Journal of Engineering Science and Technology
Vol. 12, No. 5 (2017) 1153 - 1164
© School of Engineering, Taylor's University

LEARNING VECTOR QUANTIZATION FOR ADAPTED GAUSSIAN MIXTURE MODELS IN AUTOMATIC SPEAKER IDENTIFICATION

IMEN TRABELSI*, MED SALIM BOUHLEL

Sciences and Technologies of image and Telecommunications (SETIT),
University of Sfax, Tunisia
*Corresponding Author: imen.trabelsi@enit.rnu.tn

Abstract

Speaker Identification (SI) aims at automatically identifying an individual by extracting and processing information from his/her voice. Speaker voice is a robust a biometric modality that has a strong impact in several application areas. In this study, a new combination learning scheme has been proposed based on Gaussian mixture model-universal background model (GMM-UBM) and Learning vector quantization (LVQ) for automatic text-independent speaker identification. Features vectors, constituted by the Mel Frequency Cepstral Coefficients (MFCC) extracted from the speech signal are used to train the New England subset of the TIMIT database. The best results obtained (90% for gender- independent speaker identification, 97 % for male speakers and 93% for female speakers) for test data using 36 MFCC features.

Keywords: Speaker identification, LVQ, GMM, MFCC, TIMIT.

1. Introduction

In recent times, a great deal of research has been conducted concerning the field of Automatic speaker recognition (ASR) [1]. It can be credited to the developing requirement for enhanced security in remote identity identification or verification in real-world applications such as telephone banking or online access. Automatic Speaker Recognition consists in recognizing humans from the speaker-specific information included in their voices. Each speaker has unique physiological characteristics of speech because of the different sizes of vocal fold (biometric modality).

Within speaker recognition, a distinction can be made between speaker verification (SV) and speaker identification (SI) systems. Speaker verification is

Nomenclatures

b_i	Component densities
n	Number of frames
Pi	Mixture weights
x	Support vectors

Greek Symbols

α	Adaptation coefficient
Γ	Relevance factor
μ	Mean vector
Σ	Covariance matrix

Abbreviations

ASR	Automatic speaker recognition
EM	Expectation maximization
GMM	Gaussian mixture model
LVQ	Learning vector quantization
RBF	Radial basis function
SI	Speaker identification
SV	Speaker verification
TDSI	Text-dependent speaker identification
TISI	Text-independent speaker identification
UBM	Universal Background Model

the process of verifying if the pronounced segment is matched with the identity claim of a speaker. Speaker identification is the process of identifying who is speaking from the registered speakers in the database. Speaker identification systems are typically distinguished into two categories: text-dependent speaker identification (TDSI) and text-independent speaker identification (TISI). TDSI requires an explicit identification process, usually combined with a predetermined group of words or sentences (pass-phrase or Personal Identification Number). TISI requires an implicit identification protocol, with no constraints on the speech content. When the speaker is, per example, registering a complaint, the verification is processed. Compared to TDSI, it is more convenient. Another classification in speaker identification is possible. It is based on two categories: closed set and open set. If the speaker is registered in the database, we talk about closed set speaker identification task. If not, we are talking of open-set speaker identification task. This paper is about recognizing speakers, with text-independent content and in a closed-set task. Figure 1 shows a basic architecture of the speaker identification system. Various modelling techniques are studied in the field of speaker recognition. Self-Organizing Map (SOM) [2], HMM (Hidden Markov Modeling) [3], Learning Vector Quantization (LVQ) [4,5], Gaussian Mixture Model (GMM) [6-8] and Support Vectors Machines [9] are the most studied techniques.

From this list, the most successful one is GMM. The success of GMM is not only due to its successful application in speaker identification area and the availability of sufficient data for speaker modeling but also because it encompasses many of the most used estimation methods (maximum likelihood, generalized least squares etc.). The concept of Gaussian Mixture Model–Universal Background Model (GMM–UBM) is an effective framework that has found great success in the last years [10-12]. Conceptually, UBM is a large mixture of Gaussians that covers all speakers.

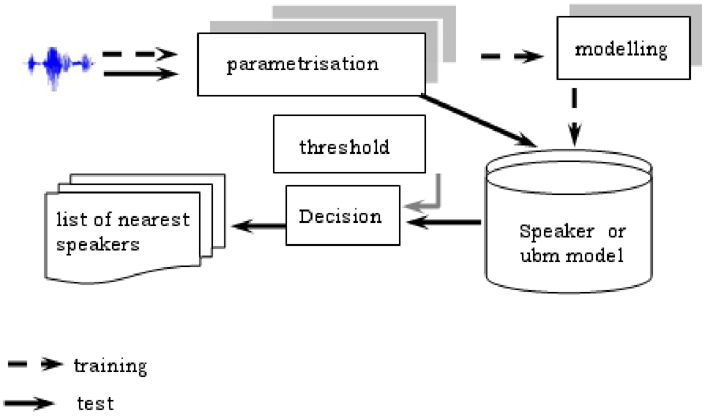

Fig. 1. Basic architecture of speaker identification system.

The concept of Gaussian Mixture Model–Universal Background Model (GMM–UBM) is an effective framework, with a great success in the last years [10-12]. Conceptually, UBM is a large number of mixture components that covers all speakers. In this paper, a novel GMM-Based approach is proposed. The Gaussian mixture model-universal background model (GMM-UBM) is combined with learning vector quantization (LVQ) classifier to further improve the performance of a robust speaker identification system. In other words, the main objective of this study is to propose an efficient approach to explore the benefits of the strong points of GMM and LVQ network. A study of the effect of gender-dependent (male/female) acoustic models is also presented. And as a final point, different feature vectors are explored. The remainder of this paper is organized as follows. The feature extraction technique is given in Section 2. In Section 3, the GMM/LVQ technique is presented. The experimental results are presented in Section 4. Conclusions are made in Section 5.

2. Feature Extraction

Mel-frequency cepstral coefficients (MFCCs), introduced in 80’s, have been used successfully as features for many speech processing applications, including speaker and emotion recognition [13-15]. The MFCC coefficients are estimated from the mel-frequency cosine transform representation of the log power spectrum. In order to improve the performance of speech recognition system, and in addition to these typical post-processing operations, MFCC features are usually augmented with temporal information; both first and second derivatives are executed [16]. These temporal parameters show a good performance in capturing the transitional characteristics of the speech, which can contribute to better identify the speakers.

3. LVQ/Gaussian Mixture Model Approach

3.1. Gaussian mixture model

As the GMM can smoothly approximate any smooth shape of the density distribution of the feature space, is then often used to test the discriminate capabilities of the short-time spectral features [17, 18].

The Gaussian mixture density is defined by:

$$p(x|\lambda)=\sum_{i=1}^{N} P_i\, b_i(x) \qquad (1)$$

A GMM is a weighted linear combination of *N* unimodal Gaussian densities density *bi(x).*

$$b_i(x)=\frac{1}{(2\Pi)^{\frac{d}{2}}\left|\sum_i\right|^{\frac{1}{2}}}\ \exp\left[-\frac{1}{2}(x-\mu_i)^T\sum_i^{-1}(x-\mu_i)\right] \qquad (2)$$

Each Gaussian component is described by a mean vector, *μi*, and a covariance matrix, *Σi. bi(x)* has also its own probability (mixture weight). Mixture weights (*pi*) which are constrained to be positive must satisfy constraint:

$$\sum_{i=1}^{N} p_i \;=1 \qquad (3)$$

In order to estimate the maximum likelihood model for a given training vector, the iterative expectation-maximization (EM) algorithm is always used [19]. Generally, the diagonal covariance matrices are often used and this is due to several reasons as mentioned in [5]. Compared to other modelling techniques, GMMs are especially used because of the relatively reduced training time and the scaled models.

3.2. Universal background models (UBM)

The concept of GMM–UBM is widely used for speaker recognition where the availability of training data is sparse. A UBM or World Model represents general background data that includes large number of speakers, languages, communication channels, recording devices, and environments. UBM model is an off-line trained GMM model. The parameters are trained through multiple loop of EM algorithm which iteratively modifies the GMM parameters to increase the model's likelihood value. In the next step, all the speaker models are estimated updated with maximum a posteriori (MAP) adaptation from the UBM. This is done because modelling the acoustic characteristics of speakers needs an important amount of training data. Also, we have to keep in mind, that the merged data are not equitable over the subsets. In this case, the final acoustic model will be biased toward the dominant subset.

The new sufficient statistics from each speaker (mixture *i*) are computed to modify the old sufficient statistics from the UBM (mixture *m*) following this equation:

$$\mu^{*}_{X,\mathrm{i}} = \alpha_i . \hat{\mu}_{X,\mathrm{i}} + (1-\alpha_i).\mu_i^{UBM} \tag{4}$$

where $\mu^{*}_{x,i}$ the final adapted mixture mean.

To control the balance between the old and the new statistics, an adaptation coefficient *αi* is used:

$$\alpha_i = \frac{n_i}{n_{i+\Gamma}} \tag{5}$$

where n_i is the number of frames for the mixture is and Γ is a fixed relevance factor of [14-16].

The use of GMM-UBM concept in speaker identification is motivated by the following reasons:

- Gives a compact representation for the speaker space since the information is embedded.
- Makes strong assumptions about the data.
- Aims at mapping a complete utterance to a fixed-length vector with the use of a UBM model.
- Simple to learn and estimate.

This is what’s called the GMM-UBM approach which becomes a popular system in the area of speaker recognition for its significant performance reported in the literature.

3.3. Learning vector quantization (LVQ)

LVQ, initially developed by Kohonen [12], is a supervised two-layer neural network that applies winner-take-all Hebbian learning. Similar to SOM classifier and kNN methods, the aim of LVQ is to globally optimize the positions of codevectors generated with unsupervised learning algorithm, with a minimized chance of being misclassified. From the input space, an input vector is randomly chosen. If the codevector and the class label of the input vector are the same, then the distance of the vectors in the same class is reduced. Otherwise, the codevector is moved away from the input vector. Applying the LVQ technique involves:

- The structure of LVQ network is designed;
- The different weights of vectors are initialized;
- A vector from the training data is presented;
- The distances between the input vector and the reference vectors are calculated;
- The nearest reference vectors are updated, according to the distance criteria;
- The steps from 3 to 5 are repeated until all patterns are correctly classified or the number of loop has been exceeded.

LVQ is appealing for several reasons: LVQ network is easy to implement and it converges fast. It defines a clustering of the data distribution by means of the prototypes. The classifier can also deal with multiclass problems without modifying the learning algorithm or the decision rule. The complexity of the resulting classifier can be controlled by the user. Furthermore, missing values will to be replaced, but will simply be ignored for the comparison between prototypes and input training vectors.

3.4. Architecture of the proposed system GMM-UBM/LVQ

The GMM-UBM system serves as a means of speaker representation for the attached LVQ network. Figure 2 gives a detailed description for respectively the training and the test process of the proposed system. The training process is achieved in two steps. The first step which is derived from the GMM-UBM approach consists in generating the UBM. The second step generates the appropriate model for each target speaker by transforming the mean of the UBM model throw MAP) criterion. Only the means are adapted using MAP adaptation, the covariance and mixture weights remain unchanged. In fact, the best overall performance is usually from adapting only the mean vectors [16]. After that, a GMM supervector is created through the concatenation of all the mean vectors of the target model. GMM supervectors can be thought of as a mapping between an utterance and a high-dimensional vector. The generated supervectors represent the input for the LVQ network.

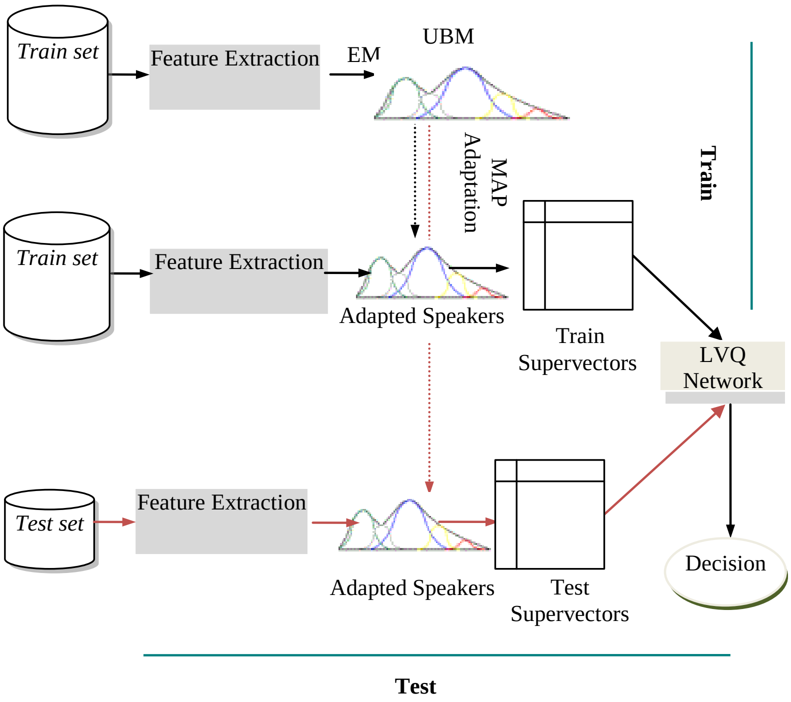

Fig. 2. GMM-UBM/LVQ architecture.

4. Results and Discussion

4.1. Experimental protocol and speech database

The study is conducted using the TIMIT database [21] for both UBM training and evaluation. TIMIT is a corpus of phonemically and lexically transcribed speech of American English speakers (428 male, 192 female), from 8 different dialect regions and was designed for developing automatic speech recognition systems.

The average quantity of speech available per-speaker is 30 seconds. The content of the recordings is exclusively read speech. The ages of the speakers are heavily biased towards younger speakers. The database was released as 16 kHz, 16 bit SPHERE format files. Experiments have been conducted under the experimental conditions described (in Table 1). In the parameterization phase, we use several number of MFCC coefficients. The speech is first pre-emphasized using a pre-emphasize filter (0.97). Then, the pre-emphasized signal is divided into 16 ms frames with 8 ms overlap. Cepstral mean normalization (the removal of the DC level obtained from the time evolution of the cepstral coefficients) is also performed. Each MFCC vector is augmented by the first and second derivatives. The feature extraction module is done by VOICEBOX [22]. For the experiments, three different 128 Gaussian component UBM models were built (UBM-gender independent, UBM for female speakers and UBM for male speakers). Individual speaker models are MAP-adapted; only mean vectors, with a relevance factor of 16.

Table 1. Experimental conditions.

Corpus	**TIMIT**
Dialect	Dr1
Region	England
Speakers	18 female, 31 male
Train utterances per speaker	8
Test utterances per speaker	2

4.2. Discussion

An important impact on the classification ability of the LVQ algorithm is related to the number of initial codebook vectors. Using multiple codebook vectors per class can describe more better the classification frontiers, but also too many may prove to be merely an overfitting and lead to poor accuracy. Table 2 summarizes results of speaker identification task using different number of codebook vectors. Note that in this experiment, the number of MFCC coefficients is set to 12.

Table 2. Speaker identification accuracy in percent (%) as a function of training codebook vectors and execution time (s) using GMM-UBM/LVQ method.

Codebook vectors	**RR (%)**	**Time (s)**
49	65	99
98	76	122
147	82	225
196	86	292

After the end of 49 training codebook vectors, the LVQ network can correctly recognize 65% of the test set. For 196 codebook vectors, the overall identification rate level increases to 86%. But, the higher recognition rate is, the lengthener training time is.

Now, for 128 gaussian and 196 codebook vectors, we analyze the effect of the MFCC number on speaker identification performance. Table 3 lists the results for 8, 12, 16 and 20 MFCC coefficients. It can be seen that the recognition rate increases as we increase the number of MFCC. But with 20 MFCC coefficients, the recognition rate tends to decrease. Increasing the accuracy of the identification system by increasing the number of parameters leads to an increase of complexity and eventually does not lead to a better result.

Table 3. Effect of coefficient order on speaker identification.

MFCC number	RR (%)
8	81
12	86
16	86
20	82

In the next experience, for 128 Gaussian and 196 codebook vectors, we analyze the effect of the MFCC number appending the first and second order derivatives (delta and double delta) of features on speaker identification performance. Table 4 shows the identification rates (%) values for each feature set with delta and double delta features. As it can be seen, Appending the delta and double delta improves the recognition accuracy, the feature vector composed of MFCC+ delta+ delta delta shows approximately 4% further improvement in the recognition rate.

From these two precedent experiments, we conclude that the greater the number of parameters in a model, the greater should be the training sequence.

Table 4. Effect of time derivatives of mfcc feature.

	Delta(MFCC)	Delta delta(MFCC)
RR (%)	85%	90%

Table 5 lists and compares the results if employing separately LVQ and LVQ/GMM. As it can be seen, the results obtained with lVQ network are poor. The performances achieved are all below 15%. The initialization step is confirmed very critical with this network.

Table 5. Comparison LVQ vs GMM-UBM/LVQ

Codebook vectors	LVQ	LVQ/GMM
49	10	65
98	12	76
147	15	82
196	15	86

Finally, gender information is believed helpful for speaker recognition. From the gender’s point of view, there are many differences between females and males either in speech production or acoustical characteristics. In this experiment, two universal background model UBM gender-dependent (male, female) are trained:

- For female-specific evaluations, we perform feature extraction, generate a UBM using the whole female training set, adapt each target female speaker model from this UBM, and then evaluate the system performance using the female test set.
- For male-specific evaluations, we perform feature extraction, generate a UBM using the whole male training set, adapt each target male speaker model from this UBM, and then evaluate the system performance using the male test set.

The results from these two independent systems are listed in Table 6. We observe that the gender-dependent systems outperform the gender-independent system. Comparing both genders, it is evident that male recognition rates are higher than female ones. This result may be due to the particular composition of speaker sets in TIMIT database. These results are consistent with these of other researchers in the same database [2, 15].

Table 6. Comparison gender-independent systems vs. gender-dependent systems.

	Gender independent	Female Speaker	Male Speaker
RR (%)	90	93	97

In this paper, numerous speaker identifications systems are presented in Table 7. With years, the research in the field of independent speaker recognition is focalizing more on the implementation of large and continuous vocabulary. Hybrid models, as in [7, 23], are the new adopted approach in ASR systems. The authors utilize GMM–SVM hybrid architectures, in order to combine the power of generative GMM and the discrimination ability of SVM. Considering feature extraction, there are different methods which are being used, such as fractional MFCC, linear predictive coefficients (LPC), Perceptual Linear Prediction (PLP) and Power-Normalized Cepstral Coefficients (PNCC), since all of these techniques accomplish good results. Apart from this, combining information from more sources is presented in many papers, in order to achieve significant benefits from the advantages of every source. A good example is the combination of MFCC and PNCC [24], where the MFCC and PNCC techniques are combined together in order to enhance the performances of the of the ASR. Another example is the combination of PLP and MFCC (MF-PLP) in [25].

Table 7. Comparison between various ASR systems.

Researchers	Year	Method	Overall performance (%)
[26]	2016	Normalized PNCC+GMM/UBM Fusion (Normalized MFCC+ Normalized PNCC) +GMM/UBM	89.17 95
[27]	2013	Fractional MFCC+ Generalized Linear Discriminant sequence kernel	90.78
[23]	2012	LPC+GMM/SVM	96,42
[28]	2010	Weighted dynamic MFCC+GMM	92.8
[25]	2010	MF-PLP + Iterative clustering approach	86
[29]	2010	GMM	92.98
[30]	2009	MFCC, delta coefficient+ Combination of MLP, SVM and decision trees	94.4

5. Conclusions

The paper addressed the issue of automatic text-independent speaker identification. The contribution of the manuscript can be broadly summarized under the following points.

- Various pre-processing stages prior to feature extraction and LVQ configuration were studied and implemented on TIMIT database.
- In the feature extraction stage, different feature extraction techniques like MFCC, Delta MFCC, Delta-Delta MFCC and their combinations are explored. The combination of MFCC, Delta MFCC, Delta-Delta MFCC features provided 90% performance against 86% for only MFCC in the initial experiment.
- In the recognition stage, different codebook vectors numbers were employed. The best configuration was for 196 vectors. The combined LVQ and GMM-UBM classifier provides 86% performance against 15% for LVQ. These experiments indicate the usefulness of this combination in enhancing the speaker identification system performance.
- Two speaker identification gender-dependent systems are built. They seem to be more accurate than the gender-independent speaker identification system. Gender information is believed helpful for speaker identification.

Although the results were positive and promising, there were some limitations. First, this study is only related to one dialect. Therefore to generalize the results for larger groups of speakers, the study should involved more speakers from all the TIMIT dialects. Second, speech which is a behavioral signal, that may not be consistently reproduced by the speaker (mental state health), is only represented in this paper by spectral and dynamic features. It will be interesting to combine these spectral features with prosodic and voice quality features. Finally, the effectiveness of the proposed method was verified using clean speech. Further evaluation will be conducted on noisy data and can be evaluated under speaker verification case.

References

1. Drygajlo, A. (2012). Automatic speaker recognition for forensic case assessment and interpretation. *Forensic Speaker Recognition*. Springer New York.
2. Hillenbrand, J.M.; and Clark, M.J. (2009). The role of f0 and formant frequencies in distinguishing the voices of men and women. *Attention, Perception, & Psychophysics*, 71(5), 1150-1166.
3. Abdel-Hamid, O.; and Jiang, H. (2013). Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 7942-7946.
4. Kekre, H.B.; and Kulkarni, V. (2010). Speaker identification by using vector quantization. *International Journal of Engineering Science and Technology* (*JESTEC*), 2(5), 1325-1331.
5. Shanmugapriya, P.; and Venkataramani, Y. (2015). FLVQ based GMM in speaker verification. *Journal of Applied Sciences*, *15*(2), 295.
6. Motlicek, P.; Dey, S.; Madikeri, S.; and Burget, L. (2015). Employment of subspace gaussian mixture models in speaker recognition. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4445-4449.

7. Trabelsi, I.; and Ben Ayed, D. (2012). On the use of different feature extraction methods for linear and non linear kernels. *Proceedings of the IEEE 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*, 797-802.
8. Trinh, T. D.; Park, M. K., Kim; J. Y.; Lee, K. R.; and Cho, K. (2015). Enhanced speaker verification using GMM-supervector based modified adaptive GMM training. *Mobile and Wireless*, Springer Berlin Heidelberg.
9. Ayoub, B.; Jamal, K.; and Arsalane, Z. (2015). An analysis and comparative evaluation of MFCC variants for speaker identification over VoIP networks. *In IEEE World Congress on Information Technology and Computer Applications Congress (WCITCA),* 1-6.
10. Sukhwal, A.; and Kumar, M. (2015). Comparative study of different classifiers based speaker recognition system using modified MFCC for noisy environment. *Proceedings of the International Conference on Green Computing and Internet of Things (ICGCIoT),* 976-980.
11. Trabelsi, I.; and Ben Ayed, D. (2013). A multi level data fusion approach for speaker identification on telephone speech. *International Journal of Signal Processing, Image Processing & Pattern Recognition*, 6(2), 33-42.
12. Aggarwal, N. (2015). Analysis of various features using different temporal derivatives from speech signals. *International Journal of Computer Applications*, 118(8), 1-9.
13. Reynolds, D.A. (2001). Automatic speaker recognition: Current approaches and future trends. *Speaker Verification: From Research to Reality*, 14-15.
14. Reynolds, D.A. (1995). Automatic speaker recognition using Gaussian mixture speaker models. In *The Lincoln Laboratory Journal*.
15. Dempster, A.P.; Laid, N.M.; and Durbin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1), 1-38.
16. Kohonen, T.; Hynninen, J.; Kangas, J.; Laaksonen, J.; and Torkkola, K. (1995). LVQ PAK: *The learning vector quantization program package. Technical report.* Laboratory of Computer and Information Science Rakentajanaukio 2 C.
17. Trabelsi, I.; and Bouhlel M-S. (2015). Feature selection for gumi kernel-based svm in speech emotion recognition. *International Journal of Synthetic Emotions (IJSE)*, 6(2), 57-68.
18. Trabelsi, I.; and Bouhlel M-S. (2016). Dynamic sequence-based learning approaches on emotion recognition systems. *Proceedings of International Conference on Automation, Control Engineering and Computer Science*, 625-631.
19. Trabelsi, I.; Ben Ayed, D.; and Ellouze, N. (2016). Comparison between GMM-SVM Sequence Kernel and GMM: Application to speech emotion recognition. *Journal of Engineering Science and Technology* (*JESTEC*), 11(9), 1221- 1233.
20. Kohonen, T.; Hynninen, J.; Kangas, J.; Laaksonen, J.; and Torkkola, K. (1996). *LVQ PAK: The learning vector quantization program package*. Technical report, Laboratory of Computer and Information Science Rakentajanaukio 2 C, 1991-1992.

21. Zue, V.; Seneff, S.; and Glass, J. (1990). Speech database development at MIT: TIMIT and beyond. *Speech Communication*, *9*(4), 351-356.
22. Brookes, M. (1997). Voicebox: Speech processing toolbox for matlab Software. Retrieved March 14, 2016, from http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.
23. Trabelsi, I.; and Bouhlel M.-S. (2016). A multi features fusion support vector machine for classification of emotion issue in the design of an audio recognition system. *International Journal of Applied Pattern Recognition* 3(2), 181-196.
24. Revathi, A.; and Venkataramani, Y. (2008). Iterative clustering approach for text independent speaker identification using multiple features. In IEEE *2nd International Conference on Signal Processing and Communication Systems (ICSPCS 2008),* 1-6.
25. Ouzounov, A. (2010). Cepstral features and text-dependent speaker identification-a comparative study. *Cybernetics and Information Technologies*, 10(1), 3-12.
26. Al-Kaltakchi, M.T.S.; Woo, W.L.; Dlay, S.S.; and Chambers, J.A. (2016). Study of fusion strategies and exploiting the combination of MFCC and PNCC features for robust biometric speaker identification. *Proceedings of IEEE 4th International Conference on Biometrics and Forensics (IWBF),* 1-6.
27. Ajmera, P.K.; and Holambe, R.S. (2013). Fractional Fourier transform based features for speaker recognition using support vector machine. *Computers & Electrical Engineering*, 39(2), 550-557.
28. Weng, Z.; Li, L.; and Guo, D. (2010). Speaker recognition using weighted dynamic MFCC based on GMM. In IEEE *International Conference on Anti-Counterfeiting Security and Identification in Communication (ASID),* 285-288.
29. Naseem, I.; Togneri, R.; and Bennamoun, M. (2010). Sparse representation for speaker identification. *In IEEE International Conference on Pattern Recognition (ICPR),* 4460-4463.
30. Boujelbene, S.Z.; Mezghani, D.B.A.; and Ellouze, N. (2009). Application of combining classifiers for text-independent speaker identification. In *IEEE International Conference on Electronics, Circuits, and Systems, 2009. (ICECS 2009),* 723-726.