Asian Journal of Computer Science and Technology
ISSN: 2249-0701 (P) Vol.7 No.2, 2018, pp.123-127

DOI: https://doi.org/10.51983/ajcst-2018.7.2.1866

Performance of Speaker Verification Using CSM and TM

S. Sathiamoorthy[1], R. Ponnusamy[2*] and R. Visalakshi[3]
[1&2]Assistant Professor, Division of Computer & Information Science,
Annamalai University, Annamalai Nagar, Tamil Nadu, India
[3]Assistant Professor, Department of Computer Science and Engineering,
Annamalai University, Annamalai Nagar, Tamil Nadu, India
E-mail: ks_sathia@yahoo.com
* Corresponding Author

(Received 15 July 2018; Revised 29 July 2018; Accepted 13 August 2018; Available online 19 August 2018)

Abstract **- In this paper, we presented the performance of a speaker verification system based on features computed from the speech recorded using a Close Speaking Microphone(CSM) and Throat Microphone(TM) in clean and noisy environment. Noise is the one of the most complicated problem in speaker verification system. The background noises affect the performance of speaker verification using CSM. To overcome this issue, TM is used which has a transducer held at the throat resulting in a clean signal and unaffected by background noises. Acoustic features are computed by means of Relative Spectral Transform-Perceptual Linear Prediction (RASTA-PLP). Autoassociative neural network (AANN) technique is used to extract the features and in order to confirm the speakers from clean and noisy environment. A new method is presented in this paper, for verification of speakers in clean using combined CSM and TM. The verification performance of the proposed combined system is significantly better than the system using the CSM alone due to the complementary nature of CSM and TM. It is evident that an EER of about 1.0% for the combined devices (CSM+TM) by evaluating the FAR and FRR values and the overall verification of 99% is obtained in clean speech.**

Keywords: **Autoassociative neural network, of Relative Spectral Transform-Perceptual Linear Prediction (RASTA-PLP); Close Speaking Microphone, Throat microphone**

I. INTRODUCTION

Human listeners can reliably recognize known voices by barely listening to the voice of the speaker. The uniqueness of one's voice can be attributed to both physical and acquired characteristics of a person. Due to the distinct shapes and sizes of the voice producing organs (e.g., vocal folds, vocal tract, larynx, etc.) and partly due to the articulators (e.g., tongue, teeth, lip etc.) .physical differences occur largely. Despite these anatomical properties, individuals can also be differentiated by their accent, vocabulary, speaking rate and other personal mannerisms that are attained over a period of time. Attaining this basic human specific capability is a key challenge for Voice Biometrics. Like human listeners, voice biometrics also uses the features of a person's voice to decide the speaker's identity [2]. The need for speedy, efficient, accurate, and robust Speaker verification is essential for rising importance for commercial, forensic, and government applications [3].

A. Speaker Verification

Speaker verification (SV) is the process of verifying the claimed identity of a speaker. It is a problem of binary classification in which the claim is either accepted or rejected based on the statistical similarity measures of a test utterance with the claimed speaker model (true class) and a selected background/impostor model (false class). SV is a one-to-one matching i.e., speaker's voice is coordinated to one template. Fig.1 shows the block-diagram of a typical SV system.

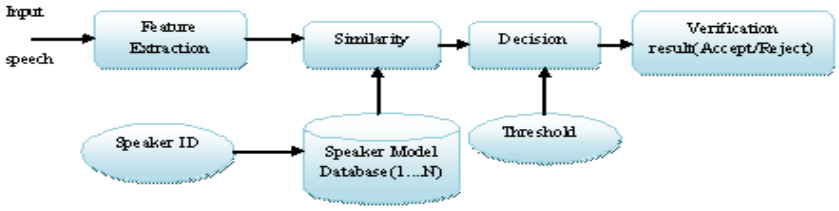

Fig. 1 Speaker Verification

1. Throat Microphone

In this paper, the Throat microphone device is used. TM also known as laryngophone will fit the TM around our neck. In TM, two transponders (mic pickups) are present and they absorb the vibrations generated by the larynx then turning them into electronic audio signals, by this means sound is transmitted directly from the throat (making solid contact with the throat). Since using transponders Throat strap at the back of our neck, it anchors the unit and keeps it in place.

The throat microphone is a transducer and is positioned in contact with the skin adjacent to the larynx near the vocal folds. The TM records clean speech in the presence of high background noise. Usually, the throat, speech is a low amplitude signal and is of high quality. In a noisy situation, the clearness of close speaking microphone speech is affected because the microphone picks up the voice as well as the background noise [4]. But the clearness of the throat microphone signal is almost same as that of the signal obtained in a noise-free environment. Therefore the throat microphone is an ideal choice for the use of speech applications, even in undesirable conditions [5]. A throat microphone and a person wearing the TM is depicted in Fig.2 (a) and (b).

2. Outline of the work

The features are computed using RASTA-PLP. The AANN model is used to confirm the identity of speakers from clean and noisy situation Speaker verification system is viewed as working in four stages namely Analysis, Features Extraction, Modeling and Testing [6]. The block diagram of speaker verification is shown in Fig.3

Fig. 2 Throat Microphone

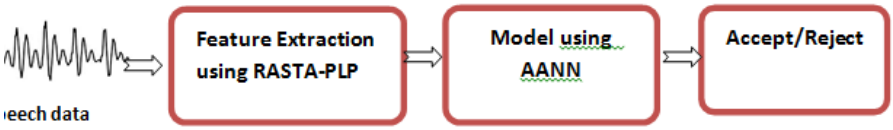

Fig. 3 Block Diagram of Speaker verification

II. ACOUSTIC FEATURE EXTRACTION TECHNIQUES

In this paper, feature extraction is based on RASTA-PLP for speaker verification and is explained as follows.

A. Pre-processing

Speaker signal must be preprocessed in order to extract the acoustic features from the speech signals and it is divided into consecutive analysis frame. In the proposed work, sampling rate of 8kHz pulse code modulation (PCM) format and 16 bit monophonic is deployed.

B. Relative Spectral Transform - Perceptual Linear Prediction(RASTA-PLP)

RASTA-PLP is an extension to PLP. PLP was proposed by Hynek Hermansky for warping spectra to minimize the differences between speakers when captring the important speech information [7]. RASTA method applies a band-pass filter to the energy in each frequency sub band in order to smooth over the short-term noise variations and to eliminate constant offset resulting from static spectral component in the speech channel, for instance, from a telephone line [8]. Our goal is evaluating the effect of the Rasta filtering on the features that are studied in this paper i.e. FF features and concatenated FF features, since the capability of RASTA processing to deal with diverse types of noise and more over it is evident that frequency filtered logFBEs can be enhanced with specific temporal filtering. As for the filter used, to start with an IIR filter with the transfer function.

$$H(z)=0.1*\frac{2+z^{-1}-z^{-3}-2z^{-4}}{z^{-4}*(1-0.98z^{-1})}$$

The lower cutoff frequency of the filter determines the fastest spectral change of the log spectrum, which is ignored in the output, whereas the high cutoff frequency of the filter decides the fastest spectral change and is preserved in the output parameters.

The high-pass part of the equivalent band pass filter is likely to assist in smoothing out some of speedy frame to frame spectral changes present in the short term spectral estimate due to analysis [8] aircraft channel. In Eqn.1, the low cut-off frequency is 0.6HZ.

III. MODELING THE ACOUSTIC FEATURES FOR SPEAKER VERIFICATION

A. Autoassociative Neural Network (AANN)

AANN contains of five network layer which extracts the distribution of the feature vector as shown in Fig.4. The input layer in the network has less number of units than the second and the fourth layers. The first and the fifth layers have more number of units than the third layer [9] [10]. The processing units in the second layer can be either linear or non-linear. But the processing units in the first and third layer are non-linear. To train the network, Back propagation neural network is incorporated [11].

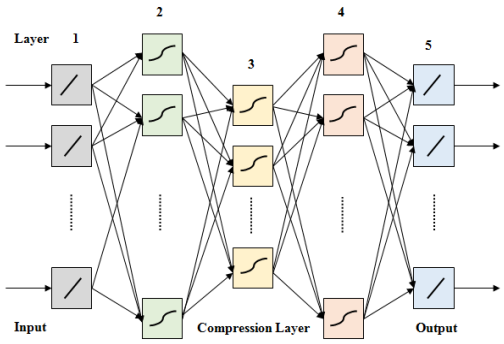

Fig. 4 A Five-Layer AANN model

IV. EXPERIMENTAL RESULTS

A. Datasets

Using CSM and TM, speech data's are collected in clean and noisy situation. In the experiments, the throat microphone, made up of piezoelectric ceramics, is placed by wearing it around neck. Since the finest spectral resolution among the contact microphones is throat microphone, we used in our experiments. For noise free situation, the corpus of speech are collected in a sound proof room from the department of linguistic, Annamalai university. For noisy situation, the corpus of speech are collected in class room, laboratories from 100 students for the duration ranging from 1 to 2 hours. All the speech signals are recorded in with sampling rate of 8 kHz, 16 bits with mono channel. Wave files ranging from 2secs to 5secs are extracted from the speaker database both for training and testing. Both the training and testing files are not dependent on text.

B. Evaluation of AANN with RASTA-PLP features using CSM and TM

1. Performance of Speaker Verification using CSM

The speech signal is segmented in successive frames, overlapping with each other. In order to reduce the effect of spectral leakage, each frame is multiplied by a hamming window [12]. Acoustic features representing the speaker information is captured from each windowed frame. These features characterize the short-time spectrum of the speech signal. The short term spectrum envelope of speech signal is attributed mainly to the shape of vocal tract. The spectral information of the same sound spoken by two person may vary owing to change in the shape of the individual's vocal tract system and the way of speech production. The calculation of RASTA-PLP features for a segment of speech signal is described in Section.2.2.

TABLE I PERFORMANCE OF SPEAKER VERIFICTAION IN TERMS OF NUMBER OF UNITS IN COMPRESSION LAYER

Microphone	13L-26N-4N-26N-13L	13L-48N-8N-48N-13L	13L-52N-12N-52N-13L
Close-Speaking-Microphone	80%	74%	72%
Throat Microphone	94%	92%	85%

The performance of speaker verification using 13 dimensional RASTA-PLP features for clean and noisy speech is evaluated using CSM and TM and by one of the pattern recognition technique called autoassociative neural network. AANN is a five layer neural network and is used to extract the distribution of the RASTA-PLP feature vectors. The performance of text-independent speaker verification system is evaluated for clean and noisy speech database.

Separate AANN models are used to capture the distribution of feature vectors of each speaker. The AANN structure *13L 26N 4N 26N 13L* achieves good performance in terms of varying structure of network in terms of number of units in compression layer. The structure is obtained from the experimental studies. The performance of speaker verification is obtained by varying the second(expansion layer) and third layer(compression layer) of AANN model and it is depicted in Table.1.

The RASTA-PLP feature vectors are given as both input and output. The weight are adjusted to transform input feature vector into the output. The output of each model is compared with input to compute the normalized squared error. The normalized squared error *e* for the feature vector *y* is given by, $e = \frac{\|y-o\|^2}{\|y\|^2}$, where *o* is the output vector given by the model. The error *e* is transformed into confidence score *c* using c=exp(-e). The average confidence score gives better performance than using confidence score for each frame.

The confidence scores for all the authentic and impostor claims are calculated and there are used to measure the performance of the system. If the confidence score is higher than a threshold, then the claim is accepted, otherwise the claim is rejected. For each threshold, the FAR and FRR are calculated using the confidence scores. The EER using thresholds. In clean and noisy speech database of 50 subjects for each, there are 50 authentic claims and 49*50 imposter claims for both CSM and TM. The structure of AANN method plays a significant role in the extraction of the distribution of the feature vectors.

An EER of 7.0% obtained by evaluating the performance in terms of FAR and FRR at threshold 0.79 and as shown in Fig.5 for clean speech. For noisy speech, an EER of 40.0% is attained by evaluating the performance in terms of FAR and FRR at threshold 0.55 and is represented in Fig.6. The performance of the speaker using RASTA-PLP features is

evaluated in terms of accuracy or Verification Rate(VR) and EER. The performance of text based speaker verification using AANN for both clean and noisy speech is shown in Table.2. It shows clean speech gives a VR of 93.0% and EER of 7.0% for noisy speech VR of 60.0% and EER of 40.0% using CSM.

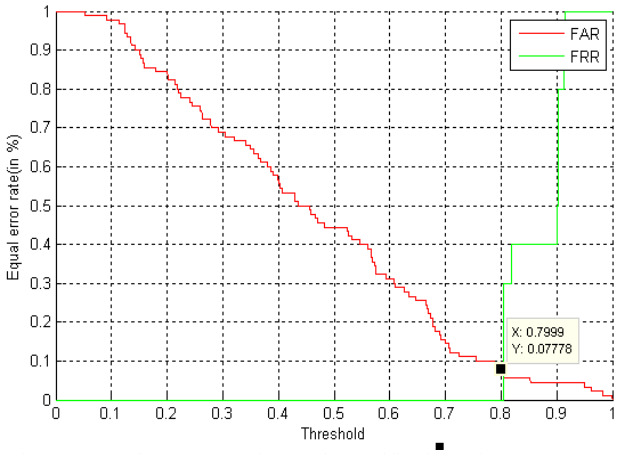

Fig. 5 FAR and FRR curves for speaker verification using RASTA-PLP and AAN for clean speech in CSM

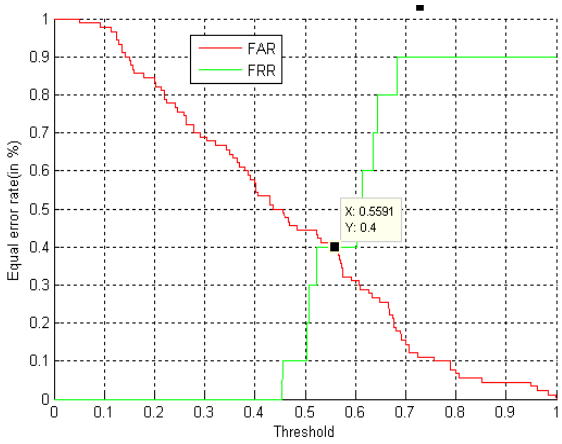

Fig. 6 FAR and FRR curves for speaker verification using RASTA-PLP and AAN for clean speech in CSM

TABLE II PERFORMANCE OF SPEAKER VERIFICTAION USING RASTA-PLP FEATURES AND AANN FOR CLEAN AND NOISY SPEECH IN CSM

Environment	VR (%)	EER (%)
Clean	93.0	7.0
Noisy	60.0	40.0

B. Performance of Speaker Verification using TM

The 13 RASTA-PLP feature vectors extracted from the clean and noisy speech and AANN is used to evaluate the performance of feature vectors extracted using TM.

An EER of 4.0% obtained by evaluating the performance in terms of FAR and FRR at threshold 0.90 and as shown in Fig.7 for clean speech using TM. For noisy speech, an EER of 10.0% attained by evaluating the performance in terms of FAR and FRR at threshold 0.72 and as shown in Fig.8.

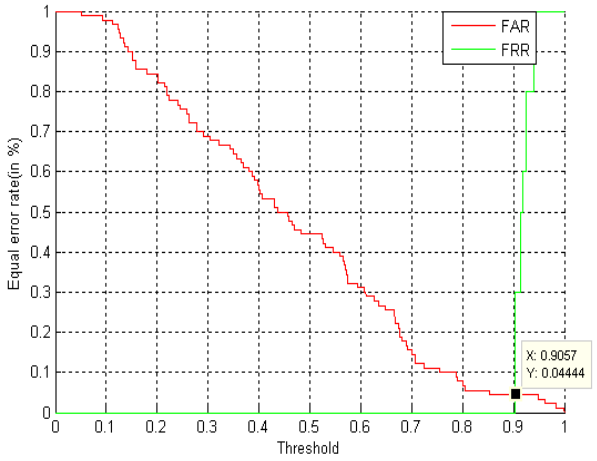

Fig.7 FAR and FRR curves for speaker verification using RASTA-PLP and AANN for clean speech in TM

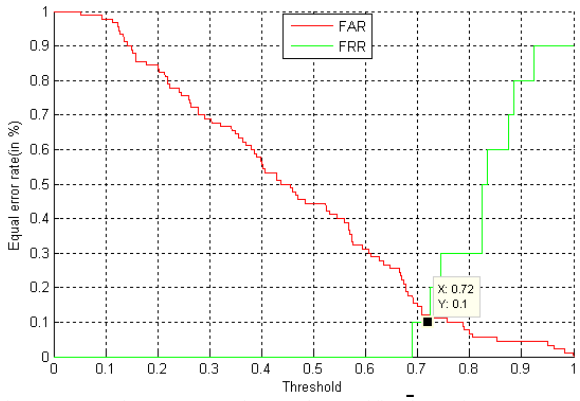

Fig. 8 FAR and FRR curves for speaker verification using RASTA-PLP and AANN for noisy speech in TM.

The performance of the speaker using RASTA-PLP features is evaluated in terms of accuracy or Verification Rate(VR) and EER. The performance of text independent speaker verification using AANN model for both clean and noisy speech is given in Table.3. It shows clean speech gives a VR of 96.0% and EER of 4.0% for noisy speech VR of 90.0% and EER of 10.0% using TM.

TABLE III PERFORMANCE OF SPEAKER VERIFICTAION USING RASTA-PLP FEATURES AND AANN FOR CLEAN AND NOISY SPEECH IN TM

Environment	VR (%)	EER (%)
Clean	96.0	4.0
Noisy	90.0	10.0

C. Combined model of CSM and TM for Clean Speech

CSM and TM are devices are combined because of complementary nature. The two devices are combined at score level using

$$c = ws_1 + (1 - w)s_2$$

where s_1 and s_2 are the scores or output of the model for AANN with RASTA-PLP features respectively and w is the

weight, $0 \leq w \leq 1$. In this work it is observed that for the weight 0.2 (w=0.2) an EER about 1.0% is achieved using AANN by combining the devices. From the evidence of CSM and TM devices an overall verification performance 99% is obtained. The verification performance of the proposed system is higher than the individual system because of complementary nature of CSM and TM. It is revealed that an EER of about 1.0% for the combined devices by evaluating the FAR and FRR values as depicted in Fig.9 and the overall verification of 99% is obtained in Table.4. The consolidated speaker verification performance for clean speech as shown in Table.5.

TABLE IV SPEAKER VERIFICTAION PERFORMANCE USING RASTA-PLP WITH AANN FOR CLEAN SPEECH

Environment	VR(%)	EER(%)
Clean	99.0	1.0

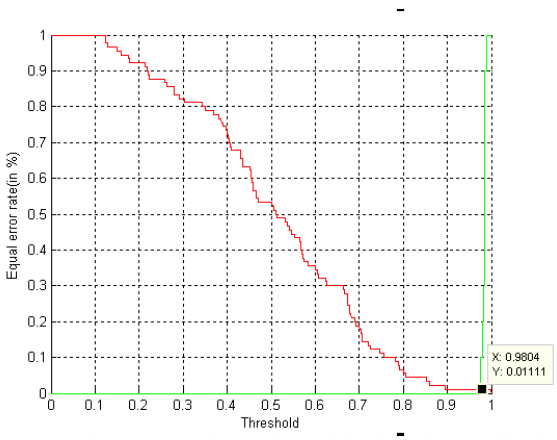

Fig. 9 FAR and FRR curves for speaker verification using combined CSM and TM for RASTA-PLP with AANN in clean speech

TABLE V PERFORMANCE OF SPEAKER VERIFICATION USING COMBINED CSM AND TM FOR RASTA-PLP FEATURES WITH AANN IN CLEAN

Measures	CSM	TM	CSM+TM
VR	93.0	96.0	99.0
EER	7.0	4.0	1.0

V. CONCLUSION

In this paper, we have proposed a speaker verification system in clean using combined CSM and TM. RASTA-PLP features are computed from the voice signal and is modeled by using AANN in order to confirm the speakers from clean and noisy surroundings. The features of text-independent speech data are collected from clean and noisy surroundings using CSM and TM. The performance of the proposed system using close-speaking microphone data degrades as the background noise increase, whereas the performance of the proposed system using throat microphone data not affected by the background noise.

The performance of text-independent speaker verification system using RASTA-PLP with AANN model in clean speech gives a VR of 93.0% and EER of 7.0% for noisy speech VR of 60.0% and EER of 40.0% using CSM. In TM, the performance of text independent speaker verification using AANN model with RASTA-PLP in clean speech gives a VR of 96.0% and EER of 4.0% for noisy speech VR of 90.0% and EER of 10.% using TM.

The verification performance of the combined system is increased than individual system due to complementary nature of CSM and TM. It is observed that an EER of about 1.0% for the combined devices (CSM+TM) by evaluating the FAR and FRR values and the overall verification of 99% is obtained in clean speech. In future, throat microphone can be used to analyze the performance of speech impaired people. Various acoustic features can be analyzed and the performance of different pattern recognition techniques can be studied.

REFERENCES

[1] Yuvan Yujin, Zhao Peihua and Zhou Qun, "Research of speaker recognition based on combination of LPCC and MFCC", *In: IEEE*, 2010.

[2] D. O Shaughnessy, *Speech Communications A Human and Machine*, Universities Press (India) Limited, 2001.

[3] Ravi P. Ramachandran, Kevin R. Farrell and Roopashri Ramachandran, "Speaker recognition - general classifier approaches and data fusion methods," *Pattern Recognition*, Vol. 35, pp. 2801–2821, December 2002.

[4] Anuradha S. Nigade and J. S. Chitode, "Throat Microphone Signals for Isolated Word Recognition Using LPC", *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 2, No. 8, August 2012.

[5] A. Shahina, B. Yegnanarayanan and M.R Kesheorey, "Throat microphone signal for speaker recognition," in *Proc. Int. Conf. Spoken Language Processing*, 2004.

[6] Li Zhu and Qing Yang, "Speaker Recognition System Based On weighted feature parameter", International conference on solid state devices and materials science, pp. 1515-1522, 2012.

[7] H. Hermansky, "Perceptual linear predictive (plp) analysis for speech," *J. Acoustic Soc. Am.*, pp. 1738–1752, 1990.

[8] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Trans. On Speech and Audio Processing*, Vol. 2, pp. 578–589, 1994.

[9] Luigi Galotto, J.O.P. Pinto, L.C. Leite, L.E.B da Silva and B.K. Bose, "Evaluation of the auto-associative neural network based sensor compensation in drive sytems," *IEEE Industry Applications Society Annual Meeting*, pp. 1–6, October 2008.

[10] P. Dhanalakshmi, S. Palanivel and V. Ramalingam, "Classification of audio signals using aann and gmm," *Applied Soft Computing*, Vol. 11, No. 10, pp. 716–723, January 2011.

[11] S. Jothilakshmi, "Spoken keywords detection using autoassociative neural networks," *Springer-International Journal of Speech Technology*, pp. 83–89, August 2014.

[12] Sondhi Benesty and Huang, *Text-dependent Speaker Recognition*, 2008.