Empirical Mode Decomposition for adaptive AM-FM analysis of Speech : A Review

Rajib Sharma[a,*], Leandro Vignolo[b], Gastón Schlotthauer[c], M.A. Colominas[c], H. Leonardo Rufiner[b], S.R.M. Prasanna[a]

[a] *Signal Informatics Laboratory,*
Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati-781039, India.
[b] *Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional – SINC(i),*
Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral, Santa Fe (3000), Argentina.
[c] *Laboratorio de Señales y Dinámicas no Lineales,*
Facultad de Ingeniería, Universidad Nacional de Entre Ríos, Oro Verde (3101), Entre Ríos, Argentina.

Abstract

This work reviews the advancements in the non-conventional analysis of speech signals, particularly from an AM-FM analysis point of view. The benefits of such an analysis, as opposed to the traditional short-time analysis of speech, is illustrated in this work. The inherent non-linearity of the speech production system is discussed. The limitations of Fourier analysis, Linear Prediction (LP) analysis, and the Mel Filterbank Cepstral Coefficients (MFCCs), are presented, thus providing the motivation for the AM-FM representation of speech. The principle and methodology of traditional AM-FM analysis is discussed, as a method of capturing the *non-linear dynamics* of the speech signal. The technique of Empirical Mode Decomposition (EMD) is then introduced as a means of performing *adaptive* AM-FM analysis of speech, alleviating the limitations of the *fixed* analysis provided by the traditional AM-FM methodology. The merits and demerits of EMD with respect to traditional AM-FM analysis is discussed. The developments of EMD to counter its demerits are presented. Selected applications of EMD in speech processing are briefly reviewed. The paper concludes by pointing out some aspects of speech processing where EMD might be explored.

Keywords: EMD, AM-FM, Wavelet, LP, MFCC, Speech Processing

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. However, it is not just a tool for communication. It is a signal which contains a multitude of information like the speaker's age, height, 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]. Even though methods for analyzing non-stationary signals like the Wavelet Transform and the Wigner-Ville Transform have been developed, they have not been popular in the speech processing community mainly because they decompose the signal in an alternate domain and introduce additional complexity to the analysis [5–7]. Thus, the *source-filter theory* of speech production has remained the backbone of speech processing [1–3]. The treatment of the speech signal as being linear and stationary for short intervals of time (10-50 ms) gives a simplistic and time-affordable analysis. Such an analysis, though, is arguable and provides an oversimplified view of the phenomena related to speech production [4, 8–10]. Thus, the *Linear Prediction* (LP) analysis of speech provides us with a noisy excitation signal as a representation of the glottal source, and a vocal tract filter which represents only the resonant cavities of the vocal tract (*the high pass filter characteristics of the lips is included in the filter*) [1–3]. Further, the oversimplification of the speech production process makes the LP analysis vulnerable to errors [11].

From the speech perception point of view, the *Mel filterbank*, which is based on the characteristics of the human ear, has been widely appreciated in speech processing [1–3]. The *Mel Filterbank Cepstral Coefficients* (MFCCs) are derived solely from the magnitude spectrum (or power spectrum) of the speech signal while neglecting the phase spectrum of speech. However, the phase spectrum of speech is equally critical to speech perception as the magnitude spectrum and has found important use in many speech processing applications [12–17]. Again, it has been observed that while the Mel filterbank may be used for a variety of speech applications, it does not always provide the optimal features, and filterbanks tuned for different applications might be more suitable for better results [18–

*Corresponding author
Email addresses: s.rajib@iitg.ernet.in (Rajib Sharma), ldvignolo@fich.unl.edu.ar (Leandro Vignolo), gschlotthauer@conicet.gov.ar (Gastón Schlotthauer), macolominas@bioingenieria.edu.ar (M.A. Colominas), lrufiner@sinc.unl.edu.ar (H. Leonardo Rufiner), prasanna@iitg.ernet.in (S.R.M. Prasanna)

22].

To overcome some of these limitations of conventional speech analysis the sinusoidal representation of speech was proposed, which models the speech signal as being constituted of a finite number of time-domain sinusoidal signals, in terms of their frequencies, amplitudes, and phases [23]. The sinusoidal model and the Teager Energy Operator (TEO) provided the impetus for the AM-FM representation of the speech signal [4, 24–28]. The concept of Multiband Demodulation Analysis (MDA) was introduced, wherein the speech signal is passed through a parallel bank of *fixed bandpass filters*, generating different time domain 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, 27, 28]. In the recent years, such a representation has been found to be useful in many areas of speech processing [4].

Though the sinusoidal and the AM-FM representation of speech are effective alternatives to the conventional analysis and processing of speech, they have some demerits as well. Neither the sinusoidal analysis nor the MDA provides a complete decomposition, i.e., a finite number of components derived from them cannot add up to be exactly the same speech signal. The sinusoidal model, again, involves short-time processing, to compute parameters pertaining to the signal components [23, 27, 28]. 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 [23, 27, 28].

If there were a method of *completely decomposing* the speech signal into a finite number of time domain components without involving any computation of parameters, and without using short-time processing of the data, it would be more appealing to the speech community. It is also desired that such a method be able to decompose the speech signal into components whose frequency spectra are dominated by the formant frequencies (*and the fundamental frequency for voiced speech*) alone. Such a decomposition would produce less but *meaningful* speech components. Ideally, the frequency spectra 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 be available. In the pursuit of such time domain speech components, we may look towards *Empirical Mode Decomposition* (EMD) of speech [7].

Empirical Mode Decomposition (EMD) is a non-linear and non-stationary data analysis technique with the ability to extract AM-FM components, called *Intrinsic Mode Functions* (IMFs), of the signal, in the time domain itself [7]. This ability of EMD to decompose a time-series data into different time-series components *without the requirement of any a priori basis* has been widely appreciated in a variety of fields [29]. In this paper, we discuss the various facets of EMD, its developments, and its applications in speech processing.

The rest of the work is organized as follows: Section 2 discusses the non-linearity and non-stationarity of speech. As such, the limitations of conventional short-time analysis of speech, with emphasis on the source-filter model and the Mel filterbank, are discussed. Some non-conventional approaches which cater to the inherent non-stationarity of the speech signal but within the "linear" framework are then briefly discussed. Section 3 presents the AM-FM analysis of speech as a means for processing the speech signal in both a non-stationary and a non-linear framework. The principal philosophy and methodology behind AM-FM analysis are reviewed. Section 4 introduces the technique of EMD as a method for *adaptive AM-FM analysis of speech*, eliminating some of its conventional drawbacks. Section 5 is dedicated to reviewing the advancements of EMD for the purpose of making it a more effective tool. Section 6 compares EMD with other non-conventional speech processing techniques. Section 7 presents some applications where EMD has been used in speech processing, thus projecting its practical utility. Section 8 summarizes this article and concludes this work.

2. Limitations of conventional short-time analysis, and utilization of non-conventional methods for speech processing

"*Much of what speech scientists believe about the mechanisms of speech production and hearing rests less on an experimental base than on a centuries-old faith in linear mathematics.*" - Teager & Teager, 1990.

To verify the validity of the source-filter theory of speech production, Teager measured the air flow at different positions inside the oral cavity. To his surprise, he found that most of the air flow was concentrated along the roof of the mouth and along the surface of the tongue. There was very little airflow at the center of the oral cavity, as depicted in Fig.1. Later, Teager also observed the presence of radial and axial airflows in the vocal tract. Thus, the air flow in the speech production system is not laminar, and hence the planar wave assumptions upon which the linear source-filter theory is based may be deemed arguable [4, 9, 30, 31].

To analyze the simplification achieved in the linear source-filter theory, we may look at Figs.2 and 3. Fig. 2 shows the detailed speech production apparatus, which includes, apart from the main vocal tract, the nasal cavity, and the cavities of *the hypopharynx* and *the piriform fossa*. This complex structure is modeled by the source-filter theory into a far simpler structure - a concatenation of multiple tubes with different cross-sectional areas, as shown in Fig.3. This drastic simplification allows the vocal

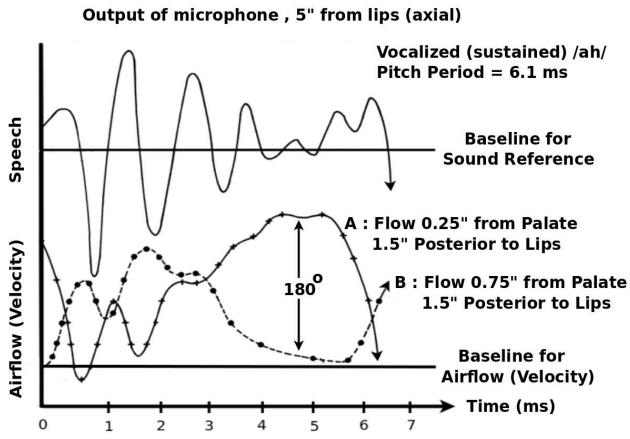


Figure 1: **Figure redrawn from [9]. Three time traces for a vocalized vowel 'ah' produced by a male speaker. The traces provide experimental verification of separated flow. The topmost waveform is that of the speech signal recorded by a microphone placed 5" from the lips. The two waveforms at the bottom, A (solid) and B (dashed), represent airflows at two different positions inside the mouth, measured simultaneously with the recorded speech. The airflow inside the mouth is measured by a 0.7mm x 0.0005cm hot wire sensor, at a temperature of 200° C, with the wire kept normal to the flow.**
< 1 > A represents air flow at a distance of 0.25" from the palate, and B represents air flow at a distance of 0.75" from the palate.
< 2 > A and B are 180° out of phase.
The waveforms show that most of the air flow occurs close to the palate, as represented by A.

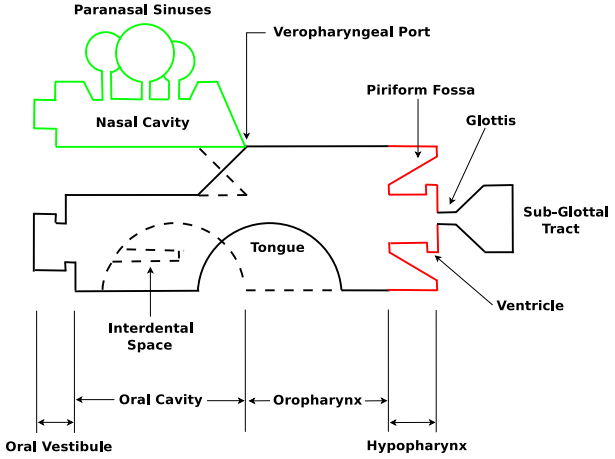


Figure 2: **Figure redrawn from [3]. Acoustic design of the vocal tract.**

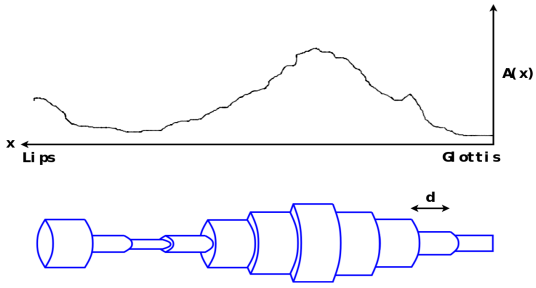


Figure 3: **Area function of the vocal tract from the glottis to the lips (above). The simplified model of the vocal tract as a concatenation of multiple tubes with different cross-sectional areas.**

tract to be modeled as a linear filter, comprising of a cascade of resonators only. A further simplification in speech analysis is obtained by considering this simplified model as being invariant for short segments of time, considering that the movements in the human vocal tract system are limited to such rates. This allows speech to be considered a quasi-stationary signal, whose short-time segments are the output of a Linear Time Invariant (LTI) system [1–3]. While these simplifications make analysis easier, it almost certainly limits the capability of capturing the information embedded in the *dynamics* or the *fine structure* of speech.

From the speech perception point of view, the *Mel filterbank* is used to imitate the characteristics of the human ear. The MFCCs, which are widely used in speech processing applications, however, do not incorporate the phase spectrum of speech. The inability to accommodate the phase spectrum of speech in the MFCCs has led to limitations in the performance of many speech processing applications [4]. Further, one may argue that while the human ear does multi-tasking, it also has an unparalleled computer with mysterious capabilities at its disposal - the brain. Henceforth, for machines to replicate the performance of the human cognitive system it may be more beneficial to construct application-oriented filterbanks, instead of using the Mel filterbank for all purposes. These limitations of conventional speech production and perception modeling and analysis are discussed below.

2.1. Limitations of Fourier Analysis

There are two basic requirements of the data for Fourier-analysis to make sense [7].

(a) The data must be stationary.
(b) The data must be generated by a linear system.

However, real-world data, like speech, seldom meet these requirements. As such Fourier analysis requires many additional harmonics to simulate non-uniformity (abrupt changes) in the data. *It spreads the energy over a wide frequency range.* A simple example is the delta function, which produces a flat Fourier spectrum.

$$\delta(t) \leftrightarrow 1, -\infty < \omega < \infty$$

As such, for non-stationary signals, Fourier analysis produces a multitude of components which combine perfectly mathematically but may not represent meaningful characteristics of the signal [7]. Also, even if the data is stationary, and is constituted of a finite number of sinusoids, Fourier analysis would make sense only if the data is of infinite duration. Lesser the duration of the data, greater is its non-uniform (abruptly changing) and non-stationary char-

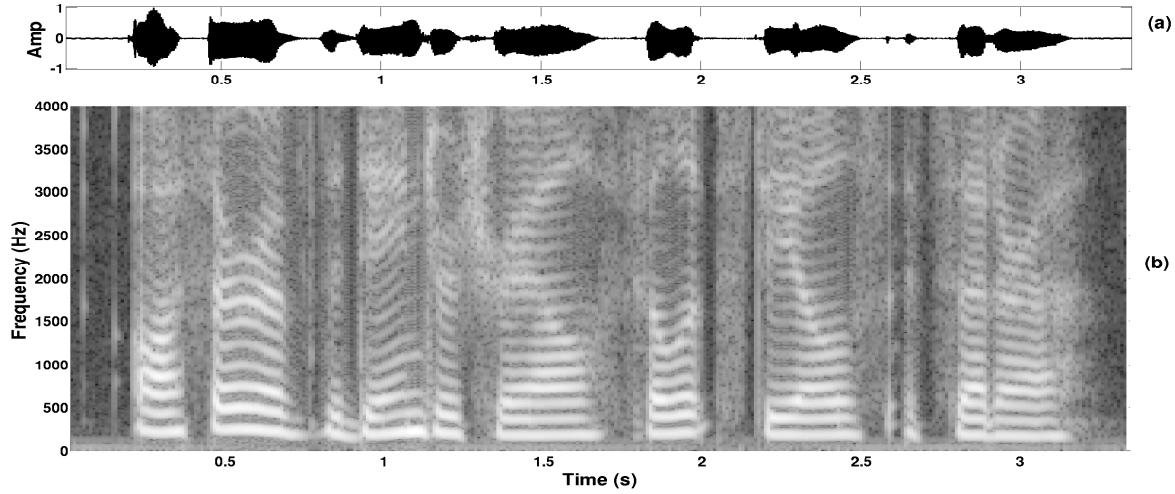


Figure 4: **(a) Speech ; (b) Spectrogram of** (a)**. Framesize of 25 ms with frameshift of 10 ms is used.**

acteristics, and wider the Fourier spectrum. Moreover, Fourier analysis cannot track the change in the frequency content of the signal, as all its components are spread over the entire time scale. As a way of countering this particular limitation, the *Short-Time Fourier Transform* (STFT) has been the utilized, particularly for speech signal processing, wherein the Fourier analysis is done for *short fixed segments* of the speech signal [1–3]. Given a continuous-time speech signal, $s(t)$, its continuous-time STFT, is given by,

$$S(\tau, f) = \int_{-\infty}^{\infty} s(t) w(t-\tau)\ e^{-j2\pi f t} dt,$$

where $w(t)$ represents a symmetric window of finite time-width, and τ the time-instant at which the window is placed. Thus, depending on the width of $w(t)$, STFT provides a fixed time and frequency resolution of the signal. Fig.4 represents the time varying STFT magnitude spectrum in the form of an image, popularly known as the *Spectrogram*, where the STFT spectrum is evaluated at every 10 ms, considering a time window of 25 ms. Clearly, STFT is not an adaptive signal analysis method [32]. There is no "correct" window size, and it varies not only with the task at hand but even within a particular signal.

2.2. Limitations of LP Analysis

The LP analysis has been the cornerstone of speech analysis based on the source-filter theory. The LP analysis is used to estimate the vocal tract resonances, or *formants*, and the excitation source of the speech signal. However, in accordance with the source-filter theory, the LP analysis does not model the antiresonances of the speech production system. Also, it does not model the resonances in the cavities of the hypopharynx and the piriform fossa, which influence the overall spectrum of speech. As a result, the LP analysis is prone to inaccurate estimation of the speech

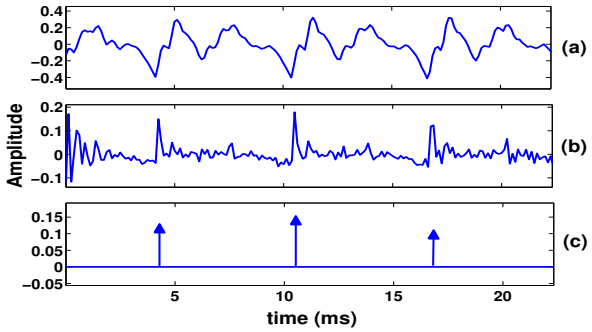


Figure 5: **From top to bottom : A voiced speech signal , its LP residual, and the ideal excitation signal.**

formants, and the excitation signal represented by the LP residual [11].

The ideally expected output of LP analysis is an LP filter with accurate estimation of the vocal tract resonances and the spectral slope of voiced speech, and an LP residual or error signal, which resembles a train of impulses separated by the time-varying pitch period of the speech signal. However, in practicality, the LP residual turns out to be a noisy signal, with relatively larger amplitudes in the vicinity of the GCIs, as reflected in Fig.5. The noisy characteristics of the LP residual may be attributed to three main factors [11] :

(i) The inaccurate estimation of the coefficients of the poles, corresponding to the resonances of the vocal tract system, which makes the LP residual to have non-zero values at time-instants other than the GCIs.

(ii) The inaccurate estimation of the phase angles of the formants, which results in significant bipolar swings in the LP residual, around the GCIs.

(iii) The presence of strong anti-resonances in the speech production system, which causes the large ampli-

tudes in the LP residual to occur at time-instants other than the GCIs.

These differences between the ideal excitation signal and the LP residual, as observed in Fig.5, reflect the mismatch between the actual characteristics of the speech production system, and that modeled by the source-filter theory using LP analysis.

2.3. Importance of phase in speech perception

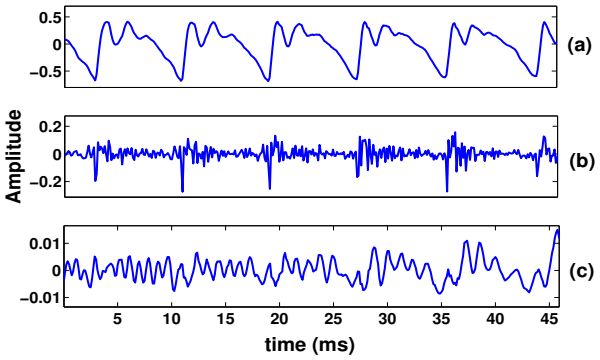


Figure 6: **(a) Speech (b) Phase-only reconstruction of speech (c) Magnitude-only reconstruction of speech. Rectangular window of 1024 ms duration, with 75% overlap between frames is used.**

Table 1: **Consonant intelligibility of Phase-only and Magnitude-only reconstructed speech for different analysis window sizes, and different window types. Values in the table are quoted from [15].**

Window	Intelligibility (%)			
	Magnitude - only		Phase - only	
	32 ms	1024 ms	32 ms	1024 ms
Hamming	84.2	14.1	59.8	88.0
Rectangular	78.1	13.2	80.0	89.3

The understanding of phase does not come as intuitively as that of energy or amplitude. This is probably the reason why the magnitude spectrum is mostly involved in analysis, whereas the phase spectrum remains neglected. Even the MFCCs do not incorporate the phase spectrum of speech. However, the phase spectrum obtained from the STFT of speech has been found to be particularly important in speech perception [15–17]. Phase-only reconstructed speech, i.e., speech signal reconstructed using only its phase spectrum while keeping the magnitude spectrum fixed to unity, is found to be highly intelligible, particularly when rectangular windows are used for analysis. Compared to this, magnitude-only reconstructed speech, i.e., speech signal reconstructed using only its magnitude spectrum while keeping the phase spectrum fixed to zero, is found to be less intelligible. The intelligibility of magnitude-only reconstructed speech is also limited to shorter analysis windows. Table 1 lists the consonant intelligibility averaged over 12 listeners, for 16 consonants spoken by Australian English speakers in *vowel-consonant-vowel* context [15, 16]. The aforementioned observations are evidenced in the table. Again, the phase-only reconstructed speech, as shown in Fig.6, is also observed to carry information about the *epochal events* or the *glottal closure instants* [11, 33–40] in voiced speech [15–17]. This is particularly evidenced for large analysis windows.

2.4. Inadequacies of the Mel Filterbank and the source-filter theory

To validate the utility of the MFCCs for characterizing speaker information vs. speech information, over the broader speech spectrum (0-8 kHz), the *Fisher's F-ratio* [20], which is a ratio of the inter-speaker variance to the intra-speaker variance, is computed for utterances of different speakers. For this experiment 60 triangular filters which are uniformly spaced in the linear frequency scale are used [20]. Every speech frame, of a given utterance, is subjected to the filterbank, to obtain 60 subband energies for the frame. Let $x^i_{k,j}$ be the subband energy of the k^{th} frequency band of the j^{th} speech frame belonging to the i^{th} speaker. The average subband energy of the k^{th} frequency band for the i^{th} speaker is given by,

$$u^i_k = \frac{1}{N^i}\sum_{j=1}^{N^i} x^i_{k,j} \ , \ k = 1, ..., 60 \ , \tag{1}$$

where N^i is the total number of speech frames for all the utterances belonging to the i^{th} speaker. Then, the average subband energy of the k^{th} frequency band for all the speakers is given by,

$$u_k = \frac{1}{M^s}\sum_{i=1}^{M} u^i_k \ , \tag{2}$$

where M^s is the total number of speakers. The F-ratio for the k^{th} frequency band, is then given by,

$$F^{ratio}_k = \frac{\frac{1}{M^s}\sum_{i=1}^{M^s}(u^i_k - u_k)^2}{\frac{1}{M^s N^i}\sum_{i=1}^{M^s}\sum_{j=1}^{N^i}(x^i_{k,j} - u^i_k)^2} \ , \ k = 1, ..., 60 \tag{3}$$

As can be seen from equation (3), the numerator of F^{ratio}_k gives the variation of energy in the k^{th} frequency band for different speakers. The denominator gives the variation of energy in the k^{th} frequency band for the same speaker. Thus, a high value of F-ratio for a given frequency band indicates the presence of high speaker-specific information in the band. Contrarily, a low value of F-ratio indicates the presence of speech-specific information in the band. Fig.7 shows the F-ratio (in dB) computed for different sessions of the NTT-VR corpus [41], and for different types of speech of the CHAINS corpus [42]. As can be seen from the figure, for the NTT-VR corpus the F-ratio is high roughly in three regions - below 500 Hz, around 5000 Hz,

and around 7000 Hz. These are the regions which carry most of the speaker-specific information. The rest of the regions, particularly between 1 - 4 kHz, have low F-ratio values and carry the message information of speech. In the case of the CHAINS corpus, again, it may be observed that the F-ratio values starts to increase after their lowest point (between 3-4 kHz), as the frequency increases. This is found to be true irrespective of the type of articulation or speaking style.

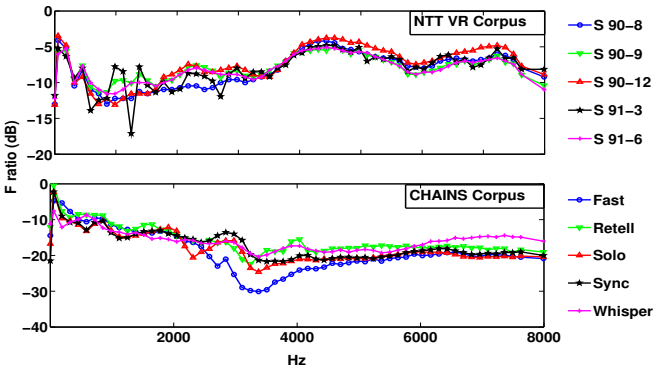


Figure 7: **F-ratio for different frequency bands of speech, evaluated on the NTT (redrawn from [20]) and CHAINS databases.**

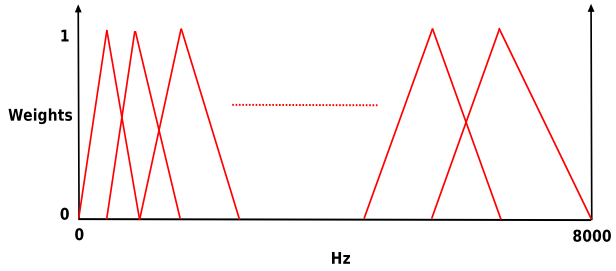


Figure 8: **Mel filterbank in the linear frequency scale.**

The three frequency bands of the F-ratio curves of the NTT-VR corpus, carrying speaker-specific information, may be attributed to three different aspects of the speech production system. The high values of F-ratio below 500 Hz signify the fundamental frequency variation or the variability of the glottal source amongst speakers. Similarly, the high values of F-ratio around 7 kHz might be attributed to the vocal tract constrictions in the production of unvoiced speech. The high speaker discrimination information between 4 - 6 kHz, however, is believed to be contributed by the resonances and antiresonances of the hypopharynx and the piriform fossa - the structures which are not included in the source-filter theory [20, 43–45]. Henceforth, attempts have been made to incorporate the velocity-to-velocity transfer functions of the hypopharynx and the piriform fossa in the source-filter theory based speech production model [43–45]. These attempts have revealed that these structures in the lower vocal tract significantly change the spectrum of voiced speech above 3.5 kHz, producing stable formants in the higher frequency spectrum of speech [43–45].

The above experiments suggest that the conventional Mel filterbank (Fig.8), which progressively deemphasizes higher frequencies, may not be optimal for speaker recognition purposes. Based on these observations, alternate avenues for speaker recognition are being explored, particularly those emphasizing the higher frequency spectrum of speech [4, 18–20, 46–49]. Some of these experiments use different filterbanks, as opposed to the conventional Mel filterbank [18–20], while others use an AM-FM representation of speech [46, 48, 49], which is discussed later, in Sec.3, in this paper. Even for speech recognition, the high resolution of the Mel filterbank at low frequencies might affect the machine recognition of speech. Also, the peaks and valleys within 1-4 kHz indicate that the Mel Filterbank may not represent the optimal filterbank for speech recognition.

2.5. Utility of non-conventional analysis and filterbank optimization for speech processing

The MFCCs, despite the limitations of the Mel filterbank, are the most widely used features for most speech processing applications like speech recognition [50], speaker verification [51], emotion recognition [52–55], language recognition [56], etc., and even for non-speech acoustic signal processing tasks, such as music information retrieval [57]. However, as discussed in the preceding subsection, it is quite ambitious to assume that they would provide the best possible performance for all applications. This is why many alternatives to the Mel filterbank have been recently introduced, allowing to improve the performance of different tasks. Most of these alternatives consist of modifications to the classical filterbank [58–64]. To improve the feature extraction process, a common strategy consists of designing filterbanks using data-driven optimization procedures [65–68]. In this direction, different methodologies based on non-stationary data analysis techniques like *Evolutionary Algorithms* (EAs) [21, 22, 69–71] and *Wavelet Transform* (WT) [5, 6, 32, 72–76] have been utilized in different speech processing applications.

An example of filterbank optimization using EAs for the extraction of cepstral features may be found in [77]. Fig.9 shows the filterbanks optimized for Hindi stressed speech corpus [78] and the FAU Aibo Emotion Corpus [79, 80], for stressed and emotional speech classification respectively. The features so obtained are called the *Evolutionary Spline Cepstral Coefficients* (ESCCs). ESCC-Eh corresponds to the filterbank optimized for the Hindi corpus, and ESCC-Ef corresponds to the filterbank optimized for the FAU Aibo corpus. It can be noticed that the optimized filterbanks differ significantly from the Mel filterbank (Fig.8).

As a way of overcoming the time and frequency resolution limitations of the STFT, the *Wavelet Transform* (WT) was introduced [5, 6, 32]. The continuous-time WT

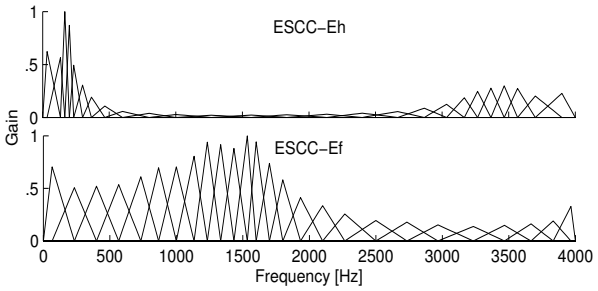


Figure 9: **Figure redrawn from [77]. Filterbanks optimized using EA for stress and emotion recognition.**

of a signal, $s(t)$, is given by,

$$\Psi_s^{\psi}(\tau, r) = \frac{1}{\sqrt{|r|}} \int_{-\infty}^{\infty} s(t)\psi^* \left(\frac{t-\tau}{r}\right) dt$$

where $\psi(t)$, called the *mother wavelet*, represents an oscillatory signal of finite time-width. Here, r represents the scale and τ the time around which the signal is analyzed. Thus, WT allows the visualization of the signal at different scales, depending on the value of r. Thus, WT, in essence, may be defined as an adjustable window STFT. The discrete version of the continuous-time WT, called the *Discrete Wavelet Transform* (DWT), is popularly used for analyzing non-stationary digital signals. For a digital speech signal, $s(n)$, the *dyadic* DWT is obtained as

$$W_s^{\psi}\left[\tau, 2^j\right] = \sum_{n=0}^{N-1} s(n)\psi_{\tau,2^j}^*(n)$$

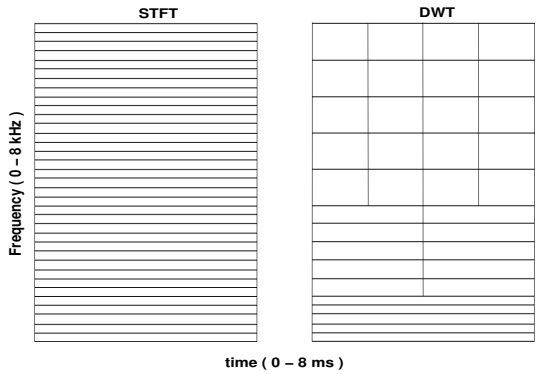


Figure 10: **Comparison between time-frequency resolution of STFT (left) and 3-level DWT (right), calculated for a 8 ms signal of sampling frequency 16 kHz.**

The comparison between the time-frequency resolutions of the STFT and the DWT may be visualized in Figure 10. The decomposition of the DWT is further extended by the Wavelet Packets Transform (WPT), which applies the filtering process of the *binary decomposition tree* to both the low-frequency and high-frequency component of the signal, at each stage of the decomposition. Then, an overcomplete dictionary is obtained by this process, providing more flexibility for the analysis of specific frequency bands. From the decomposition of the WPT, different sub-trees can be selected in order to extract the desired information. The flexibility of the WPT has been exploited in many speech processing tasks, particularly in the context of speech and speaker recognition problems [72–74]. The flexibility provided by the WPT, however, comes along with the challenging problem of choosing the optimum set of coefficients among all the possible combinations for a particular application. Usually, this is tackled by restricting the search to orthogonal basis [81–84]. For applications like speech recognition, studies have concluded that redundant representations could provide increased robustness [22, 75]. One may relate this to the fact that the analysis performed at the level of the auditory cortex is highly redundant, and therefore non-orthogonal [85]. Another concern in the design of useful wavele-based decomposition is the choice of an adequate wavelet family and associated parameters that suit the particularities of the signal of interest and the problem at hand. Many approaches are being proposed to address this issue depending on the application [86–88]. As such, in order to exploit the flexibility provided by the WPT decomposition to extract optimal speech features, new and unorthodox methodologies are being explored [75, 76].

Thus, it may be concluded that there is a definite scope for non-conventional methods like the WT or the EA in various speech processing applications. Most of these methods, in essence, tackle the problem of *non-stationarity*. The idea is to find out the frequency bands which are important for a particular application. However, while these techniques optimize the features for a particular application, the optimization is not adaptive to individual speech utterances. Further, they do not take into consideration the inherent *non-linearity* of the speech production system, or try to capture information embedded in the non-linear characteristics of speech. These techniques process the speech signal in a "linear" framework [5, 6, 32]. Henceforth, they are bound to exhibit limitations in analyzing non-linear signals [7]. Henceforth, a representation of the speech signal is desired which takes into consideration not only its non-stationarity but also its non-linear dynamics. In this direction, we must head towards the AM-FM analysis of speech.

3. AM-FM analysis of speech

As an attempt to overcome some of the limitations of traditional STFT analysis of speech, the sinusoidal representation of speech was proposed [23]. This model represented the glottal excitation source, and hence the speech signal, as being constituted of a finite number of sinusoids, as shown in the block diagram of Fig.11. The frequencies, amplitudes and phase angles of the sinusoids, for each speech frame, are derived from the peaks of its STFT

spectrum. As such, this representation tries to reduce the redundancies of the STFT spectrum but does not really tackle the problems of non-linearity and non-stationarity. Its principal demerit, however, is that it requires the evaluation of a number of parameters. Also, as the process involves peak picking of the STFT spectrum, it is, in essence, a miniature version of the STFT representation of speech [23]. Because of this reason, the sinusoidal representation of speech is not a complete decomposition, i.e., the components derived from it cannot synthesize exactly the same speech signal from which they have been derived.

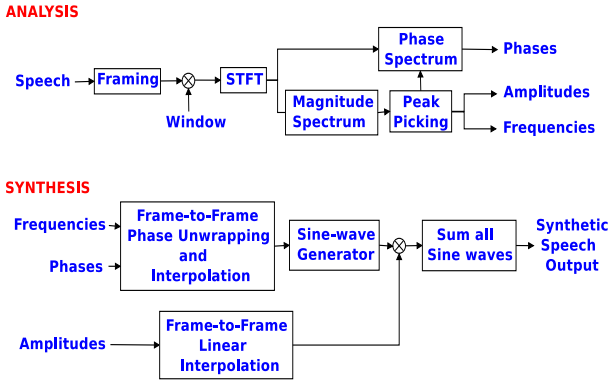


Figure 11: **Figure redrawn from [23]. Analysis and Synthesis process of the sinusoidal model of speech.**

Even though the sinusoidal model did not address the inherent non-linearity of the speech production mechanism, it aroused the possibility that the representation of the speech signal by *a small finite number of meaningful sinusoids* could be an effective mechanism for speech analysis. The next question was how to extract sinusoidal like waveforms from the speech signal without using linear and stationary analysis, i.e., the Fourier spectrum. This leads us to the next development in speech analysis - the AM-FM representation of speech. The AM-FM representation aims to represent the speech signal as the sum of a finite number of narrowband signals, with slowly varying amplitudes and frequencies. Thus, each component of a speech signal, under this representation, is an AM-FM signal, and not a sinusoid, with limited degrees of amplitude and frequency modulation. Ideally, one would want such AM-FM components to be centered around the resonances or the centers of energy of the speech signal [4, 25, 27, 28]. Thus, under AM-FM analysis, a continuous-time speech signal, $s(t)$, may be ideally represented as,

$$s(t) = \sum_{k=1}^{N} R_k(t) \ , \tag{4}$$

$$R_k(t) = a_k(t) \cos\left[2\pi\{f_k t + \int_0^t q(\tau)d\tau\} + \theta\right] \ , \tag{5}$$

where $R_k(t)$ represents an AM-FM signal having a center frequency corresponding to the speech formant frequency f_k. The amplitude and frequency modulating signals of $R_k(t)$ are given by $a_k(t)$ and $q(t)$ respectively, and θ is a constant phase. Henceforth, in order to realize the AM-FM representation, a demodulation technique is required which could estimate the instantaneous amplitude envelope and frequency of each AM-FM component of the speech signal. One of the popular techniques for this purpose is the Hilbert Transform [4–7]. The Hilbert Transform is a reliable estimator of the frequency and amplitude envelope functions of a *monocomponent* signal, provided certain conditions are met [4–7]. These conditions are :

(i) : The frequency variation should not be large, i.e., the signal should be narrowband.
(ii) : The amplitude variation should not be large.
(iii) : The rate of frequency and amplitude variation should not be large.

Assuming these conditions are satisfied, the Hilbert Transform, $H[x(t)]$, of a signal $x(t)$, is derived from its Fourier Transform as ,

$$x(t) \leftrightarrow X(\omega) \ ,$$
$$\frac{1}{\pi t} \leftrightarrow -j \ sgn(\omega) = \begin{cases} -j \ , & \omega > 0 \\ j \ , & \omega < 0 \end{cases} \ ,$$
$$H[x(t)] = x(t) * \frac{1}{\pi t} \ ,$$
$$H[x(t)] \leftrightarrow -j \ sgn(\omega)X(\omega) = \begin{cases} -jX(\omega) \ , & \omega > 0 \\ jX(\omega) \ , & \omega < 0 \end{cases}$$

The instantaneous frequency function, $f(t)$, and amplitude envelope function, $a(t)$, is derived from the analytical signal, $z(t)$, which is devoid of any negative frequency Fourier components.

$$z(t) = x(t) + jH[x(t)] = a(t)e^{j\phi(t)} \ , \tag{6}$$
$$z(t) \leftrightarrow Z(\omega) = \begin{cases} 2X(\omega) \ , & \omega > 0 \\ 0 \ , & \omega < 0 \end{cases} \ ,$$
$$a(t) = |z(t)| \ , \tag{7}$$
$$\phi(t) = \arctan \frac{\Im\{z(t)\}}{\Re\{z(t)\}} \ , \ f(t) = \frac{1}{2\pi}\frac{d\phi(t)}{dt} \tag{8}$$

Correspondingly, the *Discrete Fourier Transform* (DFT) is utilized for estimating the instantaneous frequency and amplitude envelope of any discrete time signal, $x(n)$.

3.1. The Teager Energy Operator and the proof of non-linearity in speech

Even though the Hilbert Transform can track frequency and amplitude variations, it is based on the Fourier Transform, and hence some of the limitations of Fourier analysis are also associated with it. This led to the development of the Teager Energy Operator (TEO), for tracking the instantaneous frequencies and amplitude envelopes of AM-FM signals [4, 24, 26, 27]. The TEO, $\Psi[x(n)]$, of a

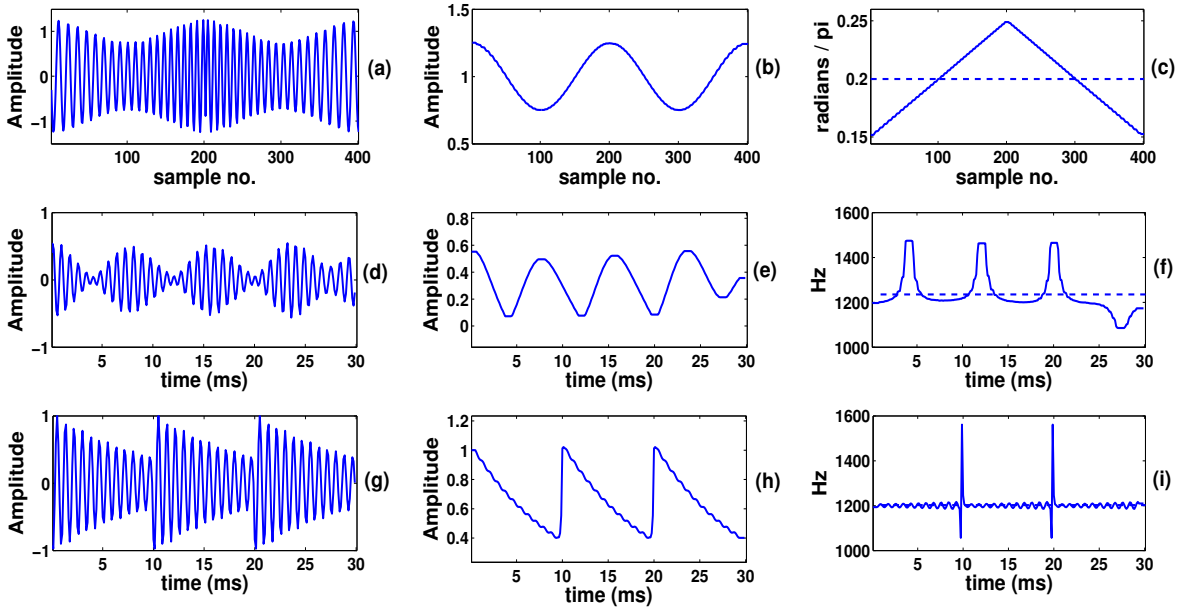


Figure 12: **(a)** $s_{mod}(n)$ **(b) Estimated amplitude envelope of** $s_{mod}(n)$ **using DESA-1 (c) Estimated instantaneous frequency of** $s_{mod}(n)$ **using DESA-1. Dashed line shows average instantaneous frequency. 11-point median filter is used to smooth the estimates ; (d)** $s_{bpf}(n)$ **(e) Estimated amplitude envelope of** $s_{bpf}(n)$ **using DESA-1 (f) Estimated instantaneous frequency of** $s_{bpf}(n)$ **using DESA-1. Dashed line shows average instantaneous frequency. 11-point median filter is used to smooth the estimates ; (g)** $s_{syn}(n)$ **(h) Estimated amplitude envelope of** $s_{syn}(n)$ **using DESA-1 (i) Estimated instantaneous frequency of** $s_{syn}(n)$ **using DESA-1. Dashed line shows average instantaneous frequency.**

discrete-time signal, $x(n)$, is an estimate of the total instantaneous energy of the process generating the signal, and is given by,

$$\Psi[x(n)] = x^2(n) + x(n-1)x(n+1) \tag{9}$$

The Discrete-time Energy Separation Algorithms (DESAs), or the Teager-Kaiser algorithms (TKs), are used to estimate the envelope and frequency functions of discrete-time AM-FM signals. Out of the many DESAs, the more popular DESA-1 algorithm is given by,

$$\omega(n) = \arccos\left\{1 - \frac{\Psi[y(n)] + \Psi[y(n+1)]}{4\Psi[x(n)]}\right\}, \tag{10}$$

$$where\ y(n) = x(n) - x(n-1)\ ,$$

$$|a(n)| \approx \sqrt{\frac{\Psi[x(n)]}{1 - \left\{1 - \frac{\Psi[y(n)]+\Psi[y(n+1)]}{4\Psi[x(n)]}\right\}^2}}\ , \tag{11}$$

where $\omega(n)$ and $a(n)$ are the instantaneous digital frequency and envelope functions estimated from $x(n)$. As with the Hilbert Transform, the same requirements are equally applicable to the DESAs for tracking AM-FM signals [27]. However, the DESAs are much simpler and efficient algorithms than the Hilbert Transform and are free from the limitations of Fourier analysis. The DESAs laid the foundation for the independent investigation of speech signals in terms of their constituent AM-FM signals, without the assumptions of the source-filter theory, and the involvement of Fourier analysis.

To evaluate whether speech is indeed the result of a linear resonator system, a simple experiment is performed [26, 27]. An arbitrary voiced speech signal, $s(n)$, with $F_s = 8$ kHz, is taken, and its formant frequencies are evaluated by a 12-order LP analysis. The signal, $s(n)$, is then band-pass filtered by a Gabor filter around one of its formant frequencies, f_{res}, to obtain a filtered output $s_{bpf}(n)$. In our case, the third formant, with f_{res} =1200 Hz, is considered, and the Gabor filter bandwidth is taken as 400 Hz. Again, a synthetic voiced speech signal, $s_{syn}(n)$, is generated by exciting a single resonator vocal tract system, $v_{syn}(n)$ (having resonant frequency $f_{res} = 1200$ Hz), with a train of impulses having a frequency of 100 Hz.

$$s_{syn}(n) = [-\sum_{-\infty}^{\infty} \delta(n - kN)] * v_{syn}(n),$$

$$V_{syn}(z) = \frac{1}{(1 - p_1 z^{-1})(1 - p_1^{\star} z^{-1})}\ ,$$

$$p_1 = 0.98 \times e^{j2\pi \times 1200/F_s}\ ,\ \frac{F_s}{N} = 100\ Hz\ ,\ F_s = 8\ kHz$$

Also, an AM-FM signal, $s_{mod}(n)$, is generated with grad-

ually varying amplitude envelope and frequency.

$$s_{mod}(n) = \begin{cases} a(n)\cos\{0.2\pi(n-100)+\pi(n-100)^2/4000\}\ , \\ \qquad n = 0, ..., 200 \\ a(n)\cos\{0.25\pi(n-200)+\pi(n-200)^2/4000+\pi\}\ , \\ \qquad n = 201, ..., 400 \end{cases}$$

$$where\ a(n) = 1 + 0.25\cos(\pi n/100)$$

The amplitude envelope and frequency functions of $s_{mod}(n)$, $s_{bpf}(n)$, and $s_{syn}(n)$, are then estimated using DESA-1. Fig.12 shows the plots of the estimates. As $s_{syn}(n)$ is generated from an LTI system, the frequency estimate of $s_{syn}(n)$ is almost constant within an excitation period. Jumps in the frequency function indicate the impulse locations. Similarly, the amplitude envelope is an exponentially decaying function within a pitch period, and jumps occur at excitation instants. In contrast to this, the amplitude envelope of $s_{bpf}(n)$ is a more like a sinusoid, even within a glottal cycle. The frequency function also increases and decreases with time within a pitch period. These characteristics are similar to the estimates obtained from the AM-FM signal $s_{mod}(n)$. These observations suggest that even within a glottal cycle, the speech signal may not be considered as the output of an LTI system, but rather as a combination of AM-FM signals. Such AM-FM components are mainly contributed by the resonances of the vocal tract system, and analyzing them individually might be useful for various speech processing tasks.

3.2. Multiband Demodulation Analysis and the Pyknogram

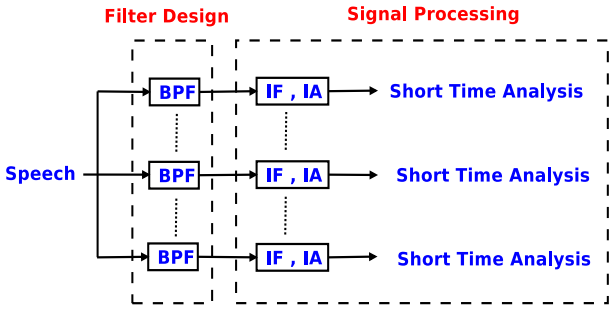


Figure 13: **Multiband Demodulation Analysis. BPF : Band-Pass Filter , IF : Instantaneous frequency , IA : Instantaneous Amplitude Envelope.**

Though the objective of AM-FM analysis is to obtain a representation of the speech signal as a sum of its resonances, which represent narrowband AM-FM signals, there are two obstacles in this formulation. Firstly, how to obtain the formant frequencies without short-time Fourier and LP analysis ? Secondly, how to ensure that the sum of the components adds up to be exactly the same speech signal ? To circumvent the first problem, the framework of Multiband Demodulation Analysis (MDA) was proposed [4, 25, 28]. In this framework, the speech signal is passed through a *fixed parallel bank of overlapping band-pass filters*, as shown in Fig.13. 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, as the filters are overlapping, and not disjoint, the output components may approximate, but will never exactly add up to be the same speech signal. So, just like in sinusoidal analysis, the synthesis is approximately true, if a large number of filters with less overlap is used, but not complete.

As seen in Fig.13, there are two stages in the MDA. The first stage involves the design of the filters, which may vary for different speech processing tasks. Three main questions are to be answered.

(i) : What filter to use, and the number of filters ?
(ii) : The center frequencies of the filters ?
(iii) : The bandwidths of the filters ?

Generally, a Gabor filter, $h_g(t)$, is used for filtering, as it has a low value of time-bandwidth product, and it does not produce sidelobes [25, 27, 28].

$$h_g(t) = e^{-\alpha^2 t^2}\cos(\omega_c t)\ ,$$

$$H_g(\omega) = \frac{\sqrt{\pi}}{2\alpha}\left[e^{-\frac{(\omega-\omega_c)^2}{4\alpha^2}} + e^{-\frac{(\omega+\omega_c)^2}{4\alpha^2}}\right]$$

The second stage involves the processing of the time-domain AM-FM signals, obtained from the band-pass filterbank. The instantaneous amplitude envelope and frequency function of each AM-FM signal is estimated using the DESA or the Hilbert Transform. They are then utilized to obtain short-time estimates of mean instantaneous frequency and bandwidth. Two most used formulations are the mean amplitude weighted instantaneous frequency, F_w, and the mean amplitude weighted instantaneous bandwidth, B_w, given by,

$$F_w = \frac{\int_{t_0}^{t_0+T} f(t)a^2(t)dt}{\int_{t_0}^{t_0+T} a^2(t)dt}\ , \tag{12}$$

$$B_w^2 = \frac{\int_{t_0}^{t_0+T}[\{\dot{a}(t)/2\pi\}^2 + \{f(t)-F_w\}^2 a^2(t)]dt}{\int_{t_0}^{t_0+T} a^2(t)dt}\ , \tag{13}$$

where t_0 and T are the start and duration of the analysis frame, and $a(t)$ and $f(t)$ are the instantaneous amplitude envelope and frequency respectively of the AM-FM signal under consideration.

Based on the philosophy and methodology discussed above, AM-FM analysis have been applied successfully in different speech processing tasks, particularly in the fields are speech and speaker recognition [4, 46, 48, 49, 89–91]. The usefulness of AM-FM analysis may be appreciated in an important time-frequency representation of the speech signal - the *speech pyknogram* [28]. Fig.14 shows a typical pyknogram constructed from a speech signal of F_s =

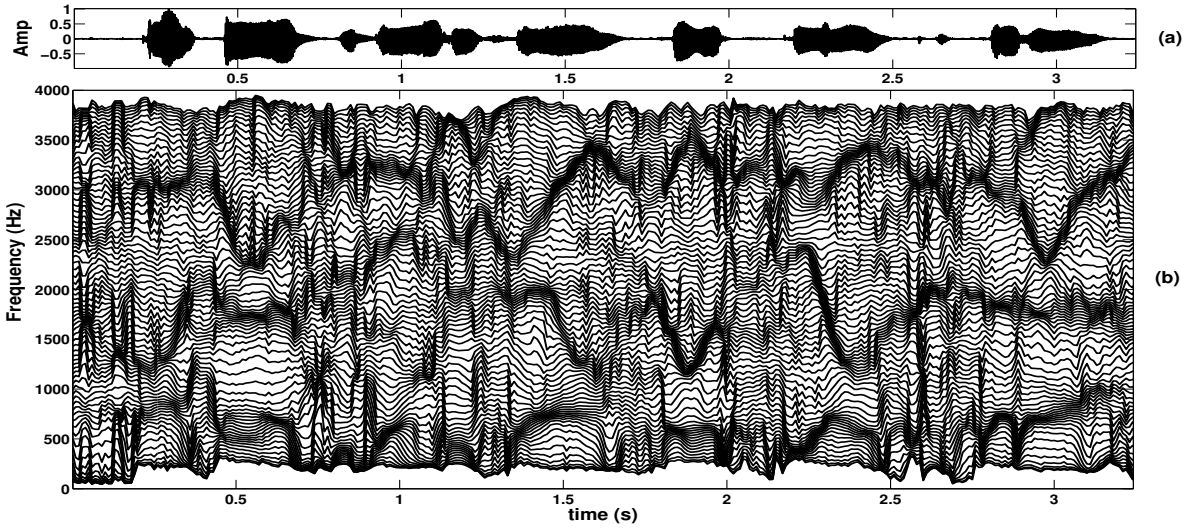


Figure 14: **(a) Speech ; (b) Pyknogram of** (a) **using 80 Gabor bandpass filters of 1000 Hz effective RMS bandwidth. Framesize of 25 ms with frameshift of 10 ms is used.**

8 kHz. It is formed by MDA of the speech signal using 80 Gabor filters uniformly spaced in the linear frequency scale. The Gabor filters have an effective RMS bandwidth of 1000 Hz. For every speech component obtained from MDA, corresponding to a Gabor filter with center frequency cf_k , $k = 1, ..., 80$, a short time frequency estimate $F_w(t_{fr}, cf_k)$ is obtained at every 10 ms time interval t_{fr}, using equation (12). The time duration of the analysis frame is taken as $T = 25\ ms$. This results in the time-frequency distribution $F_w(t, f)$ called the pyknogram. The Greek word "pykno" means dense. As can be seen from Fig.14, the dense clusters of curves in the pyknogram indicate the trajectories of different formants. It is a much more vivid representation than the spectrogram of Fig.4, for the same speech file, created using the same time resolution and analysis window size. Henceforth, the pyknogram is processed to identify regions with dense clusters for the purpose of tracking formant frequencies and their bandwidths [28].

The above discussion suggests how AM-FM analysis could provide useful information for speech processing. However, it still remains as a *fixed* analysis, determined by the design of the filterbank. As such, apart from the useful components which carry the formants information, a multitude of other components are also generated, as seen in the pyknogram. The ideal objective of AM-FM analysis, to represent the speech signal in terms of its time-varying resonances only, as encapsulated in equation (4), is not still desired. With the objective of making AM-FM analysis a truly adaptive and compact analysis, we look towards EMD of speech.

4. Empirical Mode Decomposition

The limitations of conventional short-time processing of speech, apart from the fixed time-frequency resolution, lies in the inability to capture the dynamics of the speech signal [92–95]. As such, the first derivative (Δ or *velocity* coefficients), and the second derivative ($\Delta\Delta$ or *acceleration* coefficients) are often utilized on top of the normal speech processing features, as in the case of MFCCs [92]. To minimize this limitation, efforts have also been made to utilize dynamic frame rate and length based on the time-varying properties of the speech signal [93–95]. However, none of these solutions which try to capture the dynamics of the speech signal, just like WT, cater to the problem of non-linearity of the speech signal. Though traditional AM-FM analysis, based on the MDA, can capture the non-linear dynamics of speech, it acts as a fixed filterbank, producing a lot of redundant components as well. As such, there have been efforts to break this fixed filterbank structure, to use alternative methods than the MDA, to make AM-FM analysis adaptive to the speech signal [96, 97]. But, more is desired. All the above efforts, again, are not adaptive to the every speech signal that is being analyzed. Thus, there is a definite requirement of a technique for speech processing, which is more effective than the currently available tools, yet which is adaptive, and does not complicate the analysis. The various features that such a technique is required to possess may be summarized as follows :

• **(i) Complete, compact and adaptive decomposition :** Fourier analysis gives a complete decomposition, i.e., the sum of the components add up to be exactly the same speech signal. But, it is limited in analyzing signals which are the output of non-linear and non-stationary processes. On the other hand, AM-FM analysis relies on a large fixed overlapping bank of filters to extract signal components. Thus, a non-linear and non-stationary signal decomposition technique is required, which is data adaptive, has little complexity, and produces time domain

components which add up to be the exact same speech signal. The number of such components needs to be countably finite.

• **(ii) Unique and Meaningful Components :** The information in speech resides in the various resonant structures that shape and modulate the air flow passing through the voice production apparatus, and the glottis that controls the amount of air flowing out through the apparatus. Thus, the desired decomposition technique should be able to extract time-domain components of speech, which carry the formants information, and the glottal source information of speech. The information carried by them, ideally, should not overlap. In short, the components should carry unique and meaningful information of the speech signal.

• **(iii) No short-time processing and parameter computation :** The components should be obtained from the desired decomposition without any short-time processing of the speech signal, and without any short-time computation of parameters, unlike in the case of sinusoidal analysis.

• **(iv) Reliable instantaneous frequency and amplitude estimate :** The components derived from the desired decomposition should be narrowband, and have limited fluctuations in amplitude and frequency so that reliable estimates of instantaneous frequency and amplitude envelope could be obtained from either the DESA or the Hilbert Transform.

Thus, a data-adaptive and complete analysis technique is required which can decompose the speech signal into a finite number of meaningful time domain components, without the requirement of the assumptions of short-time linearity and stationarity, such that reliable instantaneous amplitude envelope and frequency estimates could be obtained from them. With the aim of achieving these goals, we look forward towards the technique of *Empirical Mode Decomposition* (EMD) [7, 29, 98], for processing speech signals. EMD is a method that decomposes a signal into *oscillatory* or AM-FM components, called *Intrinsic Mode Function* (IMFs), in a completely data-driven manner without the requirement of any a priori basis. Fig.15 shows the flowchart of the EMD process. The pseudocode for the same is given below :

Pseudocode for EMD : Let $s(t)$ be a continuous-time speech signal.

•**(i)** Let $r_0(t) = s(t)$. We subject an *outer residue*, $r_{k-1}(t)$, to a *sifting process* to obtain an IMF, $h_k(t)$, and another *outer residue*, $r_k(t)$, from it. In other words, if k represents the index of the sifting process, then, the k^{th} sifting process decomposes the $(k-1)^{th}$ outer residue, $r_{k-1}(t)$, into the k^{th} IMF, $h_k(t)$, and the k^{th} *outer residue*, $r_k(t)$.

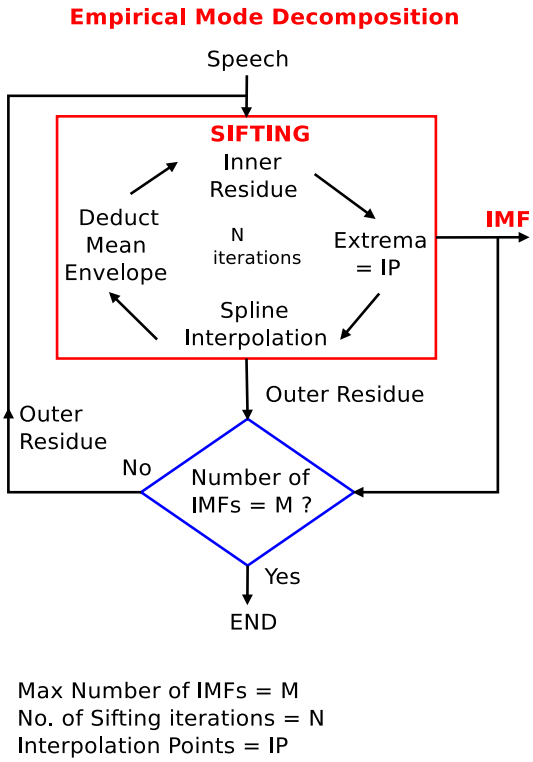


Figure 15: **Flowchart of Empirical Mode Decomposition**

The *sifting process* for EMD is given as :
Let $h^0_{k-1}(t) = r_{k-1}(t)$. Repeat the following steps for each *sifting iteration*. Let n represent the sifting iteration index, where $n = 1, ..., N$.

$\star$(**a**) Given the *inner residue* signal $h^{n-1}_k(t)$, find the maxima and minima locations of $h^{n-1}_k(t)$. These locations are to be used as x-coordinates of the *Interpolation Points* (IPs), to be used for cubic spline interpolation.

$$t_{max} = \{t : \frac{d}{dt}h^{n-1}_{k-1}(t) = 0, \frac{d^2}{dt^2}h^{n-1}_{k-1}(t) < 0\} ,$$

$$t_{min} = \{t : \frac{d}{dt}h^{n-1}_{k-1}(t) = 0, \frac{d^2}{dt^2}h^{n-1}_{k-1}(t) > 0\}$$

$\star$(**b**) Obtain the y-coordinates of the IPs from $h^{n-1}_k(t)$.

$$y_{max} = h^{n-1}_{k-1}(t_{max}) \ , \ y_{min} = h^{n-1}_{k-1}(t_{min})$$

$\star$(**c**) Create the maxima envelope $e_{max}(t)$ using cubic spline interpolation, with the IPs as $\{t_{max} \ , \ y_{max}\}$. Create the minima envelope $e_{min}(t)$ using cubic spline interpolation, with the IPs as $\{t_{min} \ , \ y_{min}\}$. Deduce the mean envelope $e(t)$ as,

$$e(t) = \frac{e_{max}(t) + e_{min}(t)}{2}$$

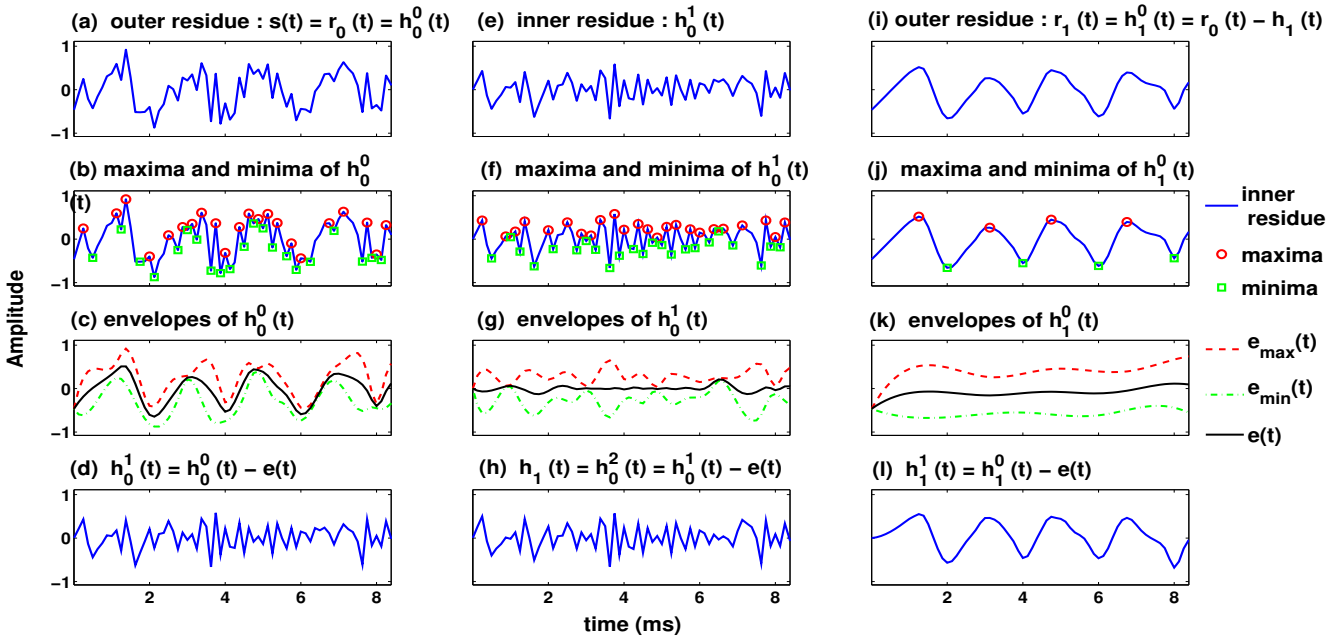


Figure 16: **Simulation of the EMD algorithm using a noisy sinusoidal signal** $x_g(t)$.

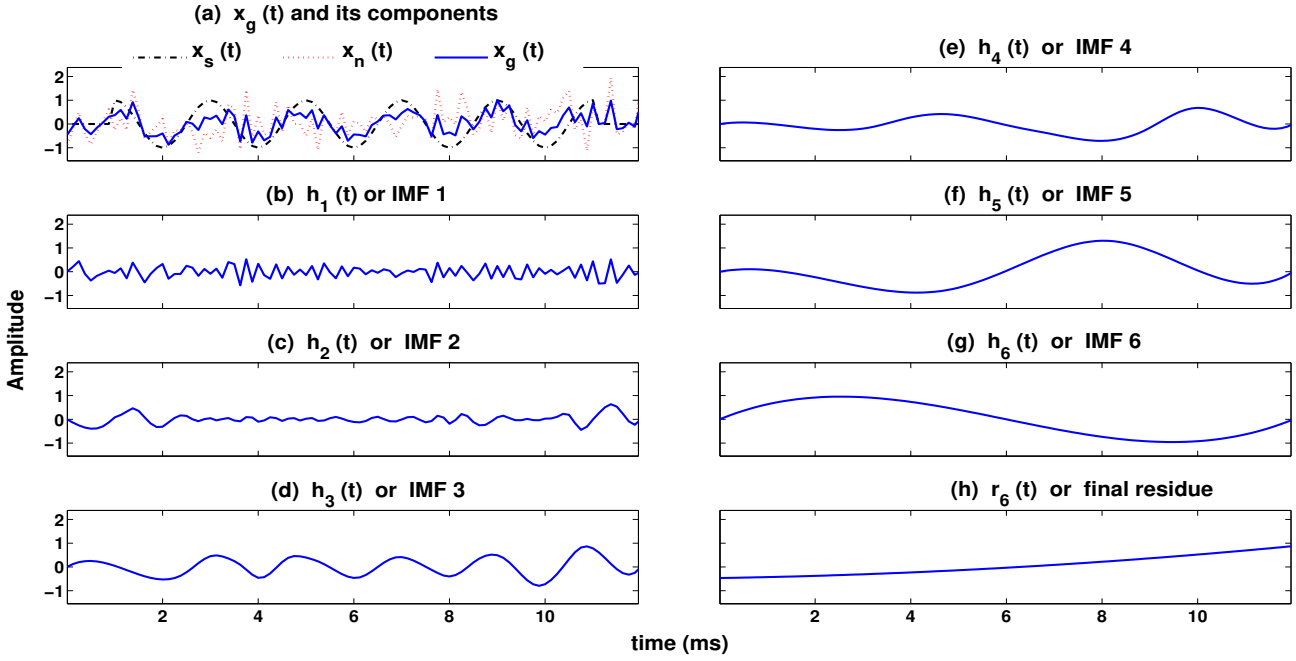


Figure 17: **IMFs obtained from EMD of** $x_g(t)$. $N = 10$ **and** $M = \infty$ **are considered.**

⋆(**d**) $h^n_{k-1}(t) = h^{n-1}_{k-1}(t) - e(t)$. Go to step (a). Stop when n = N.

•(**ii**) Set $h_k(t) = h^N_{k-1}(t)$. Obtain $r_k(t) = r_{k-1}(t) - h_k(t)$.

•(**iii**) Go to step (i). Ideally, the decomposition is to be stopped when the *outer residue* takes the form of a trend, i.e., the number of extrema in $r_k(t)$ is 2 or less [7, 29, 98]. Practically, however, the decomposition may be stopped when a user-defined maximum number (M) of AM-FM components, i.e., the IMFs, have been extracted, as shown in Fig.15.

$$s(t) = r_M(t) + \sum_{k=1}^{M} h_k(t) \tag{14}$$

For a digital speech signal, the decomposition may be rep-

resented as

$$s(n) = r_M(n) + \sum_{k=1}^{M} h_k(n) \tag{15}$$

Equations (14) and (15) represent the decomposition of the signal in terms of its IMFs and its final residue, which is a trend-like signal.

To illustrate the mechanism involved, we use a signal $x_g(t) = x_s(t) + x_n(t)$, where $x_s(t) = \cos(2\pi * 500 * t)$, and $x_n(t)$ is a zero mean Gaussian white noise signal such that the *Signal to Noise Ratio* (SNR) is 0 dB. Fig.16 shows the working mechanism of EMD on $x_g(t)$. In this example, we consider $N = 2$ sifting iterations per sifting process. The first sifting process is completely illustrated, resulting in the first IMF, $h_1(t)$, after $N = 2$ sifting iterations. A new outer residue, $r_1(t)$ is also thereby obtained. The first iteration of the sifting process applied on the new outer residue is also shown.

Fig.17 shows the IMFs obtained from EMD of $x_g(t)$, where $N = 10$ sifting iterations are used per sifting process. The decomposition is allowed to stop naturally by keeping no maximum limit ($M = \infty$) on the number of IMFs. Under this condition, the decomposition stops automatically when the final residue has insufficient extrema to construct the maxima and minima envelopes. For the noisy sinusoid, $x_g(t)$, the final residue is obtained after 6 IMFs have been extracted, as shown in the figure. One can easily observe the similarity between IMF_3 and the pure sinusoid $x_s(t)$, and that between IMF_1 and the white noise signal $x_n(t)$, which shows the ability of EMD to segregate the components of a signal.

4.1. The importance of the sifting process

As explained above, the EMD process results in a finite number of time-domain components, $h_k(t)$, $k = 1, ..., M$, called IMFs, and a final residue signal $r_M(t)$, which is the low-frequency trend of the signal [7, 29, 98]. An IMF is defined as a signal having the following properties.

(i) The number of extrema and the number of zero-crossings in an IMF must either be equal or differ at most by one.

(ii) At any point, the mean value of the envelope defined by the local maxima, and the envelope defined by the local minima, is zero.

Thus, the aim of EMD is to obtain *oscillatory functions* from the signal. The process of *sifting* is designed to achieve this purpose. As mentioned earlier, both the Hilbert Transform and the TEO require the signal to be narrowband (ideally monocomponent), with limited degrees of frequency and amplitude modulation, for accurate demodulation. The above-mentioned properties of an IMF make it locally narrowband and symmetric, which enables accurate demodulation of their instantaneous frequency and amplitude envelopes. Having applied EMD on a synthetic signal, we now apply it to a natural speech signal. Fig.18 shows the first 5 IMFs obtained from a natural speech signal, where $N = 10$ sifting iterations have been used per sifting process, and the decomposition is curtailed at $M = 9$. The second plot of the figure shows the *Electroglottograph* (EGG) signal [99, 100] corresponding to the speech signal. The EGG signal represents a measurement of the movements of the vocal folds during the production of voiced speech. As can be seen from the figure, there is a strong similarity between IMF_4 and the EGG signal, which reflects the ability of EMD to extract information about the glottal source producing the speech signal [101]. This shows the ability of EMD to extract latent information from the signal in its IMFs. Again, the entire sifting process involves no a priori basis function. Further, the process is carried out on the entire data stream, and no parameter computations are involved.

4.2. Hilbert Huang Transform as a generalized Fourier Transform

Having derived the IMFs from the signal, they are represented in terms of their instantaneous amplitude envelopes and frequencies using the Hilbert Transform. This entire process of extracting IMFs from the data, and representing them in terms of their instantaneous amplitude envelopes and frequencies, is termed as Hilbert Huang Transform (HHT) [7, 29, 98, 102]. We have, from equation (14),

$$s(t) = \sum_{k=1}^{M} h_k(t) \; + \; r_M(t) = \sum_{k=1}^{M+1} h_k(t) \tag{16}$$

Each component, $h_k(t)$, derived from the signal, can then be represented using the Hilbert Transform, using equations (6)-(8), as,

$$h_k(t) \xrightarrow{Hilbert \;\; Transform} a_k(t)e^{j\theta_k(t)} \;,$$
$$h_k(t) = \Re\Big\{a_k(t)e^{j\theta_k(t)}\Big\} \;,\; f_k(t) = \frac{1}{2\pi}\frac{d}{dt}\theta_k(t) \;, \tag{17}$$

where $a_k(t)$ and $f_k(t)$ represent the instantaneous amplitude envelope and frequency of $h_k(t)$. The signal can then be represented as,

$$s(t) = \Re\Big\{ \sum_{k=1}^{M+1} a_k(t)e^{j2\pi \int f_k(t)dt} \Big\} \tag{18}$$

The standard Fourier representation of the same signal is given by,

$$s(t) = \int_{-\infty}^{\infty} S(\omega)e^{j\omega t}d\omega = 2\pi \int_{-\infty}^{\infty} S(f)e^{j2\pi ft}df \tag{19}$$

A comparison of equations (18) and (19) shows that HHT is a generalized Fourier Transform, without its limitations. HHT represents the signal in terms of a finite number of components, unlike Fourier Transform. While

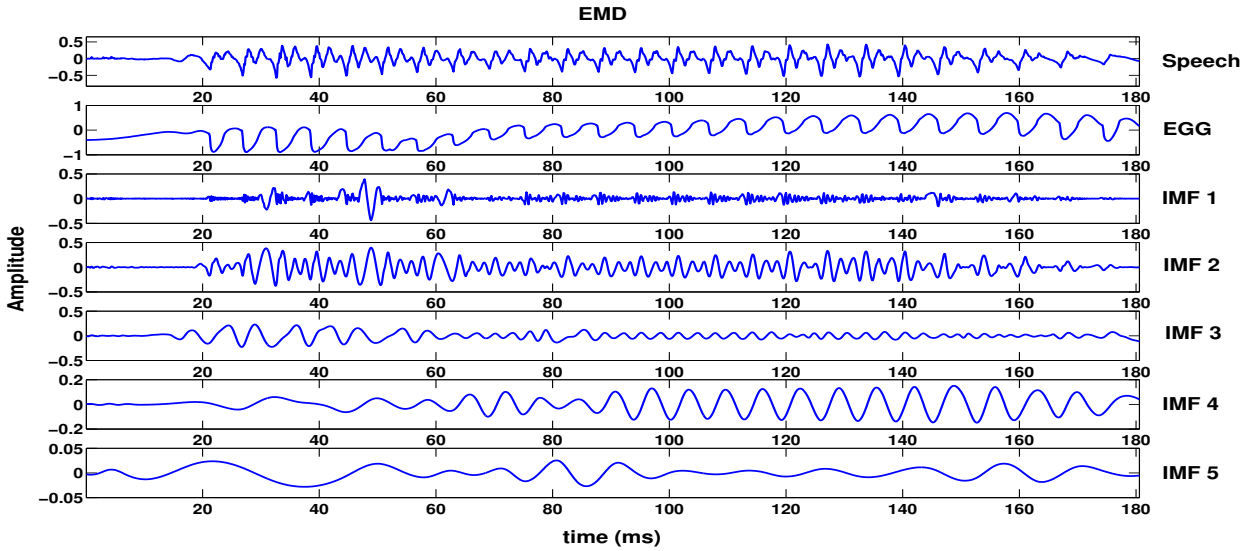


Figure 18: **(a) Speech ; (b) EGG ; (c)-(g) are IMFs 1-5 of the speech signal, obtained using EMD.**

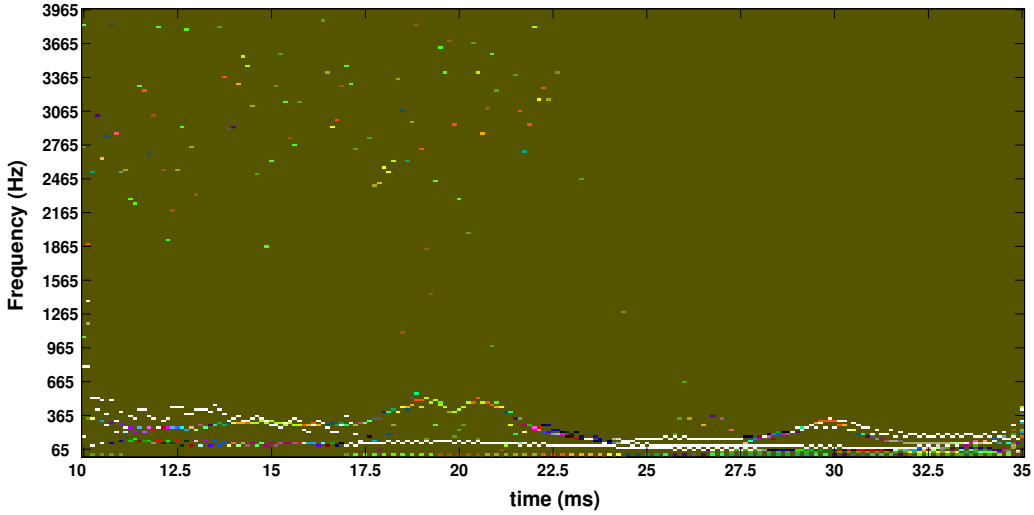


Figure 19: **Hilbert spectrum of the speech signal used in Fig18, using EMD.**

the amplitude envelope and frequency of each component in Fourier representation is constant for an infinite time duration, it is time varying in the case of HHT. HHT is a complete, compact and adaptive Fourier representation of the signal [7, 29, 98, 102]. This formulation, when presented in terms of an image, is called the Hilbert Spectrum [7, 29, 98, 102]. The Hilbert spectrum can be defined as the time-frequency distribution of the instantaneous energy envelope, which is the squared magnitude of the amplitude envelope. In general, the last few components, which are low-frequency trend-like waveforms, are excluded from the spectrum, as they have high energy and obscure the image [7, 29, 98, 102]. Fig.19 shows the Hilbert spectrum for a section of the speech signal used in Fig.18. As is evident from the spectrum, most of the energy in the spectrum lies within 60-500 Hz, which is the *pitch frequency* range, i.e., the frequency range of vibration of the vocal folds or the glottal source. As such, this spectrum can be easily post-processed to obtain the instantaneous pitch frequency [103].

$$H(f,t) = \left\{ a_k^2(t) \mid f_k(t), t \right\} \ , \ k = 1, ..., K \leq M \tag{20}$$

From the Hilbert spectrum, the marginal Hilbert spectrum can be derived as,

$$h(f) = \int_{t=0}^{T} H(f,t) dt \tag{21}$$

The marginal Hilbert spectrum gives the probability that an oscillation of frequency f could have occurred locally at some time during the entire duration (T) of the signal. Similarly, the instantaneous energy density can be computed from the Hilbert spectrum, which reflects the energy fluctuations in the signal with respect to time

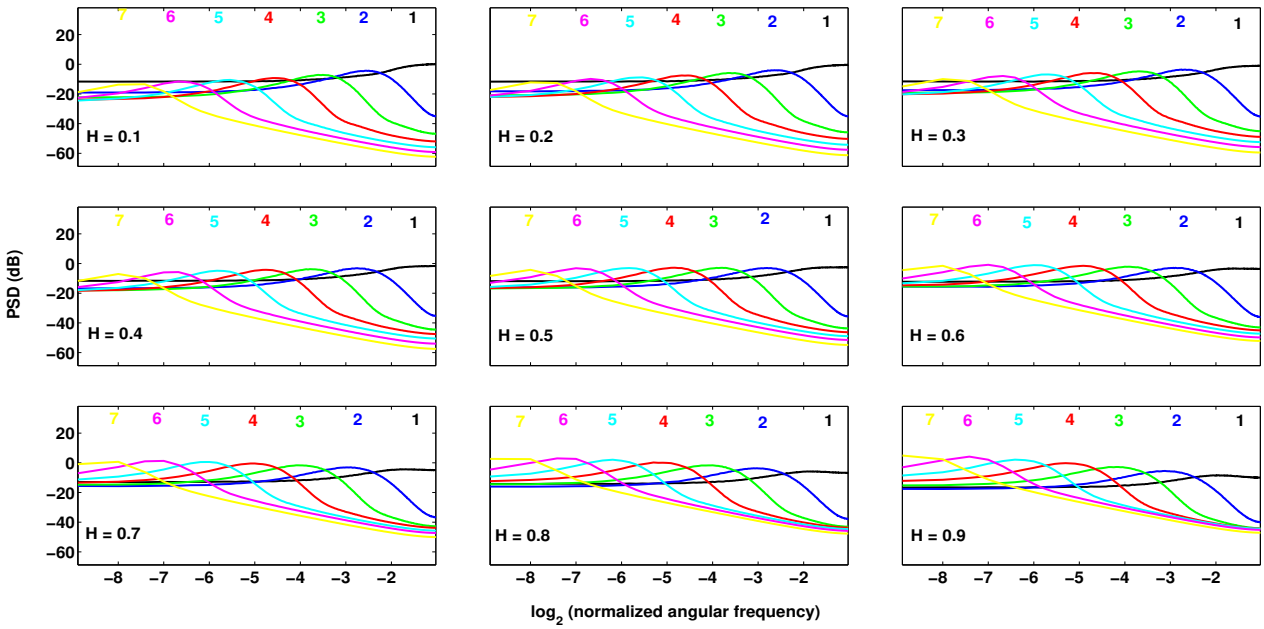


Figure 20: **IMF power spectra in the case of fractional Gaussian noise, for Hurst exponent H = { 0.1 , 0.2 , ... , 0.9 }. The estimated power spectrum densities (in dB) is plotted as a function of the logarithm of the normalized frequency for the first 7 IMFs. The IMF number is mentioned above the peak of the corresponding power spectrum. For each of the nine H values, the spectral estimates have been computed on the basis of 5000 independent sample paths of 512 data points.**

[7, 29, 98, 102].

$$IE(t) = \int_f H(f,t)df \tag{22}$$

4.3. The dyadic filterbank nature of EMD

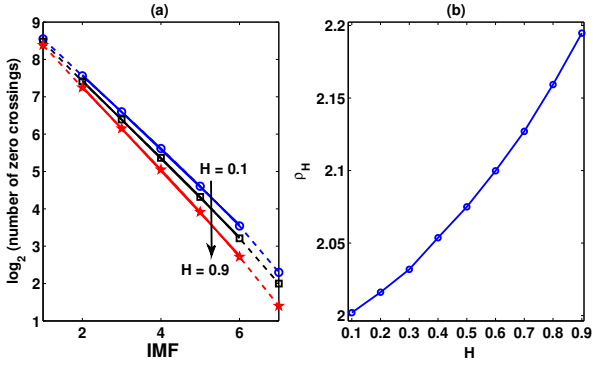


Figure 21: **(a) IMF average number of zero-crossings in the case of fractional Gaussian noise. For clarity, only those curves corresponding to H = 0.1 (bubbles), H = 0.5(squares) and H = 0.9 (stars) have been plotted in the diagram; the remaining cases lead to regularly intertwined similar curves. The superimposed solid lines correspond to linear fits within the IMF range k = 2 to 6. (b) Corresponding decrease rate of zero-crossings.**

While EMD is an effective decomposition process, it is an algorithm without a solid mathematical framework. Even though efforts have been made to provide some mathematical representation [104], deducing conclusions about its behavior is not straightforward. To have a better understanding of the behavior of the process, studies were carried out on the decomposition of noise by EMD, by Wu and Huang, and Flandrin et al., separately [29, 102, 105–109]. For the experiments, fractional Gaussian noise (fGn) was used. The autocorrelation function of a fractional Gaussian noise sequence $x_H(n)$ is given by,

$$r_H(m) = \mathbb{E}\Big[x_H(n)x_H(n+m)\Big] \ ,$$
$$r_H(m) = \frac{\sigma^2}{2}\Big\{|m-1|^{2H} - 2|m|^{2H} + |m+1|^{2H}\Big\} \tag{23}$$

As is evident from equation (23), the parameter H, $0 < H < 1$, called the *Hurst component*, controls the nature of the signal. For $H = 0.5$, fGn becomes a white Gaussian noise sequence. For $0 < H < 0.5$, the power spectrum of fGn is of high-pass nature, whereas $0.5 < H < 1$ produces a fGn sequence of low-pass nature. For the experiments, 5000 realizations of fGn were generated for each of the 9 H values given by $H = 0.1, 0.2, ..., 0.8, 0.9$. The fGn sequences were of 512 samples length. Each fGn sequence was then decomposed by EMD, and the properties of the first 7 IMFs were then examined.

Fig.20 shows the plots of the power spectra (averaged for 5000 fGn sequences) of the IMFs, for $H = \{0.1,\ 0.2,\ ...,\ 0.9\}$, The plots show that barring the first IMF, the rest of the IMFs (IMFs 2-7), for all the H values, have power spectra having band-pass nature. The frequencies corresponding to the peaks of these band-pass spectra, approximately decrease by a factor of two as the

IMF order increases. In other words, starting from the 2^{nd} IMF, EMD acts as a *dyadic filterbank* on the signal [29, 102, 105–109]. The 1^{st} IMF exhibits a weak high-pass spectrum for all the H values. The IMFs being locally zero mean signals, the number of zero crossings of an IMF could be used to estimate the dominant frequency that it represents. Fig.21 shows the plots of the number of zero crossings of the IMFs vs the IMF number. As can be seen from the figure, the curves in Fig.21(a) have an approximate slope of -1, which means that the dominant frequency reflected in the IMFs decrease by a factor of 2 with respect to the IMF order. The average decrease rate of the number of zero-crossings, ρ_H, for each of the H values, is plotted in Fig.21(b), which ascertains this observation. Thus, if $z_H(k)$ represents the number of zero-crossings of the k^{th} IMF, we have,

$$z_H(k') = \rho_H^{(k'-k)} z_H(k),\ k' > k \geq 1\ , \tag{24}$$
$$\rho_H = 2.01 + 0.2(H - 0.5) + 0.12(H - 0.5)^2\ , \tag{25}$$
$$\rho_H \approx 2 \tag{26}$$

Given that the dominant frequencies of the IMFs decrease by a factor of 2 with the IMF order, the power spectral density of the bandpass IMFs can then be approximately related to one another as,

$$S_{k',H}(f) = \rho_H^{\alpha(k'-k)} S_{k,H}(\rho_H^{[k'-k]} f)\ , \tag{27}$$
$$k' > k \geq 2,\ \rho_H \approx 2,\ \alpha = 2H - 1 \tag{28}$$

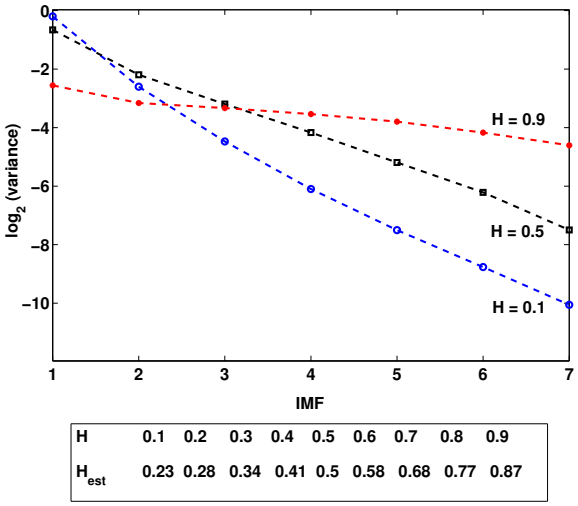
Fig.22 plots the variance of the IMFs vs the IMF number, for for $H = \{0.1,\ 0.5,\ 0.9\}$. As can be seen from the figure, the variances of the IMFs decrease with respect to the IMF number, at a rate dependent on the H value. For the case of white noise ($H = 0.5$), the slope of the curve is approximately -1, i.e., the IMF energy decreases by a factor of 2 with increasing IMF order.

$$V_H(k') = \rho_H^{(\alpha-1)(k'-k)} V_H(k)\ ,$$
$$\rho_H \approx 2,\ \alpha = 2H - 1,\ k' > k \geq 2\ ,$$
$$V_H(k') = \rho_H^{2(H-1)(k'-k)} V_H(k),\ k' > k \geq 2 \tag{29}$$

The slope (κ_H) of the log-linearized version of equation (29) can be used to estimate the Hurst component [106–108] as,

$$V_H(k') = C\rho_H^{2(H-1)k'},\ k' \geq 3\ , \tag{30}$$
$$\log_2 V_H(k') \approx \log_2 C + 2(H-1)k',\ k' \geq 3\ , \tag{31}$$
$$\log_2 V_H(k') \approx C' + \kappa_H k',\ k' \geq 3\ , \tag{32}$$
$$H_{est} = 1 + \frac{\kappa_H}{2} \tag{33}$$

The estimated Hurst component values, H_{est}, obtained by this process, are shown in Fig.22.



H	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
H_{est}	0.23	0.28	0.34	0.41	0.5	0.58	0.68	0.77	0.87

Figure 22: **Estimated IMF** $log_2(variance)$ **in the case of fractional Gaussian noise, for hurst component H = { 0.1 , 0.5 , 0.9 }. The values of the empirical (energy-based) variance estimates are given for all the 9 H values.**

4.4. Some aspects of EMD

Apart from the fact that EMD does not have a robust mathematical framework, there are some aspects of the decomposition that are to be catered to. One of them is the sampling rate of the signal. As is evident from the flowchart of EMD, Fig.15, and the pseudocode, the detection of extrema is crucial to the process. Again, the instantaneous frequency derived using the Hilbert Transform depends on differentiation with respect to time, as is evident from equations (8),(17). For these reasons, it is beneficial for the decomposition if the signal is sampled at much above the Nyquist rate [7, 29, 98, 110]. Another issue that needs to be catered to are the *"end-effects"*, which are large swings that occur at the ends of the IMFs due to cubic spline fitting. By zero-padding the ends of the signal, however, such effects could be curtailed to a certain extent [7, 29, 98]. More details on *"end-effects"*, and how to curtail it could be found in [111]. In the case of speech, which has silence regions at the beginning and end of the signal, the *"end-effects"* are not too concerning. Apart from this, the *sifting criterion* is another aspect that needs attention. If the number of sifting iterations is low, the decomposed signals would not be eligible to be IMFs, which would cause erroneous instantaneous frequency and amplitude envelope estimates. On the other hand, oversifting would result in the smoothing of the IMFs, and thus they would become more like sinusoids and may lose the information they are supposed to carry. A number of *sifting criteria* have been proposed to ascertain that the IMFs adhere to their defined properties [7, 29, 98, 112–114]. All of them need some parametric tuning, and none of them may be deemed significantly better than the other. Amongst them, one of the more recent and popular criterion is the *stopping criterion* proposed in [112]. This criterion is based on minimizing

the parameter $\gamma(t) = \left|\frac{[e_{max}(t)+e_{min}(t)]/2}{[e_{max}(t)-e_{min}(t)]/2}\right|$, for ascertaining globally small fluctuations of the mean envelope signal even for locally large fluctuations of the signal. Two thresholds θ_1 and θ_2 are used to control the number of iterations, N, in every sifting process. When $\gamma(t) < \theta_2$ for a fraction α of the duration of the signal, and $\gamma(t) < \theta_1$ for the remaining fraction of the duration of the signal, the sifting process is stopped. Default values of the parameters are : $\alpha \approx 0.05, \theta_1 \approx 0.05, \theta_1 = 10\theta_2$. In general, it has been found that the *dyadic filterbank* nature of EMD is well maintained, for fractional Gaussian noise with both flatband and skewed spectra, if the number of sifting iterations is around 10 [114, 115].

Besides the above mentioned marginal issues, there are two major aspects of EMD that need to be discussed, particularly for decomposing speech signals :

(i) Ability to separate frequency components.
(ii) Mode-mixing.

To examine the ability of EMD to separate different frequency components in the signal, a signal, $x(t)$, composed of a lower frequency sinusoid, $x_l(t)$, and a higher frequency sinusoid, $x_h(t)$, is considered [116].

$$x(t) = x_l(t) + x_h(t) \ ,$$
$$x(t) = a_l \cos(2\pi f_l t + \phi_l) + a_h \cos(2\pi f_h t + \phi_h) \ ,$$
$$f_l, f_h \ll F_s \ ,$$

where F_s is the sampling frequency. To simplify the experiment, and without any loss of generality, x(t) is considered as,

$$x(t) = x_l(t) + x_h(t) \ ,$$
$$x(t) = a\cos(2\pi f t + \phi) + \cos(2\pi t) \ , \tag{34}$$
$$f = \frac{f_l}{f_h} \in \ ]0, 1[\ , \ a = \frac{a_l}{a_h}, \ \phi = \phi_l - \phi_h$$

The signal $x(t)$ is decomposed by EMD, and then the following parameter is computed,

$$c_1^{(N)}(a, f, \phi) = \frac{||d_1^{(N)}(a, f, \phi) - \cos(2\pi t)||_2}{||a\cos(2\pi f t + \phi)||_2} \ , \tag{35}$$

where $d_1^{(N)}(a, f, \phi)$ is the first IMF obtained from the decomposition of $x(t)$, where N sifting iterations have been used in the sifting process. In the experiment $N = 10$ is used, whereas ϕ is kept constant. The parameter $c_1^{(10)}(a, f, \phi)$ is averaged over different values of $\phi \sim [0, 2\pi)$. Thus, $c_1^{(10)}(a, f, \phi)$, represents a function of the frequency and the amplitude ratios, f and a, respectively, where $||.||_2$ denotes the Eucledian norm. $c_1^{(10)}(a, f, \phi)$ gives a measure of whether EMD could successfully extract the components of the signal, $x(t)$, or not [116].

Fig.23 plots the values of $c_1^{(10)}(a, f, \phi)$ as an image, with f and a being the independent variables. The whiter regions of Fig.23 indicate that the components have been properly extracted, whereas the darker shades indicate the combinations of f and a, where proper decomposition could not be achieved by EMD. As can be seen from the figure, for EMD to successfully decompose the signal into its actual constituents, there is a dependency on both f and a. There is a hard cut-off, $f \lesssim 0.67$, irrespective of a, only below which the constituents can be adequately segregated. Also, even within this limit, the performance of segregation decreases as the lower frequency component becomes stronger than the higher frequency component. Ideally, for proper segregation of the components, $a \lesssim 1$ is required [116].

This simple experiment of segregating the sinusoidal constituents of the signal gives us an idea about the difficulties involved in extracting the *true* components of a non-linear and non-stationary signal like speech. To add to the problem, most of the energy of the speech signal is present in its voiced regions, which have a high spectral slope of -6 dB/octave [1–3]. This causes the higher frequency spectrum of speech to be overshadowed by its lower frequency spectrum. Thus, the characteristics of the speech spectrum are not in tune with the requirements of the amplitude ratio, needed for successful segregation of its components by EMD. Due to this fact, EMD is limited in extracting meaningful IMFs, which characterize the higher frequency content of speech. Henceforth, as is discussed later, most of the *vocal tract resonances* or *formants* of voiced speech are captured by the first IMF alone [117, 118]. The second IMF captures the first formant only, and the rest of the IMFs are of a lower frequency and represent the glottal source information [101, 117, 118].

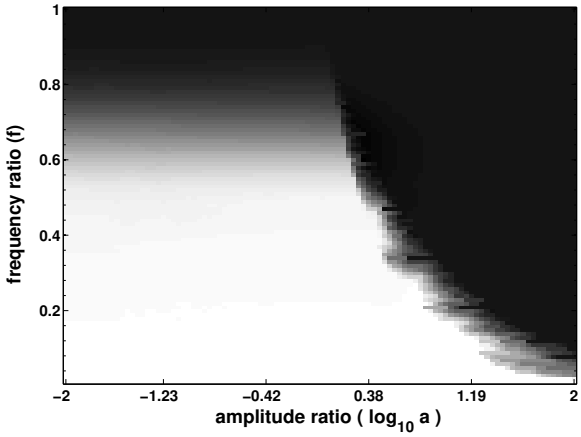


Figure 23: **2-D projection of $c_1^{(10)}(a, f, \phi)$ onto the (a, f) plane of amplitude and frequency ratios is plotted.**

Thus, like any other technique, EMD also has its due share of limitations. However, the most important phenomenon in the EMD decomposition is the phenomenon of *mode-mixing*. *Mode-mixing* may be defined as the presence of *disparate frequency scales* within an IMF, and/or the presence of the same frequency scale in multiple IMFs [7, 29, 98, 102, 112, 115]. It is vividly observed in the

case of non-stationary signals in which oscillations of *disparate frequency scales* occur intermittently. In reality, *mode-mixing* is not unexpected, as EMD is designed to locally separate a signal into low and high-frequency components. However, for many applications, this phenomenon may hinder the utility of the IMFs, and one may instead want IMFs which have a narrower frequency spectrum, as is desired ideally in AM-FM analysis. As an example of this phenomenon, we may consider the IMFs extracted from the speech signal in Fig.18. As is seen from the figure, IMF_1, which mostly consists of higher frequency oscillations, is corrupted in between by lower frequency signals. Similarly, the primary frequency scales reflected in IMFs 2-4 seem to be distributed amongst them. There are parts of IMF_3 that appear to have the same frequency scale as that of major parts of IMF_4. Similarly, parts of IMF_2 seem to carry a low amplitude oscillation, which is mainly present in IMF_3. If such frequency variations are too large within an IMF, then the instantaneous frequency and amplitude envelope obtained from it, will not be reliable, as discussed earlier.

5. Developments of EMD

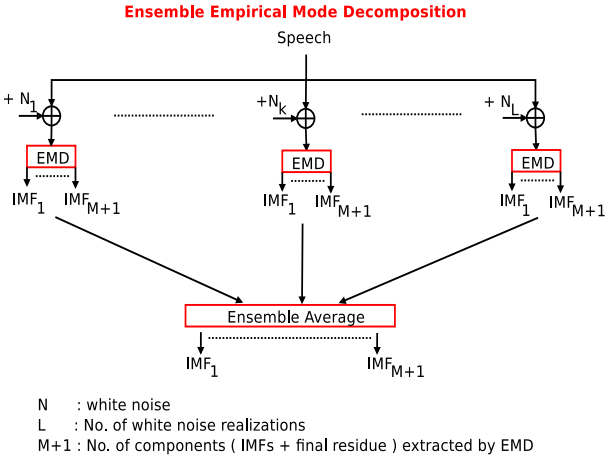


Figure 24: **Flowchart of Ensemble Empirical Mode Decomposition.**

To reduce the effects of mode-mixing in extracting IMFs from real physical signals, many modifications have been proposed to the EMD algorithm [100, 115, 119–124]. However, 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) [115]. The idea of infusing finite amplitude white noise into the signal serves an important purpose - the flat spectrum of white noise balances the skewed spectrum of speech to a certain extent. 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 - the addition of finite amplitude white noise increases the number of extrema present in the *inner residue* signal, which when used as interpolation points in a *sifting iteration*, leads to better estimates of the maxima and minima envelopes of the signal [115]. Fig.24 shows the flowchart of EEMD. The pseudocode for it is given below.

Pseudocode for EEMD : Let $s(t)$ be a continuous-time speech signal.

•(**i**) Create L noisy copies of the signal $s(t)$ using L independent realizations of finite amplitude normally distributed, $N(0,1)$, white noise.

$$s^l(t) = s(t) + \beta w^l(t),\ l = 1, ..., L\ , \tag{36}$$

where $s^l(t)$ is the l^{th} noisy copy of the signal $s(t)$, obtained by adding the l^{th} white noise sequence, $w^l(t)$, of zero mean and unit variance. The factor $\beta > 0$ controls the variance of noise with respect to the signal, which is generally taken as 10-40 % of the variance of the signal [115].

•(**ii**) Decompose each $s^l(n)$ using EMD, to obtain M IMFs from the signal. Generally, 10 sifting iterations are used in the sifting process, but may be adjusted for the task at hand and with respect to the level of added noise [115].

$$s^l(t) = \sum_{k=1}^{M} h_k^l(t) + r_M^l(t),\ l = 1, ..., L\ ,$$
$$s^l(t) = \sum_{k=1}^{M+1} h_k^l(t),\ l = 1, ..., L \tag{37}$$

•(**iii**) The final components are obtained as the ensemble average of the components obtained from each noisy copy of the signal.

$$h_k(t) = \frac{1}{L}\sum_{l=1}^{L} h_k^l(t),\ \forall\ k = 1, ..., M+1\ , \tag{38}$$

$$\hat{s}(t) = \sum_{k=1}^{M+1} h_k(t) = \frac{1}{L}\sum_{k=1}^{M+1}\sum_{l=1}^{L} h_k^l(t) \tag{39}$$

It is expected that as the number of white noise realiza-

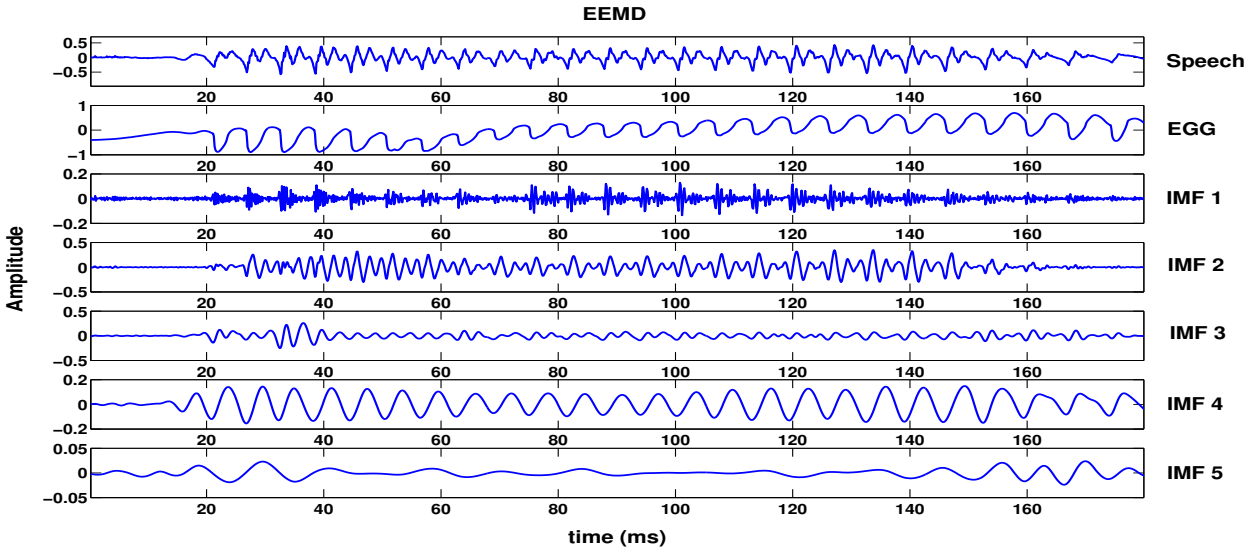


Figure 25: **(a) Speech ; (b) EGG ; (c)-(g) are IMFs 1-5 of the speech signal, obtained using EEMD.**

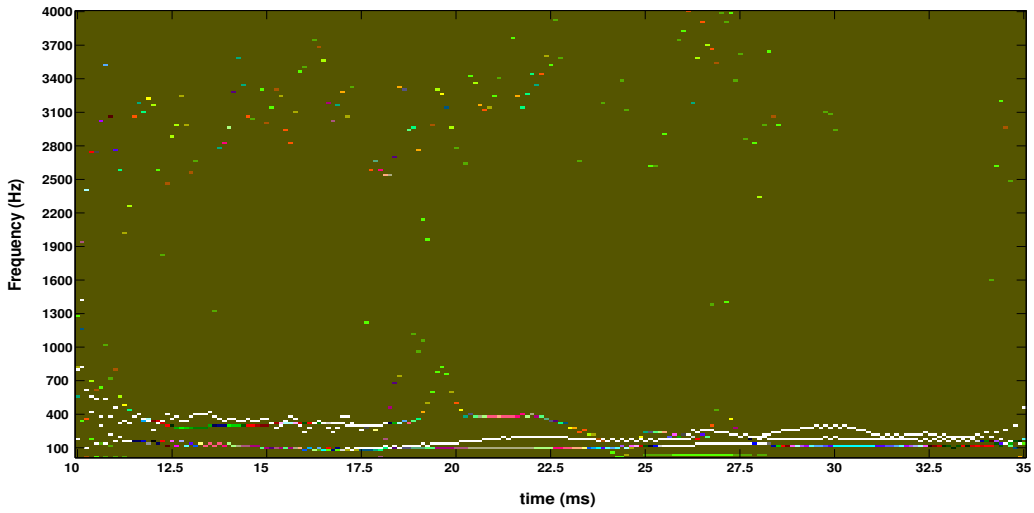


Figure 26: **Hilbert Spectrum of the speech signal, used in Fig.25, using EEMD.**

tions L is increased, the effect of noise would cancel out.

$$var[\hat{s}(t)] = var\Big[\frac{1}{L}\sum_{l=1}^{L}\sum_{k=1}^{M+1} h_k^l(t)\Big] \ ,$$

$$var[\hat{s}(t)] = var\Big[\frac{1}{L}\sum_{l=1}^{L} s^l(t)\Big] \ ,$$

$$var[\hat{s}(t)] = var\Big[\frac{1}{L}\sum_{l=1}^{L}\{s(t)+\beta w^l(t)\}\Big] \ ,$$

$$var[\hat{s}(t)] = var[s(t)] + \frac{1}{L}\beta^2 \ , \tag{40}$$

$$var[\hat{s}(t)] = var[s(t)], \ L \to \infty \ , \tag{41}$$

$$\hat{s}(t) = s(t) = \sum_{k=1}^{M+1} h_k(t), \ L \to \infty \ , \tag{42}$$

Fig.25 shows the IMFs obtained by EEMD of the same speech signal, which is decomposed by EMD in Fig.18. 10 white noise realizations have been used in the process, and the variance of noise has been kept at 20 %. $N = 10$ and $M = 9$ are used in the decomposition. It is evident from the figures that EEMD produces components with much lesser mode-mixing than EMD. Also, the IMFs of EEMD have a much better representation of the higher frequency spectrum of speech, as is reflected in the Hilbert spectrum of Fig.26. To quantify this observation, we calculate the *mean frequency* of the IMFs generated by EMD and EEMD [101]. The mean frequency of IMF$_k$ (F_k^m), gives an indication of the dominant frequency reflected in the IMF. Mathematically, it gives the *central tendency of the*

power spectrum of the IMF, and is given by,

$$F_k^m = \sum_{f=0}^{F_s/2} \frac{f \times S_k(f)}{\sum_{f=0}^{F_s/2} S_k(f)} \ , \ k = 1, ..., M+1 \ , \tag{43}$$

where $S_k(f)$ represents the power spectrum (squared magnitude spectrum) of IMF$_k$, and F_s is the sampling frequency of the speech signal. Fig.27(a) shows how the lower order IMFs of EEMD have a much higher mean frequency than that of EMD, thus giving a better representation of the higher frequency spectrum of the speech signal.

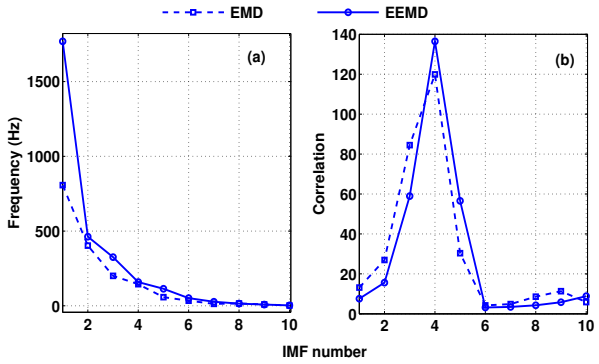


Figure 27: **(a) Mean Frequency of the IMFs - EMD vs EEMD ; (b) Maximum Correlation of the IMFs with the EGG signal - EMD vs EEMD.**

To evaluate how the lower frequency information of speech is represented by EMD and EEMD, the maximum correlation of the digital EGG signal, $e(n)$, with respect to the IMFs, obtained from both EMD and EEMD, is evaluated.

$$R_k^e = \max_m \Big\{ \sum_n h_k(n) e(n+m) \Big\} \ , \ k = 1, ..., 10 \ , \tag{44}$$

where R_k^e represents the maximum correlation of IMF$_k$ with the EGG signal. Fig.27(b) plots the values of R_k^e for both EMD and EEMD [101]. As is evident from the figure, EEMD reflects the glottal source information in a better way than EMD. Also, the source information is less distributed amongst the components of EEMD than that of EMD, a consequence of reduced mode-mixing. In general, the source information is found to be distributed almost entirely amongst two consecutive IMFs in the case of EEMD [101].

Finally, we may look at the distribution of the speech resonances in the IMFs. Fig.28 shows the magnitude spectra of the LP filters of the first 4 IMFs of a voiced speech signal of the TIMIT corpus. A 24-order LP analysis is used on the 16 kHz speech signal. The reference formant frequencies are obtained from the VTR Formants database [125]. For better visualization, the spectra are plotted only upto 4 kHz, within which the first four principal formants of the speech signal are generally confined. As can be seen from the figure, the first IMF of EMD carries all the formants, except the first formant, which is carried by the second IMF [117, 118]. In the case of EEMD, the formants structure is more evenly distributed amongst the first 4 IMFs. Thus, a better spectral segregation is achieved in the case of EEMD, compared to that of EMD. The IMFs of EEMD, hence, may be considered to be better suited for AM-FM analysis.

5.1. Improvements in EEMD

While there are many merits of EEMD, there are also certain limitations to the process. One of them is its efficiency. As is reflected in equations (41) and (42), a large number of white noise realizations are required to average out the effect of noise [115]. Again, there is no guarantee that each of the noisy signal copies would produce the same number of IMFs, which creates problems in averaging the IMFs. One way to circumvent this problem is to restrict the EMD decomposition to a fixed smaller number of IMFs (as shown in Fig.15) than that would be obtained if the decomposition is allowed to continue till a trend with only two extrema remains. In the recent years, efforts to effictively cancel out the noise infused with the signal has led to many EEMD variants [100, 123, 124].

It was observed that using white noise in pairs of opposite polarities substantially reduces the effect of the added noise in the IMFs finally derived from EEMD. This development was termed Complementary Ensemble Empirical Mode Decomposition (CEEMD). However, the problem that number of IMFs produced could still be different for the different EMD processes of an EEMD or CEEMD decomposition still existed. To circumvent this problem, an algorithm was designed, which not only decomposes the signal but also the white noise realizations. The IMFs obtained from the white noise realizations, which could be interpreted as correlated noise signals, are then fused with the residue signal, at the beginning of each sifting process. The signal IMFs are obtained progressively after averaging the results at each stage. This algorithm is termed Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN). However, it was observed that CEEMDAN sometimes produced some high-frequency and low-amplitude spurious IMFs, in the decomposition. To overcome this problem, the Improved Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (ICEEMDAN) was developed, which makes some subtle and effective modifications to the CEEMDAN algorithm. The pseudo code of ICEEMDAN is given below :

Algorithm for ICEEMDAN : Let $s(t)$ be a continuous-time speech signal. Let $E_k[.]$ be the operator which denotes the operation of extracting the k^{th} IMF from any signal $x(t)$, using EMD. Then, if $\Upsilon[x(t)]$ denotes the local mean of the signal, we have, $E_1[x(t)] = x(t) - \Upsilon[x(t)]$. Let $w^l(t)$ denote the l^{th} realization of zero mean unit variance white noise.

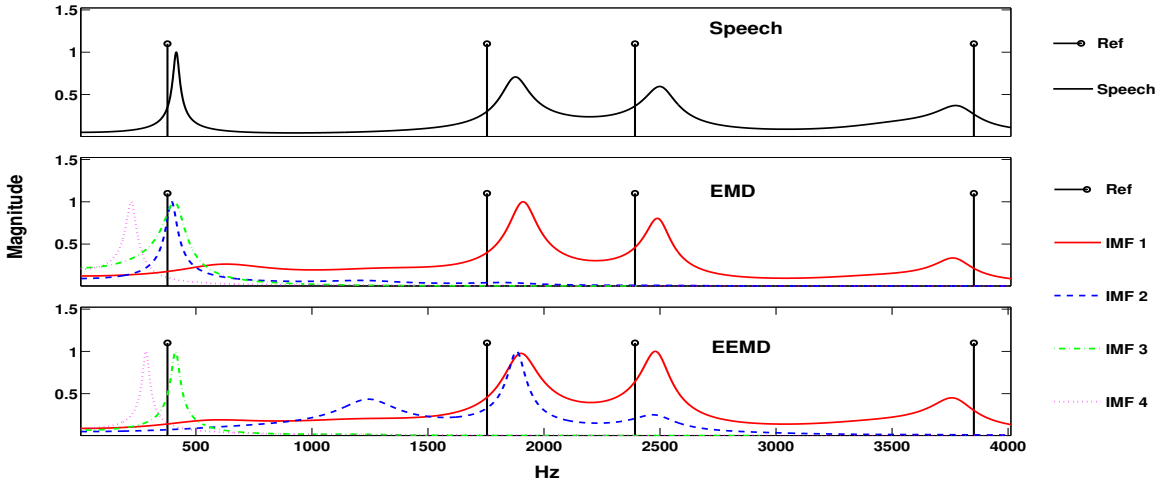


Figure 28: **Formants distribution in the IMFs - EMD vs EEMD. (a) Normalized Magnitude spectrum of LP filter of pre-emphasized voiced speech ; Normalized Magnitude spectra of LP filters of the first 4 IMFs, derived from (b) EMD of voiced speech ; (c) EEMD of voiced speech.**

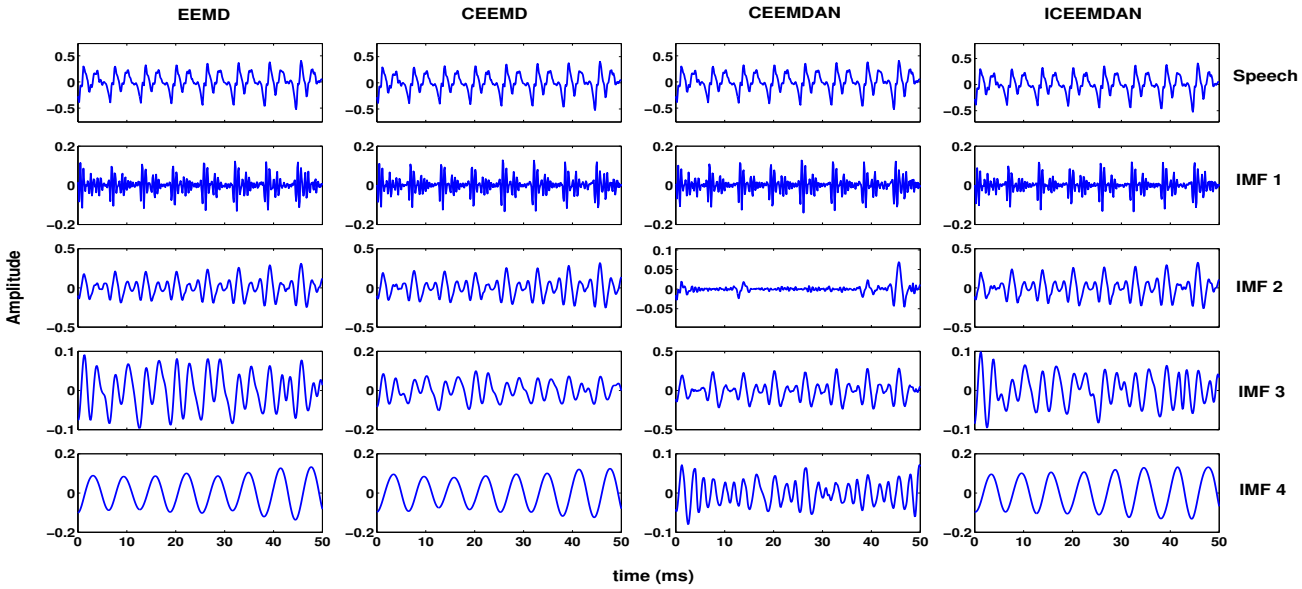


Figure 29: **The first 4 IMFs obtained from a speech signal using (from left to right column) EEMD, CEEMD, CEEMDAN, and ICEEMDAN. 20 white noise realizations (10 white noise pairs for CEEMD) are used in the processes. The number of sifting iterations are kept fixed to 10.**

For $k = 1, ..., M$, repeat the following steps.

•**(i)** Let $r_0(t) = s(t)$. The $(k-1)^{th}$ residue, $r_{k-1}(t)$ is mixed with noise as,

$$r^l_{k-1}(t) = r_{k-1}(t) + \beta_{k-1} E_k[w^l(t)],\ l = 1, ..., L\ ,$$

where β_{k-1} is used to control the SNR at each stage of the decomposition.

•**(ii)** The k^{th} IMF, $h_k(t)$, is derived as,

$$E_1[r^l_{k-1}(t)] = r^l_{k-1}(t) - \Upsilon[r^l_{k-1}(t)],\ l = 1, ..., L\ ,$$
$$r_k(t) = \frac{1}{L}\sum_{l=1}^{L} \Upsilon[r^l_{k-1}(t)],$$
$$h_k(t) = r_{k-1}(t) - r_k(t)$$

•**(iii)** Go to step (i). Stop when a maximum number of

IMFs, M, are extracted, i.e., when $k = M$.

$$s(t) = r_M(t) + \sum_{k=1}^{M} h_k(t) = \sum_{k=1}^{M+1} h_k(t),$$
$$where\ r_M(t) = h_{M+1}(t)$$

Generally, and in this work, $\beta_0 = \epsilon_0\ std(s(t))/std(E_1[w^l(t)])$. and $\beta_{k-1} = \epsilon_0\ std(r_{k-1}(t))$, $k > 2$. In this work, $\epsilon_0 = 0.2$. The number of iterations in a complete sifting process is determined by the local-global stopping criterion. The maximum number of iterations per sifting process is not allowed to exceed 15, i.e., $N \leq 15$ [100, 112, 126, 127].

Fig.29 shows the first four IMFs derived from a speech signal using EEMD, CEEMD, CEEMDAN, and ICEEMDAN. It may be observed that, in this example, there is no significant advantage of any one variant over the other [101]. In the case of CEEMDAN, a spurious mode is exhibited in IMF_2. However, the amplitude of the mode is quite less, and may or may not effect speech analysis depending on the procedures applied. Fig.30 shows the *reconstruction error* of the 4 algorithms, for the speech signal decomposed in Fig.29. Given a speech signal, $s(n)$, and its IMFs (including the final residue), $\{h_k(n), k = 1, ..., M+1\}$, the *reconstruction error*, r_e, is given by,

$$r_e = 10 \log_{10} ||s(n) - \sum_k h_k(n)||_2 \tag{45}$$

As Fig.30 shows, the reconstruction errors for the EEMD variants are lower than that of EEMD. However, it may be noted that even for EEMD, the reconstruction error is quite low. However, the processing time, for both EEMD and its variants remains quite large. Table 2 lists the time taken to extract 10 IMFs ($M = 9$) from a speech signal ($F_s = 8$ kHz) of around 3.5s duration, by EMD, EEMD, and the EEMD variants. A fixed number of sifting iterations, $N = 10$, is considered for all the methods, for a fair comparison, and the local-global stopping criterion is not used in this case. As is clear from the table, EEMD and its variants are time costly, and hence EEMD and its variants currently are limited in use in real-time applications, despite their obvious merits. Highly efficient coding, of course, could alleviate this drawback substantially. Regardless, EEMD and its variants are quite useful in applications which are not real-time. Also, in the case of applications where limited levels of noise could be tolerated, EEMD is just as useful as its variants.

6. Comparison with other speech processing approaches

As discussed in Secs.2 and 3, conventional AM-FM analysis and non-stationary signal analysis techniques like Wavelet Transform have been very effective in speech processing applications, and provide an alternative to conventional speech processing analysis. Henceforth, we need to weigh the effectiveness of EMD with respect to such techniques.

Table 2: **Computational time of EMD, EEMD, CEEMD, CEEMDAN, and ICEEMDAN, in decomposing a speech signal of around 3.5 seconds duration. 10 IMFs are extracted from the EMD variants, where 10 sifting iterations are used per sifting process. The algorithms are implemented in the GUI mode of MATLAB, on a machine having an Intel i5 quad-core processor of 3.2 GHz clock frequency, and 4 GB RAM.**

Method	EMD	EEMD	CEEMD	CEEMDAN	ICEEMDAN
Time (s)	0.83	15.62	14.94	32.88	30.11

As is discussed in Sec.2, the choice of the *mother wavelet*, $\psi(t)$, is critical to Wavelet analysis. As an illustration of this point, we may consider the case of a speech signal decomposed by 10-level DWT, using the 'Daubechies-4' and 'Biorthogonal-2.4' wavelets. The first five time-domain detail components, reconstructed by Inverse DWT of the first five detail coefficients, are shown in Fig.31. It is evident that changing the *mother wavelet* changes the decomposition. Apart from this, the WT does not tackle the problem of *non-linearity* of the speech signal [4, 8–10]. It is essentially an *adjustable window* STFT and hence is not applicable for analyzing non-linear systems [7, 29]. This is where EMD, and hence its proposed variants, scores over STFT and WT. EMD and its variants are able to extract the components of the signal, without requiring any *a priori basis*. Further, the *sifting process* is a non-linear process, and hence EMD is applicable for analyzing signals produced by non-linear systems [7, 29].

Unlike STFT and WT, AM-FM analysis or rather MDA, maybe used for dealing with both the non-stationary and the non-linear characteristics of a speech signal, as discussed in Sec.3. The basic aim of AM-FM analysis is to represent the speech signal in terms of AM-FM components, which are dominated by its resonant frequencies or formants, as reflected in equation (4), in Sec.4. But, as the speech formants are not known *a priori*, traditional AM-FM analysis uses a large bank of overlapping band-pass filters, to obtain AM-FM signals from the signal, which are used for further analysis [4, 23, 27, 28]. As such, the design of the filterbank remains an open issue. Fig.31 shows the first five high-frequency components obtained using a Gabor filterbank of only 20 uniformly spaced filters in the Hz scale, each having an effective bandwidth of 400 Hz. It is evident that AM-FM analysis would produce a significant number of redundant components, which may not be useful for analysis. Fig.31 also presents the first 5 IMFs of the speech signal, derived using EEMD, for comparison, which reflects its superiority over DWT and traditional AM-FM analysis. As the variants of EEMD perform similarly, they are not shown in the figure.

Fig.32 shows the magnitude spectra of the LP filters of the first 4 high-frequency components, obtained from DWT and AM-FM analysis, of the same voiced speech segment that is used in Fig.28. Irrespective of the type of mother wavelet used, DWT operates as a *fixed dyadic*

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
R. Sharma, L. D. Vignolo, G. Schlotthauer, M. A. Colominas, H. L. Rufiner & S. R. M. Prasanna, "Empirical Mode Decomposition for adaptive AM-FM analysis of Speech: A Review"
Speech Communication, Vol. 88, pp. 39-64, 2017.

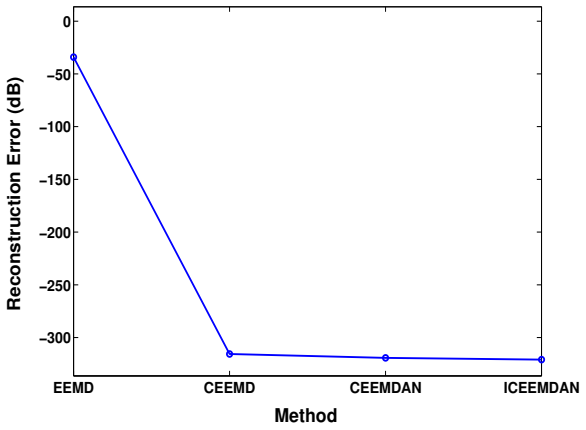


Figure 30: **The reconstruction error (in dB) of the speech signal shown in Fig.29, for the methods - EEMD, CEEMD, CEEMDAN and ICEEMDAN.**

filterbank [5, 6, 32]. The components derived clearly show the overlap of formants information, and also seem to lack precision in capturing the formants. In the case of AM-FM analysis, the components are obtained by Gabor filtering the speech segment, with a 40-filter Gabor filterbank, each filter having a bandwidth of 400 Hz. The LP magnitude spectra for only the four highest frequency components, corresponding to the Gabor filters having center frequencies below 4 kHz, are plotted. As reflected in the figure, AM-FM analysis is much more precise, but it would require a lot many components to capture the vocal tract resonances, and thus many of its components might be unuseful for speech analysis. Comparison of Figs.28 and 32 shows that out of all the techniques EEMD provides the best solution to obtaining the ideal goal of AM-FM analysis of speech - a limited number of components which encapsulates the vocal tract resonances precisely.

Finally, the time taken by DWT and AM-FM analysis in decomposing a speech signal is enlisted in Table 3, in comparison with EMD and MEMD. The same speech signal for which Table 2 is generated, is used in this case. 10 time-domain components are generated from the DWT decomposition of the speech signal using Biorthogonal 2.4 wavelet. For AM-FM analysis, a 40-filter overlapping Gabor filterbank is used. As can be observed from the table, DWT and AM-FM analysis are faster algorithms. EEMD is extremely time costly. EMD though not the fastest, it has an acceptable time-cost, which could be improved by efficient coding.

7. Some Applications of EMD in speech processing

In general, EMD has found usage in two broad areas of speech processing :

(i) Speech Analysis - Enhancement/Denoising [108, 128–134], Pitch Tracking [103, 135–137], Formant Tracking [118, 138] , Pathological Voice Analysis [139–141], etc.

(ii) Feature Extraction - Noise and audio Classification [142, 143] , Emotion Classification [144, 145], Speaker Recognition [146, 147], Voiced/Unvoiced speech classification [101, 148–150] etc.

Table 3: **Computational time of EMD, EEMD, DWT and AM-FM analysis, in decomposing a speech signal of around 3.5 seconds duration. 10 IMFs are extracted from EMD and EEMD, where 10 sifting iterations are used per sifting process. 10 components are also extracted from DWT using Biorthogonal 2.4 wavelet. 40 components are extracted from AM-FM analysis using a Gabor filterbank, each filter having a bandwidth of 400 Hz. The algorithms are implemented in the GUI mode of MATLAB, on a machine having an Intel i5 quad-core processor of 3.2 GHz clock frequency, and 4 GB RAM.**

Method	EMD	EEMD	DWT	AM-FM
Time (s)	0.83	15.62	0.42	0.40

We briefly revisit a few such works here.

EMD-based filtering (EMDF) of low-frequency noise for speech enhancement [133] :

Table 4: **Improvement in** *segmental SNR* **(sSNR) and** *Weighted Spectral Slope* **(WSS) measures of the EMDF system over the IMCRA/OMLSA system.**

Input SNR (dB)	Car Interior Noise		Babble Noise		Military Vehicle Noise	
	sSNR	WSS	sSNR	WSS	sSNR	WSS
10	3.6	-17.6	0.3	-7.2	2.3	-21.1
8	4.8	-23.1	0.5	-9.5	2.7	-27
6	5.8	-28.7	0.6	-11.7	3.1	-32.9
4	6.9	-34.6	0.7	-14	3.4	-38.4
2	7.8	-39.9	0.8	-16.7	3.8	-43.6
0	8.5	-45.1	0.9	-19.3	4	-48.3
-2	9.2	-49.5	1	-22.1	4.3	-52.5
-4	9.7	-53.4	1	-24.5	4.5	-56.4
-6	10.1	-56.7	1	-26.6	4.6	-59.6
-8	10.5	-60	1	-28.5	4.7	-62.6
-10	10.7	-62.8	1	-30.4	4.7	-65.5

This work uses an EMD based filtering (EMDF) mechanism for the purpose of eliminating *residual noise* from a speech signal which is already enhanced by some noise cancellation technique. The EMDF technique, applied as a post-processing step to an already enhanced speech signal, is found to significantly improve the performance of the speech enhancement system, particularly when the speech signal is corrupted by low-frequency noise. The speech enhancement system considered in this work, prior to applying EMDF, is the popular *Optimally Modified Log-Spectral Amplitude* (OMLSA), which uses a noise estimate determined by techniques like the *Improved Minima Controlled Recursive Averaging* (IMCRA). Fig.33 shows the block diagram of the overall system.

The EMDF mechanism is based on the observation that the variance of the IMFs of a clean speech signal, beyond

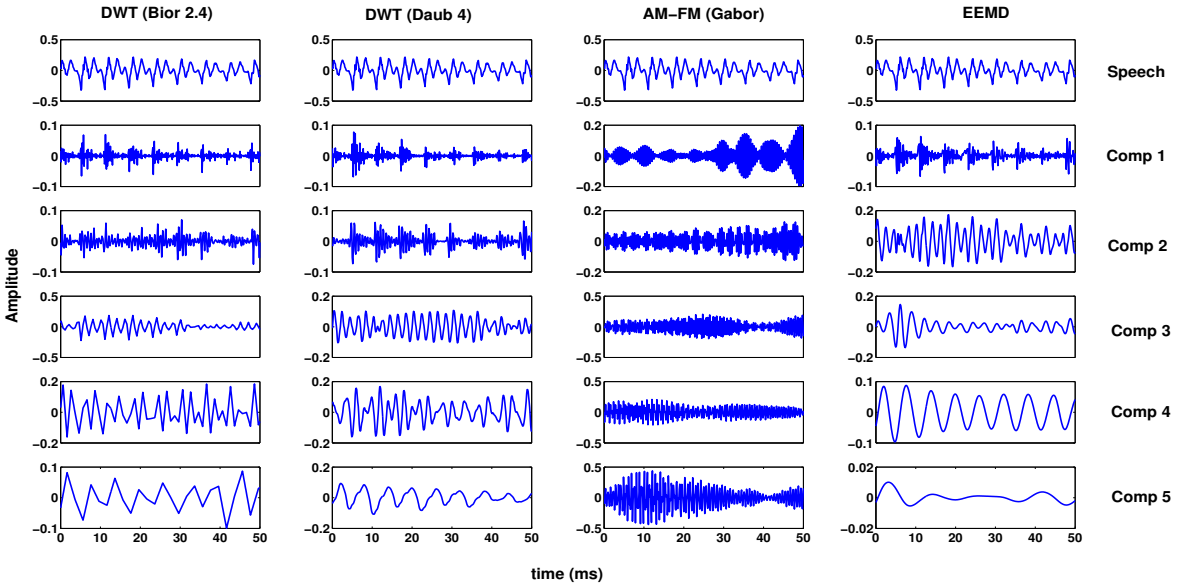


Figure 31: **The first five components (in decreasing order of frequency content) of a speech signal, as obtained from DWT (Biorthogonal 2.4 wavelet), DWT (Daubechies 4 wavelet), AM-FM analysis (20 filter linear Gabor filterbank), and EEMD.**

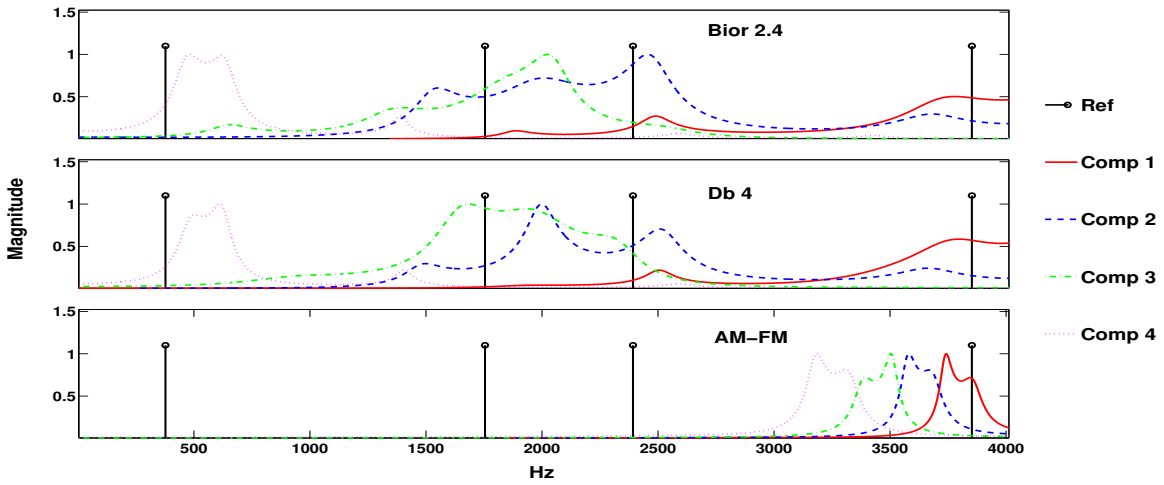


Figure 32: **Normalized Magnitude spectra of LP filters of the first four high-frequency components of the speech signal used in Fig.28. The components are obtained from (a) DWT (Biorthogonal 2.4 wavelet) ; (b) DWT (Daubechies 4 wavelet) ; (c) AM-FM analysis, using a 20 filter linear Gabor filterbank.**

the 4^{th} IMF, decreases monotonically, whereas a speech signal contaminated by low-frequency or residual noise exhibits sporadic fluctuations in the variance of its higher order IMFs, beyond IMF_4. Fig.34 shows the variance of the first 8 IMFs of a speech signal contaminated by car noise at SNR = 0 dB. As can be seen from the figure, the variances of the IMFs deviates from its decreasing trend, beyond IMF_4, and reaches a peak at IMF_7.

Let $s(n)$ be a speech signal corrupted by noise. As shown in Fig.33, the speech signal is enhanced by the OMLSA/IMCRA system to obtain the enhanced speech signal, $s_e(n)$, which is affected by residual noise. The signal, $s_e(n)$, is then processed by the EMDF system to obtain the final denoised signal, $s_D(n)$, utilizing the following steps :

(i) Decompose $s_e(n)$ by EMD. $s_e(n) = \sum_{k=1}^{M} h_k(n)$

(ii) Determine the variance, $\sigma_k^2,\ k = 1, ..., M$ of each IMF.

(iii) Detect the IMF orders of the peaks, $\{p_i \mid i = 1, 2, ...\}$ in the variance vs. IMF order curve, for IMF order $k > 4$.

(iv) Detect the IMF orders of the troughs, $\{t_i \mid i = 1, 2, ...\}$ in the variance vs. IMF order

curve, corresponding to the peaks.

(v) Compute the IMF variance buildup, $\{b_i = p_i - t_i \mid i = 1, 2, ...\}$.

(vii) Compute the index of the first occurance of the maximum buildup, $\nu = \max_i b_i,\ i = 1, 2, ...$

(vi) Compute $K = p_\nu - t_\nu$.

(vi) Partially reconstruct the enhanced speech signal to obtain $s_D(n)$.

$$s_D(n) = \sum_{k=1}^{K} h_k(n)$$

Table 4 shows the improvements in performance of the EMDF system over the IMCRA/OMLSA system, in terms of *segmental SNR* (sSNR) and *Weighted Spectral Slope* (WSS) measures. 192 speech utterances (F_s = 16 kHz), produced by 24 speakers (16 male and 8 female), obtained from the core test set of the TIMIT corpus [151] are considered for performance evaluation. The EMDF process is applied on speech frames of 512 samples, with the frames having 50 % overlap. Table 4 ascertains the utility of the EMDF process under different types of noise of varied strength. The positive values of sSNR indicate the increase in signal strength with respect to the input noise level, whereas the negative values of WSS indicate a reduction in loss of speech information.

Speech emotion recognition using novel HHT-TEO based features [144] :

In this work, the Hilbert spectrum is explored for the purpose of emotion classification using the Berlin Emotional Speech Database [152]. The Hilbert energy spectrum, $H(f,t)$, is generated from the IMFs of the speech signal, excluding the final residue. The spectrum so obtained is divided into 12 overlapping frequency bands. For each frequency band, for a given speech frame, the following two parameters are evaluated :

(i) Energy Cepstral Coefficients (ECC) : It is given by

$$ECC(f_i, t_j) = \int_{f \in f_i} H(f,t) df\ ,\ t \in t_j\ ,\ i = 1, ..., 12$$

where f_i represents a particular subband, and t_j represents a particular speech frame.

(ii) Frequency weighted Energy Cepstral Coefficients (EFCC) : It is given by

$$EFCC(f_i, t_j) = \int_{f \in f_i} f(t)\ H(f,t) df\ ,\ t \in t_j\ ,\ i = 1, ..., 12$$

where f_i represents a particular subband, and t_j represents a particular speech frame.

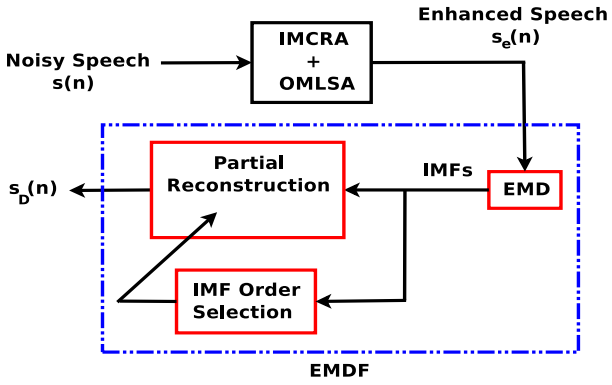


Figure 33: **EMDF based speech enhancement system.**

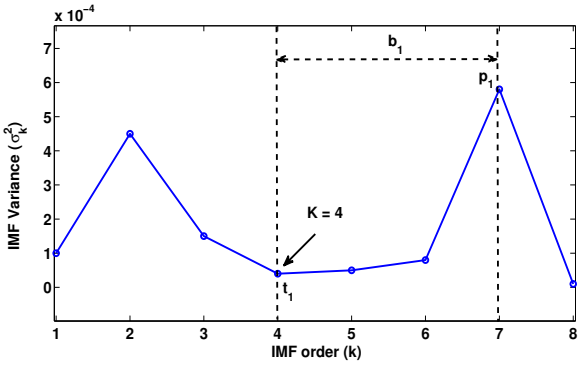


Figure 34: **Variances of the IMFs of a clean speech signal contaminated with car interior noise at 0 dB SNR.**

Thus, each speech frame produces 12 ECC and EFCC parameters. The natural logarithm of each of these 12 parameters is then taken, followed by Discrete Cosine Transform (DCT), to obtain the final 12-dimensional ECC and EFCC features per speech frame.

These parameters, in a standalone fashion, and in combination with the MFCCs, are then tested on the Berlin Emotional Speech Database. For the experiments, frames of 25ms with 50% overlap are used. The first 12 coefficients, generated from a 20-filter Mel filterbank, are used as the MFCCs. Continuous HMM using 5 states, with each state being represented by a 5 mixture GMM, is used to model the emotions. It is observed that while the ECC and EFCC features cannot outperform the MFCCs, they carry complementary information for the emotion recognition task, and significantly enhances the system performance when used in conjunction with the MFCCs. Some of the results are reported in Table 5, which confirms this observation.

8. Conclusions and Future Work

This paper emphasizes the analysis and processing of speech from a non-linear and non-stationary point of view, as opposed to the traditional short-time linear and stationary analysis of speech. Various evidences of the inherent non-linearity of speech and the limitations of conventional

Table 5: **Emotion Recognition performance (%) for MFCCs, MFCCs+ECC and MFCCs+EFCC features. A : Anger , B : Boredom , D : Disgust , F : Fear , H : Happiness , N : Neutral , S : Sadness. Values in the table are quoted from [144].**

Features	MFCCs							MFCCs + ECC							MFCCs + EFCC						
Emotion	A	B	D	F	H	N	S	A	B	D	F	H	N	S	A	B	D	F	H	N	S
A	**70**	0	0	0	30	0	0	**80**	0	0	0	15	0	0	**75**	0	0	0	5	0	0
B	0	**90**	10	10	0	30	5	0	**80**	10	0	0	15	5	0	**100**	15	0	0	5	5
D	5	0	**90**	0	10	5	5	5	0	**90**	10	5	5	5	0	0	**75**	5	5	0	10
F	5	0	0	**70**	10	0	5	5	0	0	**75**	5	5	0	15	0	0	**75**	5	5	0
H	15	0	0	0	**50**	0	0	10	0	0	0	**75**	0	0	10	0	0	0	**85**	0	0
N	5	0	0	15	0	**65**	0	0	20	0	10	0	**75**	0	0	0	10	15	0	**90**	0
S	0	10	0	5	0	0	**85**	0	0	0	5	0	0	**85**	0	0	0	5	0	0	**85**
Average	**74.29**							**80.0**							**83.57**						

speech processing mechanisms are discussed. This provides the motivation for AM-FM representation of speech, which models speech as being constituted of a finite number of AM-FM signals, centered around the vocal tract resonances. The process of EMD is then illustrated, which covers up some of the loopholes of conventional approaches of AM-FM analysis. The properties of EMD are discussed. The disadvantages of EMD and their mitigation by noise assisted EMD analysis are discussed. Finally, a few applications of EMD in speech processing are revisited.

Before concluding this article, we would like to briefly discuss some important points regarding the utility of EMD in speech processing. The obvious aim of EMD should be to eradicate short-time processing of speech to the maximum extent possible, and capture information hidden in the non-linear dynamics of the speech signal. For this purpose, efforts are required to customize EMD for speech processing applications. To be precise, if EMD could be customized to produce meaningful IMFs of speech, which are similar to that of EEMD, but at a much lesser time cost, it would become much more attractive to the speech community. Even though some efforts [122] have been made in this direction, the lack of a mathematical framework is an obstacle, and more work needs to be done in this regard. Nevertheless, even with the current state of affairs, enough work could be explored using EMD. One such exploration could be to detect the epochs or glottal closure instants (GCIs) of voiced speech [11, 33–40]. Despite the many disadvantages of the LP residual, most techniques for detecting GCIs are dependent on it, or some representation of the glottal source similar to it. However, to overcome the noisy characteristics of the LP residual, these methods strive to obtain a sinusoidal signal from it [33–35, 37, 39, 40]. The objective is to find some point in the near sinusoidal signal which can be correlated with the GCIs. In the case of EMD, we have already found that some of the sinusoid-like IMFs carry the periodicity of the glottal source [101]. Thus, an investigation needs to be carried out to see if they could be utilized to estimate GCIs. Apart from this, the scope for EMD in application to speech processing is wide open. As the method develops, becomes more adaptive and time-efficient, its usefulness as a speech processing tool will gather more recognition.

References

[1] L. R. Rabiner, R. W. Schafer, Digital processing of speech signals, Vol. 100, Prentice-hall Englewood Cliffs, 1978.

[2] L. R. Rabiner, R. W. Schafer, Introduction to digital speech processing, Foundations and trends in signal processing 1 (1) (2007) 1–194.

[3] J. Benesty, M. M. Sondhi, Y. Huang, Springer handbook of speech processing, Springer Science & Business Media, 2008.

[4] R. S. Holambe, M. S. Deshpande, Advances in Non-Linear Modeling for Speech Processing, Springer Science & Business Media, 2012.

[5] L. Cohen, Time-frequency analysis, Vol. 1406, Prentice Hall PTR Englewood Cliffs, NJ:, 1995.

[6] B. Boashash, Time frequency analysis, Gulf Professional Publishing, 2003.

[7] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, H. H. Liu, The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis, Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences 454 (1971) (1998) 903–995.

[8] J. F. Kaiser, Some observations on vocal tract operation from a fluid flow point of view, Vocal Fold Physiology: Biomechanics, Acoustics, and Phonatory Control (1983) 358–386.

[9] H. Teager, S. Teager, Evidence for nonlinear sound production mechanisms in the vocal tract, in: Speech production and speech modelling, Springer, 1990, pp. 241–261.

[10] S. McLaughlin, P. Maragos, Nonlinear methods for speech analysis and synthesis, Advances in nonlinear signal and image processing 6 (2006) 103.

[11] T. Ananthapadmanabha, B. Yegnanarayana, Epoch extraction from linear prediction residual for identification of closed glottis interval, Acoustics, Speech and Signal Processing, IEEE Transactions on 27 (4) (1979) 309–319.

[12] H. A. Murthy, Algorithms for processing fourier transform phase of signals, Ph.D. thesis, PhD dissertation, Department of Computer Science and Engineering, Indian Institute of Technology, Madras, India (1992).

[13] B. Yegnanarayana, K. Madhu Murthy, H. A. Murthy, Applications of group delay functions in speech processing, J. Inst. Elect. Telecommun. Eng 34 (1988) 20–29.

[14] K. M. Murthy, B. Yegnanarayana, Effectiveness of representation of signals through group delay functions, Signal Processing 17 (2) (1989) 141–150.

[15] K. K. Paliwal, L. D. Alsteris, Usefulness of phase spectrum in human speech perception., in: INTERSPEECH, 2003.

[16] K. K. Paliwal, L. D. Alsteris, On the usefulness of stft phase spectrum in human listening tests, Speech Communication 45 (2) (2005) 153–170.

[17] L. D. Alsteris, K. K. Paliwal, Further intelligibility results from human listening tests using the short-time phase spectrum, Speech Communication 48 (6) (2006) 727–736.

[18] S. Hayakawa, F. Itakura, Text-dependent speaker recognition using the information in the higher frequency band, in: Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on, Vol. 1, IEEE, 1994, pp. I–137.

[19] X. Lu, J. Dang, Physiological feature extraction for text independent speaker identification using non-uniform subband processing, in: Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on, Vol. 4, IEEE, 2007, pp. IV–461.

[20] X. Lu, J. Dang, An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification, Speech communication 50 (4) (2008) 312–322.

[21] L. D. Vignolo, H. L. Rufiner, D. H. Milone, J. C. Goddard, Evolutionary Splines for Cepstral Filterbank Optimization in Phoneme Classification, EURASIP Journal on Advances in Signal Proc. 2011 (2011) 8:1–8:14.

[22] L. D. Vignolo, H. L. Rufiner, D. H. Milone, J. C. Goddard, Evolutionary Cepstral Coefficients, Applied Soft Computing 11 (4) (2011) 3419–3428. doi:10.1016/j.asoc.2011.01.012.

[23] R. McAulay, T. F. Quatieri, Speech analysis/synthesis based on a sinusoidal representation, Acoustics, Speech and Signal Processing, IEEE Transactions on 34 (4) (1986) 744–754.

[24] J. F. Kaiser, On a simple algorithm to calculate theenergy'of a signal, in: Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on, 1990, pp. 381–384.

[25] A. C. Bovik, P. Maragos, T. F. Quatieri, Am-fm energy detection and separation in noise using multiband energy operators, Signal Processing, IEEE Transactions on 41 (12) (1993) 3245–3265.

[26] P. Maragos, J. F. Kaiser, T. F. Quatieri, On separating amplitude from frequency modulations using energy operators, in: Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on, Vol. 2, IEEE, 1992, pp. 1–4.

[27] P. Maragos, J. F. Kaiser, T. F. Quatieri, Energy separation in signal modulations with application to speech analysis, Signal Processing, IEEE Transactions on 41 (10) (1993) 3024–3051.

[28] A. Potamianos, P. Maragos, Speech formant frequency and bandwidth tracking using multiband energy demodulation, The Journal of the Acoustical Society of America 99 (6) (1996) 3795–3806.

[29] N. E. Huang, S. S. Shen, Hilbert-Huang transform and its applications, Vol. 5, World Scientific, 2005.

[30] W. J. Hardcastle, A. Marchal, Speech production and speech modelling, no. 55, Springer Science & Business Media, 1990.

[31] H. Teager, Some observations on oral air flow during phonation, Acoustics, Speech and Signal Processing, IEEE Transactions on 28 (5) (1980) 599–601.

[32] R. Polikar, The wavelet tutorial.

[33] R. Smits, B. Yegnanarayana, Determination of instants of significant excitation in speech using group delay function, Speech and Audio Processing, IEEE Transactions on 3 (5) (1995) 325–333.

[34] K. Sreenivasa Rao, S. Prasanna, B. Yegnanarayana, Determination of instants of significant excitation in speech using hilbert envelope and group delay function, Signal Processing Letters, IEEE 14 (10) (2007) 762–765.

[35] P. A. Naylor, A. Kounoudes, J. Gudnason, M. Brookes, Estimation of glottal closure instants in voiced speech using the dypsa algorithm, Audio, Speech, and Language Processing, IEEE Transactions on 15 (1) (2007) 34–43.

[36] K. S. R. Murty, B. Yegnanarayana, Epoch extraction from speech signals, Audio, Speech, and Language Processing, IEEE Transactions on 16 (8) (2008) 1602–1613.

[37] T. Drugman, T. Dutoit, Glottal closure and opening instant detection from speech signals., in: Interspeech, 2009, pp. 2891–2894.

[38] M. R. Thomas, J. Gudnason, P. A. Naylor, Estimation of glottal closing and opening instants in voiced speech using the yaga algorithm, Audio, Speech, and Language Processing, IEEE Transactions on 20 (1) (2012) 82–91.

[39] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, T. Dutoit, Detection of glottal closure instants from speech signals: a quantitative review, Audio, Speech, and Language Processing, IEEE Transactions on 20 (3) (2012) 994–1006.

[40] A. Prathosh, T. Ananthapadmanabha, A. Ramakrishnan, Epoch extraction based on integrated linear prediction residual using plosion index, Audio, Speech, and Language Processing, IEEE Transactions on 21 (12) (2013) 2471–2480.

[41] T. Matsui, S. Furui, Comparison of text-independent speaker recognition methods using vq-distortion and discrete/continuous hmm's, Speech and Audio Processing, IEEE Transactions on 2 (3) (1994) 456–459.

[42] F. Cummins, M. Grimaldi, T. Leonard, J. Simko, The chains corpus: Characterizing individual speakers, in: Proc of SPECOM, Vol. 6, 2006, pp. 431–435.

[43] J. Dang, K. Honda, Acoustic characteristics of the piriform fossa in models and humans, The Journal of the Acoustical Society of America 101 (1) (1997) 456–465.

[44] T. Kitamura, K. Honda, H. Takemoto, Individual variation of the hypopharyngeal cavities and its acoustic effects, Acoustical science and technology 26 (1) (2005) 16–26.

[45] K. Honda, T. Kitamura, H. Takemoto, S. Adachi, P. Mokhtari, S. Takano, Y. Nota, H. Hirata, I. Fujimoto, Y. Shimada, et al., Visualisation of hypopharyngeal cavities and vocal-tract acoustic modelling, Computer methods in biomechanics and biomedical engineering 13 (4) (2010) 443–453.

[46] C. Jankowski Jr, T. Quatieri, D. Reynolds, Measuring fine structure in speech: Application to speaker identification, in: Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on, Vol. 1, IEEE, 1995, pp. 325–328.

[47] B. Yegnanarayana, S. Prasanna, J. M. Zachariah, C. S. Gupta, Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system, Speech and Audio Processing, IEEE Transactions on 13 (4) (2005) 575–582.

[48] M. Grimaldi, F. Cummins, Speaker identification using instantaneous frequencies, Audio, Speech, and Language Processing, IEEE Transactions on 16 (6) (2008) 1097–1111.

[49] M. S. Deshpande, R. S. Holambe, Speaker identification based on robust am-fm features, in: Emerging Trends in Engineering and Technology (ICETET), 2009 2nd International Conference on, IEEE, 2009, pp. 880–884.

[50] J. R. Deller, J. G. Proakis, J. H. Hansen, Discrete-Time Processing of Speech Signals, Macmillan Publishing, NewYork, 1993.

[51] M. Sahidullah, G. Saha, A novel windowing technique for efficient computation of MFCC for speaker recognition, Signal Processing Letters, IEEE 20 (2) (2013) 149–152. doi:10.1109/LSP.2012.2235067.

[52] C. S. Ooi, K. P. Seng, L.-M. Ang, L. W. Chew, A new approach of audio emotion recognition, Expert Systems with Applications 41 (13) (2014) 5858–5869. doi:10.1016/j.eswa.2014.03.026.

[53] W. Zheng, M. Xin, X. Wang, B. Wang, A novel speech emotion recognition method via incomplete sparse least square regression, Signal Processing Letters, IEEE PP (99) (2014) 1–1.

doi:10.1109/LSP.2014.2308954.

[54] M. Reyes-Vargas, M. Sánchez-Gutiérrez, L. Rufiner, M. Albornoz, L. Vignolo, F. Martínez-Licona, J. Goddard-Close, Hierarchical clustering and classification of emotions in human speech using confusion matrices, in: Lecture Notes in Artificial Intelligence, Vol. 8113, Springer, 2013, pp. 162–169.

[55] D. Ververidis, C. Kotropoulos, Emotional speech recognition: Resources, features, and methods, Speech Communication 48 (9) (2006) 1162–1181. doi:10.1016/j.specom.2006.04.003.

[56] C.-L. Huang, S. Matsuda, C. Hori, Feature normalization using MVAW processing for spoken language recognition, in: Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific, 2013, pp. 1–4. doi:10.1109/APSIPA.2013.6694104.

[57] Z. Qin, W. Liu, T. Wan, A bag-of-tones model with MFCC features for musical genre classification, in: H. Motoda, Z. Wu, L. Cao, O. Zaiane, M. Yao, W. Wang (Eds.), Advanced Data Mining and Applications, Vol. 8346 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2013, pp. 564–575. doi:10.1007/978-3-642-53914-5_48.

[58] T. Ganchev, N. Fakotakis, G. Kokkinakis, Comparative evaluation of various MFCC implementations on the speaker verification task, in: Proceedings of the SPECOM, Vol. 1, 2005, pp. 191–194.

[59] F. Zheng, G. Zhang, Z. Song, Comparison of different implementations of MFCC, Journal of Computer Science and Technology 16 (6) (2001) 582–589. doi:10.1007/BF02943243.

[60] M. Skowronski, J. Harris, Exploiting independent filter bandwidth of human factor cepstral coefficients in automatic speech recognition, The Journal of the Acoustical Society of America 116 (3) (2004) 1774–1780.

[61] H. Yeganeh, S. Ahadi, S. Mirrezaie, A. Ziaei, Weighting of Mel Sub-bands Based on SNR/Entropy for Robust ASR, in: Signal Processing and Information Technology, 2008. ISSPIT 2008. IEEE International Symposium on, 2008, pp. 292–296.

[62] X. Zhou, Y. Fu, M. Liu, M. Hasegawa-Johnson, T. Huang, Robust Analysis and Weighting on MFCC Components for Speech Recognition and Speaker Identification, in: Multimedia and Expo, 2007 IEEE International Conference on, 2007, pp. 188–191.

[63] Y. Shao, Z. Jin, D. Wang, S. Srinivasan, An auditory-based feature for robust speech recognition, in: Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on, 2009, pp. 4625–4628. doi:10.1109/ICASSP.2009.4960661.

[64] Z. Wu, Z. Cao, Improved MFCC-Based Feature for Robust Speaker Identification, Tsinghua Science & Technology 10 (2) (2005) 158–161.

[65] Bŏril, H. and Fousek, P. and Pollák, P., Data-Driven Design of Front-End Filter Bank for Lombard Speech Recognition, in: Proc. of INTERSPEECH 2006 - ICSLP, Pittsburgh, Pennsylvania, 2006, pp. 381–384.

[66] B. Zamani, A. Akbari, B. Nasersharif, A. Jalalvand, Optimized discriminative transformations for speech features based on minimum classification error, Pattern Recognition Letters 32 (7) (2011) 948–955. doi:10.1016/j.patrec.2011.01.017.

[67] L. Burget, H. Heřmanský, Data Driven Design of Filter Bank for Speech Recognition, in: Text, Speech and Dialogue, Lecture Notes in Computer Science, Springer, 2001, pp. 299–304.

[68] T. Sainath, B. Kingsbury, A.-R. Mohamed, B. Ramabhadran, Learning filter banks within a deep neural network framework, in: Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on, 2013, pp. 297–302. doi:10.1109/ASRU.2013.6707746.

[69] T. Bäck, Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms, Oxford University Press, Oxford, UK, 1996.

[70] L. Vignolo, H. Rufiner, D. Milone, J. Goddard, Genetic optimization of cepstrum filterbank for phoneme classification, in: Proceedings of the Second International Conference on Bio-inspired Systems and Signal Processing (Biosignals 2009), INSTICC Press, Porto (Portugal), 2009, pp. 179–185.

[71] C. Charbuillet, B. Gas, M. Chetouani, J. Zarader, Optimizing feature complementarity by evolution strategy: Application to automatic speaker verification, Speech Communication 51 (9) (2009) 724–731.

[72] H. M. Torres, H. L. Rufiner, Clasificación de fonemas mediante paquetes de onditas orientadas perceptualmente, in: Anales del 1er Congreso Latinoamericano de Ingeniería Biomédica, Mazatlán 98, Vol. 1, México, 1998, pp. 163–166.
URL http://fich.unl.edu.ar/sinc/sinc-publications/1998/TR98

[73] H. M. Torres, H. L. Rufiner, Automatic speaker identification by means of mel cepstrum, wavelets and wavelets packets, in: Proceedings of the Chicago 2000 World Congress IEEE EMBS, 2000, paper No. TU–E201–02.
URL http://fich.unl.edu.ar/sinc/sinc-publications/2000/TR00

[74] A. Dabin, D. H. Milone, H. L. Rufiner, Onditas perceptualmente diseñadas para el reconocimiento automático del habla, in: Proc. 7th Argentine Symposium on Artificial Intelligence, Rosario, Argentina, 2005, pp. 249–260.
URL http://fich.unl.edu.ar/sinc/sinc-publications/2005/DMR05

[75] L. D. Vignolo, D. H. Milone, H. L. Rufiner, Genetic wavelet packets for speech recognition, Expert Systems with Applications 40 (6) (2013) 2350–2359. doi:10.1016/j.eswa.2012.10.050.

[76] L. Vignolo, H. Rufiner, D. Milone, Multi-objective optimisation of wavelet features for phoneme recognition, IET Signal Processing.
URL http://digital-library.theiet.org/content/journals/10.1049/iet-spr.2015.0568

[77] L. D. Vignolo, S. M. Prasanna, S. Dandapat, H. L. Rufiner, D. H. Milone, Feature optimisation for stress recognition in speech, Pattern Recognition Letters 84 (2016) 1–7.

[78] S. Shukla, S. Prasanna, S. Dandapat, Stressed speech processing: Human vs automatic in non-professional speakers scenario, in: Communications (NCC), 2011 National Conference on, 2011, pp. 1–5.

[79] A. Batliner, S. Steidl, C. Hacker, E. Nöth, Private emotions versus social interaction: a data-driven approach towards analysing emotion in speech, User Modeling and User-Adapted Interaction 18 (1-2) (2008) 175–206. doi:10.1007/s11257-007-9039-4.

[80] S. Steidl, Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech, Logos Verlag, 2009.

[81] D. Wang, D. Miao, C. Xie, Best basis-based wavelet packet entropy feature extraction and hierarchical eeg classification for epileptic detection, Expert Systems with Applications 38 (11) (2011) 14314 – 14320. doi:10.1016/j.eswa.2011.05.096.

[82] A. R. Ferreira da Silva, Approximations with evolutionary pursuit, Signal Processing 83 (3) (2003) 465–481.

[83] R. Coifman, M. V. Wickerhauser, Entropy-based algorithms for best basis selection, IEEE Transactions on Information Theory 38 (2) (1992) 713–718.

[84] N. Saito, R. Coifman, Local discriminant bases and their applications, Journal of Mathematical Imaging and Vision 5 (4) (1995) 337–358. doi:10.1007/BF01250288.

[85] R. Munkong, B.-H. Juang, Auditory perception and cognition, Signal Processing Magazine, IEEE 25 (3) (2008) 98–117. doi:10.1109/MSP.2008.918418.

[86] H. Rufiner, J. Goddard, A method of wavelet selection in phoneme recognition, in: Proceedings of the 40th Midwest Symposium on Circuits and Systems, Vol. 2, 1997, pp. 889–891.

[87] N. E. Saeedi, F. Almasganj, Wavelet adaptation for automatic voice disorders sorting, Computers in Biology and Medicine 43 (6) (2013) 699 – 704. doi:10.1016/j.compbiomed.2013.03.006.

[88] R. Behroozmand, F. Almasganj, Optimal selection of wavelet-

packet-based features using genetic algorithm in pathological assessment of patients' speech signal with unilateral vocal fold paralysis, Computers in Biology and Medicine 37 (4) (2007) 474 – 485. doi:10.1016/j.compbiomed.2006.08.016.

[89] F. Jabloun, A. E. Cetin, E. Erzin, Teager energy based feature parameters for speech recognition in car noise, Signal Processing Letters, IEEE 6 (10) (1999) 259–261.

[90] D. Dimitriadis, P. Maragos, A. Potamianos, Modulation features for speech recognition, in: Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on, Vol. 1, IEEE, 2002, pp. I–377.

[91] D. Dimitriadis, P. Maragos, A. Potamianos, Robust am-fm features for speech recognition, Signal Processing Letters, IEEE 12 (9) (2005) 621–624.

[92] P. Cosi, Evidence against frame-based analysis techniques, Proceedings of NATO Advance Institute on Computational Hearing (1998) 163–168.

[93] Z.-H. Tan, I. Kraljevski, Joint variable frame rate and length analysis for speech recognition under adverse conditions, Computers & Electrical Engineering 40 (7) (2014) 2139–2149.

[94] C.-S. Jung, K. J. Han, H. Seo, S. S. Narayanan, H.-G. Kang, A variable frame length and rate algorithm based on the spectral kurtosis measure for speaker verification, in: Eleventh Annual Conference of the International Speech Communication Association, 2010.

[95] Q. Zhu, A. Alwan, On the use of variable frame rate analysis in speech recognition, in: Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on, Vol. 3, IEEE, 2000, pp. 1783–1786.

[96] Y. Pantazis, O. Rosec, Y. Stylianou, Adaptive am–fm signal decomposition with application to speech analysis, IEEE Transactions on Audio, Speech, and Language Processing 19 (2) (2011) 290–300.

[97] T. F. Quatieri, T. E. Hanna, G. C. O'Leary, Am-fm separation using auditory-motivated filters, IEEE transactions on speech and audio processing 5 (5) (1997) 465–480.

[98] N. E. Huang, Empirical mode decomposition and hilbert spectral analysis.

[99] R. Sharma, K. Ramesh, S. Prasanna, Analysis of electroglottograph signal using ensemble empirical mode decomposition, in: India Conference (INDICON), 2014 Annual IEEE, IEEE, 2014, pp. 1–6.

[100] M. A. Colominas, G. Schlotthauer, M. E. Torres, Improved complete ensemble emd: A suitable tool for biomedical signal processing, Biomedical Signal Processing and Control 14 (2014) 19–29.

[101] R. Sharma, S. R. M. Prasanna, Characterizing glottal activity from speech using empirical mode decomposition, in: National Conference on Communications 2015 (NCC-2015), Mumbai, India, 2015.

[102] P. Flandrin, Some aspects of huang's empirical mode decomposition, from interpretation to applications, in: Int. Conf. Computat. Harmonic Anal. CHA, Vol. 4, 2004.

[103] H. Huang, J. Pan, Speech pitch determination based on hilbert-huang transform, Signal Processing 86 (4) (2006) 792–803.

[104] M. A. Colominas, G. Schlotthauer, M. E. Torres, An unconstrained optimization approach to empirical mode decomposition, Digital Signal Processing 40 (2015) 164–175.

[105] Z. Wu, N. E. Huang, A study of the characteristics of white noise using the empirical mode decomposition method, Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences 460 (2046) (2004) 1597–1611.

[106] P. Flandrin, G. Rilling, P. Goncalves, Empirical mode decomposition as a filter bank, Signal Processing Letters, IEEE 11 (2) (2004) 112–114.

[107] P. Flandrin, P. Goncalves, Empirical mode decompositions as data-driven wavelet-like expansions, International Journal of Wavelets, Multiresolution and Information Processing 2 (04) (2004) 477–496.

[108] G. Rilling, P. Flandrin, P. Gonçalves, Empirical mode decomposition, fractional gaussian noise and hurst exponent estimation., in: ICASSP (4), 2005, pp. 489–492.

[109] P. Flandrin, P. Gonçalves, G. Rilling, Emd equivalent filter banks, from interpretation to applications, Hilbert-Huang transform and its applications (2005) 57–74.

[110] G. Rilling, P. Flandrin, on the influence of sampling on the empirical mode decomposition., in: ICASSP (3), 2006, pp. 444–447.

[111] G. Rilling, Décompositions modales empiriques, Ph.D. thesis, PhD thesis, Ecole normale supérieure de Lyon (2007).

[112] G. Rilling, P. Flandrin, P. Goncalves, et al., On empirical mode decomposition and its algorithms, in: IEEE-EURASIP workshop on nonlinear signal and image processing, Vol. 3, NSIP-03, Grado (I), 2003, pp. 8–11.

[113] N. E. Huang, M.-L. C. Wu, S. R. Long, S. S. Shen, W. Qu, P. Gloersen, K. L. Fan, A confidence limit for the empirical mode decomposition and hilbert spectral analysis, Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences 459 (2037) (2003) 2317–2345.

[114] G. Wang, X.-Y. CHEN, F.-L. Qiao, Z. Wu, N. E. Huang, On intrinsic mode function, Advances in Adaptive Data Analysis 2 (03) (2010) 277–293.

[115] Z. Wu, N. E. Huang, Ensemble empirical mode decomposition: a noise-assisted data analysis method, Advances in adaptive data analysis 1 (01) (2009) 1–41.

[116] G. Rilling, P. Flandrin, One or two frequencies? the empirical mode decomposition answers, Signal Processing, IEEE Transactions on 56 (1) (2008) 85–95.

[117] A. Bouzid, N. Ellouze, Empirical mode decomposition of voiced speech signal, in: Control, Communications and Signal Processing, 2004. First International Symposium on, IEEE, 2004, pp. 603–606.

[118] A. Bouzid, N. Ellouze, Voiced speech analysis by empirical mode decomposition, in: Advances in Nonlinear Speech Processing, Springer, 2007, pp. 213–220.

[119] Y. Chen, M. Q. Feng, A technique to improve the empirical mode decomposition in the hilbert-huang transform, Earthquake Engineering and Engineering Vibration 2 (1) (2003) 75–85.

[120] R. Deering, J. F. Kaiser, The use of a masking signal to improve empirical mode decomposition, in: Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on, Vol. 4, IEEE, 2005, pp. iv–485.

[121] Y. Kopsinis, S. McLaughlin, Improved emd using doubly-iterative sifting and high order spline interpolation, EURASIP Journal on Advances in Signal processing 2008 (2008) 120.

[122] R. Sharma, S. M. Prasanna, A better decomposition of speech obtained using modified empirical mode decomposition, Digital Signal Processing 58 (2016) 26 – 39. doi:http://dx.doi.org/10.1016/j.dsp.2016.07.012.
URL http://www.sciencedirect.com/science/article/pii/S1051200416300975

[123] J.-R. Yeh, J.-S. Shieh, N. E. Huang, Complementary ensemble empirical mode decomposition: A novel noise enhanced data analysis method, Advances in Adaptive Data Analysis 2 (02) (2010) 135–156.

[124] M. E. Torres, M. A. Colominas, G. Schlotthauer, P. Flandrin, A complete ensemble empirical mode decomposition with adaptive noise, in: Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, IEEE, 2011, pp. 4144–4147.

[125] L. Deng, X. Cui, R. Pruvenok, Y. Chen, S. Momen, A. Alwan, A database of vocal tract resonance trajectories for research in speech processing, in: Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on, Vol. 1, IEEE, 2006, pp. I–I.

[126] http://perso.ens-lyon.fr/patrick.flandrin/emd.html.
URL http://perso.ens-lyon.fr/patrick.flandrin/emd.html

[127] http://www.bioingenieria.edu.ar/grupos/ldnlys/index.htm.

URL http://www.bioingenieria.edu.ar/grupos/ldnlys/index.htm

[128] P. Flandrin, P. Gonçalves, G. Rilling, et al., Detrending and denoising with empirical mode decompositions, Citeseer, 2004.

[129] A.-O. Boudraa, J.-C. Cexus, et al., Denoising via empirical mode decomposition, Proc. IEEE ISCCSP 4.

[130] Y. Kopsinis, S. McLaughlin, Development of emd-based denoising methods inspired by wavelet thresholding, Signal Processing, IEEE Transactions on 57 (4) (2009) 1351–1362.

[131] T. Hasan, M. K. Hasan, Suppression of residual noise from speech signals using empirical mode decomposition, Signal Processing Letters, IEEE 16 (1) (2009) 2–5.

[132] G. Tsolis, T. D. Xenos, Signal denoising using empirical mode decomposition and higher order statistics, Int J Signal Proc Image Proc Pattern Recog 4 (2011) 91–106.

[133] N. Chatlani, J. J. Soraghan, Emd-based filtering (emdf) of low-frequency noise for speech enhancement, Audio, Speech, and Language Processing, IEEE Transactions on 20 (4) (2012) 1158–1166.

[134] I. Hadhami, A. Bouzid, Speech denoising based on empirical mode decomposition and improved thresholding, in: Advances in Nonlinear Speech Processing, Springer, 2013, pp. 200–207.

[135] Z. Yang, D. Huang, L. Yang, A novel pitch period detection algorithm based on hilbert-huang transform, in: Advances in Biometric Person Authentication, Springer, 2005, pp. 586–593.

[136] G. Schlotthauer, M. E. Torres, H. L. Rufiner, A new algorithm for instantaneous f0 speech extraction based on ensemble empirical mode decomposition, in: Proc. of 17th Eur. Sign. Proces. Conf, 2009, pp. 2347–2351.

[137] G. Schlotthauer, M. Torres, H. Rufiner, Voice fundamental frequency extraction algorithm based on ensemble empirical mode decomposition and entropies, in: World Congress on Medical Physics and Biomedical Engineering, September 7-12, 2009, Munich, Germany, Springer, 2010, pp. 984–987.

[138] H. Huang, X.-x. Chen, Speech formant frequency estimation based on hilbert-huang transform, JOURNAL-ZHEJIANG UNIVERSITY ENGINEERING SCIENCE 40 (11) (2006) 1926.

[139] G. Schlotthauer, M. E. Torres, H. L. Rufiner, Pathological voice analysis and classification based on empirical mode decomposition, in: Development of multimodal interfaces: active listening and synchrony, Springer, 2010, pp. 364–381.

[140] M. Kaleem, B. Ghoraani, A. Guergachi, S. Krishnan, Pathological speech signal analysis and classification using empirical mode decomposition, Medical & biological engineering & computing 51 (7) (2013) 811–821.

[141] B. Mijovic, M. Silva, B. Van den Bergh, K. Allegaert, J.-M. Aerts, D. Berckmans, S. Van Huffel, et al., Assessment of pain expression in infant cry signals using empirical mode decomposition, Methods Inf Med 49 (5) (2010) 448–452.

[142] D. Jhanwar, K. K. Sharma, S. Modani, Classification of environmental background noise sources using hilbert-huang transform, International Journal of Signal Processing Systems Vol 1.

[143] B. K. Khonglah, R. Sharma, S. Mahadeva Prasanna, Speech vs music discrimination using empirical mode decomposition, in: Communications (NCC), 2015 Twenty First National Conference on, IEEE, 2015, pp. 1–6.

[144] X. Li, X. Li, Speech emotion recognition using novel hht-teo based features, Journal of Computers 6 (5) (2011) 989–998.

[145] L. He, M. Lech, N. C. Maddage, N. B. Allen, Study of empirical mode decomposition and spectral analysis for stress and emotion classification in natural speech, Biomedical Signal Processing and Control 6 (2) (2011) 139–146.

[146] T. Hasan, J. H. Hansen, Robust speaker recognition in non-stationary room environments based on empirical mode decomposition., in: INTERSPEECH, 2011, pp. 2733–2736.

[147] J.-D. Wu, Y.-J. Tsai, Speaker identification system using empirical mode decomposition and an artificial neural network, Expert Systems with Applications 38 (5) (2011) 6112–6117.

[148] M. K. I. Molla, K. Hirose, N. Minematsu, Robust voiced/unvoiced speech classification using empirical mode decomposition and periodic correlation model., in: INTERSPEECH, 2008, pp. 2530–2533.

[149] Z. Lu, B. Liu, L. Shen, Speech endpoint detection in strong noisy environment based on the hilbert-huang transform, in: Mechatronics and Automation, 2009. ICMA 2009. International Conference on, IEEE, 2009, pp. 4322–4326.

[150] M. K. Islam Molla, S. Das, M. E. Hamid, K. Hirose, Empirical mode decomposition for advanced speech signal processing, Journal of Signal Processing 17 (6) (2013) 215–229.

[151] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, DARPA TIMIT acoustic phonetic continuous speech corpus CDROM (1993).
URL http://www.ldc.upenn.edu/Catalog/LDC93S1.html

[152] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss, A database of german emotional speech., in: Interspeech, Vol. 5, 2005, pp. 1517–1520.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
R. Sharma, L. D. Vignolo, G. Schlotthauer, M. A. Colominas, H. L. Rufiner & S. R. M. Prasanna; "Empirical Mode Decomposition for adaptive AM-FM analysis of Speech: A Review"
Speech Communication, Vol. 88, pp. 39-64, 2017.