Turkish Journal of Computer and Mathematics Education *Vol.12 No.3(2021), 5684-5694*

Research Article

Voice Biometric: A Novel and Realistic Approach

ShashiRanjan[a], Mahesh P K[b]

[a]Assistant Professor, Dept. of ECE, DBIT, Bangalore, India,
[b]Professor and Head, Dept. of ECE, ATME College of Engineering, Mysore, India,
Email:[a]shashiranjanbe@gmail.com, [b]mahesh.k.devalapur@gmail.com

Article History: Received: 10 November 2020; Revised 12 January 2021 Accepted: 27 January 2021; Published online: 5 April 2021

Abstract: Speaker identification uses the basic speaker's wave information to recognize the speaker. The device validates the speaker's identity, which makes the person eligible for the various services the voice can provide. This will fortify every device. Attributed voice is an algorithm focused on a speaker's physiological and behavioral characteristics. Speech analysis provides it with the distinguishing characteristics of identity, allowing the speaker to be distinguished from the others.
Keywords: Attributed Voice, Efficient Decorrelator, Subband Decomposition, W-Transform.

1. Introduction

Over the previous decade, the application of wavelet analysis has been demonstrated to be useful in a variety of situations. Most importantly, in speech function extraction techniques, wavelets have been applied in two ways: Instead of applying a Discrete Cosine Transform to data before doing the wavelet transform, we use the DCT as an efficient decorrelator. W-transform is applied to the speech signal in the second method. In this instance, wavelet coefficients are considered features, however, are variable, so subband energies are used instead in particular, the wavelet transform was introduced in 1995 as a method for computing the spectrum (coherent analysis). During the period spanning 1998 to 2002, speech wavelet packet bases were used to create features of speh intelligibility variations, called mel-frequency division (MF) features were implemented in Mel-filter.

2. Methodology

Generally, the method presented here differs from other research, chiefly due to the approach taken in the pursuit of the most efficient wave packet. Other than using various Mel and wavelet filters, we have also used wavelet and corresponding conjugate mirror filters to achieve a more refined division, these have been used as simple functions for our speech recognition tasks. You will see an in-depth overview of the strategy and clarification of how the defined features relate to alternatives below.

2.1 Subband Decomposition via Wavelet Packets

The basic property of the wavelet transform holds is to be the inner product of the signal x(t) with a set of scaled-and-translated prototypes (t).

$$\psi_{a,b}(t) = \psi\left(\frac{t-b}{a}\right)dt$$

$$W_{\psi}x(a,b) = \frac{1}{\sqrt{a}}\int_{-\infty}^{+\infty} x(t)\psi^{*}\left(\frac{t-b}{a}\right)dt$$

Figure 1.1 depicts the implementation of wavelets, which iterates through a low pass or high pass filter bank followed by a sub sampling of each stage. For the wavelet transform, you have to first iterate through the low pass side and then the high pass, which is called a wavelet packet transform. This results in a tree-structured wavelet filterbank. The effect is a frequency transform that optimally reflects the signal while being capable of re-reproducing the original signal is extracted. It is also known as a wavelet decomposition or subband decomposition, as bands are found by a wavelet filterbank.

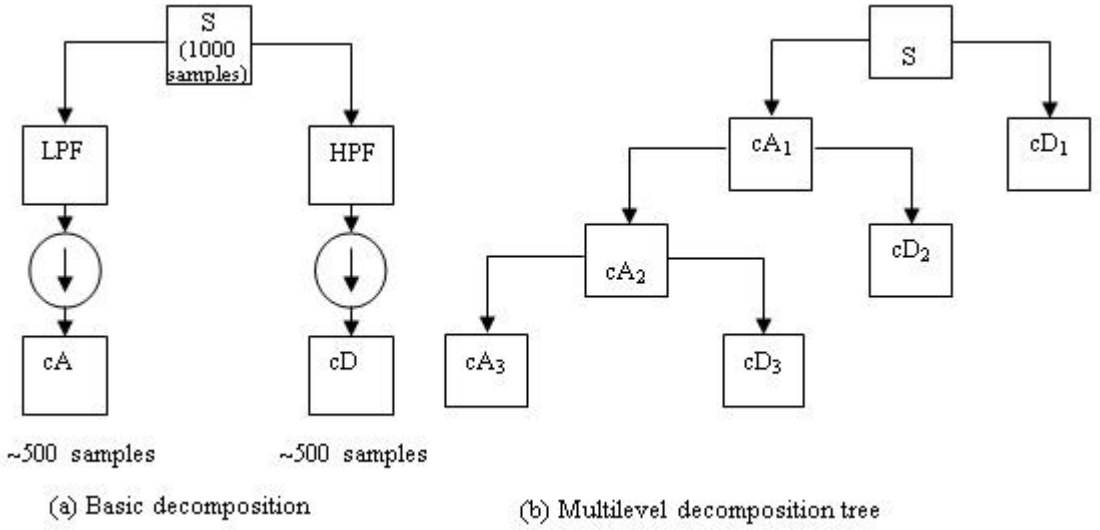


Figure 1.1: The filtering process. a: The first boxes denotes low and high pass filtering of s. ↓ denotes down sampling. b: The low and high pass filtering and down sampling symbols are left out. The approximation and detail coefficients are given indices corresponding to the decomposition level. From [MathWorks, 2004].

2.2 Design of Wavelet Packet Tree

Even though technically it is possible to use a whole-frequency set of wave packets, the amount of time required to analyze the segment imposes limits on the quantity and numerical values that can be used. We deliberately set the maximum allowable speech length to allow for our cars, computers, and telephones because our desires and priorities lie more in the short term rather than the long term. By default, 8 kHz is used for the sampling rate. To derive the Subband features with a frame length of 30ms and a subband interval of 10ms. Finally, the speaker cuts his text using the hamming window and pre-emphasizes it.

The proposed tree modalism attempts to allot more subbands to the lower-mid frequencies, while retaining approximately the same amount in the high subbands. For a given tree, the wavelet packet is calculated, which yields a sequence of subband signals. In effect, each of these subband signals has only limited frequency resolution because of the fact that the inherent bandpass cuts off all but the passband. As the length of the signal increases, the maximum frequency that is measurable by the discrete wavelet packet transform approaches the Nyquist's criterion. Since the maximum wavelet decomposition level is defined as *j=log2(N)*, the maximum resolution is defined as being twice the logarithmic Nyquist Also, the resolution is simple:

$$F_N(1/2)^j = 1/2^{j+1}$$

The Nyquist frequency is anywhere between *FN=1/2*. If *N = 256*, the highest decomposition level of j is 8 and the best possible resolution is 1/512 There are an infinite number of other possible resolutions, which are {1/256, 1/128, 1/64, 1/32, 1/16, 1/8, 1/4}. The resolutions of discrete wavelet packets which can be produced from this sampling frequency are {15.625 Hz, 31.25 Hz, 62.5 Hz, 125 Hz, 250 Hz, 500 Hz, 1000 Hz, 2000 Hz}. While taking into consideration the wide spectrum of frequencies, it must be pointed out that the critical bandwidth is within the range of [0, 4000 Hz] With these experimental findings in mind, it was concluded that:

- Starting from 0 Hz, and going up to 8000 Hz the appropriate discrete wavelet packet transform resolution is half of the critical bandwidth:

Discrete Wavelet Packet Transform resolution=$\frac{CB}{2} Hz \quad for\ f \in [0,8000] Hz$

Let's look at the wavelet tree shown in Figure 1.2. The energy of each subband is measured and transformed.

The energies of the subband signals are calculated for each frame by,

$$S_i = \frac{\sum_{mel}\left[(W_\psi)(i), m\right]}{N_i}$$

W_{ψ} : Wavelet packet transform of signal x,

i :subband frequency index (i=1,2...L),

N_i : number of coefficients in the i^{th} subband.

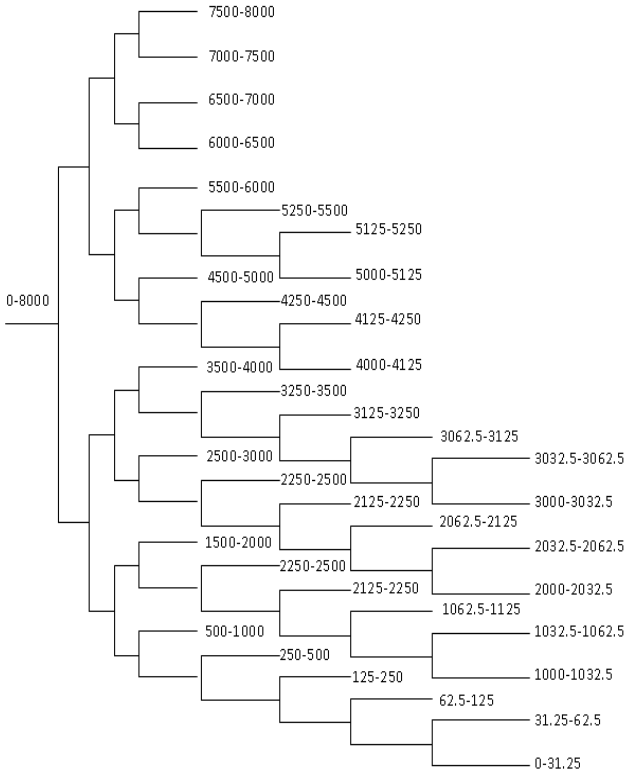


Figure 1.2: Wavelet Packet Tree

2.3 Subband based Cepstral Coefficients (SBC)

It is a common practice in computational analysis to perform the derivation of derived coefficients in two steps. Computation filterbank energies would be handled in the first level, while the decorrelation of the result would be handled in the second For Subband Based Cepstral coefficients, the wave transform is used in place of the short-long Fourier transform to extract the energy levels. these features will be demonstrated to outperform MF This is due to smooth filtering of subband signals for subband frequencies. The results by using the low-pass/high- and high-pass wavelet sub-trees are much smoother due to the equal distribution of time and frequency components. To us, this would lead to better understanding of speech and speaker recognition. To calculate the cepstral coefficients, a Discrete Cosine Transformation is applied to the subband energies.

$$SBC(n) = \sum_{i=1}^{L} \log S_i \cos\left(\frac{n(i-0.5)}{L}\pi\right), n = 1,...n'$$

In the foregoing example, L represents the number of spectral bands and n represents the number of SBC coefficients using the root-ceprel technique, they are referred to as subband coefficients.

In Figure 1.3, we present the computation of the proposed speech features

When two reactions occur, the following must occur:

- In order to minimize level drift, a first-order FIR filter is used to eliminate the signal and only the first-order harmonics of the sampled speech is preserved.
- The pre-emphasis filter, a logarithmic filter with a coefficient of $H(Z) = 1 - \alpha.Z^{-1}$, is used.
- The continuous time speech is broken into 10-millisecond segments (N = speech frames of 256 samples).
- In the process of obtaining a voiced/complete decision, we discover an important attribute. A accurate pitch-estimation algorithm was applied to our experiments, but, as usual, updated. Only frames representing the voices are retained. In other words, therefore, more background music is required.

- Wavelet is added to the articulated segments of the speech waveform The invention of this discrete wavelet packet transform gives rise to a total of B=36 frequency subbands, of which four can be discarded from previous studies. Additionally, we are thinking about how to make the system more stable.
- The next step is to find the total amount of energy in each frequency band, and then to divide it by the number of coefficients present. The detail is as explained above: The energies of the sub-band are calculated for each frame one.

$$E_p = \frac{\sum_{i=1}^{N/2^j} (Wkj(i))^2}{N/2^j}, \; Wkj \in S_1, \; p = 1, \ldots\ldots, B \qquad \circ$$

Where $W_j^k(i)$ is the *i*- th coefficient of the discrete wavelet packet transform vectorW_j^k.

- The decorrelation of the subband energies is done using a logarithmally compressed signal, and the Discrete Cosine Transformation is applied: DCT is applied on the logarithmically subband coefficients:

$$F(i) = \sum_{p=1}^{B} log_{10}(E_p).cos\left(\frac{i.(p-1/2)}{B}\right), \qquad i = 1, \ldots\ldots, r$$

The computation is based on a predefined set of *r* parameters. The SBC consists of all 36 coefficients.

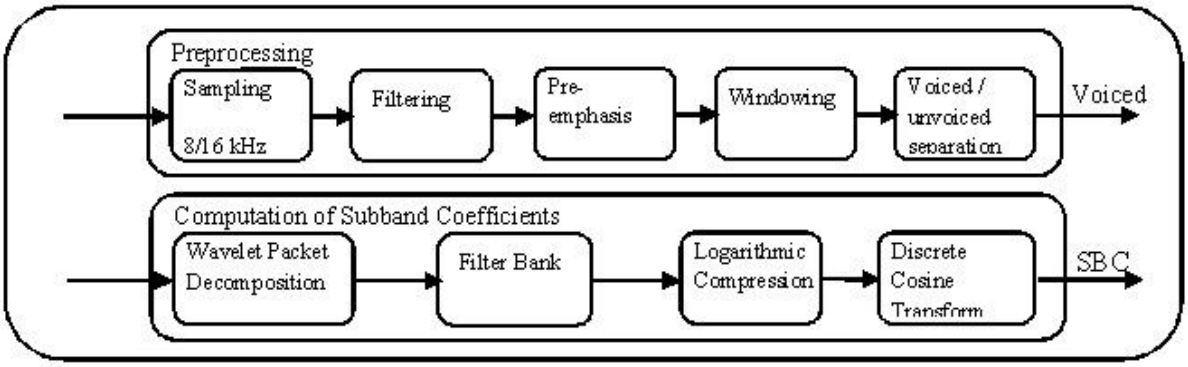


Figure 1.3: Block diagram of the speech pre-processing and the estimation of the proposed Wavelet Packets-based speech features

3.Classifications and Feature Matching

Our next task is to complete is to create a specific model for and speaker. The speaker whose name has been matched against the input will be linked to all templates in the database to make sure that his identity is not a pseudonym. Once the speaker has been found, an algorithm will be applied to that which defines their identity.

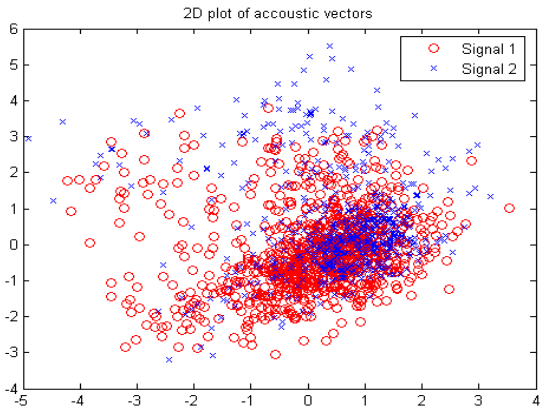


Figure 1.4 (a) Acoustic vector of same speaker

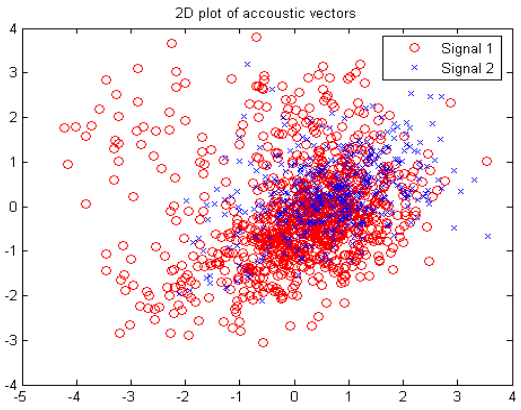
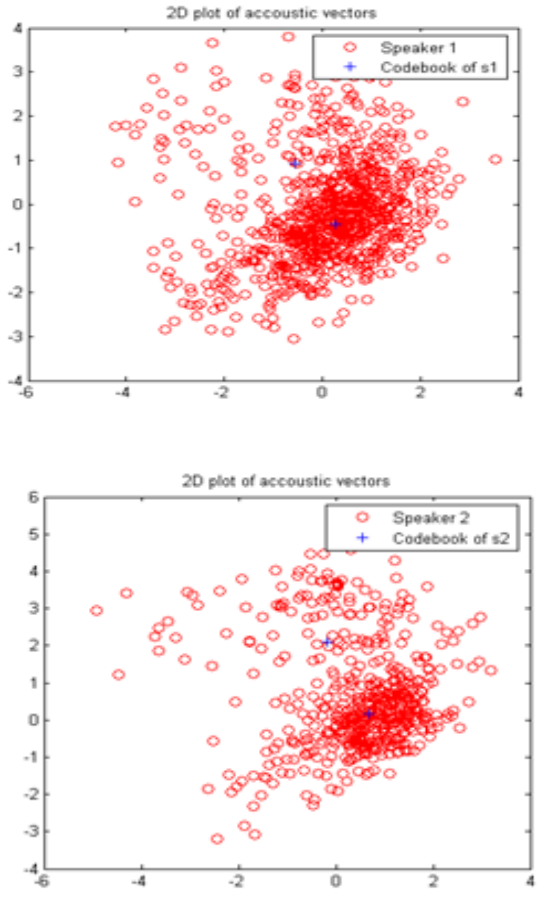
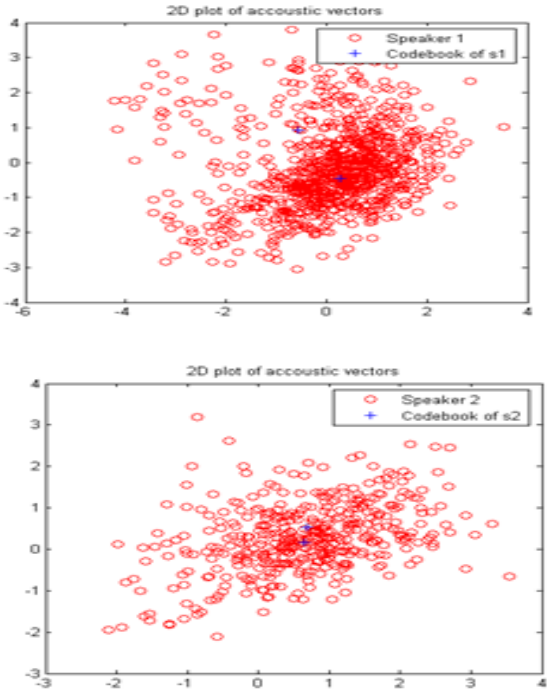


Figure 1.4 (b) Acoustic vector of different speaker



a. Acoustic Vector and Codebook of same Speaker with same Speech signal



b. Acoustic Vector and Codebook of different Speaker of same Speech signal

Figure 1.5: Acoustic Vector and Codebook

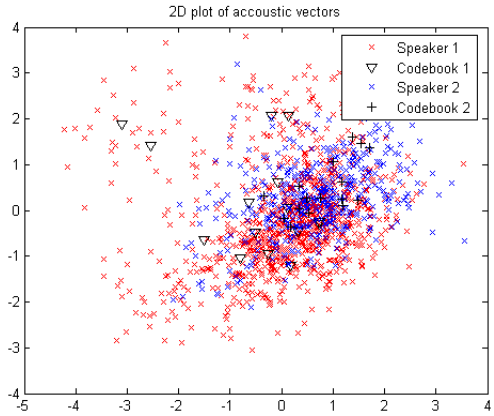


Figure 1.6: Acoustic vectors and Codebook of same speaker

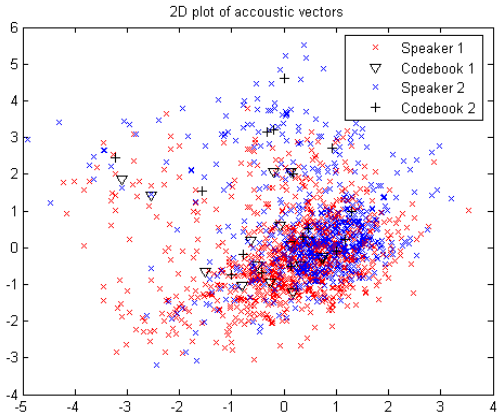


Figure 1.7: Acoustic vectors and Codebook of different Speaker

3.1 Gaussian Mixture Model (GMM)

Speaker identification is an example of an unstandardized process. Once feature vectors are classified, they seem to approximate a Gaussian distribution. It implies that each cluster is a Gaussian distributed, and features belonging to the clusters are on Gaussian scales. One problem is to be found in efficient function classification. First, it was developed because of the study of voice modal analysis, in which it was determined that voice modal frequency levels are essential for voice recognition [9].

They are:-

i. When each acoustic class is considered as a single individual model, then Gaussian is a set of different classes. Acoustic techniques illustrate vocal apparatus detail

ii. When the input to a multi-dimensional feature space Gaussian distribution gets randomized, a more normal and smoothly varying output result is generated.

Figure 1.8 gives a better understanding of what GMM really is:-

The mathematical form of an *m* component Gaussian mixture for D dimensional input vectors is,

$$P(x|M)=\sum_{i=1}^{m} a_i \frac{1}{(2\pi)^{\frac{D}{2}}\left|\sum_i\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu_i)^T \sum_i{}^{-1}(x-\mu_i)\right)$$

Where $P(x|M)$ is the probability of happening The mixture model has one non-modal density function, which takes the means of the parameters and their corresponding covariance matrices, respectively. the coefficients (or variables) may only have a positive range and are limited to amount to one are known as the “coefficients,” problem solving method, an iterative Expectation- Maximization (EM) problem can be solved using the maximum likelihood (ML) method. EM algorithm would generally need fewer than ten iterations to achieve adequate parameter convergence

The complete Gaussian mixture PDF is represented by the mean vector, covariance matrices and mixture

weights of all the component densities [9]. These parameters are collectively represented by the notation

$$\lambda=\{a_i,\mu_i,\Sigma_i\} \quad with \quad i=1,.....,M,$$

where λ is the GMM for each speaker

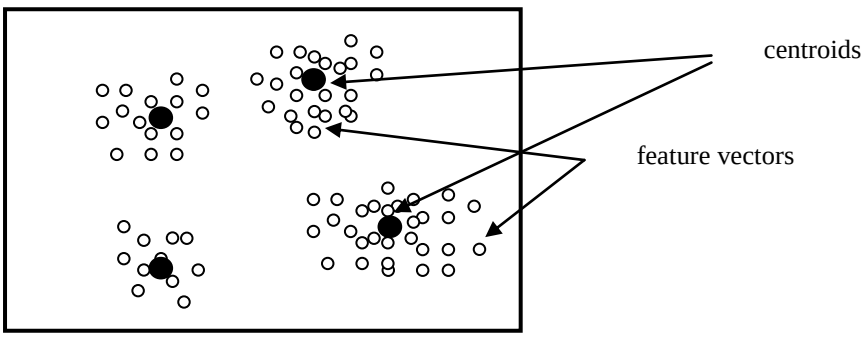


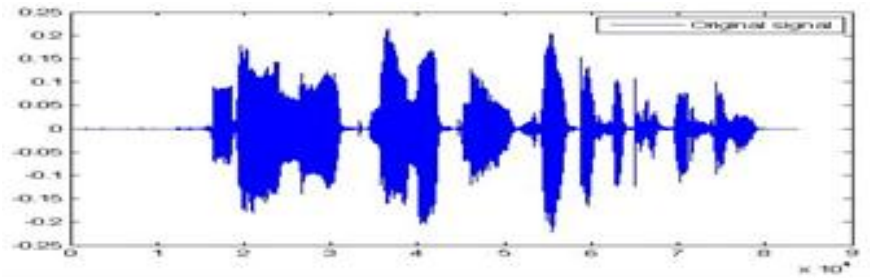
Figure 1.8: GMM model showing a feature space and corresponding Gaussian model

4. Experimental Investigations

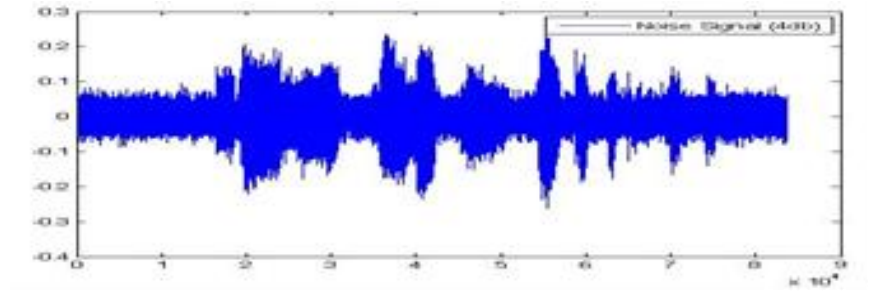
The suggested approaches have been improved by the inclusion of noise on the recording. An additive white noise of 4 dB and 8 dB is added to make it seem as if it were the sound of an environment with an SNR of 40 dB. Additionally, we've acquired three different speech files in three different scenarios.

Two new databases are developed, one for 350 distinct speech topics and one for subjects with six distinct speech attributes. The features make it stand out from the rest of the general commercial databases.

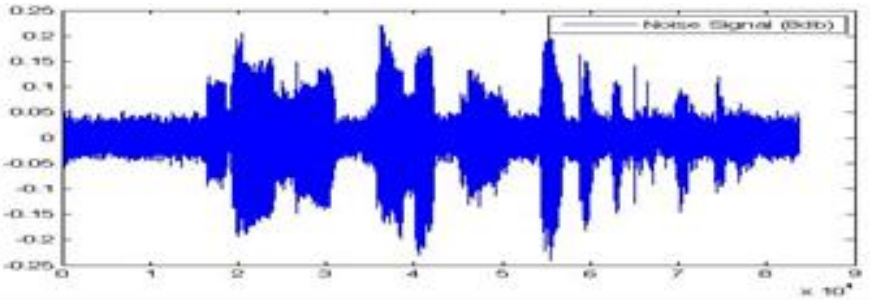
Thus, while creating the database, we have introduced noises on only test samples of speech. Thus, to evaluate methods for speech signal, we train all of all methods on good data, which are then tested on recognition system that include all types of noisy instances to ensure that they can remain clean. As can be seen in Figure 1.9, we have done the analysis with three separate data sets as well



(a)

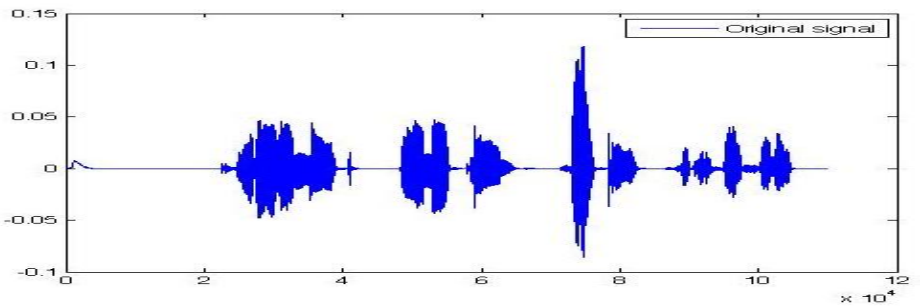


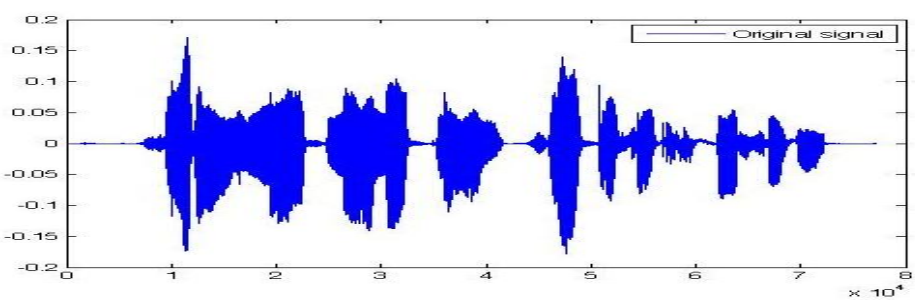
(b)



(c)

Figure 1.9: (a) Clean speech (b) Noisy speech (4 dB) (c) Noisy speech (8 dB)





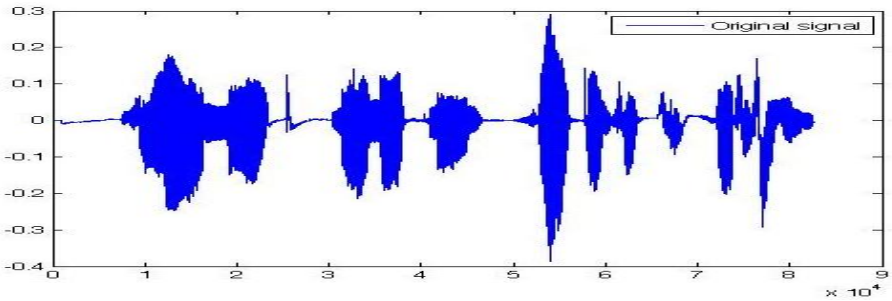


Figure 1.10: Raw data in three different condition of the same speaker

5.Result And Conclusion

5.1 Results Of Speaker Recognition On Clean Database

The results of the popularly used function extraction process, SBC (Calculating energy for 36 bands and considering the cepstral coefficients) are shown in Figure 1.11. The approach is contrasted with MFCC, LPCC, with different filter banks, and wavelet packets with differentials in regard to conventional approaches. Indicated in blue in Figure 1.12, is the efficiency of the proposed solution (indicated in blue color) is lower (indicated in red color). When clean data is emphasized, the new approach can be regarded as better.

5.2 Results of speaker recognition on Degraded Database

Since the proposed approach has been extended on an exhausted database, the robustness of this method has been demonstrated. The findings of the new approach (shown in Figure 1.12) are contrasted with the LPCC (Figure 1.12) methods like 4-to-methods as well as conventional methods such as the additive method (with added noise of 5db). Table 1.1 shows our evaluation of three separate tests done on each speaker, all under their own environmental environments, which allows us to compare their speaking rates across conditions. However, presence of background noise impacts the speech performer's ability to perform.

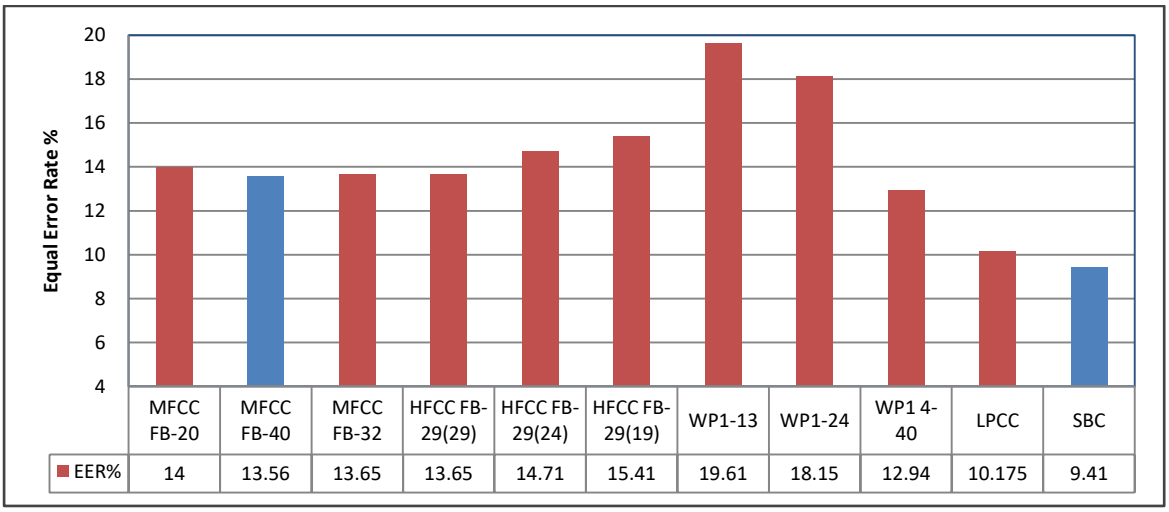


Figure 1.11 Comparison of EER for various methods for speaker recognition modality on clean biometric data

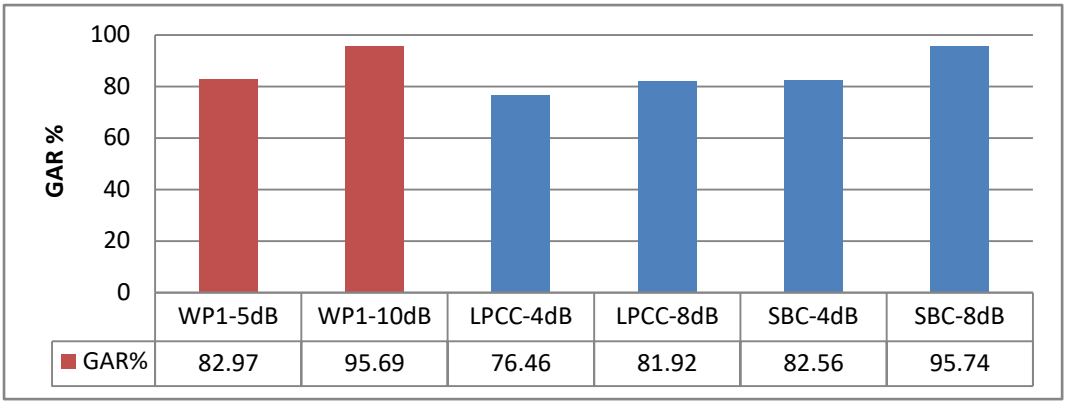


Figure 1.12 GAR comparisons for noisy database for speaker recognition

Table 1.1: Identification rate of speaker recognition in three different conditions

Modality	Method		GAR%
Speech Signal	MFCC+GMM	Condition 1	56.67
		Condition 2	87.48
		Condition 3	91.63
	SBC+GMM	Condition 1	63.34
		Condition 2	89.95
		Condition 3	93.37

References

1) Tufekci, Z., Gowdy, J.N. "Feature extraction using discrete wavelet for speech recognition". In Proceedings of the IEEE SoutheastCon 2000, Nashville, Tennessee, USA.
2) Long J.S., Datta S. "Wavelet based feature extraction for phoneme recognition". Proceedings of the ICSLP-96, Philadelphia, USA. Vol. 1, pp. 264-267
3) Sarikaya R., Hansen H.L. "High resolution speech feature parameterization for monophone-based stressed speech recognition, In IEEE Signal Processing Letters. Vol. 7, No. 7, pp. 182-285
4) Erzin E., Cetin A.E., Yardimci Y. "Subband analysis for speech recognition in the presence of car noise". In Proceedings of the ICASSP-95, Detroit, MI, USA. Vol. 1, pp. 417-420
5) Sarikaya R., Pellom B.L., Hansen H.L. "Wavelet packet transform features with application to speaker identification". In Proceedings of the IEEE Nordic Signal Processing Symposium:(NORSIG'98), Visgo, Denmark. pp. 81-84
6) Sarikaya R., Hansen H.L. "High resolution speech feature parameterization for monophone-based stressed speech recognition, In IEEE Signal Processing Letters. Vol. 7, No. 7, pp. 182-285
7) Farooq O., Datta S., "Mel-scaled wavelet filter based features for noisy unvoiced phoneme recognition". In Proceeedings of ICSLP 2002, Denver, Colorado, USA. pp. 1017-1020
8) P. Alexandre and P. Lockwood, "Root cepstral analysis: A unified view: Application to speech processing in car noise environments", Speech Communication, v.12, pp. 277-288,1993.
9) Douglas A. Reynolds, Richard C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", IEEE Transactions on Speech and Audio Processing, VOL. 3, No. 1, January 1995

10) L. Hong and A. Jain, "Integrating faces and fingerprints for personal identification", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, pp. 1295-1307, 1998.