Performance Evaluation of Wiener Filter and Kalman Filter Combined with Spectral Subtraction in Speaker Verification System

Utpal Bhattacharjeeand Pranab Das

Abstract*—This paper investigates the performance of speaker verification system in mobile environment and the techniques used to improve the robustness of the verification system. The paper demonstrates by corrupting the speech signal with additive white Gaussian noise in simulated environment. A comparative study of the three front-end noise reduction techniques namely spectral subtraction, Wiener filter and Kalman filter have been made independently as well as combining spectral subtraction with other two methods alternatively and their performances have been evaluated for the clean speech as well as contaminated speech with different level of white Gaussian noise. It has been observed that spectral subtraction plays an important role in reduction low power Gaussian noise whereas Kalman filter is efficient in reduction noise when noise power is high. Wiener filter improves the performance at all level of noise. No considerable performance improvement has been observed when spectral subtraction is combined with other two methods.

Keywords*— Wiener filter, Kalman filter, Spectral Subtraction, Speaker Verification.

I. INTRODUCTION

The major challenge in implementing speaker verification system in mobile environment is due to the mobility of the device itself, which results in highly variable acoustical environment where the device operates. Other than the white Gaussian noise, many other type of noises like harmonic noise, interfering speech, impulsive noise do have an effect on the performance of the speaker verification system. In each environment, variation in the acoustic condition and background noises corrupt the speech signal leading to intra-speaker variability. Further, people speak in a noisy environment altering their style of speaking as an attempt to improve intelligibility (Lombard effect). This factor induces further intra-speaker variability. The variability of the microphone used in the mobile handset can also have a substantial impact on performance of the speaker verification system. The most prominent factor affecting the verification system is environmental noise. Recently, much research has been conducted with a focus on improving the performance of the system in different environment conditions through filtering techniques such as spectral subtraction, Wiener filtering, Kalman filtering assuming a prior knowledge of the noise spectrum. In this paper we investigate three techniques of speech enhancement namely Spectral Subtraction, Kalman filter and Wiener filter individually and also combining spectral subtraction method with the other two methods independently.

Manuscript received on January, 2013.

Utpal Bhattacharjee, Department of computer science and Engineering, Rajiv Gandhi University, Rono Hills, Doimukh, Arunachal Pradesh, India

Pranab Das, Department of Computer Science & IT, Don Bosco College of Engineering and Technology, Guwahati, Assam, India.

The paper is organized as follows: section II describes the proposed noise enhancement scheme. The baseline Speaker Verification System has been described in section III. In section IV we describe the database used in the present study. Experimental setup and result obtained are presented in section V. The paper is concluded in section VI.

II. NOISE ENHANCEMENT TECHNIQUES

The development of proposed front end processing is based on the idea that each of the noise removal techniques has its own advantages and disadvantages in presence of environmental noise. So the idea is to build a system by combining the advantages of any two noise removal techniques and analyze their performance. Here we investigate all the three methods and then combined spectral subtraction with Wiener filter and Kalman filter as the techniques of noise reduction. The graphical representation of the proposed method is given in Fig.1.

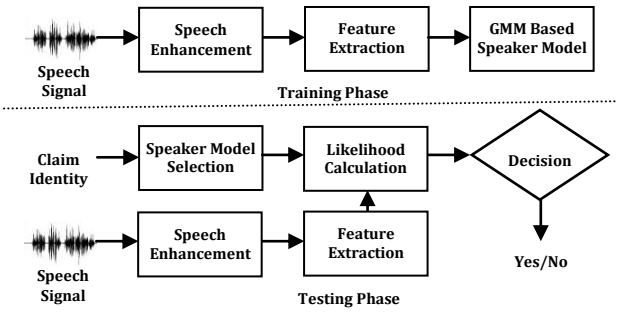

Fig.1: Speaker Verification System with speech enhancement

A. Spectral Subtraction

It is based on the principal that enhanced speech can be obtained by subtracting the estimated spectral components of the noise from the spectrum of the input noisy signal. So that average signal-to-noise ratio (SNR) is improved. It is assumed that the signal is distorted by a wide-band, stationary, additive noise d(m) to the speech signal x(m), the noisy speech y(m) can be written as,

$$y(m) = x(m) + d(m) \qquad \text{--- (1)}$$

Windowing and applying Fourier transform to both the sides result in

$$Y_w(e^{jw}) = X_w(e^{jw}) + D_w(e^{jw}) \qquad \text{--- (2)}$$

Multiplying both sides by their complex conjugates we get:

$$|Y(e^{jw})|^2 = |X(e^{jw})|^2 + |D(e^{jw})|^2 + 2|X(e^{jw})||D(e^{jw})|cos\emptyset$$

--- (3)

where Φ is the phase difference between speech and noise:

Taking the expected value of both sides we get:

$$E\{|Y(e^{jw})|^2\} = E\{|X(e^{jw})|^2\} + E\{|D(e^{jw})|^2\} + E\{2|X(e^{jw})||D(e^{jw})|cos\emptyset\}$$

$$= E\{|X(e^{jw})|^2\} + E\{|D(e^{jw})|^2\} + 2E\{|X(e^{jw})|\}E\{|D(e^{jw})|\}E\{cos\emptyset\}$$

--- (4)

in power spectral subtraction it is assumed that $E\{\cos(\emptyset)\} = 0$, hence

$$E\{|Y(e^{jw})|^2\} = E\{|X(e^{jw})|^2\} + E\{|D(e^{jw})|^2\}$$

--- (5)

$$|X(e^{jw})|^2 = |Y(e^{jw})|^2 - E\{|D(e^{jw})|^2\}$$ --- (6)

The power spectrum of noise is estimated during speech inactive periods and subtracted from the power spectrum of each frame resulting in the power spectrum of the speech. Generally Spectral subtraction is suitable for stationary or very slow varying noises in magnitude spectral subtraction it is assumed that$\{\cos(\emptyset)\} = 1$, hence:

$$E\{|Y(e^{jw})|^2\} = E\{|X(e^{jw})|^2\} + E\{|D(e^{jw})|^2\} + 2E\{|X(e^{jw})|\}E\{|D(e^{jw})|\}$$

$$= \left(E\{|X(e^{jw})|\} + E\{|D(e^{jw})|\}\right)^2$$ --- (7)

$$E\{|Y(e^{jw})|\} = E\{|X(e^{jw})|\} + E\{|D(e^{jw})|\}$$ --- (8)

$$|X(e^{jw})| = |Y(e^{jw})| - E\{|D(e^{jw})|\}$$ --- (9)

The magnitude spectrum of the noise is estimated during speech inactive periods and, again, assuming that the variations of noise spectrum are tolerable, the magnitude spectrum of speech is estimated by subtracting the average spectrum of noise from each frame[1].

B. Wiener Filter

The Wiener filter is a popular statistical approach based on the assumption that signal and the noise are stationary linear stochastic processes with known spectral characteristics that has been used for noise reduction in speech signal. Assuming that the clean speech, s(t), degraded by ad additive noise, w(t). The noisy speech, x(t) is defined as [2]

$$x(t) = s(t) + w(t)$$ --- (10)

Wiener filter is an optimal filter that minimize the Mean Squared Error (MSE). In case of Eq.(10), the filter can be defined as

$$S(\omega) = H(\omega).X(\omega)$$ --- (11)

Where ω is the frequency index and S(ω), X(ω) and H(ω) are the discrete Fourier transform of clean speech, noisy speech and that of the Wiener filter respectively. The MSE is defined as follows. The error is defined as:

$$E(\omega) = S(\omega) - \hat{s}(\omega) = S(\omega) - H(\omega).X(\omega)$$ -- (12)

The Mean Squared Error of Eq.(10) is defined as:

$$E[|E(\omega)|^2] = E[|S(\omega) - H(\omega).X(\omega)|^2]$$ --- (13)

Where E[.] stands for expectation operator. To minimize the MSE, the Wiener filter can be estimated

$$\frac{\delta E[|E(\omega)|^2]}{\delta H(\omega)} = 2H(\omega)E[|X(\omega)|^2] - 2E[|X(\omega)S^*(\omega)|] = 2H(\omega)P_{XX}(\omega) - 2P_{XS}(\omega) = 0$$ --- (14)

Where $P_{XX}(\omega)$, $P_{XS}(\omega)$ are the power spectra of noisy speech and cross power spectra between noisy speech and clean speech respectively.

If there is no correlation between the speech signal *s(t)* and additive noise *w(t)*, the power spectrum of the noisy speech and the cross power spectrum can be transformed as:

$$P_{XX}(\omega) = E[|X(\omega)|^2] = E[|S(\omega) + W(\omega)|^2] = E[|S(\omega)|^2] + E[|W(\omega)|^2] + E[|2S(\omega)W(\omega)|] = E[|S(\omega)|^2] + E[|W(\omega)|^2] = P_{SS}(\omega) + P_{WW}(\omega)$$ --- (15)

$$P_{XS}(\omega) = E[X(\omega)S^*(\omega)] = E[(S(\omega) + W(\omega))S^*(\omega)] = E[|S(\omega)|^2] = P_{SS}(\omega)$$ --- (16)

Consequently, the Wiener filter can be derived as follows:

$$H(\omega) = \frac{P_{SS}(\omega)}{P_{SS}(\omega)+P_{WW}(\omega)}$$ --- (17)

The SNR is defined by [3]

$$SNR = \frac{P_{ss}(\omega)}{P_{WW}(\omega)}$$ --- (18)

The definition can be incorporated to the Wiener filter equation as follows:

$$H(\omega) = \left[1 + \frac{1}{SNR}\right]^{-1}$$ -- (19)

C. Kalman Filter

Kalman filter is an adaptive Least Square Error (LSE) filter that provides an efficient computational recursive solution for estimating a signal in presence of Gaussian noises. It is an algorithm which makes optimal use of imprecise data on a liner (or nearly linear) system with Gaussian error to continuously update the best estimate of system's current state. Kalman filter theory is based on a state-space approach in which a state equation models the dynamics of the signal generation process and an observation equation models the noisy and distorted observation signal.

A finite dimensional linear filter is expressed as follows[4,5]:

$$x_{k+1} = F_k x_k + G_k w_k$$ --- (20)

$$y_k = H_k x_k + v_k$$ --- (21)

Where $\{w_k\}$ and $\{v_k\}$are independent zero-mean Gaussian white noises with covariance matrix $\{\Sigma_{wk}\}$ and $\{\Sigma_{vk}\}$. The Eq.(20) is called state equation and Eq.(21) is called observation equation. The parameters $\{F_k\}$, $\{G_k\}$ and $\{H_k\}$ are transition matrix, input matrix and output matrix respectively. It the parameters are known, the problem to estimate $\hat{x}_{k|k}$and $\hat{x}_{k|k-1}$in a sense of minimum variance when observed data $\{y_0, y_1 \dots \dots \dots y_k\}$ was given, is called Kalman filter problem and the solving algorithm is called

Kalman filter. Here $\hat{x}_{k|k}$ is called the estimated value of x_k at time k and $\hat{x}_{k|k-1}$ is called prediction value of x_k at time k-1. Kalman filter algorithm is given below:

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k(y_k - H_k\hat{x}_{k|k-1} \quad \text{--- (22)}$$

where K_k is Kalman gain matrix.

$$\hat{x}_{k+1|k} = F_k\hat{x}_{k|k} \quad \text{--- (23)}$$

$$K_k = \hat{\Sigma}_{k|k-1}H_k^T\left[H_k\hat{\Sigma}_{k|k-1}H_k^T + \Sigma_{vk}\right]^{-1} \quad \text{--- (24)}$$

$$\hat{\Sigma}_{k|k} = \hat{\Sigma}_{k|k-1} - K_kH_k\hat{\Sigma}_{k|k-1} \quad \text{--- (25)}$$

$$\hat{\Sigma}_{k+1|k} = F_k\hat{\Sigma}_{k|k}F_k^T + G_k\Sigma_{wk}G_k^T \quad \text{--- (26)}$$

The block diagram of Kalman filter is shown in Fig-2.

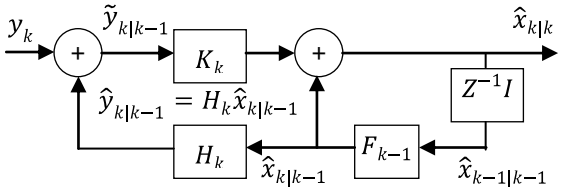

Fig 2. Block diagram of Kalman filter

III. BASELINE SPEAKER VERIFICATION SYSTEM

A baseline speaker verification system has been developed using Gaussian Mixture Model with Universal Background model (GMM-UBM) based modeling approach. Mel-frequency cepstral coefficient (MFCC)[6] has been used as feature vector in the present study. A 38-dimensional feature vector was used, made up of 19 mel-frequency cepstral coefficient (MFCC) and their first order derivatives. The first order derivatives were approximated over three samples. The coefficients were extracted from a speech sampled at 16 KHz with 16 bits/sample resolution. A pre-emphasis filter H(z)=1-0.96z^{-1} has been applied before framing. The pre-emphasized speech signal is segmented into frame of 20 microseconds with frame frequency 100 Hz. Each frame is multiplied by a Hamming window. From the windowed frame, FFT has been computed and the magnitude spectrum is filtered with a bank of 20 triangular filters spaced on Mel-scale and constrained into a frequency band of 300-3400 Hz. The log-compressed filter outputs are converted to cepstral coefficients by DCT. The 0^{th} cepstral coefficient is not used in the cepstral feature vector since it corresponds to the energy of the whole frame [7], and only 19 MFCC coefficients have been used. To capture the time varying nature of the speech signal, the first order derivative of the Cepstral coefficients are also calculated. Combining the MFCC coefficients with its first order derivative, we get a 38-dimensional feature vector. Cepstral mean subtraction has been applied on all features to reduce the effect of channel mismatch.

The Gaussian mixture model with 1024 Gaussian components has been used for both the UBM and speaker model. The UBM was created by training the speaker model with speaker's data with Expectation Maximization (EM) algorithm and finding the average of all these models [8]. The speaker models were created by adapting only the mean parameters of the UBM using maximum a posteriori (MAP) approach with the speaker specific data.

The detection error trade-off (DTE) curve has been plotted using log likelihood ratio between the claimed model and the UBM and the equal error rate (EER) obtained from the DTE curve has been used as a measure for the performance of the speaker verification system.

IV. SPEAKER RECOGNITION DATABASE

To carry out the experiments, a speaker verification database was developed and all the testing and evaluation of the speaker recognition system has been done using that database. The database consists of 30 speakers, 17 male and 13 female. Each speaker participates in 4 recording session and each session is of 5~7 minutes duration. The speakers were recorded for conversation style of speaking. The data is recorded in parallel across two devices – a table mounts microphone and a handset microphone. Recording has been done in typical office/home environment. The recording specification has been given below:

TABLE 1: Recording Specification

Number of Speakers	30 (17 male, 13 female)
Number of sessions	4 for each language.
Intersession interval	2 weeks
Data types	Speech
Type of Speech	Conversation sentences
Sampling rate	16 KHz
Sampling format	Mono-channel, 16 bits resolution
Application	Text-independent (multilingual) ASV system
Speech Duration	5~7 minutes per speaker
Languages	Assamese
Training segments	180s
Testing segments	15s, 30s, 45s
Microphones	Fixed mounted and Mobile Handset microphone
Acoustic environment	Typical Office/Home
File Format	WAV PCM

V. EXPERIMENTAL SETUP

All the experiments reported in this paper are carried out using the database described in section IV. Silence intervals from the input speech are removed based on an envelope threshold. The input signal is up-sampled, segmented to remove samples that fall below a threshold, and then re-sampled back to the original sampling rate, and filtered to smooth out the discontinuities prior to feature extraction. Only data from the handset microphone has been considered in the present study. The two available sessions were considered for the experiments. Each speaker model was trained using one complete session. The training set consists of speech data of length 180 seconds per speaker. The test set consists of speech data of length 15 seconds, 30 seconds and 45 seconds. Each speaker model is tested against 11 speakers of which one is the actual speaker and rest 10 are the imposters[9].

The test segments are now contaminated with additive Gaussian noise of SNR 5dB, 10dB 15dB and 20dB. The experiments were carried out separately for each noise level. The results of the experiments are shown in table.2(a) and table. 2(b).

TABLE 2(a): Equal Error rate for Speaker verification System at different SNR Level

SNR	BS	SS	WF	KF	WF+SS	KF+SS
Clean Speech	9.09	8.70	8.69	8.70	8.7	8.7
SNR 20 dB	16.67	10.78	10.64	13.50	11.42	11.42
SNR 15 dB	22.22	21.74	16.67	21.74	19.67	17.39
SNR 10 dB	25	25.00	19.15	24.48	20.83	20.83
SNR 5 dB	33.67	33.33	29.79	27.78	25	27.78

*BS-Baseline System; SS-Spectral Subtraction ; WF-Wiener Filter; KF- Kalman Filter; WF+SS: Wiener Filter + Spectral Subtraction; KF+SS: Kalman Filter + Spectral Subtraction.

TABLE 2: Recognition accuracy for Speaker verification System at different SNR Level

SNR	BS	SS	WF	KF	WF+SS	KF+SS
Clean Speech	90.01	91.30	90.31	91.30	91.30	91.30
SNR 20 dB	83.33	89.22	89.36	86.50	88.58	88.58
SNR 15dB	77.78	78.26	85.33	78.26	80.33	82.61
SNR 10 dB	75.00	75.00	80.85	75.52	79.17	79.17
SNR 5 dB	66.33	66.67	70.21	72.22	75	72.22

*BS-Baseline System; SS-Spectral Subtraction ; WF-Wiener Filter; KF- Kalman Filter; WF+SS: Wiener Filter + Spectral Subtraction; KF+SS: Kalman Filter + Spectral Subtraction The DET curve of the experiments are given in Fig 3(a)~(f).

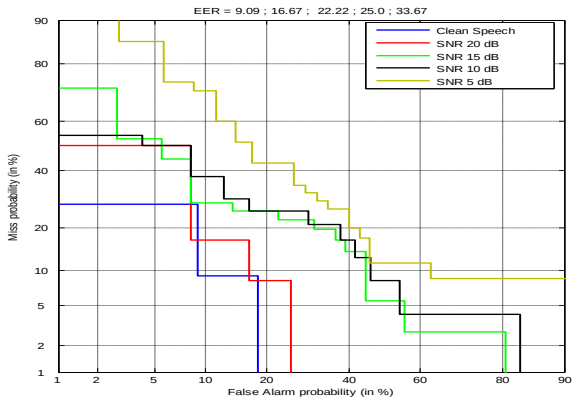

Fig 3(a): DET curve for baseline Speaker Verification System for different level of noise

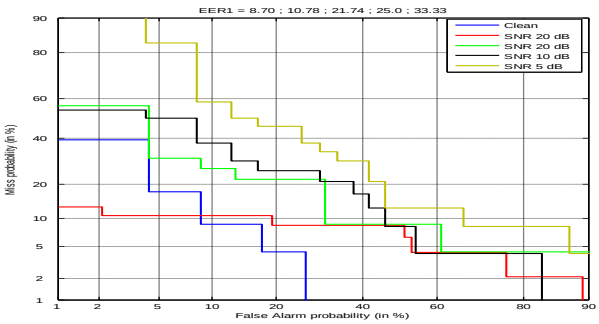

Fig 3(b): DET curve for baseline Speaker Verification System with input enhancement by Spectral Subtraction method for different level of noise

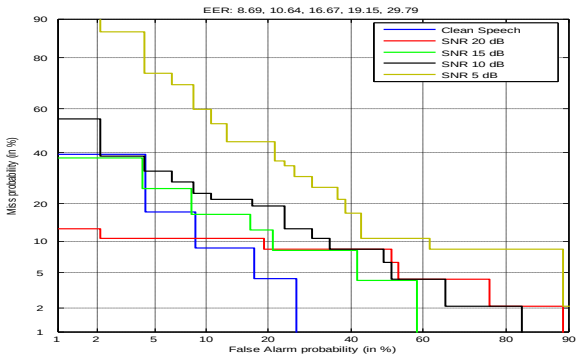

Fig 3(c): DET curve for baseline Speaker Verification System with input enhancement by Wiener filter method for different level of noise

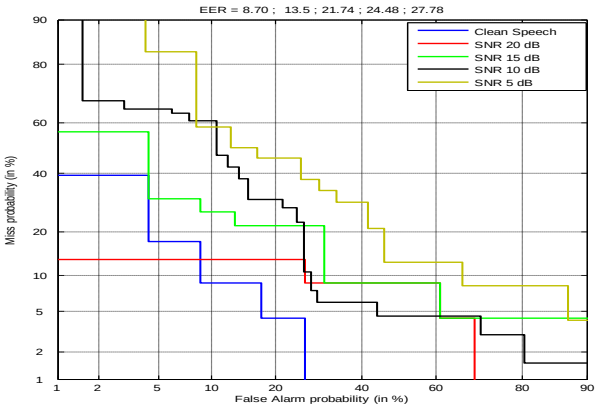

Fig 3(d): DET curve for baseline Speaker Verification System with input enhancement by Kalman filter method for different level of noise

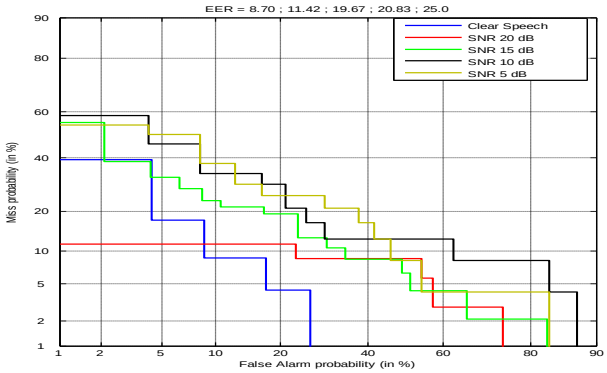

Fig 3(e): DET curve for baseline Speaker Verification System with input enhancement by Wiener filter with Spectral Subtraction method for different level of noise

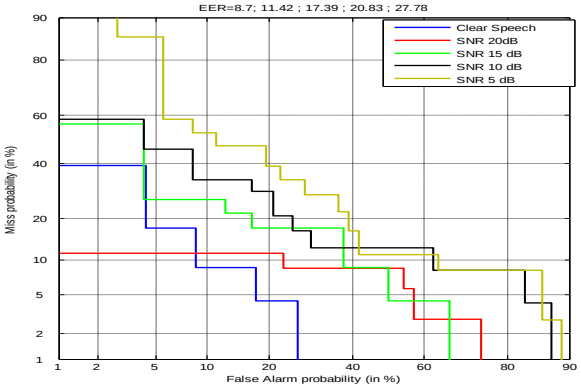

Fig 3(f): DET curve for baseline Speaker Verification System with input enhancement by Kalman filter with Spectral Subtraction method for different level of noise

VI. CONCLUSION

This paper investigated the problem of speaker verification in noisy conditions assuming absence of information about the noise. The works reported in this paper are carried out using a database contaminated by using simulated white Gaussian noise. In the present study, it has been observed that additive white noise plays an important role in degradation of the performanceof speaker verification system. Three different methods, namely spectral subtraction, Wiener filter and Kalman filter has been used for noise elimination from the speech signal. From the experiment, it has been observed that spectral subtraction is a very efficient method for elimination of white noise in high SNR condition. However, its performance rapidly degrades with reduction of SNR. Further, in the present study, it has been observed that at all SNR condition Wiener filter is an efficient noise reduction technique whereas Kalman filter is efficient in reduction noise at low SNR condition. Further, it has been observed that spectral subtraction when combined with Wiener filter and Kalman filter, the system performance does not increase considerably.

VII. ACKNOWLEDGEMENT

This work has been supported by the ongoing project grant No. 12(12)/2009-ESD sponsored by the Department of Information Technology, Government of India.

REFERENCES

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Transactions on Acoustics Speech and Signal Processing, ASSP-27(2) pp 113-120, 1979.
[2] B. Anderson and J. Moore, Optimal Filtering, Prentice Hall, 1979.
[3] J.S. Lim and A. V. Oppenheim, "Enhancement and band width compression of noisy speech", Proc. of the IEEE, Vol. 67, No. 12, 1586–1604, Dec. 1979.
[4] H. Sorenson, Kalman Filtering: Theory and Application, IEEE Press, 1985.
[5] M. Fujimoto and Y. Ariki, "Noisy Speech Recognition using Noise Reduction Method based on Kalman filter", Proc. ICASSP-2000, vol. 3, pp-1727-1730, 2000.
[6] Z. Xiaojia, S. Yang and W. DeLiang, "Robust speaker identification using a CASA front-end", Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, pp.5468-5471, 2011.
[7] N.T. Kleynhans and E. Barnard, "Language dependence in multilingual speaker verification", in Proc. of the 16th Annual Symposium of the Pattern Recognition Association of South Africa, Langebaan, South Africa, pp. 117-122, 2005.
[8] A. Reynolds, "Robust text-independent speaker identification using Gaussian mixture speaker models," Speech Communications, vol. 17, pp. 91-108, 1995.
[9] NIST 2003 Evaluation plan, http://www.itl.nist.gov/iad/mig/tests/sre/2003/2003-spkrec-evalplan-v2.2.

Utpal Bhattacharjee received his Master of Computer Application (MCA) from Dibrugarh University, India and Ph.D. from Gauhati University, India in the year 1999 and 2008 respectively. Currently he is working as an Associate Professor in the department of Computer Science and Engineering of Rajiv Gandhi University, India. His research interest is in the field of Speech Processing and Robust Speech/Speaker Recognition.

Pranab Das received his Master of Science (M.Sc.) in Computer Science from Assam University Silchar, India in the year 2005. Currently he is working as an Assistant Professor in the department of Computer Science and Information Technology of Don Bosco college of Engineering and Technology,Guwahati, India. He is alsopursuing his Ph.D. in Computer Sciencefrom Rajiv Gandhi University, India. . His research interest is in the field of Speech Processing and Robust Speech/Speaker Recognition.