https://doi.org/10.17576/apjitm-2018-0701-06

http://www.ftsm.ukm.my/apjitm
Asia-Pacific Journal of Information Technology and Multimedia
Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik
Vol. 7 No. 1, June 2018: 71 - 81
e-ISSN: 2289-2192

ROBUST SPEAKER GENDER IDENTIFICATION USING EMPIRICAL MODE DECOMPOSITION-BASED CEPSTRAL FEATURES

GHASEM ALIPOOR
EHSAN SAMADI

ABSTRACT

Automatic speaker gender identification is a field of research with numerous practical applications. However, this issue has not gained its deserved attention, in particular in the presence of environmental noises. In this paper, using the empirical mode decomposition (EMD), some new and improved mel-frequency cepstral coefficient (MFCC) features are developed to address the problem of robust speaker gender identification. In the proposed approach, EMD is employed as a filter bank to decompose the speech signal into its frequency bands. Furthermore, another variant is also developed in which the complete ensemble EMD (CEEMD) supersedes the EMD. Moreover, support vector machine (SVM) with radial basis function (RBF) kernel is employed for classification. Performance of these methods is examined for gender identification, in noise-free environments as well as in the presence of various Gaussian and non-Gaussian noises. Simulation results show that, although with fewer features used, utilizing the improved EMD-based cepstral features in noiseless situations leads to the same accuracy as that of the original MFCCs. However, in noisy environments the proposed methods outperform the conventional way of extracting the MFCCs.

Keywords: Automatic Gender Identification, *Empirical Mode Decomposition*, Mel-Frequency Cepstral Coefficients, Support Vector Machine

INTRODUCTION

In automatic speaker gender identification, gender of the speaker is identified based on the speech signal. This issue has numerous practical applications. Such systems can be helpful in sorting out the incoming phone calls on the basis of the speaker's gender in order to provide gender-oriented services. Furthermore, as a pre-processing unit, gender identification can enhance the accuracy of some recognition models, e.g. within the speaker identification (Khelif, Mombrun et al. 2017), speaker verification (Shahin 2018) and speaker diarization (Zhang, Weninger et al. 2017) systems.

Some features that have been studied in the literature for the purpose of gender identification are gender-specific features, e.g. pitch frequency (Harb and Chen 2005; Zeng, Wu et al. 2006; Levitan, Mishra et al. 2016) and formant (Childers and Wu 1991), as well as more general features such as correlation coefficients, Fourier Bessel coefficients (Spoorthy and Ramamurthy 2011) and mel-frequency cepstral coefficients (MFCCs) (Yücesoy and Nabiyev 2013; Ranjan, Liu et al. 2015; Safavi, Russell et al. 2018). Furthermore, various classification methods have been employed for this purpose, e.g. artificial neural networks (ANN) (Harb and Chen 2005), Gaussian mixture models (GMM) (Zeng, Wu et al. 2006; Yücesoy and Nabiyev 2013; Chen and Gu 2015), linear discriminant analysis(LDA) (Ranjan, Liu et al. 2015) and support vector machines (SVM) (Spoorthy and Ramamurthy 2011; Safavi, Russell et al. 2018). These studies are summarized in TABLE 1., in which the adopted features, models and datasets are included along with the achieved results. These algorithms have been

mostly studied for clean speech signals and mostly result in accuracy rates of about 95%. Despite these studies, we can say that the issue of automatic gender identification still requires more attention.

TABLE 1. A brief summary of the studies reported on gender identification

Ref.	Features	Classifier and Model	Accuracy & Achievement	Dataset
(Childers and Wu 1991)	Formant & Fundamental Frequencies	ANOVA	Formant Freq.: 98.1% Fund. Freq.: 96.2%	Personally recorded data
(Harb and Chen 2005)	Acoustic and Pitch Features	Neural Networks	93%	Recordings from French and English radio stations
(Zeng, Wu et al. 2006)	Pitch and RASTA-PLP	GMM	98%	TIMIT and some other multilingual speech samples
(Spoorthy and Ramamurthy 2011)	Fourier-Bessel coefficients	SVM	About 72.92%	Personally collected data
(Yücesoy and Nabiyev 2013)	MFCC	GMM	96.4%	TIMIT
(Chen and Gu 2015)	Tone and Energy Variations	GMM	98.9%	Lwazi
(Ranjan, Liu et al. 2015)	MFCC-Shifted Delta Coefficients	PLDA	97.63%	Fisher English (FE) and DARPA RATS corpora
(Safavi, Russell et al. 2018)	MFCC & Delta MFCC	GMM-SVM	79.18%	OGI Kids
(Levitan, Mishra et al. 2016)	Pitch & Spectral Features	logistic regression	93.8%	HMIHY

The most well-known features, vastly utilized for speech and speaker recognition, are the cepstral coefficients, in particular the MFCCs. MFCCs are short-time features extracted by applying the cosine transform on the log power spectra estimated over mel-scale-based bands. Although cepstral features have vastly and successfully utilized for some applications, the performance of these features for gender identification has not received enough attention. Moreover, our studies showed that a small amount of research has been conducted to consider the effect of noise. The main aim of the current study was to improve the performance of the celebrated MFCC features for speaker gender identification by virtue of the empirical mode decomposition (EMD). EMD is a robust adaptive time-frequency analysis method for representing a non-stationary signal as sum of components, called intrinsic mode functions (IMF), each with slowly varying amplitude and phase (Huang, Shen et al. 1998). EMD has proven to be quite versatile in a broad range of applications for extracting information from data generated in noisy nonlinear and non-stationary processes. Furthermore, it has been shown that the EMD essentially acts as a dyadic filter bank resembling those involved in wavelet decompositions (Flandrin, Rilling et al. 2004). Hence, our main idea is to first decompose the speech signal into its IMFs and then apply the mel-scale-based filter bank on the IMFs and select from the frequency bands of each mode. Moreover, support vector machine with RBF kernel is employed for classification. Performance of the proposed methods is studied in noise-free environments as well as in the presence of various Gaussian and non-Gaussian noises. Results of utilizing the proposed EMD-based cepstral features for gender identification are compared with that of the conventional way of extraction the MFCCs.

This paper is presented in the following order. We give a brief description of the EMD and SVM in sections 2 and 3, respectively. The proposed EMD-based cepstral features are

developed in section 4. Section 5 is dedicated to the simulation results. Finally some conclusion remarks are drawn in section 6.

EMPIRICAL MODE DECOMPOSITION AND ITS VARIANTS

EMD is a robust spectral decomposition method first developed by Huang *et al.* to adaptively decompose non-stationary signals into their intrinsic oscillatory components, i.e. IMFs (Huang, Shen et al. 1998). Since introduction in its original form in 2009, EMD has received some evolutions. Ensemble EMD (EEMD) was introduced as a solution to the mode mixing problem that EMD frequently suffers from (Wu and Huang 2009). Another variant is the complementary EEMD that possesses the completeness property (Torres, Colominas et al. 2011).

EMPIRICAL MODE DECOMPOSITION

Empirical mode decomposition is based on a simple premise that each signal is made of a number of intrinsic mode functions each with the following two conditions:

1. The number of extrema and the number of zero-crossings must either equal or differ at most by one.
2. At any point, the mean value of the envelopes defined by the local maxima and local minima is zero.

These IMFs are extracted through an algorithm known as the sifting process that can be summarized as follows:

1. Identify the local maxima (minima) of the signal, and then form the upper (lower) envelope by connecting all maxima (minima) by a curve, usually a cubic spline curve.
2. Form the mean envelope $m_1(t)$ by averaging these upper and lower envelopes.
3. Subtract mean envelope $m_1(t)$ form the signal to from the first probable component as:
$$h_1(t) = x(t) - m_1(t) \quad (1)$$
4. Ideally, $h_1(t)$ should be an IMF. But, if $h_1(t)$ does not satisfy the above definition of an IMF, let $h_1(t)$ be the new signal and repeat the steps 1-3 until the first IMF is extracted. Call the first IMF $c_1(t)$.
5. Let $r_1(t) = x(t) - c_1(t)$. Treat $r_1(t)$ as a new signal and repeat the steps 1-4 to extract other IMFs.
6. Repeat the above procedure K times, until $r_K(t)$ is smaller than a predetermined threshold or becomes a monotonic function that no more IMF can be extracted from.

Lastly, the signal $x(t)$ is decomposed into K IMFs $c_1(t)$ ... $c_K(t)$ and a residue $r_K(t)$ which can be either the mean trend of the signal or a constant. The signal $x(t)$ is composed by summing up all these components as:

$$x(t) = \sum_{k=1}^{K} c_k(t) + r_K(t) \quad (2)$$

This procedure can effectively sift the complex signals in time domain. IMF components provide valuable information about the signal. FIGURE *1* shows an example of decomposing a frame of speech signal into its first 5 IMFs and its residual using the EMD algorithm.

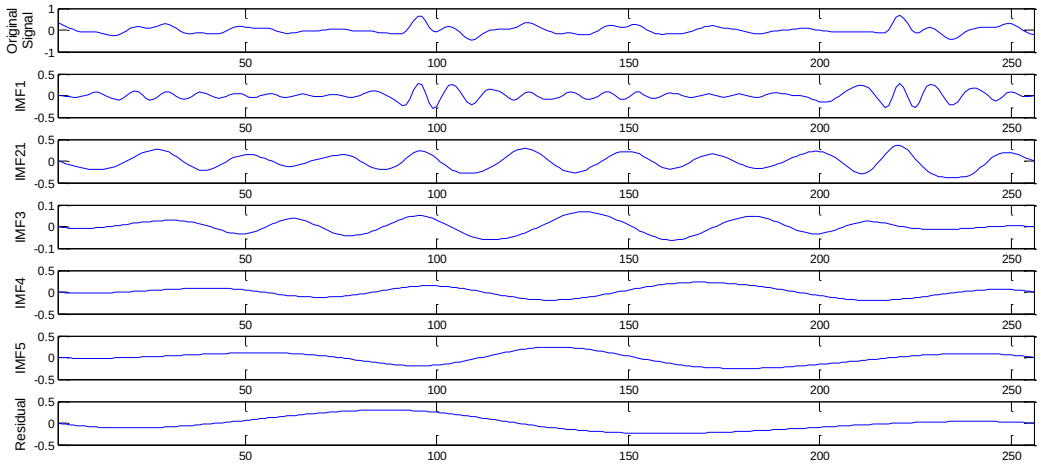

FIGURE 1. Decomposition of a speech signal into its first 5 IMFs using the EMD

COMPLEMENTARY ENSEMBLE EMPIRICAL MODE DECOMPOSITION

Ensemble EMD was introduced in 2009 to tackle the mode mixing problem of the EMD. For more information about this problem one can refer to (Wu and Huang 2009). EEMD algorithm is simple and can be summarized in the following steps:

1. Add a white noise to the signal.
2. Decompose the noisy signal into its IMFs, using the EMD procedure.
3. Repeat steps 1 and 2 with different samples of the white noise.
4. Obtain the final EEMD-based IMFs as the ensemble means of the corresponding EMD based IMFs in successive trials.

As a result of adding noise to the signal, signal reconstructed from the IMFs of the EEMD algorithm contains some residual noise. Furthermore, adding different samples of the white noise may lead to different IMFs. These issues are addressed in a modified version of the EMD, called complementary ensemble EMD. Now we try to describe this algorithm in brief. To distinguish between the IMFs resulted from the EMD and CEEDM procedures, the *k*th IMF based on these decomposition methods are indicated by $\mathrm{c}_k(t)$ and $\overline{\mathrm{c}_k(t)}$, respectively. $E_k(\cdot)$ is also defined as an operator that returns the kth IMF of a signal using the EMD algorithm. Furthermore, $\omega^m(t)$ is assumed to be the *m*th realization of a zero-mean unit-variance white noise and ε is a constant. Using these definitions, CEEMD can be summarized as follows (Torres, Colominas et al. 2011):

1. Extract the first IMF of the signal $x(t) + \varepsilon\omega^m(t)$ based on the EMD method, M times using different realizations of the noise ω. The first CEEMD-based IMF is calculated as:

$$\overline{c_1(t)} = \frac{1}{M}\sum_{m=1}^{M} c_1^m(t) \tag{3}$$

$c_1^m(t)$ is the first EMD-based IMF of the signal $x(t) + \varepsilon\omega^m(t)$ in the *m*th trial. The first residue is then calculated as:

$$r_1(t) = x(t) - \overline{c_1(t)} \tag{4}$$

2. For k=2, 3, …, the kth CEEMD-based IMF is calculated as the ensemble mean of the *k*th IMFs of the signal $r_{k-1}(t) + \varepsilon E_{k-1}\left(\omega^m(t)\right)$ for M different realizations of the noise ω, i.e.:

$$\overline{c_k(t)} = \frac{1}{M}\sum_{m=1}^{M} E_k\left(r_{k-1}(t) + \varepsilon E_{k-1}\left(\omega^m(t)\right)\right) \tag{5}$$

3. The residue is defined, at each iteration, as:

$$r_k(t) = r_{k-1}(t) - \overline{c_k(t)} \tag{6}$$

This procedure is repeated whilst the residue $r_k(t)$ has more than three extrema. The decomposed signal can now be expressed as:

$$x(t) = \sum_1^k \overline{c_k(t)} + r_K(t) \tag{7}$$

SUPPORT VECTOR MACHINE

Support vector machines are powerful tools for solving various classification problems. In addition to its sound theoretical foundation, SVM is of a good generalization performance in many real applications. In binary classification problems, depicted in FIGURE 2, consider the data set $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N\}$ in which $\boldsymbol{x}_i$s are d-dimensional feature vectors or training patterns, each with a class label $+1$ or -1. SVM searches for an optimal hyperplane H in R^d that separates this data space into two distinct sub-spaces, each corresponding to a class. This hyperplane can be described as:

$$H: \boldsymbol{x}^T\boldsymbol{w} + b = 0 \tag{8}$$

where $\boldsymbol{w}$ is the perpendicular vector to the hyperplane H and b is the bias of this function.

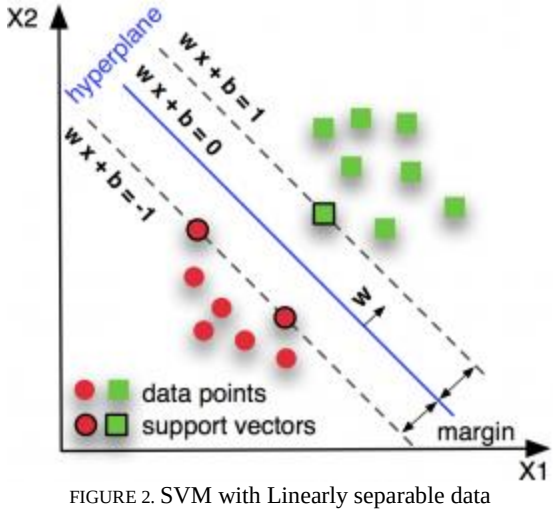

FIGURE 2. SVM with Linearly separable data

Data points that satisfy the following equalities:

$$\begin{aligned} \boldsymbol{x}_i{}^T\boldsymbol{w} + b \geq +1 \qquad & for\ \ y_i = +1 \\ \boldsymbol{x}_i{}^T\boldsymbol{w} + b \leq -1 \qquad & for\ \ y_i = -1 \end{aligned} \tag{1}$$

are on boundary hyperplanes that bounds two classes. The distance between these boundary hyperplanes is $\frac{2}{\|\boldsymbol{w}\|_2}$. According to the statistical learning theory, SVM achieves better classification ability by maximizing this distance. Hence, the goal of the SVM is to minimize the norm $\|\boldsymbol{w}\|_2^2$, subject to 0(9).

The main problem with this formulation is that if the problem is not linearly separable, there might be no solution to it. Emerging kernel methods can alleviate this problem by applying the linear algorithm on the transformed data in reproducing kernel Hilbert spaces (RKHS) that are nonlinearly related to the original input space (Rojo-Álvarez, Martínez-Ramón et al. 2014; Chen, Liu et al. 2015). It can be shown that for any RKHS $\mathcal{H}$, one can imagine a space, known as the feature space, in which the inner product can be calculated through evaluating its kernel function *K* in the input space. This mapping, denoted by ϕ and

termed feature mapping, projects the input $\boldsymbol{x} \in \mathcal{X}$ as the function $\phi(\boldsymbol{x})(\cdot) = K(\boldsymbol{x},\cdot) \in \mathcal{H}$. In other words, representing the function $\phi(\boldsymbol{x})(\cdot)$ as $\phi(\boldsymbol{x})$, the kernel K corresponds to a feature mapping ϕ for which

$$K(\boldsymbol{x}, \boldsymbol{y}) = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{y}) \rangle \tag{2}$$

Equation (10) is known as the kernel trick and states that the inner product in the feature space can be expressed in terms of the kernel function evaluation. This property makes it possible to calculate the inner products in these implicit high, or possibly infinite, dimensional spaces by means of the kernel functions evaluated in the low-dimensional input space. This in turn provides an efficient way to implicitly implement original linear algorithms, such as the SVM, in high-dimensional feature spaces while remaining in the low-dimensional input space, without direct reference to these nonlinear transformations. Therefore, in spite of linearity and convexity in RKHSs and possessing the property of universal nonlinear approximation, resultant algorithms can be solved in a reasonable complexity. One of the kernel functions often used in SVM is the Gaussian radial basis function (RBF) of the following from:

$$K(\boldsymbol{x}, \boldsymbol{y}) = exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{y}\|^2}{2\sigma^2}\right) \tag{3}$$

CEPSTRAL FEATURES

MFCCs are short-time features extracted by applying the cosine transform on the log power spectra estimated over mel-scale-based bands. Nonlinear sub-band decomposition in accordance to the mel scale is to cope with the human auditory system. This nonlinear scale relates to the linear scale in Hertz as:

$$f_{mel} = 2595\, log_{10}\left(1 + \frac{f_{Hz}}{700}\right) \tag{4}$$

Based on the commonly used procedure, MFCCs are extracted as follows:

1. The signal is first segmented into, usually overlapping, frames and each frame is then weighted by an appropriate windowing function. Over each frame, following steps are taken.
2. The spectrum of the input signal over each frame is estimated using the DFT.
3. The estimated power spectrum is mapped onto the mel scale, by applying a filter bank, with triangular overlapping frequency responses. The bandwidths as well as the central frequencies of these bandpass filters are distributed in accordance to the equation (12). The frequency responses of these filters are illustrated in FIGURE *3*.
4. Logarithm of the averaged power spectrum over each band is calculated.
5. Discrete cosine transform (DCT) is applied on the resultant log power spectra over all bands. The first coefficient is usually discarded and the remaining coefficients make up the MFCCs.

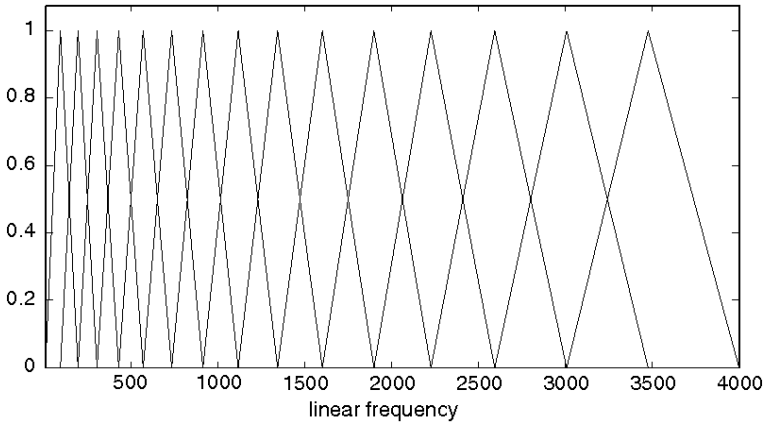

FIGURE 3. Triangular filter bank used in the original procedure of MFCC extraction

It is known that the MFFC features are very sensitive to additive noise. Studies have been done to alleviate this sensitivity and improve the robustness of these features in some applications, e.g. speaker identification (Faragallah 2018) and speech recognition (Khelifa, Elhadj et al. 2017). In this section, using the EMD, we develop some new and improved MFCC features to address this problem.

It can be shown that EMD, as well as its variants, acts as a dyadic filter bank in which the high-frequency contents of the signal reside mostly in the first IMFs and the last modes are usually of slowly-varying nature. This fact is illustrated in FIGURE 4 in which zero-crossing rates of the extracted IMFs of three signals with different frequency contents are depicted. It can be seen that going from the first IMFs to the last ones, the zero-crossing rates, as an evidence of the frequency contents of the IMFs, declines. This is in turn another proof of the fact that the EMD acts as a filter bank.

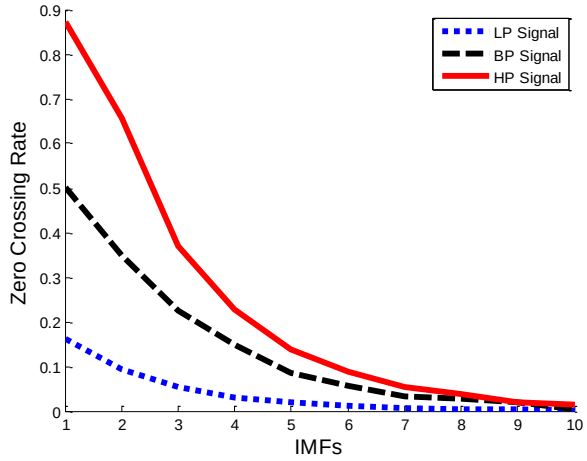

FIGURE 4. Zero-crossing rates of IMFs of three signals with different frequency contents

Based on these observations, in this paper a modification is made on the above procedure. In the proposed approach the speech signal is first decomposed into its IMFs. The EMD-based MFCCs are then extracted by applying the DCT on log power values calculated over some specific bands of the IMFs, chosen according to each mode's spectral contents. The developed algorithm can be summarized as follows:

1. Each (windowed) frame of the signal is first decomposed into 5 IMFs, using the EMD. This is because most speech signal frames can be efficiently decomposed into 5 IMFs with a negligible residue. Since common noises are usually of high frequency, the first

IMF is excluded from consideration. This exclusion improves the performance of the algorithm in presence of contaminating noises.

2. Over each IMF, steps 2-4 of the original MFCC procedure are applied. But, instead of all bands, log power values are only calculated over some selectively chosen bands.
3. The selected log power values over all IMFs are concatenated in order of the corresponding IMFs and the proposed EMD-based MFCCs are subsequently extracted by applying the DCT on these log power values.

To further improve the robustness of the EMD-based cepstral features, another scheme is also proposed in which a variant of the EMD, called complementary ensemble EMD (CEEMD), supersedes the EMD.

This procedure can be better followed using the block diagram of FIGURE 5. It should be noted that in the current study frame length is set to 256 samples and a Hamming window is used for windowing. The speech signal is decomposed into 5 IMFs, over each frame. As it has been stated above, the first IMF is discarded. Each of the 4 remaining IMFs is decomposed into 15 band, using a mel-scale triangular filter bank. On the other hand, the most efficient configuration of frequency bands as well as the total number of selected bands over all IMFs is set according to the frequency contents of each mode and validated through some experiments that will be reported in the next section.

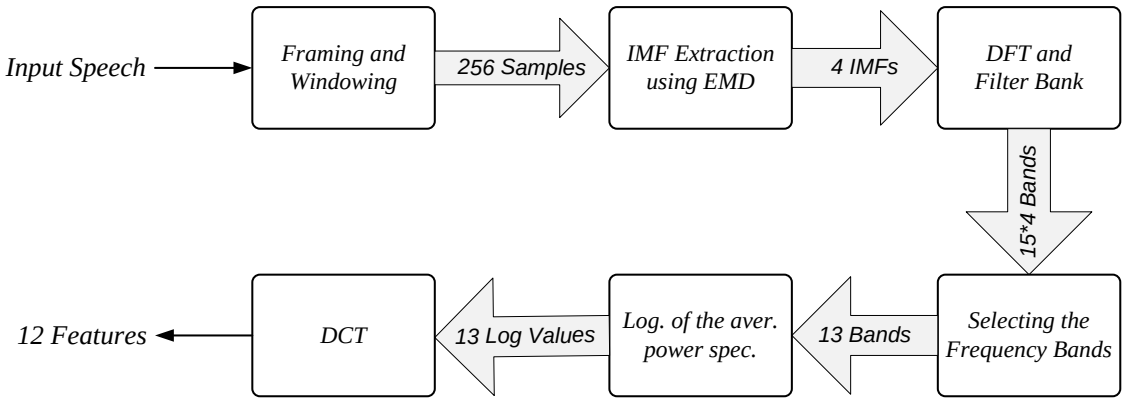

FIGURE 5. Procedure for extraction of the improved EMD-based cepstral features

RESULTS

Selected sets from train and test sections of the DARPA TIMIT database (TIMIT 1993) are used respectively for training and testing the algorithms. The selected training and testing sets contain 180 and 120 speech files, respectively. In each set, half of the speech files are uttered by female speakers and the others belongs to male speakers. Support vector machine (SVM) with radial basis function (RBF) kernel with σ=2.67 is employed for classification. For each algorithm, the classifier is first trained using all 180 training signals and then the trained model is tested over all 120 test files. Our developed methods for cepstral feature extraction using EMD and CEEMD, with M=10 and $\varepsilon = .002$, are compared with the original MFCC extraction procedure in both noise-free and noisy environments.

In step 2 of the proposed algorithm, the selection of the frequency bands is done based on the frequency contents of each IMF. In fact, over each IMF, the mel-frequency bands with a considerable level of energy are retained and other bands are discarded. Furthermore, some analytical experiments have been conducted to study the performance of various combination

forms of the frequency bands over all IMFs. Average results for some of these schemes are tabulated in TABLE 2. As one can see, the best possible result is obtained if bands 3 to 8 of the second IMF, bands 3 and 4 of the third and fourth IMFs and bands 2 and 15 of the fifth IMF are selected. These log power values construct a vector that the final proposed features are then obtained by applying the DCT on it. The resultant feature vector contains 12 coefficients.

TABLE 2. Identification accuracy for various combination forms of the frequency bands

Selected Bands					Total No. of Features	Accuracy
IMF1	IMF2	IMF3	IMF4	IMF5		
5 to 15	3 to7	4	3	1, 2 and 15	20	96.66%
9 to 14	6 to 8	5	3 and 4	1, 2 and 15	14	95.83%
---	3 to 8	3 to 4	3 and 4	1, 2 and 15	12	99.16%

Results obtained in noiseless situation for various feature extraction methods are summarized in TABLE 3. As one can see, in spite of fewer features used, utilizing the proposed EMD-based features leads to the same result as that of the commonly used MFCCs. However, for clean speech signals, the CEEMD-based feature extraction method has a slightly lower accuracy. These algorithms have also examined in the presence of various noise signals to assess the robustness of these features to environmental noises. Averaged results, obtained in the presence of white Gaussian noise as well as factory natural noise at several signal-to-noise ratio (SNR) levels, are presented in TABLEs 4 and 5. Results obtained in present of some other non-Gaussian noises, e.g. pink, babble, F16 aircraft, destroyer engine and high frequency channel noises, were overall same as that of the factory noise. These results reveal the better performance of the proposed methods in noisy environments, as compared to the original procedure for cepstral feature extraction.

TABLE 3. Identification accuracy in noise-free environment

Method	Number of Features	Accuracy
MFCC	14	99.16%
EMD-MFCC	12	99.16%
CEEMD-MFCC	12	96.66%

TABLE 4. Identification accuracy in the presence of Gaussian white noise

SNR (dB)	MFCC	EMD-MFCC	CEEMD-MFCC
0	50%	50%	64.16%
3	50%	50%	79.16%
5	51.66%	50%	80%
7	57.5%	50%	72.5%

TABLE 5. Identification accuracy in the presence of factory noise

SNR (dB)	MFCC	EMD-MFCC	CEEMD-MFCC
0	50.00%	71.66%	50.00%
3	67.50%	87.50%	69.16%
5	82.50%	92.50%	79.13%
7	89.16%	95.00%	80.00%

CONCLUSION

The main goal of the current study was to improve the robustness of the commonly used features known as MFCCs, in speaker gender identification. To address this issue two new methods were developed for extracting the cepstral-based features, using the empirical mode decomposition. It was shown that the EMD essentially acts as a dyadic filter bank. Hence, in the proposed approach the speech signal is first decomposed into its IMFs. The proposed EMD-based MFCCs are then extracted by applying the DCT on log power values calculated over some specific bands of the IMFs, chosen according to each mode's spectral contents. Moreover, to further alleviate the sensitivity of the proposed method to noise signals, another scheme was also developed based on a specific variant of the EMD algorithm, known as the CEEMD. Our simulation results showed that, in spite of fewer features, our proposed EMD-based cepstral features lead to a higher accuracy in presence of Gaussian as well as non-Gaussian noises. These results show the potential ability of the EMD, which is an adaptive time-frequency analysis method, in extracting the time-varying features of speech signals. This ability can be utilized in other application, in particular for speech-based emotion recognition, speaker identification and verification.

REFERENCES

Chen, B., W. Liu, et al. 2015. Theoretical methods in machine learning. Springer Handbook of Computational Intelligence, Springer Berlin Heidelberg: 523-543.

Chen, O. T. C. and J. J. Gu 2015. Improved gender/age recognition system using arousal-selection and feature-selection schemes. International Conference on Digital Signal Processing, DSP, Singapore, pp. 148-152.

Childers, D. G. and K. Wu. 1991. Gender recognition from speech. Part II: Fine analysis. Journal of the Acoustical Society of America, **90**(4): 1841-1856.

Faragallah, O. S. (2018). Robust noise MKMFCC–SVM automatic speaker identification. International Journal of Speech Technology, **21**(2): 185-192.

Flandrin, P., G. Rilling, et al. 2004. Empirical mode decomposition as a filter bank. Signal Processing Letters, IEEE, **11**(2): 112-114.

Harb, H. and L. Chen 2005. Voice-Based Gender Identification in Multimedia Applications. Journal of Intelligent Information Systems, **24**(2): 179-198.

Huang, N. E., Z. Shen, et al. 1998. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, The Royal Society, pp. 903-995.

Khelif, K., Y. Mombrun, et al. 2017. Towards a breakthrough speaker identification approach for law enforcement agencies: SIIP. *European Intelligence and Security Informatics Conference (EISIC)*, Athens, pp. 32-39.

Khelifa, M. O. M., Y. M. Elhadj, et al. 2017. Constructing accurate and robust HMM/GMM models for an Arabic speech recognition system. International Journal of Speech Technology, **20**(4): 937-949.

Levitan, S. I., T. Mishra, et al. 2016. Automatic identification of gender from speech. Proc. of Speech Prosody, pp. 84-88.

Ranjan, S., G. Liu, et al. 2015. An i-Vector PLDA based gender identification approach for severely distorted and multilingual DARPA RATS data. 2015 Workshop on Automatic Speech Recognition and Understanding (ASRU), IEEE, pp. 331-337.

Rojo-Álvarez, J. L., M. Martínez-Ramón, et al. (2014). A unified SVM framework for signal estimation. Digital Signal Processing: A Review Journal, **26**(1): 1-20.

Safavi, S., M. Russell, et al. 2018. Automatic speaker, age-group and gender identification from children's speech. Computer Speech & Language, **50**: 141-156.

Shahin, I. (2018). Speaker verification in emotional talking environments based on three-stage framework. *International Conference on Electrical and Computing Technologies and Applications (ICECTA)*, Ras Al Khaimah, pp. 1-5.

Spoorthy, S. and G. Ramamurthy 2011. Gender identification using significant Intrinsic Mode Functions and Fourier-Bessel expansion. International Conference on Signal Processing, Communication, Computing and Networking Technologies (ICSCCN), IEEE, pp. 86-89.
TIMIT (1993). DARPA TIMIT-Acoustic-Phonetic Continuous Speech Corpus. National Institute of Standards and Technology document NISTIR 4930.
Torres, M. E., M. Colominas, et al. 2011. A complete ensemble empirical mode decomposition with adaptive noise. International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 4144-4147.
Wu, Z. and N. E. Huang 2009. Ensemble empirical mode decomposition: a noise-assisted data analysis method. Advances in adaptive data analysis **1**(01): 1-41.
Yücesoy, E. n. and V. V. Nabiyev. 2013. Gender identification of a speaker using MFCC and GMM. 8th International Conference on Electrical and Electronics Engineering (ELECO), IEEE, pp. 626-629.
Zeng, Y.-M., Z.-Y. Wu, et al. 2006. Robust GMM based gender classification using pitch and RASTA-PLP parameters of speech. International Conference on Machine Learning and Cybernetics, IEEE, pp. 3376-3379.
Zhang, Y., F. Weninger, et al. 2017. A paralinguistic approach to speaker diarisation using age, gender, voice likability and personality traits. Proceedings of the 2017 ACM on Multimedia Conference, Mountain View, pp. 387-392.

Ghasem Alipoor
Ehsan Samadi
Electrical Engineering Department,
Hamedan University of Technology
Hamedan 5615913733, Iran.
alipoor@hut.ac.ir

Received: 27 December 2017
Accepted: 24 February 2018
Published: 26 June 2018