Importance of Higher Order Statistics for Speaker Recognition

Fidalizia Pyrtuh
[1]Department of Electronics and Communication Engineering, North-Eastern Hill University,Shillong,Meghalaya,India.

L Joyprakash Singh
[2]Department of Electronics and Communication Engineering, North-Eastern Hill University,Shillong,Meghalaya,India.

Abstract- For five decades, research in the field of speaker recognition, resulted in extensive competition amongst various methods and paradigms. Most of these methods employ frequency domain analysis. In this paper, we study and analyze the behavior of different representations of a short-phrase (Voice Password) speech signal, in the time domain, using Improved Complete Ensemble Empirical Mode Decomposition Technique (ICEEMD). Also, recognition performance of the technique, using the basic Gaussian Mixture Model (GMM) and Support Vector Machine (SVM) paradigms are shown. The database includes a set of genuine speaker speaking, exclusively, a particular password-phrase in multiple sessions.

Keywords – ICEEMDAN, Speaker recognition, GMM, SVM, Voice-password.

I. INTRODUCTION

Humans beings can decode speech signals and understand the information in speech. This concept of understanding the speech signal and recognizing the speaker, is often needed in many applications such as voice command control, audio archive indexing and audio retrieval etc[1]. Speaker recognition splits into speaker identification, speaker verification[2].Speaker identification refers in a technique of recognizing a speaker from a dataset of known speakers using a given voice sample. Speaker verification is to authenticate a known speaker. Automatic speaker recognition may be text dependent, text independent and voice password. If the user gives the content of the speech to the system then it is called text-dependent and the knowledge of the phrase is used for better recognition. In text-independent category, the system is not aware of the phrase spoken in the speech. This improves the flexibility of the recognition system but for some cases it may reduce the system efficiency. In voice-password, the text is different for different user, hence, the challenges for the system are to extract speaker specific information and speech related information[3,4,5]. In this paper, voice password based database is used for analysis.

Empirical Mode Decomposition method is a new technique, which aims for analyzing real-time nonlinear and non-stationary data. The significance of this unique technique is the ability of the method to decompose any complicated data set, into a finite and useful small number of Intrinsic Mode Functions(IMFs), that can be further processed by Hilbert transforms [6]. This new technique is adaptive, and, therefore, highly efficient. The decomposed data set is based on the instant time-scale characteristic of the data, makes it realizable to most nonlinear and non-stationary processes. This signifies that Hilbert Transform can give new physical information about nonlinear and non-stationary data series. It can also deduce the computation complexity in most dynamical changing non-stationary signal. In this method, the main concepts are the 'Intrinsic Mode Functions' which produce the instantaneous properties of the signal, which make the study of instantaneous frequency meaningful; and the processing of these frequencies, for complicated data sets, and hence makes signal processing simpler.

The rest of the paper is organized as follows. The feature extraction from voice samples, using Improved Complete Ensemble Empirical Mode Decomposition Technique (ICEEMD) and Hilbert transform are explained in section II and Section III. The feature selection and classifier used for the analysis is discussed in Section IV and Section V. Experimental results and discussions are presented in section VI.

II. IMPROVED COMPLETE ENSEMBLE EMD

Empirical mode decomposition (EMD) [7] is an adaptive (data-driven) method to analyze non-stationary signals stemming from nonlinear systems. It produces a local and fully data-driven separation of a signal in fast and slow oscillations. At the end, the original signal can be expressed as a sum of amplitude and frequency modulated (AM–FM) functions called "Intrinsic Mode Functions" (IMFs), plus a final monotonic trend. Noise-assisted versions have been proposed to alleviate the so-called "mode mixing" phenomenon, which may appear when real signals are analyzed[8], by populating the whole time–frequency space, taking advantage of the dyadic filter bank behavior of the

EMD [9,10]. The Complementary EEMD [11] significantly alleviated the reconstruction problem by using complementary (i.e., adding and subtracting) pairs of noise. However, the completeness property cannot be proven, and the final averaging problem remains unsolved since different noisy copies of the signal can produce a different number of modes. The complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) [12] proved to be an important improvement on EEMD, achieving a negligible reconstruction error and solving the problem of different number of modes for different realizations of signal plus noise. Applications of this technique can be found in areas such as biomedical engineering [13,14], seismology[15,16] and building energy consumption [17]. The Improved complete ensemble EMD (ICEEMD)[18] proposed two major improvements on the CEEMDAN method: The avoidance of the spurious modes and the reduction in the amount of noise contained in the modes are important features which grant more physical meaning to the obtained results. In this work we present improvements on this last technique, obtaining components with less noise and more physical meaning. The improved CEEMDAN's flowchart is as shown in Figure 1.

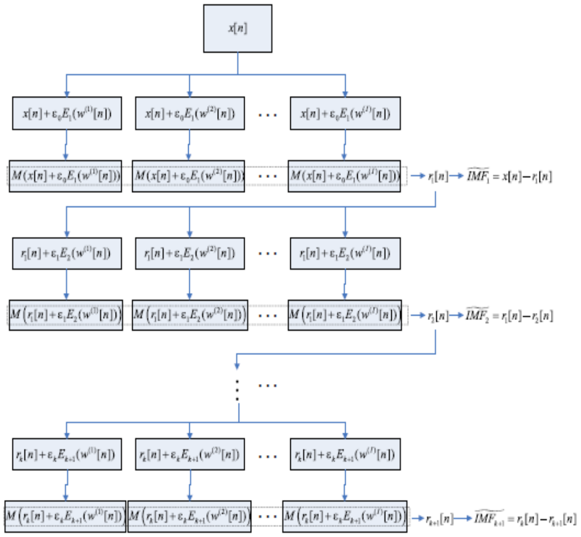
$x[n]$
$x[n]+\varepsilon_0 E_1(w^{(1)}[n])$
$x[n]+\varepsilon_0 E_1(w^{(2)}[n])$
$x[n]+\varepsilon_0 E_1(w^{(I)}[n])$
$M(x[n]+\varepsilon_0 E_1(w^{(1)}[n]))$
$M(x[n]+\varepsilon_0 E_1(w^{(2)}[n]))$
$M(x[n]+\varepsilon_0 E_1(w^{(I)}[n]))$
$r_1[n] \to \widetilde{IMF}_1 = x[n]-r_1[n]$
$r_1[n]+\varepsilon_1 E_2(w^{(1)}[n])$
$r_1[n]+\varepsilon_1 E_2(w^{(2)}[n])$
$r_1[n]+\varepsilon_1 E_2(w^{(I)}[n])$
$M(r_1[n]+\varepsilon_1 E_2(w^{(1)}[n]))$
$M(r_1[n]+\varepsilon_1 E_2(w^{(2)}[n]))$
$M(r_1[n]+\varepsilon_1 E_2(w^{(I)}[n]))$
$r_2[n] \to \widetilde{IMF}_2 = r_1[n]-r_2[n]$
$r_k[n]+\varepsilon_k E_{k+1}(w^{(1)}[n])$
$r_k[n]+\varepsilon_k E_{k+1}(w^{(2)}[n])$
$r_k[n]+\varepsilon_k E_{k+1}(w^{(I)}[n])$
$M(r_k[n]+\varepsilon_k E_{k+1}(w^{(1)}[n]))$
$M(r_k[n]+\varepsilon_k E_{k+1}(w^{(2)}[n]))$
$M(r_k[n]+\varepsilon_k E_{k+1}(w^{(I)}[n]))$
$r_{k+1}[n] \to \widetilde{IMF}_{k+1} = r_k[n]-r_{k+1}[n]$

Figure 1. The improved CEEMDAN's flowchart[18]

III. HILBERT TRANSFORM

From Hilbert transform, the IMFs yield analytic data, which are functions of time that may give sharp identifications of the speaker information. The final interpretation of the results is distinct time-dependent Hilbert amplitude or energy spectra[19]. After applying Improved complete ensemble EMD (ICEEMD) to the voice data, then Hilbert Huang Transform HHT is used on IMF obtained from ICEEMD technique, for extraction of instantaneous amplitude, instantaneous phase and instantaneous frequency[20] . First four IMFs are used for feature extraction in this study, since most frequency content and other important information, are present in these IMFs [21,22] . The Hilbert transform of a signal X(t) is Y(t) such:

$$Y(t) = H[x(t)] = \int_{-\infty}^{\infty} \frac{x(\tau)}{\pi(t-\tau)} d\tau$$

X(t) and Y(t) forms analytical signal Z(t)

$$Z(t) = X(t) + Y(t) = A(t)e^{j\theta(t)}$$

where, $A(t) = \sqrt{X^2(t) + Y^2(t)}$,

$$\theta(t) = \tan^{-1}\left[\frac{Y(t)}{X(t)}\right],$$

$$f(t) = \frac{1}{2\pi}\frac{d\theta(t)}{dt}$$

Where A(t) and $f(t)$ are instantaneous amplitude and instantaneous frequency respectively. Also, instantaneous energy is calculated for the voice samples, using Teager energy calculation[23].

y(n)=abs(x(n))^2 - x(n+1)*conj(x(n-1))

IV. SUPPORT VECTOR MACHINE

The SVM is used as a classifier in this paper, as it is less prone to overfitting and also give sparse solution, as compared to neural networks. It involves training and testing the voice sample data (genuine and imposter samples). During training, the set of data is divided into two attributes (genuine and imposter) with class label [-1 1]. This classifier aims to create a model to for the given voice samples and latter classify the train and test data according to their attributes and class label[24,25,26]. The two classes will be separated by a hyperplane as shown in Figure 2.Optimal separating hyperplane is at a distance of (-b/||w||) from the origin. The variable, ξ, measures the amount of misclassification among the two non- separable classes and ξ/||w|| provides the misclassification distance from the optimal separating hyperplane.

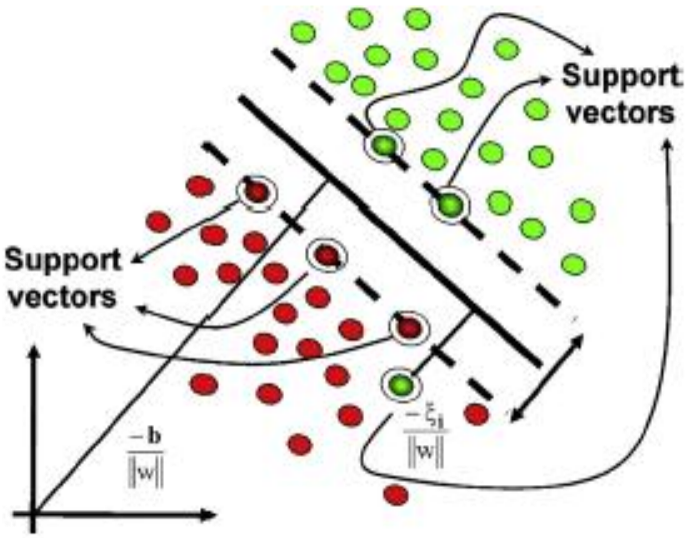

Figure 2. Example of Support Vector Machine

V. METHODOLOGY

The proposed methodology involves three major stages: feature extraction, feature selection and classification. The block diagram of speaker verification system is shown in Figure 3. The voice samples are collected from 4 females , where the first two data samples are consider as genuine candidate and the other two females data samples are consider to be imposter of the first two samples. Since, we are analyzing for voice-password technique, the first two females, will be given two separate text, while the imposter will be using the same text as the genuine speakers. Next, all four female speakers voice samples are passed through ICEEMD algorithm, which decomposed the signal into a series of mono-component signals called Intrinsic Mode Functions (IMFs). Amongst all the IMFs generated by the algorithm, only the first four IMFs are used, since they contained most of the useful information for further processing. Then Hilbert Huang Transform (HHT) is applied to the generated 4 IMFs for calculation of instantaneous amplitude, phase and frequency measurements. The block diagram of feature extraction method is as shown in Figure 4.

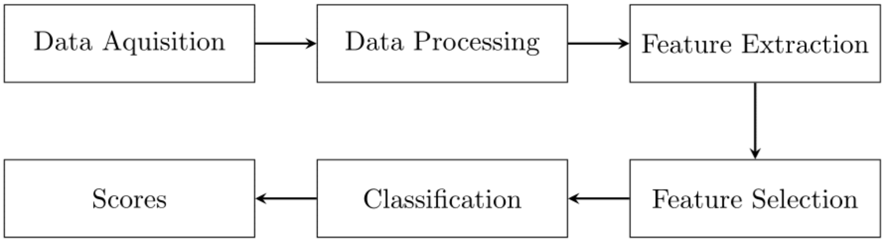

Figure 3. Block diagram of Speaker Recognition System

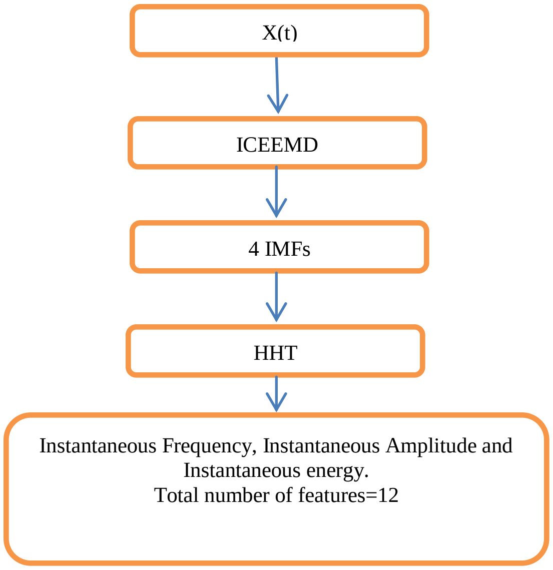
X(t)
ICEEMD
4 IMFs
HHT
Instantaneous Frequency, Instantaneous Amplitude and Instantaneous energy.
Total number of features=12

Figure 4. Feature extraction block diagram.

VI. RESULTS AND DISCUSSIONS

The training and testing data are processed using Matlab2015a on a Windows 7 platform. A box plot of the Instantaneous Frequency and Instantaneous Amplitude is as shown in Figure 5 and Figure 6, respectively. The plot of genuine candidate is marked as IMF1_P1 and for imposter, it is named as IMF1_P2. From Figure 5 and Figure 6, the difference between the IMFs of genuine and imposter candidate is not prominent. Hence, it is difficult to classify the two data accordingly. We have tried using GMM classifier and the result of misclassification is 52%.

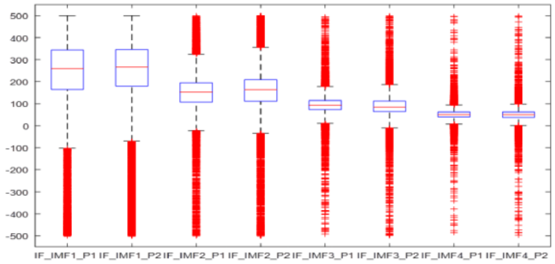

Figure 5. Box Plot of Instantaneous Frequency

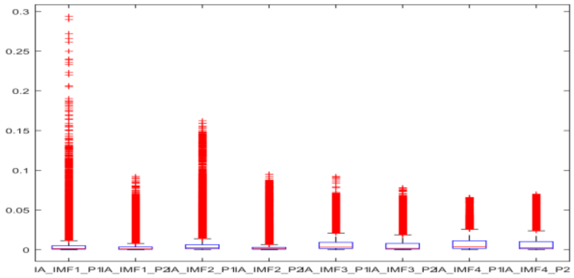

Figure 6. Box Plot of Instantaneous Amplitude

On the other hand, a box plot of higher order statistics, like, Kurtosis, Skewness and variance, is also as shown in Figure 7, Figure 8 and Figure 9, respectively. From the plot, we can conclude, that the difference between a genuine and imposter candidate, is quite prominent, hence classifying the two classes will become possible. So, the higher order statistics information will be used for further processing, to determine the speaker recognition, using Support Vector Machine classifier. Various kernels are available for SVM but the most commonly used are RBF (Gaussian Radial Basis Function) and KNN (k-Nearest Neighbor) classifier.

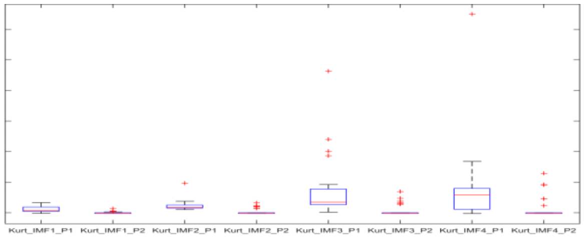

Figure 7. Box Plot of Kurtosis of Instantaneous Frequency

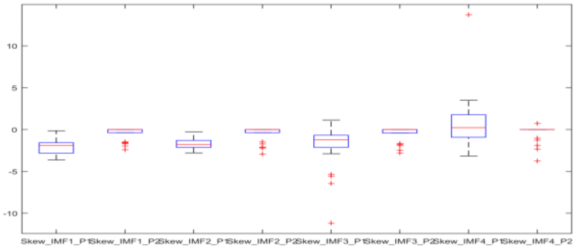

Figure 8. Box Plot of Skewness of Instantaneous Frequency

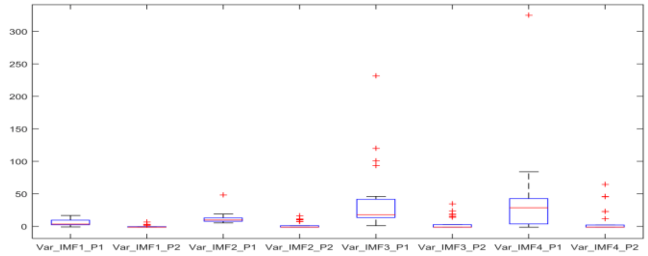

Figure 9. Box Plot of Variance of Instantaneous Frequency

The performance of the proposed voice data samples are tabulated in Table I and Table II, using SVM and SVM-KNN Classifier, respectively. Set 1 and Set 2 corresponds to the a set of genuine and imposter candidate.

SVM Parameters	Kernel Function	% of samples classified correctly	% of samples misclassified	Overall accuracy
Set 1	RBF	85.28	14.72	17.26%
Set 2		71.61	28.39	39.65%

Table I. The SVM classifier Performance

Table Ii. The Svm-Knn Classifier Performance

Sets	Kernel Function	% of samples classified correctly	% of samples misclassified	Overall accuracy

Set 1	NumNeighbors=5,
	Distance= minkowski
		84.66	15.34	18.11%

Set 2		96	4	4.16%

Hence, from the above observations, the Gaussian Radial basis function gives more consistent results, whereas the KNN classifier gives a more variable result. Further, preprocessing can be done to the raw speech data. Also, a better kernel SVM classifier can be used to enhance the recognition rate.

VII REFERENCES

[1] G.R. Doddington, "Speaker Recognition-Identifying People by their voices", Proc. Of IEEE, pp. Vol 73 no 11, 1985.

[2] S.L. Agnitio, "BATVOX: The Leading Tool for Forensics Speaker Recognition", 2008.

[3] R.K. Das, S.Jelil, S.R.Mahadeva Prasanna, "Development of Multi-Level Speech based Person Authentication System" , J Sign Process Syst, 88: 259, 2017.

[4] Muda Lindasalwa, Begam Mumtaj, I. Elamvazuthi , "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", Journal of Computing, volume 2, issue 3, March 2010.

[5] H. Abdelmajid, H. Mansour, Gafar Zen, Alabdeen Salh, A. Mohammed Khalid, "Voice Recognition using Dynamic Time Warping and Mel-Frequency Cepstral Coefficients Algorithms", International Journal of Computer Applications, Volume 116 – No. 2, April 2015.

[6] P. Flandrin, G. Rilling, P. Goncalvès, "Empirical mode decomposition as a filterbank", IEEE Signal Process. Lett. 11 (2), 112-114,2004.

[7] N.E. Huang, Z. Shen, S.R. Long, M.C. Wu, H.H. Shih, Q. Zheng, N. Yen, C.C. Tung, H.H. Liu, "The empirical mode decomposition and the Hilbert spectrum for non-linear and non-stationary time series analysis", Proc. R. Soc. Lond. A 454,903–995,1998.

[8] X. Navarro, F. Poree, G. Carrault, "ECG removal in preterm EEG combining empirical mode decomposition and adaptive filtering",Proc. of the 37th IEEE Int. Conf. on Acoust., Speech and Signal Process, ICASSP 2012, pp.661–664 IEEE, 2012.

[9] Y. Lei, Z. He, Y. Zi, "Application of the EEMD method to rotor fault diagnosis of rotating machinery", Mech. Syst. Signal Process, 23(4), 1327–1338,2009.

[10] F. Plante, G.F. Meyer, W.A. Ainsworth, "A pitch extraction reference database", 4th Eur. Conf. on Speech Communic. and Technol., Madrid, pp.837–840, Spain, 1995.

[11] J.R. Yeh, J.S. Shieh, N.E. Huang, "Complementary ensemble empirical mode decomposition: a novel noise enhanced data analysis method", Adv. Adap. Data Anal. 2 (02), 135–156, 2010.

[12] M.E. Torres, M.A. Colominas, G. Schlotthauer, P. Flandrin, "A complete ensemble empirical mode decomposition with adaptive noise", Proc. 36th IEEE Int. Conf. on Acoust., Speech and Signal Process, ICASSP 2011, Prague, CzechRepublic, pp. 4144–4147, 2011.

[13] X. Navarro, F. Poree, G. Carrault, "ECG removal in preterm EEG combining empirical mode decomposition and adaptive filtering" Proc. of the 37th IEEE Int. Conf. on Acoust., Speech and Signal Process, ICASSP 2012, pp.661–664,IEEE, 2012.

[14] Rajib Sharma, S. R. M. Prasanna, Hugo Leonardo Rufiner, Gastón Schlotthauer, "Detection of the Glottal Closure Instants Using Empirical Mode Decomposition Circuits Syst Signal Process, DOI 10.1007/s00034-017-0713-4.

[15] J. Han, M. van der Baan, Empirical mode decomposition for seismic time–frequency analysis, Geophysics 78 (2), O9–O19,2013.

[16] A. Hooshmand, J. Nasseri, H.R. Siahkoohi, Seismic data denoising based on the complete ensemble empirical mode decomposition, in: Int. Geophysical Conf.and Oil & Gas Exhibition, Istanbul, Turkey, pp. 1–4, 2012.

[17] R. Fontugne, J. Ortiz, N. Tremblay, P. Borgnat, P. Flandrin, K. Fukuda, D. Culler,H. Esaki, Strip, bind, and search: a method for identifying abnormal energy consumption in buildings, in: Proc. of the 12th Int. Conf. on Inf. Process. inSens. Netw., ACM, pp. 129–140, 2013.

[18] A. Marcelo, B Colominasa, B Gastón Schlotthauera, E María, "Torres Improved complete ensemble EMD: A suitable tool for biomedical signal processing", Biomedical Signal Processing and Control, 19-24, 2014.

[19] Y. Lei, Z. He, Y. Zi, Application of the EEMD method to rotor fault diagnosis of rotating machinery, Mech. Syst. Signal Process. 23 (4), 1327–1338, 2009.

[20] N.E. Huang, Z. Shen, S. Long, " A new view of nonlinear water waves: The Hilbert Spectrum", Annual Review of Fluid Mechanics, 417-457, 1999.

[21] N. Ramesh Babu, B. Jagan Mohan, "Fault classification in power systems using EMD and SVM", Ain Shams Engineering Journal, 8, 103–111, 2017.

[22] Rajib Sharma, Ramesh Bhukya, S.R.M Prasanna, "Analysis of the Hilbert Spectrum for Text-Dependent Speaker Verification. Speech Communication" , Journal of Speech Communication,2017.

[23] J. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal", Proc IEEE Intl Conf Acoustics, Speech and Signal Processing, ICASSP.1990, 381–384, vol.1, Apr. 1990.

[24] Steve R. Gunn, "Support Vector Machines for Classification and Regression", Ph.D. Dissertation, University of Southampton, UK:1997

[25] Nello Cristianini, John Shawe Taylor, "An introduction to Support Vector Machines and other kernel-based learning methods". Cambridge University Press;2000.

[26] Q Zhang, YQ Yang. Research of the kernel function of support vector machine. Electrical Power Science Engineering 2012:28(5):42-5