Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 55 (2013) 119–134

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Acoustic analysis and feature transformation from neutral to whisper for speaker identification within whispered speech audio streams ☆

Xing Fan, John H.L. Hansen *

Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, Richardson, TX, USA

Received 8 August 2011; received in revised form 6 July 2012; accepted 24 July 2012
Available online 25 August 2012

Abstract

Whispered speech is an alternative speech production mode from neutral speech, which is used by talkers intentionally in natural conversational scenarios to protect privacy and to avoid certain content from being overheard or made public. Due to the profound differences between whispered and neutral speech in vocal excitation and vocal tract function, the performance of automatic speaker identification systems trained with neutral speech degrades significantly. In order to better understand these differences and to further develop efficient model adaptation and feature compensation methods, this study first analyzes the speaker and phoneme dependency of these differences by a maximum likelihood transformation estimation from neutral speech towards whispered speech. Based on analysis results, this study then considers a feature transformation method in the training phase that leads to a more robust speaker model for speaker ID on whispered speech without using whispered adaptation data from test speakers. Three estimation methods that model the transformation from neutral to whispered speech are applied, including convolutional transformation (ConvTran), constrained maximum likelihood linear regression (CMLLR), and factor analysis (FA). a speech mode independent (SMI) universal background model (UBM) is trained using collected real neutral features and transformed pseudo-whisper features generated with the estimated transformation. Text-independent closed set speaker ID results using the UT-VocalEffort II corpus show performance improvement by using the proposed training framework. The best performance of 88.87% is achieved by using the ConvTran model, which represents a relative improvement of 46.26% compared to the 79.29% accuracy of the GMM-UBM baseline system. This result suggests that synthesizing pseudo-whispered speaker and background training data with the ConvTran model results in improved speaker ID robustness to whispered speech.

Keywords: Speaker identification; Whispered speech; Vocal effort; Robust speaker verification

1. Introduction

Whispered speech is a natural speech production mode, employed in public situations in order to protect privacy and to avoid certain content from being made public. For example, a customer might whisper to provide information regarding their date of birth, credit card information, and billing address in order to make hotel, flight, or car reservations through a machine interface over the telephone, or a doctor might whisper when entering a voice memo in order to discuss patient medical records in public. Aphonic individuals, as well as those with low vocal

☆ This project was funded by AFRL through a subcontract to RADC Inc. under FA8750-09-C-0067 (Approved for public release, distribution unlimited), and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. Hansen.

* Corresponding author. Address: 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. Tel.: +1 972 883 2910; fax: +1 972 883 2710.

E-mail address: John.Hansen@utdallas.edu (J.H.L. Hansen).

URL: http://crss.utdallas.edu (J.H.L. Hansen).

0167-6393/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2012.07.002

capability, such as heavy smokers, also employ whisper as a primary form of oral communication. In this study, the term "neutral speech" refers to speech produced at rest in a quiet sound-booth whose "voiced" phonemes, such as sustained vowels, contain glottal based vocal fold movement that represents periodic excitation.

There are significant differences between whisper and neutral speech production mechanisms, which result in the absence of voiced excitation, shifted formant locations and change in formant band width (Ito et al., 2005; Zhang and Hansen, 2007; Morris and Clements, 2002; Matsuda and Kasuya, 1999; Jovicic, 1998). Zhang and Hansen (2007) revealed that the change of vocal effort in test data ranging from whisper through shouted has a significant impact on automatic speaker identification (speaker ID) performance, with whisper resulting in the most serious loss in performance. Similar results were reported in other studies on automatic speech recognition (Ito et al., 2005) and speaker recognition (Jin et al., 2007) systems as well.

Past work on automatic speaker ID systems for whispered speech can be grouped into two main categories: front-end processing (Fan and Hansen, 2009; Fan and Hansen, 2008) and back-end model adaptation (Jin et al., 2007). Both methods have resulted in improvements in system accuracy. However, new front-end processing methods involve feature re-extraction and model re-training for neutral speech, which increases computational requirements and may hurt system performance on neutral test speech. For back-end model adaptation, as in Jin et al. (2007), a simple maximum a posteriori (MAP) adaptation of the original model trained with neutral speech can provide satisfactory performance under the prerequisite of a fair amount of speaker-dependent (SD) whispered adaptation data. However, in real applications, whispered adaptation data from test speakers is generally not available. Also, while it is possible to collect additional whispered data from other speakers, the fact that the total amount of real whispered data is usually much smaller compared with the available neutral data means that it is still very difficult to train a speech mode independent (SMI) universal background model (UBM). Therefore, the focus of this study is to explore efficient model training techniques that rely solely on a limited set of whisper data from *non-target speakers* for modeling whispered speech. In this study, *non-target speakers* are those speakers whose speech is not seen in the test set for closed-set speaker ID.

A similar strategy was first considered by Bou-Ghazale and Hansen (1998), where HMMs were used to statistically model characteristics needed for generating pitch contour and spectral slope patterns in order to modify the speaking style from neutral to stressed speech. In this study, the statistical information contained in a UBM trained with whispered data set collected from *non-target speakers* is employed for a transformation estimation to generate whisper features from neutral data. The convectional Mel-frequency cepstral coefficients (MFCCs) (Davis and Mermelstein, 1980), which are employed for most state-of-the-art speech systems, are used here as the front-end features throughout this study and our compensation is applied in the corresponding MFCC domain. The generated whispered features will be referred as "pseudo-whisper features" in the rest of this study.

Formulating a model training method for this task requires understanding two critical facets of the problem. One is the difference between whispered and neutral speech in the resulting front-end feature domain. In particular, the MFCCs represent information regarding the smoothed spectral envelope in the Mel domain, hence, the differences between whisper and neutral in the linear frequency domain (Ito et al., 2005; Morris and Clements, 2002), might be distorted and represented in a different way in the Mel domain. The other facet is the consistency of the differences among speakers and phonemes. For example, if the spectral differences between whispered and neutral speech are consistent across speakers, a transformation estimated using whispered adaptation data from several *non-target speakers* could be applied directly to all whispered enrollment and test data for automatic speaker ID. On the other hand, if spectral differences between whispered and neutral speech are inconsistent across speakers (i.e., the way someone "whisper" may be speaker dependent), it is necessary to explore alternative methods that could estimate the particulars of a given enrollment speaker's whispered speech. If the spectral differences are phoneme or phoneme-class dependent, the problem will be even more complex since a unique mapping will be needed for each phoneme or phoneme-class. Past studies (Ito et al., 2005; Jovicic, 1998; Matsuda and Kasuya, 1999; Eklund and Traunmuller, 1996) provided comparison results for the average differences between whispered and neutral speech across phonemes in the linear frequency domain. However, those studies have not examined individual speaker differences in terms of the variations of those differences in the linear frequency or the Mel domain.

This study first compares the smoothed spectral envelope of whispered and neutral speech using a maximum likelihood transformation estimation. The dependence of the estimated transformation on speakers and phonemes is analyzed. Based on the analysis results, this study proposes a method that models the differences between whispered and neutral speech by a convolutional filter with zero mean additive noise (ConvTran). The parameters of the ConvTran transformation are estimated using a first order vector Taylor series (VTS) approximation and the expectation maximization (EM) algorithm. Pseudo-whisper features generated with the proposed ConvTran model are used to train a SMI-UBM, which will include equal amounts of neutral and pseudo-whispered speech. Also, because the proposed method keeps some level of speaker-dependent information in the resulting pseudo-whisper features, after the SMI-UBM is trained, a speaker dependent model can be further obtained by adaptation of the SMI-UBM with both neutral and selected pseudo-whisper features. Constrained maximum likelihood linear