Speech Communication 91 (2017) 1–16

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Analysis of the Intrinsic Mode Functions for Speaker Information

Rajib Sharma*, S.R.M. Prasanna, Ramesh K. Bhukya, Rohan Kumar Das

Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati, 781039, India

ARTICLE INFO

Article history:
Received 4 October 2016
Revised 9 April 2017
Accepted 11 April 2017
Available online 27 April 2017

Keywords:
EMD
MEMD
IMFs
MFCCs
SV
i-vector
GMM

ABSTRACT

This work explores the utility of the time-domain signal components, or the *Intrinsic Mode Functions* (IMFs), of speech signals', as generated from the *data-adaptive filterbank* nature of *Empirical Mode Decomposition* (EMD), in characterizing speakers for the task of text-independent *Speaker Verification* (SV). A *modified version of EMD*, denoted as MEMD, which extracts IMFs with lesser *mode-mixing*, and provides a better representation of the higher frequency spectrum of speech, is also utilized for the SV task. Three different features are extracted over 20 ms frames, from the IMFs of EMD and MEMD. They are, then, tested individually, and in conjunction with the *Mel Frequency Cepstral Coefficients* (MFCCs), for SV. Two corpora - the NIST SRE 2003 corpus, and the CHAINS corpus - are used for the experiments. The results evaluated on the NIST SRE 2003 database, using the *i-vector* framework, reveal that the features extracted from the IMFs, in conjunction with the MFCCs, enhances the performance of the SV system. Further, it is observed that only a small set of lower-order IMFs is useful and necessary for characterizing speaker-specific information. The combination of the features with the MFCCs is also found to be useful when *short speech utterances* of $\leq$10 s are used for testing. Similarly, the results evaluated on the CHAINS corpus, using the conventional *Gaussian Mixture Model* (GMM) framework, reveal that the features, in combination with the MFCCs, enhance the performance of the SV system, not only for *normal* speech, but also for *fast* and *whispered* speech. Again, it is observed that only the first few IMFs are needed and useful for achieving such enhanced performance.

1. Introduction

The *Mel Frequency Cepstral Coefficients* (MFCCs) have been the cornerstone for speaker modeling for a long stretch of time (Rabiner and Schafer, 1978; 2007; Benesty et al., 2008). The MFCCs are obtained by the application of the *Mel filterbank* on the short-time Fourier magnitude spectrum of the speech signal. The Mel filterbank is a non-uniform filterbank, which is based on how the human ear perceives sound using critical bands (Benesty et al., 2008). The final outcome, the MFCCs, are found to be effective in capturing speaker information in a compact manner. Even though the MFCCs were not motivated to capture the production aspects of the speech signal, they are believed to capture the vocal tract information embedded in the signal (Rabiner and Schafer, 1978; 2007; Benesty et al., 2008). The vocal tract information modeled includes static and dynamic vocal tract shape information. The vocal tract shape and its dynamics represent one aspect of speaker information. The excitation source also contributes to the speaker information (Prasanna et al., 2006; Das and Mahadeva Prasanna, 2016). There are also attempts to use long term information for speaker modeling, like idiolect, prosody, modulation, etc. Fazel and Chakrabartty (2011); Petrovska-Delacrétaz et al. (2007); Jin and Zheng (2011); Yegnanarayana et al. (2005).

The performance of the speaker recognition system is found to degrade significantly under different practical challenges (Beigi, 2012; Neustein and Patil, 2011). These include degraded acoustic environments, change in speaking styles, short utterances for training and testing, etc., which have always demanded to look for more features representing speaker information. One way is to look for different components of speaker information like the glottal source, the vocal tract system, and long term features as mentioned earlier. An alternative way is to explore new signal processing approaches for modeling speaker information. The focus of this work is to investigate the capability of the *Amplitude Modulated-Frequency Modulated* (AM-FM) components of the speech signal, as obtained from the *Empirical Mode Decomposition* (EMD) (Huang and Shen, 2005; Huang and Pan, 2006; Sharma et al., 2017), in characterizing speaker-specific information. The speech signal has been shown to be constituted of AM-FM signals, even for short time intervals amounting to a few *pitch* periods

* Corresponding author.
E-mail addresses: s.rajib@iitg.ernet.in (R. Sharma), prasanna@iitg.ernet.in (S.R.M. Prasanna), r.bhukya@iitg.ernet.in (R.K. Bhukya), rohankd@iitg.ernet.in (R. Kumar Das).

http://dx.doi.org/10.1016/j.specom.2017.04.006
0167-6393/© 2017 Elsevier B.V. All rights reserved.

(Hardcastle and Marchal, 1990; Teager, 1980; Teager and Teager, 1990; Kaiser, 1990; Maragos et al., 1992; 1993; Holambe and Deshpande, 2012). Henceforth, the concept of AM-FM analysis of speech was developed, wherein the speech signal is represented as the sum of a finite number of narrowband signals, with slowly varying amplitudes and frequencies (Holambe and Deshpande, 2012; Bovik et al., 1993; Maragos et al., 1993; Potamianos and Maragos, 1996). Thus, each component of the speech signal, under this representation, is an AM-FM signal, with limited degrees of amplitude and frequency modulation, which hopefully capture the *key characteristics* associated with the signal.

As the speech formants are not known *a priori*, the simple framework of *Multiband Demodulation Analysis* (MDA) has found popular use for decomposing the speech signal into its AM-FM components (Holambe and Deshpande, 2012; Bovik et al., 1993; Potamianos and Maragos, 1996). In this framework, the speech signal is passed through a *large fixed parallel bank of overlapping band pass filters*. This ensures that even if the formant frequencies vary with time, one of the filters will pick up the speech resonances at any given instant. However, this makes it a *fixed* analysis, determined by the design of the filterbank. Also, apart from the useful components which carry the vocal tract resonances and the *glottal source* information, a multitude of other components are also generated. The ideal objective of AM-FM analysis, to represent the speech signal in terms of its time-varying resonances only (also the glottal source characteristics), is still left desired. Hence, if there were a method of performing AM-FM analysis such that it could *adaptively decompose* the speech signal into a *finite number of meaningful* time domain components, it would be more appealing to the speech community. It is on these lines that the technique of EMD has gained attention in the speech processing domain (Sharma et al., 2017).

EMD circumvents the problem of designing the filterbank required for extracting the AM-FM speech components. It decomposes the speech signal into a much smaller set of AM-FM signals, called *Intrinsic Mode Functions* (IMFs), compared to traditional AM-FM analysis, while ensuring that the IMFs carry latent and meaningful information about the processes generating the signal (Huang et al., 1998; Huang and Shen, 2005; Boudraa and Cexus, 2007; Sharma and Prasanna, 2015; Sharma et al., 2017). This is made possible by the *data-adaptive filterbank* nature of EMD - its inherent characteristic (Wang et al., 2010; Wu and Huang, 2009). This implies that the IMFs represent signals which have their power spectra which are dependent on the speech signal being decomposed - they are adaptive not pre-defined. EMD is known to exhibit a phenomenon, called *mode-mixing*, which limits the utility of the IMFs extracted from it. *Mode-mixing* may be defined as the presence of *disparate frequency scales* within an IMF, or the distribution of the same frequency scale amongst many IMFs. This makes it difficult to draw conclusions about the set of IMFs which carry important information, and which do not. While many advancements of EMD exist for reducing *mode-mixing*, they consume a lot of time, and hence cannot be used for tasks that demand efficiency (Wu and Huang, 2009; Yeh et al., 2010; Torres et al., 2011; Colominas et al., 2014). Recently, however, a *modified version of EMD*, (MEMD) (Sharma and Prasanna, 2016), has been proposed, which reduces *mode-mixing*, but with time-efficiency. As such, in this work, apart from EMD, MEMD is also used for extracting the IMFs of the speech signal. Additionally, MEMD provides a better segregation of the speech spectrum, and thus a better representation of the vocal tract resonances of the speech signal (Sharma and Prasanna, 2016).

In this work, we apply short time processing on the components of speech, derived from EMD/MEMD, to derive speaker information. The objective of this work is to determine whether the adaptive nature of EMD is beneficial in any way for extracting speaker-specific cues. This work investigates whether energy or cepstral-like features, derived from the IMFs, could represent speaker information in a different manner that the MFCCs. Three different features - *Sum Log Squared Amplitude* (Wu and Tsai, 2011), and *Log Sum Squared Amplitude* (Rabiner and Schafer, 1978; 2007; Benesty et al., 2008), and *Entropy* (Papoulis and Pillai, 2002) - are extracted over 20 ms frames, from the IMFs, and utilized in the task of text-independent *Speaker Verification* (SV) (Rabiner and Schafer, 1978; 2007; Benesty et al., 2008). The first two features represent cepstral-like features, similar to the MFCCs, but obtained from the filterbank manifested in the IMFs of EMD/MEMD. The Entropy feature is a measure of the information content in the IMFs. We hypothesize that the IMFs are AM-FM signals which vary adaptively with the speech signal, and as such they might capture certain aspects of the signal, which a fixed structure like the Mel filterbank may not. Though the Mel filterbank is believed to capture the vocal tract information, it was principally designed from the perspective of speech perception (Rabiner and Schafer, 1978; 2007; Benesty et al., 2008). The IMFs, however, have been shown to capture the formants structure (Huang and Chen, 2006; Bouzid and Ellouze, 2007; Sharma and Prasanna, 2016) and the glottal source characteristics (Yang et al., 2005; Huang and Pan, 2006; Schlotthauer et al., 2009; 2010a; Sharma and Prasanna, 2015) of the speech signal, which are production aspects of the speech signal. Just like the perception aspects, the production aspects carry critical information not only of the speech signal, but also the speaker uttering the signal. Based on these observations, we expect the features obtained from EMD/MEMD filterbank to augment the performance of the SV system.

The principal question of the range of IMFs that is useful for SV, is investigated in this paper. The three experimental features are concatenated with 39-dimensional MFCCs, and evaluated in the SV task. The performances of these combinations are compared with 51-dimensional MFCCs to evaluate whether the features really capture speaker information in a different and useful manner. The same three features extracted from the IMFs are also extracted from 20 AM-FM signals obtained from a *20-filter uniform Gabor filterbank* applied on the speech signal. The performance of the SV system for these 20-dimensional features is compared with the features (much lesser dimension than 20) obtained from the IMFs. This enables us to verify the utility of the adaptive nature of the EMD/MEMD filterbank.

Two databases - the NIST SRE 2003 corpus (NIST, 2003), and the CHAINS corpus (Cummins et al., 2006), are used in this work in an attempt to find the relevant answers. Both the databases use standard quality speech, having *sampling frequency* (F_s) of 8 kHz. The state-of-the-art *i-vector* framework (Dehak et al., 2011) is used for the NIST corpus, whereas the standard *Gaussian Mixture Model* (GMM) (Bishop, 2006) is used to train and test the speakers for the CHAINS corpus. Also, extending the experiments, we investigate whether the features derived are useful when applied to *fast* or *whispered* speech (Grimaldi and Cummins, 2008; Jin et al., 2007; Fan and Hansen, 2011; 2009; Sarria-Paja et al., 2013), i.e, when the rate or nature of speech is drastically changed from normal speech. Also, we explore how effective the features derived from the IMFs may be in the case of *limited test-data conditions*, i.e, when *short test-utterances*, of 10 s - 2 s, are used for verifying the claim of the speaker (Das et al., 2014; 2015). The performances of the SV system under these scenarios are relevant from the perspective of practical deployment. The salient contributions of this work may be summarized as :

- Exploration of features from the IMFs of the speech signal for representing speaker specific information.
- Investigate which subset of the IMFs of the speech signal represent speaker information effectively.