Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 50 (2008) 312–322

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification

Xugang Lu *, Jianwu Dang

Japan Advanced Institute of Science and Technology, 1-1, Asahidai, Nomi, Ishikawa 923-1292, Japan

Received 18 June 2007; received in revised form 23 October 2007; accepted 27 October 2007

Abstract

The features used for speech recognition are expected to emphasize linguistic information while suppressing individual differences. For speaker recognition, in contrast, features should preserve individual information and attenuate the linguistic information at the same time. In most studies, however, identical acoustic features are used for the different missions of speaker and speech recognition. In this paper, we first investigated the relationships between the frequency components and the vocal tract based on speech production. We found that the individual information is encoded non-uniformly in different frequency bands of speech sound. Then we adopted statistical Fisher's *F*-ratio and information-theoretic mutual information measurements to measure the dependencies between frequency components and individual characteristics based on a speaker recognition database (NTT-VR). From the analysis, we not only confirmed the finding of non-uniform distribution of individual information in different frequency bands from the speech production point of view, but also quantified their dependencies. Based on the quantification results, we proposed a new physiological feature which emphasizes individual information for text-independent speaker identification by using a non-uniform subband processing strategy to emphasize the physiological information involved in speech production. The new feature was combined with GMM speaker models and applied to the NTT-VR speaker recognition database. The speaker identification using proposed feature reduced the identification error rate 20.1% compared that with MFCC feature. The experimental results confirmed that emphasizing the features from highly individual-dependent frequency bands is valid for improving speaker recognition performance.
© 2007 Elsevier B.V. All rights reserved.

PACS: 01.30.−y

Keywords: Speaker identification; Physiological features; Speech production; Fisher's *F*-ratio; Mutual information; Frequency warping

1. Introduction

Linear prediction coefficient (LPC) and Mel frequency Cepstral coefficient (MFCC) features are widely used as acoustic features for speech recognition. The state of the art of text-independent speaker identification algorithms are also based on modeling the LPC or MFCC feature using Gaussian mixture model (GMM) (Reynold, 1995). However, the purpose of speech recognition is quite different from that of speaker recognition, the former task needs to emphasize linguistic information and suppress individual information, while the later task needs to preserve individual information. This contradiction suggests that either LPC or MFCC may not meet the requirements of both speech and speaker recognition tasks.

1.1. Speaker feature extraction based on signal processing

For speaker recognition, the problem is how to extract and utilize the information that characterizes individual speakers. Generally speaking, individual information of speakers results mainly from two factors: physiological and social factors. The former are related to the speaker's gender, age, and oral morphology, which are inborn

* Corresponding author.
E-mail addresses: xugang@jaist.ac.jp (X. Lu), jdang@jaist.ac.jp (J. Dang).

0167-6393/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2007.10.005

characteristics; the latter concerns the speaker's dialect, idiolect, occupation, and so on, which result from his/her social environment. In this paper, we focus on the former factors, and investigate their acoustic characteristics in speech. In other words, we attempt to extract individual information which is involved in morphological details and acoustic characteristics, and implement it in speaker recognition.

Usually, when producing a speech sound, speakers' physiological and morphological features are encoded in acoustic characteristics of the sound. The diverse articulators contribute different physical properties in the acoustic spectrum, which are personalized in individual morphologies. In order to extract that information, some speech feature representations have been developed. The LPC feature can well model the vocal tract properties by using an all-pole model, which reflects the main vocal tract resonance property in the acoustic spectrum (Rabiner and Juang, 1993; Stevens, 1998). This feature emphasizes the formant structure that concerns major individual differences of the speakers, while some significant details of individuals such as the nasal, piriform fossa and other side branches are ignored. In contrast, the MFCC feature takes the mechanism of the auditory nonlinear frequency resolution into consideration, which improves the representation robustness (Rabiner and Juang, 1993). For extracting more direct physiological features, the fundamental frequency or pitch which reflects the vocal cord information of speakers is often used (Atal, 1976). The LPC residual signal has also been proposed for describing the speakers' glottal information (He and Liu, 1995). When these features were used for speaker recognition, the performance was improved to some extent (Atal, 1976; He and Liu, 1995). For speaker recognition, the essential goal is to find the non-linguistic information which is highly correlated with individual characteristics from speech sounds, those studies discussed above endeavored to extract the intrinsic individual information for speaker recognition task. However, how to find and extract the intrinsic individual-related acoustic features is still a difficult problem.

1.2. Speaker feature investigation based on speech production

The main focus of this study is to provide a systematic analysis of the relationship between acoustic frequency components and individual characteristics from both the speech production point of view, and the statistical information processing point of view, and to propose a physiological feature extraction method for speaker recognition. Actually, in essence, most of the previous studies try to extract the features for speaker recognition from the main vocal tract physical property, which is usually described by the phoneme-dependent dynamics of vocal tracts. However, the speaker-dependent features are more important for speaker recognition which is expected to be invariant in the articulation dynamics. In the vocal tract, during speech production, there are number of side branches, such as the nose, piriform fossa, etc., which have less variation during speech and introduce invariant features in some specific frequency regions (Suzuki et al., 1990; Dang and Honda, 1994, 1996a,b). The frequency regions concerned with the side branches may be related to paralinguistic or non-linguistic information. One important characteristic of those side branches is that they show large variation across speakers, but have small changes during speech production for the same speaker. In addition, they are not easily changed by conscious efforts. In other words, the frequency regions produced by those side branches are not easily disguised in the speech. The acoustic features around those frequency regions should be suitable for individual description.

Based on the analysis above, in this paper, we investigate new features which reflect the important speaker-specific information in frequency domain, design a subband processing strategy for feature extraction, and apply it to the speaker identification task. This paper is organized as follows. Section 2 analyzes encoding of speaker individual information from speech production point of view, i.e., analyzing the individual physiological features caused by speaker individual speech organs' morphology, which shows the non-uniform frequency dependency of individual characteristics. Section 3 introduces two methods, i.e., Fisher's *F*-ratio (Wolf, 1972) and mutual information (Cover and Thomas, 1991), to quantify the dependencies between frequency components and speaker identities. Based on the results from Sections 3 and 4 describes the proposed non-uniform subband processing algorithm, and extracts the new speaker physiological feature. In Section 5, the speaker model using HTK is first introduced, then speaker identification experiments are performed to test the new feature extracted based on the algorithm in Section 4. In addition, in Section 5, the performance of the system with the proposed feature is compared with those using traditional features. Finally, Section 6 gives conclusions and future directions.

2. Encoding of speaker individual information in speech production

In order to extract intrinsic speaker features, we must know where and how the speaker features are encoded during speech production. In this section, we try to clarify distributions of speaker features in frequency domain, and explain the intrinsic connections between the frequency components and physical factors for speaker individual information representation based on several speech production studies (Suzuki et al., 1990; Dang and Honda, 1994, 1997).

From the view point of speech production, speech sound results from a sound source modulated by the vocal tract filter. The speaker-specific information should be involved in the invariant characteristics of the sources and the vocal tracts of the speakers. In this study, we are more concerned

Download English Version:

https://daneshyari.com/en/article/566040

Download Persian Version:

https://daneshyari.com/article/566040

Daneshyari.com