Journal of Information Security, 2012, 3, 335-340
doi:10.4236/jis.2012.34041 Published Online October 2012 (http://www.SciRP.org/journal/jis)

Text Independent Automatic Speaker Recognition System Using Mel-Frequency Cepstrum Coefficient and Gaussian Mixture Models

Alfredo Maesa[1], Fabio Garzia[1,2], Michele Scarpiniti[1], Roberto Cusani[1]
[1]Department of Information, Electronics and Telecommunications Engineering, University of Rome, Rome, Italy
[2]Wessex Institute of Technology, Southampton, UK
Email: fabio.garzia@uniroma1.it

Received July 21, 2012; revised August 14, 2012; accepted September 3, 2012

ABSTRACT

The aim of this paper is to show the accuracy and time results of a text independent automatic speaker recognition (ASR) system, based on Mel-Frequency Cepstrum Coefficients (MFCC) and Gaussian Mixture Models (GMM), in order to develop a security control access gate. 450 speakers were randomly extracted from the *Voxforge.org* audio database, their utterances have been improved using spectral subtraction, then MFCC were extracted and these coefficients were statistically analyzed by GMM in order to build each profile. For each speaker two different speech files were used: the first one to build the profile database, the second one to test the system performance. The accuracy achieved by the proposed approach is greater than 96% and the time spent for a single test run, implemented in Matlab language, is about 2 seconds on a common PC.

Keywords: Automatic Speaker Recognition; Access Control; Voice Recognition; Biometrics

1. Introduction

In last decades, an increasing interest in security systems has arisen. These systems are very useful since they allow managing security in a very efficient way, reducing the need of human resources. Most of them implement an access control system [1-4]. In particular, a huge number of research efforts were directed to speaker recognition problem [5-15]. In fact, many strategic places are of vital importance to the assessment of involved people. A simple way to verify people identity can consist in analyzing its voice. In fact, voice based recognition systems represent biometric systems that allow the access control in a very fast and low intrusive way, requesting a reduced collaboration of the people.

The human voice is peculiar to each person and this is due to the anatomical apparatus of phonation. The vocal tract consists of three main cavities: the oral cavity, the nasal cavity and the pharyngeal cavity [16]. The nasal cavity is essentially bony, hence static in time; furthermore it can be isolated through the soft palate. The oral cavity is formed by the bony structure of the palate and soft palate; its conformation can be altered significantly by the movement of the jaw, lips and tongue. The pharyngeal cavity extends to the bottom of the throat and it can be compressed retracting the base of the tongue towards of the wall of the pharynx. In the lower part it ends with the vocal cords: a couple of fleshy membranes traversed by the air coming from the lungs. During the production of a sound, the space between the membranes (glottis) can be completely opened or partially closed.

Due to the peculiarity of the voice formation apparatus, it can be possible to recognize a particular individual from its voice. In addition, this operation can be evaluated in an automatic approach [13-15]. In literature, this problem is addressed as Automatic Speaker Recognition (ASR) [17], and it is widely discussed by the research community [13-15].

Speaker recognition is classified as a hybrid biometric recognition approach, as it has two components: the physical one related to the anatomy of the vocal apparatus, and the behavioral component, pertinent to the mood of the speaker just in the recording moment [15].

There are several approach to ASR based on features, vector quantization, score normalization, pattern matching, etc., but the most of them are text dependent [6,7,9-11, 13,14].

In this paper, we propose text independent ASR system based on Mel-Frequency Cepstrum Coefficients (MFCC) [18,19] and Gaussian Mixture Models (GMM) [20-22]. Then the model parameters are estimated with the maximum similarity making use of the Expectation and Maximization (EM) algorithm [23,24]. The novel com-

bination of these two techniques, allows the system to reach high recognition rates and high operative velocities, as shown in the following, allowing to use the proposed system in real security context. In addition, unlike other works on ASR presented in literature, because the recorded speaker signal could be corrupted by environmental additive noise, a spectral subtraction algorithm [25,26] is also used. Comparisons with the state of the art demonstrate the effectiveness of the proposed approach in terms of accuracy rate.

The data acquisition can be performed through simple microphones which are well spread and their cost is negligible. However cheap instrumentation may be more affected by disturbances such as background noise and the spectral subtraction algorithm could be no more sufficient for efficient noise suppression.

The paper is organized as follows: Section 2 describes the ASR problem. Section 3 introduces the MFCC technique, while Section 4 introduces the GMM models. Section 5 describes the proposed ASR system and Section 6 shows some interesting experimental results. Finally Section 7 concludes the work.

2. System Description

A biometric recognition system generally consists of:

- A sensor which makes acquisition of data and its subsequent sampling: in the specific case the sensor is a microphone, possibly with a high Signal to Ratio (SNR) value. Since the input signal is essentially speech, the sampling rate is usually set to 8 kHz;
- A step of preprocessing that in the voice context is constituted by the signal cleaning: simply denoising algorithm can be applied to recorded data after a normalization procedure. In order to clean recorded speech signal from environmental additive noise, a spectral subtraction algorithm is used [23,24] in this paper;
- The extraction of the peculiar characteristics (feature extraction): in this stage Mel frequency cepstral coefficients are evaluated using a Mel filter bank after a transformation of the frequency axis in a logarithmic one;
- The generation of a specific template for each speaker: in this work we have decided to use the Gaussian Mixture Models (GMM) where model parameters are estimated with the maximum similarity making use of the Expectation and Maximization (EM) algorithm;
- In case of the user is registering (enrollment) for the first time to the system, this template will be added to the database, using some database programming techniques;
- Otherwise, in case of test among users already present in the database, a comparison (matcher) determines which profile matches the generated template of the test speech. The matcher utilizes a similarity test, obtaining by a ratio value that can be accepted if it is higher than a decision threshold.

The typical ASR system is shown in **Figure 1**.

The technologies used for the development of the biometric system are the MMFCC for the extraction of the characteristics and the GMM for the statistical analysis of the data obtained, for the templates generation and for the comparison.

3. Mel Frequency Cepstral Coefficient

The term "cepstrum" is a pun where the first letters of the term "spectrum" are reversed. It was described in 1963 by Bogert *et al*. [27]. Cepstrum is defined as the inverse Fourier transform of the logarithm of the spectrum of a signal [28,29]:

$$x_c(n) = DFT^{-1}\left\{\log\left|DFT\left\{x(n)\right\}\right|\right\} \tag{1}$$

The cepstrum transform the signal from the frequency domain into the quefrency domain.

When cepstrum is applied to the voice, its strength is to be able to divide excitation and transfer function. In a signal $y(n)$ based on the source-filter model, in this specific context, respectively the vocal cords and the vocal tract, cepstrum allows separation in $y(n) = x(n) * h(n)$, where the source $x(n)$ passes through a filter described by the impulse response $h(n)$. The spectrum of $y(n)$ obtained by the Fourier transform is $Y(k) = X(k)\,H(k)$ where k index of discrete frequencies, *i.e.* the product of two spectra, respectively the source and the filter one. Separating these two spectra is complicated. On the contrary, it is possible to separate the real envelope of the filter from the remaining spectrum by formulating all the phase at the beginning. The cepstrum is based on the properties of the logarithm that can transform the product of the argument in sums of logarithms.

Starting from the logarithm of the modulus of the spectrum:

$$\begin{aligned}&\log\left|Y(k)\right| \\ &= \log\left(\left|X(k)H(k)\right|\right) = \log\left(X(k)\right) + \log\left(H(k)\right)\end{aligned} \tag{2}$$

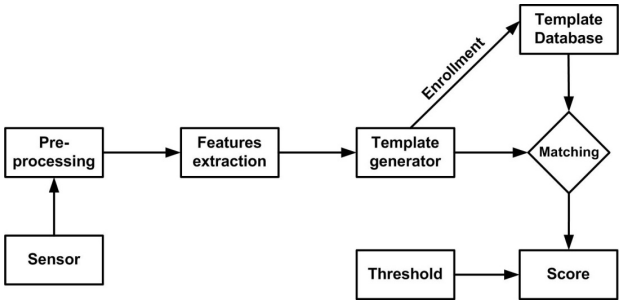

Figure 1. A typical ASR system.

it is possible to separate the fast oscillating component from the slow one, respectively by means of a high and low pass filter, obtaining:

$$\begin{aligned} c(n) &= DFT^{-1}\left(\log\left|Y(k)\right|\right) \\ &= DFT^{-1}\left(\log\left|X(k)\right|\right) + DFT^{-1}\left(\log\left|H(k)\right|\right) \end{aligned} \tag{3}$$

that is the signal cepstrum in the quefrency domain. In the low quefrencies are described the transfer function information, in the high quefrencies there is data about excitation.

Hence the initial wave of percussion created by the vocal cords and shaped by the throat, nose and mouth can be analyzed as a sum of a source function (given by the excitation of the vocal cords) and a filter (throat, nose, mouth). The separation between high and low quefrency, can be obtained by a high pass lifter (filter) for the fast oscillation and a low pass lifter for the slow one.

Psychoacoustic studies [30-32] have shown that the mind perception of the frequency content of the sound follows a nearly logarithmic scale, the Mel scale, which is linear up to 1 kHz and logarithmic thereafter:

$$\text{mel}(f) = \begin{cases} f \text{ if } f \le 1\,\text{kHz} \\ 2595\log\left(1+\dfrac{f}{7000}\right) \text{ if } f > 1\,\text{kHz} \end{cases}$$

The Mel scale is shown in **Figure 2**, where it is clear the compression of the Mel scale (reported in *y*-axis) with respect the Hertz scale (in *x*-axis) for frequencies greater than 1 kHz. In this scale pitches are judged by listeners to be equal in distance from one another.

Mel-cepstrum estimates the spectral envelope of the output of the filter bank. Let Y_n represent the logarithm of the output energy from channel n, applying the discrete cosine transform (DCT) we obtain the cepstral coefficients MFCC through the equation:

$$c_k = \sum_{n=1}^{N} Y_n \cos\left[k\left(n-\frac{1}{2}\right)\frac{\pi}{N}\right] \forall k = 0,\ \cdots,\ K \tag{4}$$

The simplified spectral envelope is rebuilt with the first K_m coefficients, with $K_m < K$:

$$C(\text{mel}) = \sum_{k=1}^{K_m} c_k \cos\left(2\pi k \frac{\text{mel}}{B_m}\right) \tag{5}$$

where B_m is the bandwidth analyzed in Mel domain and $K_m = 20$ is a typical value assumed by K_m. c_0 is the mean value in d*B* of the energy of the filter bank channels, hence it is in direct relation with the energy of the sound and it can be used for the estimation of the energy.

Schematically, the coefficients are derived in the following way: the spectrum of the original signal is computed with the Fourier transform; the obtained spectrum is mapped in Mel making use of appropriate overlapping windows; for each obtained function the logarithm is calculated; the discrete cosine transform is calculated (DCT); the coefficients are the amplitudes of the resulting spectrum. In order to emphasize the low quefrencies DCT is chosen.

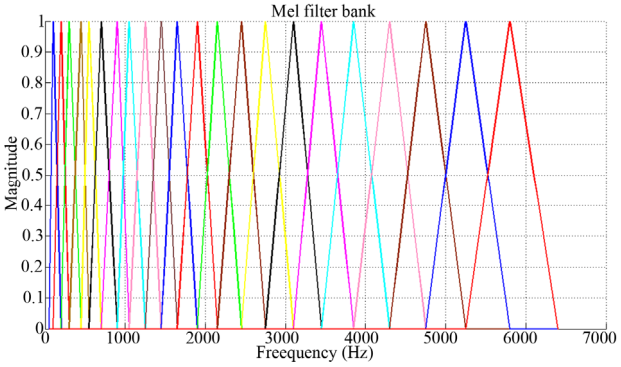

Figure 2. Mel filter bank.

4. Gaussian Mixture Model

Each arbitrary probability density function (pdf) can be approximated by a linear combination of unimodal Gaussian density [20]. Under this assumption, Gaussian mixture models have been applied to model the distribution of a sequence of vectors $X = x_1, x_2, \cdots, x_t, \cdots, X_T$ each one of dimension *D*, containing data on the characteristics extracted from the voice of a subject, according to:

$$p(x_t|\lambda) = \sum_{i=1}^{M} w_i\, p_i(x_t) \tag{6}$$

$$p(X_t|\lambda) = \prod_{t=1}^{T} p(s > t|\theta) \tag{7}$$

where w_i are the weights of the corresponding mixtures to the unimodal Gaussian densities p_i with $i = 1, \cdots, M$ and:

$$p_i(x_t) = \left(\frac{1}{\sqrt[D]{2\pi}\sqrt{\det(\Sigma_i)}}\right)^{-\frac{1}{2}\left((x_t-\mu_i)^T \Sigma^{-1}(x_t-\mu_i)\right)} \tag{8}$$

The weights of the mixtures satisfy the constraint:

$$\sum_{i=1}^{M} w_i = \tag{9}$$

Each speaker is identified by a λ model obtained from GMM analysis. In particular lambda is defined as:

$$\lambda = \{w_i, \mu_i, \Sigma_i\} \tag{10}$$

where μ_i is the mean vector and Σ_i is the covariance matrix.

Given a characteristic vector sequence of the speaker to be identified, the model parameters are estimated with the

maximum similarity λ making use of the Expectation and Maximization algorithm [23,24]. The λ model is compared with a characteristic vector *X* by calculating the log-likelihood similarity [23]:

$$\log P(X|\lambda) = \sum_{T} \log P(x_t|\lambda) \quad (11)$$

In order to decide, it is utilized a similarity test obtained by the following ratio:

$$\frac{P(X|\text{ Speaker})}{P(X|\text{ Other Speaker})} > \sigma \quad (12)$$

where σ is the dec on the contrary, a collection of models of different speakers. The final score of a certain subject S_c over an *X* vector containing the voice features of the test is given by:

$$\log L(x) = \log p(X|S = S_c) - \log \sum_{S \in \text{pop}} p(X|S \neq S_c) \quad (13)$$

where $L(X)$ represents the similarity value of *X* vector with respect to S_c compared with the characteristics of other individuals in the database (pop), excluding the one taken into account.

5. System Implementation

In the pre-processing phase, the signal has been improved using spectral subtraction [33,34] and segmented into frames partially overlying (50%) and relatively small. Frames not containing voice were skipped. The size of each frame is less than 20 ms in order to make the contained wave stationary. Each frame has been subjected to the Hamming window to minimize the discontinuities at the edges of the frame. For each frame 20 MFCC were calculated. The obtained data represents the characteristics of a speaker. This information, organized in a matrix containing a vector of Mel-Cepstral coefficients for each frame, is analyzed by the GMM using 32 mixtures. The result is a set of statistical data characterized by a mean vector, a covariance matrix and a weight vector which constitute the template itself. The template is employed when a speaker is added into the system or for the test step among the users already registered.

The public voice database *Voxforge.org* [35] was used in order to validate the system. *Voxforge* is an internet community including researchers and "donors" of human voice. The preset aims are to support who intends to realize and test an automatic speaker recognition system, a speech recognition engine, or any application related to analysis, to the recognition and more generally to the study of the human voice. Anyone can register on the website and send his own voice recordings to be made available to the whole community. For this study 450 speaker utterances were randomly extracted from Voxforge website. For each speaker two speeches were employed: the first one in order to perform the training phase and the second one to test the system. Since the recognition system is text independent, each speech contains different words (typically reads paragraphs of popular books).

In the training step each template generated from the analysis of the speakers' utterances is stored into the system. This set of information represents the knowledge base of the system obtained in the training phase. The test stage was made utilizing the test templates of each speaker compared to the whole knowledge base of the system, *i.e*. all the templates stored in the training phase. This comparison was performed using the criterion of log-likelihood previously described. The output of the test phase is a matrix containing the similarity estimation of each test with respect to each profile stored in the system. This matrix is structured in this way: the rows represent the *i*th test and the column the *j*-training. Hence in position (*i*,*j*) is contained the value representing the similarity likelihood of test speaker *i* with respect to training speaker *j*. Since the comparison is made by log-likelihood, for each row (test) the system nominates the column (speaker in the system) containing the maximum value as the owner of the speech.

6. Experimental Results

As shown in **Table 1** there were 433 identifications on 450 subjects, this means that accuracy rate is 96.22%. Since the system creates a hierarchy of candidates owners of each test, if the top five were accepted as good results, it would be achieved a recognition rate of 97.78%.

With regard to temporal performances, it should be taken into account that the complete computation test involves the training data processing, the test data elaboration and the comparison from training and test data. Obviously it is also possible perform a single test and compare it to profiles in the system. These performance results in terms of time required, are specific to the database used, since the system developed can run with audio files containing variable size, speech length and sampling. The temporal performances are exposed in **Table 2**.

7. Comparison with the State of Art

This section discusses about the main speaker recognition systems found in scientific literature. In 1995 Reynolds [36] implemented an identification system based on spectral variability obtaining a 96.80% accuracy rate with 49 speakers. In 2009 Revathi, Ganapathy and Venkataramani [37] through an iterative clustering ap-

Table 1. Accuracy performances.

	Speakers	Hit	Accuracy
1st	450	433	96.22%
In top 5	450	440	97.78%

proach, PLP (Perceptual Linear Predictive cepstrum) and MF-PLP (Mel Frequency PLP) achieved 91% accuracy rate with 50 speakers randomly chosen from TIMIT database [38]. In 2009 Chakroborty and Saha [39] combining MFCC and IMFCC (Inverted MFCC) based on gaussian filter, reached 97.42% accuracy rate with 131 subjects of YOHO database [40]. In 2010 Saeidi, Mowlaee, Kinnunen and Zheng-Hua [41] through Kullback-Leibler divergence achieved 97% accuracy rate with 34 speakers. In 2011 Gomez [42] implemented an identification system based on novel parametric neural network, reaching 94% accuracy with 40 speakers. In 2011 Rao, Prasada and Nagesh [43] made a study comparing GMM, HMM (Hidden Markov Models) and MFCC. The accuracy rate obtained in best test condition was 99% on 200 subjects taken from TIMIT database.

Table 3 summarizes the accuracy rates reached by the previous approaches.

8. Conclusions

In this paper we have introduced an ASR system based on MFCC and GMM. The accuracy of the proposed system is greater than 96% and with 450 speakers.

*I*th, as shown as a high recognition rate on a wide number of subjects, together with a high operative velocity, make it useful for real security access control applications.

Table 2. Time performances.

	Time (min:sec)
Whole computation	17:03
Training	00:48
Test & comparison	16:15
Single test	00:02

Table 3. Comparison with the state of the art.

Approach	Accuracy rate
Reynolds [36]	96.80%
Revathi *et al*. [37]	91%
Chakroborty and Saha [39]	97.42%
Saeidi *et al*. [41]	97%
Gomez [42]	94%
Rao *et al*. [43]	99%
Proposed	97.98%

REFERENCES

[1] F. Garzia, E. Sammarco and R. Cusani, "The Integrated Security System of the Vatican City State," *International Journal of Safety & Security Engineering*, Vol. 1, No. 1, 2011, pp. 1-17. doi:10.2495/SAFE-V1-N1-1-17

[2] G. Contardi, F. Garzia and R. Cusani, "The Integrated Security System of the Senate of the Italian Republic," *International Journal of Safety & Security Engineering*, Vol. 1, No. 3, 2011, pp. 219-247. doi:10.2495/SAFE-V1-N3-219-247

[3] F. Garzia and R. Cusani, "The Safety/Security/Communication System of the Gran Sasso Mountain in Italy," *International Journal of Safety & Security Engineering*, Vol. 2, No. 1, 2012, pp. 13-39.

[4] F. Garzia, E. Sammarco and R. Cusani, "Vehicle/People Access Control System for Security Management in Ports," *International Journal of Safety & Security Engineering*.

[5] H. Beigi, "Fundamentals of Speaker Recognition," VDM Verlag, Saarbrücken, 2011. doi:10.1007/978-0-387-77592-0

[6] R. J. Mammone, X. Y. Zhang and R. P. Ramachandran, "Robust Speaker Recognition: A Feature-Based Approach," *IEEE Signal Processing Magazine*, Vol. 13, No. 5, 1996, pp.1290-1312. doi:10.1109/79.536825

[7] F. Soong, A. Rosenberg, L. Rabiner and B. Juang, " A Vector Quantization Approach to Speaker Recognition," *Acoustics*, *Speech and Signal Processing* (*ICASSP*), 1985, pp. 387-390.

[8] R. Auckenthaler, M. Carey and H. Lloyd-Thomas, "Score Normalization for Text-Independent Speaker Verification Systems," *Digital Signal Processing*, Vol. 10, No. 1-3, 2000, pp. 42-54. doi:10.1006/dspr.1999.0360

[9] S. Furui, "Recent Advances in Speaker Recognition," *Pattern Recognition Letters*, Vol. 18, No. 9, 1997, pp. 859-872. doi:10.1016/S0167-8655(97)00073-1

[10] S. Pruzansky, "Pattern-Matching Procedure for Automatic Talker Recognition," *JASA*, Vol. 26, No. 1, 1963, pp. 403-406.

[11] P. D. Bricker and S. Pruzansky, "Effects of Stimulus Content and Duration on Talker Identification," *JASA*, Vol. 44, No. 3, 1968, pp. 1596-1607.

[12] D. Jurafsky and J. H. Martin, "Speech and Language Processing," Prentice Hall, Boston, 2008.

[13] M. Farrùs, "Prosody in Automatic Speaker Recognition: Applications in Biometrics and Voice Imitation," VDM Verlag, Saarbrücken, 2010.

[14] D. A. Reynolds, "An Overview of Automatic Speaker Recognition Technology," *Acoustics*, *Speech and Signal Processing* (*ICASSP*), 2002, pp. 4072-4075.

[15] J. Mariani and F. Bimbot, "Language and Speech Processing," John Wiley & Sons, Chichester, 2010.

[16] I. R. Titze, "Principles of Voice Production," Prentice Hall, Boston, 1994.

[17] N. Morgan, H. Bourlard and H. Hermansky, "Speech Processing in the Auditory System, Chapter Automatic Speech Recognition: An Auditory Perspective," Springer,

Berlin, 2004.

[18] F. Zheng, G. Zhang and Z. Song, "Comparison of Different Implementations of MFCC," *Journal of Computer Science & Technology*, Vol. 16, No. 6, 2001, pp. 582-589. doi:10.1007/BF02943243

[19] M. Sahidullah and G. Saha, "Design, Analysis and Experimental Evaluation of Block Based Transformation in MFCC Computation for Speaker Recognition," *Speech Communication*, Vol. 54, No. 4, 2012, pp. 543-565. doi:10.1016/j.specom.2011.11.004

[20] C. Bishop, "Pattern Recognition and Machine Learning," Springer, Berlin, 2006.

[21] D. A. Reynolds, "Gaussian Mixture Models," Technical Report, MIT Lincoln Laboratory, Cincinnati, 2001.

[22] M. A. T. Figueiredo and A. K. Jain, "Unsupervised Learning of Finite Mixture Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 3, 2002, pp. 381-396. doi:10.1109/34.990138

[23] D. A. Reynolds, T. F. Quatieri and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, Vol. 10, No. 2, 2000, pp. 19-41.

[24] L. Xu and I. Jordan, "On Convergence Properties of the EM Algorithm for Gaussian Mixtures," *Neural Computation*, Vol. 8, No. 1, 1996, pp. 129-151. doi:10.1162/neco.1996.8.1.129

[25] S. V. Vaseghi, "Advanced Digital Signal Processing and Noise Reduction," John Wiley & Sons, Chichester, 2006.

[26] S. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 27, No. 2, 1979, pp. 113-120. doi:10.1109/TASSP.1979.1163209

[27] B. P. Bogert, J. R. Healy and J. W Tukey, "The Quefrency Analysis of Time Series for Echoes: Cepstrum, Pseudo-Autocovariance, Cross-Cepstrum, and Saphe Cracking," *Proceedings of the Symposium on Time Series Analysis*, 1963, pp. 209-243.

[28] C. Roads, "The Computer Music Tutorial," MIT Press, Cincinnati, 1996.

[29] J. G. Proakis and D. G. Manolakis, "Digital Signal Processing," Prentice Hall, Boston, 2007.

[30] D. O'Shaughnessy, "Speech Communication: Human and Machine," Addison-Wesley, Boston, 1987.

[31] S. Stevens, J. Stanley, J. Volkman and E. B. Newman, "A Scale for the Measurement of the Psychological Magnitude Pitch," *Journal of the Acoustical Society of America*, Vol. 8, No. 3, 1937, pp. 185-190. doi:10.1121/1.1915893

[32] M. Gold, "Speech and Audio Signal Processing," John Wiley & Sons, Chichester, 2002.

[33] M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of Speech Corrupted by Acoustic Noise, *Proceedings of IEEE International Conference on Acoustics, Speech, and signal Processing* (*ICASSP*1979), 1979, pp. 208-211.

[34] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *Speech and Audio Processing*, Vol. 9, No. 5, 2001, pp. 504-512. doi:10.1109/89.928915

[35] "Voxforge Database," www.voxforge.org

[36] R. Reynolds, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Transactions on Speech and Audio Processing*, 1995, pp. 72-83.

[37] A. Revathi, R. Ganapathy and Y. Venkataramani, "Text Independent Speaker Recognition and Speaker Independent Speech Recognition Using Iterative Clustering Approach," *International Journal of Computer Science & Information Technology*, Vol. 1, No. 2, 2009, pp. 30-42.

[38] "TIMIT Speech Database," http://www.ldc.upenn.edu

[39] S. Chakroborty and G. Saha, "Improved Text-Independent Speaker Identification Using Fused MFCC and IMFCC feature Sets Based on Gaussian Filter," *International Journal of Signal Processing*, Vol. 5, No. 1, 2009, pp. 11-19.

[40] "Yoho Speech Database," http://www.ldc.upenn.edu

[41] R. Saeidi, P. Mowlaee, T. Kinnunen and Z. H. Tan, "Signal-to-Signal Ratio Independent Speaker Identification for Co-Channel Speech Signals," *Proceedings of International Conference on Pattern Recognition* (*ICPR*2009), 2009, pp. 4565-4568.

[42] P. Gomez, "A Text Independent Speaker Recognition System Using a Novel Parametric Neural Network," *Proceedings of International Journal of Signal Processing, Image Processing and Pattern Recognition*, December 2011, pp.1-16.

[43] R. R. Rao, V. K. Prasad and A. Nagesh, "Performance Evaluation of Statistical Approaches for text-Independent Speaker Recognition Using Source Feature," *InterJRI Computer Science and Networking*, Vol. 2, No. 1, 2010, pp. 8-13.

