An experimental comparison of modelling techniques for speaker recognition under limited data condition

H S JAYANNA and S R MAHADEVA PRASANNA

Department of Electronics and Communication Engineering, Indian Institute of Technology Guwahati, Guwahati 781 039
e-mail: {h.jayanna, prasanna}@iitg.ernet.in

Abstract. Most of the existing modelling techniques for the speaker recognition task make an implicit assumption of sufficient data for speaker modelling and hence may lead to poor modelling under limited data condition. The present work gives an experimental evaluation of the modelling techniques like Crisp Vector Quantization (CVQ), Fuzzy Vector Quantization (FVQ), Self-Organizing Map (SOM), Learning Vector Quantization (LVQ), and Gaussian Mixture Model (GMM) classifiers. An experimental evaluation of the most widely used Gaussian Mixture Model–Universal Background Model (GMM–UBM) is also made. The experimental knowledge is then used to select a subset of classifiers for obtaining the combined classifiers. It is proposed that the combined LVQ and GMM–UBM classifier provides relatively better performance compared to all the individual as well as combined classifiers.

Keywords. Speaker recognition; limited data; CVQ; FVQ; SOM; LVQ; GMM; GMM–UBM.

1. Introduction

The objective of speaker recognition is to recognize the speaker from the speech signal (Atal 1976). State-of-the-art speaker recognition systems assume the availability of sufficient data for speaker modelling and testing. However, there are certain applications in practice where the available speech data is limited. For instance, speaker recognition in non-cooperative scenario and speaker tracking and segmentation. In the present work, *sufficient data* is used to denote the case of having speech data of a few minutes (more than one minute). Alternatively, *limited data* denotes the case of having speech data of a few seconds (less than 15 seconds) (Angkititrakul & Hansen 2007). The significance of the amount of speech data for speaker modelling and testing has been studied earlier (Prasanna *et al* 2006). This study experimentally demonstrates that when the speech data for training is less, then the performance is poor due to poor speaker modelling and also speech data is insufficient to make reliable decision during testing. Therefore, the objective of speaker recognition under limited data condition is to obtain as good and reliable performance as possible.

Speaker recognition can be classified as speaker verification and speaker identification. Speaker verification deals with validating the identity claim of the speaker. Speaker identification deals with identifying the most likely speaker of the test speech data. Speaker identification can be further classified into closed-set or open-set modes. Closed-set speaker identification refers to the case where the speaker is a member of the set of N enrolled speakers. In open-set speaker identification, the speaker may also be from outside the set of N enrolled speakers. Speaker recognition can be operated in either text-dependent or text-independent mode. In text-dependent mode, speech for the same text is used for both training and testing. No such restrictions are imposed in text-independent mode. The present work focuses on text-independent, closed-set, speaker identification.

Speaker recognition system may be considered to consists of four stages. They include: speech analysis, feature extraction, speaker modelling and speaker testing. Speech analysis involves analysing the speech signal using suitable frame size and shift for the feature extraction. Feature extraction involves extracting speaker-specific features from the speech signal at reduced data rate. The extracted features are further combined using modelling techniques to generate speaker models. The speaker models are then tested using the features extracted from the test speech signal. The improvement in the performance can be achieved by employing new or improved techniques in one or more of these stages. Earlier we have demonstrated that the Variable Frame Size and Rate (VFSR) analysis under limited data condition improved performance over the existing Fixed Frame Size and Rate (FFSR) analysis (Jayanna & Prasanna 2006). Similarly, it may be possible to develop new modelling techniques suitable for limited data condition and use them instead of existing modelling techniques. This may also improve the speaker recognition performance. Hence the motivation for the present work.

State-of-the-art speaker recognition systems employ various modelling techniques like Crisp Vector Quantization (CVQ), Fuzzy Vector Quantization (FVQ), Self-Organizing Map (SOM), Learning Vector Quantization (LVQ) and Gaussian Mixture Model (GMM). The success of each of the modelling techniques depends on the principle employed for clustering. Among theses modelling techniques, the widely used one is GMM (Reynolds 1995). The success of GMM is due to the availability of sufficient data for speaker modelling (Angkititrakul & Hansen 2007). Recently, some attempts have been made to recognize the speakers under limited data condition using the concept of Gaussian Mixture Model–Universal Background Model (GMM–UBM) (Angkititrakul & Hansen 2007) (Prakash & Hansen 2007). In another attempt the authors have proposed that by selecting those feature vectors which provide good speaker discrimination, it is possible to identify speakers under limited data (Kwon & Narayanan 2007).

In this work, first we evaluate the performance of various modelling techniques and then combine some of them to improve the performance. The modelling techniques may offer different information about the patterns to be classified due to the difference in the working principle employed and hence could be used to improve the performance in a combined modelling system (Kittler *et al* 1998). This is the motivation for combining the different models. For instance, in case of CVQ the feature vectors are clustered into non-overlapping clusters, whereas in case of FVQ, the feature vectors are clustered into overlapping clusters. Thus since the principle of clustering is different, it may be possible to combine these modelling techniques to obtain a combined modelling technique. The rest of the paper is organized as follows: Speech database details for the study are discussed in section 2. In section 3, the speaker recognition studies using different modelling techniques are discussed. Section 4 presents the proposed combined modelling techniques for speaker recognition. Summary of the present work and the possible future directions are mentioned in section 5.

2. Speech database

In order to evaluate the performance of the speaker recognition system, the YOHO (Campbell Jr. 1995) and the TIMIT (Zue & Glass 1990) databases are used. The YOHO database consists of speech data from 138 speakers. The speech data is sampled at 8 kHz and stored with 16 bits/sample resolution. The training data for each speaker includes 96 speech files, each of about 3 sec duration. The testing data for each speaker includes 40 speech files each of about 3 sec duration. Since the original database is not meant for limited data condition, we have taken one, two, four and eight speech files of each speaker to create modified database.

The TIMIT database consists of speech data from 462 speakers in the training set and 168 speakers in the test set. The speech data is collected over microphone, sampled at 16 kHz and stored with 16 bits/sample resolution. Since most of the speech information is present up to 4 kHz, the speech database is resampled to 8 kHz. The speech data for each speaker includes 10 speech files, each of about 3 sec duration. In this work, we have used one set of first 30 speakers and another set of first 138 speakers from the test set of the TIMIT database. The first 5 speech files of each speaker are used for training and the remaining for testing. This database is also not meant for limited data condition and hence we have taken one, two, four and five speech files of each speaker to create modified database.

3. Speaker recognition studies

In this work, the initial studies are conducted using the training data and test data of one file (3 sec) from each of the first 30 speakers of the YOHO database. These studies are later extended to the data of all the 138 speakers from the YOHO database and to the data of first 30 and first 138 speakers from the test set of the TIMIT database. In all our experiments, speech is analyzed in frames of 20 ms with shifts of 10 ms. For each frame, excluding c_0, 13 dimensional Mel-Frequency Cepstral Coefficients (MFCC) are extracted as feature vectors (Deller *et al* 1993). Cepstral Mean Subtraction (CMS) is applied to remove the linear channel effect. Silence and low energy speech frames are removed using an energy-based frame selection technique (Deller *et al* 1993). The threshold used for selection of the speech frames is 0·1 times the average frame energy. The extracted features are used for modelling the speakers by different modelling techniques. While testing, each feature vector of the test speech data is compared with all the speakers models. The speaker model which has the minimum distance (Euclidean distance) or maximum *a posteriori* probability is recognized as the tentative speaker of the speech frame. The speaker with the assignment of maximum number of frames is recognized as the final speaker of the test speech data. The speaker recognition rate depends on the amount of training and testing data and also on the codebook size or the number of Gaussian mixtures. The speaker recognition system is therefore evaluated for different values of these parameters.

3.1 *Speaker modelling by Direct Template Matching (DTM)*

When the amount of available data is small, the number of feature vectors is also small. For instance, assuming about 80% speech frames, for 3 sec of speech signal there are about 240 feature vectors. Since the number of feature vectors is insufficient, we can use direct template matching to find the speaker recognition rate. In DTM technique, during the identification phase, the test feature vector of an unknown speaker is compared with all the reference training feature vectors to identify tentative speaker of the speech frame. This process is repeated

Table 1. Speaker recognition rates (%) for the 30 speakers of the YOHO database using *3 sec training and testing data* for CVQ modelling technique.

Modelling technique	Codebook size			
	16	32	64	128
CVQ	63·33	66·67	**70·00**	60·00

for all the testing frames. The speaker with maximum number of frames is identified as the speaker of the test speech data. In the 30 speakers case of the YOHO database, the recognition rate of 63·33% is obtained for one speech file (3 sec) of training and testing data. Though this technique is simple and easy to perform, the recognition rate is poor. The poor recognition rate is due to large intraspeaker and inter speaker variability. This shortcoming may therefore be reduced using different modelling techniques. The objective of modelling technique is to better cluster or capture the distribution of the feature vectors according to the speaker information. Speaker models built contain the feature vectors from different sound units, but from the same speakers. This may enable dominance of speaker information over speech information. This aspect is verified in limited data condition using the following speaker modelling techniques.

3.2 *Speaker modelling using CVQ*

CVQ is also termed as Vector Quantization (VQ) (Gray 1984). VQ involves finding a subset of feature vectors termed as *Codevectors* from the whole set, which can act as representative vectors. To find the codevectors for a given speaker, CVQ clusters all the feature vectors in the feature space into non-overlapping clusters with crisp boundaries and hence the name. The lookup table of codevectors is termed as *codebook*. The codebooks of different sizes are built using the binary split and *k-means* clustering procedures during training (Gray 1984). The *k-means* clustering involves grouping the input feature vectors into non-overlapping *k*-clusters. For given size, one codebook is built for each speaker in the database. The feature vectors of the test speech data are compared with the codebooks of different speakers to find out the most likely speaker of the test speech signal. The experimental results using CVQ for the 30 speakers of the YOHO database using one speech file (3 sec) for training and testing data are given in table 1. The highest recognition rate of 70% is achieved using a codebook size of 64.

As we have already mentioned 3 sec data provides 240 frames. This number is too small for forming a CVQ codebook of size 64, since as a thumb rule, the number of feature vectors should be about 10 times the number of non-overlapping clusters (Rabiner & Juang 1993). Accordingly, CVQ with codebook of size 16 seems to be optimum. However, as per the result obtained, even higher codebook of sizes like 32 and 64 give higher recognition rates. This implies that for limited data condition about 5 times the codebook size may be kept as thumb rule, while deciding the codebook size. Accordingly, further increase in the codebook size to 128 gives poor recognition rate. Better recognition rate of CVQ compared to DTM implies that, it may be better to use some techniques for modelling. A maximum recognition rate of only 70% is due to the limited training and testing data and also the modelling technique employed. Therefore, for given training and test data, to increase the recognition rate, we can explore alternate modelling techniques.

Table 2. Speaker recognition rates (%) for the 30 speakers of the YOHO database using *3 sec training and testing data* for FVQ modelling technique.

Learning rate	Codebook size			
	16	32	64	128
1·30	**70·00**	60·00	60·00	60·00
1·31	70·00	63·33	66·67	56·67
1·32	70·00	63·33	60·00	56·67
1·33	70·00	73·33	66·67	60·00
1·34	63·33	70·00	66·67	60·00
1·35	63·33	70·00	63·33	60·00
1·38	63·33	70·00	70·00	60·00
1·39	60·00	**76·67**	66·67	60·00
1·40	60·00	76·67	70·00	66·67
1·45	66·67	63·33	66·67	63·33
1·50	60·00	76·67	**73·33**	**70·00**
1·55	63·33	73·33	73·33	66·67
1·60	56·67	66·67	60·00	60·00

3.3 *Speaker modelling using FVQ*

FVQ is an alternative to CVQ and employs fuzzy logic principle for clustering. The basic principle of fuzzy logic is that a given feature vector can be assigned to more than one cluster with certain degree of association to find the codevectors for a given speaker. FVQ clusters all the feature vectors in the feature space into overlapping clusters with fuzzy boundaries and hence the name (Bezdek & Harris 1978). In FVQ each feature vector is assigned to all the clusters, but with different degrees of association, as dictated by the membership function. The merit of FVQ compared to CVQ is that, since all the feature vectors are associated with all the clusters, there are relatively more number of feature vectors for each cluster and hence the codevectors may be more reliable. The codebooks of different sizes are built using binary split and fuzzy *c-means* clustering procedures during training (Bezdek & Harris 1978). Fuzzy *c-means* clustering involves grouping the input feature vectors into overlapping *c-clusters*. The nature of clustering depends strongly on the learning rate parameter hence it needs to be tuned for better recognition rate. The feature vectors of the test speech data are compared with the codebooks of different speakers as in the case of CVQ. The experimental results using FVQ for the 30 speakers of the YOHO database using one speech file (3 sec) for training and testing data and different learning rate parameter are given in table 2. The highest recognition rate of 76·67% is achieved for a codebook of size 32 using a learning rate parameter of 1·39. For the same amount of speech data (3 sec), we are able to further increase the recognition rate from 70% of CVQ to 76·67%. This can be attributed to the fuzzy *c-means* clustering employed in FVQ.

The better recognition rate by FVQ suggests that by increasing the number of elements for clustering, the recognition rate also increases. This is achieved by associating the same set of feature vectors to different clusters, of course, with different membership functions. This improvement in recognition rate is at the cost of increased computational complexity of tuning the learning rate parameter. However, it is still preferable due to the small amount of data. On the similar lines we can also explore other VQ modelling techniques based on neural networks.

Table 3. Speaker recognition rates (%) for the 30 speakers of the YOHO database using *3 sec training and testing data* for SOM modelling technique.

Iterations	*h*	*η*	Codebook Size (CS)			
			16	32	64	128
500*CS	1	0·01	60·00	60·00	53·33	66·67
500*CS	1	0·02	53·33	70·00	**70·00**	66·67
500*CS	1	0·03	60·00	70·00	66·67	66·67
500*CS	1	0·04	56·67	60·00	63·33	70·00
500*CS	1	0·05	**70·00**	60·00	60·00	63·33
500*CS	1	0·06	66·67	70·00	70·00	**73·33**
500*CS	1	0·07	70·00	56·67	63·33	53·33
500*CS	1	0·08	66·67	63·33	63·33	53·33
500*CS	1·1	0·06	66·67	70·00	56·67	56·67
500*CS	1·2	0·06	70·00	66·67	56·67	66·67
500*CS	1·4	0·06	66·67	60·00	66·67	56·67
550*CS	1	0·06	73·33	**73·33**	63·33	56·67
600*CS	1	0·06	53·33	70·00	53·33	70·00
650*CS	1	0·06	63·33	66·67	70·00	60·00

3.4 *Speaker modelling using SOM*

A neural network counter part of VQ, but with unsupervised learning can be realized using SOM. The approach for identifying the codevectors is by learning in an unsupervised way. The clustering is therefore influenced by the actual distribution of feature vectors and hence the modelling may be different. SOMs are a special class of neural networks based on competitive learning (Kohonen 1990). Thus, the performance of the SOM depends on the parameters such as neighbourhood (h), learning rate (η) and number of iterations. The recognition rate for 30 speakers of the YOHO database using one speech file (3 sec) for training and testing data and different values of h, η and number of iterations are given in table 3. The highest recognition rate of 73·33% is achieved for a codebook of size 32 using $h = 1$, $\eta = 0.06$, and the number of iterations equal to 500 times the codebook size.

The recognition rate of 73·33% by SOM using unsupervised learning implies that, even the feature vectors from limited data provide speaker information in the feature space. Further, each speaker has a unique distribution of feature vectors which is learnt by SOM. It may be possible to improve the recognition rate of SOM by using the LVQ.

3.5 *Speaker modelling using LVQ*

LVQ developed by Kohonen (1990) is used to globally optimize the codebooks after they are generated with unsupervised learning algorithm like SOM. LVQ is a supervised learning technique that uses class information to optimize the positions of codevectors obtained by SOM, so as to improve the quality of the classifier decision regions. An input vector is picked at random from the input space. If the class label of the input vector and the codevector agree, then the codevector is moved in the direction of the input vector. Otherwise the codevector is moved away from the input vector. Therefore, due to this fine tuning there may be improved recognition rate compared to SOM. The recognition rate for the 30 speakers of the YOHO database using one speech file (3 sec) for training and testing data and different η and iterations are given in table 4. The LVQ gives the best recognition rate of 80% for a codebook of size 32 which is better than that of all other VQ techniques discussed so far.

Table 4. Speaker recognition rates (%) for the 30 speakers of the YOHO database using *3 sec training and testing data* for LVQ modelling technique.

Iterations	η	Codebook Size (CS)			
		16	32	64	128
500*CS	0·01	63·33	60·00	66·67	60·00
500*CS	0·02	60·00	66·67	63·33	53·33
500*CS	0·03	60·00	66·67	**73·33**	60·00
500*CS	0·04	53·33	63·33	66·67	**63·33**
500*CS	0·05	63·33	76·67	66·67	66·67
500*CS	0·06	70·00	60·00	70·00	60·00
500*CS	0·08	63·33	70·00	63·33	60·00
550*CS	0·05	63·33	53·33	70·00	63·33
550*CS	0·06	**73·33**	**80·00**	60·00	60·00
575*CS	0·06	56·67	66·67	66·67	53·33
600*CS	0·06	63·33	70·00	66·67	63·33
700*CS	0·06	60·00	63·33	73·33	60·00

The improvement in the recognition rate compared to SOM implies that employing supervised learning over initially obtained unsupervised codevectors indeed improves the recognition rate. Thus the fine tuning by LVQ is beneficial under limited data condition also. The aforementioned modelling techniques are based on non-parametric clustering approach. Speaker modelling by parametric probabilistic approach like GMM and GMM–UBM can also be explored.

3.6 *Speaker modelling using GMM*

The GMM is the most widely used probabilistic modelling technique in speaker recognition. The GMM needs sufficient data (at least one minute) to model the speaker well to yield good recognition rate (Reynolds & Rose 1995). Unlike the centroids design, as we discussed in the above modelling techniques, in GMM system the distribution of feature vectors is modelled by the parameters like weight, mean and covariance (Reynolds & Rose 1995). Since in our experimental conditions training and testing data are limited, GMM may not be the best choice. However, we conducted the experiment using VQ initialized GMM to see its effectiveness under limited data condition. The experimental results for the 30 speakers of the YOHO database using one speech file (3 sec) for training and testing data and different Gaussian mixtures are given in table 5. The GMM yields the highest recognition rate of 73·33% using 16 Gaussian mixtures.

The recognition rate of GMM-based system is better compared to CVQ, but poor compared to all other VQ modelling techniques. This means that the data may be too sparse to model

Table 5. Speaker recognition rates (%) for the 30 speakers of the YOHO database using *3 sec training and testing data* for GMM modelling technique.

Modelling technique	Gaussian mixtures			
	16	32	64	128
GMM	**73·33**	40·00	36·67	13·33

Table 6. Speaker recognition rates (%) for the 30 speakers of the YOHO database using *3 sec training and testing data* for GMM–UBM modelling technique.

Modelling technique	Gaussian mixtures			
	16	32	64	128
GMM–UBM–NIE	60·00	60·00	63·33	**76·67**
GMM–UBM–IE	60·00	66·67	73·33	**83·33**

by the Gaussian mixtures. To alleviate this problem to some extent the concept of Universal Background Model (UBM) can be used along with GMM.

3.7 *Speaker modelling using GMM–UBM*

The concept of GMM–UBM (Reynolds *et al* 2000) is widely used for speaker recognition where the availability of training data is sparse (Angkititrakul & Hansen 2007, Prakash & Hansen 2007). In case of GMM–UBM system, speech data collected from large number of speakers is pooled and the UBM is trained. The UBM model parameters represent the characteristics of all speakers and hence UBM acts as a speaker-independent model. The speaker-dependent model can be created by performing maximum *a posteriori* (MAP) estimation from the UBM using speaker-specific training speech. The UBM training can be done in two ways: (i) Speech data pooled from the other database, not used for the speaker recognition study, provided speech data is collected from the same environment. (ii) Same speech data for both UBM training and evaluation, provided the speakers set used for recognition is not included in UBM training (Angkititrakul & Hansen 2007, Prakash & Hansen 2007, Reynolds *et al* 2000). We conducted the study using the YOHO database for both UBM training and evaluation. Since our experimental study considers evaluation set of first 30 and 138 speakers, experiments are conducted with Not Including Evaluation set (NIE) and Including Evaluation set (IE) in UBM training. In (Reynolds *et al* 2000), it is mentioned that the number of speakers and amount of data to train the UBM are randomly selected. We trained the UBM with roughly two hours of speech data, equally contributed by speakers selected from the YOHO database. The experimental results for the 30 speakers of the YOHO database using one speech file (3 sec) for training and testing data and different Gaussian mixtures are given in table 6. The GMM–UBM yields the highest recognition rate of 76·67% and 83·33% using 128 Gaussian mixtures for NIE and IE, respectively.

The recognition rate obtained by not including the speakers in building UBM i.e. GMM–UBM–NIE is the actual result for GMM–UBM. The recognition rate of 76·67% by the same implies that UBM does not seem to provide any benefit in terms of improving the recognition rate. The higher recognition rate of 83·33% for GMM–UBM–IE is due to the data of each of the speakers used in building the GMM–UBM–IE. Hence there is bias in the UBM towards each of the speakers.

4. Speaker modelling using combined modelling techniques

The modelling techniques discussed so far are different with respect to their working principle and hence may be combined to further improve the recognition rate. The proposed combination technique works as follows: Let $x_1, x_2, \ldots, x_N$ be the frame scores obtained for the

Table 7. Best individual and combined modelling speaker recognition rates (%) for the 30 speakers of the YOHO database using *3 sec training and test data* for different modelling techniques.

Modelling techniques	Codebook size/Gaussian mixtures			
	16	32	64	128
CVQ	63·33	66·67	**70·00**	60·00
FVQ	70·00	**76·67**	73·33	70·00
SOM	73·33	**73·33**	70·00	73·33
LVQ	73·33	**80·00**	73·33	66·67
GMM	**73·33**	40·00	36·67	13·33
GMM–UBM–NIE	60·00	60·00	63·33	**73·33**
GMM–UBM–IE	60·00	66·67	73·33	**83·33**
LVQ–FVQ		**80·00**		
LVQ–GMM		**80·00**		
LVQ–GMM–UBM–NIE		**83·33**		
LVQ–GMM–UBM–IE		**86·67**		

test data of a speaker using the modelling technique M_1 with N. Similarly, $y_1, y_2, \ldots, y_N$ are the frame scores obtained for the same test data using the modelling technique M_2. Then, the corresponding speaker frame scores are linearly added which results in $z_1, z_2, \ldots, z_N$. The speaker with the combined highest frame score is recognized as the final speaker of the test speech data. The frame score specifies the score of each speaker for the test speech data and hence the speaker who scores the highest frames will be the recognized speaker of the test data. On the other hand, the recognition rate specifies the total number of correctly identified speakers out of N speakers considered for the study. The best recognition rate of individual models and the recognition rate of different combined models are given in table 7. Among the combined modelling techniques, the LVQ–GMM–UBM–IE and LVQ–GMM–UBM–NIE systems yield the highest recognition rate of 83·33% and 86·67%, respectively. The improvement in the recognition rate is due to the different working principles employed in LVQ and GMM–UBM. That is, the supervised learning over unsupervised learning involved in LVQ and other speakers data used as UBM in GMM–UBM. Moreover, LVQ modelling technique is based on non-parametric approach, whereas GMM–UBM based on parametric approach and hence this combination gives the best recognition rate. Further, in the other combined techniques like LVQ–FVQ and LVQ–GMM the working principles are different. However, the FVQ and LVQ are fine tuned using only the speaker-specific speech data which may not be optimum and hence the combination techniques using these modelling techniques yield lower recognition rate compared to LVQ–GMM–UBM.

For the other data sizes of 6, 12 and 24 sec we conducted the study only with LVQ, GMM–UBM and the combined LVQ–GMM–UBM modelling techniques. The experimental results are shown in figure 1a. It is evident from the figure that the recognition rate of GMM–UBM–NIE below 10 sec of training and testing data is less than that of LVQ and GMM–UBM–IE. This means that the available training data is insufficient to train the speaker-dependent model in GMM–UBM. Under such conditions the combined system gives better recognition rate than the individual systems. Also, GMM–UBM–IE recognition rate is higher than that of the other individual techniques even for data of less than 10 sec duration. This is due to the availability of speaker-specific sufficient data while training the UBM model. The proposed combined modelling technique shows significant improvement in the recognition rate up to
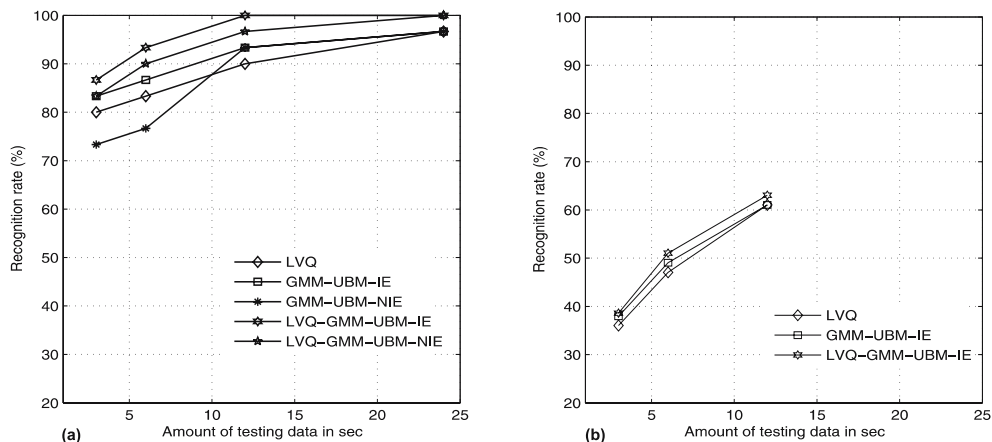
Figure 1. Speaker recognition rates is based on LVQ, GMM–UBM and LVQ–GMM-UBM modelling for different sizes of training and testing data for **(a)** first 30 and **(b)** 138 speakers taken from the YOHO database.

12 sec and above 12 sec the recognition rate of all modelling techniques approach one another. Therefore, in order to verify the recognition rate for the whole database, the experiment is carried out up to 12 sec training and testing data and the results are shown in figure 1b. The trend in the experimental results shown in figure 1b resemble with figure 1a which imply that the proposed combined modelling technique shows a similar behaviour for the large database also.

To verify the robustness of the proposed combined modelling technique, we conducted the experiments on the TIMIT database also. In the GMM–UBM modelling technique, we used the TIMIT training set to train the UBM roughly for 2 hours of data. The speaker recognition experiments are conducted on the TIMIT test set. Experimental studies are conducted as the YOHO database set-up and the results are shown in figures 2a and b for a set of first 30 and 138 speakers, respectively. The experimental results for the TIMIT database also resemble those for the YOHO database irrespective of speaker population and amount of data. Hence, the LVQ–GMM–UBM can be used as a modelling technique for speaker recognition under limited data condition.

5. Summary and conclusions

In this paper, we explored the different modelling techniques and then proposed combined LVQ–GMM–UBM modelling technique for speaker recognition under limited data condition. First, we discussed the working principles and the efficiency of different modelling techniques. Then, we combined different modelling techniques to see the effectiveness. As a result, we found that the combined LVQ–GMM–UBM model gives better recognition rate than the individual and other combined modelling techniques. Therefore, LVQ–GMM–UBM model can be used for speaker modelling.

In the present work, effectiveness of combined LVQ–GMM–UBM model is verified using clean speech data. The effectiveness of the combined modelling needs to be verified on noisy speech data. Further, this work used linear combination of frame scores obtained for the same test data using different modelling techniques to identify a speaker. Different combination
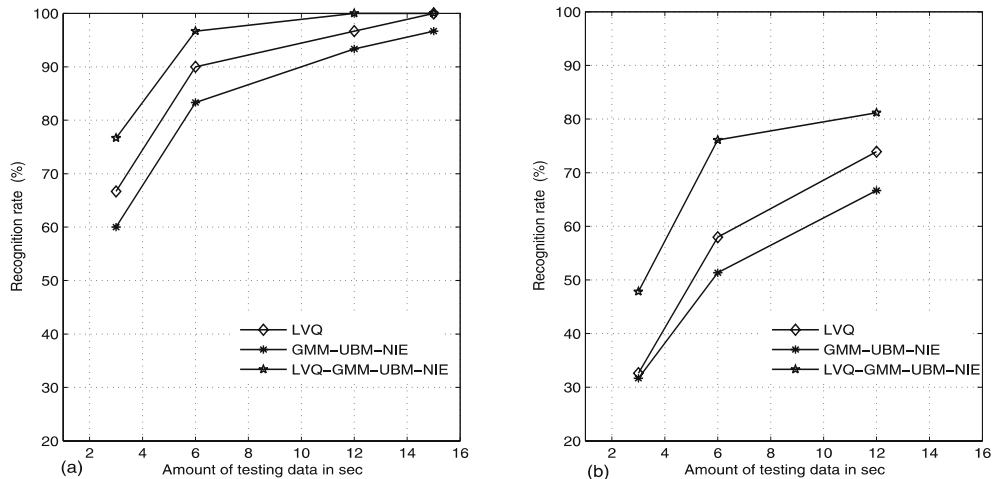
Figure 2. Speaker recognition rate based on LVQ, GMM–UBM–NIE and LVQ–GMM–UBM–NIE modelling for different sizes of training and testing data for a set of **(a)** first 30 and **(b)** first 138 speakers taken from the TIMIT test set.

techniques need to be explored to improve the speaker recognition rate. Also, this work can be extended further to improve the recognition rate by developing new techniques in speech analysis, feature extraction and testing stages of the speaker recognition system.

References

Angkititrakul P, Hansen J H L 2007 Discriminative In-Set/Out-of-Set speaker recognition. *IEEE Trans. Audio Speech Language Process.* 15(2): 498—508

Atal B S 1976 Automatic recognition of speakers from their voices, *Proc. IEEE* 64(4): 460—475

Bezdek J C, Harris J D 1978 Fuzzy portions and relations; an axiomatic basis for clustering. *Fuzzy Sets and Systems* 1: 111—127

Campbell Jr J P 1995 Testing with the YOHO CD-ROM voice verification corpus. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* Detroit, Michigon 341—344

Deller J, Hansen J, Proakis J 1993 *Discrete Time Processing of Speech Signals*, 1st ed. IEEE Press

Gray R 1984 Vector quantization *IEEE Acoust., Speech, Signal Process. Mag.* 1: 4—29

Jayanna H S, Prasanna S R M 2006 Variable segmental analysis based speaker recognition in limited data condition. In *Proc. IEEE-Int. Conf. Signal, Image Process* vol. 2 Karnataka, India

Kittler J, Hatef M, Duin R P W, Matas J 1998 On combining classifiers. *IEEE Trans. Patt. Anly. Machine Intelligence* 20(3): 226-–239

Kohonen T 1990 The self-organizing map. *Proc. IEEE* 78(9): 1464-–1480

Kwon S, Narayanan S 2007 Robust speaker identification based on selective use of feature vectors. *Patt. Recog. Lett.* 28: 85—89

Prakash V, Hansen J H L 2007 In-Set/Out-of-Set speaker recognition under sparse enrollment. *IEEE Trans. Audio Speech Language Process* 15(7): 2044—2051

Prasanna S R M, Gupta C S, Yegnanarayana B 2006 Extraction of speaker-specific excitation information from linear prediction residual of speech. *Speech Communication* 48: 1243—1261

Rabiner L, Juang B H 1993 Fundamentals of Speech Recognition. (Singapore: Pearson Education)

Reynolds D A, Rose R C 1995 Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Trans. Speech Audio Process* 3(1): 72—83

Reynolds D A 1995 Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication* 17: 91—108.

Reynolds D A, Quateri T F, Dunn R B 2000 Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing* 10: 19—41

Zue S S V, Glass J 1990 Speech database development at MIT:TIMIT and beyond. *Speech Communication* 9: 351-–356