MFCC VQ based Speaker Recognition and Its Accuracy Affecting Factors

Satyanand Singh
Associate Professor
Department of ECE
St Peter's Engineering
College, Dhoolapally,
Hyderabad.

Dr. E.G Rajan
Founder President and
Director Pentagram Research
Centre (P) Limited, 1073,
Road No. 44, Jubilee Hills,
Hyderabad.

ABSTRACT
The present study was conducted to evaluate the accuracy affecting factors of a Mel-Frequency Cepstral Coefficients (MFCC) and Vector Quantization (VQ) based speaker recognition system. This investigation analyses the factors that affecting recognition accuracy using speech signal from day to day life in surrounding environments. It was studied the mismatch affects of text-dependency, voice sample length, speaking language, speaking style, mimicry, the quality of microphone, utterance sample quality and surrounding noise. The corpuses of 10 people of 20 utterance subjects were collected which were indicate that any mismatch degrades recognition accuracy. It was found that most dominating factors that degrades the accuracy of speaker recognition systems were surrounding noise, quality of microphone by which voice sample were collected, disguise, and degrading of the sample rate and quality. Speech-related factors and sample length were less critical.

General Terms: Speaker recognition, Speaker recognition based ATM machine, Phone banking, Database services and Man machine interface.

Keywords: GF, Triangular Filter, Subbands, Correlation, MFCC, inverted MFCC, Vector Quantization

1. INTRODUCTION
A speaker recognition system mainly consists of two main modules, speaker specific feature extractor as a front end followed by a speaker modeling technique for generalized representation of extracted features [1, 2]. Since long time MFCC is considered as a reliable front end for a speaker recognition application because it has coefficients that represents audio, based on perception [3, 4]. In MFCC the frequency bands are positioned logarithmically (on the Mel-scale) which approximated the human auditory systems response more closely than the linear spaced frequency bands of FFT or DCT. This allows for better processing of data. Fig.1 shows the speaker recognition system used in this investigation. Accuracy of automatic speaker recognition is known to degrade severely when there is acoustic mismatch between the training and testing material [5, 6]. The mismatch can be due to the person himself (health, attitude, surrounding environment), due to technical reasons (microphone, transmission channel), or due to the recording environment (additive noise, echo, healthy or unhealthy environment). In this work, our main motivation is to gain more understanding on factors affecting MFCC and VQ based speaker recognition performance [7]. Over the years, MFCC modeled on the human auditory system has been used as a standard acoustic feature set for speech related applications. In this work we collected corpus of speakers in English and Telugu from the different location of Andhra Pradesh (India) and experimented. We studied the following parameters that were going to affect the efficiency of speaker recognition systems.

1.1 Technical Factors:
- Additive Noise
- Sampling Rate
- Quality of Microphone
- Distance to Microphone

1.2 Speaker Dependent Factors:
- Text Reading vs Spontaneous
- Deliberate Confusing

1.3 Voice Sample Related Factors:
- Text Dependent vs Text Independent
- Sample Length
- Speaking Languages

From the technical factors, we studied the effect of quality of microphone used in voice corpus collection and microphone mismatch, because the role of microphone has been systematically reported in literature to be one of the main factors that degrade the efficiency of speaker recognition systems. We also studied the effect of distance from the microphone. A close talking microphone is expected to be more accurate, but less user-convenient and about the optimum distance of microphone. We also studied the effect of sampling rate and additive environmental background noise where voice corpus has been collected for training and testing. From the speaker related factors, we studied whether the speech guidance is spontaneous or reading done by the person affects performance. In addition, we studied concealing outfit done by the speaker, i.e. speaker does not want to be recognized as himself by the system and deliberately wanted to changes natural speaking style. Regarding voice sample related factors, we studied whether the text content were important in applications, text-independent were more convenient for the speakers. The length of the sample was considered important, and we should confirm this in result. Regarding the speaking language, we studied if the speaker tries to speaks in his mother tongue language were identified better than in foreign language (English). We investigated if there were

difference between the models which were trained in native and non-native speech.

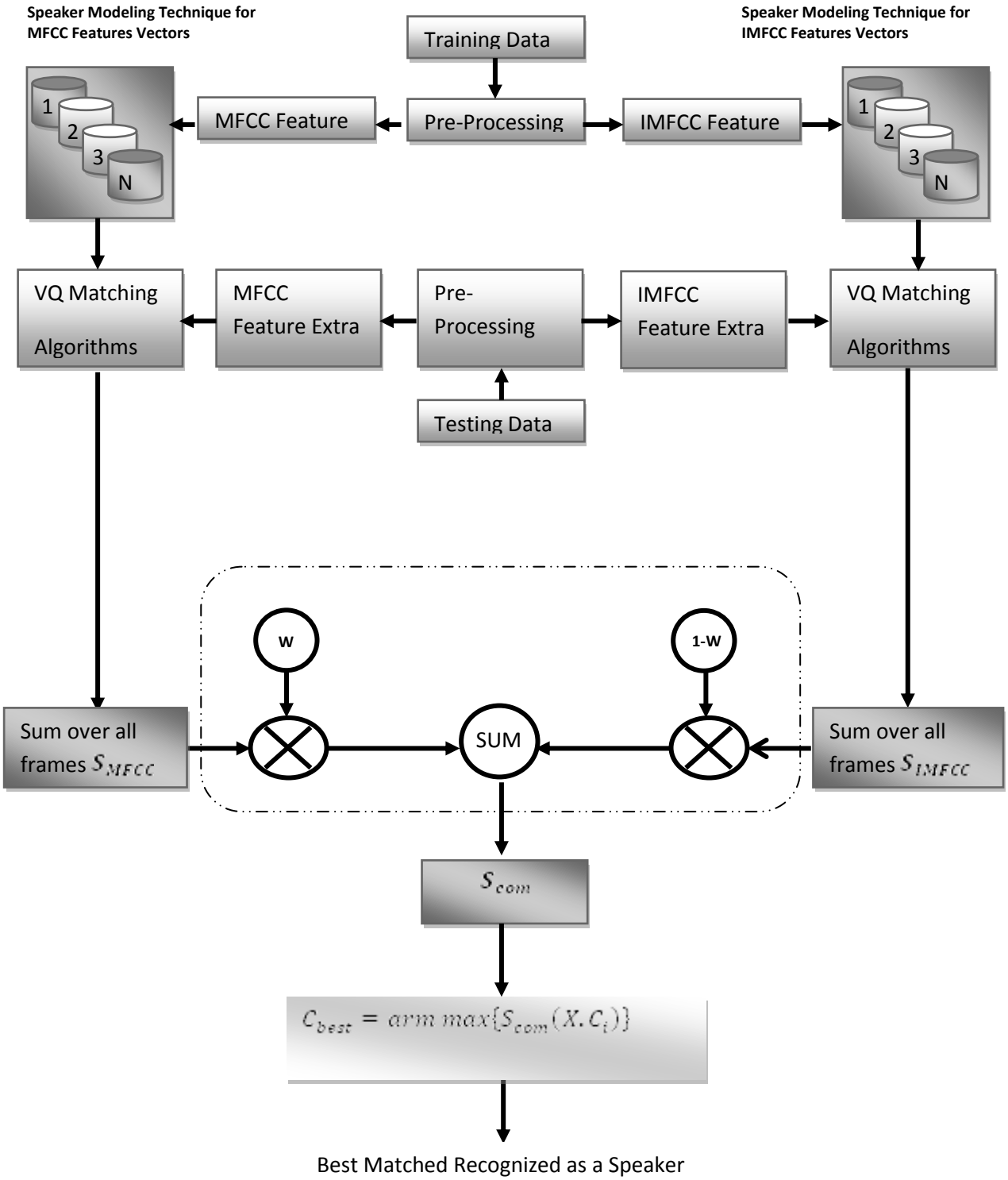

Fig 1: Speaker recognition system

2. TEST SETUP

2.1 Recording Apparatus

The HP Compaq dc 7800 integrated work centre stand (GN783AA) desktop has been used in the recording of speaker voices, with a 44.1KHz and 16 bits per second sampling rate. After recording voices of all speakers it has been stored as PCM encoded s1.wav, s2.wave…, sN.wav files. The recording volume was adjusted during the voice collection of speakers but the speaking style variation changes the loudness in many samples which has been recorded. The desktop has inbuilt sound cards with a Realtek High Definition Audio Codec and two different micro phones have been used to collect the voice corpus in this investigation. Microphones were used in recording were as follows:

- **R1**- Sony ECM-909A stereo microphone
- **R2**- Built-in microphone of the HP Compaq dc 7800

R1 was unidirectional, it has a noise-cancellation function, and the distance to speaker mouth was fixed to 3–4 cm (headset). R2 is omnidirectional and the distance may vary between 50–70 cm as per the comfort of speakers.

2.2 Voice Sample Collection Subject and Task

We recorded speech in an acoustic anechoic room at the DSP lab of St Peter's Engineering College, Hyderabad (India) in Electronics and Communication Engineering Department. This place was intended to provide as high signal-to-noise ratio as possible without excessive "pops" due to breath noises (some pops still occur in the recordings). The recording gain was kept constant across all recordings. We have recorded utterances for this investigation were at one sitting for each speaker. The text for the utterances was randomly selected by speaker. The main voice recordings consist of seven male and three female speakers of ten utterance of each using sampling rate of 8.0 kHz with 16 bits/sample. Voice sample were recorded at 3–4 cm away from R1 and at 50–70 cm away from R2. We also studied the effect of disguise as well as the spontaneous speech against text reading. Therefore, some subjects also completed additional speaking tasks where they were told to change their voice deliberately in order to be not recognized correctly, or speak spontaneously on an ordinary theme, such as the weather or personal feelings, what's your plan for today's evening.

In this investigation our main objective was to get the speaker familiar with the tasks, read the paper with the sentences, and answer the questions but not worry about the accent or the translation of the sentences. The speaker must concentrate more on the speech itself instead of the fact that he or she was recorded. For this reason, the spontaneous samples were recorded last, and the first 2 seconds of these samples are not used in the training and testing purpose.

2.3 Materials and Methods used in this Investigation

The speakers were asked to speak in English and in their native language. All speakers spoke the different sentences as per their comfort to English language. Sentences were chosen with a particular interest in the occurrence of common English phonemes. The 20 voice sample of all speaker's has been taken for babble, knock, traffic, television, tick R1, R2 and all voice sample related factors. Later we refer to "short" and "long" samples correspondingly. The spontaneous speech recordings were more than 90 seconds long. Speaker models were always trained from speech material consisting of the voice s1.wav, s2.wav…..sN.wav. We distinguished between text-dependent and text-independent utterances. All of the comprehensive recognition tests were based on text-independent sentences.

2.4 Voice Sample Preparation

After recording, we have prepared the samples for the test runs. Each sample is trimmed by removing silence from both ends of the sample, signal is down sampled, and finally noise is added. In this investigation we have used Microsoft sound recorder tool to remove the silence part of voice and SoX software for re-sampling and quantizing the files using to:

A-Quality: 44.1 kHz, 16 bits,

B-Quality: 22.05 kHz, 16 bits,

C-Quality: 8 kHz, 8 bits.

In this investigation, we recorded voice samples of five different types of noise using the A-quality, 6 seconds each: babble, knock, television, heavy traffic, and ticks. The babble noise simulates background talk. The knock is a repeated impulse every 0.740 s (knock on a wood desk). The television noise is loud television sound with a Hindi Serial on a lab corner. The heavy traffic noise is a sound of a vehicle passing by, a repeated pattern on a noisy background. Ticks consist of random knocks and ticks on a wood desk with a 3 dB cut down rain noise in the background. Each sample has low- and high-volume versions with 6 dB intensity difference. The low ones are mixed with the R2 samples and the high ones with the R1 samples. The average duration of the training samples was 6 seconds per speaker and out of twenty utterances one is used for training. For matching purposes remaining 19 voice corpus of the length 6 seconds, which was further divided into four different subsequences of the lengths 6 s (100%), 3 s (50%), 2s (33%) 1s (16 %) and 0.5s(8%) . Therefore, for 10 speakers we put 10X19X5 = 950 utterance under test and evaluated the efficiency. These speakers were speaking for the most-part English language, but a small number of speakers of other dialects of English and non-native speakers were included. The voice sample corrupted by different types of noises is shown in Fig.2.

2.5 Methodological Parameters Setup

In the investigation Gaussian Filter(GF) were used as the averaging bins instead of triangular for calculating MFCC as well as inverted MFCC in a typical speaker recognition application [8, 9]. There are three main inspiration of using GF. First inspiration is GF can provide much smoother transition from one subbands to other preserving most of the correlation

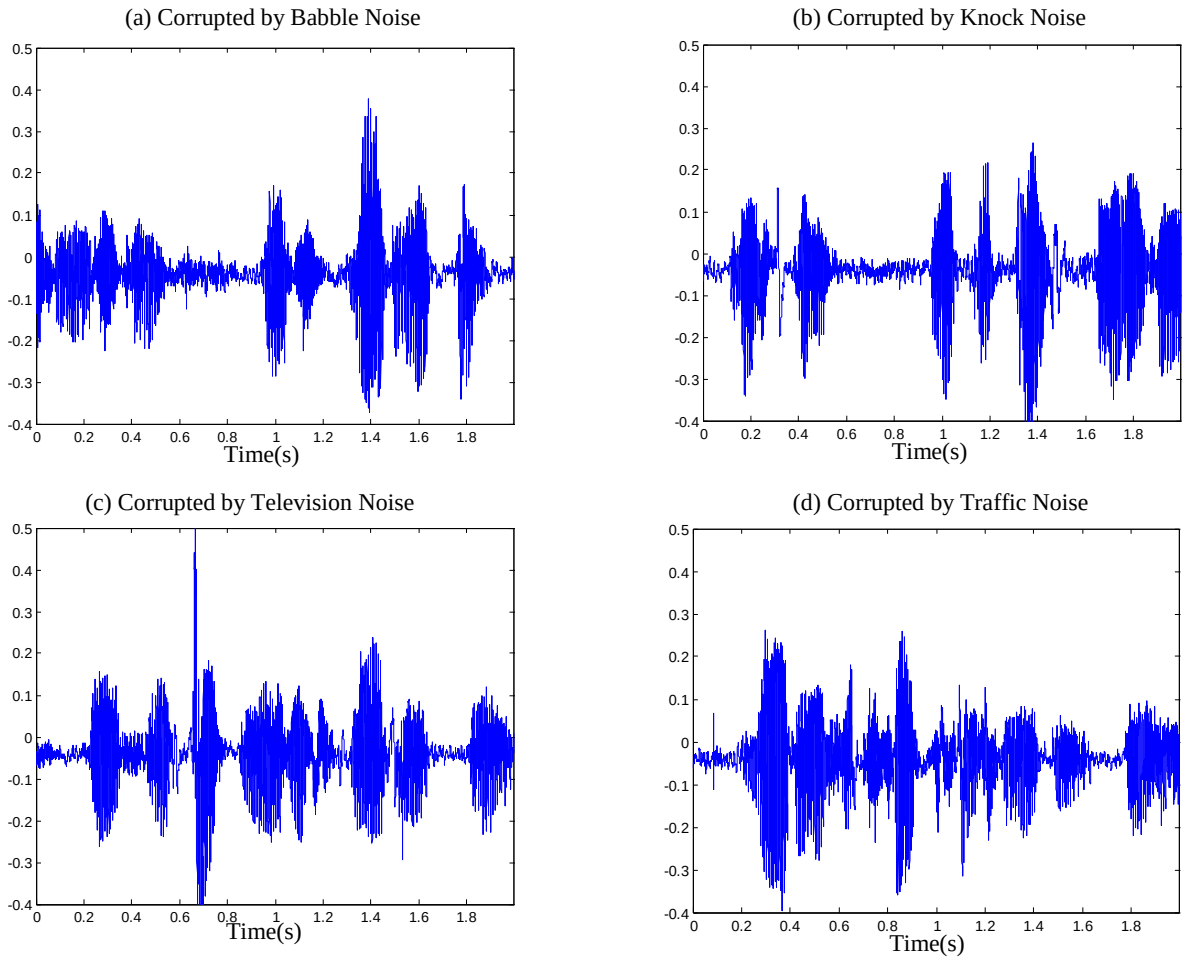

Fig 2: Same speaker's voice corrupted by different noises

between them. Second inspiring point is the means and variances of these GFs can be independently chosen in order to have control over the amount of overlap with neighboring subbands. Third inspiring point is the filter design parameters for GF can be calculated very easily from mid as well as end points located at the base of the Original TF used for MFCC and inverted MFCC. In this investigation both MFCC and inverted MFCC filter bank are realized using a moderate variance where a GF's coverage for a subbands and the correlation is to be balanced. Results show that GF based MFCC and inverted MFCC perform better than the conventional TF based MFCC and inverted MFCC individually. Results are also better when GF based MFCC & inverted MFCC is combined together. Their model scores in link to the results that are obtained by combining MFCC and inverted MFCC feature set realized using usual TF [10]. All the implementations have been done with VQ-Linde Buzo Gray (LBG) algorithm as speaker modeling paradigm [11]. According to psychophysical studies human perception of the frequency content of sounds follows a subjectively defined nonlinear scale called the Mel scale [12]. MFCC is the most commonly used acoustic features for speaker recognition. MFCC is the only acoustic approach that takes human perception (Physiology and behavioral aspects of the voice production organs) sensitivity with respect to frequencies into consideration, and therefore is best for speaker recognition [13].

3. RESULT AND DISCUSSION

3.1 Technical parameters that affecting the efficiency

In this investigation we found that the noise has the strongest effect in the performance of speaker recognition efficiency. We investigated closely into five different noise named as babble, knock, television, traffic and ticks types while using two different microphones R1 and R2. . Table.1 shows the summary of identification rate for different types of noises and table.2 shows the summary of identification rate for microphones. The television noise is having much more impact on the speaker recognition efficiency and babble noise were least impact on speaker recognition efficiency. The effect of microphone was not significant, except for the mismatch of clean sample training samples and recognition from samples contaminated by impulsive noise.

Table 1. Summary of identification rate for noises

Noise	No Of Utterances	Correct Identification	Efficiency %
Babble	950	915	96.31
Knock	950	910	95.78
Television	950	880	92.63
Traffic	950	889	93.57
Ticks	950	911	95.89

Table 2. Summary of identification rate for Microphones

Apparatus	No Of Utterances	Correct Identification	Efficiency %
R1	950	942	99.15
R2	950	936	98.52

3.2 Speaker Dependent Factors

Deliberate cheating was possible, recognition rates are 95–100 %. Error rates are similar when recognizing speakers from spontaneous speech, when the database is constructed from text reading. Recognition fails mostly, with or without noise mismatch.

3.3 Voice Sample Related Factors

The linguistic and data-related factors which we studied in this investigation were language mismatch between training and recognition, text dependence, and the length of the voice sample that were used in training and testing. The effects of voice sample related factors in recognition error rate listed in table 3, with varying training and recognition voice sample, text dependent and text independent sample. In this investigation it has been observed that the impact on performance of speaker recognition rate were not more dependent on the length of voice sample that was being used in training and testing. Fig. 3 shows that the overall performance affecting factors and its impact on efficiency of speaker recognition system and fig.4 shows the plot of different accuracy affecting factors of speaker recognition systems.

Table3. Summary of identification rate for voice sample related factors

Noise	No Of Utterances	Correct Identification	Efficiency %
Text - dependant	950	945	99.47
Text-Independent	950	930	97.89
Short sample	950	920	96.84
Long sample	950	948	99.78
Language mismatch	950	945	99.47

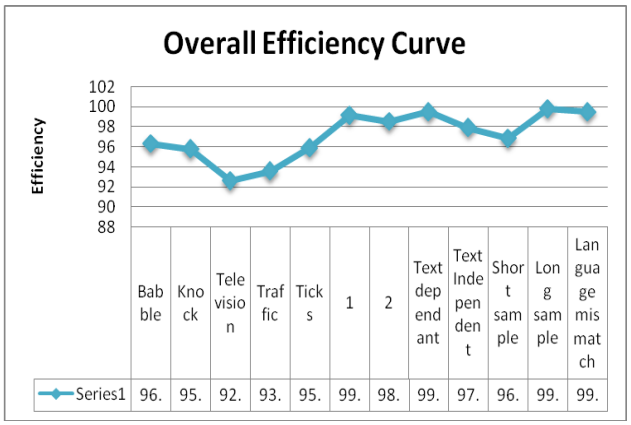

Fig 3: Performance affecting factors and efficiency

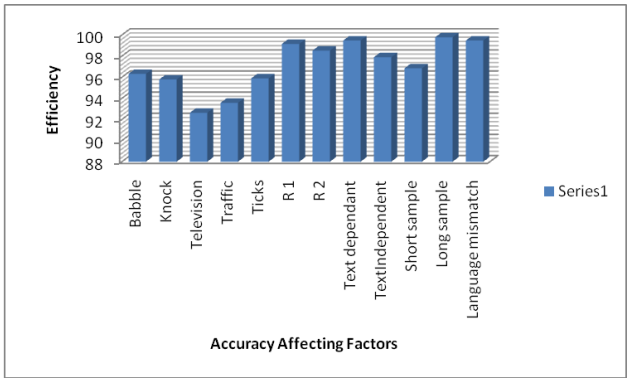

Fig 4: Plot of efficiency and accuracy affecting factors

4. SUMMARY

The parameters affecting performance of the speaker recognition are listed in descending order as follows:

1. Noise,
2. Quality and distance of microphones,
3. Disguise,
4. Quality of sample in training and testing,
5. Length of the voice sample that were used in training and testing,
6. Spoken language used in training and testing,
7. Text-dependency.

Mostly performance of the speaker recognitions are affected by many factors simultaneously. Computing factor specific effects would be misleading. The interpretation is in detailed that were driving the speaker recognition performance described below.

Noise- The background noise is the most significant factor for the speaker recognition accuracy, which is high for the clean samples but deteriorates quickly for noisy samples. Only babble noise has no significant influence. Results were better without mismatch.

Microphones- Results were best without mismatch the microphone quality itself is insignificant and it has not much more impact on the efficiency of speaker recognition systems.

Disguise- Deliberate cheating is possible, the recognition fails in most of the cases but it has no much more impact on speaker recognition systems.

Quality of the Voice Sample- For clean voice samples, the higher quality of microphones leads to the better results. However, ordinary quality of microphone gives almost the same results as good quality microphones. However, B quality gives almost the same results as A. For noisy samples, the C quality voice sample gives better performance, especially with R2.

Voice Sample Length- In general it was assumed that longer samples improve the speaker recognition efficiency but could not be verified in this investigation.. There is no significant difference to clean samples. In background noise the short samples provide better results but the difference is within the confidence level.

Language used in training and testing- There is no much more impact on the efficiency of speaker recognition in native language speech. For the noisy samples, the English language samples give better results.

Text-dependency- In this investigation we have used MFCC and VQ based speaker recognition and the role of text is insignificant.

5. CONCLUSION

The most important conclusion in this investigation was that the voice samples used in training and testing conditions should match. The most significant single factors that affect speaker recognition performance, among the tested ones, was the noise. When the training and recognition data contain different types of noise, the error rate is very high in most cases. When the noise conditions match, the error rate is systematically below 10 %. Speech and language factors are less important than technical factors but deliberate cheating makes an exception: cheating is possible.

6. REFERENCES

[1] Gatica-Perez, G. Lathoud, J.-M. Odobez and I. Mc Cowan. 2007 Audiovisual probabilistic tracking of multiple speakers in meetings, IEEE Transactions on Speech and Audio Processing, 15(2), pp. 601–616.

[2] J. P. Cambell, Jr. 1997 Speaker Recognition A Tutorial Proceedings of the IEEE, 85(9), pp. 1437-1462.

[3] Faundez-Zanuy M. and Monte-Moreno E. 2005 State-of-the-art in speaker recognition , Aerospace and Electronic Systems Magazine, IEEE, 20(5), pp. 7-12.

[4] K. Saeed and M. K. Nammous. 2005 Heuristic method of Arabic speech recognition, in Proc. IEEE 7th Int. Conf. DSPA, Moscow, Russia, pp. 528–530.

[5] Lamel, L.F. and Gauvain, J.L., 2000. Speaker Verification over the Telephone, Speech Communication, pp. 141–154.

[6] Ortega-Garcia, J., Gonz´alez-Rodriguez, J., et al., May 1998 AHUMADA: A large speech corpus in Spanish for speaker identification and verification, IEEE Intl. Conf. on Acoust. Speech and Signal Proc, pp. 773–776.

[7] Singh Satyanand, Dr. E.G Rajan. March 2011 Vector Quantization Approach for Speaker Recognition Using MFCC and Inverted MFCC, International Journal of Computer Applications,17(1), pp. 1-7 .

[8] Yegnanarayana B., Prasanna S.R.M., Zachariah J.M. and Gupta C. S. 2005 Combining evidence from source suprasegmental and spectral features for a fixed-text speaker verification system , IEEE Trans. Speech and Audio Processing, 13(4), pp. 575-582.

[9] J. Kittler, M. Hatef, R. Duin, J. Mataz. 1998 On combining classifiers, IEEE Trans, Pattern Anal. Mach. Intell, 20(3), pp. 226-239.

[10] He, J., Liu, L., Palm, G. 1999 A Discriminative Training Algorithm for VQ-based Speaker Identification , IEEE Transactions on Speech and Audio Processing, 7(3), pp. 353-356.

[11] Laurent Besacier and Jean-Francois Bonastre. 2000 Subband architecture for automatic speaker recognition, Signal Processing, 80, pp. 1245-1259.

[12] Ganchev, T., Fakotakis, N., and Kokkinakis, G. 2005 Comparative Evaluation of Various MFCC Implementations on the Speaker Verification Task, Proc. of SPECOM Patras, Greece, pp. 1191-194.

[13] Zheng F., Zhang, G. and Song, Z. 2001 Comparison of different implementations of MFCC, J. Computer Science & Technology 16(6), pp. 582-589.