Published Online 4-05-2021

A COMPREHENSIVE FRAMEWORK FOR VOICE BIOMETRICS

Shashi Ranjan[1], Mahesh P K[2]
[1]*Assistant Professor, Dept of ECE, DBIT, Bangalore, India,*
[2]*Professor and Head, Dept of ECE, ATME College of Engineering, Mysore, India,*
[1]*shashiranjanbe@gmail.com,* [2]*mahesh.k.devalapur@gmail.com*

Abstract: Recognizing someone by his/her voice is called Automatic Speaker Recognition (ASR). Biometric protection comes under Speaker ID biometric is concerned with individual characteristics Voiceprint or biometric authentication is used to identify a person. It depends on an individual's voice activity or physiology. Behavioral biometrics includes speech, signature, keystroke, and typing while physiological biometrics includes face, retina, fingerprints, and DNA now biometric authentication is cutting-edge.Several forms of speech recognition and feature extraction systems were employed in this paper. A novel feature extraction method was proposed using Sub-band based SBC, and energy calculation was also covered. For a specific combination of 350 users, our approach achieves recognition rate above 94% with EER less than 6%. The output is measured on two separate sets of data, clean and degraded.

***Keywords*: ASR, Feature vector , GMM, MFCC, Packet tree Wavelet**

1. Introduction

Voice biometrics is the method where the speaker's voicewave information may be added to the mission. It may be possible to classify and authenticate the speaker by the voice alone An additional protective shield is now covers will be attached to any sensitive unit. A computerized speech-recognitioning system depends on the physical characteristics of the speaker's voice. Though available as a standard add-on, voice biometrics offers advanced abilities such as speaker identification and authentication.

Speech recognition should be distinct from speech detection; where the objective is to understand what is being said, speech, speech recognition is to identify the speaker The voice section provides information about how to find speakers from voice cues. preprocessing is covered in Section 2 Section 3 provides a detailed walkthrough of the feature extraction process. Section 4 explains the Identification and Matching algorithms are presented below. It is to be found in Section 5.

2. Preprocessing

Figure 1.1 indicates the pre-phasing and post-phasing methods. in this research, we evaluate speech at 8 kilohertz It has a wide dynamic range and resists both white and additive noise. In order to reduce the sensitivity range, focus is applied after truncation. This smoother has a first-order low-pass filter applied to it.In the time domain, with input (*x[n]*) and *0.9 ≤ (a) ≤ 1.0*, the equation is:

$y[n] = x[n]{-}a\ .\ x[n{-}1]$ 1.1

And the transfer function of the FIR filter in z-domain is:

$H\ (\ Z\ ) = 1 - \alpha.Z^{-1},\ 0.9 \leq \alpha \leq 1.0$ 1.2

Where alpha is the improvement parameter.

Pre-emphasis is added as a fixed value or as a derivative from the auto-correlation.

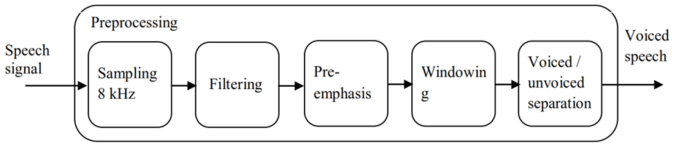


Figure 1.1 Preprocessing steps

The next stage is likely to produce more energy in the ultrashort frequency range. Furthermore, this is what is caused by glottal fricatives. tuning the timbre of a voice helps increase the artificial neural network's accuracy pre-emphasis filter prior to windowing decreases the emphasis of the input signal At thirty milliseconds, the filter technique is applied to consider the ideological position. The process described in Figure 1.2.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

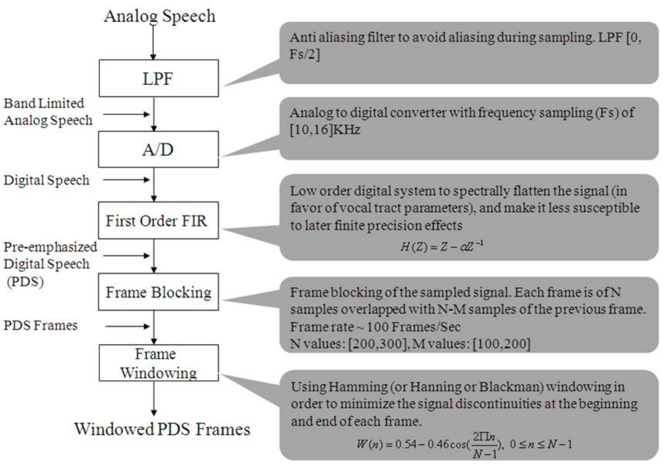

Analog Speech
LPF
Band Limited Analog Speech
A/D
Digital Speech
First Order FIR
Pre-emphasized Digital Speech (PDS)
Frame Blocking
PDS Frames
Frame Windowing
Windowed PDS Frames
Anti aliasing filter to avoid aliasing during sampling. LPF [0, Fs/2]
Analog to digital converter with frequency sampling (Fs) of [10,16]KHz
Low order digital system to spectrally flatten the signal (in favor of vocal tract parameters), and make it less susceptible to later finite precision effects
$H(Z) = Z - \alpha Z^{-1}$
Frame blocking of the sampled signal. Each frame is of N samples overlapped with N-M samples of the previous frame. Frame rate ~ 100 Frames/Sec
N values: [200,300], M values: [100,200]
Using Hamming (or Hanning or Blackman) windowing in order to minimize the signal discontinuities at the beginning and end of each frame.
$W(n) = 0.54 - 0.46\cos(\frac{2\Pi n}{N-1}),\ 0 \le n \le N-1$

Figure 1.2 Flowchart of pre-processing steps

When Figure 1.3 is selected, it will show the details in the numbers at the bottom of the image. The next move in Figure 1.4 will enable you to provide it. Figure 1.5 is a 256 x 64 bit image representing the frequency and logarithmic power spectrums.

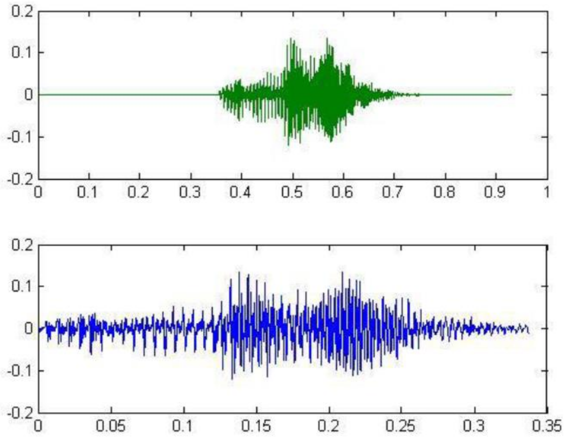


Figure 1.3 Truncated version of original signal

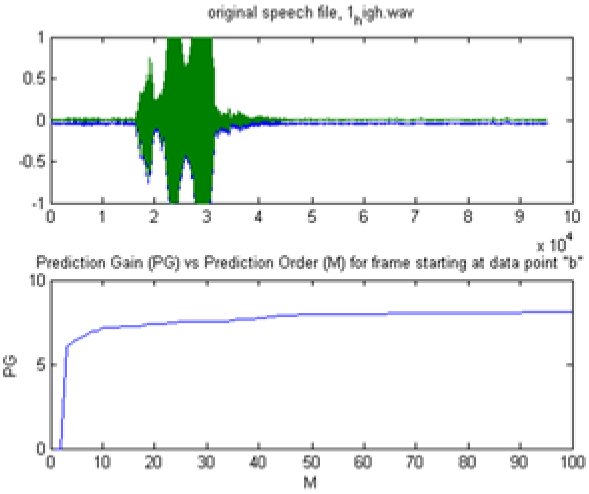


Figure 1.4 Prediction Gain versus Prediction Order for frame starting at data point ‘b’

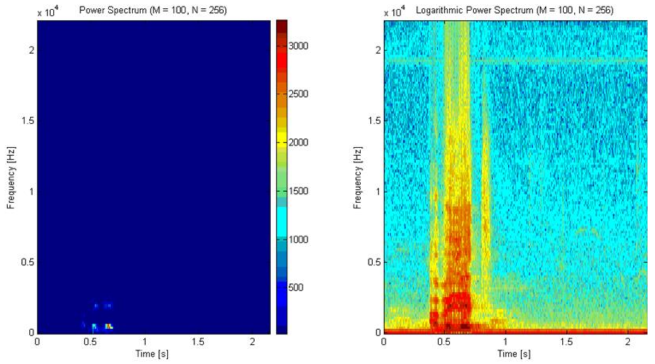


Figure 1.5 Power spectrum and Logarithmic Power Spectrum when M=100 and N=256

In this step, we need to decompose the speech into N-frame subphrases. As a consequence, the first sample is made up of three distinct components. The second ripple begins at N, moves on to M-N and goes back to the first frame. As you can see, the signal in 1.6 represents the short feed network.

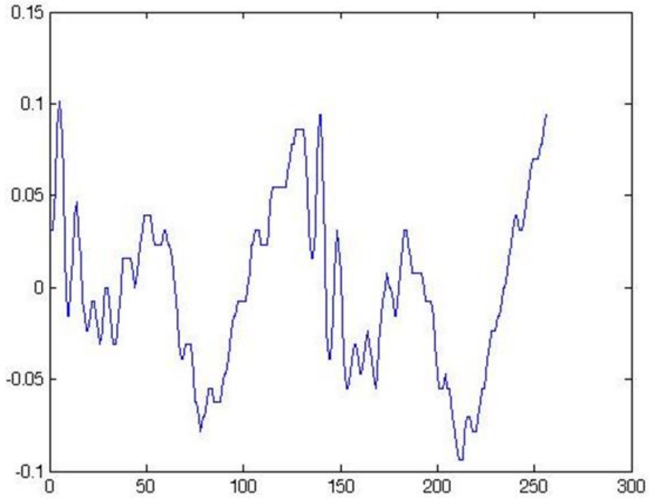


Figure 1.6 Frame output of truncated signal

Fourier transform can be applied to each pulse to create a moving image This figure shows a truncated sine wave with a window size of 400 samples, as in Figure 1.7(a). This graph is the product of the FFT analysis of the given example. Figure 1.7 (b) shows how the power spectrum changes if a Hamming window reduces the data.

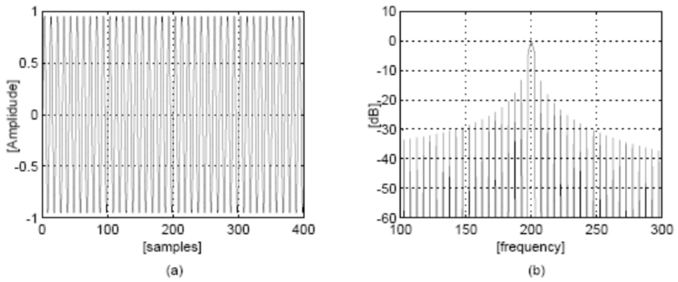


Figure 1.7 a) A 200 Hz sine wave, truncated by a rectangular window. b) FFT of the sine wave

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Hamming window function, defined by Equations (1.3) and (1.4), can be well described by plotting it on 256 lines.

$$x[n] = 0.54 - 0.46cos\left(\frac{2\pi n}{N-1}\right)$$ 1.3

$$W_{Hamming}[n] = \begin{cases} x[n] & N.r \le n < N(r+1), \quad r = 0,1,2,3,\ldots\ldots, M-1 \\ 0 & otherwise \end{cases}$$ 1.4

3. Feature Extraction

An important aspect of speech processing component is to be able to extract function vectors from the speech signal Small-unit sampling (or the function selection) is a technique of gathering small piece units of expression. For effective speech recognition, a waveform should have all of these properties. To represent the speech signals in such a way that the problem can be solved with the features.

The algorithm most commonly used in this research is MFCC, with modifications.

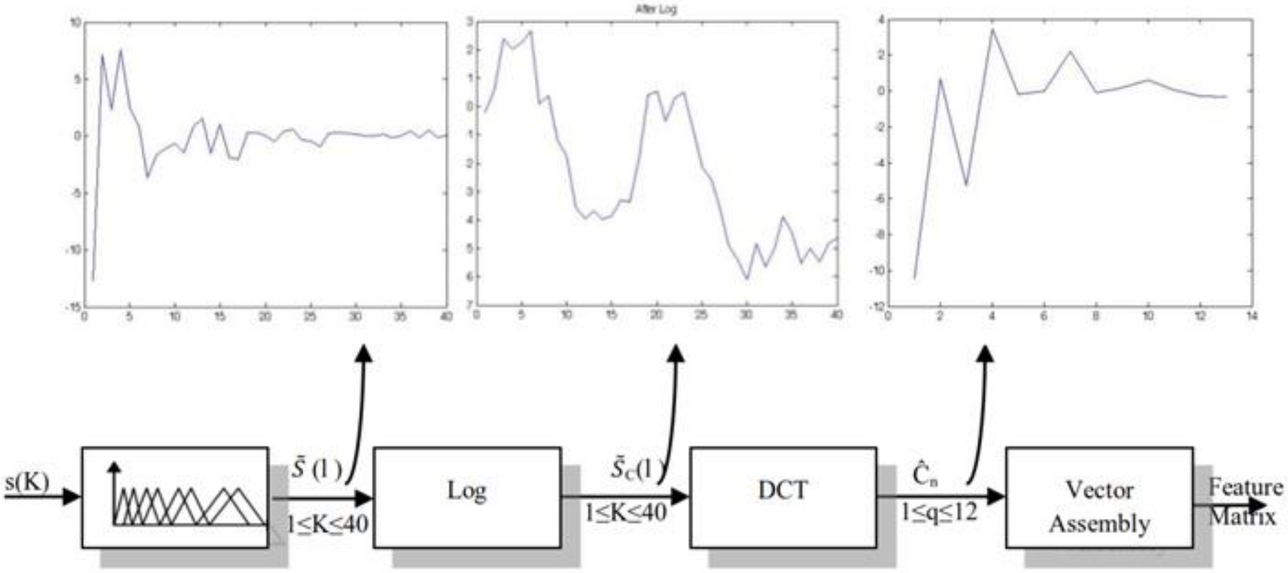


Figure 1.8. The steps in the frequency domain involved in calculating the final feature vectors and the steps in the mel-cepstrum domain involved in calculating the final feature vectors.

3.1 Mel-Frequency Cepstral Coefficients (MFCC)

Mel-Frequency Cepstral Coefficients provide some details on the strength of voice, and offer a basis against which it can be measured. couplings are computed from the FT (voice-speaker recognition) with a logarithmic equation MFCC based spectral analysis' The Coefficient can process many data.

Even when calculating the Mel Coefficients, the results are strikingly the same. the Mel scale has been focused on hearing acuity and hearing the tone preference'Mels' weight'

The mel-scale is approximated by

$$Mel(f) = 2595\log\left(1+\frac{f}{700}\right)$$ 1.5

where $Mel\ (f)$: the frequency in mels

f : the input frequency in Hertz

With selected filters of Mel decomposition bank, it processed with about the same efficiency as human ear reaction.

$$\tilde{S}(l) = \sum_{k=0}^{N/2} S(k)M_l(k)$$ 1.6

Where :

$\tilde{S}$(l) :Mel spectrum.

S(K) :Original spectrum.

M(K) :Mel filterbank.

L=0 ,1 , ..., L-1 , Where L is the total number of Mel filter banks

N/2 = Half FFT size.

The MFCCs may be calculated using following equation:

$$\tilde{C}_n = \sum_{k=1}^{K}(\log\tilde{S}_k)\left[n\left(k-\frac{1}{2}\right)\frac{\pi}{K}\right] \qquad where\ n = 1,2,\ldots\ldots,K$$ 1.7

The system will use 40 linear filters and 40 logarithmic filters 12 MFCCs from each speaker Each computational step of the procedure can be seen in Figure 1.8.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Wavelet analysis is shown to be successful in signal processing. In speech recognition, feature analysis has been applied twice with wavelets for the first, we use the wavelet transform instead of the DCT The signal is processed using an algorithm. Concerning this case, either the wavelet coefficients with high energy are used or the subband energies are used.

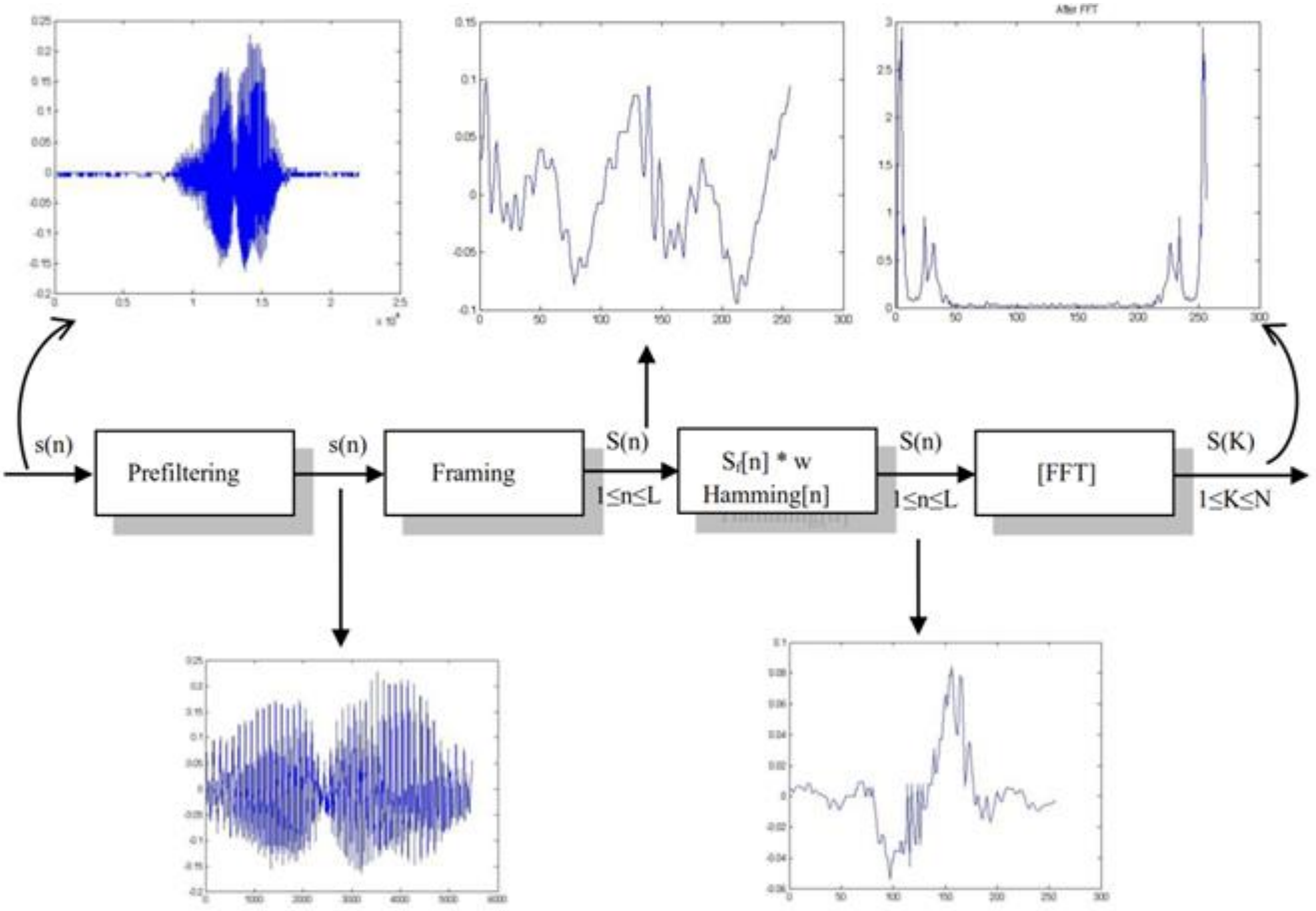


Figure 1.9 The steps in the time domain involved in calculating the final feature vectors. The input signal is an utterance of the English word "nine" sampled at a frequency of 8000 Hz.

The Wavelet Transform is the internal product of the x(t) signal with a set of square base wavelets in which copies of the prototype wavelet are scaled and interpreted by the base wavelets $\psi(t)$.

$$\psi_{a,b}(t) = \psi\left(\frac{t-b}{a}\right)dt \qquad 1.8$$

$$W_{\psi}x(a,b) = \frac{1}{\sqrt{a}}\int_{-\infty}^{+\infty} x(t)\psi^*\left(\frac{t-b}{a}\right)dt \qquad 1.9$$

3.2 Development of the Packet Tree Wavelet

speech words have a 20- to 26-millisecond timing precision The number of cycles was quantified in this equation. For Subband extraction of the features, a Cepstralfeatures of 30ms is used. In the previous and concluding sections of the voice, the screen is further punctuated with bold text to reinforce the important points. When the signal is interpreted using the full wavelet transform, the highest resolution occurs.The maximum resolution of the DWPT is bound by the maximum decomposition stage for a speech frame of j sample sets, with a lower limit of $j=\log_2(N)$. The most suitable plan of action is:. Consequently, the safest resolution is:

$$F_N(1/2)^j = 1/2^{j+1} \qquad 1.10$$

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

Wavelet trees is apparent in the figure in Figure 1.10. The energies of the subband are multiplied by the number of transformation coefficients.The subband signal energies are computed for each frame as,

$$S_i = \frac{\sum_{mel}\left[(W_{\psi})(i), m\right]}{N_i} \qquad 1.11$$

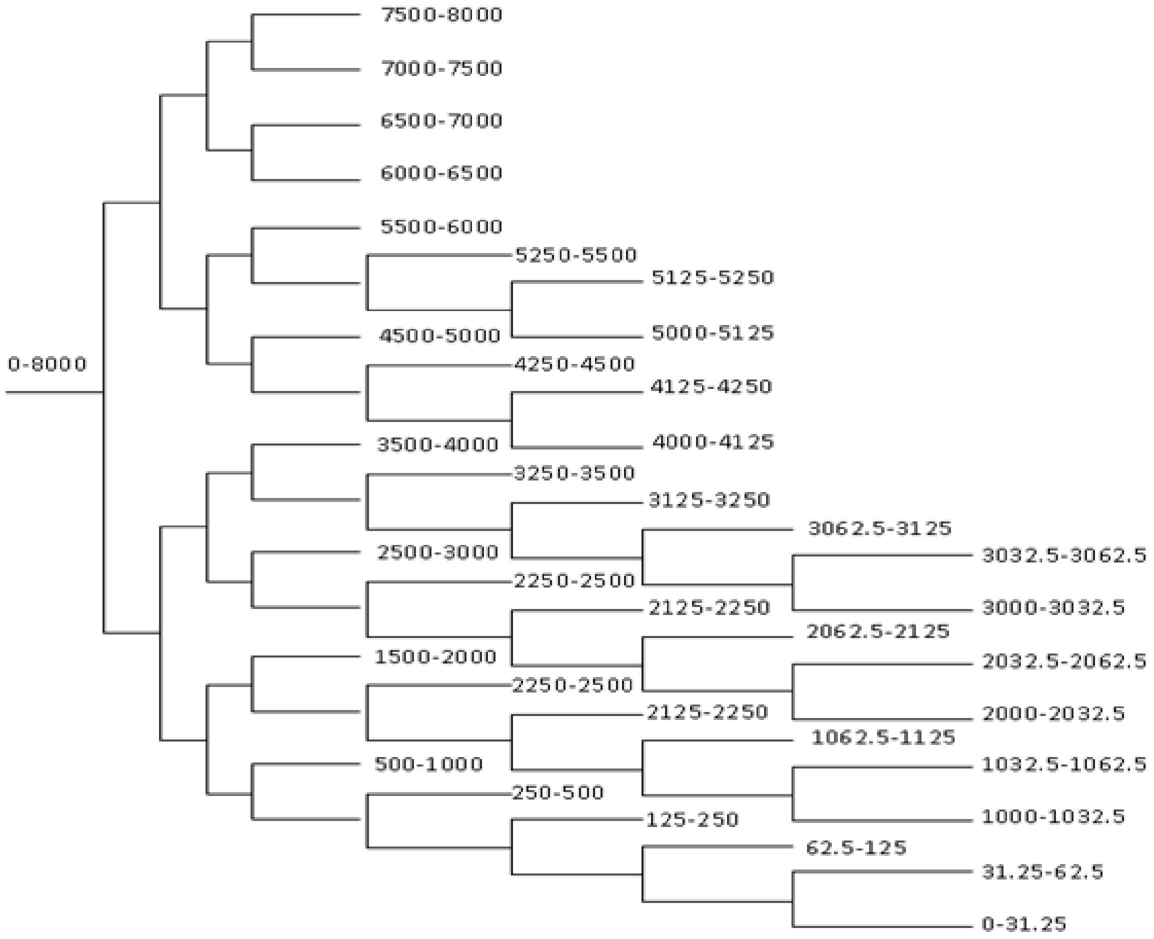


Figure 1.10 Wavelet Packet Tree

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.

The subband associated Cepstral coefficients are obtained by applying the Discrete Cosine Transformation from subband energies

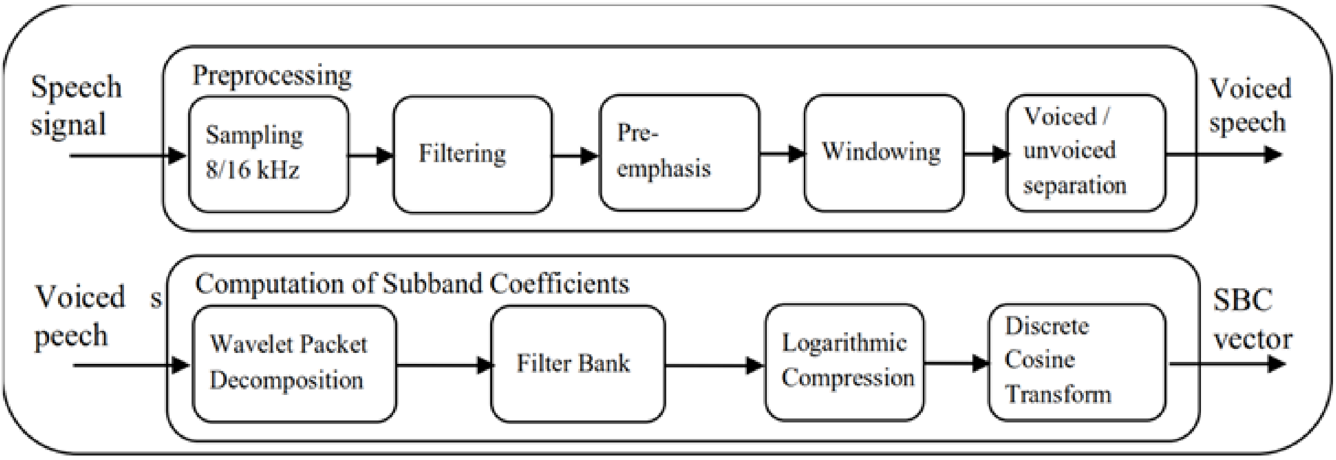


Figure 1.11 Block diagram of the speech pre-processing and the estimation of the proposed Wavelet Packets-based speech features

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

$$SBC(n)=\sum_{i=1}^{L}\log S_i \cos\left(\frac{n(i-0.5)}{L}\pi\right), n=1,\dots n' \quad 1.12$$

In Figure 1.11, the calculation of the suggested speech characteristics, SBC, is shown

4. Gaussian Mixture Model (GMM)

The method we use is highly non-parametric because it doesn't depend on model parameters. Creative phrase:plotting: When Feature Vectors are seen after the clustering, they look like the Gaussian distribution in several ways.

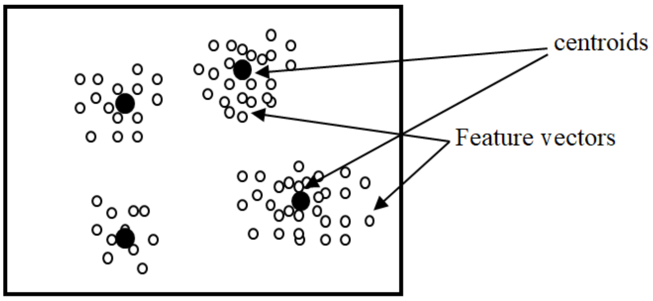


Figure 1.12 GMM model showing a feature space and corresponding Gaussian model

Two details inspire the use of Gaussian mixing density for the detection of speakers[182]. They are:-

i. The individual Gaussian samples represent different acoustic classifications Acoustic groups serve as models of the vocal tract.

ii. The gaussian mixture density does a better job than the polynomial spacing for approximating the multi-dimensional vectors

Figure 1.12 gives a better understanding of what GMM really is:-

The mathematical shape of a Gaussian component for dimensional input vectors is,

$$P(x|M)=\sum_{i=1}^{m} a_i \frac{1}{(2\pi)^{\frac{D}{2}}\left|\sum_i\right|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu_i)^T \sum_i^{-1}(x-\mu_i)\right) \quad 1.13$$

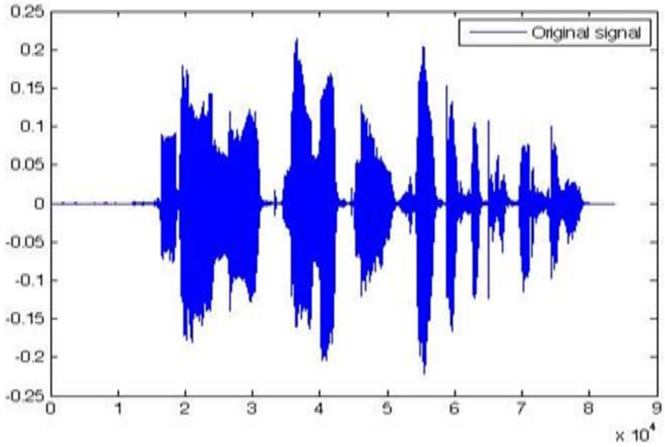
Where $P(x|M)$ is the possibility of the mixture model, The Gaussian PDF is given by the mean and covariance matrices These parameters are summarized by the notation

$$\lambda=\{a_i,\mu_i,\Sigma_i\} \quad with \quad i=1,......,M, \quad 1.14$$

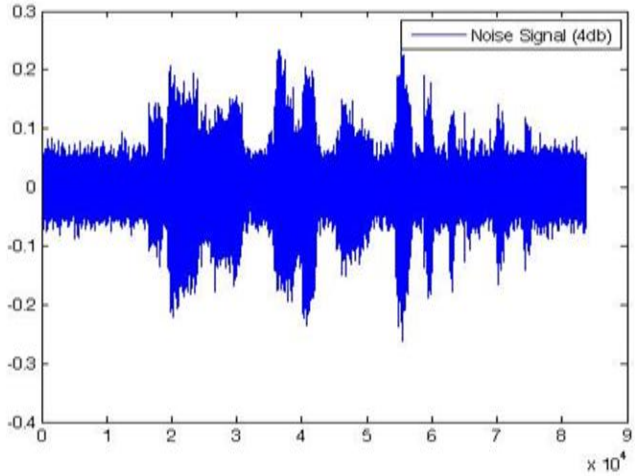
where λ is the GMM for each speaker

5. Experimental Results

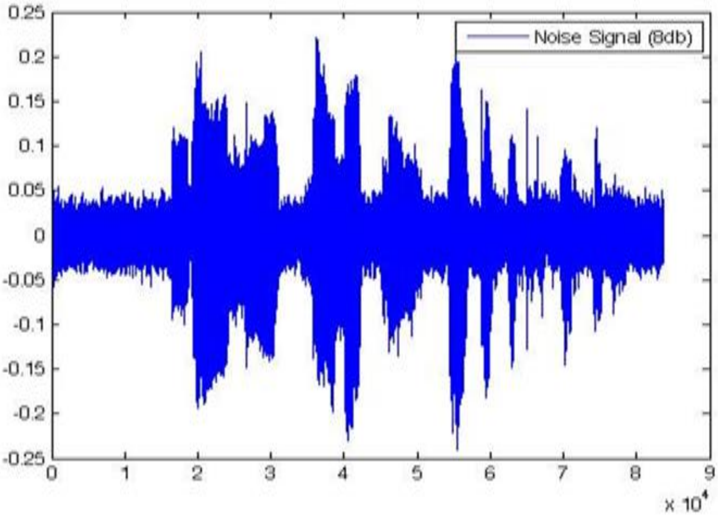
For the Voice modality, we add noise to the dataset. An additive white noise is applied along with two separate SNR of 4 dB and 8dB sampled for three conditions to make the speech appear to be at normal rate. The samples shown in Figure 1.13 are all noisy. While building the database, we've applied noises only to speech samples. For these methods, we test all the methods on both clean and noisy examples to emulate real-world conditions. We also imported the speech data in Figure 1.14 into three separate states, as shown in the figure above.



(a)



(b)



(c)

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Figure 1.13 (a) Clean speech. (b) Noisy speech (4 dB). (c) Noisy speech (8 dB)

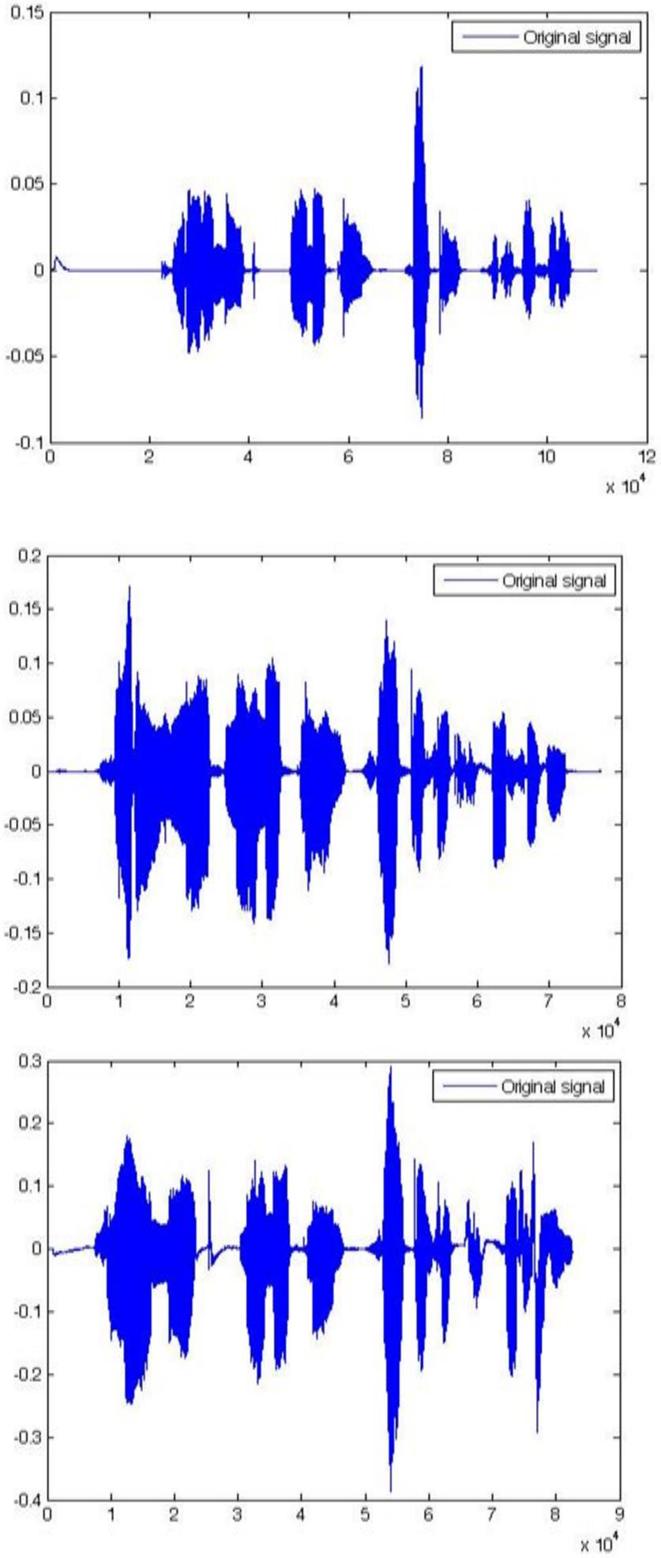


Figure 1.14 Raw data in three different condition of the same speaker

When we compare Figures 1.15 and 1.16, we see that while the form varies between the two speakers, the phoneme used differs also.

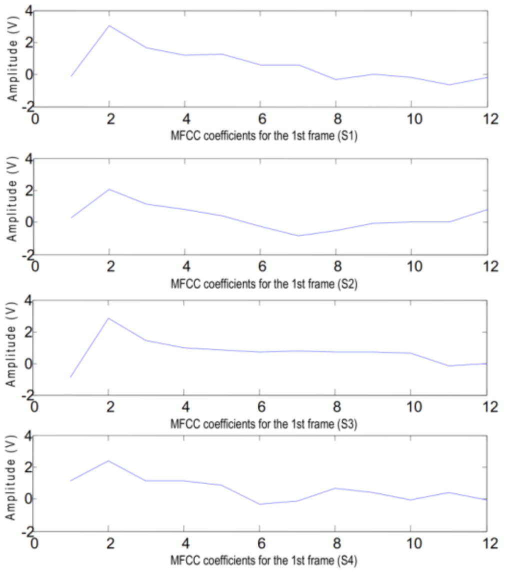


Figure 1.15 Four different utterances for the same speaker considering 12 MFCC coefficients

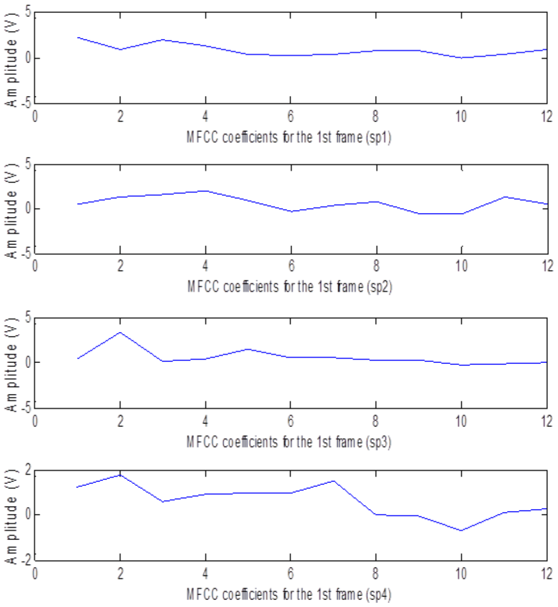


Figure 1.16The same utterance by four different speakers considering 12 MFCC coefficients

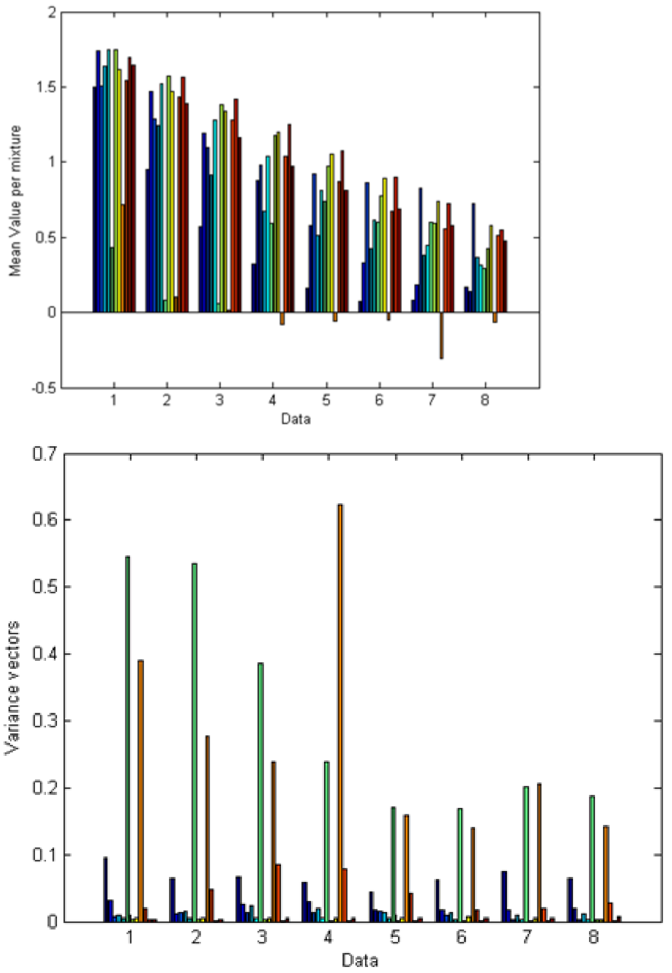
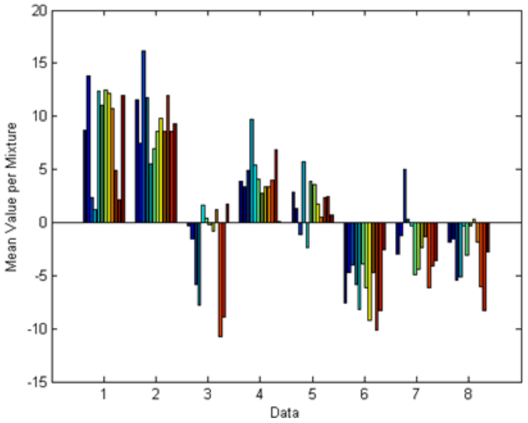


Figure 1.17 (a) Mean vectors per Gaussian mixture of MFCC (b) Variance vectors per Gaussian mixture of MFCC

The assumptions made when defining a GMM assume that the feature vector space is divided unique components dependent on feature vector clustering and Gaussian feature vectors, Thesefigures display the results after running GMM on a mixture of MFCC and SBC features.



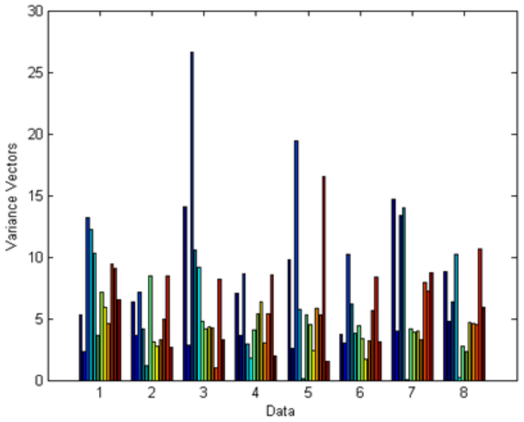
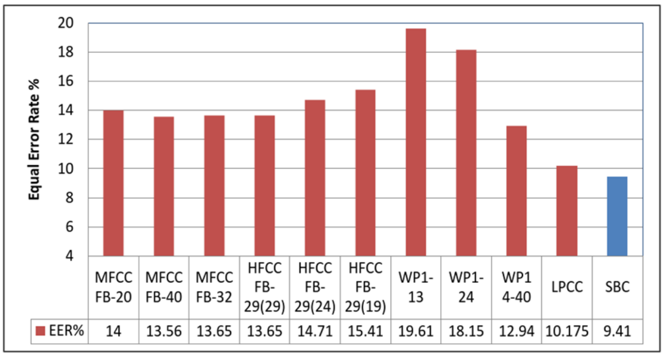


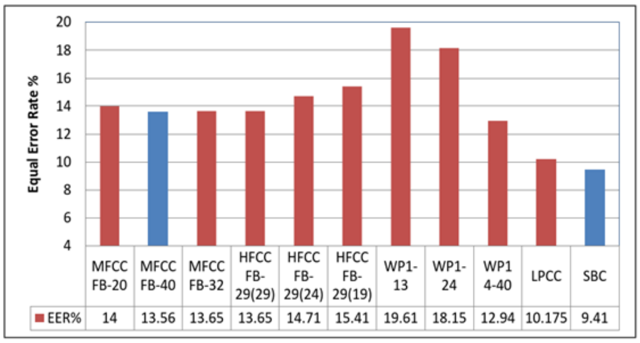
Figure 1.18 (a) Mean vectors per Gaussian mixture of SBC (b) Variance vectors per Gaussian mixture of SBC



	MFCC FB-20	MFCC FB-40	MFCC FB-32	HFCC FB-29(29)	HFCC FB-29(24)	HFCC FB-29(19)	WP1-13	WP1-24	WP1 4-40	LPCC	SBC
EER%	14	13.56	13.65	13.65	14.71	15.41	19.61	18.15	12.94	10.175	9.41

Figure 1.19 Genuine Acceptance Rate (GAR) of traditional and proposed approaches

The proposed method is contrasted with conventional methods, including MFCC, LPCC, and HFCC with different filter banks.



	MFCC FB-20	MFCC FB-40	MFCC FB-32	HFCC FB-29(29)	HFCC FB-29(24)	HFCC FB-29(19)	WP1-13	WP1-24	WP1 4-40	LPCC	SBC
EER%	14	13.56	13.65	13.65	14.71	15.41	19.61	18.15	12.94	10.175	9.41

Figure 1.20 Comparison of EER for various methods for speaker recognition modality on clean biometric data.

To prove the validity of the proposed solution, the degraded database is employed. The results of the proposed method (added noise of 4dB and 8dB) are compared to methods like Wavelet (added noise of 5dB and 10dB) and LPCC (which display better performance in Figure 1.21).

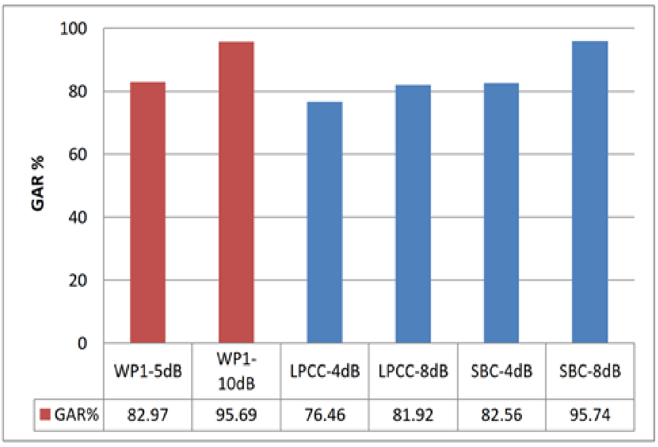


Figure 1.21 GAR comparison for noisy database for speaker recognition.

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

6. Conclusion:

We proposed a novel feature extraction technique that includes Sub-using the known subband energy as well as the widely popular36Extract Spectral Band Scores (SC), and we have further discussed which featuresets which are currently in use (MFCC). Some spread-spectrum signal modulation techniques are employed for MFCC (as well as SBC). With these results, we now in hand, we can more specifically identify the speaker characteristics required for new ELSD thus making it an integral component of speaker recognition.

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..

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731