33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Telephone-quality Pathological Speech Classification using Empirical Mode Decomposition

M.F. Kaleem, B. Ghoraani, A. Guergachi, S. Krishnan

Abstract— **This paper presents a computationally simple and effective methodology based on empirical mode decomposition (EMD) for classification of telephone quality normal and pathological speech signals. EMD is used to decompose continuous normal and pathological speech signals into intrinsic mode functions, which are analyzed to extract physically meaningful and unique temporal and spectral features. Using continuous speech samples from a database of 51 normal and 161 pathological speakers, which has been modified to simulate telephone quality speech under different levels of noise, a linear classifier is used with the feature vector thus obtained to obtain a high classification accuracy, thereby demonstrating the effectiveness of the methodology. The classification accuracy reported in this paper (89.7% for signal-to-noise ratio 30 dB) is a significant improvement over previously reported results for the same task, and demonstrates the utility of our methodology for cost-effective remote voice pathology assessment over telephone channels.**

I. INTRODUCTION

Pathological speech refers to speech problems that result from damage to or malfunction of the human speech organs. The traditional way of diagnosing pathological speech is based on listening to a patient's voice. However, such approaches are subjective to the experience of the specialist performing the diagnosis. Automated speech pathology detection is a very active field of research, since automation of this task improves the accuracy of assessment, and allows long distance identification and monitoring of pathological speech for patients in distant geographical areas. In this regard, telemedicine has an important role to play, since transmitting the voice over telephone channels allows a low-cost alternative for automated voice pathology assessment and vocal quality monitoring [1].

The main goal of automated pathological speech detection systems is to be able to characterize any input voice as either normal or pathological. These systems use signal processing tools, such as temporal and spectral methods, or the more recently introduced time-frequency methods, for characterization of the input voice signals. A classifier is then applied to descriptors obtained from this characterization in order to discriminate between normal and pathological speech.

Most of the temporal and spectral methods assume stationarity of pathological speech over 10-30 ms intervals, and use stationary signal processing tools. Non-stationary signal analysis techniques, on the other hand, can better reveal non-stationary behaviour of signals such as trends, discontinuities and repeated patterns, and hence match pathological speech more closely. Empirical Mode Decomposition (EMD) [2] is a recent technique for non-stationary signal analysis, which has found extensive application in broad areas of biomedical engineering (e.g. [3]). Compared to time-frequency non-stationary signal analysis techniques which have been used extensively in pathological voice classification (e.g. [4]), EMD is computationally simple, and also allows extraction of instantaneous descriptors from time and frequency domains without the need for elaborate signal transformations.

EMD has also been applied to classification of normal and pathological speech [5], but, like many other studies, in the context of sustained vowels, and not continuous speech. Sustained vowels can offer a controlled way of measuring voice characteristics and may produce good results, however they do not incorporate important vocal function attributes, for example rapid voice onset and termination, fundamental frequency and amplitude variations, and voice breaks. These attributes, though, are relevant as to how the voice quality is perceived in everyday life. Descriptors extracted from continuous speech samples can be more suitable for correct prediction of abnormal voice quality, since they can better represent the attributes mentioned previously, and properly reflect a person's real world experience.

Furthermore, there is sparse literature on telephone-quality pathological speech classification. The work described in [1] uses descriptors such as pitch and amplitude measures and harmonic-to-noise ratio extracted from sustained vowels to classify telephone-quality pathological speech with an accuracy of 74.2 %. In this paper, we present a methodology based on application of EMD on continuous speech samples to extract features for classification between telephone quality normal and pathological speech. Using four unique features, among them two based on instantaneous temporal and spectral measures not readily available in other popular approaches for speech analysis, we will classify telephone quality speech under different noise conditions with classification accuracy higher than that reported in [1]. Given the computational simplicity of our approach, the methodology presented in this paper elucidates the possibility of remote detection and assessment of voice pathology.

II. 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 func-

M.F. Kaleem (corresponding author), B. Ghoraani and S. Krishnan are with the department of Electrical Engineering, Ryerson University, Toronto, Canada. m2kaleem@ryerson.ca

A. Guergachi is with the Ted Rogers School of Information Technology Management, Ryerson University, Canada.

978-1-4244-4122-8/11/$26.00 ©2011 IEEE

tion (IMF) is a function $f : \mathbb{R} \rightarrow \mathbb{R}$ such that [2]: 1) the number of maxima, which are strictly positive, and the number of minima, which are strictly negative, are either equal, or differ at most by one, and 2) the mean value of the envelope, as defined by the maxima and the minima, is zero.

The technique for decomposition of the data into IMFs is known as *sifting* [2], which reduces the signal to be decomposed into K IMFs and a residue. Therefore, at the end of the decomposition, we can represent $x(n)$ as the sum of K IMFs and a residue r_K, as $x(n) = \sum_{i=1}^{K} c_i(n) + r_K(n)$.

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 [2]. First, the analytic signal z_i corresponding to each IMF is obtained as $z_i(n) = c_i(n) + jH[c_i(n)]$, where $H[c_i(n)]$ is the Hilbert transform of an IMF $c_i(n)$. Writing $z_i(n)$ in the polar form we have: $z_i(n) = a_i(n)e^{j\theta_i(n)}$, where a_i represents the instantaneous amplitude, and θ_i the instantaneous phase, corresponding to IMF c_i, and are given by: $a_i(n) = \sqrt{c_i^2(n) + H^2[c_i(n)]}$ and $\theta_i(n) = tan^{-1}\frac{H[c_i(n)]}{c_i(n)}$, respectively. The instantaneous frequency, ω_i, is then given by: $\omega_i(n) = \frac{d\theta_i(n)}{dn}$. 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: $x(n) = \sum_{i=1}^{K} a_i(n)exp(j \sum \omega_i(n)dn)$. This expression is a representation of the amplitude and frequency of each IMF as a function of time. This allows a time-frequency-amplitude distribution, $H(\omega, n)$, which is called the Hilbert spectrum. From the Hilbert spectrum, it is also possible to calculate the marginal spectrum, $h(\omega)$, which measures the total amplitude contribution from each frequency value over the whole signal length. The marginal spectrum is defined as: $h(\omega) = \sum_{n=1}^{N} H(\omega, n)$.

III. Method

The next subsections will present different aspects of the methodology for classification of telephone-quality pathological speech in more detail.

A. Data

The data used in this study consisted of the Massachusetts Eye and Ear Infirmary (MEEI) voice disorders database [6], which consists of 51 normal and 161 pathological speakers whose disorders span a variety of organic, neurological, traumatic and psychogenic factors. All the speech signals are sampled at 25 kHz and quantized at 16 bits/sample.

1) Telephone-quality speech simulation: In order to simulate telephone quality speech from the original data, distortions introduced by telephone transmission were incorporated into the original speech data, for both normal and pathological speech [1]. First, all the speech signals were down-sampled to 8 kHz, with the effective bandwidth being 4 kHz. Care was taken to low-pass filter the signals in order to prevent aliasing. The down-sampled data was then band-pass filtered with a linear filter having the frequency range 300 - 3400 Hz, which is the bandwidth of telephone transmission. The final manipulation of data included addition of additive white Gaussian noise (AWGN). The AWGN was added with 3 different signal-to-noise (SNR) ratios, which are 50 dB, 40 dB and 30 dB. This allowed us to test our methodology with different levels of noise present in the telephone quality speech. It should be mentioned that the 30 dB SNR was chosen to compare our results with those presented in [1]. Also, any mention of speech signals in this paper is with reference to telephone quality speech with additive noise as presented in this section, unless specifically mentioned otherwise.

B. Decomposition

EMD was used to decompose the telephone-quality normal and pathological speech signals into IMFs. The speech signals are of different lengths, and for each type of signal, we used the first 0.8 seconds, corresponding to 6400 time samples. This length of continuous speech is the same for both normal and pathological signals, and demonstrates the strength of our approach, since it does not rely on segmentation of data, and the non-stationarities in the signals are preserved and reflected in the extracted features. For the length of 0.8 seconds, the signals are decomposed into between 12 and 14 IMFs. The IMFs resulting from the application of EMD to all speech signals were rigorously examined. It was found that IMFs 11 and higher exhibited insignificant temporal or spectral structure, with extremely low signal amplitudes and number of signal oscillations. Therefore, IMFs 1 till 10 for all normal and pathological signals were selected for further analysis and feature extraction.

IV. Extraction and Evaluation of Features

In the next subsections we will describe the features that were extracted from IMFs 1-10 of both normal and pathological speech signals, and the rationale behind the choice. Also, as explained in Sec. II, it is possible to extract both temporal and spectral descriptors from IMFs. To take advantage of this, and keeping in view good classification results obtained previously using features extracted from the joint time-frequency domain (e.g. [4]), we extracted both temporal and spectral features.

A. Energy of intrinsic mode functions

The energy of the IMFs is larger for pathological speech signals as compared to normal speech signals. This can be explained in terms of the coherence of the normal speech, which means that normal speech signals are decomposed faster than pathological speech signals, and the amplitude of IMFs of normal signals also have lesser variations. Fig. 1, which shows IMF 6 of normal and pathological signals, is representative of this observation. Therefore the energy, E_i, of the 10 IMFs (1 till 10) of the normal and pathological signals is selected as a feature. The energy of the IMFs represents a temporal feature, and is calculated as: $E_i = \sum_{n=1}^{N} c_i^2(n), i = 1$ to 10.

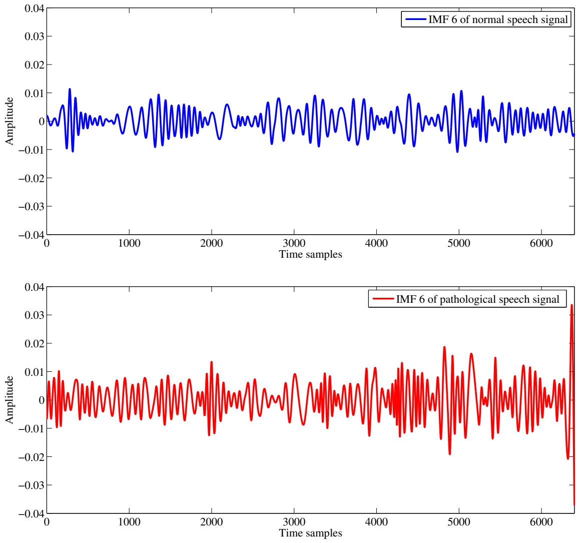

Fig. 1. IMF 6 of representative telephone quality normal (upper figure) and pathological (lower figure) speech signals (SNR=30 dB), showing larger amplitude variations of pathological speech IMFs, hence signifying larger energy content compared to normal speech IMFs.

B. Variance of instantaneous amplitude and frequency

In general, the frequency structure is found to be more even in IMFs of normal speech signals compared to IMFs of pathological speech signals, as the time-frequency plot of IMF 4 for normal and pathological speech in Fig. 2 demonstrates. The frequency structure of pathological signals has a larger spread along the frequency axis, which means that different time samples are more likely to have a different frequency content for an IMF of pathological speech. This can be explained by the presence of discontinuities in pathological speech. This also demonstrates the advantage of using continuous speech samples in the analysis, and the use of a non-stationary signal analysis method such as EMD, since the true structure of the speech signals is used to extract meaningful descriptors. Mode-mixing, which is a known issue with EMD [7], also has a role to play in this. Mode-mixing occurs when components of different frequencies intermittently come into existence, or disappear from a signal. This is more likely to happen for pathological speech signals. Mode-mixing is manifested by components with the same frequency spilling over into different IMFs. We use mode-mixing to our advantage by extracting discriminative features from IMFs which may have been mode-mixed, instead of using methods to eliminate mode-mixing, which are required in other approaches [5]. At the same time, this helps us keep our methodology computationally simple. Pathological speech signal characteristics also differ from normal speech in terms of larger variation of the instantaneous amplitude. The instantaneous amplitude represents the energy of the instantaneous frequency values, and the relatively even structure of instantaneous amplitude for normal speech signals can be explained in terms of stronger and more distinguishable formant frequencies in normal signals.

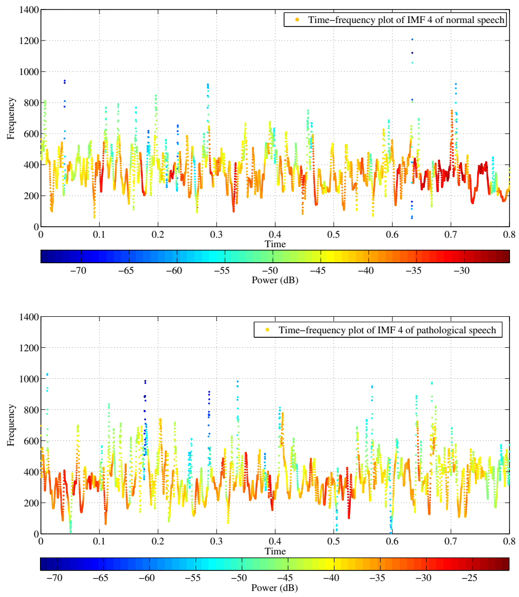

Fig. 2. Time-Frequency plot of IMF 4 of telephone quality normal (upper figure) and pathological (lower figure) speech signals (SNR=30 dB), showing difference in frequency structure of both types of speech signals.

TABLE I

AVERAGE CLASSIFICATION ACCURACY WITH DIFFERENT LEVELS OF NOISE ADDED

SNR (dB)	50	40	30
Avg. classification accuracy	89.7 %	91.3 %	92.6 %

As features, therefore, we use the variance, $Var(a_i)$ (temporal feature), of the instantaneous amplitude a_i, and the variance, $Var(\omega_i)$ (spectral feature), of the instantaneous frequency ω_i, obtained from each of the 10 IMFs. The instantaneous amplitude and instantaneous frequency are calculated by means of the analytic signal z_i as explained in Section II.

C. 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. Visual analysis of the marginal spectrum of all IMFs showed that beyond a frequency threshold, the 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 till 6, is used as a spectral feature, and is given by:$\sum_{f=f_{th_i}}^{f_{max_i}} h_i(\omega), i =$ 1 to 10, 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 frequency thresholds in Hz, for IMFs 1 to 6, are, respectively, 2000, 1500, 1100, 900, 500, 300. The frequency thresholds were

TABLE II

CLASSIFICATION RESULTS FOR ONE RUN OF EXPERIMENT IN TERMS OF MEASURES DEFINED IN [1]. SNR: SIGNAL-TO-NOISE RATIO. ACC: ACCURACY. SENS: SENSITIVITY. SPEC: SPECIFICITY. POS PRED: POSITIVE PREDICTIVITY. NEG PRED: NEGATIVE PREDICTIVITY.

SNR (dB)	Acc(%): (TP+TN)/TP+TN+FP+FN	Sens(%): TP/(TP+FN)	Spec(%): TN/(TN+FP)	Pos Pred(%): TP/(TP+FP)	Neg Pred(%): TN/(TN+FN)
50	93.8	96.9	84.3	95.1	89.6
40	92.5	98.8	72.5	91.9	94.9
30	91.5	98.1	70.6	91.3	92.3

selected using the Wilcoxon rank sum statistical test (not described here due to space limitation) so as to obtain values that provide maximum separation between the sum of marginal spectrum for normal and pathological speech signals beyond the frequency threshold values.

V. RESULTS

The objective of the work presented in this paper is to automatically classify telephone quality pathological speech signals using the features described in Section IV. We used 10 IMFs (1-10) of all the speech signals in our analysis, and the feature vector was constructed from the features extracted from these IMFs. This way the feature vector consisted of 36 elements, given that 3 features were extracted from each of the 10 IMFs, and 1 feature from the first 6 IMFs. In order to employ a simple classification scheme for classification of normal and pathological speech signals, a linear discriminant analysis (LDA) based classifier was employed by using the PASW Statistics 18 software [8]. The average classification results for different levels of noise added to the telephone quality signals are presented in Table I. Given the addition of AWGN to the speech signals after down-sampling and filtering, the experiment consisting of application of EMD to continuous speech samples (downsampled, filtered, noise-added), extraction of feature vector, and classification, was performed 5 times for each value of the SNR. The classification accuracy obtained for each value of SNR was then averaged over the 5 runs of the experiment. From Table I, it can be observed that as the SNR is decreased from 50 dB to 30 dB, the average classification accuracy decreases by less than 3% only. The classification accuracy of 89.7% at SNR of 30 dB compares very favorably with the classification accuracy of 74.2 % presented in [1].

Furthermore, we illustrate our results using the measures defined in [1]. These measures are defined in terms of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). The classification results in terms of these measures for one run of the experiment for each value of SNR used are shown in Table II. In terms of these measures as well, results from our methodology show significant improvement. The decrease in classification accuracy with an increase in the noise level in the speech signals can be explained by the fact that addition of noise to the normal speech signals makes their characteristics similar to pathological speech, as also illustrated by the increase in false positives.

VI. CONCLUSIONS

We presented a methodology based on EMD to discriminate between telephone quality normal and pathological speech with different levels of noise added. The main strength of our approach, as compared to other works, is the computational simplicity and effectiveness of our method, with respect to decomposition of the signals, the features extracted from the decomposed signals, and classification. Another strength of our approach lies in the use of continuous speech samples instead of sustained vowels. We extracted 4 features, 2 temporal and 2 spectral, from a large subset of the IMFs obtained after decomposing speech signals using EMD. The robustness of the features was demonstrated by achieving a high classification accuracy for different levels of noise present in the speech. The telephone quality pathological speech classification results are a considerable improvement over previously known results, also demonstrating, in particular, the efficacy for this task of instantaneous descriptors of signals, such as instantaneous amplitude and frequency, and in general, of EMD, using which such descriptors may be extracted with ease.

REFERENCES

[1] R. Moran, R. B. Reilly, P. de Chazal, and P. Lacy, "Telephony-based voice pathology assessment using automated speech analysis," *IEEE Trans. on Biomedical Engineering*, vol. 53, no. 3, pp. 468–477, March 2006.

[2] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N. C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. R. Soc. Lond. A*, vol. 454, no. 1971, pp. 903–995, 1998.

[3] M. F. Kaleem, L. Sugavaneswaran, A. Guergachi, and S. Krishnan, "Application of empirical mode decomposition and Teager energy operator to EEG signals for mental task classification," in *Proc. of the 32st Annual Intl. Conf. of the IEEE EMBS*, 2010, pp. 4590–4593.

[4] B. Ghoraani and S. Krishnan, "A Joint Time-Frequency and Matrix Decomposition Feature Extraction Methodology for Pathological Voice Classification," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, p. 11 pages, 2009.

[5] G. Schlotthauer, M. E. Torres, and H. L. Rufiner, "Pathological voice analysis and classification based on empirical mode decomposition," *Development of Multimodal Interfaces: Active Listening and Synchrony*, vol. 5967, pp. 364–381, 2010.

[6] "Massachusetts Eye and Ear Infirmary Voice Disorders Database, Version 1.03 (CDROM)," Lincoln Park, NJ:Kay Elemetrics Corporation (1994).

[7] M. F. Kaleem, A. E. Cetin, A. Guergachi, and S. Krishnan, "Using a variation of empirical mode decomposition to remove noise from signals," in *Proc. of 21st Intl. Conf. on Noise and Fluctuations (ICNF 2011)*, 2011, pp. 127–130.

[8] PASW Statistics 18 software. [Online]. Available: http://www.spss.com