An Enhanced Speaker Recognition System Using a Combined Approach of Speech Signal and EGG Signal

Kunal Anand
Department of Computer Science & Engineering
College of Engineering Bhubaneswar, Bhubaneswar, Odisha, India

Subhasish Mohapatra
Department of Computer Science & Engineering
College of Engineering Bhubaneswar, Bhubaneswar, Odisha, India

Abstract- Speaker Recognition (SR) is the process of detecting a speaker based on the features obtained from the speaker's voice sample. The traditional SR system is based on the recorded speech signal as the only source of information. The decision to authenticate or reject the speaker is taken using the MFCC features extracted from the speech sample obtained from the speaker. This is done using Gaussian Mixture Model (GMM). However, this speech signal approach has few limitations like it performs well in a noiseless environment, but the performance degrades in noisy environments due to the effect of noise on the speech signal. An alternate approach to overcome the limitation of speech signal approach is the SR system using excitation source characteristics using Electroglottography (EGG) signal. This paper is intended to do a comparative study of the performance of SR system using speech signal and EGG system, along with discussing the benefits and shortcomings of each approach to suggest a suitable approach to build an efficient SR system. The paper also enlightens the concept of score level fusion of two approaches to get an enhanced and efficient SR system. The paper concludes with a glipmse on the future works that might be carried for further performance enhancement.

Keywords – Speaker Recognition (SR), MFCC, EGG, GMM, Excitation source

I. INTRODUCTION

In today's scenario identifying an authorized person is a very important aspect in order to avoid any unauthorized access to a system. There are some certain physical as well as behavioral characteristics based on which one human being is distinguished from another. These characteristics are referred to as Biometric Measures [1]. Now, some features like facial and vocal features can be perceived very easily, whereas some features like finger prints, iris pattern, DNA structure etc. are difficult to extract distinguishing features and so they require experts for this purpose. Combined, they are known as biometric security features which are applied for security, surveillance and forensic application to authorize individuals and restrict unauthorized access. The process of designating or identifying an individual based on the characteristics extracted from his/her voice sample is known as Speaker Recognition (SR) [2]. This designation of an individual based on the voice sample is also known as Voice Recognition [3-4]. Using this technique, the speaker's voice can be used to affirm their identity and control their access to the services like voice based banking, shopping over phone, voice based security systems, database related services, voice post etc. Generally Speaker recognition is defined in following three tasks named as Speaker Identification, Speaker Verification and Speaker Detection. Speaker Identification is the process in which the features obtained from a voice sample of an unknown speaker is compared with a set of known speaker models. The distinguished speaker is decided on the basis of the label with whom the best matching score is obtained. In Speaker verification process, an identity claim is also provided as input along with the voice sample. Hence in this case, the comparison of the unknown speaker voice sample is only performed with the speaker model of the claimed identity. The provided identity claim is accepted if the comparison gives the satisfactory result; otherwise the call is rejected. [5]. In recent years, Speaker Detection [6][7] is delineated in NIST (National Institute of Standards and Technology) evaluations. It is an open set speaker identification task to identify a particular speaker from a given conversational voice sample involving multiple speakers. Now, since speaker detection involves a conversational voice sample as input, it requires an additional task known as speaker tracking to be done. Speaker tracking is the procedure of defining the intervals in the test sample during which the target speaker is speaking.

Apart from above classification, SR tasks, based along the character of input speech, can likewise be separated into two classes named as Text dependent SR system and Text independent SR system [8][13].

1.1 Basic Structure of Speaker Recognition (SR) System

The basic structure of a speaker recognition system consists of two distinct phases known as Training phase and Testing phase as shown in fig 1.1

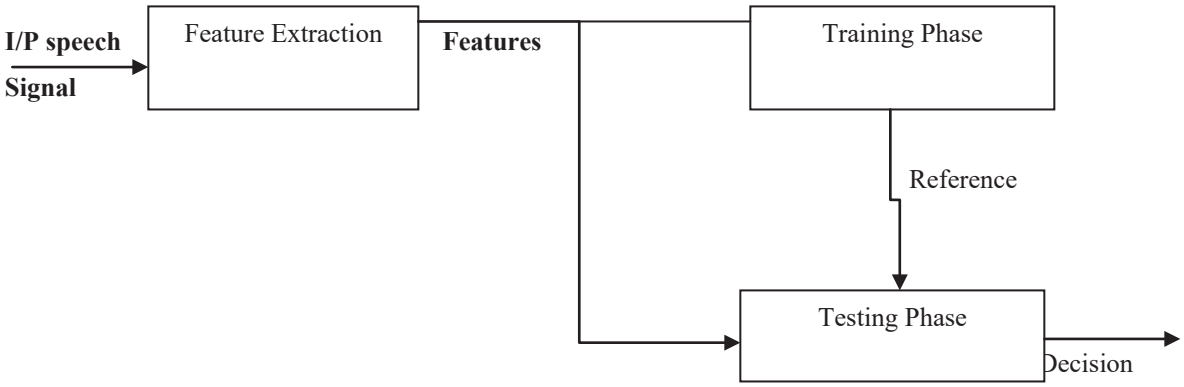

Fig. 1.1: Block diagram of SR system

A speech sample obtained from an unknown speaker acts as the input to the system. The obtained speech signal is preprocessed sampling, filtering etc. followed by the analysis that is typically some kind of short-term spectral analysis. This captures the features sensitive to a speaker as identified subsequently. These extracted features are compared with prototype features compiled into the models of known speakers. A matching process is invoked to compare the sample features and the model features. If the identification process is of closed-set then the model which provides the best matching score is assumed to be the distinguished speaker. On the other hand, in case of open-set identification process, if the matching score does not satisfy a threshold test then a no match decision is taken [5]. All SR systems have to serve following two different phases named as Training or preparation phase and testing phase. During Training stage, the voice sample from each registered speaker is collected to build a reference for that speaker. After completion of the preprocessing Cepstral analysis (MFCC, in our case) is performed to extract speaker distinguishing feature. These extracted features are used to build a speaker model using a suitable modeling technique (GMM, in our case) for each speaker and it is used as a reference model. During the testing phase also, the feature extraction part is repeated for test data and these characteristics are examined against the speaker model and the decision is made about recognition.

II. SPEAKER RECOGNITION SYSTEM- A REVIEW

A speech signal contains basically two sorts of data named as Low level information that are easier to extract from the speech signal as they are immediately accessible from the speech signal. Example: Formant location, Bandwidths, pitch periodicity. Another data is termed as High level information that is difficult to extract directly from the speech signal. Example: Perception of words and their meaning, syntax etc. The most common and popular SR system is based on the speech signal recorded from a speaker. This system is commonly based on some features like linear predictive Cepstral coefficients (LPCC), Mel-frequency Cepstral coefficients (MFCC) [10], or log area ratio (LAR). Out of the referred features, MFCC features are widely used technique for SR system. Gaussian mixture model (GMMs) [11] has been one of the most generally used approaches for modeling in speaker recognition applications. Nevertheless, this technique works fine only in noiseless environment. The performance gets degraded when the atmosphere becomes noisy. When the speech sample gets corrupted by noise, the dispersion of the speech feature vectors is also damaged and it produces poor recognition performance.

Electroglottograph (EGG) Signal: During speech production, the action of vocal folds is an important aspect for several speech applications. When twin in folding tissue layers vibrate due to the air flowing out from the windpipe into the larynx during expiration, the vocal folds are created.. Electroglottography (EGG) [9] is a non-invasive technique to examine the action of the vocal folds. An example of implantation of EGG technique is a throat microphone which consists of two electrodes. A high frequency modulated current is transmitted through the subject's thyroid cartilage by a couple of electrodes. The time variation of the level of contact between vibrating vocal folds during voice production is measured by the EGG device. Typical waveforms of the audio signal and corresponding EGG signals are shown on Figure 2.1 as below:

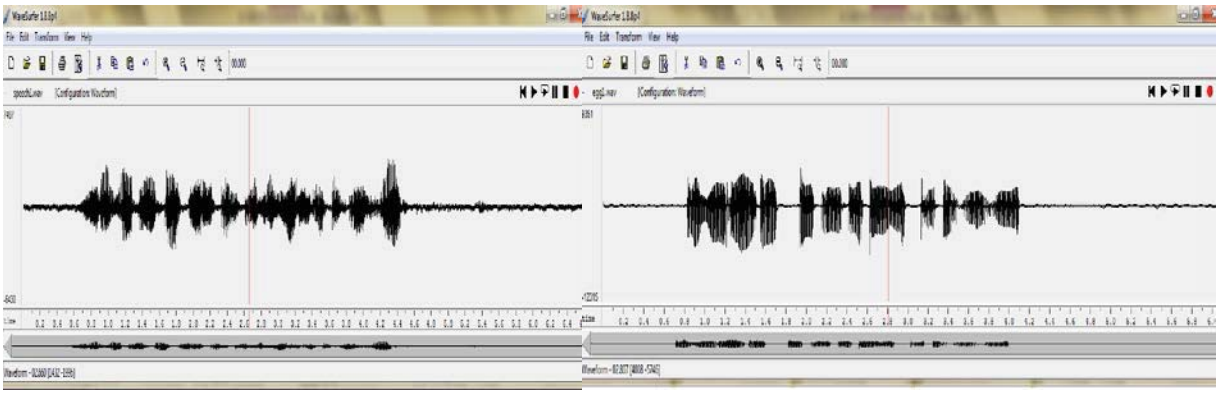

Fig. 2.1: A speech segment represented by audio waveform (left) and EGG waveform(right)

The EGG waveform consists of slow changes as well as low-level high-frequency noise which is easily distinguished in the case of unvoiced region.

The conventional SR system is based on speech signal for identification purpose. The limitation of this approach is that the system gives good performance only if the environment is noiseless. The moment it comes into the contact of the noise the performance of the system sees a sharp decline because the speech signal recorded using a normal microphone also captures the noise from the background which degrades the character of speech signals. Due to this the classification is not trustworthy. Now, if we consider the excitation source characteristics then we can find that the vocal fold vibration varies from one person to another during speech output. This variation could be in the extent of closure or in the manner and pace of blockage. If EGG device is used along with the conventional microphone then these speaker specific vocal fold actions could be captured better. Here, one important significance of EGG signal is that, since it observes the vibration of vocal folds happening inside the torso, it is literally not or less struck by the interference environment which constitutes it a safer option to perform speaker recognition tasks efficiently in noisy environment also. A comparative study of the speech signal and EGG signal in different environment is discussed as below using figure 2.2:

Speech Signal Analysis (Noiseless Environment)

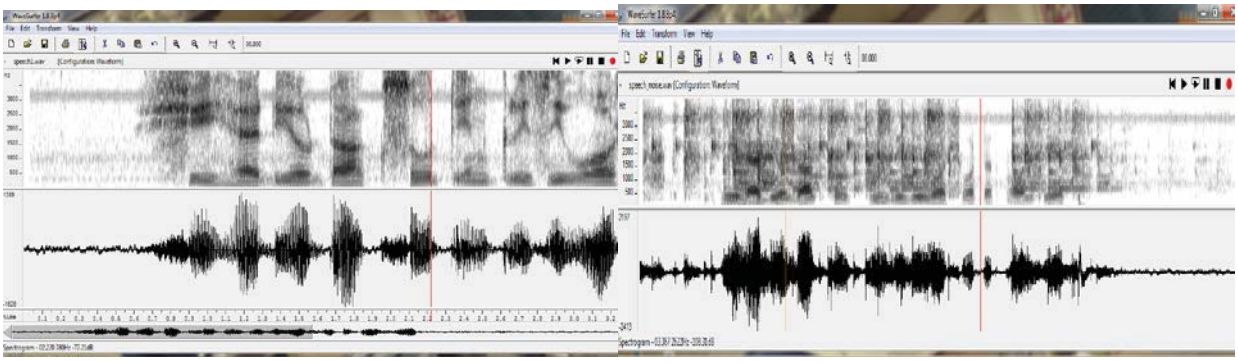

Fig 2.2: Speech signal with its spectrogram in noiseless environment(left) and noisy (right)

From fig 2.2 we can see that the dominant frequencies of speech signal are easily identified in case of noiseless environment. On the other hand, the dominant frequencies get damaged when the speech signal comes into the contact of noise.

EGG signal

Fig 2.3: Waveform of EGG signal along with the spectrogram in noiseless as well as noisy environment

On the other hand EGG signal is not affected by noise as the EGG signal is recorded by the electrodes connected to the subject's throat.

Time domain aspect

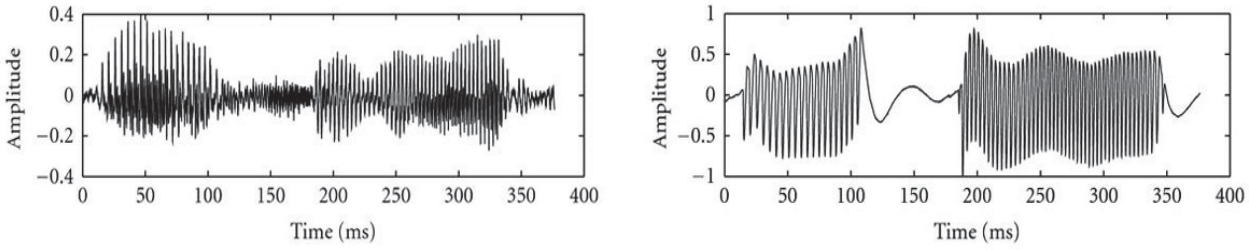

Fig 2.4 : A speech segment represented by: (left) speech and (right) EGG waveform

In fig 2.5, we can see that the EGG waveform contains slow changes and very low-level high-frequency noise for unvoiced segment. On the other hand, in the case of speech signal, in unvoiced segment also there is some low level energy present.

III. Proposed work: Speaker recognition system using a combined approach

In the earlier works carried out, it has been shown that the performance gets degraded in presence of the noise [12]. The implementation of SR system using EGG signal may be an alternate approach. In this work, we have first analyzed the performance of the SR system using EGG signal.

3.1 EGG Parameterization

The EGG signal is regarded as "almost periodic" in voicing segments. Fig 3.1 shows the characteristics of an EGG signal in one pitch period.

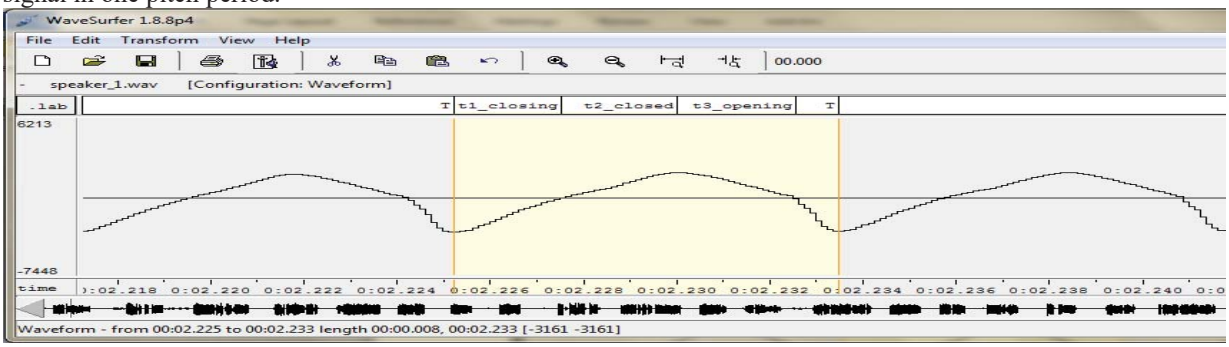

Fig 3.1**:** One pitch period of EGG waveform, showing opened phase (T), closing phase (t1), closed phase (t2) and opening phase (t3)

One pitch period represents a glottal cycle. On that point are four stages involved in a glottal cycle. These are keyed out as open phase, closing phase, closed phase and opening stage. The vocal fold is separated during open phase,

time interval T, due to which air flows into vocal tract. This phase records low and flat admittance. The lower tissue layers of vocal fold starts making contact gradually by the end of open phase. Due to this the admittance increases marginally and it is recorded by the EGG device. Now, during the beginning of closing phase i.e. at the completion of open phase the contact gradually increases and it proceeds to the middle and upper tissue layer of the vocal fold. Due to this the admittance gets increased during the closing phase. In fig 3.1, t1 represents "closing phase". The vocal fold remains in contact for the entire duration of the closed phase. The admittance may increase or decrease marginally during this period due to elastic collision between tissues. As soon as the closed phase begins to end, the vocal folds start preparing for separation. In fig 3.1, t2 represents "closed phase". When opening phase begins, the separation of vocal folds layers start and it gradually goes to the middle and the upper tissue layers. Towards the close of the opening phase The vocal folds gets fully separated due to which air flows at maximum into the vocal tract from the windpipe. When the vocal folds start separating, the important thing to notice here is the sharp fall in admittance, which is because the admittance dips at an increased rate compared to decrease in admittance during the terminal level of closed phase. In fig 3.1, t3 represents "opening phase". [12] The four different phases in a glottal cycle are represented as Glottal opening instant (GOI), Glottal closure instant (GCI), Glottal closed instant (GCDI) and Glottal opened instant (GODI) shown in the fig 3.2.

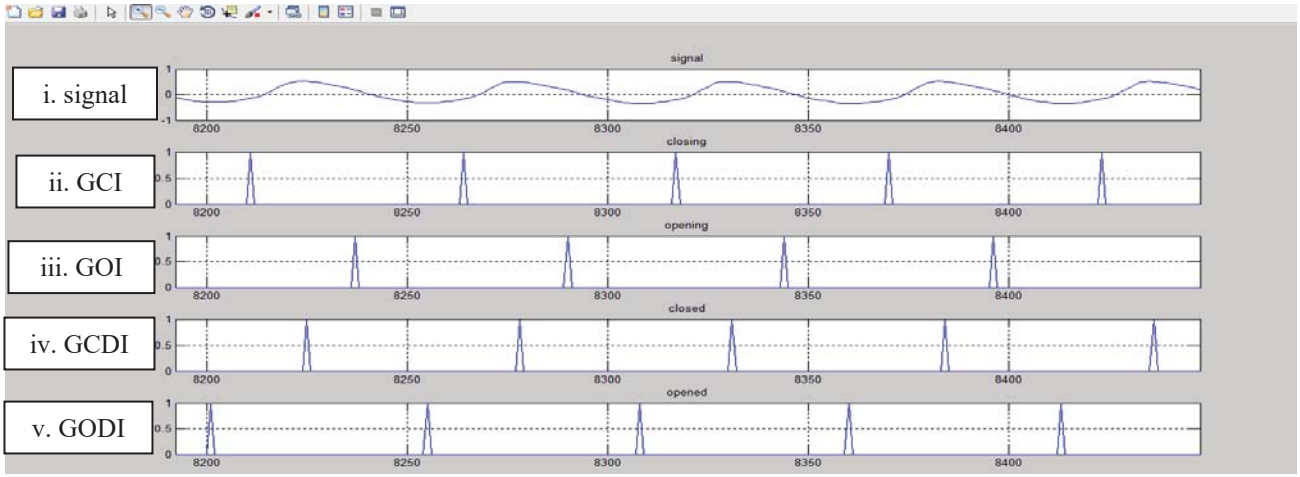

Fig: 3.2 Time domain features of EGG signal: (i) Signal, (ii) Glottal Closure Instant, (iii) Glottal Opening Instant, (iv) Glottal Closed Instant, (v) Glottal Opened Instant

3.2 Creation of Speaker Database

To go with the proposed work, first of all a speaker database of 100 speakers (58 male speakers and 42 speakers) was created. During the database creation, the speech utterance text was prepared which contained 50 text utterances borrowed from "TIMIT" database. Out of 50, 30 text utterances have been used as training data and rest 20 text utterances has been used as test data. The transcription was performed in a laboratory environment using "AUDACITY version 2.0.3" software with the aid of a microphone and EGG device.

3.3 MFCC Feature Extraction

The best known and most popular technique is MFCC which has been used for feature extraction in this work. The feature extraction was performed for both training data and testing data using MFCC technique [10] for speech signal.

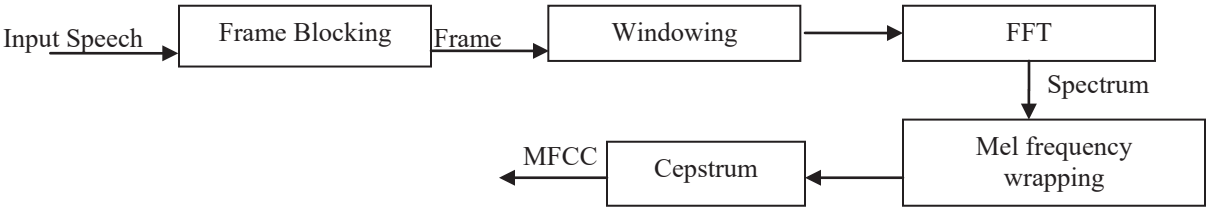

Fig 3.3: MFCC Feature Extraction

Frame Blocking: Here, the speech signals are blocked into several frames of N samples. The samples are adjacent and are separated by M where, M<N. The first frame consists of the first N samples. The second frame begins M samples after the first inning, and overlaps it by N - M samples and hence along. Windowing: Once frame blocking is accomplished, windowing is done to minimize the effect of discontinuities at the start and the end point of each frame. If we set the window as w (n), 0≤n≤N-1, where N is the number of samples in each form, and so the effect of windowing is the sign. Typically the Hamming window is used.

$$y_i(n) = x_i(n)\, w(n), \qquad 0 \le n \le N-1$$

Fast Fourier Transform (FFT): In the next step Fast Fourier Transform is performed, where each frame is converted from time domain into frequency domain. The FFT is a fast algorithm to implement the Discrete Fourier Transform (DFT), which is defined on the set of N samples {xn}, as follow:

$$X_k = \sum_{n=0}^{N-1} x_n\, e^{-j2\pi n/N},\ k=0,1,2,3,\ldots..,N-1$$

In general Xk's are complex numbers and we only consider their absolute value (frequency magnitudes). The solution after this step is frequently consulted to as a spectrum.
Mel-frequency wrapping: It has been demonstrated through some studies that human perception of the frequency contents of sounds for speech signals does not conform to a linear scale. Therefore, a subjective pitch is measured along a scale known as 'Mel' scale for each tone with an actual frequency, f, measured in Hz. The Mel-frequency scale is the linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. A filter bank, which is spaced uniformly on the Mel-scale can be applied to simulate the subjective spectrum. That filter bank has a triangular band pass frequency response, and the spacing as well as the bandwidth is paid off by a constant Mel frequency interval. Cepstrum: The log Mel spectrum is converted back into time. The effect is known as the Mel frequency Cepstrum coefficients, commonly referred as MFCC. As the Mel-spectrum coefficients (and so their logarithm) are literal numbers, they can be converted into the time domain using the Discrete Cosine Transform (DCT). Since the first component from the DCT represents the mean value of the input signal, which carried little speaker specific information, is excluded.

3.4 Gaussian Mixture Modeling (GMM)

Gaussian Mixture Models (GMM) [11], have shown to be very effective because it is simple to implement and computationally cheap in such scenario where the system does not have any information about the text utterance in advance. A probability density function which is a mixture of Gaussians provides the feature vectors of training utterance X. Passed on the enrollment data X, the maximum likelihood estimates of the λ can be obtained using the expectation-maximization (EM) algorithm. A likelihood ratio test statistic of the form as mentioned below establishes the speaker verification process.

$P(Y|\lambda) / p(Y|\lambda_{bg})$; where λ is the speaker model and λ_{bg} represents a desktop model.

In our work, speaker models are built for different Gaussians like 2, 4, 8, 16, 32 and 64 for each speaker using speech signal as well as an EGG signal in noiseless condition using the GMM modeling technique.

3.5 Test Phase

The last phase of SR system consists of testing phase where the test speech sample is provided as the input to the system. The SR system does the comparison of the feature extracted from the speech sample with the models built using GMM for each speaker and produces the result in the form of the best matching score for speaker identification process and accept/reject for speaker verification process. The block diagram of the testing stage is shown in fig 3.4 as under:

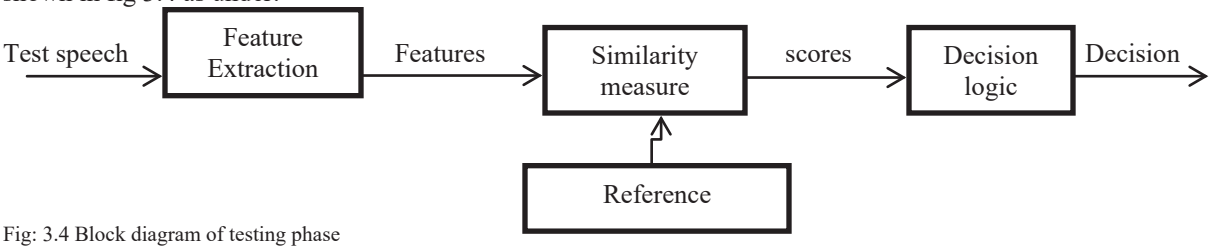

Fig: 3.4 Block diagram of testing phase

3.6 Speaker Recognition System: A combined Approach

In earlier systems, Speaker recognition was done using speech signal or EGG signal using different attacks. In that approach, feature extraction was performed for both string and trial data, after which model building was executed and then using test data features the identification task was performed and the determination was reached. The above approach gave different performances under different circumstances which have been hashed out in section 4. In society to heighten the performance of SR system, a mixed approach is adopted where before taking the last decision the features from different attacks were merged together. This combined feature is further used for recognition purpose. The block diagram for this combined approach is shown in fig 3.5.

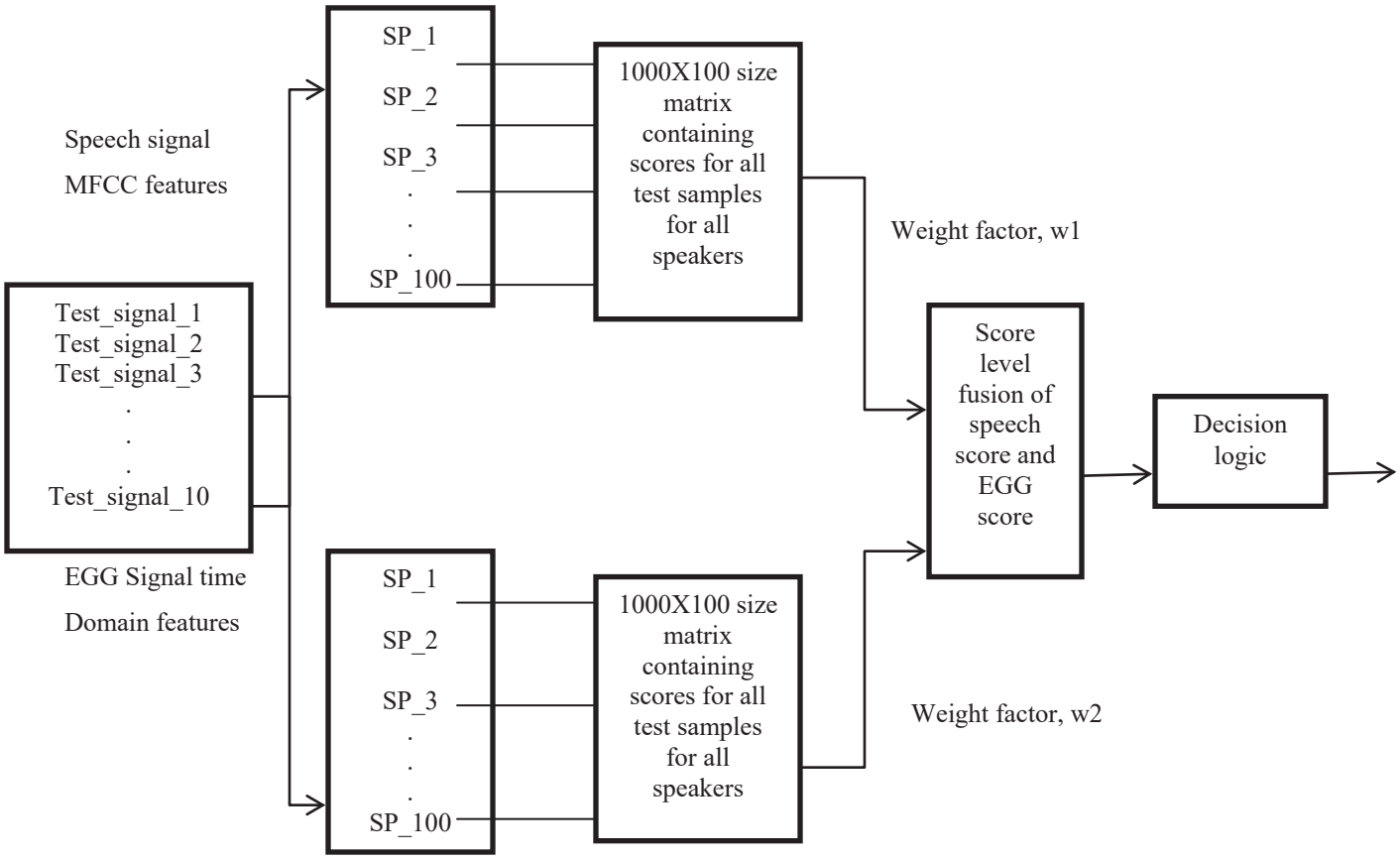

Fig 3.5: SR System: A combined approach

If we analyze the fig 3.5, we adopted a combined approach using MFCC approach for speech signal and time domain approach for EGG signal. In above fig 3.5, for speaker_1, the MFCC feature from speech signal for test_sample_1 has been examined on the model for each speaker. The exam yields a probability score for that test sample against each speaker. So, 100 scores are obtained. Likewise, time domain features from the EGG signal for test_sample_1 has been examined on the model for each speaker. The exam yields a probability score for that test sample against each speaker. So, 100 scores are obtained. At present there are 100 speakers and each speaker has 10 test samples, then the size of the two tables will be 1000X100 i.e. A total of 100000 probability scores is obtained for speech signal approach and EGG signal approach. After that, the score level fusion is performed where different weight factors lying between 0.1 and 0.9 is used against the speech probability score and the EGG probability score. The obtained result will be a combined score which takes in some quantity of both MFCC features of the speech signal and time domain feature from EGG signal. This combined score is used for recognition purpose and is believed that this approach will raise the operation of the organization.

IV. PERFORMANCE EVALUATION

4.1 Performance of SR System using Speech Signal in Noiseless Environment

In this approach, speaker recognition was done using speech signal. Foremost of all, MFCC feature extraction was performed for training data as well as testing data obtained for each speaker. In one case the feature extraction part was completed, models of different modes for each speaker were made using a GMM modeling technique. In testing part, 10 test samples of 5 second duration were applied. Hence, a total of 1000 test samples were examined on the models created during model building stage. The effect obtained is mentioned in table 4.1.

Speaker Models	Speaker Recognition Performance (in %)
Model_2	69
Model_4	71.90
Model_8	77
Model_16	78.10
Model_32	79.10
Model_64	78.10

Table 4.1 Speaker recognition performance using the speech signal in noiseless environment

Observation: From the above table 4.1, we can observe that the SR system based on speech signal gives a significant performance in noiseless condition. With the increase in the number of Gaussian, the performance gets better i.e. For 2 Gaussian model the system gives 69% whereas for the 32 Gaussian model it gives the peak performance of 79.10%. However, the performance starts to decrease when 64 gaussians are used, which means that the performance of the system is best with number of gaussians not very high not very low.

4.2 Performance of SR System using Speech Signal in Noise Environment

In this section the aim was to determine the measure of debasement in the performance of system as speech signal feature vectors get damaged when the signal comes into the contact of noise. For this, the experiment was performed using three types of noises like babble noise, factory noise and white interference. For each noise, the experiment was done using different signal to noise ratio (SNR) and reflections were named. The percentage degradation in the carrying out of the speaker recognition system was calculated using following formula.

Performance degradation (in %) = ((A-B) / A) * 100

Where,

A = Performance percentage in noiseless

B = Performance percentage in -5 SNR

SR system performance was observed in presence of noises like babble, factory and white noise obtained from "NOISEX 92" database. Feature extraction was performed using the mixture test samples and testing was performed. The received solutions are noted in table 4.2 as pictured at a lower place:

Speaker recognition system performance in babble noise							
Models/ SNR	**Noiseless (A)**	**20 (in dB)**	**15 (in dB)**	**10 (in dB)**	**5 (in dB)**	**-5 (dB) (B)**	**Performance Degradation (in %)**
Model_2	69	56.90	40.20	17.40	5.90	2	97.10
Model_4	71.90	61.80	41	18.20	6.10	1.30	98.10
Model_8	77	64.40	49.80	23.40	9.70	3.10	95.97
Model_16	78.10	68.20	53.10	24	9.20	2.70	96.54
Model_32	79.10	69.90	57.10	27.30	10.10	2.40	96.96
Model_64	78.10	70.90	57.20	27.50	9	3.50	95.51
Speaker recognition system performance in factory noise							

Models/ SNR	Noiseless (A)	20 (in dB)	15 (in dB)	10 (in dB)	5 (in dB)	-5 (dB) (B)	Performance Degradation (in %)
Model_2	69	46.30	29.50	8	2.10	1	98.55
Model_4	71.90	53.20	34.20	7.60	1.50	1	98.66
Model_8	77	54.40	36.50	7.20	1.90	1.10	98.57
Model_16	78.10	59.60	41.30	8.50	1	1	98.72
Model_32	79.10	62.80	42.90	10.30	1.20	1	98.73
Model_64	78.10	63.70	43.10	11.50	1.20	0.9	98.84
Speaker recognition system performance in white noise							
Models/ SNR	**Noiseless (A)**	**20 (in dB)**	**15 (in dB)**	**10 (in dB)**	**5 (in dB)**	**-5 (dB) (B)**	**Performance Degradation (in %)**
Model_2	69	17.70	10	3.60	1.80	1	98.55
Model_4	71.90	18.80	9.50	2.80	1.20	1	98.66
Model_8	77	21.70	7.60	2.50	1.40	1	98.70
Model_16	78.10	24.10	8.30	3.20	2	1.60	97.95
Model_32	79.10	22.60	6	3.60	2.30	0.80	98.98
Model_64	78.10	21.40	6.10	3.70	2.30	0.30	99.61

Table 4.2: Performance of SR system using speech signal in the presence of different noises

Observation: From table 4.2, we can see that the performance of SR system using speech signal degrades in the presence of noise. The experimentation was done using three different noises babble, factory and white interference. For each noise, the carrying out of the system was highly degraded for -5 SNR whereas performance goes on improving when the SNR is increased i.e. for 20 SNR the performance is more gamey.

4.3 Performance of SR System using EGG signal (Time Domain Approach)

In this advance, four time domain parameters named as Glottal opening instant (GOI), Glottal closure instant (GCI), Glottal closed instant (GCDI) and Glottal opened instant (GODI) were applied for performing the speaker identification task. These characteristics are discussed earlier, in detail, in sub-section 3.1. Using these features, for EGG signal feature extraction was performed for train data followed by model building for each speaker using different gaussians. Afterward that the speaker identification task was performed and the obtained result is indicated in table 4.3 as under:

Speaker Models	Speaker Recognition Performance (in %)
Model_2	30
Model_4	26.90
Model_8	29.60
Model_16	25.40
Model_32	15.30

Table 4.3: Performance of SR system using EGG signal in time domain approach

4.4 Performance of SR system using combined approach

Model / test duration	Speaker Recognition performance				
	MFCC technique using speech	A time domain approach using EGG	Weight factor (w1)	Weight factor (w2)	Combined Approach
Model 8	77	29.60	0.1	0.9	39.90
			0.2	0.8	52.10
			0.3	0.7	58.80
			0.4	0.6	62.90
			0.5	0.5	68.30
			0.6	0.4	72.80
			0.7	0.3	76.30
			0.8	**0.2**	**78.40**
			0.9	0.1	78

Table 4.4: Performance of SR system using a combined approach

In our experimentation, we have considered model with 8 gaussians. For model_8, MFCC approach for speech signal gives a speaker recognition performance of 77%. On the other hand, the time domain approach, for EGG signal, gives a speaker recognition performance of 29.60%. When speaker recognition task was performed using the above discussed combined approach, the system gives the performance of 78.40%, i.e. the performance of the SR system using the combined technique is improved by 1.78% as compared to the SR system based on speech signal and 62.24% as compared to the SR system using EGG signal.

V. CONCLUSION

Identifying speakers by voice was originally investigated for applications in speaker authentication.. This research work is meant to explore different ways to perform speaker recognition task and determine the best possible choice. In the beginning stage of the work, speaker recognition was done using speech signal in noiseless as well as noisy environment. The experiment made an observation that the system performs reasonably well in a noiseless environment only whereas, the performance declines in noise. The following stage of the study concentrated on the use of EGG signal to perform speaker recognition as previous research works has shown that EGG signal does not become affected by interference. The advantage of EGG signal is its robustness from noise which gets it a desirable choice for speaker recognition as we have seen that SR system performance degrades severely (approx. 95% degradation) when a speech signal comes into the contact of noise. The final section of the workplace implemented a combined approach using MFCC approach for oral communication and time domain approach for EGG signal. In this blended approach, the probability scores were computed for both speech signal and EGG signal approach. The two lots were combined using score level fusion before taking any decision. This combined score was applied to do the speaker identification task and the obtained result gave an enhanced performance (78.40% for 8 gaussians) as compared to the earlier used approaches.

This paper work may be further extended to other aspects of SR system like speaker verification and detection.

REFERENCES

[1] J.S Dunn, F. Podio. : Biometrics Consortium website, 2007.

[2] Kinnunen, Tomi; Li, Haizhou (1 January 2010). "An overview of text-independent speaker Recognition: From features to supervectors". Speech Communication 52 (1): 12–40. Do: 10.1016/j. specimen. 2009.08.009

[3] Jean-Francois Bonastre, F Bimbot, Louis-Jean Boe, Joseph P. Campbell, Douglas A. Reynolds, Ivan Magrin-Chagnolleau: "Person Authentication by Voice: A Need for Caution"]. Person Authentication by Voice: A Need for Caution. 8th European Conference on Speech Communication and Technology. Geneva, Switzerland, September' 03

[4] Van Lancker and Kreiman (July 3, 1984). "Familiar voice recognition: Patterns and Parameters. Part I: Recognition of backward voices". Journal of Phonetics. Pp. 19–38, 1984

[5] Rabiner L.R, Juang B.H, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, N.J., 1993.

[6] S. Furui. : "An overview of speaker recognition technology", ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, pp. 1-9, 1994.

[7] M.A. Przybocki, A.F. Martina, "speaker recognition, evaluation, using summed two channels Telephone data for speaker detection and speaker tracking, Eurospeech" 1999 Proceedings (1999) pp. 2215–2218.

[8] Rosenberg A.E, Bimbot F, Parthasarthy S. : "Overview of Speaker Recognition" In J. Benesty, M. Mohan Sondhi & Y. Huang (Eds.), Springer Handbook of Speech Processing, Page 725-741, Berlin

[9] FLE Lecluse, MP Brocaar, J. Vershurre The electoglottography and its relation to glottal activity. Fol Phoniatr 1975; 27: 215-24.

[10] Md. Rashidul Hasan, Jamil Mustafa, Md. Golam Rabbani, Md. Saifur Rehman :"Speaker Identification using MFCC" 3rd International Conference on Electrical & Computer Engineering ICECE 2004, 28-30 December 2004, ISBN 984-32-1804-4 565

[11] A.R. Douglas, C.R. Richard: "Robust text-independent speaker identification using Gaussian mixture speaker models, " IEEE Transactions On Speech And Audio Processing, , vol. 3, No. 1, pp. 72-83, January, 1995.

[12] Cirovic Zoran, Banjac Zoran, Milosavljevi Milan: "Multimodal Speaker Verification Based On Electroglottograph Signal and Glottal Activity Detection" EURASIP Journal on Advances in Signal Processing January 2010.

[13] D A reynolds, R C. Rose: "Robust text-independent speaker identification using Gaussian Mixture speaker models" IEEE trans. Speech audio processing 3, 72-83 (1995).