Digital Signal Processing 58 (2016) 26–39

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

A better decomposition of speech obtained using modified Empirical Mode Decomposition

Rajib Sharma *, S.R. Mahadeva Prasanna

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

ARTICLE INFO

Article history:
Available online 22 July 2016

Keywords:
EMD
IPs
Mode mixing
Dyadic filterbank
LP
Formants

ABSTRACT

The objective of this work is to obtain *meaningful time domain components*, or *Intrinsic Mode Functions* (IMFs), of the speech signal, using Empirical Mode Decomposition (EMD), with reduced *mode mixing*, and in a time-efficient manner. This work focuses on two aspects – firstly, extracting IMFs of the speech signal which can better reflect its higher frequency spectrum; and secondly, to get a better representation and distribution of the *vocal tract resonances* of the speech signal in its IMFs, compared to that obtained from standard EMD. To this effect, modifications are proposed to the EMD algorithm for processing speech signals, based on the critical nature of the interpolation points (IPs) used for cubic spline interpolation in EMD. The effect of using different sets of IPs, other than the extrema of the residue – as used in standard EMD – is analyzed. It is found that having more IPs is beneficial only upto a certain limit, after which the characteristic *dyadic filterbank* nature of EMD breaks down. For certain sets of IPs, these modified EMD processes perform better than EMD, giving better frequency separability between the IMFs, and an enhanced representation of the higher frequency content of the signal. A detailed study of the distribution of the *formants*, in the IMFs of the speech signal, is done using *Linear Prediction* (LP) analysis of the IMFs. It is found that the IMFs of the EMD variants have a far better distribution of the formants structure within them, with reduced overlapping amongst their filter spectrums, compared to that of standard EMD. Henceforth, when subjected to the task of formants estimation of voiced speech, using LP analysis, the IMFs of the modified EMD processes cumulatively exhibit a superior performance than that of standard EMD, or the speech signal itself, under both clean and noisy conditions.

1. Introduction

Speech is the principal method of communication amongst human beings. It is a signal generated by a complex psycho-acoustic process, developed as a result of thousands of years of human evolution. But, it is not just a tool for communication. It is a signal which contains a multitude of information like the speaker's age, emotion, accent, health and physiological disorders, identity, etc. – which give rise to the various fields of Speech Processing today [1–3]. However, speech is a highly *non-linear and non-stationary signal*, and hence unearthing such information is not a trivial task [3,4]. In the pursuit of many noble speech processing tasks, the decomposition of the speech signal, thus becomes necessary, so that information required for specific tasks can be *segregated, extracted, enhanced or captured* from the plethora of information that the signal carries. On the basis of the popular *source-filter theory* of speech production, the *Short Time Fourier Transform* (STFT) has been the principally utilized for decomposing short segments (10–50 ms) of speech, for further analysis and processing [1–3]. Apart from the fact that STFT can provide only a fixed time and frequency resolution, irrespective of the local frequency content of the signal, the treatment of short speech segments as being produced by a linear system is arguable, and strong contrary evidence exists, regarding the non-linearity of speech, even for short time intervals pertaining to a few glottal or excitation cycles [4–8].

As a way of overcoming the time and frequency resolution limitations of STFT, the *Wavelet transform* (WT) was introduced [9–11]. However, in essence, WT is nothing but an *adjustable window* Fourier Transform, which may provide a solution for non-stationarity, but is a linear transform [12,13]. As such, WT, or advanced filter design techniques [14,15], are applicable for linear system analysis only, and may provide limited insights in the analysis of non-linear signals like speech. Further, the problem of selection of suitable *mother wavelet* exists in the case of WT. As a means of enhancing the accuracy of the STFT analysis of speech, the sinusoidal representation of speech was proposed [16]. This

* Corresponding author.
E-mail addresses: s.rajib@iitg.ernet.in (R. Sharma), prasanna@iitg.ernet.in (S.R. Mahadeva Prasanna).

http://dx.doi.org/10.1016/j.dsp.2016.07.012
1051-2004/© 2016 Elsevier Inc. All rights reserved.

representation models the speech signal as being constituted of a finite number of prominent sinusoidal signals. This representation was, in essence, a miniature version of the STFT analysis, and hence, it carried the fundamental limitations of the STFT analysis. Moreover, the sinusoidal model, required the computation of parameters pertaining to the sinusoidal components, for every short speech segment [16–18]. Nevertheless, the sinusoidal model, along with the Teager Energy Operator (TEO), provided the impetus for the AM-FM representation of the speech signal [4,19–21,17,18]. The AM-FM representation envisages the speech signal as being constituted of a finite number of narrowband AM-FM signals, centered around the *vocal tract resonances* (also called *formants*), or the centers of energy of the speech signal [4,20,17,18]. As the speech resonances are not known a priori, the technique of Multiband Demodulation Analysis (MDA), is applied, wherein the speech signal is subjected to a parallel bank of overlapping bandpass filters, generating different time domain AM-FM signals from the speech signal. These signals are then represented in terms of their instantaneous frequencies and amplitudes, as estimated from the Hilbert Transform or the TEO [4,17,18]. In the recent years, such a representation has been found to be useful in many areas of speech processing [4]. However, the AM-FM representation, and so also the sinusoidal model, are not complete decompositions, i.e, a finite number of components derived from them may approximate but never add up to be the exact same speech signal. Further, apart from the signal components which carry the *vocal tract resonances* and the *glottal source* information, a multitude of other components are also generated by sinusoidal analysis or MDA of speech, which may not be useful for analysis [16–18].

Thus, there is a desire and requirement of a technique which can *completely decompose* the speech signal into a small finite number of time domain components, without involving any computation of parameters, and without using short-time processing of the data. It is also desired that such a method be able to decompose the speech signal into components whose frequency spectrums are dominated by the formant frequencies or the fundamental frequency alone. Such a decomposition would produce less, but *meaningful* speech components. Ideally, the frequency spectrums of the components so generated should not overlap, and each component should carry information about a single formant or the glottal source only. Such components, then, may be considered narrowband with respect to the speech signal, and therefore the piecewise linearity and stationarity assumptions might be more applicable to them. Thus, even conventional short-time analysis based on the source-filter theory might be more effective provided such speech components are available. In the pursuit of such time domain speech components, we may look towards *Empirical Mode Decomposition* (EMD) of speech [12].

Empirical Mode Decomposition (EMD) is a non-linear and non-stationary data analysis technique, with the ability of extracting components, called *Intrinsic Mode Functions* (IMFs), of the signal, in the time domain itself [12]. This ability of EMD to decompose a time-series data into different time-series components has been widely appreciated in a variety of fields, including speech processing [13,22–25]. Principal to EMD is the detection of extrema (maxima and minima) of the signal and cubic spline interpolation using the extrema as the points of interpolation. Thus, the precise detection of extrema is essential to the process, which requires that the signal be sampled at many times the Nyquist rate [26]. Apart from this, there are two central issues that effect EMD – *mode-mixing* and *end-effects* [12,13]. For a speech signal, which has silence regions at the beginning and end of the signal, *end-effects* are not significant. However, *mode-mixing*, which is the intermingling of *frequency-scales* within the IMFs, is a bottleneck. A simplified visualization of the phenomenon of mode-mixing can be done by considering a signal, $s(t)$, which is the sum of two sinusoids.

$$s(t) = s_l(t) + s_h(t) = a_l \cos(2\pi f_l t) + a_h \cos(2\pi f_h t),$$

where $s_l(t) = a_l \cos(2\pi f_l t)$ is the sinusoid of lower frequency, having frequency f_l and amplitude a_l, and $s_h(t) = a_h \cos(2\pi f_h t)$ is the higher frequency sinusoid, with frequency f_h and amplitude a_h. It has been shown that EMD is able to decompose $s(t)$ into $s_l(t)$ and $s_h(t)$, only if the ratio $f = f_l/f_h \lesssim 0.67$ [27]. Also, even if the frequency ratio $f \lesssim 0.67$ is satisfied, for satisfactory segregation, the lower frequency component must not override the higher frequency component, i.e, $a = a_l/a_h \lesssim 1$ [27].

Recent research suggests that the higher frequency spectrum of speech contains invaluable information for speech processing applications, particularly pertaining to speaker characteristics [4,28–33]. As such, there may be unprecedented benefits in having better time domain representation of the higher frequency content of speech, and a better estimation of the higher frequency *vocal tract resonances* or *formants* [4,1–3,34]. Unfortunately, most of the energy in speech is contained in its *voiced* regions, which exhibit a spectral slope of −6 dB/octave, thus subduing its higher frequency content [1–3]. This is the primary reason why the higher frequency spectrum of speech has remained under-utilized in speech processing, and the accurate estimation of the higher frequency formants still remains a challenging task [4,1–3,34]. The spectral slope also contradicts the requirement of EMD for extracting meaningful IMFs of the speech signal, and inhibits its ability to characterize its higher frequency content. As such, most of the formants of voiced speech are captured by the first IMF alone [25]. The second IMF captures the first formant only, and the rest of the IMFs are of lower frequency, and represent the glottal source information [25,35].

To reduce the effects of mode-mixing in extracting IMFs from real physical signals, many modifications have been proposed to the EMD algorithm [36–42]. But, the best results have come by the infusion of noise to the signal. It was observed that by combining the signal with finite amplitude white noise, before feeding it to EMD, mode-mixing could be curtailed satisfactorily. This development was termed Ensemble Empirical Mode Decomposition (EEMD) [39]. The idea of infusing finite amplitude white noise into the signal serves an important purpose – it lends energy to the subdued higher frequency spectrum of speech, which makes the extraction of the higher frequency content of the speech signal much more feasible for EMD. Simultaneously, another effect occurs – adding white noise to the signal increases the number of extrema present in the *inner residue* signal, which when used as interpolation points (IPs), in a *sifting iteration* [12,13,39], leads to better estimates of the maxima and minima envelope's of the signal. However, this approach, though very effective, requires a very high number of white noise realizations to eventually cancel out the effect of noise, thus making it inefficient [39]. In the recent years, many different ways to cancel out the noise incorporated in EEMD has been proposed [40–42]. These developments, while very promising, are still very time-consuming, and hence have had limited use.

The fact that mode-mixing could be curbed by *increasing the number of extrema in the inner residue in a certain way*, which are used as interpolation points (IPs) for cubic spline interpolation in EMD, forms the motivation behind our work. If it were possible to increase the number of IPs in the residue signal, effectively, by some simple signal processing method, rather than the infusion of noise, it might be possible to obtain a better signal decomposition, compared to that of EMD, with little additional time overhead. Though such a decomposition might still be inferior to EEMD and its variants, it might be more applicable to speech signal analysis, and other real world data analysis, where time is of precious value. Thus, given how vital the IPs are in the decomposition process of EMD, we adopt three different ways, to change the IPs from