Med Biol Eng Comput (2013) 51:811–821
DOI 10.1007/s11517-013-1051-8

ORIGINAL ARTICLE

Pathological speech signal analysis and classification using empirical mode decomposition

Muhammad Kaleem · Behnaz Ghoraani · Aziz Guergachi · Sridhar Krishnan

Received: 30 October 2012 / Accepted: 15 February 2013 / Published online: 5 March 2013

Abstract Automated classification of normal and pathological speech signals can provide an objective and accurate mechanism for pathological speech diagnosis, and is an active area of research. A large part of this research is based on analysis of acoustic measures extracted from sustained vowels. However, sustained vowels do not reflect real-world attributes of voice as effectively as continuous speech, which can take into account important attributes of speech such as rapid voice onset and termination, changes in voice frequency and amplitude, and sudden discontinuities in speech. This paper presents a methodology based on empirical mode decomposition (EMD) for classification of continuous normal and pathological speech signals obtained from a well-known database. EMD is used to decompose randomly chosen portions of speech signals into intrinsic mode functions, which are then analyzed to extract meaningful temporal and spectral features, including true instantaneous features which can capture discriminative information in signals hidden at local time-scales. A total of six features are extracted, and a linear classifier is used with the feature vector to classify continuous speech portions obtained from a database consisting of 51 normal and 161 pathological speakers. A classification accuracy of 95.7 % is obtained, thus demonstrating the effectiveness of the methodology.

Keywords Empirical mode decomposition · Speech signal analysis · Feature extraction · Pathological speech classification

1 Introduction

Pathological speech refers to speech problems that result from damage or malfunction of the human speech organs. These problems are exacerbated in people who use their voice professionally, for example professors, lawyers, actors and singers. The traditional way of diagnosing pathological speech is based on listening to a patient's voice. However, such approaches are subjective to the training and expertise of the specialist performing the diagnosis [1]. Significant attention has thus been paid to objective assessment of speech pathology, making automated speech pathology detection an active field of research [2–5].

The main goal of automated pathological speech detection systems is to enable characterization of any input voice as either normal or pathological. These systems use signal processing tools, such as temporal, spectral and cepstral methods (e.g., [6–8]), or the more recently introduced joint time–frequency methods (e.g., [9, 10]), as a means for accurate and discriminatory feature extraction. A classifier is then applied to the features thus extracted in order to discriminate between normal and pathological speech.

An important aspect to consider in pathological speech classification is whether the features have been extracted

M. Kaleem (✉) · S. Krishnan
Department of Electrical and Computer Engineering,
Ryerson University, Toronto, Canada
e-mail: farhat@gmail.com

S. Krishnan
e-mail: krishnan@ee.ryerson.ca

B. Ghoraani
Chemical and Biomedical Engineering Department,
Rochester Institute of Technology, Rochester, USA

A. Guergachi
Ted Rogers School of Information Technology Management,
Ryerson University, Toronto, Canada

from sustained vowels or continuous speech. Although sustained vowels offer a more controlled way of quantifying voice characteristics, and in general produce good classification results, they are not representative of speaking as they are of singing [11]. On the other hand, continuous speech signals capture important attributes of speech, such as rapid voice onset and termination, changes in voice frequency and amplitude, as well as sudden voice breaks. These attributes are also important in perception of voice quality of a speaker in everyday life [11], as has been assessed in recent works [9, 10].

While extraction of features from sustained vowels and their effectiveness in classification of speech signals has been studied extensively in literature, similar studies using continuous speech signals are fewer in number. Moreover, the use of sustained vowels has been motivated by the resulting simple acoustic structure, hence resulting in simpler and consistent voice quality assessment [11]. Studies analyzing continuous speech signals rely in general on segmentation of speech to identify the voiced, unvoiced and silent periods in speech. This is required for features used to quantify periodicity and regularity in speech, such as the harmonic-to-noise ratio, cepstral peak prominence and pitch amplitude, which hold only for voiced regions of speech. Speech signals, however, have non-stationary dynamics, and stationarity of signals cannot be assumed over short time intervals. It is possible for important signal characteristics to be lost if segmentation occurs at intervals containing non-stationary dynamics [10]. One of the reasons for the scarcity of works on feature extraction from continuous speech is the challenge of segmenting the speech signal into voiced, unvoiced/silent periods prior to feature extraction [9]. The study in [11] has pointed out the challenge involved in the analysis of continuous speech due to the inherent non-stationarity of the signal. This makes non-stationary signal analysis techniques, which do not require segmentation of the signals, and which can better reveal non-stationary behavior of signals such as trends, discontinuities and repeated patterns, more attractive for the analysis and automatic classification of speech signals.

A newer method for non-stationary signal analysis is empirical mode decomposition (EMD) [12], which has found extensive application in many areas of science and engineering, including biomedical engineering, e.g., [13, 14]. EMD decomposes signals into functions of time, from which spectral information may also be obtained. Therefore, EMD lends itself well to extraction of temporal as well as spectral descriptors from the original signal. However, application of EMD to analysis of speech signals is scarce. A noise-enhanced algorithm of EMD has been used to extract the fundamental frequency from sustained vowels as described in [15], whereas EMD-based classification of normal and pathological voices is presented in [16]. However, the work in [16] is also based on sustained vowels, and not continuous speech, thereby not realizing the full potential of a non-stationary signal analysis technique like EMD.

With this background, the goal of the work presented in this paper is to apply EMD in its simplest form to randomly selected continuous speech portions of both normal and pathological signals so as to extract meaningful and well-defined temporal and spectral features. With an emphasis on simplicity and effectiveness of the proposed methodology, we demonstrate the efficacy of the features presented in this paper by applying a simple classification scheme to classify normal and pathological signals with high accuracy. To the best of our knowledge, pathological voice classification using randomly selected continuous speech portions has not been done before. In this context, this work can also be seen as a preliminary step towards text-independent pathological voice classification.

2 Methods

2.1 Empirical mode decomposition

Empirical mode decomposition is an adaptive technique that allows decomposition of non-linear and non-stationary data into *intrinsic mode functions*. An intrinsic mode function (IMF) satisfies the following two conditions [12]:

1. The number of maxima, which are strictly positive, and the number of minima, which are strictly negative, for each IMF, are either equal, or differ at most by one.
2. The mean value of the envelope, as defined by the maxima and the minima, for each IMF, is zero.

The technique for decomposition of the data into IMFs is known as sifting, a brief description of which follows:

1. For a given discrete time signal $x[n]$, all the local minima and maxima of $x[n]$ are identified.
2. The upper envelope E_U is calculated by using a cubic spline to connect all the local maxima. Similarly, the lower envelope E_L is calculated from the local minima. The upper and lower envelopes should cover all the data in $x[n]$ between them.
3. The mean $E_{mean} = (E_U + E_L)/2$ of the upper and lower envelopes is calculated, and $x[n]$ is updated by subtracting the mean from it $x[n] \leftarrow x[n] - E_{mean}$.
4. The previous three steps are executed till $x[n]$ is reduced to an IMF $c_1[n]$, which conforms to the properties of IMFs described previously. The first IMF contains the highest oscillation frequencies found in the original data $x[n]$.

5. The first IMF $c_1[n]$ is subtracted from $x[n]$ to get the residue $r_1[n]$.
6. The residue $r_1[n]$ is now taken as the starting point instead of $x[n]$, and the previously mentioned steps are repeated to find all the IMFs $c_i[n]$ so that the final residue r_K either becomes a constant, a monotonic function, or a function with a single maximum and minimum from which no further IMF can be extracted.

Therefore, at the end of the decomposition, we can represent $x[n]$ as the sum of K IMFs and a residue r_K

$$x[n] = \sum_{i=1}^{K} c_i[n] + r_K[n] \tag{1}$$

2.2 Instantaneous amplitude and frequency

The IMFs obtained through decomposition of the signal using EMD lend themselves well to calculation of instantaneous amplitude and the instantaneous frequency through application of the Hilbert transform on each IMF [12]. First, the analytic signal z_i corresponding to each IMF is obtained, as shown below:

$$z_i[n] = c_i[n] + jH(c_i[n]), \tag{2}$$

where $H(c_i[n])$ is the Hilbert transform of an IMF $c_i[n]$. Writing Eq. 2 in the polar form we have:

$$z_i[n] = a_i[n]e^{j\theta_i[n]}, \tag{3}$$

where a_i represents the instantaneous amplitude, and θ_i the instantaneous phase, corresponding to IMF c_i. The instantaneous amplitude and instantaneous phase are given by:

$$a_i[n] = \sqrt{c_i^2[n] + H^2(c_i[n])} \tag{4}$$

$$\theta_i[n] = \tan^{-1}\frac{H(c_i[n])}{c_i[n]} \tag{5}$$

The instantaneous frequency, ω_i, can then be obtained as:

$$\omega_i[n] = \frac{\mathrm{d}\theta_i[n]}{\mathrm{d}n} \tag{6}$$

It is important to realize here that these instantaneous values represent the amplitude and frequency varying trigonometric function, as given in Eq. 3, that best fits the original signal at the local level, where each IMF represents the signal at the local level corresponding to a particular time-scale.

Once the Hilbert transform has been applied to all IMFs, and the instantaneous amplitudes and frequencies calculated, the original signal $x[n]$ can be expressed as follows:

$$x[n] = \sum_{i=1}^{K} a_i[n] \exp\left(j \sum \omega_i[n]\mathrm{d}n\right) \tag{7}$$

Equation 7 is a representation of the instantaneous amplitude (hence instantaneous energy) and frequency contribution by the locally occurring frequency components in each IMF. The real part of Eq. 7 allows a time–frequency–amplitude distribution called the Hilbert spectrum, $H[\omega, n]$, which is expressed as:

$$H[\omega, n] = \; Re \sum_{i=1}^{K} a_i[n] \exp\left(j \sum \omega_i[n]\mathrm{d}n\right) \tag{8}$$

2.3 Marginal spectrum

From the Hilbert spectrum (Eq. 8), it is also possible to calculate the marginal spectrum, $h(\omega)$, which measures the total amplitude (and hence energy) contribution from each frequency value over the whole signal length. The marginal spectrum is defined as:

$$h(\omega) = \sum_{n=1}^{N} H[\omega, n] \tag{9}$$

2.4 Speech data

The data used in this study have been obtained from the Massachusetts Eye and Ear Infirmary (MEEI) voice disorders database from 1994 [17]. Comprehensive details of the database and a list of studies using this database have been presented in [3]. Many studies, however, have only used a subset of the database, whether sustained vowels (e.g., 53 normal and 173 pathological [1], 53 normal 53 pathological [16]) or continuous speech (e.g., 51 normal and 161 pathological [9, 10, 18]). In this study, we have also used a subset of the database available to us, consisting of continuous speech signals from 51 normal and 161 pathological speakers. This allows us a good reference to compare our results with those obtained using other approaches on the same database. All of the speakers spoke the same sentence, which is: “when the sunlight strikes rain drops in the air, they act like a prism and form a rainbow”. The speech signals are sampled at 25 kHz and quantized at 16 bits/sample. Depending on the speaker, and also on the extent of voice pathology, the amount of time to speak the same sentence is different for different speakers. Therefore all speech signals, normal and pathological, are of different lengths.

2.5 Selection of speech portion length

Using the whole signal length for decomposition and feature extraction would have been computationally prohibitive, hence we decided to use shorter portions of continuous speech selected randomly from different parts of the speech signals. The length of the speech portions was chosen based on the following design criteria:

1. Using a portion length long enough to incorporate actual speech, and not consisting entirely of unvoiced/silence period.
2. Having a portion length which preserves the non-stationarity present in the speech signals.
3. Achieving a fair compromise between classification accuracy and computational time required by the methodology.

Regarding Point 1, it has been suggested in [19], which uses the same speech database as in this work, that a good classification score between normal and pathological speech may be achieved by using the silence parts of the database, which demonstrates the differences between the recordings. Therefore, the length of the speech portion should be a fair compromise that mitigates differences between normal and pathological speech that may arise due to reasons other than speech pathology, while at the same time reducing the overall computation time.

Point 2 relates to the non-stationarity aspects of portions of speech signals. Stationarity of speech signals is accepted over 10–30 ms segments [20], and any portion of continuous speech having a much longer length is able to capture the non-stationarity present in the signal. Though the 10–30 ms limit has been studied for normal speech, the presence of more transients, and larger spread of formants in pathological speech, as mentioned in this paper (Sect. 2.7) and elsewhere (e.g., [10]), gives reasons for confident assumption of non-stationarity for pathological speech segments of length much greater than 10–30 ms, especially if the speech portions have been obtained from different parts of the signal. In the same context, confidence in the use of EMD as a means of temporal and spectral feature extraction can be had due to the demonstrated superiority (e.g., [21]) of time–frequency resolution of the EMD-based method, which is independent of the length of signal used.

To objectively select a speech portion length representing a compromise between classification accuracy and computational burden, as listed in Point 3, while also taking Points 1 and 2 into consideration, different lengths of the speech signals were chosen, and the classification accuracy checked for each length using the features described in Sect. 2.7. Based on this analysis, the results of which are presented in Sect. 3, we selected a speech portion length of 800 ms, which corresponds to 20,000 samples, for further analysis and feature extraction. These 800-ms portions were extracted randomly from different parts of the normal and pathological speech signals.

2.6 Selection of IMFs

For the experiments in this study, the publicly available Matlab implementation [22] of EMD was used. Figure 1 shows 20,000 samples of randomly chosen continuous normal and pathological speech signals obtained from the database used in this study (Sect. 2.4) and first 10 of their corresponding IMFs obtained after application of EMD to the signals. The IMFs bring out the differing characteristics of the normal and pathological signals at different time-scales, which can be quantified in features useful for classification of normal and pathological speech.

Application of EMD to 20,000 samples of each speech signal results in the signals being decomposed into between 11 and 15 IMFs. The minimum number of IMFs was 11 for normal speech signals, and 13 for pathological speech signals. Since normal speech is in general more coherent than pathological speech, normal speech signals are expected to be decomposed faster than pathological speech.

In order to have uniformity in comparison between all speech signals, we decided to use the same number of IMFs from all signals, making 11 the highest number of IMFs to be used for further analysis. Similar to the approach taken for selecting the length of the speech portion described in Sect. 2.5, the number of IMFs to be used for analysis and feature extraction was selected objectively. For this purpose, the accuracy of classification between normal and pathological speech signals was measured using features extracted (Sect. 2.7) from 11 IMFs obtained from both normal and pathological speech signals. Thereafter, the IMFs were excluded from feature extraction one-by-one, starting from IMF 11, and the classification accuracy measured. In the same way, the classification accuracy was measured with IMFs removed one-by-one, but starting from IMF 1.

Based on the results of this approach, more details of which are provided in Sect. 3, it was decided to use IMFs 1–10 for feature extraction, as the highest classification accuracy was obtained using this combination of IMFs.

2.7 Feature extraction

The first 10 IMFs obtained from each of the normal and pathological speech signals were carefully analyzed, and discriminative temporal and spectral features extracted from the IMFs. These features are described in this section.

2.7.1 *Energy of intrinsic mode functions*

The energy of the analyzed IMFs c_i of the normal and pathological signals is selected as temporal feature. The energy, E_i, is calculated as:

$$E_i = \sum_{n=1}^{N} c_i^2[n], i = 1{-}10 \qquad (10)$$

where N is the length of the time samples.

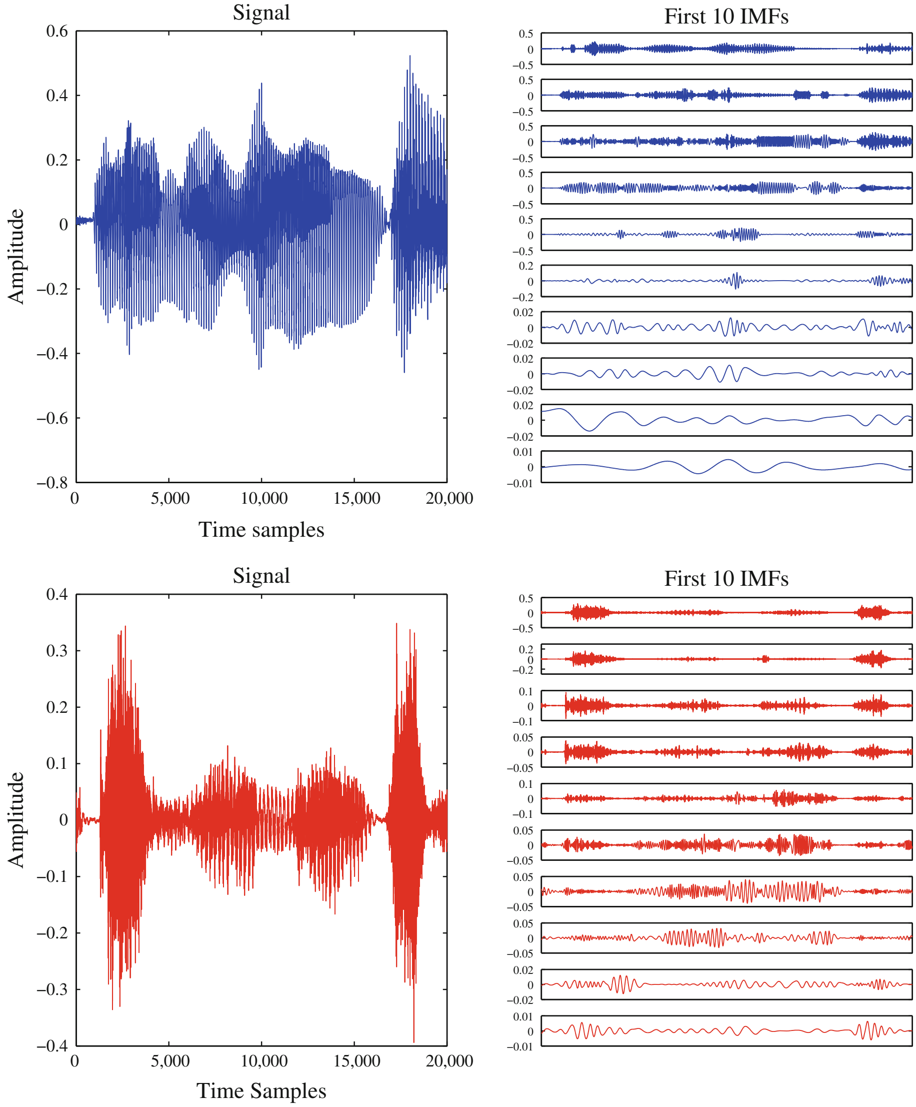


Fig. 1 800-ms portions of randomly chosen normal (*upper figure*) and pathological (*lower figure*) speech signals and their corresponding first 10 IMFs

Due to the normal speech being more coherent than pathological speech, the amplitude of normal speech signals in general has lesser variations than pathological signals. This difference is manifested in the decomposition process, whereby the amplitude variations of the pathological signals are spread over a larger number of IMFs as compared to normal speech signals, as can also been seen in Fig. 1. Therefore, the energy values E_i of IMFs i are expected to capture the spreading of the pathological signals' energy over a larger number of IMFs. It is also for this reason that using the energy values of the IMFs was found to be a better choice than using the energy of the signals themselves, as the distribution pattern of the signal energy into different IMFs is generally different for normal and pathological signals, which together with the other features described next helps increase the classification accuracy.

2.7.2 *Instantaneous joint time–frequency features*

The instantaneous frequency represents the time-varying value of the frequency component occurring at a time instant corresponding to a local time-scale extracted in an IMF. The instantaneous amplitude acts like a weighting coefficient for the instantaneous frequency values, and represents the energy value of a particular frequency component at a given time instant occurring at local time-scale in an IMF. In this context, the instantaneous amplitude can also be seen as the instantaneous energy density of the frequency components in an IMF.

Pathological speech is characterized by the more likely presence of transients and discontinuities, and more noisiness at higher frequencies. Normal speech generally has stronger and more distinguishable formant frequencies, whereas the formants in pathological speech are more spread and less structured. These differences are expected to be reflected in the instantaneous frequency and amplitude attributes. A time–frequency plot of IMF 7 from both normal and pathological speech signals is shown in Fig. 2.

We extract a set of two temporal features, and a set of two spectral features from the instantaneous amplitude and instantaneous frequency values obtained from each IMF. The set of instantaneous temporal features measures the following two values for each IMF i:

1. The spread of the instantaneous temporal energy density, SP$_i$(ITED). This is given by:

$$\mathrm{SP}_i(\mathrm{ITED}) = \frac{1}{N}\sum_{n=1}^{N}\left(a_i^2[n] - \bar{a}_i{}^2[n]\right). \tag{11}$$

2. The deviation of the instantaneous temporal energy density, D_i(ITED). This is given by:

$$D_i(\mathrm{ITED}) = \frac{1}{N}\sum_{n=1}^{N} n \cdot a_i[n]. \tag{12}$$

In both equations above, a_i represents the instantaneous amplitude of IMF i, and N is the length of the time samples.

Similarly, the set of instantaneous spectral features contains the following two feature values for each IMF i:

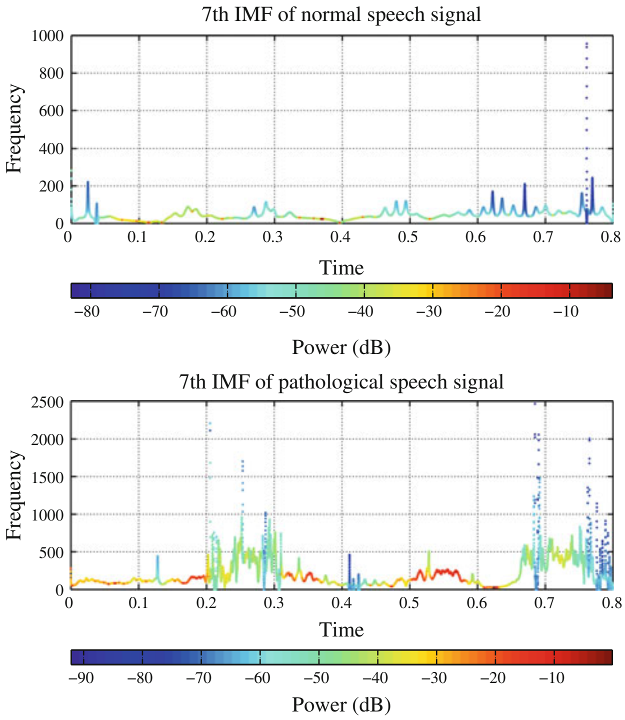


Fig. 2 Time–frequency plot of IMF 7 of both normal (*upper figure*) and pathological (*lower figure*) speech signals, showing a more even instantaneous frequency structure for normal speech signals

1. The spread of the values of the instantaneous frequency, SP(ω_i). This is given by:

$$\mathrm{SP}(\omega_i) = \frac{1}{N}\sum_{n=1}^{N}\left(\omega_i^2[n] - \bar{\omega}_i{}^2[n]\right). \tag{13}$$

2. The deviation of the instantaneous spectral energy density, D_i(ISED).

$$D_i(\mathrm{ISED}) = \frac{1}{N}\sum_{n=1}^{N} a_i[n]\omega_i[n]. \tag{14}$$

In the two equations above, a_i and ω_i represent the instantaneous amplitude and frequency, respectively, of IMF i, whereas N is the length of the time samples.

2.7.3 Sum of marginal spectrum

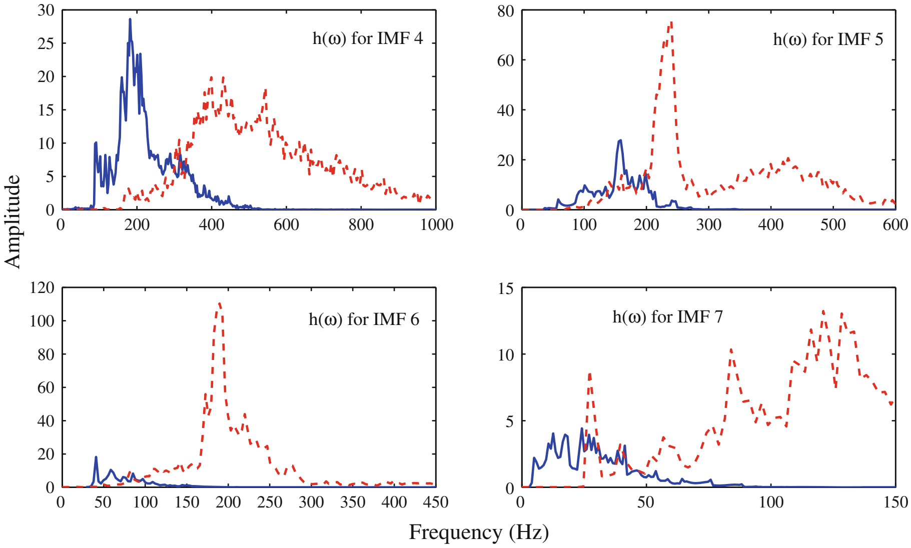
Due to the noisy and irregular nature of pathological speech, there are more components with higher energy at higher frequencies for pathological speech signals. This effect also depends on the level of speech pathology present in the speech, and can be exploited by extracting a feature from the marginal spectrum of the speech signals. As described in Sect. 2.3, the marginal spectrum represents the total amplitude (and hence energy) contribution from each frequency value over the whole signal length. It can be observed from Fig. 3 that beyond a frequency threshold, values of which for each IMF are explained later in this section, difference between the marginal spectrum amplitude for normal and pathological speech is discernable. Therefore, the sum of the marginal spectrum, $h(\omega)$, above a frequency threshold, for IMFs 1–10, is used as a spectral feature, and is given by:

$$\sum_{f=f_{\mathrm{th}_i}}^{f_{\max_i}} h_i(\omega), \quad i = 1\text{–}10 \tag{15}$$

where f_{th_i} represents the frequency threshold for each IMF used in the feature, $f_{\max_i}$ is the maximum value of the frequency in each IMF, and $\omega = 2\pi f$. The selection of the frequency threshold for each IMF is described next.

An experiment using a non-parametric statistical method, namely the Wilcoxon rank sum test [23], was designed to determine the frequency threshold beyond which there is discernable difference in the marginal spectrum of both normal and pathological speech signal. Firstly, frequency threshold values for each IMF based on visual analysis of the marginal spectrum plots of normal and pathological signals were selected. Then, four values, at fixed intervals, greater than each of these frequency thresholds, and four value less than the thresholds at the same intervals were used in the test to obtain that value of threshold which provides maximum separation between the

Fig. 3 Figures showing marginal spectrum $h(\omega)$ of IMFs 4, 5, 6 and 7 of both normal (*solid lines*) and pathological (*dashed lines*) speech signals. These figures illustrate that the marginal spectrum for pathological speech has higher amplitude beyond the frequency thresholds for the respective IMFs as compared to marginal spectrum for normal speech



marginal spectrum of normal and pathological speech signals. To carry out the test, first the sum of marginal spectrum was calculated for each IMF for each of the 9 frequency threshold values for 25 randomly selected normal and pathological signals. Then the Wilcoxon rank sum test was performed using the sum of marginal spectrum for all the 25 normal and pathological speech signals for each frequency threshold value to test the separation between the sum of marginal spectrum values for normal and pathological speech. The frequency threshold value with the lowest p value was then chosen as the threshold for that particular IMF. Table 1 lists the frequency thresholds decided by visual analysis of marginal spectrum plots, and the thresholds selected by the experiment described in this section. These thresholds were used to obtain the feature values represented by Eq. 15.

2.8 Classification

In order to employ a simple classification scheme for classification of normal and pathological speech signals, a linear discriminant analysis (LDA)-based classifier [24] was employed. A LDA-based classifier works in the following way: the feature vector is transformed into canonical discriminant function values given by

Table 1 Thresholds for calculating the sum of marginal spectrum as a discriminative feature

IMF no.	Frequency threshold by visual inspection (Hz)	Frequency threshold by statistical test (Hz)
1	5,000	5,100
2	2,100	2,100
3	1,550	1,575
4	900	925
5	500	500
6	300	400
7	150	125
8	100	20
9	50	40
10	25	15

$$f = a + b_1v_1 + b_2v_2 + b_3v_3 + b_4v_4 + b_5v_5 + b_6v_6 \qquad (16)$$

where $\{v\}$ represents the set of temporal and spectral features described in Sect. 2.7, $\{b\}$ represents the coefficients, and a is a constant. The posterior probability of each sample that occurs in each of the two groups is calculated using the discriminant scores and the prior probability values of each group. Based on the posterior probability, the sample is assigned to the group with the highest posterior probability.

The classification accuracy for pathological speech classification was calculated using stratified tenfold cross-validation on the whole speech data set. This technique randomly divides the training set into 10 disjoint subsets, where each subset has roughly equal size and same class proportions as in the training set. One subset is removed, and the classifier is trained using the other 9 subsets. The trained classifier is then used to classify the removed subset. This is repeated by removing each of the 10 subsets one at a time.

3 Results

3.1 Speech portion length and IMF number selection

The analysis and feature extraction of normal and pathological signals previously described was based on 10 IMFs extracted from continuous speech portions of 800 ms length chosen randomly from different parts of the signals. This length for the speech portions was selected as a result of the objective methodology described in Sect. 2.5, which

Table 2 Classification accuracy (rounded to nearest whole number) obtained using different continuous speech portion lengths

Portion length (ms)	Classification accuracy (%)
80	83
160	82
240	83
320	87
400	87
480	85
560	87
640	84
720	89
800	96
1,200	98

had the aim of finding a portion length offering a fair balance between classification accuracy and computational time. In this context, the classification accuracy obtained for different continuous portion lengths is shown in Table 2.

It can be seen from Table 2 that a portion length of 800 ms results in the highest classification accuracy compared to any smaller portion length. The decrease in accuracy for a shorter length of speech portion can be explained by the greater likelihood of silent periods, which have noise-like characteristics, affecting the correct classification of normal signals. At the same time, a portion length of 1,200 ms led to greater classification accuracy, due to the fact that more discriminatory characteristics of normal and pathological signals can be included in longer portion lengths, but at the cost of a higher computational time for signal decomposition and features extraction.

The objective methodology for selection of the number of IMFs used for feature extraction was also described in Sect. 2.6. The focus of this methodology was to find the number of IMFs which lead to the highest classification accuracy. Table 3 lists the classification accuracy obtained for different number of IMFs used, with features extracted from IMFs 1–10 having the highest accuracy. The classification accuracy for <6 IMFs was lower than 90 %. Also, removing the IMFs one-by-one, but starting from IMF 1, resulted in a classification accuracy of <90 %. This demonstrates that the lower index IMFs contribute more to the discrimination than the higher index IMFs.

Table 3 Classification accuracy (rounded to nearest whole number) obtained using different number of IMFs

Number of IMFs	Classification accuracy (%)
1–11	93
1–10	96
1–9	94
1–8	95
1–7	92
1–6	92

The empirical methods for selecting the speech portion length and the number of IMFs work well for different signal lengths, as well as for different number of signals, therefore forming objective methodologies than can be used in different experimental setups.

3.2 Effectiveness of extracted features

To check the effectiveness of the extracted features in discriminating between normal and pathological speech, we used unpaired *t* tests of the null hypothesis that the feature values obtained from each of the 10 IMFs for both normal and pathological speech have the same mean. The *p* values thus obtained are shown in Table 4 for each of the six features, and IMFs 1–10. It can be observed from Table 4 that the null hypothesis is rejected for all cases except for IMFs 9 and 10 for the feature E_i, and IMFs 4 and 10 for the feature SP_i(ITED). The combination of all the extracted features determines the effectiveness in discrimination between normal and pathological speech.

3.3 Classification results

The accuracy of classification between normal and pathological speech signals is calculated with a LDA-based classifier on the whole speech data set. The confusion matrix obtained using stratified tenfold cross-validation on the speech data set is shown in Table 5. From this table it can be seen that out of 51 normal speech signals, 49 were correctly classified, whereas 2 were misclassified as pathological. Out of 161 pathological speech signals, 154 were correctly classified, and 7 were misclassified as normal. The overall total classification accuracy represented by these results was calculated using the definition of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) as (TP + TN)/(TP + TN + FP + FN) [26]. The classification accuracy thus obtained is 95.7 %. The misclassified pathological signals were analyzed in detail, and it was found that these speech portions, extracted randomly from different parts of pathological signals, were more similar to normal speech in their characteristics, as demonstrated by a lack of noisiness, as well as of transients and discontinuities. Although the pathological signals from which these portions were extracted were perceptually quite similar to normal, we may speculate the resulting false negatives as a consequence of random selection of speech portions and a limited speech portion length. This, however, is corroborated by the finding discussed earlier in Sect. 3.1 that an increase

Table 4 *p* values obtained by performing unpaired *t* tests of the null hypothesis that the feature values obtained for each of 10 IMFs of normal and pathological signals used in the study have the same mean

Feature	IMF 1	IMF 2	IMF 3	IMF 4	IMF 5	IMF 6	IMF 7	IMF 8	IMF 9	IMF 10
E_i	0.00365	3.85×10^{-9}	1.38×10^{-7}	0.03045	0.01508	4.29×10^{-10}	3.43×10^{-9}	4.84×10^{-8}	0.0526	0.75809
D_i(ITED)	5.72×10^{-5}	0.01088	0.00575	0.00651	0.00156	0.00172	0.00589	0.00075	0.00788	0.00365
D_i(ISED)	4.22×10^{-42}	2.95×10^{-27}	1.78×10^{-22}	4.94×10^{-19}	3.60×10^{-16}	9.46×10^{-16}	1.67×10^{-17}	5.17×10^{-19}	1.78×10^{-20}	5.34×10^{-17}
$h(\omega)$	9.47×10^{-22}	9.60×10^{-10}	1.50×10^{-17}	1.48×10^{-24}	1.27×10^{-22}	2.23×10^{-20}	4.58×10^{-23}	6.34×10^{-15}	1.41×10^{-19}	1.61×10^{-16}
SP_i(ITED)	0.00428	9.57×10^{-9}	1.64×10^{-6}	0.10234	0.01351	1.03×10^{-9}	1.03×10^{-12}	4.97×10^{-6}	0.01415	0.08220
$\mathrm{SP}(\omega_i)$	1.44×10^{-11}	9.12×10^{-12}	2.93×10^{-12}	2.60×10^{-14}	1.75×10^{-13}	8.49×10^{-14}	1.04×10^{-12}	9.82×10^{-12}	1.43×10^{-9}	1.72×10^{-16}

Table 5 Classification results using linear discriminant analysis with tenfold cross-validation

		True classification	
		Pathological	Normal
Predicted classification	Pathological	154 (TP)	2 (FP)
	Normal	7 (FN)	49 (TN)

TP true positive, *TN* true negative, *FP* false positive, *FN* false negative

in speech portion length leads to better classification accuracy.

To further ensure no bias in the classification accuracy due to the unequal class sizes, the experiment was repeated with 50 normal signals, as well as 50 pathological signals randomly selected from the 161 available. In this case, the classification accuracy using tenfold cross-validation was 95.8 %, thereby demonstrating the robustness of the methodology.

4 Discussion

In this paper, we presented a methodology based on EMD to automatically discriminate between normal and pathological speech. A main strength of our approach is the simplicity and effectiveness of the methodology with respect to decomposition, feature extraction, and classification. Also important in our methodology is the ability to extract meaningful features from both the time and frequency domains. In this regard, extraction of true instantaneous features at different time-scales distinguishes our methodology from other approaches. The efficacy of instantaneous features in capturing the discriminatory information hidden in normal and pathological speech signals is demonstrated by the high classification accuracy obtained in this work.

Another strength of our approach lies in the use of continuous speech samples instead of sustained vowels. Compared to approaches that rely on sustained vowels or segmented portions of speech signals, our approach offers a more realistic automated assessment of voice quality, and demonstrates the advantage of using a non-stationary signal analysis technique to classify signals with non-stationary dynamics. An important point in this regard is that the use of continuous speech does not require an extra processing step to separate the sustained vowels from the speech. This is an advantage, as any error in the sustained vowel separation will propagate to the pathological speech classification stage. Due to the nature of pathological speech, which has a less periodic structure compared to normal speech, errors in sustained vowel detection and separation could be very common.

We used 800-ms portions of continuous speech selected randomly from different parts of the speech signals. This length of speech portion, which was selected objectively, corresponds to 20,000 samples and can effectively capture the non-stationarity present in speech. At the same time, extracting different portions from different parts of the signals contributes to the robustness of the methodology. We also selected a large subset of the IMFs for analysis and feature extraction using an objective methodology. This is important, since the useful temporal and spectral properties of signals are spread into a large subset of all IMFs.

We may also compare our methodology favorably with another EMD-based approach for pathological speech classification [16] which uses speech signals from the same database as used in this work. The approach in [16], based on sustained vowels instead of continuous speech, uses only 3 IMFs, numbers 2–4, without a rationale for excluding the rest of the IMFs. Also, the approach in [16] uses the maximum power spectral density and the corresponding frequencies of these 3 IMFs as the feature vector. Since the use of EMD can result in mode-mixing, which is a known issue with EMD [25], features dependent on frequency values alone might not be robust to a wide range of speech signals. In case the noise-assisted EMD [25] is used to reduce the effect of mode-mixing, the methodology becomes computationally prohibitive. Furthermore, as already discussed in Sect. 1, EMD is a non-stationary signal analysis technique, and should be taken advantage of for analyzing non-stationary signals, such as continuous portions of speech signals.

We can also compare our results with some recent approaches using a similar methodology and the same speech database as used in this work. The time–frequency matrix-decomposition approach [10] using the same voice database and an LDA classifier achieves a correct classification result of 98.6 % using speech segments of 80 ms. In comparison, our methodology achieves comparable classification accuracy using text-independent portions of speech while using a less complex method.

Another approach based on adaptive time–frequency decomposition using matching pursuit [9] using the same voice database achieves a correct classification result of 93.4 % using an LDA classifier. Features extracted from the ambiguity domain [18], but using segmented speech signals, reach a classification accuracy of 97.5 %, again using the same database and classifier as used in this paper. The work in [18] does not provide details about the segmentation.

An aspect of EMD decomposition that needs to be kept in view is the processing time required to extract all the IMFs. Since EMD is an iterative algorithm, the length of the signal to be decomposed adds to the processing time. In this regard, however, using shorter portions of speech signals is also an engineering compromise to keep check on the processing time required by the decomposition process. In the methodology presented in this paper, by far the largest chunk of the overall processing time is represented by the decomposition process, which consumes on average 10 s for each speech signal consisting of 20,000 samples. Extraction of the 6 features takes on average 2 s per speech signal. We would like to mention that these numbers have been calculated on a laptop computer (AMD processor, with 4 GB of RAM) running the current version of MATLAB software version 7.10.0, and are provided as a representative measure only. Other works we compared our methodology with in this section [9, 10, 16, 18] do not provide estimates of processing times.

As further work, we plan to incorporate the methodology presented in this paper in an automatic tool which can classify short portions of continuous speech. We have applied EMD for classification of telephone-quality pathological speech [27], and the results have been encouraging. This encourages the view that an automated pathological voice classification tool which can classify ordinary or telephone-quality speech with high accuracy can have clinical utility.

References

1. Henriquez P, Alonso JB, Ferrer MA, Travieso CM, Godino-Llorente JI, Diaz-de-Maria F (2009) Characterization of healthy and pathological voice through measures based on nonlinear dynamics. IEEE Trans Audio Speech Lang Process 17(6):1186–1195
2. Parsa V, Jamieson DG (2000) Identification of pathological voices using glottal noise measures. J Speech Lang Hear Res 43(2):469–485
3. Saenz-Lechona N, Godino-Llorentea JI, Osma-Ruiza V, Gomez-Vilda P (2006) Methodological issues in the development of automatic systems for voice pathology detection. Biomed Signal Process Control 1(2):120–128
4. Gelzinis A, Verikas A, Bacauskiene M (2008) Automated speech analysis applied to laryngeal disease categorization. Comput Methods Programs Biomed 91(1):36–47
5. Schlotthauer G, Torres ME, Jackson-Menaldi MC (2010) A pattern recognition approach to spasmodic dysphonia and muscle tension dysphonia automatic classification. J Voice 24(3):346–353
6. Godino-Llorente JI, Gomez-Vilda P (2004) Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors. IEEE Trans Biomed Eng 51(2):380–384
7. Shama K, Krishna A, Cholayya NU (2007) Study of harmonics-to-noise ratio and critical-band energy spectrum of speech as acoustic indicators of laryngeal and voice pathology. EURASIP J Adv Signal Process. doi:10.1155/2007/85286
8. Markaki M, Stylianou Y, Arias-Londono JD, Godino-Llorente JI (2010) Dysphonia detection based on modulation spectral features and cepstral coefficients. In: Douglas S, Kehtarnavaz N (eds) Proceedings of the 2010 IEEE international conference on acoustics, speech, and signal processing, Dallas, Texas, USA, pp 5162–5165
9. Umapathy K, Krishnan S, Parsa V, Jamieson DG (2005) Discrimination of pathological voices using a time–frequency approach. IEEE Trans Biomed Eng 52(3):421–430
10. Ghoraani B, Krishnan S (2009) A joint time–frequency and matrix decomposition feature extraction methodology for pathological voice classification. EURASIP J Adv Signal Process. doi: 10.1155/2009/928974
11. Parsa V, Jamieson DG (2001) Acoustic discrimination of pathological voice: sustained vowels versus continuous speech. J Speech Lang Hear Res 4(2):327–338
12. Huang NE, Shen Z, Long SR, Wu MC, Shih HH, Zheng Q, Yen NC, Tung CC, Liu HH (1998) The empirical mode decomposition

and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc R Soc Lond A 454(1971):903–995

13. Kaleem MF, Sugavaneswaran L, Guergachi A, Krishnan S (2010) Application of empirical mode decomposition and Teager energy operator to EEG signals for mental task classification. In: Armentano R, Monzon JE, Sacristan E, Lovell N (eds) Proceedings of the 2010 annual international conference of the IEEE engineering in medicine and biology society (EMBC), Buenos Aires, Brazil, pp 4590–4593
14. Mijovic B, De Vos M, Gligorijevic I, Taelman J, Van Huffel S (2010) Source separation from single-channel recordings by combining empirical mode decomposition and independent component analysis. IEEE Trans Biomed Eng 57(9):2188–2196
15. Schlotthauer G, Torres ME, Rufiner HL (2009) Voice fundamental frequency extraction algorithm based on ensemble empirical mode decomposition and entropies. In: Doessel O, Schlegel WC (eds) IFMBE proceedings, world congress on medical physics and biomedical engineering, vol 25/4, Springer, Berlin, pp 984–987
16. Schlotthauer G, Torres ME, Rufiner HL (2010) Pathological voice analysis and classification based on empirical mode decomposition. In: Esposito A et al (eds) Development of multimodal interfaces: active listening and synchrony; LNCS 5967, pp 364–381
17. Kay Elemetrics Corporation (1994) Massachusetts eye and ear infirmary voice disorders database. Version 1.03 (CDROM), Lincoln Park, NJ, USA
18. Sugavaneswaran L, Umapathy K, Krishnan S (2010) Exploiting the ambiguity domain for non-stationary biomedical signal classification. In: Armentano R, Monzon JE, Sacristan E, Lovell N (eds) Proceedings of the 2010 annual international conference of the IEEE engineering in medicine and biology society (EMBC), Buenos Aires, Brazil, pp 1934–1937
19. Malyska N, Quatieri TF, Sturim D (2005) Automatic dysphonia recognition using iologically-inspired amplitude-modulation features. In: Petropulu AP, Bystrom M (eds) Proceedings of the 2005 IEEE international conference on acoustics, speech, and signal processing, Philadelphia, Pennsylvania, USA, vol 1, pp 873–876
20. Furui S (1986) On the role of spectral transition for speech perception. J Acoust Soc Am 80(4):1016–1025
21. Adam O (2006) Advantages of the Hilbert Huang transform for marine mammals signal analysis. J Acoust Soc Am 120(5): 2965–2973
22. Flandrin P et al. (2007) Matlab codes for empirical mode decomposition algorithm. http://perso.ens-lyon.fr/patrick.flandrin emd.html. Accessed 25 Jan 2013
23. Hettmansperger TP, McKean J (2010) Robust nonparametric statistical methods, 2nd edn. Chapman and Hall/CRC Monographs on Statistics and Applied Probability, CRC Press, New York
24. Duda RO, Hart PE, Stork DG (2001) Pattern classification, 2nd edn. Wiley and Sons, New York
25. Wu Z, Huang NE (2009) Ensemble empirical mode decomposition: a noise assisted data analysis method. Adv Adapt Data Analysis 1(1):1:41
26. Moran RJ, Reilly RB, de Chazal P, Lacy PD (2006) Telephony-based voice pathology assessment using automated speech analysis. IEEE Trans Biomed Eng 53(3):468–477
27. Kaleem MF, Ghoraani B, Guergachi A, Krishnan S (2011) Telephone-quality pathological speech classification using empirical mode decomposition. In: Bonato P, Laine A, Lovell N (eds) Proceedings of the 2011 annual international conference of the IEEE engineering in medicine and biology society (EMBC), Boston, MA, USA, pp 7095–7098