Robust Text-Independent Speaker Identification Using Artificial Immune System

Dr. Rand EL-Kouatly [*]

Abstract

The description of the Artificial Immune System (AIS) with the major Gaussian Mixture Model (GMM) based speaker Identification system is introduced In this paper. The new proposed menthod improves the clustering of the speech features vectors of the trained speech signals using the Supervised Clonal Selection methods based on the AIS. The experiments show that the proposed algorithm produced improved results when companed to the conventional GMM algorithm.

Keywords: Speaker Identification, Gaussian Mixture Model, Artificial Immune System, Clonal Selection.

[*] Faculty of Information Technology, Networking and System Department Damascus University

1. Introduction

Speech is one of the most natural mean of exchanging information among humans. This has spawned a growing interest in developing machines that can accept speech as input and act appropriately based on the information conveyed.

Speech consists of complex patterns, the ability of humans to recognize complex patterns and classify them is considered as a very sophisticated cognitive action. Tremendous amount of research has been done to create machines which can learn to classify and recognize patterns. Depending on the domain of the application, patterns are characterized by certain attributes (features), which allow the machine to classify them into the different categories (classes). [1]

Automatic Speaker Recognition (ASR) is a pattern recognition problem where the classes are the speaker identities. The main objective of speaker recognition systems is to secure access and identify the users. Automatic Speaker Recognition (ASR) systems generally fall into one of two categories: Automatic Speaker Identification (ASI) systems which aim to answer the question “who is the speaker?”, and the Automatic Speaker Verification (ASV) which aim to answer the question “is the speaker the one who claims to be?”.

Automatic Speaker Recognition (ASR) may be text dependent where each user has certain text to utter then the system can recognize the user, or may be text independent where a user can say partially any sentence and the system should be able to identify the speaker.

There is a huge amount of information present in a speech signal and speech can be described as having a number of different levels of information. At the top level, we have lexical and syntactic features, such as language use and sentence construction [2]. These require a lot of intelligence to understand and interpret, and automating this process requires high computation cost as the proposed systems.

On the other hand prosodic features represent information such as: intonation, stress, and rhythm of speech [3]. These require high computation cost and do not present a good results is ASR systems. Further below these are Phonetic features which represent the sound of individual syllables, and at the most basic level "low-level acoustic features" [4] which generally give information on the system that creates the sound, such as the speaker’ vocal tract may give speaker dependent information, but it is text or time dependent and thus not obviously suited for ASR systems. It is likely that the human brain uses a combination of these levels of information when identifying a speaker.

Most previous works relied on the use of low-level acoustic features. Mel-Frequency Cepstral Coefficients (MFCCs) have been particularly popular in recent years as they give a highly compact representation of the spectral envelope of a sound [5]. Line Spectrum Pairs have also been popular, as they have the related perceptual linear prediction values which have been shown to be more robust in noisy environments [6]. It is both conceivable and probable that different features have a different level of importance in characterizing different voices.

The two most popular methods were used in the previous works are vector quantization (VQ) and Gaussian Mixture Models (GMM). In the VQ method [6] [7] [8], speaker models are formed by clustering the speaker’s feature vectors in K non overlapping clusters. Each cluster is represented by its centroid (average vector) and the resulting collection of K centroids is referred to as its code book and serves as a model for the speaker. The two considerations when using this method are: what method to use to perform the clustering and what size codebook to use.

The GMM method [9] differs from the VQ method in that it is a parametric method: A GMM consists of K Gaussian distributions parameterized by their a priori probabilities, mean vectors and covariance matrices. The parameters are typically estimated by maximum likelihood estimation [10] [11].

The present paper is organized as follows; in next section the speaker identification based on vector quantizer is described. In section 3 an introduction to the Artificial Immune System (AIS) is presented. In section 4 the Gaussian Mixture Model for speaker identification is explained in more details [9]. In section 5 the Unsupervised Clonal Selection (UCSC) [12] algorithm is described. In section 6 the description of the proposed algorithm is presented. The implementation and experimented results are presented in section 7 and 8 respectively. Finally, section 9 presents summary and conclusion.

2. Vector Quantization

A vector quantizer is a system for mapping a sequence of continuous or discrete vectors into a digital sequence suitable for communication or storage in a digital channel. The goal of such a system is data compression: to reduce the bit rate so as to minimize communication channel capacity or digital storage memory requirements while maintaining the necessary fidelity of the data [6]. Vector Quantization (VQ) is a lossy data compression method based on the principle of block coding [13].

VQ may be seen of as an approximator. Figure 1 shows an example of a 2- dimensional VQ. Here, every pair of numbers falling in a particular region are approximated by a star associated with that region.

In Figure 1, the stars are called code vectors and the regions defined by the borders are called encoding regions.

The General scheme based on VQ system for Speaker Identification is shown in Fig. 2. Test and reference patterns (feature vectors) are extracted from speech utterances statistically or dynamically. At the training stage, reference models are generated (or trained) from the reference patterns by various methods. A reference model (or template) is formed by obtaining the statistical parameters from the reference speech data. A test pattern is compared against the reference templates at the pattern matching stage. The comparison may be conducted by probability density estimation or by distance (dissimilarity) measure. After comparison, the test pattern is labeled to a speaker model at the decision stage. The labeling decision is generally based on the minimum risk criterion [14].

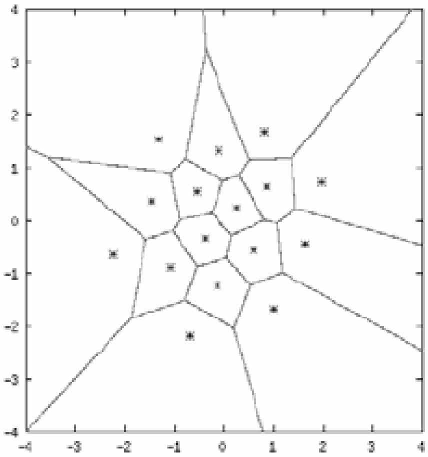


Figure 1. An example of a 2-dimensional VQ [6].

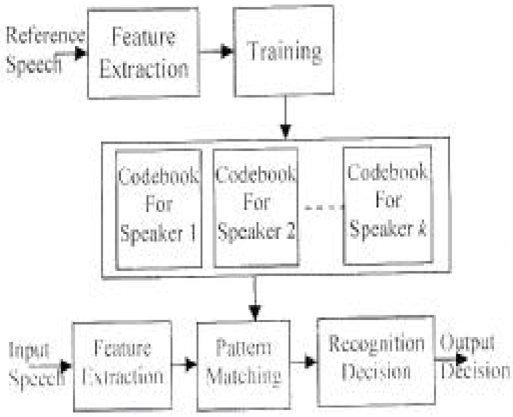


Figure 2. Speaker Identification System based on VQ [14].

3. Artificial Immune Systems

Artificial Immune Systems (AIS) is a field of study devoted to the development of computational models based on the principles of the biological immune system, the biological immune system is a network of cells, tissues, and organs that work together to defend the body against attacks by "foreign" invaders. So AIS is an emerging area that explores and employs different immunological mechanisms to solve computational problems [12].

There are various mechanisms in the artificial immune system such as clonal selection, affinity maturation, somatic hyper-mutation, receptor editing and negative selection. A lot of immune algorithms were developed aiming to find solutions to a broad class of complex problems. Applications of AIS have included the following areas: clustering and classification, anomaly detection, optimization, control, computer security, learning, bio-informatics, image processing, robotics, virus detection and web mining. Many of the immune algorithms use principles inspired by the clonal selection theory of acquired immunity. The clonal selection principle is used by the immune system to describe the basic features of an immune response to an antigenic stimulus. It establishes the idea that only those cells that recognize the antigens proliferate, thus being selected against those which do not. The process of proliferating called clonal expansion [12]. The selected cells are subject to an affinity maturation process which improves their affinity to the selective antigens [15]. It is beyond the scope of this paper to accurately describe the AIS process; readers unfamiliar with this process should refer to[15]or [16].

4. Gaussian Mixture Models

For text-independent speaker identification, where there is no prior knowledge of what the speaker will say, the most successful function has been used is Gaussian mixture models. A Gaussian mixture density is a weighted sum of *M* component densities as shown in Fig .3. For a *D*-dimensional feature vector, *x*, the mixture density $p(x/l)$ for the model l used for the likelihood function is defined as [17]:

$$p(x/l)=\sum_{i=1}^{M} w_i\, p_i(x) \qquad (1)$$

Where w_i is the mixture weights vector satisfying the constraint $\sum_{i=1}^{M} w_i = 1$[9], and $p(x_i)$ is the Gaussian probability density function given by:

$$p_i(x)=\frac{1}{(2p)^{D/2}\,|\Sigma_i|^{1/2}}\exp\left\{-\frac{1}{2}(x-m_i)^T(\Sigma_i)^{-1}(x-m_i)\right\} \qquad (2)$$

Where each $p(x_i)$ is parameterized by a mean $D´1$ vector, μ_i, and a $D´\ D$ covariance matrix Σ_i, The complete Gaussian mixture density is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. These parameters are collectively represented by:

$$l=\{p_i, m_i, \Sigma_i\} \qquad i=1,....,M \qquad (3)$$

So, each speaker, in the training or identification is represented by a GMM and is referred to its λ model.

It has been shown [17], that the GMM will model some underlying set of acoustic classes, such as vowels, nasals or fricatives which may reflect speaker dependent vocal tract, on the other hand the spectral shape of the *i*th classes can be represented by the mean μ_i, and the variations of the average spectral shape can be represented by covariance matrix Σ_i [18].

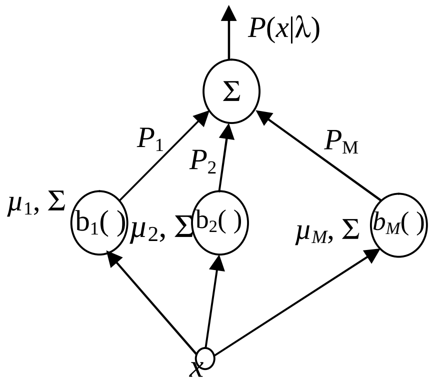
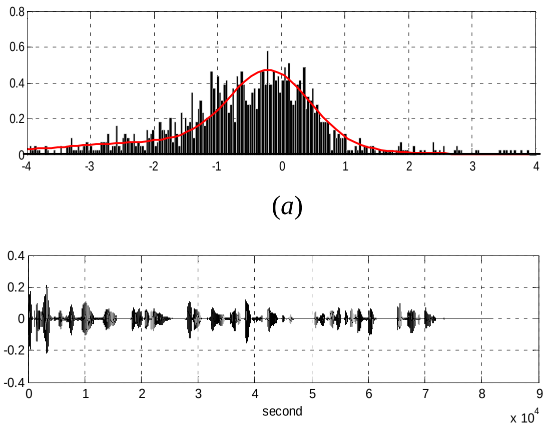


Figure 3. Description of M Component of Gaussian Mixture Model [17].

In the GMM-Universal Background Model GMM-UBM system [9] use a single, speaker-independent background model to represent $p_i(x,l)$. The UBM is a large GMM trained to represent the speaker-independent distribution of features. Specifically, the goal is to select speech that is reflective of the expected alternative speech to be encountered during recognition.

Figure 4 represent an example of the Estimated Gaussian Mixture Model of the Melcepstral Coefficients of the 10 second speech signal shown in Figure 4(a), and the histogram of the Melcepstral Coefficients with estimated GMM in Figure 4(b).



(*a*)

(*b*)

Figure 4: An example of the Estimated Gaussian Mixture Model of the Melcepstral Coefficients of the speech Signal, (a) Speech Sample, (b) Histogram of the Melcepstral Coefficients with estimated GMM.

Figure 5, Shows an example of the first and second Melcepstral features vectors mapped with Expectation Maxima (EM), and the variance of the trained data, computed by using the same spoken sentence of 10 second duration, one by using a male speaker and the other by female speaker. As noticed the Gaussian Mixture Model of the two speakers are quite different even the spoken sentence are the same, this indicate that the GMM of the Melcepstral features are related to the speakers himself.

5. Unsupervised Clonal Selection Classification (UCSC)

In Unsupervised Clonal Selection Classification (UCSC) [12], clustering problem is considered as optimization problem and the objective is to find the optimal partitions of data where the resulting clusters tend to be compact as possible. A simple criterion which is the within cluster spread is used in UCSC, this criterion needs to be minimized for good clustering. UCSC uses the sum of the Euclidean Distances of the points from their respective cluster centroids as clustering metric and uses clonal selection algorithm as clustering algorithm which ensures finding the global optima. The number of clusters *K* is supposed to be known and the appropriate cluster centers m_1, m_2,…,m_k have to be found such that the clustering metric *J* is minimized. Mathematically, the clustering metric *J* for the *K* clusters $C=\{C_1, C_2, \ldots, C_K\}$ is given by the following equation:

$$J(G,\mathrm{M})=\sum_{i=1}^{k}\sum_{j=1}^{N} g_{ij}\left\|x_j-m_i\right\| \qquad (4)$$

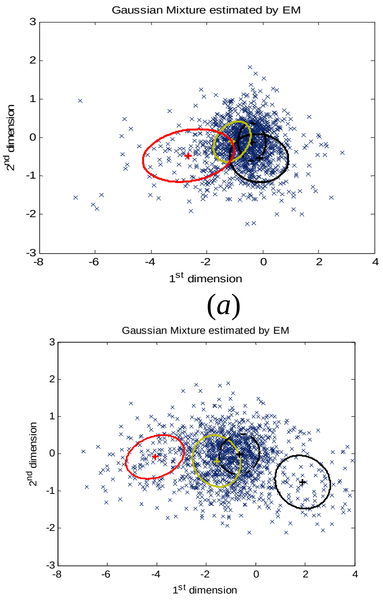
where $x_j \in \Re^d$, $j = 1,\ldots, N$ are data points, $\Gamma=\{g_{ij}\}$ is a partition matrix,

$$g_{ij}=\begin{cases}1 & if \quad x_j \in C_i\\ 0 & otherwise\end{cases} \text{ with } \sum_{i=1}^{K} g_{ij}=1 \quad \forall j \; ,$$

$\mathrm{M}=[m_1, m_2, \ldots, m_k]$, and M is centroid matrix and

$$m_i=\frac{1}{N_i}\sum_{j=1}^{N} g_{ij}x_j, \qquad i=1,..,k$$

is the mean for the C_i cluster with N_i points. The task of clonal selection algorithm is to search for the appropriate cluster centers wherefore *J* is minimized [12].



(*a*)

(*b*)

Figure 5: An example of the first and second Melcepstral features vectors mapped with GMM of the 10 second training speech ignal with same spoken sentence, (a) Female Speaker, (a) Male Speaker.

6. UCSC algorithm and the speaker identification

The main goal to be achieved using UCSC algorithms is to optimize the GMM trained data before using it in speaker verification. The UCSC is used in the present work to optimize the clustering and the clustering centroids of the constructed sets of detectors for a given speaker voice signal deviation from normal behaviour, The algorithm generates detectors from a segmented version of the original data (the self set N_s) whose representation differs from one application to another, where the proposed Clonal Selection algorithm tend to ensure the global optima of the clustering using equation (4), where x_i represents the data sets of the trained and tested speech features. The M matrix will represent the new GMM of the features if trained tested speech signal. Accordingly, the

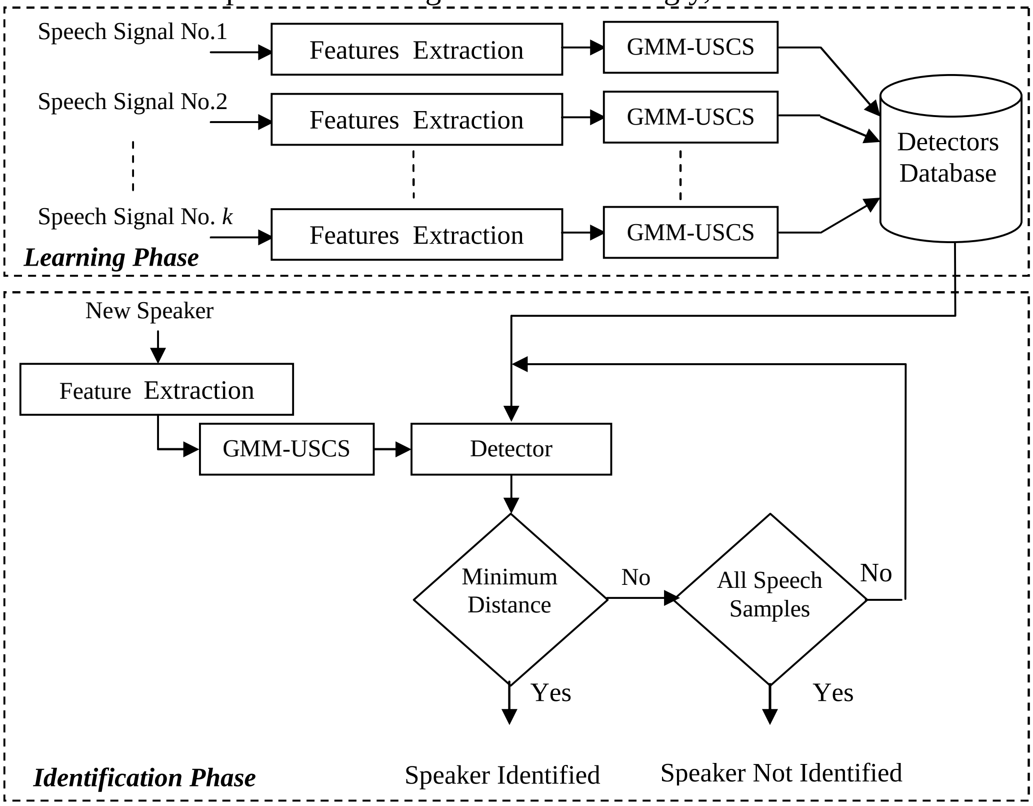


Figure 6: Speaker Recognition system based on the USCS algorithm.

USCS algorithm optimizes the GMM data. The generated detectors are then used as a voiceprint to monitor the acquired new voice signals (identification phase). If the signal is produced by the same speaker, the form and the data distribution must be very similar to the original one (used to generate detectors), so a very low anomalies rate will be obtained (null in the ideal case). According to the obtained anomalies rate, the automatic speaker recognition system decides of the new voice speaker identity. To achieve that, a database of voiceprints of different speakers is used. A voiceprint is given by the set of detectors obtained when applying the negative selection algorithm to the corresponding voice signal (the learning phase), where a non-selected detectors are omitted. If the lowest obtained anomalies rate is higher than a fixed threshold ∂ (witch a parameter of the system that determine the highest accepted value of the detected anomalies rate), the system decides that this voice signal does not belong to any speakers of the database, and so the speaker is not identified. Else, the speaker is identified as the one with the lowest anomalies rate (lower than ∂). Figure 6 resumes how the recognition system operates.

Figure 7, shows an example of the first and second Melcepstral features vectors mapped with GMM and GMM-USCS methods, computed by using of 10 second of male trained speech samples. The shown example shows a change in the computed mean between the two used different methods, the changes in Clustering and Centroids are due to the changes in the criteria used when using Immune System instead of using *K*-mean mathematical methods [12]. It is noticed also that the Expectation Maxima (EM) and the Variance of 1st and 2nd order vectors are almost the same for the same male shown in Figure .5(b) and Figure 7, even with different spoken sentence. This will verify the ideas of using Melceptral computation with GMM are more speaker dependent and less speech independent.

7. Implementation

The experiments were conducted using a collection of speech database [19] from 39 male speakers, with a 6 sentences for each speaker. The experiments used five session per speaker with three sessions for training and two sessions for testing the data [20].

First, the speech is segmented into frames by a 20-ms window progressing at a 10-ms frame rate. The speech frame was pre-emphasised using a pre-emphasis coefficient of 0.95 prior to being windowed by a Hamming window. A speech activity detector is then used to discard silence–noise frames. The speech activity detector is a self-normalizing, energy based detector that tracks the noise floor of the signal and can adapt to changing noise conditions. Next, mel-scale cepstral feature vectors with 12 order mel-coefficients are extracted from the speech frames. The mel-scale cepstrum is the discrete cosine transform of the log-spectral energies of the Speech segment *Y*, where all cepstral coefficients except its zeros value (the DC level of the log-spectral energies) are retained in the processing. Finally, delta cepstra are computed using a first order orthogonal polynomial temporal fit over ±2 feature vectors as described in [17].

Then, the sequence of feature vectors was divided in overlapping segments of T feature vectors, using the following [17]:

$$\overbrace{\vec{x}_1, \vec{x}_2, \mathbf{L}\ \vec{x}_T}^{\text{Segemnt 1}}, \vec{x}_{T+1}, \vec{x}_{T+2}, \mathbf{L}$$

$$\vec{x}_1, \overbrace{\vec{x}_{2,} \vec{x}_3, \mathbf{L}\ \vec{x}_{T+1}}^{\text{Segemnt 2}}, \vec{x}_{T+2}, \vec{x}_{T+3}, \mathbf{L} \qquad (5)$$

The identified speaker of each segment was compared to actual speaker of the test utterance and the number of segments which were correctly identified was recorded, this steps were repeated for each test utterances from the data sets. The final performance evaluated by computing the higher percentage of the correct test utterance using the following equation [17]:

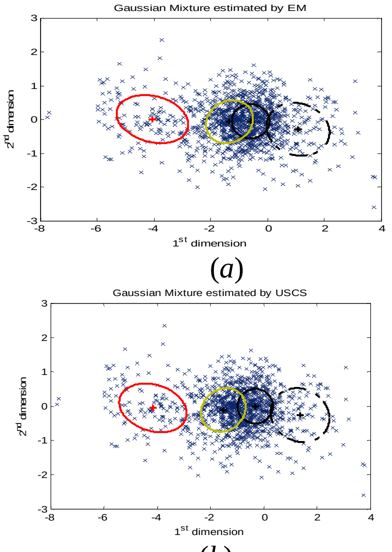
$$\text{Correct identficat ion (\%)} = \frac{\text{Number of correctly identficat ion segment}}{\text{Total number of segments}} \times 100 \qquad (6)$$

The evaluation was repeated for different speech utterance length for tested and trained speech from different male and female speakers. The trained database was generated in two phases, the first phase is by using GMM methods as in sets [18], the second phase the trained database was generated using GMM of the trained vectors optimized by artificial immune systems using USCS algorithm [12] using detection threshold ∂ selected to be 0.4 as in [20].

8. Analysis of Tests Results

It is so difficult to characterize the performance of speaker identification system due to complexity and testing scenarios [21]. The best and most used ranking is by using the detection error tradeoff (DET) curvess. The DET indicates the tradeoff between the false rejection (or nondetection) error happens when a valid identity claim is rejected. A false acceptance (or false alarm) error consists in accepting an identity claim from an impostor.

The speaker recognition with GMM and modified GMM-USCS system was programmed using MATLAB program. The experiment are shown using the detection error trade-offs (DETs) curve, shown in Figure 8. The experiment was conducted first without using the UCSC algorithms, the results was shown in the Figure.8 which shows almost identical results obtained in [9]. The experiment was repeated using the proposed GMM–UCSC, the obtained results shows an improvement in the DET curves as shown in Figure 8.



(*a*)

(*b*)

Figure 7: An example of the first and second Melcepstral features Vectors mapped with GMM of the 10 second training speech signal with same spoken sentence computed by, (a) GMM, (b) GMM-USCS.

In order to examine the improvement of the speaker recognition system performance, the Equal Error Rate points was computed (The EER is a summary performance which indicates the points on DET curve where the false rejection *fr*, and false alarm *fa* points are equals *fr=fa*) [22]. The system are also evaluated by computing the minimum Detection Cost Figure minDCF, which represent the two error rates weighted by their respective costs, that is $C = C_{fa}P_{fa} + C_{fr}P_{fr}$ [11]. In this equation, C_{fa} and C_{fr} are the costs given to false acceptances and false rejections, respectively. The cost function is minimal if the threshold is correctly set to the desired operating point.

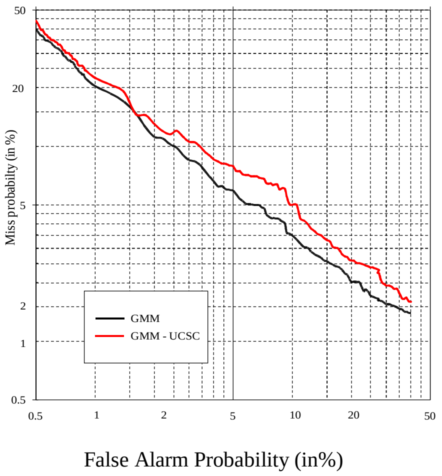


Figure 8: DET Curves for tow compositions: Pooled male and female data using GMM, and pooled male and female models using GMM with USCS.

Table 1 reports the performance of the EER (%) using GMM and GMM-UCSC, when testing 39 sentences of male and female speakers. The performance of the tested speech corpora [19] is comparable with the performance shown in [9]. The results shows that an improvement in detection performance of about 2.34%.

Table.1: Performance evaluation of EER(%) and Minimum DET for the Basic GMM and GMM with USCS method, for Tested Male and Female Speakers.

	EER (%)	Min DCT
GMM	5.89	2.26
GMM-UCSC	8.23	3.48

9. Conclusion

Accuracy in ASR over recent years has improved significantly in the past decades, but still the accuracy of speaker verification under the required performance especially when text independent used in the verification. In this study a new approach was introduced based on Artificial Immune System, the ideas is to use a combination of GMM system with USCS algorithm in order to improve the clustering and the expectation maxima that are more relatives to speaker features and more independent from the speech itself. The obtained results show an improvement in the detection of about 2.34%.

References

[1] Lawrence Rabiner and Biing-Hwang Juang, Fundamental of Speech Recognition", Prentice-Hall, Englewood Cliffs, N.J., 1993.

[2] S. Mohammad and T. Pedersen, "Combining Lexical and Syntactic Features for Supervised Word Sense Disambiguation", in the Proceedings of the Conference on Computational Natural Language Learning (CoNLL), pp. 225-32, Boston, MA, May 6-7, 2004.

[3] J. Adell, A Bonafonte, and D. Escudero, "Analysis of prosodic features: towards modeling of emotional and pragmatic attributes of speech", in the Proceedings of the Natural Language No. 35 pp. 277-283, Spain, 2005.

[4] Randy Allen Harris, "Voice Interaction Design: crafting the new conversational speech systems", Morgan Kaufmann Publishers, Elsevier, SAN Francisco, 2005.

[5] M.R. Hasan, M. Jamil, M. G. Rabbani, and S. Rahman, "Speaker Identification Using Mel Frequency Spestral Coefficients", in 3rd International Conference on Electrical & Computer Engineering ICECE 2004, Dhaka, Bangladesh, 28-30 December 2004.

[6] H. B. Kekre, V. Kulkarni, "Speaker Identification by using Vector Quantization", International Journal of Engineering Science and Technology Vol. 2(5), pp 1325-1331, 2010.

[7] W.C. Ching-T. Hsieh, C. H. Hsu, "Robust Speaker Identification System Based on Two-Stage Vector Quantization", Tamkang Journal of Science and Engineering, Vol. 11, No. 4, pp. 357_366, 2008.

[8] T. Matsui, and S. Furui, C. H. Hsu, "Comparison of Text-Independent Speaker Recognition Methods Using VQ-Distortion and Discrete Continuous HMM's", IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 2, NO. 3, pp 456-459, JULY 1994.

[9] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", in Digital Signal Processing Vol. 10, Nos. 1–3, pp 446-456, January/April/July 2000.

[10] B. Xiang, and T. Berger, "Efficient Text-Independent Speaker Verification with Structural Gaussian Mixture Models and Neural Network", IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 11, NO. 5, pp 446-456, SEPTEMBER 2003.

[11] F. BimbotB et al, "A Tutorial on Text-Independent Speaker Verification", EURASIP Journal on Applied Signal Processing, Vol:4, pp 430–451. 2004.

[12] M.T. Al-Muallim, R. El-Kouatly, "Unsupervised Classification Using Immune Algorithm", International Journal of Computer Applications (0975 – 8887), Vol. 2 – No.7, pp 44-48, June 2010.

[13] H.B. Kekre, T. K. Sarode, "Vector Quantized Codebook Optimization using K-Means", International Journal on Computer Science and Engineering Vol.1(3), pp 283-290, 2009.

[14] R. M. Gray, ``Vector Quantization," IEEE ASSP Magazine, pp. 4--29, April 1984.

[15] L. N. de Castro and J. Timmis, "Artificial Immune Systems: A New Computational Intelligence Approach", Springer, 2002.

[16] D. Dasgupta and L. F. Niño, "Immunological Computation Theory and Applications. Boca Raton", CRC Prees, Taylor & Francis Group, 2009.

[17] Reynolds, D. A. and Rose, R. C., "Robust text-independent speaker identification using Gaussian mixture speaker models", IEEE", IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. 3 (1995), 72–83.

[18] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", in Digital Signal Processing Magazine 10, Nos.

1–3, January/April/July, pp 19–41, 2000.

[19] Speech DATAbase available in http://www.voxforge.org/home", 2010. Or http://www.ldc.upenn.edu/Catalog/nonMember.html.

[20] K. M. Faraoun, A. Boukelif, "Artificial Immune Systems for text-dependent speaker recognition", In Scientific Commons, http://en.scientificcommons.org, July 06, 2006,

[21] Douglas A. Reynolds, "An overview of Automatic Speaker Recognition technology", in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp IV-4072 - IV-4075 May 2002.

[22] D. E. Struim, W. M. Camphell, D. A. Reynolds, "Classification Methods for Speaker Recognition", in Lecture Notes in Computer Science, Volume 4343/2007, pp 278-297, Springer link, 2007.·

· Received 18/8/2012