Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 57 (2014) 87–100

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Detection of speaker individual information using a phoneme effect suppression method

Songgun Hyon[a,b], Jianwu Dang[a,c,d,*], Hui Feng[d,e], Hongcui Wang[a,d], Kiyoshi Honda[a,d]

[a] School of Computer Science and Technology, Tianjin University, China
[b] School of Computer Science, KimIlSung University, Democratic People's Republic of Korea
[c] School of Information Science, Japan Advanced Institute of Science and Technology, Japan
[d] Tianjin Key Laboratory of Cognitive Computing and Application, China
[e] School of Liberal Arts and Law, Tianjin University, China

Available online 25 September 2013

Abstract

Feature extraction of speaker information from speech signals is a key procedure for exploring individual speaker characteristics and also the most critical part in a speaker recognition system, which needs to preserve individual information while attenuating linguistic information. However, it is difficult to separate individual from linguistic information in a given utterance. For this reason, we investigated a number of potential effects on speaker individual information that arise from differences in articulation due to speaker-specific morphology of the speech organs, comparing English, Chinese and Korean. We found that voiced and unvoiced phonemes have different frequency distributions in speaker information and these effects are consistent across the three languages, while the effect of nasal sounds on speaker individuality is language dependent. Because these differences are confounded with speaker individual information, feature extraction is negatively affected. Accordingly, a new feature extraction method is proposed to more accurately detect speaker individual information by suppressing phoneme-related effects, where the phoneme alignment is required once in constructing a filter bank for phoneme effect suppression, but is not necessary in processing feature extraction. The proposed method was evaluated by implementing it in GMM speaker models for speaker identification experiments. It is shown that the proposed approach outperformed both Mel Frequency Cepstrum Coefficient (MFCC) and the traditional *F*-ratio (FFCC). The use of the proposed feature has reduced recognition errors by 32.1–67.3% for the three languages compared with MFCC, and by 6.6–31% compared with FFCC. When combining an automatic phoneme aligner with the proposed method, the result demonstrated that the proposed method can detect speaker individuality with about the same accuracy as that based on manual phoneme alignment.

Keywords: Speaker identification; Frequency warping; MFCC; Speech production; Phoneme-related effects

1. Introduction

Speaker recognition and speech recognition have separate goals and consequently focus on separate properties of utterances. In speech recognition, detected features emphasize linguistic information to strengthen semantic differences and suppress differences across speakers. The goal of speaker recognition, in contrast, is to extract a speaker's individual information from speech and suppress linguistic information (Campbell, 1997). For speaker recognition, the problem is how to extract and utilize the information that characterizes individual speakers. If we can extract speaker individual information correctly, it is not only useful for speaker recognition but also beneficial for speech recognition because such information can then be suppressed. In general, speaker information is strongly dependent on specific characteristics of the morphology of speech organs as well as idiosyncratic articulations. This

[*] Corresponding author at: School of Computer Science and Technology, Tianjin University, China.
E-mail addresses: h_star1020@yahoo.com (S. Hyon), jdang@jaist.ac.jp (J. Dang), fenghui@tju.edu.cn (H. Feng), hcwang@tju.edu.cn (H. Wang), khonda@sannet.ne.jp (K. Honda).

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.09.004

study, therefore, attempts to develop a novel method to extract speaker individual information by focusing on the factors of articulation as well as the vocal tract morphology.

1.1. Studies of speaker-specific features

Individual differences in speech sounds can be divided into sociolinguistic and biophysical categories. Both are realized by the human speech production system either for the sound source or the dynamic and/or static properties of the vocal tract (Dang and Honda, 1996a). For example, if we focus on the geometrical properties of the vocal tract, side branches such as the nasal cavity and piriform fossa contribute to speaker individuality. Thus, extraction of speaker information is equivalent to sorting out the acoustic components derived from these factors.

The speaker-specific sound production properties are naturally linked to human auditory function. The human auditory system has a non-linear response to sound in the frequency domain, with higher resolution in lower frequencies, and vice versa. To deal with this property, several non-linear frequency conversion methods have been proposed, and non-linear transformations such as Mel, Bark, and Equivalent Rectangular Bandwidth (ERB) are commonly applied to the preprocessing in speech recognition and speaker recognition. In particular, Mel-Frequency Cepstrum Coefficients (MFCC) have been widely used as feature parameters in both speech recognition and speaker recognition methods. The mapping between the physical frequency scale and Mel scale is approximately linear below 1 kHz and logarithmic above 1 kHz. A popular formula to convert the linear scale into a Mel scale is

$$F_{Mel} = 2595\,\lg(1 + F/700), \tag{1}$$

where F is physical frequency and F_{Mel} is Mel-frequency (O'Shaughnessy, 1987; Davis and Mermelstein, 1980) have shown that their speaker recognition system based on MFCC outperforms those systems based on such features as LFCC, LPCC, LPC and RC. The MFCC feature basically reflects the mechanism of the auditory nonlinear frequency resolution, which improves the representation robustness (Yu and Wang, 2005).Besides the Mel scale, the Bark scale (Zwicker and Terhardt, 1980) and ERB scale (Moore and Glasberg, 1996) have also been used. Although these three frequency scales are different in the shape of transformation curves, they have basically the same purpose, i.e., to make the short-time spectrum information conform to the characteristics of the human auditory system. In other words, the lower the frequency components emphasized in the original frequency resolution, the higher the frequency components compressed in the frequency axis, which is a key reason why MFCC parameters outperform those with linear frequency scales in speech recognition. Since the Mel nonlinear frequency scale is not designed to describe speech production properties, the superiority of MFCC in speaker recognition is not as prominent as that in speech recognition. Therefore, it is necessary to develop a set of new features to describe the properties of speech production.

For many years, researchers have attempted to develop physiologically related features for speaker recognition to overcome the limitations of MFCCs. Hayakawa and Itakura (1995) investigated the contribution of different frequency bands to individual properties by using LPC analysis and showed that speaker individual information exists in high frequency region. Miyajima et al. (1999) used a monotonic warping function slightly different from the Mel frequency warping function to process the speech spectrum, and gained some improvement, though the distribution of speaker individuality in the frequency domain is not monotonic. Orman and Arslan (2001) analyzed the contribution of different frequency sub-bands to speaker recognition performance and proposed a feature extraction method based on a set of filters. Miyajima et al. (2001) extracted features by using second-order all-pass function to normalize the frequency spectrum, and obtained some improvement. In order to efficiently represent a speaker's individuality in the short-time spectrum, Yu et al. (2008) proposed a non-linear frequency transform function and feature extraction algorithm which is based on the analysis of the contribution of the short-time spectrum in different frequency sub-bands to speaker recognition and on the technique of least-squares polynomial curve fitting.

Feature extraction associated with speaker individuality has also been studied based on the human speech mechanism. Kitamura and Saitou (2007) investigated perceptual contributions of frequency-dependent acoustic properties to speaker individuality, and they found that humans perceive speaker characteristics by focusing on more invariant acoustic properties. Lu and Dang (2008) adopted the Fisher's F-ratio method (Wolf, 1972) to explore the effects of vocal tract morphologies on speaker individual information. They found that the side branches of the vocal tract such as the piriform fossa providing relatively invariant acoustic properties. Their results confirmed that speaker information is not distributed uniformly in each frequency band, and demonstrated that the adaptive frequency scale transformation based on the F-ratio score was highly capable of extracting speaker information.

1.2. Influence of phonemes on speaker individual information

In all languages the properties of phonemes are determined by a set of manners and places of articulation coupled to sound sources, with a certain range of phonetic variability from one context to another. The manners and places of articulation endow the phonemes with not only different linguistic properties, but also individual speaker variation. Therefore it is necessary to examine phoneme-specific speaker information and to separate as much as possible speaker information from speech information.

For the reason mentioned above, many researchers have attempted to find some breakthrough points from the