A Robust Speaker Identification System for Natural and Whispered Speech

Garima Sood[1]
[1]Bachelor of Technology,
Electronics and Communication Engineering Department
National Institute of Technology, Hamirpur
Himachal Pradesh, India
[1]garima.sood.06@gmail.com

Sidharth Pancholi[2], Amit M. Joshi[3]
[2]Research Scholar, [3]Assistant Professor
Electronics and Communication Engineering Department
National Institute of Technology, Jaipur
Rajasthan, India
[2]sid.2592@gmail.com,[3]amjoshi.ece@mnit.ac.in

***Abstract*—Speaker Identification, using both speech and whisper, is an emerging research topic. The main challenge lies in improving the robustness of the system in highly noisy environment. In this paper, different identification algorithms for both normal and whispered speech have been compared to check the robustness. Mel frequency cepstral coefficient (MFCC) method, Gabor filter-bank methods and an Empirical Mode Decomposition (EMD) based AM-FM approach have been employed for feature extraction. The extracted features have been classified with various classifiers such as Support Vector Machine, Fine K Nearest Neighbor (KNN) and Weighted KNN. A database of 16 subjects has been created, both in normal, as well as in whispered mode. It is observed that the separate Gabor filter-bank method provides the best accuracy (98.1% with Fine KNN) for whispered speech, and the AM-FM approach offers the best accuracy (98.9% with Fine KNN) for normal speech.**

Keywords— Empirical Mode Decomposition (EMD); Gabor filter; k-NN; SVM; Speaker Identification

I. INTRODUCTION

Speaker Identification based on speech has gained a lot of popularity in security systems. This can be used in automation, banking, and even in patient monitoring. Speech based security systems provide better user convenience than other biometric systems like fingerprint, retina, face recognition, DNA. Both normal speech as well as whispered speech can be used to solve the purpose, but whisper based systems have an edge over the other systems, as a whisper cannot be forged easily, unlike speech, which can be recorded secretly to fool the system [1]. As compared to normal or neutral speech, a whisper is different in the sense that it has low signal to noise ratio (SNR), and its lower frequency formants shift to higher frequencies [2]. While whispering, there is a lack of glottal excitation, and no vibrations are generated in the vocal chord [2, 3].

There are basically two main approaches to Speaker Identification: a closed set approach, and an open set approach. Closed set Speaker Identification involves the comparison of the test audio data with the available data, and the speaker identity of the closest match is returned. Whereas, open set identification is a combination of closed set identification as well as speaker verification [4]. However, a closed set approach is considered in the paper. Fig.1 shows the waveforms of normal and whispered speech. In normal speech, the amplitude of the signal is high as compare to whispered speech. There are many peaks noticed in the normal speech because of the vocal chord vibration [2].

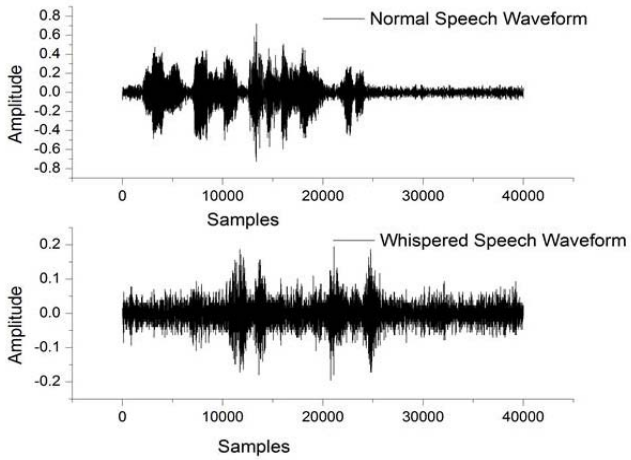

Fig.1 Waveforms of normal and whispered speech of the same subject

There have been various approaches observed in literature for Speaker Identification, with different normalization techniques such as the MFCC (Mel frequency cepstral coefficients [5, 6], 2D Gabor filter-bank features (in spectral and temporal domain) [7], separate 1D Gabor filter-bank features (separately in spectral and temporal domains) [8], AM-FM based approach (where instantaneous features serve as features) [9]. All of the above methodologies are examined for normal as well as whispered speech and results are compared in this paper.

The subsequent paper is organized as follows: Section I gives a brief overview of natural and whisper based speaker identification system. Section II covers the methods employed for calculating features from the speech signal. Section III describes the proposed speaker identification methods in detail. In Section IV, the experimental evaluation and analysis of results is performed and the conclusion of this work is derived in Section V.

978-1-5386-4318-1/17/$31.00 ©2017 IEEE

II. FEATURE EXTRACTION

In this section, an overview of the various methods which are used for extraction of features for speaker identification has been discussed.

A. Mel Frequency Cepstral Coefficients

MFCC is the most popular feature extracting technique in short time spectral analysis [10]. This method uses two types of filters: linearly spaced and logarithmic filters. The signal is expressed in Mel frequency scale (having a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz) to capture the important characteristics of speech [11]. The following formula is used to compute the Mels for a given frequency f in Hz [12].

$$\text{Mel (f)} = 2595*\log10\ (1+f/700) \tag{1}$$

B. Gabor Filter-bank Features

A 2D Gabor filter shows a spectro-temporal receptive field. The temporal and spectral modulation frequencies need to be paired. These filters are aligned to certain spectro-temporal modulation patterns that are present in speech. The paired spectro-temporal modulation frequencies determine the filter's shape and the pattern which is capable of acquiring the strongest response in that filter. The main parameters of these 2D Gabor filters are the spectro-temporal center modulation frequencies and the spectral and temporal modulation bandwidths [13].

C. Separate (1D) Gabor Filter

This approach involves the usage of two 1D Gabor Filters in spectral and temporal domain separately, instead of using a single 2D spectro-temporal Gabor Filter. Separate 1D filters are used instead of inseparable 2D filters. Separate 1D filter has an advantage of being modular in structure, where overall 2D filter operation is performed by two separate 1D operations. The following equation gives the 1D Gabor Filter, $g_{\omega,v(x)}$.

$$h_w = \begin{Bmatrix} 0.5-0.5\cos\left(\frac{2\pi x}{w}\right), & -\frac{w}{2}<x<\frac{w}{2} \\ 0, & else \end{Bmatrix} \tag{2}$$

$$s_\omega(x) = i\omega x \tag{3}$$

$$g_{\omega,v(x)} = s_\omega(x).h_{\frac{v}{2\omega}}(x) \tag{4}$$

Where, $s_\omega(x)$ is the carrier which is a sinusoidal signal and $h_{\frac{v}{2\omega}}(x)$ is the Hann envelope of the filter. Here w is the width and ω are the radian frequency [13].

D. Empirical Mode Decomposition

AM-FM based extraction is involved to extract the instantaneous frequencies (IFs) using Empirical Mode Decomposition (EMD). Subsequently, the Hilbert Transform is applied for feature extraction. The integration of EMD and Hilbert Transform is collectively known as Hilbert Huang Transform (HHT) as designated by NASA [14-16].

The EMD method reduces the given signal to Intrinsic Mode Functions (IMF). An IMF can be understood as a zero-mean waveform, where its number of zero-crossings differs from its number of extrema at most by one. The number of these zero-crossings roughly indicates each mode's mean frequency [17]. The process of generating IMFs is known as sifting which involves the following steps:

1. For the input signal x(t), the upper envelope u(t) and lower envelope l(t) are formed using the cubic spline function.
2. The mean of the envelopes is calculated as m(t):

$$m(t) = \frac{u(t) + l(t)}{2} \tag{5}$$

3. The proto-IMF h(t) is the difference between the main signal and the mean.

$$h(t) = x(t) - m(t) \tag{6}$$

4. If the stoppage criterion [18] and IMF definition are satisfied, then the proto-IMF is assigned as an IMF component c (t) else the above procedure is repeated until the desired result is obtained. Hence the first IMF is obtained as $h_{1,k}$ defined by eq. (7). The Standard Deviation (SD) for the stoppage criterion is normally set as 0.3 [7].

$$h_{1,k}(t) = h_{1,k-1}(t) - m_{1,k}(t) \tag{7}$$

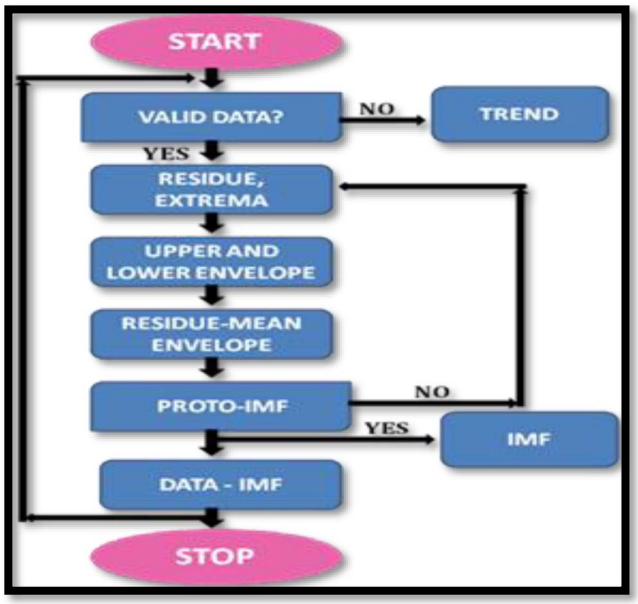

Fig. 2 Block diagram for generation of IMF

5. Similarly, other IMFs are calculated from the remaining signal r (t).

$$r(t) = x(t) - c(t) \tag{8}$$

The flowchart of the above steps is shown in Fig. 2 which is used to generate IMF.

Now, the obtained IMFs are Hilbert Transformed to obtain the Instantaneous Frequencies (IFs). For any signal [7], its Hilbert Transform is given as:

$$H(t) = \frac{1}{\pi} P \int_{\infty}^{\infty} \frac{x(\tau)}{t-\tau} d(\tau) \tag{9}$$

Where, P is the Cauchy principle value of the singular integral.

If x (t) is the message signal, then an analytic signal a (t) can be formulated as:

$$a(t) = x(t) + j.\hat{x}(t) \quad (10)$$

$$f(t) = \frac{1}{2\pi}\frac{d\phi(t)}{dt} \quad (11)$$

Where, $\hat{x}(t)$ is the Hilbert Transform of x(t). The instantaneous frequency f (t) is given as (11).

III. PROPOSED ALGORITHM FOR SPEAKER IDENTIFICATION

This section covers the steps of proposed algorithm for speaker Identification using both neutral and whispered speech. The speech signal is first windowed into frames, and then the signal is further processed.

Step1: First, a database of 8 males and 8 females is created, in which their voice samples are recorded both in whisper and neutral speech for 30, 10 and 5 secs duration. The sampling frequency is 8 KHz. The recordings are saved in .wav format.

Step2: The voice signal is now preprocessed and the sample is divided into frames. This is known as windowing of the input speech signal.

Step3: Now, the signal is passed through the respective filters for the extraction of features. For the MFCC, Gabor and Separate Gabor approach, the signal is first passed through triangular filters to obtain log mel spectral features.

Step4: The log Mel spectral features are then processed by Discrete Transform (DCT) to obtain MFCC features. 2D Gabor and separate Gabor features are extracted from the log mel spectral features by passing through the 2D spectro-temporal Gabor filter-bank and separate 1D Gabor filters respectively. The obtained features are then normalized using the Histogram Equalization (HEQ) technique [20] and only their absolute values are considered.

Step5: For the AM-FM approach, the windowed signal is processed through Empirical Mode Decomposition (EMD) and then it is applied to Hilbert Transform to obtain Instantaneous Frequencies (IFs) which are considered as the features.

Step6: Various classification models are employed (SVM, Fine K-NN and Weighted K-NN) and the accuracies for both whisper and speech is obtained and analyzed.

IV. RESULT AND ANALYSIS

The dataset was created for the experimental evaluation and analysis for 16 speakers, 8 males and 8 females, who are aged between 20-35 years. Speech samples were recorded for some finite duration (i.e. 5 sec, 10 sec and 30 sec) in neutral mode as well as in whispered mode (5 Sec and 10 sec). The subjects were asked to read out sentences s10-s16 of the CHAINS corpus [19]. The samples were recorded at the lab itself, under no special conditions and at a frequency of 8000 hertz, by using the inbuilt microphone of the PC. The speaker was positioned at a rough distance of 50 centimeters from the device. All the work was carried out in MatLab2015a simulation platform. Firstly, the log Mel spectral features are calculated by setting an amplitude spectrogram frame length of 25ms and temporal resolution of 100 frames/sec. A Mel scale is used using 31 equidistant filters of triangular nature whose Centre frequencies are between 0-8000 Hz [21].

Now, 2D Gabor filter bank features are extracted from log Mel spectrogram by spectro-temporal Gabor filter bank consisting of five filters with spectral and temporal modulation frequencies ($\omega_s\left(\frac{cycles}{channel}\right), \omega_t(hertz)$) as (0.000, 0.0), (0.029, 6.2), (0.060, 9.9), (0.122, 15.7), (0.250, 25.0). The maximum extension of filters in spectral domain was thrice of the Mel-bands (i.e 93) and in temporal domain it extended upto 400 ms. Similarly, individual 1D gabor filters in spectral and temporal domain of the same modulation frequencies as of 2D filters are applied on the log Mel spectral features to extract the separate gabor features. All the other parameters are also kept same as that in 2D gabor filter-bank features. Then HEQ normalization is done on all the features and their absolute value is considered. Fig.3 shows all the extracted feature waveforms for whispered speech. The y axis represents log Mel Spectrum, MFCC, Gabor Filter bank and Seperate gabor filter respectively.

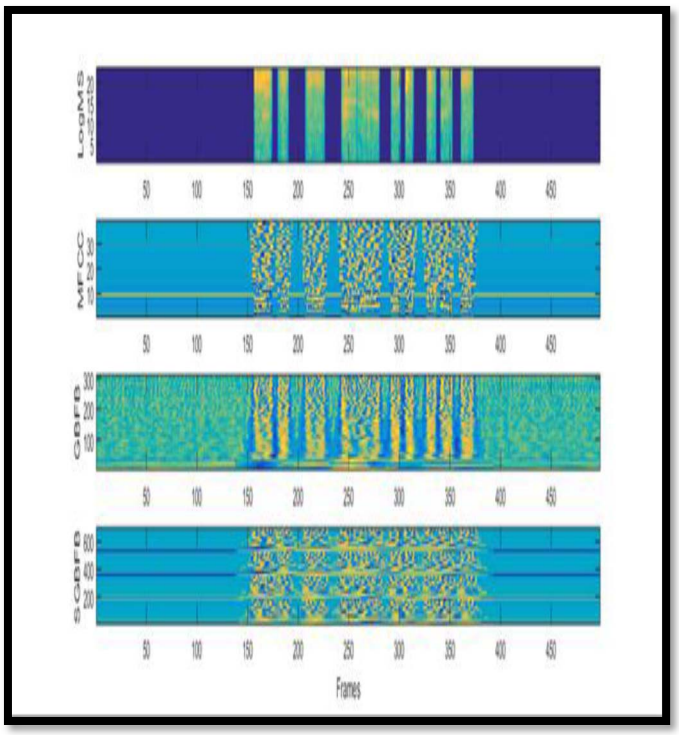

Fig.3 The various features for whispered speech.

For the EMD based approach a frame size of 1024 samples is taken with 50% overlap in adjacent frames. The first seven IMFs are considered empirically, and their Hilbert transform is used to obtain the IFs whose absolute values serve as features [7]. Fig.4 shows the first seven IMFs of a speech signal.

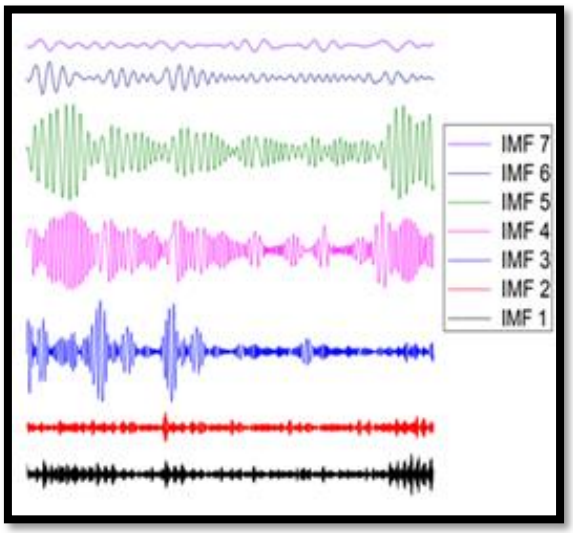

Fig.4.The first seven IMFs of speech signal.

In the experiment, 30% of the data was utilized for testing purpose and the rest was used for the training of the system. For training and testing by normal speech, the MFCC features have classification accuracy of 78.10%, 44.90% and 35.30% with a linear SVM, Fine K-NN and Weighted K-NN respectively. The Gabor filter features acquire an accuracy of 81.40% with Linear SVM, 92.50% with Fine K-NN and 80.0% with Weighted K-NN. The separate Gabor filter bank features provide an accuracy of 81.70%, 97.60% and 80.50% when identification was done using SVM, Fine K-NN and Weighted K-NN respectively. The AM-FM based model for normal speech yield results as 91.00%, 98.90% and 97.50% with Linear SVM, Fine K-NN and Weighted K-NN respectively. Fig.5 shows a bar chart of the obtained results for normal speech.

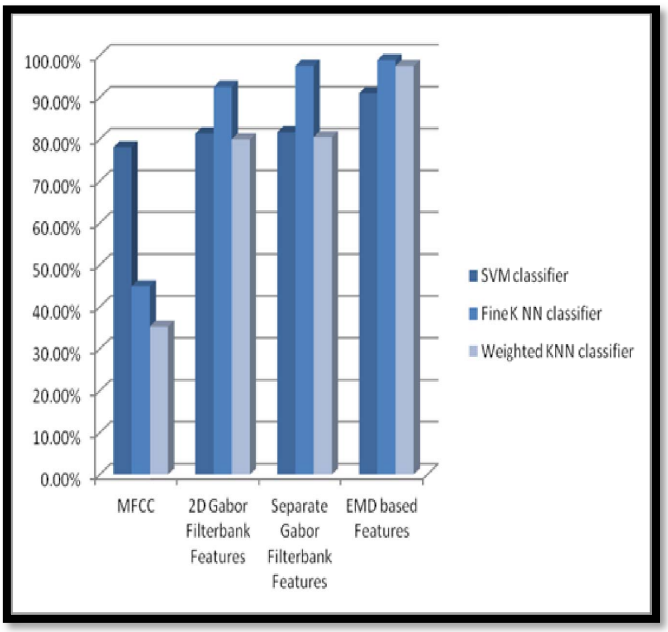

Fig.5 Results Obtained For Normal Speech

For training and testing by whispered data, the MFCC features obtain an accuracy of 71.10% with Linear SVM, 58.30% with Fine K-NN and 38.00% with Weighted K-NN. The 2D Gabor features result in an accuracy of 78.50%, 94.40% and 83.50% with Linear SVM, Fine K-NN and Weighted K-NN respectively whereas the separate Gabor filter-bank features give 80.00%, 98.10% and 81.50% accuracies for the same. The AM-FM approach gave the results as 84.60% with Linear SVM, 94.60% with Fine K-NN and 91.40% with Weighted K-NN for whispered speech. The results are shown in Fig.6 via various approaches for whispered speech. However, it has been observed that Fine k-NN classifier has better accuracy as compared to other classifiers. In Fine k-NN Euclidean distance is considered as one with its neighbors. Fig.7 shows the confusion matrix for normal speech with Fine k-NN classifier. The confusion matrix gives correctly classified and the misclassified values and shows that most of the time, correct predictions were made to have an accurate and robust model which correctly identifies the speaker.

The above results show that for normal speech, the AM-FM based feature extraction model provide better accuracy than others for all the three identification methods. However, for whispered speech, the EMD features give the highest accuracy only in the Linear SVM and Weighted K-NN classification methods, the Separate Gabor filtered features show better accuracy for Fine K-NN.

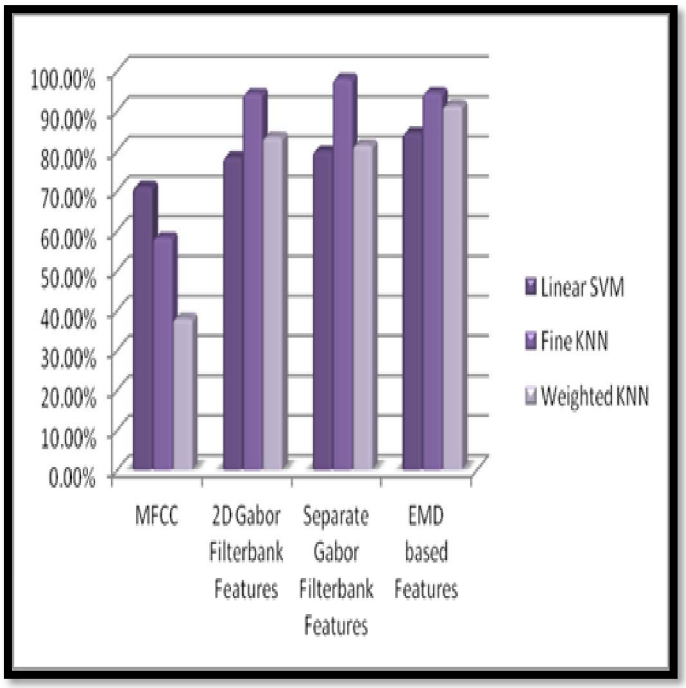

Fig.6 Results Obtained For Whispered Speech

It is also inferred that the AM-FM model and the Separate Gabor Filter banks for spectral and temporal domain has better results for identifying the speaker. However, the results are good in both cases through normal speech and whisper, than the conventional MFCC and 2D Gabor filter approach. On the whole, Fine K-NN yields best results for all the feature extraction techniques except for the MFCC where Linear SVM proved to be better.

V. CONCLUSION & FUTURE SCOPE

This work shows a comparison between the various methods employed for speaker recognition in both normal and whisper mode to have a robust system. The paper presents the feature extraction of signals (normal speech and whisper) through MFCC and different types of filters namely 2D Gabor in spectro-temporal domain, 1D gabor in spectral and temporal domain individually and via Hilbert Huang Transform to obtain IFs. Then the identification was done by various classifiers using SVM and K-NN and it was observed that Fine K-NN outperforms all the other methods in almost all feature extraction models except for MFCC where Linear SVM proves better comparatively. Also, the best feature extraction technique was the AM-FM based IF extraction methodology which acquire a maximum accuracy of 98.90% and 94.60% for normal and whispered data respectively.

In future, the accuracy may be improved by collecting the samples (for normal speech and whisper) in a special recording studio. A future work could include real-time training and speaker identification system which may be useful for an IOT based applications

Act. / Pred.	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16
S1	695	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
S2	0	700	0	3	0	0	0	0	0	1	1	0	2	1	1	0
S3	1	0	698	0	0	0	0	1	0	2	2	0	0	1	0	0
S4	3	0	2	693	0	0	0	0	0	0	0	0	0	0	1	0
S5	0	0	0	4	700	4	1	3	1	0	2	4	0	1	2	0
S6	1	0	0	0	0	696	0	0	1	0	0	2	2	0	3	0
S7	0	0	0	0	0	0	697	0	0	1	0	0	0	0	0	0
S8	0	0	0	0	0	0	0	696	0	0	0	0	0	0	1	0
S9	0	0	0	0	0	0	0	0	696	0	0	0	0	0	3	0
S10	0	0	0	0	0	0	0	0	0	696	2	0	0	0	0	0
S11	0	0	0	0	0	0	2	0	2	0	693	0	2	0	1	0
S12	0	0	0	0	0	0	0	0	0	0	0	694	0	0	1	0
S13	0	0	0	0	0	0	0	0	0	0	0	0	694	0	0	0
S14	0	0	0	0	0	0	0	0	0	0	0	0	0	697	0	0
S15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	186	0
S16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	289

Fig. 7 Confusion matrix for k-NN Classification Normal speech

REFERENCES

[1] Fan, Xing, and John HL Hansen. "Acoustic analysis for speaker identification of whispered speech." *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010.

[2] Xu, Juan, and Heming Zhao. "Speaker identification with whispered speech using unvoiced-consonant phonemes." *Image Analysis and Signal Processing (IASP), 2012 International Conference on*. IEEE, 2012.

[3] Ghaffarzadegan, Shabnam, Hynek Boril, and John HL Hansen. "UT-VOCAL EFFORT II: Analysis and constrained-lexicon recognition of whispered speech." *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014.

[4] Wang, Gang, et al. "On intrinsic mode function." *Advances in Adaptive Data Analysis* 2.03 (2010): 277-293. Dahake, Prajakta P., Kailash Shaw, and P. Malathi. "Speaker d

[5] ependent speech emotion recognition using MFCC and Support Vector Machine." Automatic Control and Dynamic Optimization Techniques (ICACDOT), International Conference on. IEEE, 2016.

[6] Jo, Jihyuck, Hoyoung Yoo, and In-Cheol Park. "Energy-Efficient floating-point MFCC extraction architecture for speech recognition systems." *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 24.2 (2016): 754-758.

[7] Wang, Jia-Ching, et al. "Speaker identification with whispered speech for the access control system." *IEEE Transactions on Automation Science and Engineering* 12.4 (2015): 1191-1199.

[8] Schädler, Marc René, and Birger Kollmeier. "Separable spectro-temporal Gabor filter bank features: Reducing the complexity of robust features for automatic speech recognition." The Journal of the Acou ical Society of America 137.4 (2015): 2047-2059.

[9] Sharma, Rajib, et al. "Empirical mode decomposition for adaptive AM-FM analysis of speech: A review." *Speech Communication* (2017).

[10] Hasan, Md Rashidul, et al. "Speaker identification using mel frequency cepstral coefficients." *variations* 1.4 (2004).

[11] Lawrence Rabiner and Biing-Hwang Juang, Fundamental of Speech Recognition", Prentice-Hall, Englewood Cliffs, N.J., 1993.

[12] Jr., J. D., Hansen, J., and Proakis, J. Discrete-Time Processing of Speech Signals, second ed. IEEE Press, New York, 2000.

[13] Schädler, Marc René, Bernd T. Meyer, and Birger Kollmeier. "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition." *The Journal of the Acoustical Society of America* 131.5 (2012): 4134-4151.

[14] N. E. Huang, 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 nonstationary time series analysis," Proc. R. Soc. London, Series A, vol. 454, pp. 903–995, 1998.

[15] N. E. Huang, S. R. Long, and Z. Shen, "The mechanism for frequency downshift in nonlinear wave evolution," Adv. Appl. Mech., vol. 32, pp. 59–117, 1996.

[16] N. E. Huang, Z. Shen, and S. R. Long, "A new view of nonlinear water waves-the Hilbert spectrum," Ann. Rev. Fluid Mech., vol. 31, pp. 417–457, 1999.

[17] Wang, Gang, et al. "On intrinsic mode function." *Advances in Adaptive Data Analysis* 2.03 (2010): 277-293.

[18] M. W. Mak and S. Y. Kung, "Estimation of elliptical basis function parameters by the EM algorithm with application to speaker verification," IEEE Trans. Neural Netw., vol. 11, no. 4, pp. 961–969, Jul. 2000.

[19] Cummins, Fred, et al. "The chains speech corpus: Characterizing individual speakers." *Proc of SPECOM*. 2006.

[20] De La Torre, Angel, et al. "Histogram equalization of speech representation for robust speech recognition." *IEEE Transactions on Speech and Audio Processing* 13.3 (2005): 355-366.

[21] Schroder, Jens, et al. "On the use of spectro-temporal features for the IEEE AASP challenge 'detection and classification of acoustic scenes and events'." Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on. IEEE, 2013.