Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 66 (2015) 107–117

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

An advanced entropy-based feature with a frame-level vocal effort likelihood space modeling for distant whisper-island detection [☆]

Chi Zhang, John H.L. Hansen [*]

Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering and Computer Science, Dept. of Electrical Engineering, University of Texas at Dallas, 2601 N. Floyd Road, EC33, Richardson, TX 75080-1407, USA

Received 17 October 2013; received in revised form 11 August 2014; accepted 9 September 2014
Available online 26 September 2014

Abstract

A challenging research problem which has received limited attention in the speech research community is whisper-island detection. Effective whisper island, or VECP-Vocal Effort Change Point, detection is the first step needed to ensure the engagement of effective subsequent speech processing steps to address whisper. In this study, we first propose an improved entropy-based feature from a previous study which is integrated within a model-less whisper-island detection algorithm. The improved 3-D WhID feature shows better discrimination properties between whisper and neutral speech, resulting in a 0.00% MDR (miss detection rate), lower FAR (false alarm rate), MMR (mismatch rate) and collectively a reduced MES (multi-error score). With improved VECP detection results and no need for a prior trained GMM, the BIC-based vocal effort clustering algorithm attains a 100% detection rate of whisper-islands. In this study, a more challenging task of distant whisper-island detection is also addressed using a proposed frame-based vocal effort likelihood space modeling algorithm (model-base). A corpus named UT-VE-III consisting of spontaneous and read whisper embedded neutral speech using a microphone array from various distances in a real-world conference room is developed. For the whisper embedded neutral speech of UT-VE-III at 1-m, 3-m and 5-m distance using a Lavalier microphone and distant microphone, the proposed algorithm sustains consistent performance for VECP detection and whisper classification rates.

Keywords: Vocal effort; Distant whisper; Detection; BIC; T^2-BIC; Segmentation; Clustering

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 vibration/excitation, speech content and structure is significantly altered. It is noted that whispered speech may be intentional, or caused by a temporary or long-term change in the vocal fold structure, or muscle control due to disease of the vocal system, such as functional aphonia (Koufman, 1991), laryngeal cancer (Gavidia-Ceballos and Hansen, 1996). 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 and therefore from being overheard by remote listeners in public settings (Ito et al., 2005). 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

[☆] This project was funded by AFRL under Contract FA8750-12-1-0188 and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen.

[*] Corresponding author.

http://dx.doi.org/10.1016/j.specom.2014.09.004
0167-6393/© 2014 Elsevier B.V. All rights reserved.

shouted, has the most dramatic loss in terms of vocal effort for speech processing systems (Zhang and Hansen, 2007). 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.). For example, an automatic phone answering system needs to detected the whispered personal information, such as credit card or social security number which should not be overheard by other people, such that it can deploy the corresponding ASR/Spkr-ID algorithm for whispered speech to recognize the needed information. In another case, if a phone call or VoIP call is made under the circumstances that loud/normal volume speech is prohibited, after detecting the speech is in whispered mode, the corresponding speech enhancement or speech conversion algorithm can be performed to transform the whispered speech to be more intelligible in the other side of the call. In fact, several algorithms have been developed to compensate for the negative influence of whispered speech on speech recognition (Ito et al., 2005) or speaker identification applications (Fan and Hansen, 2011; Fan and Hansen, 2009) while assuming knowledge of the location of whispered speech within an audio stream. Furthermore, whispered speech has a high probability of conveying confidential or sensitive information. For spoken document retrieval or in call center monitoring, detection and identification of whispered islands in speech 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 (Zhang and Hansen, 2011a; Wenndt et al., 2002; Zhang and Hansen, 2010). In Zhang and Hansen (2011a), 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 (VECP) detection and vocal effort classification.

In this study, the previous WhID feature set is analyzed and improved, followed by a combination with a newly proposed model-less whisper-island detection algorithm. To further explore the task of distant whisper-island detection, a new framework deploys the proposed discriminative feature set entitled “Vocal Effort Likelihood(VEL)” to detect the vocal effort change point between distant whisper and neutral speech, and therefore detect the distant whisper-island. It is noted that distance based speech processing of normally phonated speech remains fundamentally different than distance based whispered speech processing, since again the whispered speech is absent of all periodic excitation and the audio signal is further reduced in amplitude due to distant speech capture. The remainder of this paper is organized as follows. First, the algorithm proposed in Zhang and Hansen (2011a) is reviewed in Section 2. The improvement of the WhID feature set and the proposed model-less whisper-island detection algorithm are introduced in Section 3. Later in Section 4, the performance of improved WhID feature set and the model-less whisper-island detection algorithm are evaluated through experimental results. Further exploration of distant whisper-island detection algorithm (model-base) and evaluation are illustrated in Sections 5 and 6 respectively. Finally, discussion and conclusions are presented.

2. Review of previous work

In Zhang and Hansen (2011a,b), an algorithm was proposed to detect whisper-islands embedded within an audio stream of neutral speech. The two steps of the algorithm in Zhang and Hansen (2011a) is illustrated in Fig. 1.

In the segmentation step, an entropy based speech feature is deployed using a model-less BIC/T^2-BIC segmentation algorithm (Zhou and Hansen, 2005; Huang and Hansen, 2006) to detect the vocal effort change points (VECP) between whispered and neutral speech. The proposed BIC/T^2-BIC algorithm is a feature-based acoustical change detector. Based on the discriminative property of the input acoustical feature, the BIC/T^2-BIC algorithm can detect the corresponding acoustical changes within the audio stream (Zhang and Hansen, 2011a; Zhou and Hansen, 2005; Huang and Hansen, 2006). In Zhang and Hansen (2011a), the proposed feature set WhID is used to detect the vocal effort changes between whispered and neutral speech. The construction of the WhID feature set can be shown in Eq. (1).

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

Later in the classification step, a GMM based vocal effort classifier is used to identify the vocal effort of the segments obtained in the segmentation step. The acquisition of speech data with consistent vocal effort is required for training of the GMMs of vocal efforts in whisper and neutral.

3. Model-less whisper-island detection algorithm

3.1. The formulation of WhID feature set

In Zhang and Hansen (2011a), experimental results showed that the proposed WhID feature set is discriminative between vocal efforts of whisper and neutral speech. However, the discriminative property of WhID has not been analyzed.

For WhID, the two frequency bands (450–650 Hz and 2800–3000 Hz) for the 1-D entropy ratio calculation are suggested in Wenndt et al. (2002) to calculate the energy ratio for differentiating whisper and neutral speech. The