Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 81 (2016) 54–71

www.elsevier.com/locate/specom

Significance of analytic phase of speech signals in speaker verification

Karthika Vijayan[*], Pappagari Raghavendra Reddy, K. Sri Rama Murty

Department of Electrical Engineering, Indian Institute of Technology Hyderabad, India

Received 1 July 2015; received in revised form 22 January 2016; accepted 6 February 2016
Available online 26 February 2016

Abstract

The objective of this paper is to establish the importance of phase of analytic signal of speech, referred to as the analytic phase, in human perception of speaker identity, as well as in automatic speaker verification. Subjective studies are conducted using analytic phase distorted speech signals, and the adversities occurred in human speaker verification task are observed. Motivated from the perceptual studies, we propose a method for feature extraction from analytic phase of speech signals. As unambiguous computation of analytic phase is not possible due to the phase wrapping problem, feature extraction is attempted from its derivative, i.e., the instantaneous frequency (IF). The IF is computed by exploiting the properties of the Fourier transform, and this strategy is free from the phase wrapping problem. The IF is computed from narrowband components of speech signal, and discrete cosine transform is applied on deviations in IF to pack the information in smaller number of coefficients, which are referred to as IF cosine coefficients (IFCCs). The nature of information in the proposed IFCC features is studied using minimal-pair ABX (MP-ABX) tasks, and t-stochastic neighbor embedding (t-SNE) visualizations. The performance of IFCC features is evaluated on NIST 2010 SRE database and is compared with mel frequency cepstral coefficients (MFCCs) and frequency domain linear prediction (FDLP) features. All the three features, IFCC, FDLP and MFCC, provided competitive speaker verification performance with average EERs of 2.3%, 2.2% and 2.4%, respectively. The IFCC features are more robust to vocal effort mismatch, and provided relative improvements of 26% and 11% over MFCC and FDLP features, respectively, on the evaluation conditions involving vocal effort mismatch. Since magnitude and phase represent different components of the speech signal, we have attempted to fuse the evidences from them at the i-vector level of speaker verification system. It is found that the i-vector fusion is considerably better than the conventional scores fusion. The i-vector fusion of FDLP+IFCC features provided a relative improvement of 36% over the system based on FDLP features alone, while the fusion of MFCC+IFCC provided a relative improvement of 37% over the system based on MFCC alone, illustrating that the proposed IFCC features provide complementary speaker specific information to the magnitude based FDLP and MFCC features.

Keywords: Analytic phase; Instantaneous frequency; Feature extraction; MP-ABX tasks; t-SNE visualization; Speaker verification.

1. Introduction

Speaker verification is the task of verifying the claimed identity of a person from his/her voice. It is an important task in the field of speech processing, finding applications in the areas of voice access control, telephone banking and forensics (Kinnunen and Li, 2010). Speaker verification system requires extraction of speaker-specific information from the speech signal. In addition to information about the speaker identity, the speech signal conveys rich information related to textual message, language of communication, emotion and health state of the speaker. From such a composite signal, the extraction of speaker-specific features that help in discriminating the speakers well, has to be realized. Speaker-specific characteristics are mainly a result of anatomical structure, like vocal tract shape and size, and learned speaking habits, like dialect and prosody. The anatomical structure plays an important role in characterizing the speaker and hence, we need to analyze the speech signal and extract features representing the anatomical structure of the speech production mechanism.

One of the major goals of signal analysis is to infer the characteristics of the underlying system from the signal. In the case of speech analysis, we need to extract information

[*] Corresponding author. Tel.: +91 9581145556.
E-mail addresses: ee11p011@iith.ac.in, karthikavijayan@gmail.com (K. Vijayan), ee12m1023@iith.ac.in (P. Raghavendra Reddy), ksrm@iith.ac.in (K. Sri Rama Murty).

http://dx.doi.org/10.1016/j.specom.2016.02.005
0167-6393/© 2016 Elsevier B.V. All rights reserved.

about the vocal tract system (VTS) and excitation source from the observed speech signal. Since speech is a natural signal, it is not amenable for closed-form mathematical representation. However, natural signals can be analyzed by expanding them using a complete set of basis functions, having precise mathematical representation. From a mathematical point of view, there are several ways of achieving this signal decomposition. The Fourier transform is a prominent approach for signal decomposition, in which an arbitrary signal is expressed as a linear combination of complex sinusoids (Oppenheim et al., 1999). The set of coefficients of these complex sinusoids represents relative contributions of different frequencies, and is called the spectrum. The spectrum, in general, is complex-valued and it is often advantageous to express it in terms of its magnitude and phase. In the case of speech signals, prominent peaks in the magnitude spectrum, referred to as formants (Quatieri, 2001), convey information about the resonances of the VTS. The locations of the formants are influenced by the anatomical structure of the VTS, and hence, are important for speaker recognition. For example, location of the first formant, being inversely proportional to the length of the vocal tract, might be helpful in inferring the height of a speaker (Greisbach, 1999). Most of the state-of-the-art speaker recognition systems use features extracted from the magnitude spectrum of the speech signal. Mel frequency cepstral coefficients (MFCCs) (Davis and Mermelstein, 1980) and linear prediction cepstral coefficients (LPCCs) (Makhoul, 1975), which represent the gross envelope of the magnitude spectrum, are the commonly used features for speaker recognition (Kinnunen and Li, 2010). Though there were attempts to extract features from the phase spectrum of the speech signal, they were not as popular as their magnitude counterparts (Alsteris and Paliwal, 2007; Picone, 1993). Since Fourier transform is a weighted average of the signal, over the entire duration, the Fourier spectra of signals with time-varying frequency content is not physically meaningful (Cohen, 1995). For example, the Fourier transform of a Gaussian modulated chirp signal in Fig. 1(a) does not offer any insight into its time-varying characteristics. Hence, short-time Fourier transform (STFT) is used to analyze the time-varying characteristics of the VTS (Rabiner and Schafer, 1978).

Amplitude modulated and frequency modulated (AM-FM) signal decomposition provides an alternative way of analyzing time-varying frequency content in a signal. In this analysis, a narrowband (NB) signal (predominantly mono-component) is decomposed into instantaneous amplitude and phase components. Several methods have been proposed to accomplish such a decomposition (Cohen, 1995; Gianfelici et al., 2007; Griffiths, 1975; Kumaresan and Rao, 1999; Maragos et al., 1993a, 1993b; Potamianos and Maragos, 1999; Quatieri, 2001; Quatieri et al., 1997). The analytic signal representation obtained through the Hilbert transform is the most commonly used method for AM-FM decomposition (Boashash, 1992; Cohen, 1995). Fig. 1(b) and (c) shows the instantaneous amplitude and analytic phase variations obtained from the analytic signal representation of a Gaussian modulated chirp signal in Fig. 1(a).

Another well known algorithm for AM-FM decomposition of a NB signal is energy separation algorithm (ESA) (Maragos et al., 1993a, 1993b). This method utilizes nonlinear Teager-Kaiser energy operator, which calculates energy of a mono-component signal as the product of its squared amplitude and frequency (Kaiser, 1990). The instantaneous characteristics of the signal are then obtained by applying the ESA algorithm (Maragos et al., 1993a). Comprehensive comparison between Hilbert transform method and ESA method can be found in Vakman (1996) and Potamianos and Maragos (1994). Phase locked loops and extended Kalman filters have also been explored for demodulation of NB signals (Gill and Gupta, 1972; Pai and Doerschuk, 2000; Pantazis et al., 2011).

Generalization of AM-FM demodulation techniques to multi-component wideband (WB) signals, like speech, is not straightforward. It is not physically meaningful to interpret the instantaneous amplitude and phase of a multi-component signal (Boashash, 1992). The most common solution for AM-FM decomposition of a WB signal is to pass the signal through a bank of NB filters, and then apply the preferred NB decomposition algorithm on the output of each filter (Potamianos and Maragos, 1996). This strategy is called multiband demodulation analysis, and is similar to phase vocoder in speech processing (Quatieri, 2001). Another popular approach for AM-FM separation of WB signals is the empirical mode decomposition (Huang et al., 1998), which uses the extrema of the signal to obtain intrinsic mode functions (IMF). Although this method provides highly accurate signal representations, straightforward implementation of sifting procedure produces mode mixing (Huang et al., 1998). That is, a specific signal may not be separated into the same IMFs every time. This problem makes it hard to implement feature extraction, model training and pattern recognition, since a feature is no longer fixed at one labeling index.

The AM-FM analysis, especially the amplitude component, has been used in speech processing applications. The locations of the formants were estimated from AM-FM decomposition of speech signals using linear adaptive or fixed filter-banks (Atal and Shadle, 1978; Potamianos and Maragos, 1996; Rao and Kumaresan, 2000). Features extracted from instantaneous amplitude envelopes of NB components were used for speech and speaker recognition (Gowda et al., 2015; Kinnunen, 2006; Sadjadi et al., 2012; Shannon et al., 1995). Frequency domain linear prediction (FDLP) features, derived from all-pole models of amplitude envelopes, were found to be either comparable or better than the conventional MFCC features for speech processing applications (Athineos and Ellis, 2007; Ganapathy et al., 2014).

Both AM and FM components are essential for exact reconstruction of the original signal. Perceptual studies also assert that FM component is important for human perception of speech signals, especially in noisy conditions (Wolfe et al., 2009; Won et al., 2014; Zeng et al., 2005). However, the FM component has received lesser prominence than the AM component in mainstream speech processing. This could be due to the inevitable phase wrapping problem associated with the computation of FM component, or the analytic phase