ELSEVIER

Speech Communication 39 (2003) 301–310

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

# Speaker-specific mapping for text-independent speaker recognition

Hemant Misra [a,*], Shajith Ikbal [b], B. Yegnanarayana [b,1]

[a] *Department of Electrical Engineering, Indian Institute of Technology, Madras 600 036, India*
[b] *Department of Computer Science and Engineering, Indian Institute of Technology, Madras 600 036, India*

Received 31 July 2000; received in revised form 27 August 2001; accepted 14 January 2002

**Abstract**

In this paper, we present the concept of speaker-specific mapping for the task of speaker recognition. The speaker-specific mapping is realized using a multilayer feedforward neural network. In the mapping approach, the aim is to capture the speaker-specific information by mapping a set of parameter vectors specific to linguistic information in the speech, to a set of parameter vectors having linguistic and speaker information. In this study, parameter vectors suitable for speaker-specific mapping are explored. Background normalization for score comparison and network error criterion for frame selection are proposed to improve the performance of the basic system. It is shown that removing the high frequency components of speech results in loss of performance of the speaker verification system. For all the 630 speakers of the TIMIT database, an equal error rate (EER) of 0.5% and 100% identification is achieved by the mapping approach. On a set of 38 speakers of the dialect region "dr1" of NTIMIT database, an EER of 6.6% is obtained.


**Résumé**

Dans ce papier, nous présentons une approche de reconnaissance du locuteur basée sur une projection spécifique à chaque utilisateur. Cette projection est réalisée au moyen d'un réseau de neurones multi-couches. Le but de la projection est de capturer les informations spécifiques au locuteur en transformant un ensemble de paramètres représentant l'information linguistique en un ensemble de paramètres caractérisant l'information linguistique ainsi que l'information propre au locuteur. Dans cette étude, les paramètres les plus appropriés pour faire cette tranformation sont également évalués. On montre aussi que la normalisation des scores, ainsi que l'utilisation du critère d'erreur du réseau de neurone pour la sélection des vecteurs acoustiques, augmentent les performances du système. Nous montrons également que le fait de laisser tomber les composantes haute fréquence du signal résulte en une déterioration des performances du système. Sur un ensemble de 630 locuteurs de la base de données TIMIT, un égal taux d'erreur de 0.5% et 100% d'identification sont obtenus par l'approche proposée ici. Sur un ensemble de 38 locuteurs de la région dialectale "dr1" de la base de données NTIMIT, un égal taux d'erreur de 6.6% est obtenu.


*Abbreviations:* MLFFNN: multilayer feedforward neural network; Ne criterion: network error criterion; EER: equal error rate; ANN: artificial neural network; LI: linguistic information; SI: speaker information; LSI: linguistic and speaker information; GMM: Gaussian mixture model; HMM: hidden Markov model; BG: background.

* Corresponding author.
*E-mail address:* yegna@iitm.ernet.in (B. Yegnanarayana).
[1] Tel.: +91-44-2354591; fax: +91-44-2350509.

0167-6393/02/$ - see front matter 
PII: S0167-6393(02)00046-8

*Keywords:* Speaker recognition; Artificial neural network; Speaker-specific mapping; Linguistic information; Speaker information; Background normalization; Network error criterion; Equal error rate

## 1. Introduction

Speech signal contains information about speaker, emotional state of the speaker, and the linguistic information (LI) that corresponds to the message part of the speech signal. All the above information is embedded in the speech signal. Human beings seem to perform effortlessly the task of extracting the relevant information of each component. It is necessary to extract features specific to a speaker for recognizing the speaker. LI and emotional state of the speaker have also clues for speaker identity. For example, some speakers may use some phrases or words more often than others.

There are two types of tasks in speaker recognition: identification (Gish and Schmidt, 1994) and verification (Rosenberg, 1976). In the identification task, given a test utterance, the goal is to find the identity of the test speaker. In contrast, in the verification task, a test speaker claims himself as a speaker enrolled with the system, and the goal is to check the validity of the speaker's claim. The output of a speaker verification system is binary, i.e., the claim of a test speaker is either accepted or rejected. This decision is made based on similarity of the test utterance to the target model. A threshold is required to check the level of similarity (Matsui et al., 1996).

Speaker recognition can be performed in three modes: text-dependent, text-prompted and text-independent. In a text-dependent system (Furui, 1981) the training and testing utterances are same, and matching is done usually at the acoustic level. In a text-prompted system, system prompts a user to speak a combination of digits (or words) randomly selected from a pre-stored set of digits (or words) at the time of testing. A speech recognition system is used to verify whether the user has uttered the same digits (or words) as prompted by the system. If the user utters the correct digits (or words) then only his/her claim is verified. In a text-independent system (Gish and Schmidt, 1994), training and testing utterances need not be the same. Some statistical characteristics of speech features are used to arrive at a decision. The four main issues in speaker recognition are the following: feature extraction from speech signal, generation of speaker model, matching and decision logic.

In this paper a mapping approach is proposed for the task of text-independent speaker recognition. In this approach the mapping property (Funahashi, 1989; Hornik, 1991) of a multilayer feedforward neural network (MLFFNN) is used to generate a model for each speaker (Haykin, 1999; Lippmann, 1987; Yegnanarayana, 1999). In Section 2, the background (BG) for the mapping approach is presented. Analysis of the mapping approach is given in Section 3. The choice of parameters/features and their suitability for speaker-specific mapping are discussed in Section 4. In Section 5, the issue of BG normalization is addressed in the context of the proposed mapping approach. A method for selecting speaker-specific frames is proposed in the same section. In Section 6, the importance of the high frequency components for speaker recognition is examined.

## 2. Mapping approach

Gong and Haton (1992) proposed a "nonlinear vectorial interpolation function" in their text-dependent speaker recognition studies. They used the mapping property of a MLFFNN to obtain the "interpolation vector" for each speaker. For these studies, transcription of the utterances was required during training and testing. Hermansky and Malayath (1998) suggested speaker-specific mapping approach for text-independent speaker recognition. Cepstral coefficients derived from the perceptual linear prediction (PLP) (Hermansky et al., 1992) were used as features, and the results were comparable to the approach based on Gaussian mixture model (GMM) (Reynolds, 1994, 1995).

In the mapping approach the goal is to capture the speaker-specific information. This may be accomplished by mapping a set of parameter vectors specific to LI in the speech, to a set of parameter vectors having linguistic and speaker information (LSI). Such a mapping function is captured for every speaker. The choice of mapping from LI to LSI is better than mapping from LSI to LI, as discussed later in this section.

Linear prediction (LP) (Makhoul, 1975) analysis provides some clues to obtain parameters that contain predominantly LI or LSI. The order of the LP analysis determines the number of peaks of an all-pole system. Each complex pole-pair accounts for one resonance peak, and the real poles account for the roll-off of the spectrum. A low (4–8) order LP analysis captures the gross features of the envelope of speech spectrum. Speaker information (SI) may be lost in such a representation, while LI may be preserved. In contrast, a higher (>12) order LP analysis captures both the gross and finer details of the envelope of the spectrum, thus preserving both linguistic as well as speaker-specific information. But if the order of the LP analysis is very high, then the model may also capture spurious peaks in the spectrum. Generally it is difficult to determine the correct LP analysis order (and model) for capturing the LSI.

Fig. 1 shows some results of qualitative analysis to demonstrate the speaker-specific features in the LP spectrum. Speech utterances sampled at 16 kHz were collected from two speakers (1 female and 1 male) over a microphone. Four utterances for the vowel /a/ were collected from each speaker. A frame size of 20 ms was considered for analysis. The samples were Hamming windowed after differencing (pre-emphasis). Figs. 1(a) and 1(b) show the LP spectra for the female speaker obtained using 6th and 14th order LP analysis, respectively. Each spectrum pair was obtained for the same segment of the utterance. Similarly, Figs. 1(c) and 1(d) are the LP spectra for the male speaker. We observe the following: For the 6th order LP analysis, the spectra for different utterances of the same speaker are similar, and the spectra of the two speakers are also similar. For the 14th order LP analysis, the spectra for different utterances of the same speaker are similar, and the spectra of the two speakers are significantly different. These differences can be attributed to the speakers because the LI is same for both the speakers.

Fig. 2 shows the LP spectra for four different speakers. Speech for the utterance of the vowel /a/ was collected from four speakers (1 female and 3 males). As in the previous experiment, the steady portions of the utterances were considered for analysis. Figs. 2(a) and 2(b) show the spectra obtained with 6th and 14th order LP analysis, respectively, for the utterances of the four speakers. Each spectrum pair belongs to one speaker, and the lower and higher order LP analysis was performed on the same segment of the utterance. For the 6th order LP analysis, the spectra for all the four speakers are similar, except for small differences in the high frequency region (Fig. 2(a)). But for the 14th order LP analysis (Fig. 2(b)), the spectra of the speakers are significantly different. The 6th order LP analysis qualitatively validates the assumption that the parameter vector contains LI, since for the same vowel sound the spectra are similar for all the speakers. For the 14th order LP analysis, the spectra of the four speakers are significantly different. These variations can be attributed to SI, since the LI is same in all the four cases. Thus the low order LP spectrum (6th order in the present case) captures the LI, whereas the higher (14th) order LP spectrum captures both LSI. This study shows that by choosing suitable order for LP analysis, parameter vectors containing LI or LSI can be extracted from speech signals.

It appears that a mapping function can be derived by mapping the LI feature vector to LSI feature vector or vice versa. But mapping from LSI to LI is likely to be less effective because of the nonuniqueness of the projection from a higher resolution spectrum to a lower resolution spectrum. We will also show experimentally that this conjecture is true.

## 3. Analysis of mapping approach

Let $\mathscr{F}(\cdot)$ denote the nonlinear mapping from LI to LSI. For $N$ input–output vector pairs, the average mapping error is given by

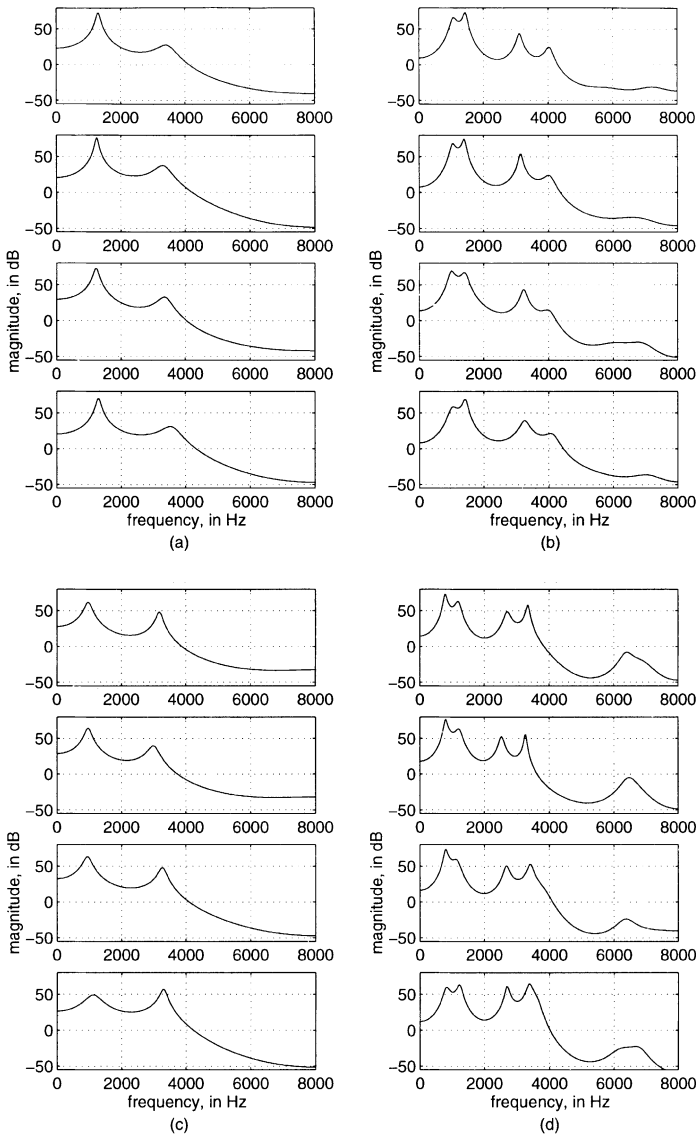


Fig. 1. (a) and (b) are the 6th and 14th order LP spectra for four different utterances of the same vowel sound (/a/) by a female speaker, respectively. (c) and (d) are similar spectra for a male speaker. Segment size is 20 ms and sampling frequency is 16 kHz.

$$E(\mathscr{F}) = \frac{1}{N} \cdot \sum_{n=1}^{N} |\boldsymbol{o}_n - \mathscr{F}(\boldsymbol{i}_n)|^2, \tag{1}$$

where the vector $\boldsymbol{i}_n$ contains the LI, the vector $\boldsymbol{o}_n$ contains the LSI, and $|\cdot|$ is the Euclidean distance between the two vectors. The aim is to find the mapping function such that $E(\mathscr{F})$ is minimized. The assumption in this analysis is that $(\boldsymbol{i}_n, \boldsymbol{o}_n)$, $n = 1, \ldots, N$, are related by the mapping function $\mathscr{F}(\cdot)$. As $\boldsymbol{i}_n$ represents the LI and $\boldsymbol{o}_n$ the LSI, the mapping function that minimizes $E(\mathscr{F})$ should be specific to the speaker. The task is to derive this speaker-specific mapping function using $(\boldsymbol{I}, \boldsymbol{O})$, where $\boldsymbol{I} = \{\boldsymbol{i}_1, \boldsymbol{i}_2, \ldots, \boldsymbol{i}_N\}$ and $\boldsymbol{O} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_N\}$.

Let us assume that there exists a function $\mathscr{F}_k(\cdot)$ for the $k$th speaker that performs a nonlinear

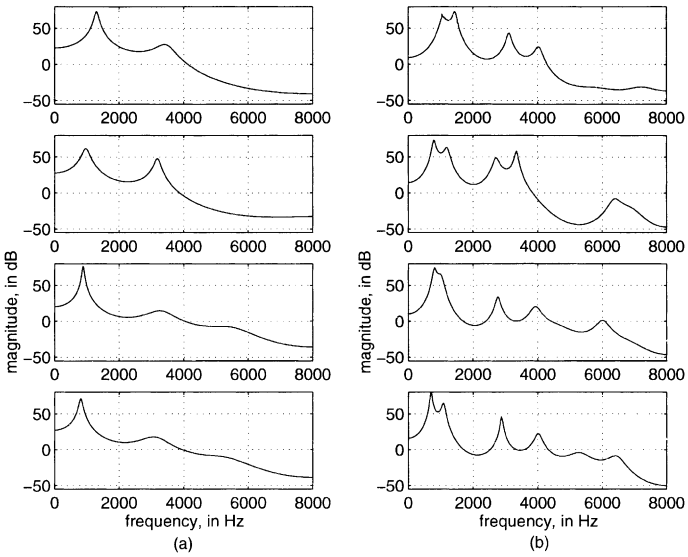


Fig. 2. (a) Sixth order LP spectra for the vowel sound (/a/) spoken by four different speakers and (b) 14th order LP spectra for the same segment as in part (a). Segment size is 20 ms and sampling frequency is 16 kHz.

mapping between $\boldsymbol{I}$ and $\boldsymbol{O}$. The free parameters of $\mathscr{F}_k$ are adjusted by minimizing $E(\mathscr{F}_k)$ to obtain the $k$th speaker model ($M_k$). In the case of identification, the speaker model ($M_k$) that gives the least error for a test utterance is adjudged as the identified speaker.

After selecting the parameter vectors suitable for mapping, the next task is to derive the mapping function itself. The nonlinear speaker-specific mapping function can be captured using an MLFFNN. In MLFFNN, the mean-square-error is minimized using a gradient descent algorithm. In the present studies, the parameter vectors obtained from the training data are used to adjust the weights using backpropagation learning law (Haykin, 1999; Yegnanarayana, 1999). Models ($M_k$) are derived for each speaker ($k$) using all the frames in the training data of that speaker. For testing, the input parameter vector is given to each of the MLFFNN. The difference between the desired output vector and the actual output vector of the MLFFNN is used as distance for that frame. The total accumulated distance averaged over all the test frames gives an indication of the proximity of the test utterance to the speaker model. Test data different from the training data is used to study the performance of the speaker models for identification and verification. For identification, the parameter vectors of the test utterance of each speaker are given to each of the models. The speaker model that gives the least distance is the identified speaker. For verification, the performance is expressed in terms of equal error rate (EER). The EER is computed using the $K \times K$ matrix of distances, obtained when the test utterance of each speaker is evaluated with models of all the speakers, where $K$ is the number of registered speakers. For computation of EER, the minimum and maximum distance values are found from the distance matrix, and the threshold for acceptance is incremented from the minimum value to the maximum value in small steps. At each threshold value, false acceptance and false rejection rates are computed. For the threshold value where the false acceptance rate is equal to the false rejection rate, the corresponding error is marked as EER.

## 4. Parameters for speaker-specific mapping

To derive the parameter vectors suitable for the mapping approach, we have used the LPC derived cepstral coefficients (Furui, 1981; Rabiner and

Juang, 1993). In the following experiments, 38 speakers (14 female and 24 male) of the dialect region "dr1" of the NTIMIT (Jankowski et al., 1990) database are used. Out of the 10 sentences of each speaker, 2 sentences are same for all the speakers. The remaining 8 sentences are different for each speaker. We have used the common two sentences plus five other sentences for training the speakers' models. The remaining three sentences, used for testing, are different for each speaker. We have used the average distance, obtained from concatenation of all the three test utterances, to calculate the performance. This provides a text-independent evaluation of the system.

A frame size of 20 ms and a frame shift of 10 ms are used in these experiments. Speech is pre-emphasized and Hamming windowed. Durbin's algorithm is used to extract the LP coefficients (Rabiner and Juang, 1993). Cepstral coefficients are obtained from the LPCs using the following recursive relation:

$$c_0 = \ln \sigma^2,$$

$$c_m = a_m + \sum_{k=1}^{m-1} \frac{k}{m} \cdot c_k \cdot a_{m-k} \quad 1 \leqslant m \leqslant p$$

$$= \sum_{k=1}^{m-1} \frac{k}{m} \cdot c_k \cdot a_{m-k} \quad m > p,$$

where $\sigma^2$ is the gain term in the LPC model, $c_n$ is the $n$th cepstral coefficient, and $a_m$ is the $m$th LP coefficient. Linearly weighted cepstral coefficients (Yegnanarayana and Reddy, 1979) are used as a parameter vector in these experiments.

Let $m$ and $n$ be the orders of the LP analysis for the input and output vectors, respectively. From each set of LPCs, 20 weighted cepstral coefficients are obtained. The first coefficient, being zero, is ignored. The remaining 19 coefficients are used as a parameter vector. Even though the dimensions of the input and output parameter vectors are the same (19), the corresponding LP analysis orders are different. A dimension higher than the order of the LPCs is used to represent the weighted cepstral parameter vector, as this will improve the resolution of the formants in the spectral domain corresponding to the parameter vector. Note that normally a large dimension weighted cepstral vector is required to represent the LP spectrum accurately.

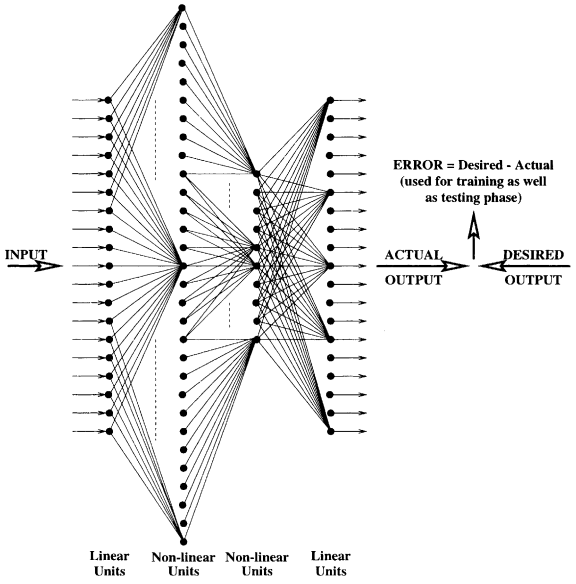


Fig. 3. Structure of an MLFFNN to capture speaker-specific mapping feature.

The structure of the MLFFNN used in these studies is 19L30N10N19L, as shown in Fig. 3, where L denotes linear units and N denotes non-linear units. The nonlinear activation function of each unit is given by $(16/9)\tanh(2x/3)$, where $x$ is the input activation value. The number before L (or N) denotes the number of units in that layer. The first layer is the input layer (19L), while the last one is the output layer (19L). The first hidden layer has 30 nonlinear units and the second hidden layer has 10 nonlinear units. In the experiments, the order ($p$) of LP analysis is $m$ for the input and $n$ for the output parameter vectors, respectively, whereas the size of the input and output parameter vectors is 19. The weights of the network are initialized with randomly generated numbers in the range $-0.5$ to 0.5.

Table 1 shows the performance of the mapping approach for the verification task in terms of EER. The entries in the table are the performance for a pair of LP orders used for deriving the input and output vectors. The upper triangular matrix of the table shows that, for a given order ($m$) of the LP analysis for deriving the *input* parameter vector, as

Table 1
Verification performance by mapping approach

| LP order for input | LP order for output | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 6 | 8 | 10 | 12 | 14 | 16 | 18 | 20 |
| 6 | 27.3 | 24.1 | 21.1 | 15.8 | 13.2 | 15.8 | 18.5 | 20.4 |
| 8 | 21.1 | 31.2 | 24.5 | 15.8 | 18.5 | 21.1 | 23.7 | 29.0 |
| 10 | 23.7 | 22.6 | 33.6 | 23.7 | 23.5 | 26.4 | 28.5 | 29.0 |
| 12 | 20.3 | 20.7 | 22.9 | 36.7 | 26.3 | 26.4 | 30.9 | 29.0 |
| 14 | 23.7 | 25.8 | 18.5 | 20.3 | 32.7 | 28.6 | 29.5 | 29.0 |
| 16 | 24.5 | 20.5 | 21.1 | 19.3 | 28.7 | 31.6 | 28.6 | 27.8 |
| 18 | 27.9 | 23.7 | 17.9 | 16.3 | 21.9 | 28.4 | 36.3 | 29.0 |
| 20 | 27.0 | 22.7 | 23.0 | 22.4 | 23.7 | 27.5 | 29.0 | 26.4 |

Input–output parameters are LPC derived cepstral coefficients of dimension 19 (1st coefficient is ignored); number of speakers is 38; network structure is 19L30N10N19L.

the order ($n$) of the LP analysis for the *output* is increased, initially the EER decreases, reaches a minimum value, and then starts increasing. We can interpret that in this case the LI is kept constant at the input, and the SI is progressively increased at the output. For very high order of the LP analysis for deriving the output vector, the LP spectrum starts picking up spurious peaks, masking the speaker-specific information. That is why the EER performance is decreased when $n$ is increased beyond 14. On the other hand, for a given order of the LP analysis for the *output* parameter vector, as the order of the analysis for the *input* increases, the EER increases in the upper triangular matrix of the table. This is because as the difference between the input and output orders is reduced, the network cannot capture any speaker-specific information. That is why the performance of the system is least when there is no difference between the input and output order, as can be seen from the diagonal values.

In the lower triangular matrix, the input to the network can be interpreted to contain LSI, and the desired output to contain LI. Although from the ER values we can see that even in this case the network captures some speaker-specific information when the difference between the input and the output orders is more, the performance in these cases is relatively poorer compared to the results in the upper triangular matrix corresponding to mapping from LI to LSI. This is because in the case of mapping from LSI to LI, the network has to produce a smooth output spectrum compared to the input spectrum. The desired output is also a smooth spectrum, and hence the difference between the desired and the actual smooth spectra will not be able to bring out the discrimination between speakers. Thus we cannot expect symmetry in the performance matrix. The best performance of 13.2% EER is obtained when the order of the LP analysis is 6 for the input and 14 for the output. This result is significant in the sense that it validates our conjecture that mapping from LI to LSI indeed captures the speaker-specific information. We use this mapping in the speaker recognition experiments discussed in the following sections.

## 5. Background normalization and frame selection

So far the Euclidean distance between the output of the network and the desired output parameter vector was used for evaluating the performance of a speaker model relative to the models of other speakers. The concept of relative score with respect to a reference BG model is known to improve the performance of a speaker verification system (Heck and Wientraub, 1997). Therefore a BG model is generated using the parameter vectors extracted from speech utterances of a large number of speakers. These speakers are different from the speakers registered with the system. An MLFFNN is trained with the pooled input–output parameter vectors from all these speakers. The weights of the BG model are used as

initial weights to train each speaker model. This is to avoid any bias the choice of arbitrary initial weights may introduce while generating a speaker model. The structure of the speaker model and the orders of the LP analysis for deriving the input and output parameter vectors are same as those for the BG model. A speaker model is thus fine tuned to a given speaker, over and above the BG model. The relative score for the test utterance of a speaker is obtained using the difference between the average distance for the BG model and the speaker model.

For generating the BG model, all the 76 speakers of the "dr2" set of NTIMIT are used. The first 7 utterances from each of the 76 speakers are taken, and the input–output parameter vector pairs are extracted. A frame size of 20 ms and a frame shift of 10 ms are used. Speech signal is pre-emphasized and Hamming windowed. The orders of the LP analysis for deriving the input and output parameter vectors are 6 and 14, respectively. The total number of input–output vector pairs obtained from the 7 utterances of the 76 speakers is 160442. From this pool of parameter vector pairs, one sixth of the vector pairs are taken at random to train the BG model. The experimental conditions are same as used in Section 4. The same 38 speakers of "dr1" of NTIMIT are considered for enrollment. Out of the 10 utterances of each speaker, 7 are used for training and 3 for testing. The MLFFNN structure used for generating the BG model as well as the speaker model is 19L30N10N19L.

The performance in terms of EER using the relative score is shown in Table 2 for different number of iterations used for training the BG model. It is to be noted that poor training (say using 1 iteration) of BG model may be viewed as random initialization for further training to generate a speaker model. This gives an EER of about 18%. This is because poor initialization leads to poor normalization. On the other hand, proper training of the BG model using 10 or more iterations results in improved performance due to better normalization. The use of the BG model for normalization of the scores reduced the EER significantly from 13.2% (in Table 1) to 7.1% (in Table 2). It is interesting to note that the performance is relatively invariant to the number of iterations used for training the BG model.

Table 2
Effect of number of iterations for BG model on the performance of speaker verification

| Number of iterations | 10 | 20 | 30 | 40 | 50 |
|---|---|---|---|---|---|
| EER | 7.8 | 7.9 | 7.5 | 7.1 | 7.1 |

The EER values are for a set of 38 speakers. The network structure is 19L30N10N19L.

In speaker recognition, all the frames of a speech utterance may not be equally important. Some frames may contain significant speaker-specific information (Eatock and Mason, 1990). If such frames are used, then the performance of the system can be improved. In the present work, we suggest a *network error* (Ne) criterion to select frames having speaker-specific information. This criterion involves training the speaker model initially for a fixed number of iterations (about 50) using all the frames of the speaker data. At this stage the frames that give the lower distance are termed as good frames, and are segregated from the bad high distance frames. The initially trained speaker model is further trained using the good frames. The number of iterations for training the model in the second phase of training is nearly 600 in the present study. The EER for the 38 speakers set is reduced from 7.1% to 6.6% when the Ne criterion is used.

## 6. Significance of high frequency components

For the studies in the previous sections, the NTIMIT database was used. The NTIMIT database was derived from the TIMIT database (Jankowski et al., 1990). In this section the results obtained on both TIMIT and NTIMIT databases are compared.

The same 38 speakers of the dialect region "dr1" of the TIMIT and NTIMIT are used for comparison. Two experiments are conducted in this comparative study. In the first experiment, the standard TIMIT database is used. In the second experiment, the TIMIT utterances are passed through a finite impulse response (FIR) low pass

Table 3
Performance of mapping approach for 38 speakers set

| Database | Processing | EER |
|---|---|---|
| TIMIT[a] | Bandwidth of 8 kHz | 0.5 |
| TIMIT[b] | Bandwidth of 3.6 kHz | 6.1 |
| NTIMIT[c] | Telephone channel | 6.6 |

[a] For TIMIT speech.
[b] For low pass filtered TIMIT speech.
[c] For NTIMIT speech.

filter (LPF) of order 33 with a cut-off at 3560 Hz. The objective of the experiment is twofold: (1) to see the effect of removing the high frequency components on the performance of the system, and (2) to check the effect of channel distortions on the performance of the mapping approach. The results of these two experiments are compared with the results of the experiment with the same 38 speakers set from the NTIMIT database that was used in the previous studies.

The results of the three experiments are shown in Table 3. The EER obtained on the TIMIT database is low as expected. When the clean speech of the TIMIT data is low-pass filtered, the performance is degraded to an EER value of 6.1%, which is comparable to the performance of 6.6% EER for the NTIMIT data. This shows that speaker-specific information is available in the higher frequencies (above 4 kHz), and it is lost in the low-pass filtered TIMIT data as well as in the NTIMIT data.

Finally, the performance of the mapping approach is obtained using all the 630 speakers of the TIMIT database. The results are that we get an identification of 100% and an EER of 0.5%. These results are similar to the performance obtained using GMM (Reynolds et al., 1995), indicating that the proposed mapping approach indeed captures speaker-specific mapping.

## 7. Summary and conclusion

In this paper we have shown that speaker-specific information can be captured by suitable mapping of parameter vectors. The aim was to find out the parameter vector pair suitable for the mapping approach. The parameter vector pair derived from the 6th and 14th order LP analysis for the input and output, respectively, was found to be most suitable. The input vector may be considered as representing the LI and the output vector representing the LSI.

We proposed a BG normalization technique to improve the performance of the system. An Ne criterion was proposed to select frames having significant speaker-specific information. An EER of 6.6% was obtained when both the BG normalization and Ne criterion were used together, compared to an EER of 13.2% without using them.

We have shown experimentally that the high frequency components are important for a speaker verification system. The mapping approach was shown to perform as well as the GMM-based approach for all the 630 speakers of the TIMIT database.

The speaker-specific mapping was captured from the training data itself, and no assumption about the underlying probability density functions was made. Another advantage in the proposed method is that the number of free parameters is significantly less compared to a GMM-based approach. The number of free parameters (weights of the network) in the present studies are 1119, whereas in a GMM they are typically 19456 for a 1024 mixture model and for a parameter vector dimension of 19, assuming that the variances of the Gaussians are held constant.

## References

Eatock, J.P., Mason, J.S., 1990. Automatically focusing on good discriminating speech segments in speaker recognition. In: Proc. Internat. Conf. on Spoken Language Process., Kobe, Japan, pp. 133–136.

Funahashi, K.-I., 1989. On the approximate realization of continuous mappings by neural networks. Neural Networks 2, 183–192.

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

Gish, H., Schmidt, M., 1994. Text-independent speaker identification. IEEE Signal Process. Magazine (October), 18–32.

Gong, Y., Haton, J.-P., 1992. Non-linear vectorial interpolation for speaker recognition. In: Proc. IEEE Internat. Conf. on

Acoust. Speech Signal Process., San Francisco, California, USA, Vol. 2, pp. II173–II176.

Haykin, S., 1999. Neural Networks: A Comprehensive Foundation. Prentice-Hall International, NJ.

Heck, L.P., Wientraub, M., 1997. Handset-dependent background models for robust text-independent speaker recognition. In: Proc. IEEE Internat. Conf. on Acoust. Speech Signal Process., Munich, Germany, Vol. 2, pp. 1071–1074.

Hermansky, H., Malayath, N., 1998. Speaker verification using speaker-specific mapping. In: Proc. RLA2C, Avignon, France.

Hermansky, H., Morgan, N., Bayya, A., Kohn, P., 1992. RASTA-PLP speech analysis technique. In: Proc. IEEE Internat. Conf. on Acoust. Speech Signal Process., San Francisco, CA, USA, Vol. 1, pp. 1121–1124.

Hornik, K., 1991. Approximation capabilities of multilayer feedforward neural networks. Neural Networks 4, 251–257.

Jankowski, C., Kalyanswamy, A., Basson, S., Spitz, J., 1990. Ntimit: a phonetically balanced, continuous speech, telephone bandwidth speech database. In: Proc. IEEE Internat. Conf. on Acoust. Speech Signal Process., Albuquerque, NM, USA, Vol. 1, pp. 109–112.

Lippmann, R.P., 1987. An introduction to computing with neural nets. IEEE ASSP Magazine (April), 4–22.

Makhoul, J., 1975. Linear prediction: A tutorial review. Proc. IEEE 63 (April), 561–580.

Matsui, T., Nishitani, T., Furui, S., 1996. Robust methods to updating model and a-priori threshold in speaker verification. In: Proc. IEEE Internat. Conf. on Acoust. Speech Signal Process., Atlanta, GA, USA, pp. 97–100.

Rabiner, L., Juang, B.-H., 1993. Fundamentals of Speech Recognition. Prentice Hall, Englewood Cliffs, NJ.

Reynolds, D.A., 1994. Speaker identification and verification using Gaussian mixture speaker models. In: Proc. ESCA Worksh. of Automatic Speaker Recognit., Identification and Verification. Martigny, Switzerland, April, pp. 27–30.

Reynolds, D.A., 1995. Speaker identification and verification using gaussian mixture models. Speech Communication 17 (1–2), 91–108.

Reynolds, D.A., Zissman, M.A., Quatieri, T.F., O'Leary, G.C., Carlson, B.A., 1995. The effects of telephone transmission degradations on speaker recognition performance. In: Proc. IEEE Internat. Conf. on Acoust. Speech Signal Process., Detroit, MI, USA, pp. 329–332.

Rosenberg, A.E., 1976. Automatic speaker verification: A review. Proc. IEEE 64 (4), 475–487.

Yegnanarayana, B., 1999. Artificial Neural Networks. Prentice Hall India, Connaught Circle, New Delhi.

Yegnanarayana, B., Reddy, D.R., 1979. A distance measure based on the derivative of linear prediction phase spectrum. In: Proc. IEEE Internat. Conf. on Acoust. Speech Signal Process., Washington DC, USA, pp. 744–747.