*Sādhanā* Vol. 35, Part 5, October 2010, pp. 525–546. © Indian Academy of Sciences

# Limited data speaker identification

H S JAYANNA[1,*] and S R MAHADEVA PRASANNA[2]

[1]Department of Information Science and Engineering, Siddaganga Institute of Technology, Tumkur 572 103
[2]Department of Electronics and Communication Engineering, Indian Institute of Technology Guwahati, Guwahati 781 039
e-mail: jayannahs@sit.ac.in; prasanna@iitg.ernet.in

MS received 19 November 2009; revised 13 August 2010; accepted 17 August 2010

**Abstract.** In this paper, the task of identifying the speaker using limited training and testing data is addressed. Speaker identification system is viewed as four stages namely, analysis, feature extraction, modelling and testing. The speaker identification performance depends on the techniques employed in these stages. As demonstrated by different experiments, in case of limited training and testing data condition, owing to less data, existing techniques in each stage will not provide good performance. This work demonstrates the following: multiple frame size and rate (MFSR) analysis provides improvement in the analysis stage, combination of mel frequency cepstral coefficients (MFCC), its temporal derivatives ($\Delta$, $\Delta\Delta$), linear prediction residual (LPR) and linear prediction residual phase (LPRP) features provides improvement in the feature extraction stage and combination of learning vector quantization (LVQ) and gaussian mixture model – universal background model (GMM–UBM) provides improvement in the modelling stage. The performance is further improved by integrating the proposed techniques at the respective stages and combining the evidences from them at the testing stage. To achieve this, we propose strength voting (SV), weighted borda count (WBC) and supporting systems (SS) as combining methods at the abstract, rank and measurement levels, respectively. Finally, the proposed hierarchical combination (HC) method integrating these three methods provides significant improvement in the performance. Based on these explorations, this work proposes a scheme for speaker identification under limited training and testing data.

**Keywords.** Speaker identification; limited training and testing data; MFSR; LPR; LPRP; LVQ; GMM–UBM and combining.

## 1. Introduction

Speaker recognition aims at recognizing people from their voice (Atal 1976). Speaker recognition can be either identification or verification depending on the task objective.

*For correspondence

The speaker verification involves accepting or rejecting the identity claim of a speaker. In speaker identification since there is no identity claim, the system identifies the most likely speaker of the test speech signal. Speaker identification can be further classified into closed-set and open-set. The task of identifying a speaker who is known *a priori* to be a member of the set of $N$ enrolled speakers is known as closed-set speaker identification (Rosenberg 1976). On the other hand, speaker identification system that is able to identify the speaker, even from outside the set of $N$ enrolled speakers is known as open-set speaker identification (Rosenberg 1976). Depending on the mode of operation, speaker recognition is either text-dependent or text-independent type. Speech for the same text is used for both training and testing in the text-dependent case and no such restrictions are made in the text-independent case. This work focuses on closed-set, text-independent, speaker identification using limited training and testing data.

Speaker identification system may be viewed as consisting of four stages, namely; analysis, feature extraction, modelling and testing. It is preferable to have speaker identification technology that provides good performance using limited speech data for training and testing. Such a solution increases the usability of this technology for person identification as part of multi-factor authentication systems in e-commerce applications (Bimbot *et al* 2004). There are practical situations like forensic investigations where the availability of speech data is limited and technology is needed to identify the person using only limited data (Campbell *et al* 2005). Also, if we have technology that provides good performance using limited training and testing data, then we can develop multiple models, perform multiple testings and finally combine them to identify the speaker in normal speaker identification tasks. This may improve the performance as well as robustness of speaker identification system.

To achieve good performance, especially under limited training and testing data, efficient techniques are essential in each of the stages. The state-of-the-art speaker identification systems assume the availability of sufficient data for training and testing. In this work, sufficient and limited data notionally denote the case of more than one minute and less than 15 seconds, respectively (Angkititrakul & Hansen 2007; Prakash & Hansen 2007). Except for the task objective, the speaker identification study considered in this work is equivalent to speaker verification task of 10 sec training and testing data followed in NIST speaker recognition evaluations (NIST 2003). Existing speaker modelling techniques, mostly based on clustering (vector quantization (VQ)) (Gray 1984) and statistical modelling (Gaussian mixture modelling (GMM)) (Reynolds & Rose 1995) work well under sufficient data condition. Under limited data, the modelling may not be reliable due to the effect of either sparse distribution for clustering or insufficient data for the estimation of statistical model parameters. Further, during testing, the decision may not be reliable due to limited number of feature vectors. However, the objective of speaker identification under limited data condition is to obtain as good performance as possible. One approach for achieving the same is to use efficient techniques in each of the speaker identification stages. The efficient strategy should be from the perspective of alleviating the difficulty arising out of limited data. The efficient techniques can be initially explored independently to observe their potential in improving the performance. These techniques can then be integrated in different ways to develop multiple integrated systems. The evidences from all these integrated systems can be combined to obtain a speaker identification system that provides improved performance under limited data. Hence the motivation for this work.

A few attempts have been made earlier to study and provide solution to the problem of limited data in the speaker recognition task. A study based on multimodal speaker models for text-independent speaker identification using small utterances was attempted (Li 1985).

In this study, it was demonstrated that the multimodal speaker model provides better performance compared to the unimodal speaker model. Kimball *et al* (1997) studied the use of hidden markov model (HMM) for text-dependent speaker recognition under limited data and mismatched channel conditions. The speaker models were built using the broad phonetic category (BPC) and HMM. The HMM for each speaker obtained by maximum likelihood linear regression (MLLR) adaptation technique has shown to provide relatively better modelling. The universal background model (UBM) based GMM was proposed for text-independent speaker identification (Angkititrakul & Hansen 2007), where the speech data from a large pool of speakers were used to build a speaker independent GMM–UBM model. The maximum *a posteriori* (MAP) adaptation technique was used to create speaker dependent models. Experiments were conducted for Inset/out of set speaker identification using GMM–UBM. Alternatively, a cohort UBM model was built for each speaker by pooling the data from the acoustically close speakers (Prakash & Hansen 2007). The performance using cohort concept was shown to give better performance compared to the earlier single speaker independent UBM system in (Angkititrakul & Hansen 2007). In order to improve the performance, kernel eigenspace-based maximum likelihood linear regression (KEMLLR) adaptation technique was proposed for speaker verification with limited training data (Man-Wai Mak *et al* 2006). In that study, it was demonstrated that KEMLLR adaptation has shown to give better performance compared to MAP adaptation when the training data is less than 8 sec.

As reviewed above, most of the studies treat the problem of limited data from the modelling perspective. This direction for alleviating the effect of limited data is obvious. This is because, as mentioned earlier, the effect of limited data is the sparse distribution for clustering and limited features for statistical modelling. However, apart from improved modelling, it may be possible to improve the speaker recognition performance by developing new speech analysis, feature extraction and combination techniques. For instance, if we have a speech analysis technique to generate more number of feature vectors from limited data, then it may improve the performance. The present work demonstrates the significance of multiple frame size and rate (MFSR) analysis in generating more number of feature vectors. In the existing solution for limited data, most methods use mel frequency cepstral coefficients (MFCC) as feature vectors. MFCCs mostly represent speaker-specific vocal tract information. It may therefore be possible to use additional features representing speaker-specific excitation source aspect. This work demonstrates the gain that can be achieved by combining the speaker information from MFCC, its temporal derivatives ($\Delta$, $\Delta\Delta$), linear prediction residual (LPR) and linear prediction residual phase (LPRP) features. Among the different modelling techniques proposed for limited data, GMM–UBM by MAP approach is found to be most effective. It may be possible to explore other modelling techniques having different working principles. The present work illustrates the same using the learning vector quantization (LVQ) technique. Further, the combined modelling using LVQ and GMM–UBM is demonstrated to provide better performance than either LVQ or GMM–UBM modelling.

The proposal mentioned above concentrates on selecting efficient techniques suitable for analysis, feature extraction and modelling stages of the speaker identification system. To further improve the performance, all these techniques can be integrated in the respective stages to obtain what are called *Integrated Systems*. As will be demonstrated later, different integrated systems are possible based on the choice of techniques in different stages. Finally, the evidences from the integrated systems can be combined to improve the performance. There are several methods in the literature for combining the decision of multiple classifiers to improve the performance (Xu *et al* 1992; Ho *et al* 1994; Lee *et al* 2006; Rahman & Fairhurst 2000). In (Xu *et al* 1992), attempts have been made to combine individual classifiers using methods

like Bayesian formalism, voting method and Dempster-Shafer (D–S) theory for handwriting recognition. In (Ho *et al* 1994), to make a combined decision using multiple classifier outputs for machine printed word and character recognition, ranked voting and logistic regression methods were proposed. In (Lee *et al* 2006), voting method was used for speaker identification based on the results of various resolution filterbanks. The combination of decision by majority voting and divide and conquer was proposed for pattern classification in (Rahman & Fairhurst 2000). It was shown in all these studies that the combined framework gives better performance than the individual method. In this work, we try with some of the existing combination methods to combine evidences from different integrated systems. In addition, new combination techniques are proposed to improve the performance. The present work proposes strength voting (SV), weighted borda count (WBC), supporting systems (SS) and hierarchical combination (HC) methods for combining the evidences. All these are demonstrated to provide improved performance.

The novelty of the present work may be summarized as follows: MFSR as speech analysis technique, combination of MFCC, $\Delta$, $\Delta\Delta$, LPR and LPRP as feature, combined LVQ–GMM–UBM as model, all for speaker identification under limited training and testing data. The second contribution is the development of integrated systems using the above techniques to further improve the identification performance. The third contribution includes proposing different combination schemes like strength voting, weighted Borda count, supporting systems and hierarchical combination for combining evidences from the integrated systems. Finally, a scheme is proposed for speaker identification under limited training and testing data.

The rest of the paper is organized as follows: in section 2, the databases and the experimental set-up for the study are discussed. Section 3 describes the speaker identification using MFSR analysis of speech. In section 4, speaker identification using combination of features is demonstrated. Section 5 presents speaker identification using combined modelling. In section 6, development and evaluation of integrated systems are made. The different combination techniques and their potential use for the present work are described in section 7. A scheme for speaker identification using limited training and testing data is given in section 8. Summary and conclusions of this study with scope for future work are *n* discussed in section 9.

## 2. Databases and experimental set-up

In this work the performance of the speaker identification system is evaluated using the YOHO (Campbell 1995), TIMIT (Zue *et al* 1990) and 2003 NIST speaker recognition evaluation (NIST–SRE–2003) databases (NIST 2003). The YOHO database consists of speech data from 138 speakers (106 male and 32 female) (Campbell 1995). The speech files are of type *combination lock phrases* (e.g. 36–24–36). 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 database is not meant for limited training and testing data condition, we have taken one, two and four wave files of each speaker to create the database for the present work. In this study, first we perform the speaker identification studies for the set of first 30 speakers. Based on the results and inferences, the study is then extended to the whole database of 138 speakers.

The TIMIT database consists of speech data from 462 speakers (326 male and 136 female) in the training set and 168 speakers (112 male and 56 female) speakers in the testing set of total 630 speakers. 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 database is resampled to 8 kHz. The speech data for each speaker includes 10 speech files, each of about 3 sec duration. The speech are for the phonetically balanced sentences (e.g. She had your dark suit in greasy wash water all year) and hence different from those of YOHO database. However, the quality of data is almost same in both the cases. In this work, we have used one set of first 30 speakers and another set of the first 138 speakers from the testing set of the TIMIT database. This choice is to keep the experimental protocol similar to that of YOHO database. The first five speech files are used for training and the remaining for testing. We have taken one, two, four and five speech files for each speaker to create the database for the present work. The evaluation of the proposed techniques using TIMIT database reconfirms the observations made in the earlier studies using YOHO database.

The NIST–SRE–2003 database consists of speech data from 356 speakers (149 male and 207 female). The training and testing data consist of spontaneous speech collected over cellular phone, sampled at 8 kHz and stored with 16 bits/sample resolution. The speech data range between a few seconds and a minute. A detailed description of the database can be found in the NIST–SRE–2003 plan (NIST 2003). The speech quality will be very poor compared to the YOHO and TIMIT databases due to nature of transducer, coding employed and channel characteristics. Evaluation on this database shows the robustness of the improvement trends in the proposed techniques. Since the database is also not meant for limited data condition, we have taken three, six and twelve seconds of each speaker data to create the database for the present work. To have comparative study, the experiments are again conducted as in the YOHO database set-up using selected 138 out of the 356 speakers.

## 3. MFSR analysis for speaker identification under limited data

Speech signals analysed with fixed frame of size 20 ms and rate of 10 ms is termed as single frame size and rate (SFSR) analysis. In the limited data condition, available training and testing data is small. If we use SFSR analysis, then it will not provide sufficient feature vectors to train and test the speaker. The studies (Nagarajan 2004; Sarada *et al* 2004a; Sarada *et al* 2004b) demonstrated that the spectral variations in speech can be captured by combining multiple frame size (MFS) and multiple frame rate (MFR). It was shown that combined MFS and MFR gives better performance compared to single frame size (SFS) for language and speech identification task. In this study, we demonstrate the use of MFS, MFR and MFSR analysis techniques for speaker identification.

In the case of MFS, the MFCC feature vectors are computed for each frame of size 12, 14, 16, 18 and 20 ms with shift of 10 ms. Since MFS is effectively a multiresolution analysis, feature vectors extracted with different frame sizes are considerably different due to different frequency resolutions (Rabinar & Juang 1993). Moreover, by varying the frame size we can vary the spectral information and hence feature vectors with different speaker-specific information. In the case of MFR, the MFCC feature vectors are extracted for each frame shift of 2, 4·5, 6·5, 8·5 and 10·5 ms with a constant frame size of 20 ms. Speaking rate as well as pitch are different for each speaker and also vary for the same speaker depending on the contextual information during speech production. The rate of change of spectral information can be captured by analysing the same speech data at different frame rates. Since MFR analysis is effectively a multi-shifting technique, the set of speech samples involved in the analysis of speech are different at each rate. As a result, feature vectors representing vocal tract information may be different. In order to gain the advantages of MFS and MFR, we combine

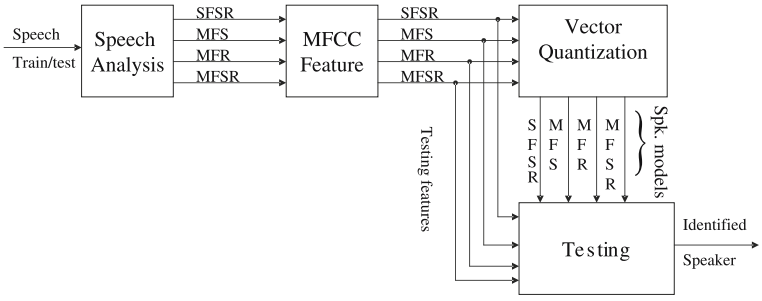


**Figure 1.** The SFSR and proposed MFSR based speaker identification system for limited data condition.

them and call it as MFSR (Jayanna & Prasanna 2009). In all the methods, MFCC feature vectors are computed using 35 filters in the filter bank and the feature vectors excluded the first coefficient $c_o$. Cepstral mean subtraction (CMS) is applied to the MFCC to remove the linear channel effect. Silence and low energy speech parts are removed using an energy-based frame selection technique (Deller *et al* 1993). The threshold used for the speech frames selection is 0·1 times the average frame energy. The modelling technique used is the vector quantization (VQ) and the testing technique used is the majority voting with Euclidean distance similarity measure. The steps involved in the SFSR and proposed MFSR based speaker identification system for limited data condition are shown in the block diagram in figure 1.

The experimental results obtained for different codebook sizes for the first 30 speakers taken from the YOHO database, each having 3 sec training and testing data are shown in table 1. The results show that MFSR yields the highest performance of 90% for codebook of size 128. The performance is higher than that of SFSR which provides 70% for 64 codebook size. The MFSR performance is also higher than that of the MFS and MFR that provide 73% and 87% for codebook of size both 32 and 128, respectively. The same trend is observed for different data sizes of 6 sec and 12 sec for the 30 and 138 speakers, and is shown in figures 2a and b, respectively. To verify the robustness of the analysis techniques, we conducted the experiments on TIMIT and NIST–2003–SRE databases also using SFSR and MFSR analysis techniques. The experimental results of the TIMIT database for a set of first 30 and 138 speakers are shown in figures 2c and d, respectively. Since the database is of good quality, speaker identification performance is similar to that of the YOHO database. The experimental

**Table 1.** Speaker identification performance (%) for a set of first 30 speakers of the YOHO database, each having *3 sec training and test data* for SFSR and MFSR analysis.

| | | | Codebook size | | | |
|---|---|---|---|---|---|---|
| Analysis | Feature | Modelling | 16 | 32 | 64 | 128 |
| SFSR | MFCC | VQ | 63 | 67 | 70 | 60 |
| MFS | MFCC | VQ | 73 | 73 | 70 | 73 |
| MFR | MFCC | VQ | 73 | 80 | 77 | 87 |
| MFSR | MFCC | VQ | 80 | 80 | 87 | 90 |

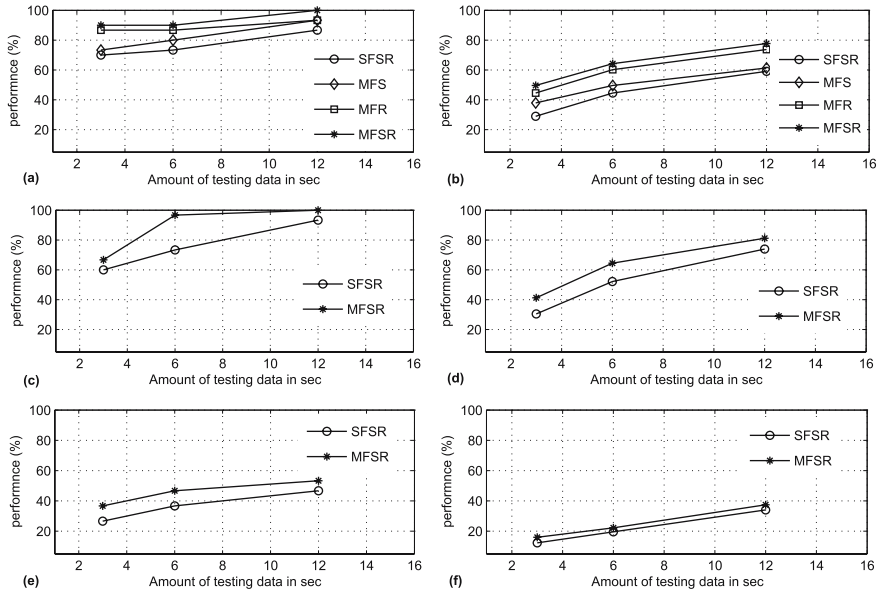


**Figure 2.** Performance of the speaker identification system based on SFSR and MFSR analysis techniques for different sizes of training and testing data; **(a)** for a set of first 30 speakers taken from the YOHO database; **(b)** for 138 speakers taken from the YOHO database; **(c)** for a set of first 30 speakers taken from the TIMIT database; **(d)** for a set of first 138 speakers taken from the TIMIT database; **(e)** for 30 speakers taken from the NIST–2003–SRE database and **(f)** for 138 speakers taken from the NIST–2003–SRE database.

results of the NIST–SRE–2003 database for a set of 30 and 138 speakers are shown in figure 2e and f, respectively. The experimental results are relatively poor compared to those of YOHO database. This is expected due to the poor quality of speech data as a result of transducer, coding and channel characteristics. Though the experimental results are poor, the general trend, that is, MFSR providing better performance than SFSR, resemble those for the YOHO database, irrespective of speaker population and amount of data. This shows that MFSR is preferable over SFSR as an analysis technique for speaker identification under limited data condition.

## 4. Combination of features for speaker identification under limited data

The studies by (Prasanna *et al* 2006; Yegnanarayana *et al* 2005; Murthy & Yegnanarayana 2006) used the combination of vocal tract (MFCC) and source information (LPR and LPRP) for speaker recognition under sufficient data condition. The combined system was shown to give better performance. In this study, we explore the effectiveness of the combined system to the limited data case. Since the amount of data is small, any one feature extraction technique may not provide enough features for modelling and testing. We evaluate different feature extraction techniques under limited data condition using SFSR in the analysis stage and VQ in the modelling stage. The different feature extraction techniques are MFCC, $\Delta$, $\Delta\Delta$, LPR and LPRP to know the effectiveness of the speaker information present in them under limited data. The MFCC are known to give good performance for speaker recognition (Davis & Mermelstein 1980). The MFCC feature vectors do not capture the transition characteristics of

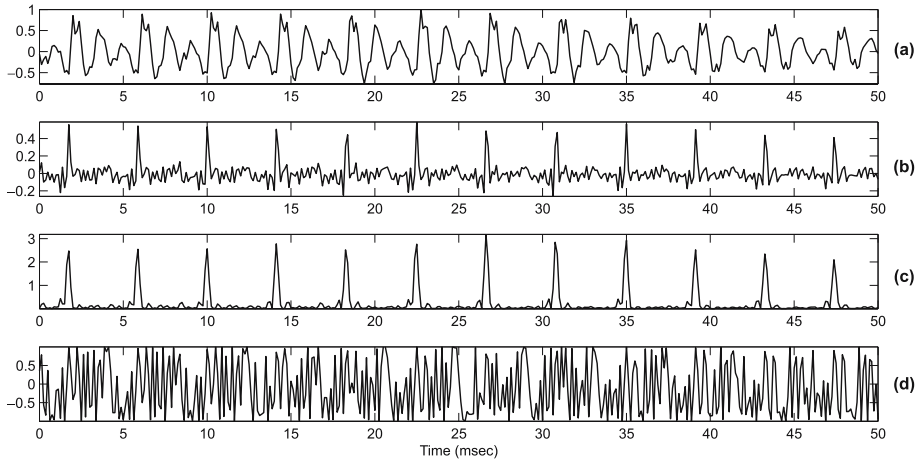


**Figure 3.** Difference between LPR and LPRP: **(a)** a segment of speech signal, **(b)** corresponding LP residual, **(c)** Hilbert envelop (HE) of LP residual and **(d)** corresponding LP residual phase.

the speech signal which also contains speaker-specific information (Furui 1981; Furui 1986). The transition characteristics can be captured by computing the $\Delta$ and $\Delta\Delta$ coefficients, obtained respectively from first and second order time derivatives of MFCC. The derived features can be used for speaker recognition as either concatenated to MFCC (Furui 1981; Furui 1986) or individual features (Kinnunen *et al* 2003). In this study, since the number of feature vectors are small, to minimize the effect of curse of dimensionality, we considered them as independent evidences to determine the level of information in them.

The studies made in (Yegnanarayana *et al* 2005; Murthy & Yegnanarayana 2006) show that the LP residual and LP residual phase, respectively, also contain speaker-specific excitation information that can be used for speaker recognition. The LP residual is obtained by first predicting the vocal tract information using linear prediction coefficients (LPC), and then suppressing them from the speech signal using inverse filter formulation (Yegnanarayana *et al* 2005). The LPRP is obtained by dividing the LP residual by its Hilbert envelope (Murthy & Yegnanarayana 2006). The Hilbert envelope is the magnitude of the analytic signal of a given real signal. The difference between LPR and LPRP is shown in figure 3. From the figure, we can understand that the LPRP mostly contains speaker-specific sequence information whereas the LPR contains excitation source information related mainly to glottal closure instants (GCIs) due to the dominance of energy around GCIs. It may be possible that these two features may have different aspect of speaker-specific excitation information. They may therefore be combined to gain advantage.

The effectiveness of LP residual and LPR phase is studied using the vector-pulse code modulation (V–PCM) concept (Cuperman & Gersho 1985). The difference between V–PCM and VQ is that the former directly quantizes the group of signal values and the later quantizes the parameter vectors derived from the signal. In V–PCM study, for both LPR and LPRP processing, blocks of samples are used for quantization. The block size is chosen to be 3 and 5 ms with the objective of keeping its size less than pitch period to avoid the dominance of pitch information (Yegnanarayana *et al* 2005). All these features are then combined to obtain better representation of speaker. As a result, the combined (MFCC$+\Delta+\Delta\Delta+$LPR$+$LPRP) feature gives better performance than the individual feature. The combination is done at the scoring level, in which the highest performed codebook size of each feature extraction

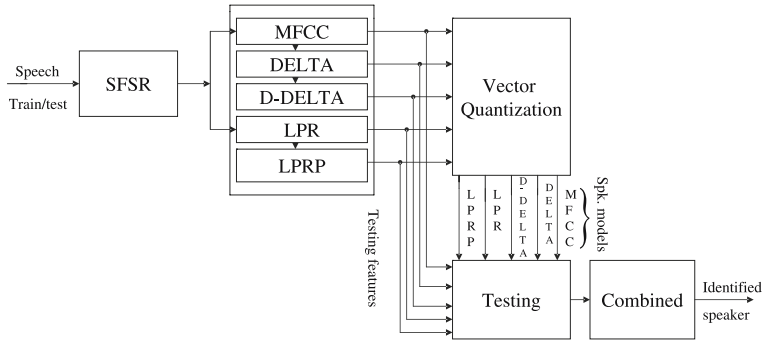


**Figure 4.** Combination of feature based speaker identification system for limited data condition.

technique is considered and their frame ratios are linearly combined. The frame ratio of a speaker specifies the ratio of number of frames scored to total number of frames of test speech data. The speaker who scores the highest frame ratio will be the identified speaker.

The steps involved in the combination of feature based speaker identification system for limited data condition are shown as block diagram in figure 4. The experimental results obtained for different codebook sizes of first 30 speakers taken from the YOHO database, each having 3 sec training and testing data are shown in table 2. The results show that the MFCC feature yields the highest identification performance of 70% for the codebook of size 64. The $\Delta$ and $\Delta\Delta$ provide 37% and 33%, respectively for the 64 codebook size. The performance of both LPR and LPRP is 47% for the 128 and 64 codebook sizes, respectively. Though the performance of derivatives of MFCC, LPR and LPRP are less than the MFCC, combining MFCC with them may improve the performance owing to different information present in them. As a result, the combined MFCC $+ \Delta + \Delta\Delta +$ LPR $+$ LPRP features yield the highest performance of 87%. This performance is better than any of the combinations and the individual features. The same trend is observed for data sizes of 6 and 12 sec for the 30 and 138 speakers, and the results are shown in figures 5a and b, respectively. The experimental

**Table 2.** Different feature extraction techniques performance (%) for the 30 speakers of YOHO database, each having 3 sec training and testing data.

| | | | Codebook size | | | |
|---|---|---|---|---|---|---|
| Analysis | Feature | Modelling | 16 | 32 | 64 | 128 |
| SFSR | MFCC | VQ | 63 | 67 | 70 | 60 |
| SFSR | $\Delta$ | VQ | 30 | 23 | 37 | 27 |
| SFSR | $\Delta\Delta$ | VQ | 20 | 20 | 33 | 23 |
| SFSR | LPR | VQ | 13 | 23 | 30 | 47 |
| SFSR | LPRP | VQ | 30 | 30 | 47 | 23 |
| | MFCC $+ \Delta + \Delta\Delta$ | | | | 77 | |
| | MFCC $+ \Delta + \Delta\Delta +$ LPR | | | | 80 | |
| | MFCC $+ \Delta + \Delta\Delta +$ LPRP | | | | 80 | |
| | MFCC $+ \Delta + \Delta\Delta +$ LPR $+$ LPRP | | | | 87 | |

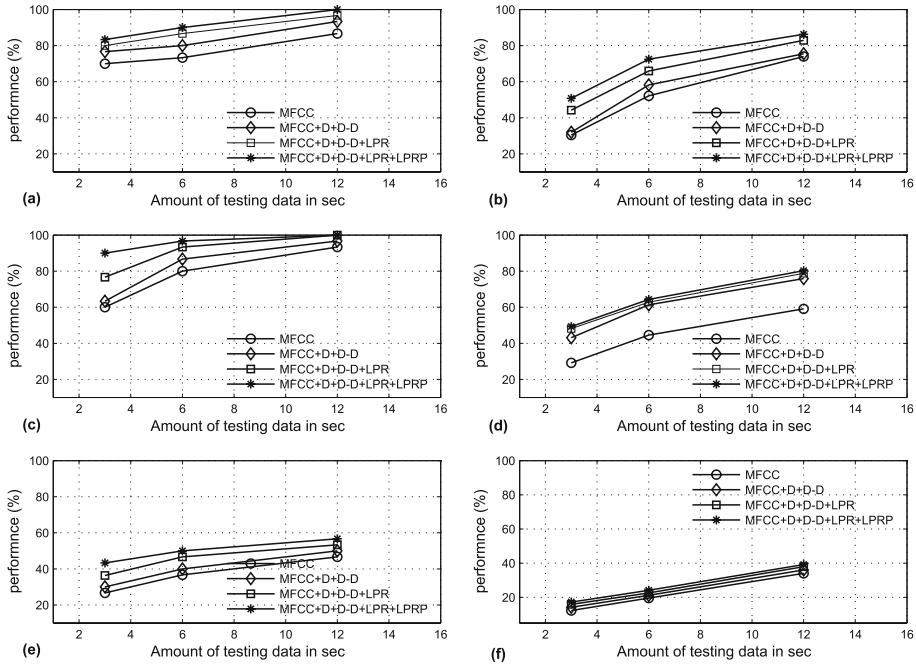


**Figure 5.** Performance of the speaker identification system based on combination of features for different sizes of training and testing data; **(a)** for a set of first 30 speakers taken from the YOHO database; **(b)** for 138 speakers taken from the YOHO database; **(c)** for a set of first 30 speakers taken from the TIMIT database; **(d)** for a set of first 138 speakers taken from the TIMIT database; **(e)** for 30 speakers taken from the NIST–2003–SRE database and **(f)** for 138 speakers taken from the NIST–2003–SRE database.

results of the TIMIT database for the first 30 and 138 speakers are shown in figures 5c and d, respectively. The experimental results of the NIST–SRE–2003 database for the 30 and 138 speakers are shown in figures 5e and f, respectively. The experimental results for the TIMIT and NIST–SRE–2003 databases also show the same trend, that is, combination of features providing the best performance, as in the case of YOHO database. This shows that the combination of features can be used for speaker identification under limited data condition.

## 5. Combined modelling for speaker identification under limited data

State-of-the-art speaker identification uses GMM–UBM as a modelling technique when the training data is sparse (Angkititrakul & Hansen 2007; Prakash & Hansen 2007). In this study, apart from the GMM–UBM, we evaluate other modelling techniques like crisp vector quantization (CVQ) (Gray 1984), fuzzy vector quantization (FVQ) (Bezdek & Harris 1978), self-organizing map (SOM) (Kohonen 1990), learning vector quantization (LVQ) and gaussian mixture model (GMM) (Reynolds & Rose 1995). This study uses SFSR in the analysis stage and MFCC in the feature extraction stage. Based on the performance of the individual models, the study proposes the combined LVQ and GMM–UBM (LVQ–GMM–UBM) model to provide relatively better performance than the individual models. The combination is done at the scoring level in which the highest performed codebook size of the modelling techniques

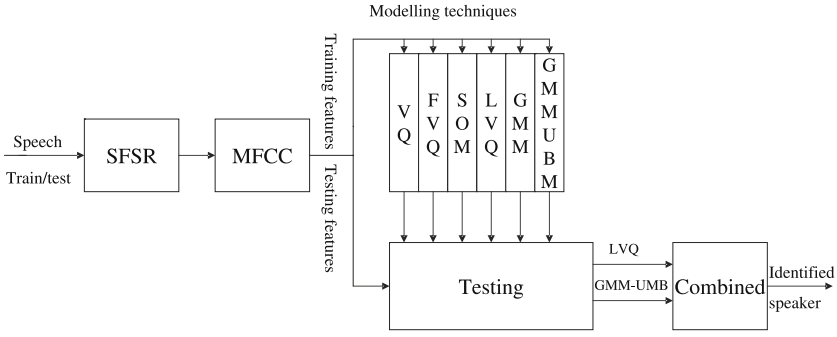


**Figure 6.** Combined modelling based speaker identification system for limited data condition.

are considered, and their frame ratios are linearly combined. The steps involved in the LVQ–GMM–UBM modelling based speaker identification system for limited data condition are shown in the block diagram in figure 6.

The experimental results for different codebook sizes for the 30 speakers taken from the YOHO database are shown in table 3. It shows that the CVQ yields an identification performance of 70% for the codebook of size 64. The FVQ gives an identification performance of 77% for the codebook size of 32 (weighting factor = 1·39) which is higher than that of CVQ. The difference between CVQ and FVQ is that the former uses hard decision processing (Gray 1984), and the later uses soft decision processing (Bezdek & Harris 1978) while designing the codebook. The SOM gives an identification performance of 73% for the codebook of size 32 (iterations = 16000, neighbourhood = 1 and learning rate = 0·06). The LVQ gives the identification performance of 80% for the codebook of size 32 (iterations = 17600 and learning rate = 0·06). This performance is higher than that of SOM. This is because SOM employs unsupervised learning whereas LVQ employs supervised learning over initially

**Table 3.** Individual and combined modelling based speaker identification performance (%) for the 30 speakers of YOHO database, each having *3 seconds training and testing data*.

| | | | Codebook size | | | |
|---|---|---|---|---|---|---|
| Analysis | Feature | Modelling | 16 | 32 | 64 | 128 |
| SFSR | MFCC | CVQ | 63 | 67 | 70 | 60 |
| SFSR | MFCC | FVQ | 70 | 77 | 73 | 70 |
| SFSR | MFCC | SOM | 73 | 73 | 70 | 73 |
| SFSR | MFCC | LVQ | 73 | 80 | 73 | 67 |
| SFSR | MFCC | GMM | 73 | 40 | 37 | 13 |
| SFSR | MFCC | GMM–UBM–NIE | 60 | 60 | 63 | 77 |
| SFSR | MFCC | GMM–UBM–IE | 60 | 67 | 73 | 83 |
| | | LVQ–FVQ | | | 80 | |
| | | LVQ–GMM | | | 80 | |
| | | LVQ–GMM–UBM–NIE | | | 83 | |
| | | LVQ–GMM–UBM–IE | | | 87 | |

obtained unsupervised codevectors from SOM, and hence improved the performance. The GMM yields 73% for Gaussian mixtures of 16.

The GMM–UBM experiments are conducted in two contexts: (i) not including evaluation set (NIE) in which speakers used for speaker identification study are not included for UBM training. In this case, first 30 speakers are used for speaker identification study and the remaining 108 speakers of each 21 utterances for UBM training. As a result, it provides an identification performance of 77% for a Gaussian mixture of 128 as is shown in table 3. (ii) including the evaluation set (IE) in which speakers used for speaker identification study are also included in the UBM training. In this case, all the 138 speakers are used for speaker identification study and for each speaker 21 utterances are used for UBM training. As a result, it provides an identification performance of 83% for Gaussian mixture of 128 as shown in table 3.

We then combined different modelling techniques like LVQ–FVQ, LVQ–GMM, LVQ–GMM–UBM–NIE and LVQ–GMM–UBM–IE to see the effectiveness under limited data. We found that the combined (LVQ–GMM–UBM–NIE) and (LVQ–GMM–UBM–IE) systems yield the highest performance of 83% and 87%, respectively as shown in table 3. The combined system performance is higher than that of the individual and other combined modelling techniques. The improvement in the performance may be attributed to the different modelling principle employed in LVQ (Kohonen 1990) and GMM–UBM (Reynolds *et al* 2000). That is,

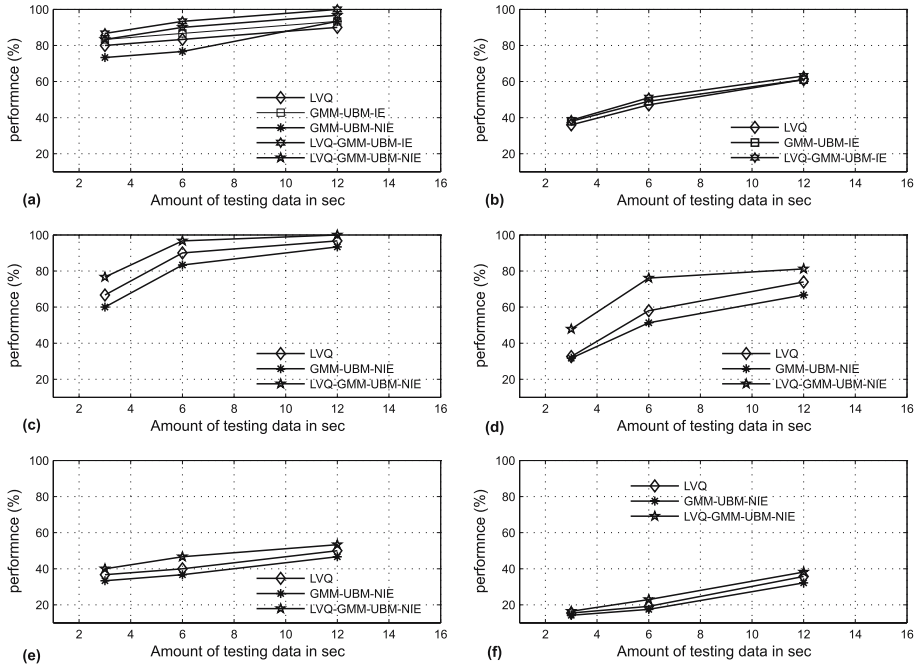


**Figure 7.** Performance of the speaker identification system based on combined modelling techniques for different sizes of training and testing data; **(a)** for a set of first 30 speakers taken from the YOHO database; **(b)** for 138 speakers taken from the YOHO database; **(c)** for a set of first 30 speakers taken from the TIMIT database; **(d)** for a set of first 138 speakers taken from the TIMIT database; **(e)** for 30 speakers taken from the NIST–2003–SRE database and **(f)** for 138 speakers taken from the NIST–2003–SRE database.

the supervised learning over unsupervised learning involved in LVQ and other speakers data used to fill the acoustic space in GMM–UBM. Moreover, the LVQ modelling technique is based on non-parametric approach, whereas the GMM–UBM is based on parametric approach. Hence this combination gives the best performance. The same trend is observed for different data sizes of 6 and 12 secs for the 30 and 138 speakers, and the results are shown in figures 7a and b, respectively. In case of TIMIT database, for UBM training, we have used 112 male and 40 female speakers of each 21 utterances taken from TIMIT training set. The speaker identification experiments are conducted on the TIMIT test set. The experimental results are shown in figures 7c and d for the 30 and 138 speakers, respectively. In the case of NIST–SRE–2003, for UBM training, we have used 130 male and 40 female speakers of each one minute data taken from NIST–SRE–2002 database, which are not part of the 138 speakers considered for the speaker identification study. The experimental results are shown in figures 7e and f for the 30 and 138 speakers, respectively. The experimental results for the TIMIT and NIST–SRE–2003 databases also resemble those for the YOHO database, that is, the combined LVQ–GMM–UBM providing the best performance, irrespective of speaker population and amount of data. We therefore suggest that the combined LVQ–GMM–UBM modelling can be used for speaker identification under limited data condition.

## 6. Integrated systems for speaker identification using limited training and testing data

The above discussed studies are made individually. That is, the proposed technique was used in the respective stage and the existing techniques in the remaining stages. The effectiveness of these techniques may be exploited well by integrating them together. The study in this section integrates the techniques discussed earlier, and develops combined speaker identification systems for limited data condition termed as *Integrated Systems*. The scheme for integrating the techniques for the development of the proposed integrated speaker identification systems are shown in the block diagram in figure 8. In this system, the MFSR is used for speech analysis to extract the MFCC, $\Delta$, $\Delta\Delta$, and SFSR is used to extract the LPR and LPRP features.

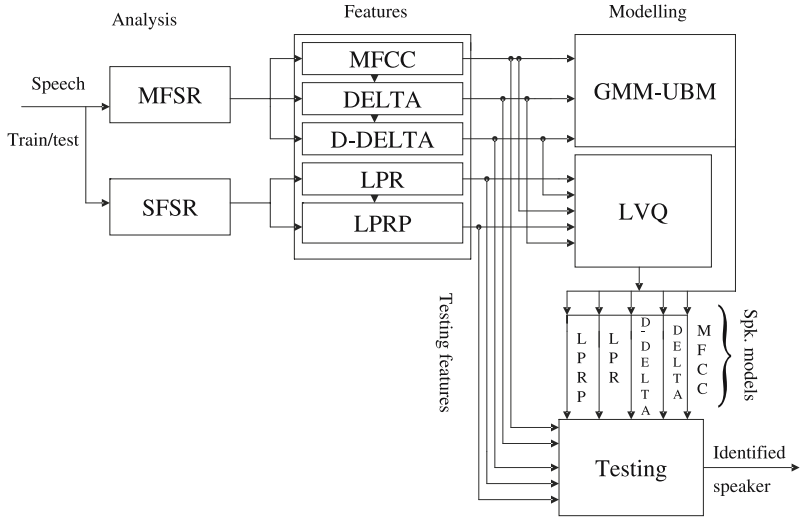


**Figure 8.** Integrated systems for speaker identification under limited data condition.

**Table 4.** Performance of the different integrated systems.

| Integrated systems | Analysis | Features | Modelling | Performance (%) |
|---|---|---|---|---|
| $S_1$ | MFSR | MFCC | LVQ | 93 |
| $S_2$ | MFSR | MFCC | GMM–UBM | 80 |
| $S_3$ | MFSR | $\Delta$ | LVQ | 57 |
| $S_4$ | MFSR | $\Delta\Delta$ | LVQ | 43 |
| $S_5$ | MFSR | $\Delta$ | GMM–UBM | 27 |
| $S_6$ | MFSR | $\Delta\Delta$ | GMM–UBM | 33 |
| $S_7$ | SFSR | LPR | LVQ | 47 |
| $S_8$ | SFSR | LPRP | LVQ | 47 |

The extracted features are individually modelled using LVQ and GMM–UBM and then combined for speaker identification. Since the performance of the LVQ depends on the parameters such as learning rate ($\eta$) and number of iterations (iter.), we have fine tuned them to avail the best performance. There are totally 8 different possibilities considered for the study, and we denote each of them as integrated systems $S_1$ to $S_8$ as shown in table 4.

The experimental results for the 30 speakers of YOHO database are given in table 4. The results show that the performance of MFSR–MFCC–LVQ is 93% ($\eta = 0{\cdot}05$ and iter. $=$ 76800) and MFSR–MFCC–GMM–UBM is 80%. These results are higher than that of the result of MFSR–MFCC–VQ that provides 90% and SFSR–MFCC–GMM–UBM–NIE that provides 73%, given in tables 1 and 3, respectively. The table 4 also shows that the performance of MFSR–$\Delta$–LVQ is 57% ($\eta = 0{\cdot}06$ and iter. $=$ 70400) and MFSR–$\Delta\Delta$–LVQ is 43% ($\eta = 0{\cdot}05$ and iter. $=$ 72960). These results are also higher than the results of SFSR–$\Delta$–VQ that provides 37% and SFSR–$\Delta\Delta$–VQ that provides 33%, given in table 2. Since we have not used GMM–UBM in the feature extraction stage, MFSR–$\Delta$ and $\Delta\Delta$–GMM–UBM results are not compared with the corresponding SFSR–$\Delta$ and $\Delta\Delta$–VQ results.

The performance of both LPR–LVQ ($\eta = 0{\cdot}06$ and iter. $=$ 38400) and LPRP–LVQ ($\eta =$ $0{\cdot}06$ and iter. $=$ 19200) is 47% and given in table 4. This result is same as that of LPR–VQ and LPRP–VQ given in table 2. The MFSR analysis is not done for LPR and LPRP. This is because, in these two cases the residual samples are compared in the time domain, and hence may not be effective. Moreover, the MFSR analysis is developed from the frequency domain perspective. Further, the GMM–UBM modelling technique is not done for LPR and LPRP. We are not doing any parameterization for extracting LPR and LPRP and hence may not be effective (Reynolds *et al* 2000). Thus, as demonstrated here we gain by integrating more than one technique proposed for speaker identification. We therefore find that integrated systems are preferred over individual systems.

## 7. Combining evidences from integrated systems for speaker identification under limited data

In the previous section we have demonstrated the potential of integrating the proposed MFSR, combination of features and combined modelling techniques for speaker identification. This resulted in 8 different integrated systems. Most of them provided improved performance over individual systems. Since each of the integrated systems are developed by considering different combination in the analysis, feature extraction and modelling stages, they may be

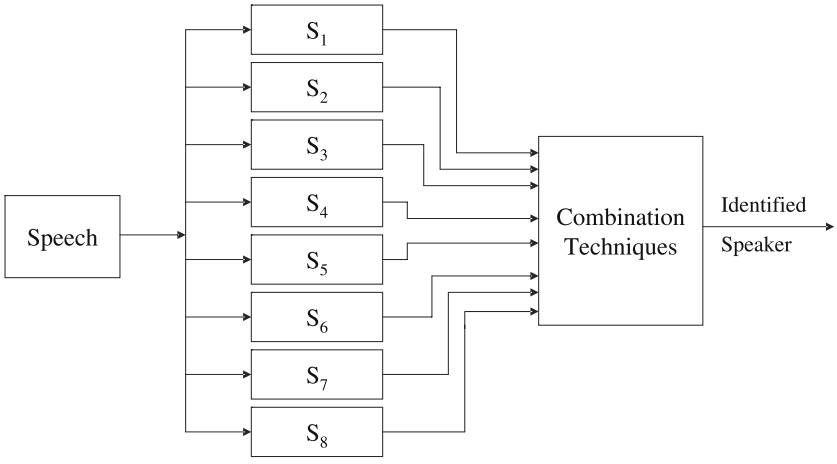


**Figure 9.** Combining evidences from the integrated systems for speaker identification under limited data condition.

treated as different systems that may have different speaker information. The evidences from them may therefore be further combined to improve the performance for speaker identification under limited data. The scheme for combining the evidences from the integrated systems for the development of the proposed speaker identification system is shown as block diagram in figure 9. Most of the existing combination schemes may be broadly grouped into three categories, namely, abstract level, rank level and measurement level combination schemes (Xu *et al* 1992).

### 7.1 *Strength voting (SV)*

This scheme is proposed under the abstract level combination. The approach in this scheme provides more voting power to the best performing system, and accordingly lower the voting power to the poorly performing systems. However, it also needs to be ensured that every system will have a voting power, irrespective of their performance. To take care of both the issues, the following scheme is proposed: Among all the integrated systems, the systems with the lowest performance is given a voting power of one vote. This ensures voting power to all the systems. Further, the other systems are given voting power of more than one vote based on their performance. In particular, the number of votes to a system is derived based on the ratio of its performance to the lowest performance system, and rounded the ratio to the nearest next higher integer value. Thus the system with higher performance is given more voting power and hence this combination scheme is termed as strength voting (SV). The number of votes from the strength voting scheme for each of the system are given in table 5. The SV works as follows: For the given test data, all the individual systems will output the most likely speakers. Each of these output speakers are voted according to the voting power of the individual systems. The speaker who gets the maximum votes is declared as the identified speaker. The identified speakers and the performance of SV are shown in table 6. It gives 97% performance for a set of the 30 speakers. The performance is higher than any of the best performing integrated systems. Although this combined system gives good identification performance, vote tie problem may occur. We have broken the tie using a strict linear ordering (Ho *et al* 1994).

**Table 5.** Number of votes based on the performance of the system.

| Systems | Performance (%) | No. of votes |
|---|---|---|
| $S_1$ | 93 | 4 |
| $S_2$ | 80 | 3 |
| $S_3$ | 57 | 2 |
| $S_4$ | 43 | 2 |
| $S_5$ | 27 | 1 |
| $S_6$ | 33 | 2 |
| $S_7$ | 47 | 2 |
| $S_8$ | 47 | 2 |

### 7.2 *Weighted borda count (WBC)*

This scheme is proposed under the rank level combination. In case of Borda count method, speaker who has the highest frame ratio gets top rank, and each subsequent speaker gets one vote less. In case of WBC also, the ranking of speaker is same as that of Borda count method. However, the decision is not only based on the rank of speaker, but also based on the performance of individual systems. We computed the weighted borda count (WBC) of each speaker $R_{xw}$ considering all the systems and their performance using the equation.

$$R_{xw} = \sum_{i=1}^{N} R_{xi} \frac{P_i}{\sum_{j=1}^{N} P_j} \quad x = 1 \ldots M,$$

where $M$ is the number of speakers, $R_{xi}$ is the rank of speaker $x$ in the system $i$, $P_i$ is the performance of the system $i$ and $N$ is the number of systems. The speaker who gets the highest $R_{xw}$ is chosen as the speaker of the test data. As a result, this combination scheme gives the identification performance of 97% for a set of first 30 speakers. The identified speakers and the performance are shown in table 6. The performance is higher than that of the best performing individual systems.

### 7.3 *Supporting systems (SS)*

This scheme is proposed under measurement level combination. For each speaker, we can identify the subset of the eight systems that support him/her. This depends on the nature of the speaker-specific information and hence modelling by the respective systems. For instance, to some of the speakers, the dominant speaker characteristics might have manifested in the MFCC, $\Delta$ and $\Delta\Delta$ features. Similarly, for some other speakers, the speaker characteristics might have manifested in the LPR and LPRP features. In that condition we benefit more by combining only those integrated systems for the combination. The supporting systems for each of the speakers may be identified by looking at the output of all the eight integrated systems. For instance, as shown in table 6 the systems $S_1$, $S_2$, $S_3$, $S_4$ and $S_7$ support the *speaker*-2. Therefore, for *speaker*-2 only the evidences from these systems are combined. Similarly, evidences from $S_1$, $S_2$, $S_3$ and $S_4$ are combined for *speaker*-3. In this way, for each of the speaker in the population, the supporting systems (SS) are identified *a priori*. Hence it is termed as supporting systems combination scheme. It should be noted that, there are as

**Table 6.** Performance of the individual and proposed speaker identification systems with the identified speakers for the 30 speakers. In the table, ✓ indicates speaker identified, ∗ indicates speaker not identified and $P$ indicates the performance.

| Systems | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | $P$(%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| S1 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ∗ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ∗ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 93 |
| S2 | ∗ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ∗ | ∗ | ✓ | ✓ | ✓ | ✓ | ∗ | ✓ | ✓ | ✓ | ✓ | ✓ | ∗ | ∗ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 80 |
| S3 | ∗ | ✓ | ✓ | ✓ | ✓ | ∗ | ∗ | ✓ | ✓ | ✓ | ∗ | ∗ | ✓ | ✓ | ✓ | ∗ | ∗ | ✓ | ✓ | ∗ | ∗ | ✓ | ∗ | ✓ | ✓ | ∗ | ✓ | ✓ | ∗ | ∗ | 57 |
| S4 | ∗ | ✓ | ✓ | ∗ | ✓ | ✓ | ∗ | ✓ | ∗ | ✓ | ∗ | ∗ | ∗ | ∗ | ✓ | ✓ | ✓ | ✓ | ∗ | ✓ | ∗ | ✓ | ∗ | ∗ | ✓ | ∗ | ∗ | ∗ | ∗ | ∗ | 43 |
| S5 | ∗ | ∗ | ∗ | ∗ | ✓ | ✓ | ∗ | ✓ | ∗ | ∗ | ∗ | ∗ | ∗ | ∗ | ∗ | ∗ | ∗ | ✓ | ∗ | ∗ | ✓ | ∗ | ∗ | ∗ | ∗ | ∗ | ✓ | ∗ | ✓ | ✓ | 27 |
| S6 | ∗ | ∗ | ∗ | ∗ | ✓ | ∗ | ✓ | ✓ | ∗ | ✓ | ∗ | ∗ | ✓ | ∗ | ∗ | ∗ | ∗ | ✓ | ∗ | ✓ | ∗ | ✓ | ∗ | ∗ | ✓ | ∗ | ∗ | ∗ | ✓ | ∗ | 33 |
| S7 | ∗ | ✓ | ∗ | ∗ | ✓ | ∗ | ✓ | ∗ | ∗ | ✓ | ∗ | ∗ | ∗ | ∗ | ✓ | ∗ | ∗ | ✓ | ✓ | ✓ | ✓ | ✓ | ∗ | ✓ | ✓ | ∗ | ✓ | ∗ | ✓ | ∗ | 47 |
| S8 | ∗ | ∗ | ∗ | ✓ | ✓ | ∗ | ∗ | ✓ | ✓ | ✓ | ∗ | ∗ | ∗ | ∗ | ∗ | ∗ | ✓ | ✓ | ∗ | ✓ | ∗ | ✓ | ∗ | ✓ | ✓ | ∗ | ✓ | ∗ | ✓ | ✓ | 47 |
| SV | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ∗ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 97 |
| WBC | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ∗ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 97 |
| SS | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 97 |
| HC | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 100 |

many supporting systems combinations as that of the number of speakers in the population. The speaker who provides the highest score for the test data is identified as the speaker of the test data. The identified speakers and the performance of SS are given in table 6. It gives 97%, better than the best performing integrated system $S_1$ which is 93%.

### 7.4 *Hierarchical combination (HC)*

The proposed combination schemes in the abstract, rank and measurement level are hierarchically combined to improve the performance. The proposed combination schemes are strength voting (SV), weighted borda count (WBC) and supporting systems (SS), respectively. For the test data of a speaker, this technique first picks up only the top 50% (4) speakers who get the highest votes based on the SV method. Then, the WBC is computed only for the selected speakers and 50% of the selected speakers who get less WBC are pruned out. Next, out of the 25% (2) speakers, the speaker who gets the highest frame ratio in the SS would be the recognized speaker of the test data. The identified speakers and the performance of HC are shown in table 6. It gives 100% performance. The performance is higher than that of the individual systems and combination schemes.

The experimental results for 138 speakers case of YOHO database are given in table 7. The results show that the combination schemes used in the 30 speakers study provide improved result for large database also. The HC scheme gives identification performance of 70%, higher than the integrated system MFSR–MFCC–GMM–UBM–IE which provides 60%. The other combined techniques like SV, WBC and SS also give relative improvement in the performance compared to the integrated system MFSR–MFCC–GMM–UBM–IE. The results of experiments conducted for different data sizes of 6 and 12 secs are shown in figures 10a and b, respectively. The results show the same trend as those of the 30 speakers for 3 seconds data. The experimental results of the TIMIT database are shown in figures 10c and d, respectively. The experimental results of the NIST–2003–SRE database are shown in figures 11a and b, respectively. The experimental results for the TIMIT and NIST–SRE–2003 databases also resemble those for the YOHO database.

## 8. Speaker identification system under limited data

The studies presented so far proposed techniques for improving the performance of speaker identification system using limited training and testing data. There is a significant

**Table 7.** Speaker identification performance for 138 speakers for different combination techniques.

| Combination techniques | Performance (%) |
|---|---|
| MFSR–MFCC–GMM–UBM–IE | 60 |
| Voting | 49 |
| Borda count (BC) | 47 |
| LCFR | 44 |
| Weighted LCFR (WLCFR) | 44 |
| Strength voting (SV) | 66 |
| Weighted borda count (WBC) | 62 |
| Supporting systems (SS) | 67 |
| Hierarchical combination (HC) | 70 |

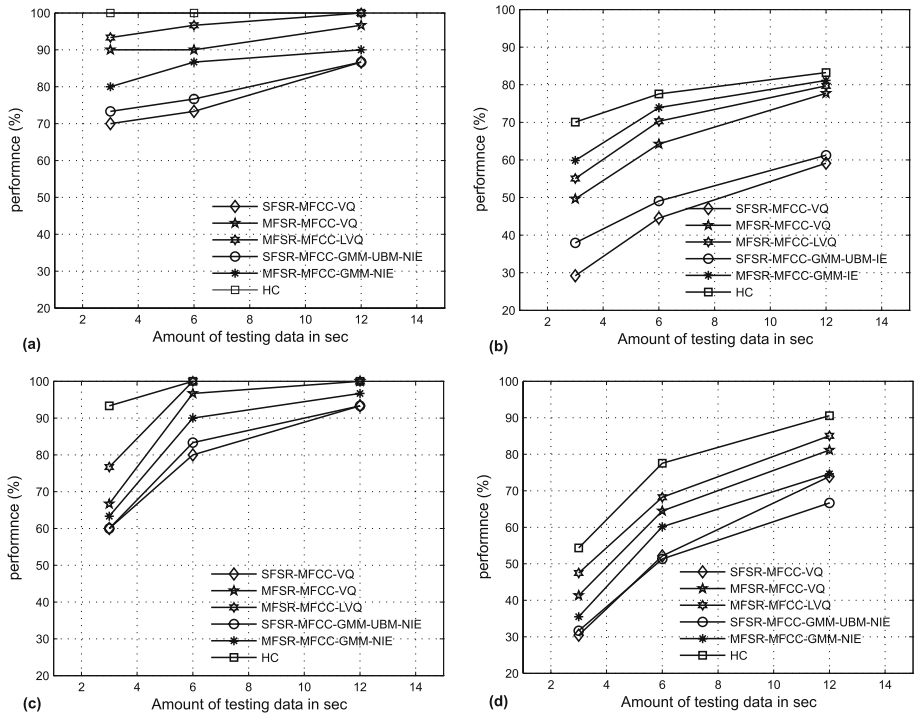


**Figure 10.** Performance of the speaker identification system based on individual, integrating and hierarchical combination systems for different sizes of training and testing data; **(a)** for a set of first 30 speakers taken from the YOHO database; **(b)** for 138 speakers taken from the YOHO database; **(c)** for a set of first 30 speakers taken from the TIMIT database; **(d)** for a set of first 138 speakers taken from the TIMIT database.

improvement in the performance of speaker identification system starting from the system based on SFSR–MFCC–VQ to the proposed system with hierarchical combination. Based on the explorations made so for we propose a scheme for speaker identification under limited

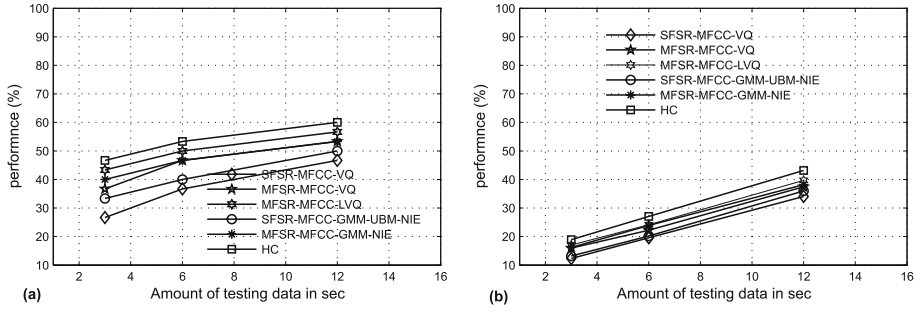


**Figure 11.** Performance of the speaker identification system based on individual, integrating and hierarchical combination systems for different sizes of training and testing data for a set of **(a)** 30 and; **(b)** 138 speakers taken from the NIST–2003–SRE database.

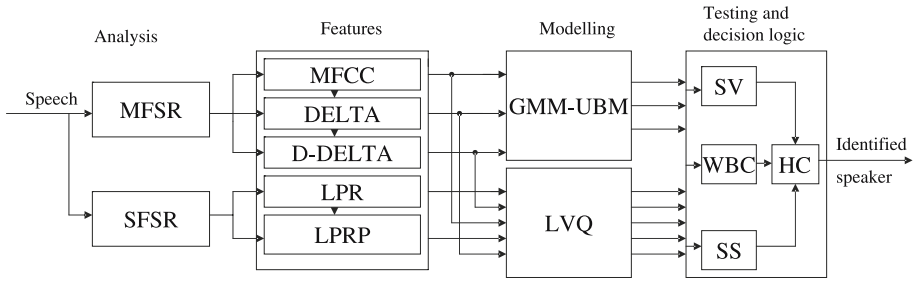


**Figure 12.** Proposed speaker identification system for limited data condition.

training and testing and is shown in figure 12. Thus as proposed initially in the introduction section, it is indeed possible to improve the speaker identification system performance by improving methods in each stage and properly integrating and combining them.

## 9. Summary and conclusions

In this paper, we first demonstrated different techniques for speaker identification under limited data. The first study demonstrated the significance of MFSR analysis of speech. The second study showed that the combination of features (MFCC, $\Delta$, $\Delta\Delta$, LPR and LPRP) provides better performance. The third study evidenced that the combined LVQ–GMM–UBM modelling technique gives better performance. We then integrated these techniques in such a way that the MFSR and SFSR analysis are used to extract the various speech features and model them separately using LVQ and GMM–UBM. As a result several integrated systems are developed and most of the integrated systems provided improved performance over individual systems. We then combined the evidences from these integrated systems using various combination schemes.

Though this work used new technique for speech analysis, known speech features are extracted and used for speaker identification. New features that have better discriminating capability are required. The state-of-the-art speaker recognition systems use GMM–UBM as a prominent modelling technique. We too used the same and in addition LVQ to improve the performance. What is really required is a modelling technique that is able to model a speaker without expecting large amount of data either for background or foreground. Even though the performance for YOHO database is good, the performance is poor for NIST speaker recognition database. This infers that apart from limited data, population size, the poor quality of speech due to nature of transducer, coder and channel severally limits the performance. The future work should focus on improving the performance to take care of these factors.

## 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): 460–47

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

Bimbot F, Bonastre J F, Fredouille C, Gravier G, Chagnolleau I M, Meignier S, Merlin T, Garcya J O, Delacretaz D P, Reynolds D A 2004 A tutorial on text-independent speaker verification. *EURASIP J. Applied Signal Process.* 4: 430–451

Campbell W M, Reynolds D A, Campbell J P, Brady K J 2005 Estimating and evaluating confidence for forensic speaker recognition. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* 1: 717–720

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

Cuperman V, Gersho A 1985 Vector predictive coding of speech at 16 kbits/s. *IEEE Trans. Communications* 33(7): 685–696

Deller J, Hansen J, Proakis J 1993 *Discrete Time Processing of Speech Signals*. IEEE Press, First edition

Davis S B, Mermelstein P 1980 Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust., Speech, Signal Process.* ASSAP-28(4): 357–366

Furui S 1981 Cepstral analysis technique for automatic speaker verification. *IEEE Trans. Acoust., Speech, Signal Process.* 29(2): 254–272

Furui S 1986 Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Trans. Acoust., Speech, Signal Process.* ASSP-34: 52–59

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

Ho T K, Hull J J, Srihari S N 1994 Decision combination in multiple classifier systems. *IEEE Trans. Pattern Analysis Machine Intelligence* 16(1): 66–76

Jayanna H S, Prasanna S R M 2009 Multiple frame size and rate analysis for speaker recognition under limited data condition. *IET Signal Processing* 3(3): 189–204

Kimball O, Schmidt M, Gish H, Waterman J 1997 Speaker verification with limited enrollment data. *Proc. European Conf. Speech Commun. and Tech. (EUROSPEECH'97)* Rhodes, Greece

Kinnunen T, Hautamaki V, Franti P 2003 On the fusion of dissimilarity-based classifiers for speaker identification. *Proc. EUROSPEECH* Geneva, Switzerland

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

Lee B-J, Yoon S-W, Kang H-G, Youn D H 2006 On the use of voting methods for speaker identification based on various resolution filterbanks. *Proc. Int. Conf. Acoust., Speech, Signal Process.* 917–920 Toulouse, France 917–920

Li K P 1985 An approach to text-independent speaker recognition with short utterances. *Proc. IEEE Int. Conf. Acoust., Speech, signal Process.* San Diego, California

Man-Wai Mak, Hsiao R, Mak B 2006 A Comparison of various adaptation methods for speaker verification with limited enrollment data. *Proc. IEEE Int. Conf. Acoust., Speech, signal Process.* Toulouse, France 1–4

Murthy K S R, Yegnanarayana B 2006 Combining evidence from residual phase and MFCC features for speaker recognition. *IEEE Signal Process. Lett.* 13(1): 52–56

Nagarajan T 2004 *Implicit systems for spoken language identification*. Indian Institute of Technology Madras, Dept. of Computer Science, Chennai, India

NIST 2003 http://www.itl.nist.gov/iad/mig//tests/sre/2003/2003-spkrec-evalplan-v2.2.pdf[online]

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)

Rahman A, Fairhurst M 2000 Decision combination of multiple classifiers for pattern classification: Hybridisation of majority voting and divide and conquer techniques. *Proc. IEEE-Workshop Appl., Computer Vision* California 58–63

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

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

Rosenberg A E Automatic speaker verification: A review. *Proc. IEEE* 64(4): 475–487

Sarada G L, Hemalatha N, Nagarajan T, Murthy H A 2004a Automatic transcription of continuous speech using unsupervised and incremental training. *Proc. Int. Conf. Spoken Language Process.* Jeju Island, Korea

Sarada G L, Nagarajan T, Murthy H A 2004b Multiple frame size and multiple frame rate feature extraction for speech recognition. *Proc. Int. Conf. Signal Process. Communication* Bangalore, India

Xu L, Krzyzak A, Suen C Y 1992 Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Trans. Syst., Man Cybern.* 22(3): 412–435

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

Zue V, Seneff S, Glass J 1990 Speech database development at MIT: TIMIT and beyond. *Speech Communication* 9(4): 351–356