.

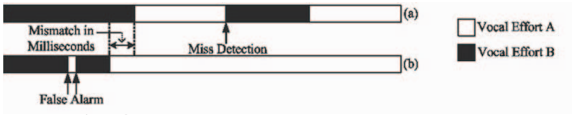

Fig. 4. Three Types of Segmentation Error

The calculation of MES can be illustrated by the Eq. 7.

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

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 [6]. 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%.

4.2. Experimental Results in MES

The audio streams consists of 40 TIMIT sentences alternatively read in whisper and neutral mode by 10 females subjects at different distances were used in this study to explore the effect of the microphone array on whisper-island detection for distance speech. Each audio stream was manually labeled for VECPs and vocal efforts in transcript files. 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. The speech audio recorded from lavalier microphone worn by subject, from one channel of microphone array and the enhanced speech audio were experimented for whisper-island detection. Thus in total 9 experimental scenarios were performed in this study. Since the speech signal collected from lavalier microphone worn by subject denotes the best quality of signal in our corpus, the experimental results obtained by using lavalier microphone signal indicates how good the performance of VCEP detection can attain. The experimental results in MES are shown in Fig. 6 for each scenario. The

MDR	Channel		
	Subject	2	Enhanced
1-m	0.0	0.24	**0.0**
3-m		0.77	0.77
5-m		1.03	**0.51**

FAR	Channel		
	Subject	2	Enhanced
1-m	13.4	23.88	28.59
3-m		25.28	**22.28**
5-m		32.64	**31.32**

MES	Channel		
	Subject	2	Enhanced
1-m	23.56	35.87	40.78
3-m		39.75	**37.26**
5-m		49.01	**46.03**

MMR	Channel		
	Subject	2	Enhanced
1-m	5.08	5.62	6.09
3-m		6.07	6.33
5-m		6.65	**6.58**

Fig. 5. Multi-Error Score for Each Experimental Scenario

(i) channel subject, (ii) channel 2 and (iii) channel enhanced, represent the speech signal used for VECP detection are from the lavalier microphone worn by the subject, the 2nd microphone of microphone array, and the enhanced speech signal respectively. It can be observed in Fig. 6 that the although in the 1 meter case, the MES of enhanced speech signal is not smaller than the MES of signal from channel 2, the MDR of channel enhanced is 0.0 which denotes all the VECP had been detected. Furthermore, as long as the distance increases, the enhanced signal has a better result with smaller MES compared with channel 2. In addition, it should be noted that the relative amount of false alarm VECP can be compensated by merging successive segments of identical vocal effort in the classification step.

4.3. Experimental Results of System

In the classification step, GMM training are using a rolling-robin scheme to obtain open test speaker and open set test speech. Since the close talk speech data is the most frequent data in the research area(especially for whispered speech), the speech data from lavalier microphone are used for GMM training for the model of vocal effort of neutral and whispered speech. The overall system performances of 9 scenarios were compared based on the detection rate(DR) and false alarm rate(FAR) of whisper-island within neutral speech audio streams, which are formulated as Eq. 8 & 9 respectively.

$$\text{Detection Rate} = \frac{\text{Correct \#}}{\text{True \#}} \times 100\% \tag{8}$$

$$\text{False Alarm Rate} = \frac{\text{False Alarm \#}}{\text{Total Detected \#}} \times 100\% \tag{9}$$

The same audio streams used in last subsection were employed here. Thus, with 20 whisper-islands for each audio stream, there are 200 potential whisper-islands in total to detect for each scenario. Experimental results of all 9 experimental scenarios are presented in Fig. 4.3 It can be observed

DR (%)	Channel		
	Subject	2	Enhanced
1-m	95.09	87.2	**92.42**
3-m		61.5	**74.5**
5-m		61.5	**66.00**

FAR (%)	Channel		
	Subject	2	Enhanced
1-m	4.12	11.96	**8.02**
3-m		28.07	**19.46**
5-m		30.51	**27.07**

Fig. 6. Detection Rate and False Alarm Rate for Each Experimental Scenario

that, the detection rates of enhanced speech are higher than that of channel 2 speech at all three distances. The FARs of enhanced speech are less than the FARs of channel 2 speech as well.

5. DISCUSSION AND CONCLUSION

From the MES results in Fig. 6, it can be concluded that as long as the distance increases, the microphone array enhanced speech has a lower MES, which indicates the microphone array enhancement can improve VECP detection performance. Furthermore, from the detection rate and FAR results in Fig. 4.3, we can conclude that the microphone array enhancement improves whisper-island detection. The informal subjective listening test also indicated that the enhanced speech has better subjective quality than the unprocessed speech from channel 2.

In this study, we formulated a probe system consisting of microphone array technique and whisper-island algorithm for whisper-island detection of distance speech and experimented the system performance for whisper embedded speech from different distances. The decreased MES and increased DR indicate that the microphone array technique can improved the whisper-island detection for distance speech. Although the improvements are encouraging, they cannot achieve the detection results as effective as that for the close-talking microphone.

6. REFERENCES

[1] J. Koufman, "The spectrum of vocal dysfunction," *The Otolaryngologic Clinics of North America. Voice Disorders*, vol. 24(5), pp. 985–988, Oct. 1991.

[2] 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.

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

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

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

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

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

[8] H. L. Van Trees, *Optimum Array Processing*, Wiley, New York, 2002.

[9] 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.