Hierarchical Speaker Identification based on Latent Variable Decomposition

Sabyasachi Patra
School of Computer Engineering
KIIT University
Bhubaneswar, Orissa, India

Subhendu Kumar Acharya
School of Computer Engineering
KIIT University
Bhubaneswar, Orissa, India

ABSTRACT
In this paper, a novel hierarchical speaker identification method based on Latent Variable Decomposition (LVD) has been proposed. Firstly, we got a coarse decision by a fast scan all registered speakers using LVD based features and GMM classifier to find R possible target speakers, and then MFCC or PCA based features were used to make final decision. LVD has another advantage: reduction of the feature vectors dimensions, and the noise is removed from speech simultaneity. So, it can reduce the computational complexity and improve the performance of speaker identification. The experimental results showed that the proposed method could improve recognition accuracy of system remarkably and the system has better robustness by comparing with the traditional speaker identification method.

General Terms
Speaker Identification, Feature Extraction, Dimension Reduction

Keywords
MFCC, PCA, GMM, LVD, SNR, Features, Classifiers

1. INTRODUCTION
Speaker identification is an important embranchment of speaker recognition. Automatic speaker identification is a process in which a machine identifies a person from his or her voice characteristic [1][2]. Recent advances in research and development in speaker recognition and identification system have resulted in the speaker identification becoming one of the most trusted methods for authorization and in forensics. However the field deployments of such speaker identification system call for their ability to work in noisy environment with minimal time complexity. The challenge of designing such efficient and robust speaker identification system has been receiving enhanced attention from the research community and this has been identified as the focus of this work.

The process of speaker identification is divided into two main phases e.g. the enrollment phase (training) and the identification phase (testing). Feature extraction is common in both the phases. In current states of arts, the information contained in the power spectral envelope (in the form of Cepstral coefficients) has been used as features for speaker identification. Though the Cepstral coefficients perform quite well for speaker identification, its performance can further be improved by using the hidden features of the speech signal. The Latent Variable Decomposition is applied on the magnitude spectral vector of the speech signal to extract the hidden features embedded in it. In this method the distribution of the spectral vectors is represented in terms of the mixture multinomial distribution of apriori probability of some fixed number of hidden class and the multinomial distribution of frequency beam given the hidden class, which constitute the elements of the transform matrix used to obtain the new feature vectors.

In this paper both MFCC features and LVD based features are used for hierarchical speaker identification. Once the features extraction is over, the next stage in enrollment phase is speaker modeling. GMM classifier is proved to be best classifier for speaker modeling in last 15 years. In the proposed approach two models are created for each speaker using GMM classifier one with MFCC feature vectors (or PCA based features) and the other with LVD based features during training. A hierarchical model is used for feature matching during identification phase. In the first matching using LVD based model 20% of the total speakers are chosen as probable speakers for the next stage. This stage is named as candidate selection stage. Then feature matching is performed on the candidate selected in the first stage using MFCC (or PCA) features based model. Once the scores are obtained for the selected candidates in both stages opinion fusion is used to take final decision.

2. DATABASE
TIMIT (Texas Instrument Massachusetts Institute of Technology) [3] database is used in this paper. The TIMIT database consists of 630 speakers, out of which 70% are male and 30% are female from 10 different dialect regions in America. Each speaker has approximately 30 seconds of speech spread over 10 utterances. The speech is recorded using a high quality microphone in a sound proof booth at a sampling frequency of 16 KHz, with no session intervals between recordings. The speech of each speaker consists of 2 dialect sentences (SA), 3 phonetically compact sentences (SX) and 5 phonetically diverse sentences (SI). In this paper 200 speakers are taken out of which 100 are male and 100 are female. The experiments are conducted by adding white Gaussian noise on clean speech of TIMIT database in different SNR. White noise is dynamically generated using MATLAB toolbox.

3. SPEAKER IDENTIFICATION SYSTEM
In this paper experiments are based on closed set speaker identification. During experiments training data is prepared by concatenating 2 SA sentences, 3 SI sentences and 3 SX

sentences to produce a 24 seconds utterance containing 8 sentences for each speaker. The remaining two SX sentences are used as two independent tests segments. In this experiment 200 speakers (100 males and 100 females) are selected alphabetically from the TIMIT database. During training each speaker is trained by clean speech of TIMIT database where as the testing is done on TIMIT database contaminated with white Gaussian noise. Having acquired the testing and training utterances, it is now the role of the feature extractor to extract the acoustic features from the speech.

3.1 Feature Extraction and Parameter Estimation

MFCC feature vectors are used as original features. The MFCC feature extractor converts an utterance into a sequence of MFCC feature vectors [4]. It involves three steps, namely pre emphasis, frame blocking and windowing sections. In windowing, the input speech signal cuts into overlapping windows of equal length. Throughout the experiment a Hamming window of 16 ms length with the overlapping of 8 ms is fixed. The spectrum is calculated by using an FFT algorithm and the number of points used in the FFT algorithm is taken as the power of 2 greater than or equal to the frame size. The resulting power spectrum is windowed by a set of 26 triangular filters (mel filters) which are equally spaced apart by 1500 mels and each one having width of 3000 mels. Energy content of the speech signal is calculated across all triangular filters and then the MFCC coefficients are obtained by applying discrete cosine transform over it. To enhance the performance of the speaker identification system, time derivatives are added to the basic static parameters which are called delta coefficients and delta-delta coefficients. In this paper 13 cepstral coefficients are used along with their delta and delta-delta coefficients, results in 39-dimensional MFCC feature vectors are used for speaker modelling.

3.2 Speaker Modeling

Each speaker is modeled using one Gaussian Mixture Model (GMM) [5] with 32 mixture components. Each mixture component is characterized by its weight, mean vector and (diagonal) covariance matrix. The GMMs are trained using the EM algorithm [6] with an approximate model derived by a K-means algorithm. This algorithm converges in around 30 iterations which are used in this paper. During identification phase these models are used to identify the speaker from the given test utterance based on log likelihood score.

4. DIMENSION REDUCTION OF FEATURE VECTORS USING LVD

In the process of feature extraction, all speech signals are converted to sequences of magnitude spectral vectors through a short-time Fourier transforms. So the input speech signal is converted into a sequence of magnitude spectral vectors. These feature vectors are modeled as the outcome of a discrete random process [7] [8]. The process is modeled using a mixture of multinomial distributions, such that the mixture weights of the component multinomial vary from frame to frame. However the component multinomial themselves are assumed to be fixed for any speaker and they are learnt from training signals for each speaker through an EM algorithm.

A latent variable z governs the generation of a frequency f. The conditional probabilities for f (component multinomial) are assumed to be constant for any speaker; however the apriori probability of the latent variable z (mixture weights) varies from frame to frame. Thus the overall mixture multinomial distribution model for the spectrum of the t[th] frame is given by

$$p_t(f) = \sum_{z=1}^{Z} p_t(z)\, p(f \mid z) \qquad 1$$

Where $p_t(z)$ represents the a priori probability of z in the t[th] frame and $p(f \mid z)$ represents the multinomial distribution of f conditioned on the latent variable z. f takes the values of the discrete frequencies of the FFT for the frame. The magnitude spectral vectors are stored in the matrix XTF and the f[th] component of the t[th] feature vector is represented by XTF_{tf} which is equivalent to $p_t(f)$. Suppose T feature vectors are there and the dimension of each vector is 128. Each component of the feature vectors represents the energy content in a particular frequency band. LVD transforms these feature vectors into LVD space where each component of the new feature vector represents the energy content corresponding to a hidden class. The dimension of the new feature vectors depends on the number of hidden classes which is taken as 20 for this paper work. The transformation matrix is named as PFZ, whose elements PFZ_{fz} are represented by the multinomial distribution of f conditioned on the latent variable z i.e. $p(f \mid z)$. The matrix PTZ stores the apriori probability of z in the t[th] frame $p_t(z)$. The components of the mixture multinomial distribution of Equation 1 are initialized randomly and re-estimated through iterations governed by the following equations, which are derived through the expectation maximization algorithm:

$$p_t(z \mid f) = \frac{PTZ_{tz}\, PFZ_{fz}}{\sum_{z=1}^{Z} PTZ_{tz}\, PFZ_{fz}} \qquad 2$$

$$PFZ_{fz} = \frac{\sum_{t \in T} XTF_{tf}\, p_t(z \mid f)}{\sum_{t \in T} \sum_{f \in F} XTF_{tf}\, p(z \mid f)} \qquad 3$$

$$PTZ_{tz} = \frac{\sum_{f \in F} XTF_{tf}\, p_t(z \mid f)}{\sum_{z \in Z} \sum_{f \in F} XTF_{tf}\, p_t(z \mid f)} \qquad 4$$

Where Z is the number of hidden class, F is the dimension of feature vectors and T is the number of feature vectors. After the estimation of the matrix PFZ, it is used as a transformation

matrix which transforms the original features x_n into LVD space by the following relationship

$$y_n = V^T(x_n - \bar{x})$$

Where $V = \{v_1, v_2, ..., v_q\}$, are the latent vectors taken from the rows of the matrix PFZ and y_n is the output feature vector. The dimension of the output feature vector reduced from F to Z.

5. HIERARCHICAL SPEAKER IDENTIFICATION BASED ON LVD

The proposed hierarchical identification method has two identification stages. The R possible target speakers are first obtained by using the LVD features based GMM classifier, and then the target speaker is finally found out from these R speakers by MFCC or PCA features based GMM classifier.

5.1 LVD based Features for Coarse Decision

In the figure 1, dotted line represents the ENROLLMENT phase and solid line represents the CANDIDATE SELECTION phase. The first block is the feature extraction block which is common for both the ENROLLMENT phase and CANDIDATE SELECTION phase. Feature extraction block is meant for the extraction of 128 dimensional magnitude spectral vectors from the input speech signal. Next block in the ENROLLMENT phase is "LVD: Get PFZ" and its function is to get the transform matrix PFZ using LVD method from the input feature vectors. Its outputs are the transform matrix PFZ and the original feature vectors. The function of "LVD: Transform" is to transform the original feature vectors to new ones using latent vectors V, according to the equation: $y_n = V^T(x_n - \bar{x})$. At the same time, PFZ is stored in the enrollment database associated with the ID of the new speaker. At last, the new feature vectors obtained after LVD transform are used to obtain the trained Model Parameters (MP) and the trained model is stored in the database associated with the ID of the new speaker too. When to select the resembling candidates of an input voice sample, the processes are as follows: First, general feature extraction is made similar to that of the ENROLL process. Then, the feature vectors are input to "LVD: Transform" component. The Test Parameters (TP) are calculated from the output feature vectors, which are then compared with trained parameters of each speaker enrolled in the enrollment database. The resembling candidates are selected on the basis of highest likelihood of test parameters.

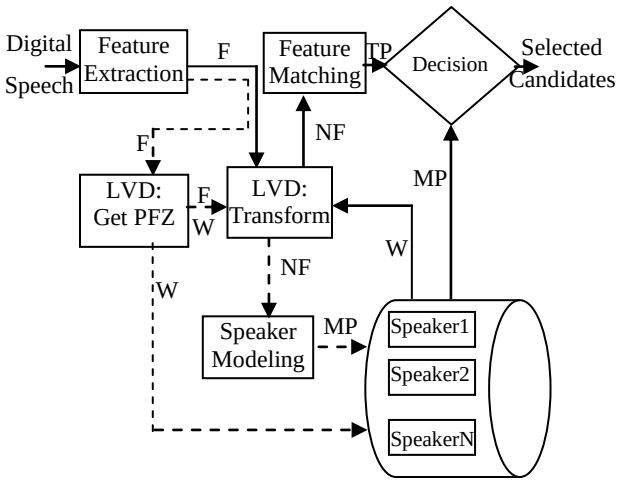
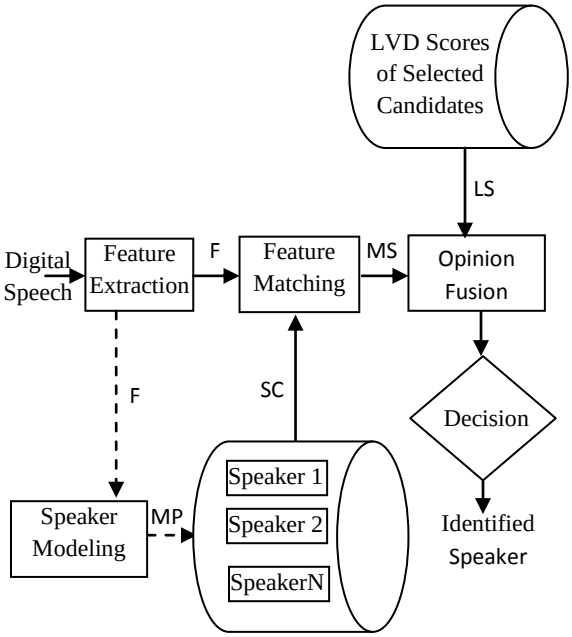


Fig 1: Candidate selection phase using LVD

5.2 Final decision using MFCC or PCA Features

The block diagram of final decision is presented in Figure 2. In this figure, dotted line represents the ENROLLMENT phase and solid line represents the IDENTIFICATION phase. The first block is the feature extraction block which is common for both the ENROLLMENT phase and IDENTIFICATION phase. Feature extraction block is meant for the extraction of 24 dimensional MFCC feature vectors or PCA feature vectors from the input speech signal. The next stage of ENROLLMENT phase is speaker modeling in which GMM classifier is used for training and the trained model is stored in the database associated with the ID of the new speaker too. When to identify the speaker of an input voice sample, the processes followed is as follows: First, general feature extraction is made similar to that of the ENROLL process. Then, the feature vectors are compared with the selected models screened out in the candidate selection process and log likelihood score is calculated. The scores obtained are combined with the scores of candidate selection stage using opinion fusion to take the final decision.



F=Feature

MP=Model Parameter

SC=Selected Candidates

MS=MFCC or PCA scores of selected candidates

LS= LVD scores of selected candidates

Fig 2: Final decision using Opinion Fusion

6. EXPERIMENTAL RESULTS

To investigate the relative performance of the proposed hierarchical method using LVD transform with the standard methods, the Speaker Identification system is applied on an artificially made noisy database prepared by using clean speech of the TIMIT database, i.e. the white Gaussian noise. The experiment is conducted with 200 speakers (100 males and 100 females) selected alphabetically from the TIMIT database. The model for each speaker is trained by clean speech of approximately 24 seconds containing 8 sentences (formed by the concatenation of 2 SA sentences, 3 SI sentences and 3 SX sentences). The remaining two SX sentences contaminated with the white noise used as two independent tests segments. During the candidate selection process 128 dimensional magnitude spectral vectors are transformed into 24 dimensional features by using LVD transform and these features are used to screen out the resembling candidates. Around 20% of speakers are selected in this stage. During the final decision process the system is tested using 32 mixture components per speaker using 24 MFCC coefficients. The scores obtained in both sages are combined in different ration to take the final decision. The results obtained by proposed method are compared with the performance of MFCC and PCA features.

6.1 Opinion Fusion

Decision fusion is an important and effective method to improve identification rates. In this subsection, our aim is to combine the scores of two stages of hierarchical classifier, which is known as opinion fusion [9]. We have used the weighted sum combination rule [10] to combine classifier scores. The weighted sum combination is defined as: $D = \beta d_1 + (1-\beta) d_2$

Where d1 and d2 are the decision scores of each classifier and β is the weighting or combination factor. It can be easily seen that a high value of β provides greater emphasis on classifier 1. For instance, when combining the scores of LDA and MFCC based classifiers, it is clear that β should be assigned a value greater than 0.5 to emphasize the contribution of LDA based classifier.

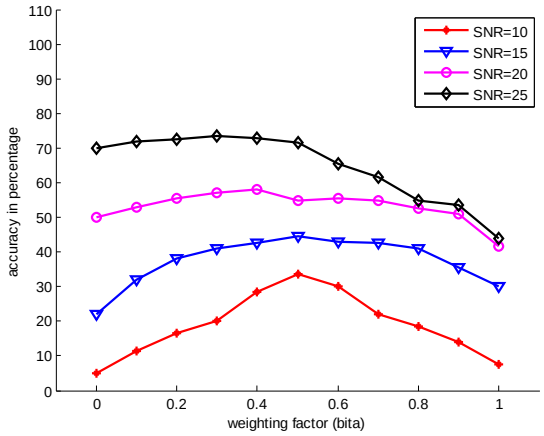


Fig 3: Speaker Identification rate vs weighting factor for MFCC features and LDA features

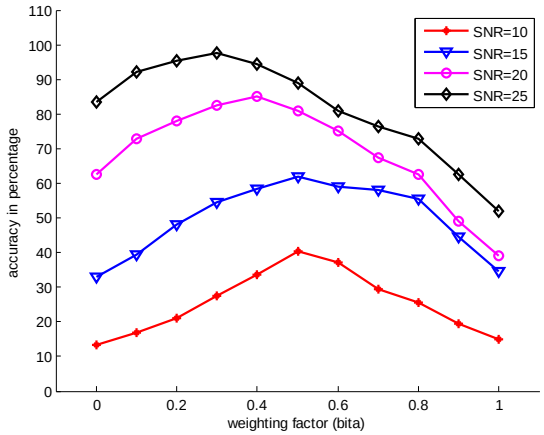


Fig 4: Speaker Identification rate vs weighting factor for PCA features and LDA features

Figure 3 and Figure 4 show the identification rates versus β for the MFCC-LDA and PCA-LDA features, respectively. The value of β is determined such that the identification rate is maximized. Table 1 shows how the optimal β varies with SNR. Therefore the performance of the speaker identification system will be optimal only when we choose different weight factor β at

different SNR. However experimental result shows that a fixed β=0.4 gives a competitive score for all SNR.

Table 1. Optimum value of β for opinion fusion

SNR in db	10	15	20	25	30
Optimum β for MFCC	0.5	0.5	0.4	0.3	0.3
Optimum β for PCA	0.5	0.5	0.4	0.3	0.2

6.2 Performance in Noisy Data

In this experiment the performance of hierarchical classifier is measured in noisy database where 20% of the total number of speakers is selected in the candidate selection stage. Two sets of experiments have been done for each SNR using fixed β in one and optimum β in other. The results obtained by taking different signal to noise ratio are listed in the following Tables 1 and 2.

Table 2. Speaker Identification Rate for fixed weighting factor (β=0.4)

Type of features	SNR in dB				
	10	15	20	25	30
MFCC	3.5	11.5	32.5	63	86.5
PCA	6.5	24.5	73	99	100
LVD+MFCC	28.5	42.5	58	73	86.5
LVD+PCA	33.5	58.5	85	94.5	90

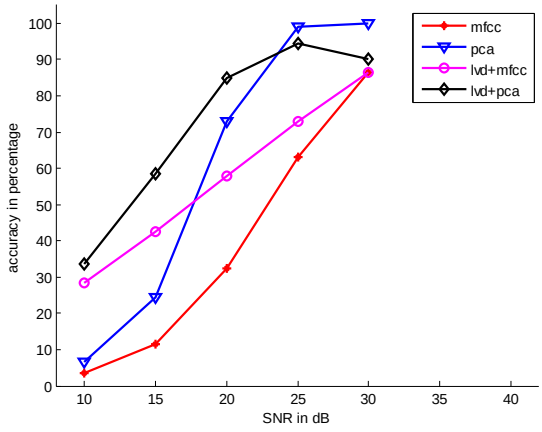


Fig 5: Speaker identification rate for fixed weighting factor (β)

Table 3. Speaker Identification Rate for optimum weighting factor (β)

Type of features	SNR in dB				
	10	15	20	25	30
MFCC	3.5	11.5	32.5	63	86.5
PCA	6.5	24.5	73	99	100
LVD+MFCC	33.5	44.5	58	73.5	79
LVD+PCA	40.5	62	85	97.5	99

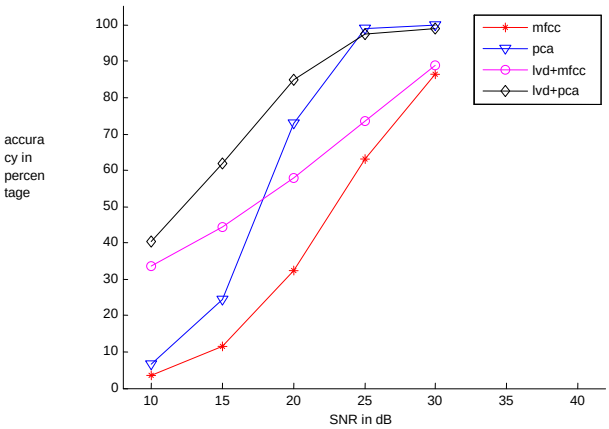


Fig 6: Speaker identification rate for optimum weighting factor (β)

From the above experiments it can be clearly seen the dominant nature of hierarchical classifier at low SNR. At high SNR the performance is slightly degraded using fixed weight. However the result is always better while optimum weight factor is taken for all SNR.

6.3 Performance with Increase Population

Identification accuracy for a population size S is computed by performing speaker identification tests on S speakers randomly selected from 200 speakers and averaging the result over 100 such iterations. This helps in averaging out the bias of a particular population composition.

Figure 7 shows the effect of population size on the speaker identification system in noisy database. We can observe a nearly constant performance of hierarchical classifier across all population where as the performance of hierarchical classifier degraded with increase in population size. Due to the above nature of hierarchical classifier it can be used for large dataset.

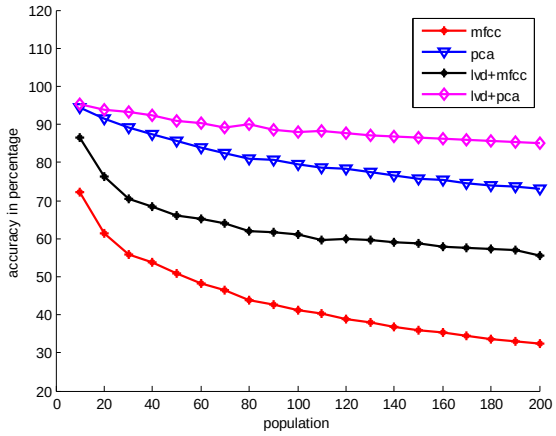


Fig 7: Speaker identification rate with increase population

7. CONCLUSION

In this paper we have designed a hierarchical classifier along with dimension reduction of feature vectors using LVD transformation. After LVD transformation the new feature vectors obtained are frequency independent and hence are immune to noise. The performance of the proposed classifier out-performs traditional classifier at low SNR where as it is competitive at high SNR. Though there are two stages in the proposed classifier, the computational complexity is not high in compared to traditional one as in the second stage only selected candidates are tested. Computational complexity can be further reduced by reducing the dimension of feature vectors in the candidate selection stage. Another significant feature of this hierarchical classifier is its stationary nature against increase population.

8. REFERENCES

[1] Atal, B. S., Automatic recognition of speakers from their voices, Proc. IEEE, Vol. 64, pp. 460-475, 1976.

[2] Reynolds, D. A., An overview of automatic speaker recognition technology, ICASSP, pp. 4072-4075, 2002.

[3] Garofolo, J. S. et al, DARPA TIMIT: Acoustic-Phonetic Continuous Speech Corpus, New Jersey: NIST Publications, 1993.

[4] Wei Han, Cheong–Fat Chan, Chiu-Sing Choy, Kong-Pang Pun, An Efficient MFCC Extraction Method in Speech Recognition, IEEE ISCAS, pp-4, September 2006.

[5] Reynolds, D. A., and Rose, R.C., Robust Text-Independent Speaker Identification using Gaussian Mixture Speaker Models, IEEE Trans. Speech and Audio Processing, 3(1):72-83, 1995.

[6] Reynolds, D. A., Experimental evaluation of features for robust speaker identification, IEEE Trans. Speech Audio Processing, Vol. SAP-2, No. 4, pp. 639-643, 1992.

[7] Hoffman, T., Unsupervised learning by probabilistic latent semantic analysis, Machine Learning, vol. 42, pp. 177-196, 2001.

[8] Bhiksha Raj, Paris Smaragdis, Latent Variable Decomposition of Spectrograms for Single Channel Speaker Separation, 2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp.17 - 20, 16-19 Oct 2005.

[9] Kittler, J., Hatef, M., Duin, P.W., and Matas, J., On Combining Classifiers, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 20(3), 226-239, 1998.

[10] Switzerl, M.V., Conrad, S., and Paliwal, K.K., Information Fusion and Person Verification Using Speech and Face Information, IDIAP Research Report, pp. 1-37, 2002.