Imputations of Hostile Conditions in Automatic Speaker Recognition Performance

J.V. Thomas Abraham, A. Nayeemulla Khan

Abstract*: Automatic Speaker Recognition (ASR) is a process in which the person is identified or the claim made by the person is verified. In the last three or four decades lot of researches have been done in this field and it has evolved a lot over these period. But in a real world scenario, performances of these speaker recognition systems have failed in hostile conditions. Building a robust speaker recognition system is a challenging task and should address all types of distortions. In this paper, the performance of a speaker recognition system in hostile conditions is analysed and presented. Especially how the environmental noise imputes the speaker recognition system is studied using the MSR Identity toolbox. Test was conducted with clean speech signals and noisy speech signals at various SNRs. The outcome of the test clearly indicates that the accuracy of the ASRs is degraded in hostile conditions. The results may be used to come up with more robust ASR systems.

Keywords: *Speaker Identification/Verification, MFCC, GMM-UBM, i-Vectors, Noise speech, robust speaker recognition*

I. INTRODUCTION

An Automatic Speaker Recognition System (ASR) recognizes the person from his/her speech signals. This system can be used as one of the biometric authentication system. A Speaker Recognition (SR) System can be an identification system or verification system. Speaker Identification (SI) is finding out the person to whom the unknown voice belongs to [3]. On the other hand, a speaker verification system (SV) is accepting or rejecting the claim made by a person that unknown the voice is his/her voice [9]. Either the identification or the verification system can be from an open set or closed set of speakers. In the closed set, the speakers will be from known population whereas in an open set, the speaker can be from unknown population as well.

A typical speaker recognition system has two stages: Enrollment and Testing stage. In an enrollment stage, a speaker's voice samples are recorded and transformed into set of features that represents the voice signal. These feature set is then modelled describing the speakers' identity. In testing phase, the same set of features are determined for an unknown voice signal and then a similarity measure is calculated between the unknown voice with the models that are already created in enrolment phase.

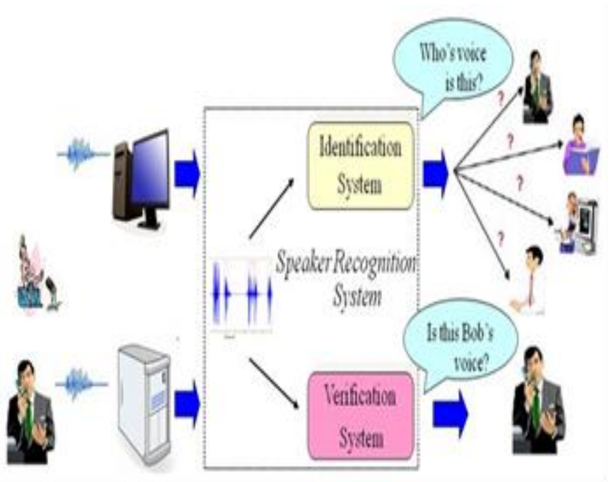


Figure 1: Speaker Recognition System

The recognition system produces good result in situation where the enrolment and testing environment are same and the voice signals are not changed much. But in a real time scenario, this is not possible because the voice signal of a person may vary in enrolment and testing phase due to several reasons like aging, illness etc. Or background may change between the two phases or the voice capturing device characteristics may vary. One or more of these intra- or inter-person changes drastically affect the speaker recognition system's performance [2][5]. So in this paper, we have explored how the hostile conditions in environment impute the recognition systems' performance. An overview of a baseline system in explained in section 2 and the experiment conducted for a baseline system in a clean and noisy environment are described in section 3. Section 4 states the recommended system for a robust system.

II. OVERVIEW OF A BASELINE SYSTEM

The baseline system captures the Mel Frequency Cepstral Coefficients (MFCC) features of a voice signal and creates a Gaussian mixture model (GMM) for each speaker [1][6]. Though other features like Linear Predictive Cepstral Coefficients (LPCC), Perceptual Linear Prediction (PLP) coefficients are used in SR systems, experiments show that MFCC performs better than other features. Once the MFCC features are extracted, a speaker specific GMM is created for each speaker. The rest of this section explains the MFCC feature extraction and GMM modelling processes.

A voice signal is quasi-stationary signal, meaning that its characteristics vary from time to time.

Revised Manuscript Received on May 06, 2019

J.V. Thomas Abraham, School of Computing Science and Engineering, VIT University Chennai Campus, Chennai, India.
A. Nayeemulla Khan, School of Computing Science and Engineering, VIT University Chennai Campus, Chennai, India.

So it is difficult to perform a signal analysis over the entire voice signal. But the characteristics of speech signal are stationary over a small period of time and hence it would be better to extract the features from a smaller period [4]. In order to extract the MFCC features, the voice signal is first broken down into small interval overlapping segments, called frames. A frame length can be from 20ms to 30ms, with an overlapping period of 10ms. Before splitting the signal into segments, the speech signal is pre-emphasized to increase the amplitude of the signal at higher frequency. In order to have a smooth end at both sides of the frame, a windowing function is applied and hamming window is widely used windowing function in a speaker recognition system.

After this, the windowed signal is given as input to the Discrete Fourier Transform (DFT) to extract the spectral information for a discrete frequency bands at discrete time. Applying DFT to the windowed signal produces a complex number that represents the magnitude and phase of that frequency component in the original signal.

The next stage is warping the signal to a mel scale and this mapping between frequency in Hertz and the mel scale is linear below 1000 Hz and logarithmic above 1000 Hz. While calculating MFCC features, this is implemented by creating a bank of filters which collect energy from each frequency band, with 10 filters spaced linearly below 1000 Hz, and the remaining filters spread logarithmically above 1000 Hz. Then the logarithmic value of each mel spectrum is determined.

Finally, the cepstrum is calculated by taking inverse DFT on the above signal.

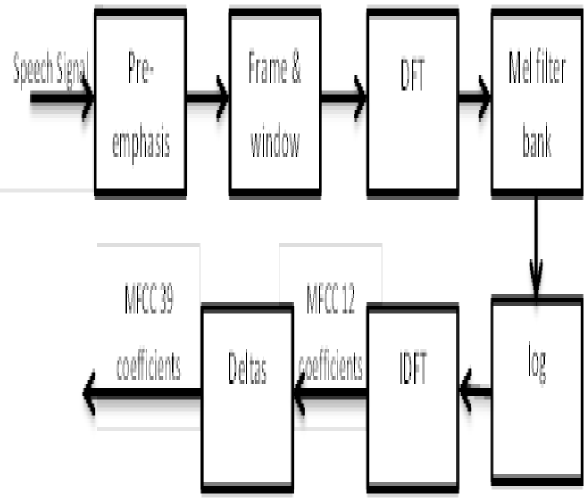


Figure 2: MFCC Feature Extraction

A Gaussian Mixture Model (GMM) was created using this 39 dimension feature vector. First, an Universal Background Model (UBM) was created from the set of speakers who were not part of the test group. Then for each test speaker, the GMM model was created by adapting the individual models to the UBM model. This will result in 'n' GMM models, one for each speaker. An unknown speech utterance is then compared with the existing models to find the best match.

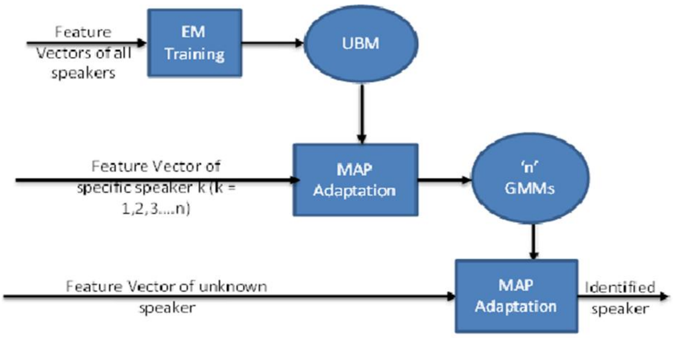

Feature Vectors of all speakers
EM Training
UBM
Feature Vector of specific speaker k (k = 1,2,3....n)
MAP Adaptation
'n' GMMs
Feature Vector of unknown speaker
MAP Adaptation
Identified speaker

Figure 3 : GMM-UBM Framework

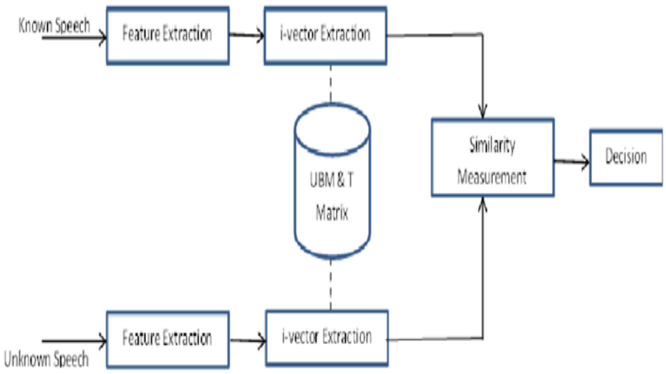


Figure 4: iVector Framework

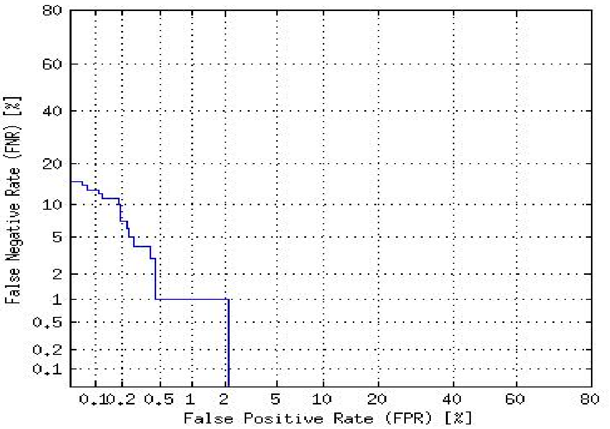
III. Experiment & Results

The experiment was conducted using MSR toolkit and TIMIT dataset which contains clean speech. The dataset contains 630 speakers and 10 speech utterances for each speaker. In this 530 speakers (530 * 10 = 5300 utterances) were taken for background model training and 100 speakers were taken for testing. For speaker specific model, 9 out of 10 speech utterances were used and remaining 1 utterance was used for testing. Verification trails were conducted for all model-test combinations, resulting in 10000 trials (100 target and 9900 imposter).

Experiments were conducted in three conditions: Firstly clean speech was used for speaker recognition. Then, secondly, the clean speech was distorted by adding some Gaussian white noise to clean speech. Finally, the distorted noise is enhanced by spectral subtraction method [8]. In all three conditions, we performed speaker recognition using MFCC and GMM modelling and repeated with the state of art i-vector modelling.

The results are given below and one can clearly understand that, the speaker verification system performs well under clean environment and recognition rate is also high. But in a real world situation this is highly impossible and the distorted speech signal influences the SR rate. Hence we need to enhance the speech signal and a robust SR system that can improve the recognition rate.

Figure 5 DET curve for GMM-UBM system with noisy



speech

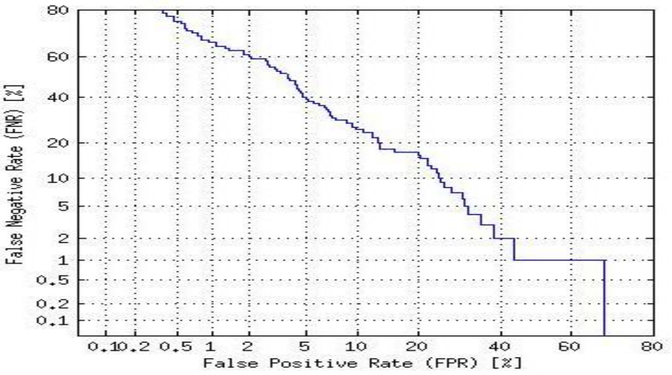


Figure 1 DET curve for GMM-UBM system with clean speech

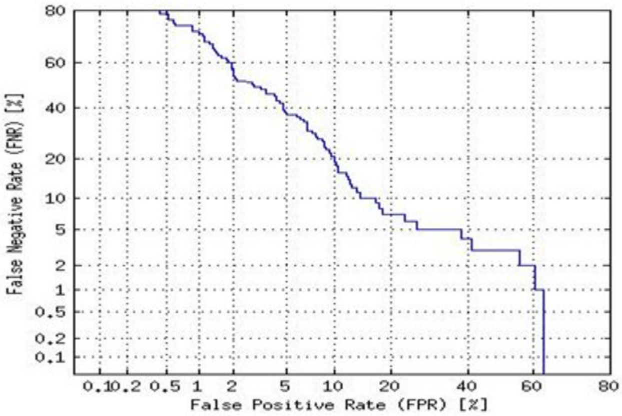


Figure 2 DET curve for GMM_UBM system with enhanced speech

Table 1 Clean Speech Signal Performance Measures

GMM (no of components)	EER%	DCF08	DCF10
256	1.1717	0.63	0.0341
512	0.6869	0.45	0.025
1024	0.3131	0.29	0.023

GMM (no of components)	EER%	DCF08	DCF10
256	15.1818	7.41	0.093
512	14.2222	4.51	0.092
1024	17.1313	7.59	0.095

Table 2 Noisy Speech Signal Performance Measures

GMM (no of components)	EER%	DCF08	DCF10
256	14.8384	6.9200	0.0970
512	15.9697	6.9100	0.0950
1024	13.0000	6.8200	0.0910

Table 2 Enhanced Speech Signal Performance Measures

Data/Model	GMM-UBM	i-vector-PLDA
Clean Speech	1.1717	0.4242
Noisy Speech	52.00	49.1818

Table 3 EER for GMM and i-Vector models

IV. CONCLUSION

In this work, we have compared the MFCC feature with GMM-UBM model and state-of-art i-vector with PLDA model under three different data sets. The results were impressive for a clean speech but adverse noisy mismatch conditions can cause major drop in performance. Developing a speaker recognition system with speech signals obtained within lab environment is completely different from developing a system in hostile conditions to get good performance results. Only when tested in hostile conditions, robust speaker recognition systems can be used in daily applications.

REFERENCES

1. Reynolds, Douglas A., Quatieri, Thomas F., and Dunn, Robert B., Speaker Verification Using Adapted Gaussian Mixture Models, Digital Signal Processing 10 (2000), 19–41.
2. Assmann, P., Summerfield, A., 2004. The perception of speech under adverse conditions. In: Greenberg, S., Ainsworth, W.A., Popper, A.N., Fay, R.R. (Eds.), Speech Processing in the Auditory System. Springer-Verlag, New York.
3. Li, Q., Huang, Y., 2010. Robust speaker identification using an auditory based feature. In: Proc. IEEE ICASSP. Dallas, TX, pp. 4514–4517.
4. Sadjadi, S.O., Hasan, T., Hansen, J.H.L., 2012. Mean Hilbert envelope coefficients (MHEC) for robust speaker recognition. In: Proc. INTERSPEECH. Portland, OR, pp. 1696–1699.

5. Mitra, V., McLaren, M., Franco, H., Graciarena, M., Scheffer, N., 2013. Modulation features for noise robust speaker identification. In: Proc. INTERSPEECH. Lyon, France, pp. 3703–3707.
6. Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi “voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) Techniques” Issue 3,March 2010, ISSN 2151-9617.
7. Md. Rashidul Hasan, Mustafa Jamil, Md. Golam Rabbani Md. Saifur Rahman, “Speaker identification using mel frequency cepstral coefficients” ICECE 2004, 28-30 December 2004, Dhaka, Bangladesh.
8. Boll, S., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoustics Speech Signal Process. 27 (2), 113–120.
9. Reynolds, D.A.: A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification. PhD thesis, Georgia Institute of Technology (1992).