Pattern Recognition 44 (2011) 2749–2759

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Text-independent speaker identification using Radon and discrete cosine transforms based features from speech spectrogram

Pawan K. Ajmera [a,*], Dattatray V. Jadhav [b], Raghunath S. Holambe [a]

[a] S.G.G.S. Institute of Engineering and Technology, Vishnupuri, Nanded, India
[b] Bhivarabai Sawant College of Engineering and Research, Pune, India

ARTICLE INFO

Article history:
Received 14 May 2010
Received in revised form
5 February 2011
Accepted 12 April 2011
Available online 22 April 2011

Keywords:
Speaker recognition
Spectrogram
Feature extraction
Radon transform
Discrete cosine transform

ABSTRACT

This paper presents a new feature extraction technique for speaker recognition using Radon transform (RT) and discrete cosine transform (DCT). The spectrogram is compact, efficient in representation and carries information about acoustic features in the form of pattern. In the proposed method, speaker specific features have been extracted by applying image processing techniques to the pattern available in the spectrogram. Radon transform has been used to derive the effective acoustic features from the speech spectrogram. Radon transform adds up the pixel values in the given image along a straight line in a particular direction and at a specific displacement. The proposed technique computes Radon projections for seven orientations and captures the acoustic characteristics of the spectrogram. DCT applied on Radon projections yields low dimensional feature vector. The technique is computationally efficient, text-independent, robust to session variations and insensitive to additive noise. The performance of the proposed algorithm has been evaluated using the Texas Instruments and Massachusetts Institute of Technology (TIMIT) and our own created Shri Guru Gobind Singhji (SGGS) databases. The recognition rate of the proposed algorithm on TIMIT database (consisting of 630 speakers) is 96.69% and for SGGS database (consisting of 151 speakers) is 98.41%. These results highlight the superiority of the proposed method over some of the existing algorithms.

1. Introduction

In recent years, biometric-based authentication systems have been widely used in many applications. Various human characteristics such as the face, speech, fingerprint, iris, etc. have been considered as discriminative features for automatic biometric recognition [1,2]. A biometric system is essentially a pattern-recognition system that recognizes a person based on a feature vector derived from a specific physiological or behavioral characteristic the person possesses. Multimodal biometric systems, which combine multiple biometric samples, or characteristics derived from samples, are developed in order to meet requirement of real-world applications [1,3]. Voice is the most natural and economical biometric modality for person identification.

Speaker recognition is to recognize persons from their voice. No two individuals sound identical because their vocal tract shapes, larynx sizes, other parts of their voice production organs, manner of speaking including the use of a particular accent, rhythm, intonation style, pronunciation pattern and choice of vocabulary are different. The speech features are broadly categorized into short-term spectral features, voice source feature, spectro-temporal features, prosodic features and high-level features. Short-term spectral features are computed from short frames of about 20–30 ms duration. They are usually descriptors of the short-term spectral envelop, which contain resonance properties of the vocal tract. The voice source features characterize the glottal flow. Intonation and rhythm are part of prosodic and spectro-temporal features. Conversational-level characteristics of speaker are captured in high-level features [4]. State-of-the-art speaker recognition systems use number of these features in parallel, attempting to cover different aspects and employing them in complementary ways to achieve more accurate recognition.

One of the biggest challenges in automatic speaker recognition is invariance across varying operating conditions (different microphone, transmission coding and background noise). Number of approaches has been proposed for tackling the invariance problem. These include robust feature extraction [5], feature normalization [6], model transformation [7,8] and match score normalization [9,10]. Many text-independent speaker recognition approaches use cepstral mean subtraction (CMS) in the context of cepstral features. In CMS all the feature vectors have been translated by subtracting the means from both the training and

* Corresponding author. Tel.: +91 9421820253; fax: +91 20 24280926.
E-mail addresses: ajmera.pawan@gmail.com (P.K. Ajmera), dvjadhao@yahoo.com (D.V. Jadhav), rsholambe@sggs.ac.in (R.S. Holambe).

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.04.009

the testing vectors. This improves recognition performance in channel mismatch condition; however, the recognition performance gets degraded for clean data (no channel mismatch) [4].

In speaker identification system, high dimension feature set is preferred to enhance the performance. However, increased feature dimension requires more computational time and storage space [11]. The classifier using high dimension feature set also requires more parameters to characterize a speaker model, e.g. Gaussian Mixture Model (GMM) [12]. This increases computational complexity, making real-time implementation more difficult. Furthermore, a large amount of data is required for the training. An alternative approach to this is to extract effective and efficient feature vectors.

Mel frequency cepstral coefficients (MFCC) [13] and linear prediction cepstral coefficients (LPCC) [5] are the two most common feature extraction techniques in speaker identification. MFCC is generally used because of its robustness in speaker identification [5,14]. Since the elements of feature vectors are generally correlated [15–17], a large number of mixtures with full covariance matrix are necessary to provide good approximation [18]. The GMM with diagonal covariance matrix is used for both speaker identification and verification [19] because of its computational simplicity.

Contextual variations in speech are better represented using a spectrogram and hence it is widely used as a tool for speech analysis [20]. A spectrogram is a graphical display of the squared magnitude of the time-varying spectral characteristics of speech [21]. It is compact and efficient in representation carrying information about energy, pitch, fundamental frequency, formants and timing. Spectrogram reading techniques have revealed that a speech spectrogram contains rich acoustic features that could be valuable in an automatic speech and speaker recognition system [22]. The choice of spectrogram reading stems from the assumption that since spectrograms may be read with 93% accuracy in a speaker-independent mode [23], mostly all the information on the acoustic phonetic level that is necessary to decode the message is present on the spectrograms. These parameters are the acoustic features of speech used in automatic stress and emotion recognition systems [24–26]. The most of these systems analyze each parameter separately and then combine them into a set of feature vectors. All of these characteristics can be captured by analyzing the spectrogram. Spectrogram preserves the important underlying dependencies between different parameters. Spectrogram of the word "zero" is shown in Fig. 1(a). Speech features caused by specific psycho-physiological effects appearing under certain emotions and stress are also captured from the spectrogram [20,23].

In [27,28] two-dimensional (2-D) Gabor filter bank has been applied on mel-spectrogram and the resulting outputs of the Gabor filters were concatenated into one-dimensional (1-D) vector as feature in speech recognition experiments. A similar approach for speech discrimination and enhancement is presented in [29]. Speech harmonicity, formants, vertical onsets/offsets, noise and overlapping simultaneous speakers have been represented using 2-D Gabor filter bank on the spectrograms[30–32].

It has been established that the phonetic information can be recovered by examining the spectrogram [20,22,23] in a visual domain rather than the conventional audio domain. Visual domain working is better because it is easier to "verbalize" speech spectrogram process than to verbalize hearing process. Speech spectrogram readers interpret spectrograms based on prior knowledge. This includes pictorial scene changes, acoustic knowledge, phonetic knowledge to these scene changes, and by associating prior linguistic knowledge to such scene changes.

This paper presents a computationally efficient text-independent speaker recognition technique. The essence of this technique lies in formulating the speaker recognition problem into pattern recognition of images and resolving it using machine learning tools [33]. The technique computes the Radon projections of the speech spectrogram in different directions to derive the speaker's voice pattern. The Radon projection of spectrogram of the word "zero" for projection at an angle of 90° is shown in Fig. 1(b). Discrete cosine transform (DCT) of Radon projection reduces the feature vector dimension to derive effective and efficient speaker

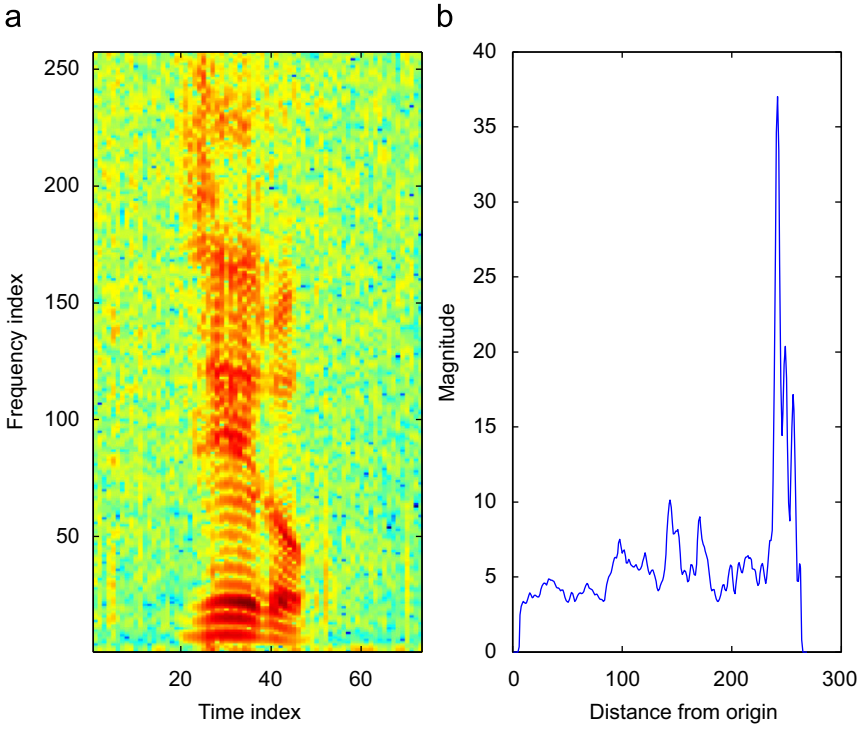

Fig. 1. (a) Spectrogram of word "zero" and (b) Radon projection of spectrogram at an angle of 90°.

features. The technique is computationally efficient, text-independent, robust to session variations and insensitive to additive noise.

The paper is organized as follows. The proposed method along with preprocessing, Radon transform and discrete cosine transform is described in Section 2. The databases used and performance evaluation are presented in Section 3. Conclusions based on the experimental results are presented in Section 4.

2. Proposed method

In pattern recognition the extracted features should have minimum intra-class variance, which means that features derived from different samples of the same class should be close while the inter-class separation should be large, i.e., features derived from samples of different classes should differ significantly. Further, features should be independent of the size, orientation and location of the pattern.

Fig. 2 shows the proposed feature extraction technique. The steps in the first column are for the training process, while second column illustrates steps for the recognition process. The preprocessing block transforms the raw speech signal into a speech spectrogram. As the spectrogram is the squared magnitude of the time-varying spectral characteristics of speech, the Fourier transform eliminates the circular shift effect in the feature domain by taking the spectrum magnitude of the Fourier coefficients. The technique uses seven Radon projections of the spectrogram in different orientations. Significant DCT coefficients (30%) of Radon projections are concatenated to derive the speech feature vector. Due to excellent data compaction property of DCT, its use in the proposed approach helps in reduction of the feature vector dimension. In the recognition process, the feature vector extracted from the utterance of an unknown person is compared against the feature vectors in the database on the basis of Euclidean distance using the nearest neighbor classifier to make the final decision.

2.1. Speech spectrogram

The input speech waveform $x(n)$ is pre-emphasized by a first order filter with transfer function

$$H(z) = 1 - az^{-1}, \quad 0.9 \le a \le 1. \tag{1}$$

The most common value for a is around 0.95 as it boosts the magnitude by 32 dB [34]. Pre-emphasis boosts the higher frequencies whose intensity would otherwise be very low due to a downward sloping spectrum caused by glottal voice source [4,34]. The process of pre-emphasis flattens the signal and makes it less susceptible to finite precision.

It is well known that the speech signal is non-stationary in nature. However, it is assumed that the speech signal remains stationary over a short duration of 20–30 ms. Hence, the pre-emphasized signal $y(n)$ is segmented into M frames of 20 ms duration with a 10 ms overlap between two consecutive frames to retain a good quality of the signal and to avoid loss of information [4,34].

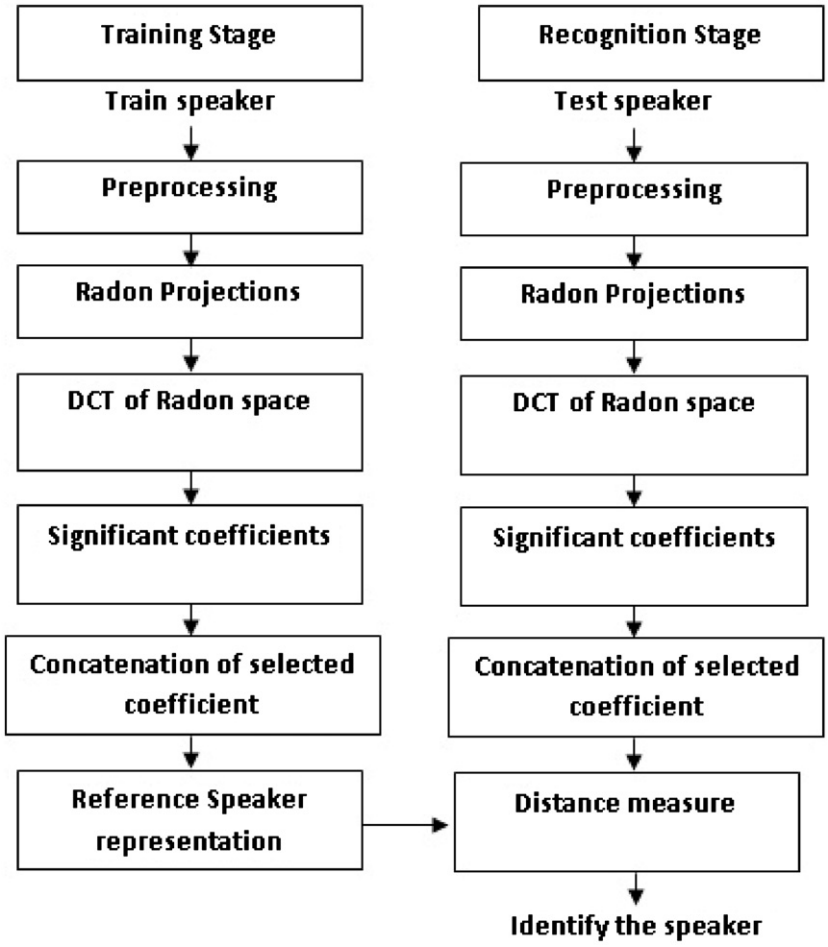

Fig. 2. Block diagram of the proposed speaker recognition algorithm.

Windowing is carried out to reduce the edge effects at the beginning and the end of the frame. In our study Hamming window is multiplied with each frame. Thus,

$$\tilde{x}_i(n) = y_i(n)w(n), \quad i = 1,2,\ldots,M \tag{2}$$

where

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{T-1}\right). \tag{3}$$

T is the number of samples in each frame.

Fourier transform of each frame is computed to produce an estimate of the short-term frequency content of the signal, called as spectrogram. The spectrogram is the squared magnitude of the time-dependent Fourier transform versus time. N length DFT of a windowed frame is computed to obtain the power spectrum as below. In our study, we have used $N=512$.

$$S_i(k) = (\mathrm{Re}\{\tilde{X}_i(k)\})^2 + (\mathrm{Im}\{\tilde{X}_i(k)\})^2, \quad k = 0,1,\ldots,N-1, \text{ and } i = 1,2,\ldots,M, \tag{4}$$

where $\tilde{X}_i(k)$ is the kth component of DFT of $\tilde{x}_i(n)$ (windowed frame). Re {.} and Im {.} indicate real and imaginary parts, respectively. The spectrums of these frames $S_i(k)$ are concatenated to construct the speech spectrogram $f(k,i)$.

$$f(k,i) = \begin{bmatrix} S_1(0) & \ldots & S_M(0) \\ . & . & . \\ . & . & . \\ S_1(N-1) & \ldots & S_M(N-1) \end{bmatrix}. \tag{5}$$

In the proposed approach, the speech spectrogram is treated as an image. Contextual variations in speech images are similar to real-world changes in scene analysis. These variations can be captured by applying image processing techniques to these patterns. The existing speaker recognition systems tend to look for specific features and take action based on the presence or absence of the expected features. It is very likely that a "misplaced" feature means an absence of that particular feature. However, the spectrogram preserves all significant and contextual features.

2.2. Radon transform

Radon transform is based on the parameterization of lines and the evaluation of integrals of an image along these lines. Due to inherent properties of Radon transform, it is a useful tool to capture the directional features of an image. Basically, the Radon transform adds up the pixel intensity values in the given image (spectrogram) or time frequency distribution along a straight line in a particular direction at a specific displacement [35]. The Radon transform of 2-D signal $f(x,y)$ is defined as

$$R(r,\theta) = \Re[f(x,y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\delta(r - x\cos\theta - y\sin\theta)dx\,dy, \tag{6}$$

where r is the distance of a line from the origin, $\theta \in [0,\pi]$ is the angle between the distance vector and x-axis [36] and $\delta(\cdot)$ is the Dirac function. The symbol $\Re$ denotes the Radon transform operator.

The spectrogram represents acoustic features like energy, pitch, fundamental frequency, formants and time in the form of a pattern [21,24–26]. The Radon transform effectively captures these features in the pattern by projecting it onto different orientation slices. The Radon projection is obtained by summing all the intensity values of those pixels that are within the circle surrounding the pattern to be recognized and on the line that is perpendicular to the ridge. For a given ridge, every pixel within the circle will be projected onto it along the perpendicular direction. This gives a rise to one Radon slice in the Radon domain. The proposed technique computes Radon projections of the spectrogram in different orientations. In each projection, the variations of the pixel intensities are preserved evenly, though the pixels are far from the origin. Therefore, the method does not measure intensity variations based on the location in the image. Since the Radon transform is linear by definition, geometric properties like straight lines or curves can be made explicitly by the Radon transform which concentrates on energies (loci of intersection of several sinusoidal curves) from the image in few high-valued coefficients in the transformed domain.

Another advantage of using Radon transform in the proposed approach is its insensitivity to additive noise. Let us consider that the noise corrupted image $f(x,y)$ is observed as

$$\hat{f}(x,y) = f(x,y) + \eta(x,y), \tag{7}$$

where, $\eta(x,y)$ is zero mean white noise. The Radon transform of $\hat{f}(x,y)$ is

$$\Re[\hat{f}(x,y)] = \Re[f(x,y)] + \Re[\eta(x,y)]. \tag{8}$$

Since the Radon transform is line integrals of the image, for the continuous case, the Radon transform of white noise is constant for all of the points and directions and is equal to the mean value of the noise (if integrated over infinite axis), which is assumed to be zero [36,37]. In our case, the signal (spectrogram—digital image) is discontinuous and composed of a finite number of pixels. Hence, $\Re[\eta(x,y)]$ in Eq. (8) does not become zero. However, as derived in [36] the relation between signal-to-noise ratio (SNR) of Radon projection (SNR_{proj}) and SNR of image (SNR_{image}) is

$$SNR_{proj} = SNR_{image} + 1.7N\,SNR_{image}. \tag{9}$$

In practice, N, which is the radius in pixels, is sufficiently large; hence SNR_{proj} is much higher than SNR_{image}. This property of Radon transform makes the proposed algorithm insensitive to zero mean white noise.

Fig. 3(a)–(c) shows the speech waveforms of the same sentence uttered by three different speakers (inter-class) while Fig. 3(d)–(f) shows their corresponding spectrograms. Fig. 3(g)–(i) shows Radon projections of the respective spectrograms at an angle of 90°. All three curves represent the same sentence but are uttered by three different speakers. These curves have different shapes as the speech pattern is produced by excitation of the vocal tract with quasi-periodic pulses of air caused by the vibration of vocal cords. As vocal tract shape, larynx size and other parts of voice producing organs are different for each speaker, their corresponding spectrograms will also differ (despite utterance of the same sentence). As Radon projections of the spectrograms of different speakers yield significant variations in shape (irrespective of the text uttered), inter-class variations are maximized.

Fig. 4(a)–(c) shows the speech waveforms of different sentences uttered by the same speaker (intra-class) at different time; Fig. 4(d)–(f) shows their corresponding spectrograms; Fig. 4(g)–(i) shows the Radon projections of respective spectrograms at an angle of 90°. As acoustic characteristics in the spectrogram remain almost unchanged for the same vocal tract configuration (irrespective of the text uttered), intra-class variations are minimized.

Fig. 5(a)–(c) shows the error signals between Radon projections of the same sentence uttered by three different speakers (inter-class) as shown in Fig. 3(g) and (h), (h) and (i), (i) and (g), respectively. The mean square errors (MSE) of these error signals are 12.602, 17.217 and 15.158, respectively. Fig. 6(a)–(c) shows the error signals between Radon projections of different sentences uttered by the same speaker (intra-class) at different time as shown in Fig. 4(g) and (h), (h) and (i), (i) and (g). The MSE of these error signals are 5.178, 5.842 and 6.052.

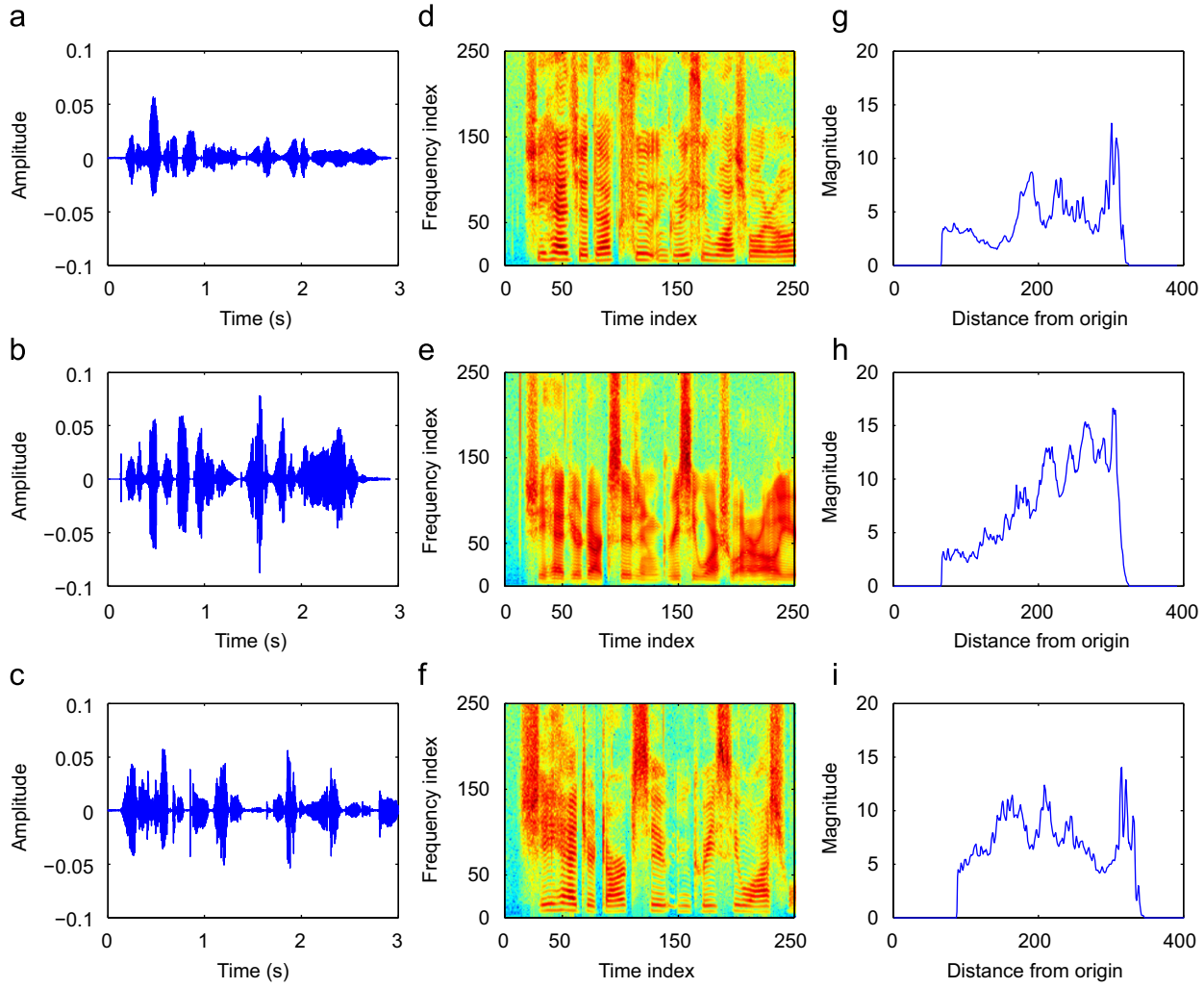

Fig. 3. Typical speech waveforms, spectrograms and Radon projections of the same sentence uttered by three different speakers: (a)–(c) speech waveforms; (d)–(f) corresponding spectrogram and (g)–(i) Radon projections of respective spectrograms at an angle of 90°.

Fig. 7 shows root mean square error (RMSE) of the curves obtained from Radon projections of the spectrograms for inter-class variations and intra-class variations. RMSE takes into account both the bias and the variance. The curves show that the proposed approach minimizes intra-class variations and maximizes inter-class variations. To compute RMSE, one reference speaker and five other speakers are selected. The RMSE curves are obtained as follows:

$$RMSE = \left[\frac{1}{N}\sum_{i=1}^{N}\{r_i(l)-\overline{R}(l)\}^2\right]^{1/2}, \quad l=0,1,\ldots,L-1. \tag{10}$$

where $\overline{R}(l)=(1/N)\sum_{i=1}^{N}R_i(l)$, $l=0,1,\ldots,L-1$.

$R_i(l)$ is Radon projection of the spectrogram of ith speech sample of a reference speaker taken from the training set, and $r_i(l)$ is Radon projection of the spectrogram of ith speech sample of a speaker taken from the testing set. To compute intra-class RMSE curve, $r_i(l)$ is used from the testing set of the reference speaker uttering five different sentences and for inter-class RMSE curve, $r_i(l)$ is used from the samples of five different speakers uttering the same sentence. N is the number of speech samples of speaker (we used $N=5$), L is the distance from the origin of the Radon projection.

The small values of MSE and RMSE for intra-class speakers as compared with those of inter-class speakers indicate that the proposed algorithm maximizes the inter-class variations and minimizes the intra-class variations.

2.3. Discrete cosine transform

DCT is a well-known signal analysis tool used in data compression due to its compact representation capability. It has an excellent energy compaction property for highly correlated data. This helps in reduction of the feature vector dimension.

Let $f(x,y)$ be represented as a discrete image $f(m,n)$ of size $M\times N$. The 2-D DCT of an image $f(m,n)$ is given as

$$B_{pq}=\alpha_p\alpha_q\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}f(m,n)\cos\frac{\pi(2m+1)p}{2M}\cos\frac{\pi(2n+1)q}{2N}, \tag{11}$$

$$0\le p\le M-1,\ 0\le q\le N-1$$

$$\alpha_p=\frac{1}{\sqrt{M}},\quad p=0=\sqrt{2/M},\quad 1\le p\le M-1,$$

$$\alpha_q=\frac{1}{\sqrt{N}},\quad q=0=\sqrt{2/N},\quad 1\le q\le N-1,$$

where α_p and α_q are normalization factors; p and q denote the frequencies; $M$and N denote the size of row and column of $f(m,n)$, respectively [38].

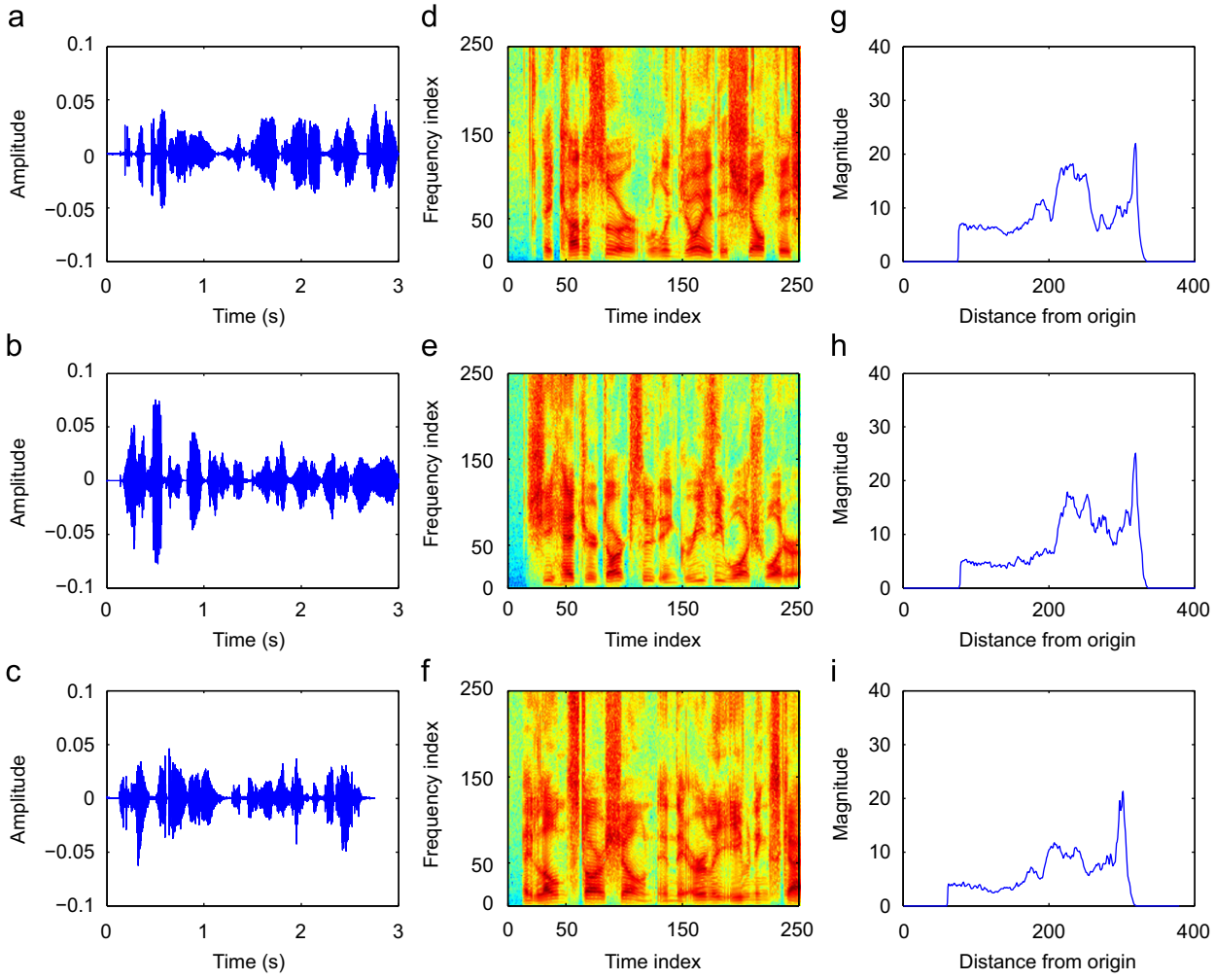

Fig. 4. Typical speech waveforms, spectrograms and Radon projections of different sentences uttered by the same speaker at different time: (a)–(c) speech waveforms; (d)–(f) corresponding spectrograms and (g)–(i) Radon projections of respective spectrograms at an angle of 90°.

3. Performance evaluation

In this section, the proposed algorithm has been analyzed and evaluated by performing various experiments on the speaker recognition task using the Texas Instruments and Massachusetts Institute of Technology (TIMIT) and our own created Shri Guru Gobind Singhji (SGGS) databases. A comparative performance has been carried out against some of the existing speaker recognition schemes, which include MFCC [13], spectrogram modulation [4] and temporal discrete cosine transform (TDCT) [39] techniques. The recognition rate in all the experiments are computed for the correct number of matches out of the total speakers used for testing. recognition rate is defined as follows:

$$\text{Recognition Rate} = \frac{\textit{Number of correct matches}}{\textit{Total number of test speaker}} \times 100\%.$$

3.1. Speaker databases

The speaker recognition experiments presented in this paper have been performed using the TIMIT and SGGS linguistic databases. The TIMIT corpora have been used because of large number of speakers in the database. This corpus contains 6300 sentences, 10 sentences spoken by each of 630 (438 male and 192 female) speakers from eight major dialect regions of the United States. The text material consist of two dialect sentences (the SA sentences) designed at Stanford Research Institute, 450 phonetically compact sentences (the SX sentences) designed at Massachusetts Institute of Technology, and 1890 phonetically diverse sentences (the SI sentences) selected at Texas Instruments. The SA sentences were meant to expose the dialectal variants of speakers and were read by all 630 speakers. The SX sentences were designed to provide a good coverage of pairs of phones, with extra occurrences of photonic contexts thought to be either difficult or of particular interest. Each speaker reads five of these sentences and each text was spoken by seven different speakers. The SI sentences were from Playwright dialog so as to add a diversity in sentence type and phonetic context. Each speaker reads three of these sentences, with each sentence being read by a single speaker. The speech signal is recorded through a high quality microphone with a sampling frequency of 16 kHz in quiet environment. Ten sentences of each speaker consist of two SA sentences, five SX sentences and three SI sentences. Each sentence is of approximately 3 s duration.

Our own created English language SGGS database contains 1510 sentences, 10 sentences spoken by each of 151 (100 males and 51 females in the age group of 16–50 years) speakers from five major dialect regions of Maharashtra State, India. Out of 10 speech files per speaker, five files consist of utterances recorded by subjects reading a prepared text and the other five files consist of dialog prompted on their own from a set of hundred sentences

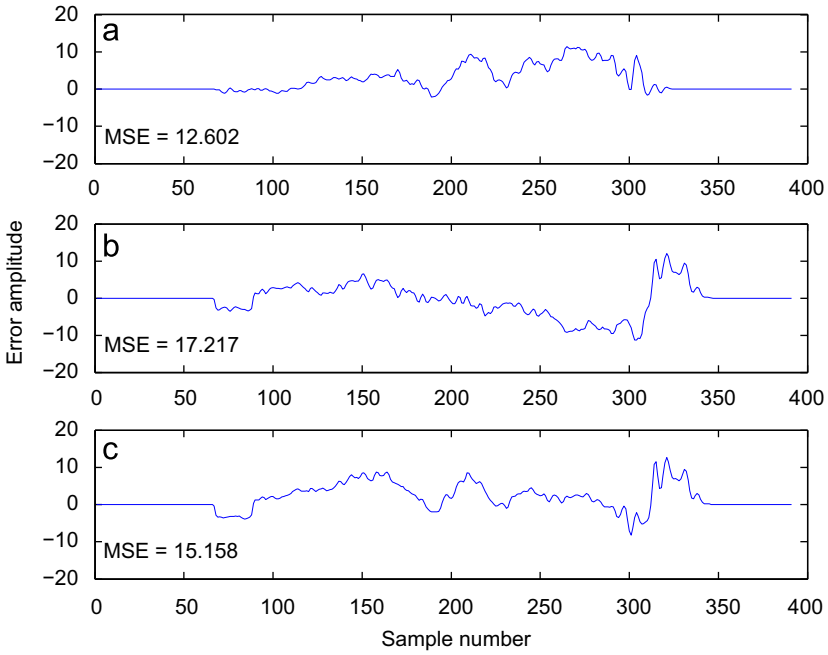

Fig. 5. Inter-class error signals of Radon projections shown in Fig. 3: (a) error signal between (g and h); (b) error signal between (h and i) and (c) error signal between (g and i).

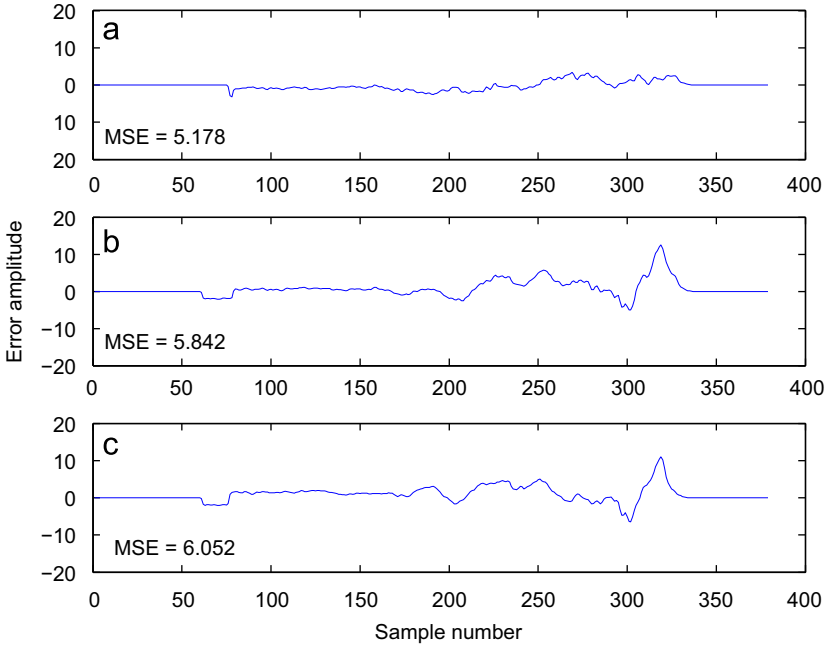

Fig. 6. Intra-class error signals of Radon projections shown in Fig. 4: (a) error signal between (g and h); (b) error signal between (h and i) and (c) error signal between (g and i).

in order to add a diversity in sentence type, pitch and rhythm. Each sentence is of approximately 3 s duration and has been recorded in a laboratory using the software 'Sound Forge Version 5.0' at a sampling frequency of 22,050 Hz.

Our own created English language SGGS-2 database consists of 360 sentences, 10 sentences spoken by each of 36 (all males) speakers obtained in two different sessions within a span of 2 months. Out of 10 speech files per speaker, five speech files were recorded in a silent environment using a condenser microphone and the other five speech files were recorded in the second session that was carried out in a classroom with a head-mounted microphone connected to a compact flash recorder. Since the recording time and medium were different for the utterance samples, session and channel variations have been included in the database. These speech sentences were recorded from subjects reading a prepared text. Each sentence is of approximately 3 s duration.

3.2. *Effect of number of Radon projections on the recognition rate*

In the first set of experiment, the performance of the proposed algorithm was evaluated using 100 speakers in TIMIT and 50 speakers in SGGS databases. Five sentences of each speaker were randomly selected for training and the remaining five sentences were used for testing from both databases. The recognition rates were computed for different number of Radon projections for angles between 0° and 180°. Fig. 8 shows the recognition rate for different number of

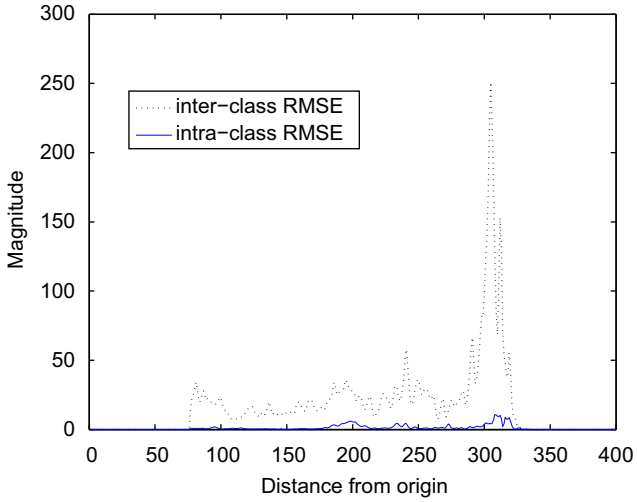

Fig. 7. Inter-class and intra-class RMSE curves.

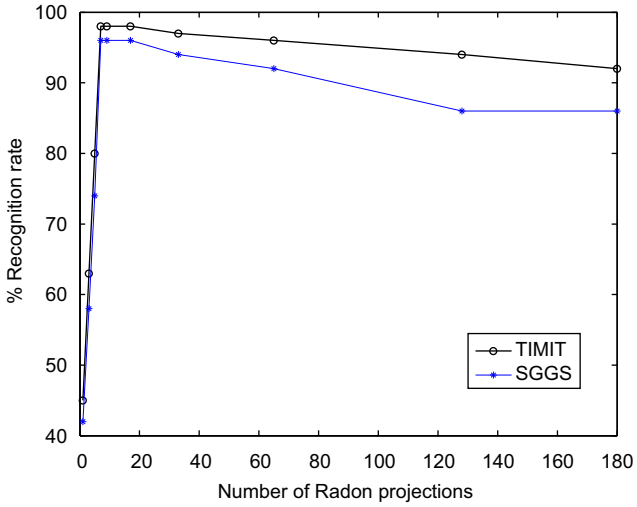

Fig. 8. Effect of number of Radon projections on the recognition performance.

Radon projections. It has been observed that the recognition rate increases with an increase in the number of projections (up to seven) and remains almost constant for any increase in the number of projections thereafter. This is because of an addition of the redundant information by the increased number of Radon projections. Hence in the subsequent experiments only seven Radon projections have been used [22.5°, 45°, 67.5°, 90°, 112.5°, 135°, 157.5°].

3.3. Effect of number of DCT coefficients on the recognition rate

The performance of the proposed approach was investigated for variation in the number of DCT coefficients selected in the feature vector. The experiment was performed using the same set of databases used for computing the number of Radon projections (Section 3.2). Fig. 9 shows the recognition rates of this approach for different percentages of DCT coefficients used. It reveals that there is a significant improvement in the recognition rate as the use of DCT coefficients increases up to 30%. Any further increase in the coefficients does not improve the performance significantly. Hence we have selected 30% coefficients of DCT as significant coefficients in all subsequent experiments. The results of the proposed approach presented in the following sections are for seven Radon projections and 30% DCT coefficients. In all the experiments the nearest neighbor classifier has been used.

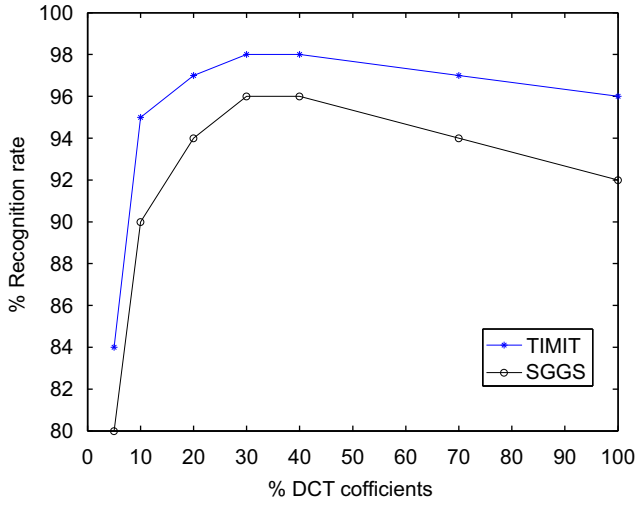

Fig. 9. Percentage of DCT coefficients used versus recognition rates (seven Radon) projections.

3.4. Comparative performance

In this experiment, the performance of the proposed approach has been evaluated with seven Radon projections and 30% DCT coefficients. The database used in the experiment consists of 630 speakers in TIMIT, 151 speakers in SGGS and 36 speakers in SGGS-2 databases with 10 sentences each. Five different combinations per speaker for training as well as for testing were used. These sentences (approximately 5×3 s$=15$ s duration) were selected randomly. Training and testing data in each set were not overlapping. The performance of the proposed approach was compared with MFCC [13], spectrogram modulation [4] and temporal discrete cosine transform (TDCT) [39]. The average recognition rate, standard deviation and 95% confidence interval level using five combinations for all the above algorithms were computed.

Twelve MFCCs from a 30-channel mel-frequency filter bank were computed. The MFCC trajectories were smoothed with relative spectral (RASTA) filtering [3] followed by first order (Δ) and second order ($\Delta\Delta$) time derivative feature computation. This was followed by voice activity detection and utterance-level mean and variance normalization. In modulation spectrogram [37], a frame length of 30 ms with a 7.5 ms overlap, Hamming window and pre-emphasis filter $H(z)=1-0.97z^{-1}$, FFT of length 256, 30 mel filters and three DCT coefficients were used. TDCT features were derived from MFCC. In this method, each cepstral coefficient was considered as an independent "signal" which is windowed in blocks of length 8. Each block is transformed into DCT coefficients. The lowest three DCT coefficients of all MFCCs were stacked to form a TDCT feature vector. Assuming that the original MFCC frame length is of L ms and frame advance (overlap) is S ms, single Δ vector spans over $S+L$ ms($L=30$ ms, $S=20$ ms). Similarly, a single $\Delta\Delta$ vector spans over $2S+L$ ms. Thus, a typical MFCC$+\Delta+\Delta\Delta$ vector contains information over an interval of 70 ms. GMM using expectation maximization (EM) [19] with Bayes' classifier was used to compute the recognition rate for MFCC, modulation spectrogram and TDCT. For GMM, $M=32$ has been used as the Gaussian mixture density. We have selected 32 feature vectors randomly and used them as initial mean vectors of GMM to derive 32 diagonal covariance matrices as initial guess. This was followed by the EM algorithm which yields maximum probability from the utterance used for training.

In the proposed approach the speech signal is of 3 s duration, which is divided into frames of 20 ms with a 10 ms overlap. FFTs of the frames with length 512 are concatenated to produce a

spectrogram of size $N \times M$, where N is the positive side length of FFT (256) and M is the total number of frames (256). Total number of frames varies according to the length of the speech signal (262–299). However, to generate a square image for computation of Radon projections, the number of frames has been restricted to 256. This spectrogram is projected using seven Radon projections in particular orientations having angles [22.5°, 45°, 67.5°, 90°, 112.5°, 135°, 157.5°] to derive feature.

Tables 1–3 show the results of these experiments using the TIMIT, SGGS and SGGS-2 databases, respectively. The proposed algorithm yields the better result on all the databases. This is because of boosting of low frequency components, which are useful in recognition as the speaker specific features are in low frequency region and removal of redundant information using DCT. Standard deviation is the most useful criterion for evaluating the effect of selection of training samples on the recognition performance; lower the standard deviation, lesser is the effect of choosing different training sample sets. The results show that the proposed algorithm yields maximum recognition rate with minimum standard deviation. Thus the proposed approach provides effective and efficient features. The narrow width of the confidence interval of the proposed algorithm implies that the performance of the proposed algorithm is less affected by changes in the training sets. The results presented in Table 3 on SGGS-2 database reveal that the proposed approach is invariant to session and channel variations.

Table 1
Average recognition rate, standard deviation and confidence interval for the different approaches (TIMIT database-5 sets, 630 speakers).

Algorithm	Average recognition rate	Standard deviation	Confidence interval	
MFCC	95.8413	0.3790	95.4127	96.2699
Modulation spectrogram	88.7302	0.4628	88.2068	89.2536
TDCT	90.5397	0.3654	90.1264	90.9530
Proposed approach	96.6984	0.2070	96.4643	96.9325

Table 2
Average recognition rate, standard deviation and confidence interval for the different approaches (SGGS database-5 sets, 151 speakers).

Algorithm	Average recognition rate	Standard deviation	Confidence interval	
MFCC	98.2781	0.5923	97.6082	98.9481
Modulation spectrogram	89.9338	0.8635	88.9572	90.9104
TDCT	95.0993	0.7551	94.2453	95.9533
Proposed approach	98.4106	0.3627	98.0003	98.8208

Table 3
Average recognition rate, standard deviation and confidence interval for the different approaches (SGGS-2 database-5 sets, 36 speakers).

Algorithm	Average recognition rate	Standard deviation	Confidence interval	
MFCC	95.5556	1.5215	93.8348	97.2763
Modulation spectrogram	83.8889	2.3241	81.2604	86.5174
TDCT	90.5556	3.7268	86.3406	94.7705
Proposed approach	97.7778	1.2423	96.3728	99.1828

Table 4
Effect of white noise on the performance of speaker recognition algorithms (168 speakers TIMIT database).

Algorithm	% Recognition rate with added noise of signal to noise level in the test utterance			
	No noise	15 dB	10 dB	5 dB
MFCC	98.8	77.8	54.1	16.6
Modulation spectrogram	88.0	70.2	33.3	10.1
TDCT	91.0	73.8	41.0	13.1
Proposed approach	98.8	92.2	81.9	48.6

3.5. Noise robustness

The insensitiveness of the proposed approach to additive noise has been tested using the TIMIT test speaker set of 168 speakers. The test speaker set was obtained by including the sentences from all speakers that read any of the SX sentences. White noise with different variances was added electronically to each test utterance to get signal to noise ratio (SNR) between 5–15 dB (increasing 5 dB every step). No noise was added during the training phase. The performance of different algorithms on the noisy speeches is shown in Table 4. The results show that the proposed algorithm is almost insensitive to additive noise. This is because of the capability of Radon transform [36]. Fig. 10(a)–(c) shows the speech signals for 'no noise', 15 dB and 5 dB SNR; Fig. 10(d)–(f) shows the corresponding speech spectrograms; Fig. 10(g)–(i) shows the Radon projections of respective spectrograms at an angle of 90°; and Fig. 10(j)–(l) shows the DCT coefficients. It is observed that with the addition of noise, the spectrogram changes. The high frequency noise components in the spectrogram are averaged out in the Radon projection because of line integrals (low-pass filter). Hence DCT coefficients of noisy signal and the performance of the algorithm are not affected by the noise. However, performance is affected if signal-to-noise ratio drops below 10 dB.

3.6. Computational complexity

All the experiments have been performed using a Pentium-IV PC, with CPU speed of 2.0 GHz, 1 GB RAM and MATLAB 7.0. The time taken by each algorithm for training as well as testing 100 speakers in TIMIT database was computed and is given in Table 5. The proposed method yields the feature vector of a very small dimension, which results in reduced computational complexity (less training and testing time) with a high recognition rate making the proposed approach suitable for real time applications.

4. Conclusions

In this paper, we have proposed a novel speaker recognition technique, which provides effective and efficient features using a combination of Radon transform and DCT of the spectrogram. The spectrogram is compact and efficient in representation carrying information about energy, pitch, fundamental frequency, formants and time of speech signal in the form of a pattern. As Radon transform adds up the pixelvalues in the given image (spectrogram) or time frequency distribution along a straight line in a particular direction at a specific displacement, it captures the effective speaker specific features from the spectrogram. Because of the use of limited number of Radon projections (seven only) and significant DCT coefficients (30%) in the feature vector, the approach yields a low dimensional feature vector, making it computationally efficient. The feasibility of the proposed approach has been successfully tested using the TIMIT and SGGS

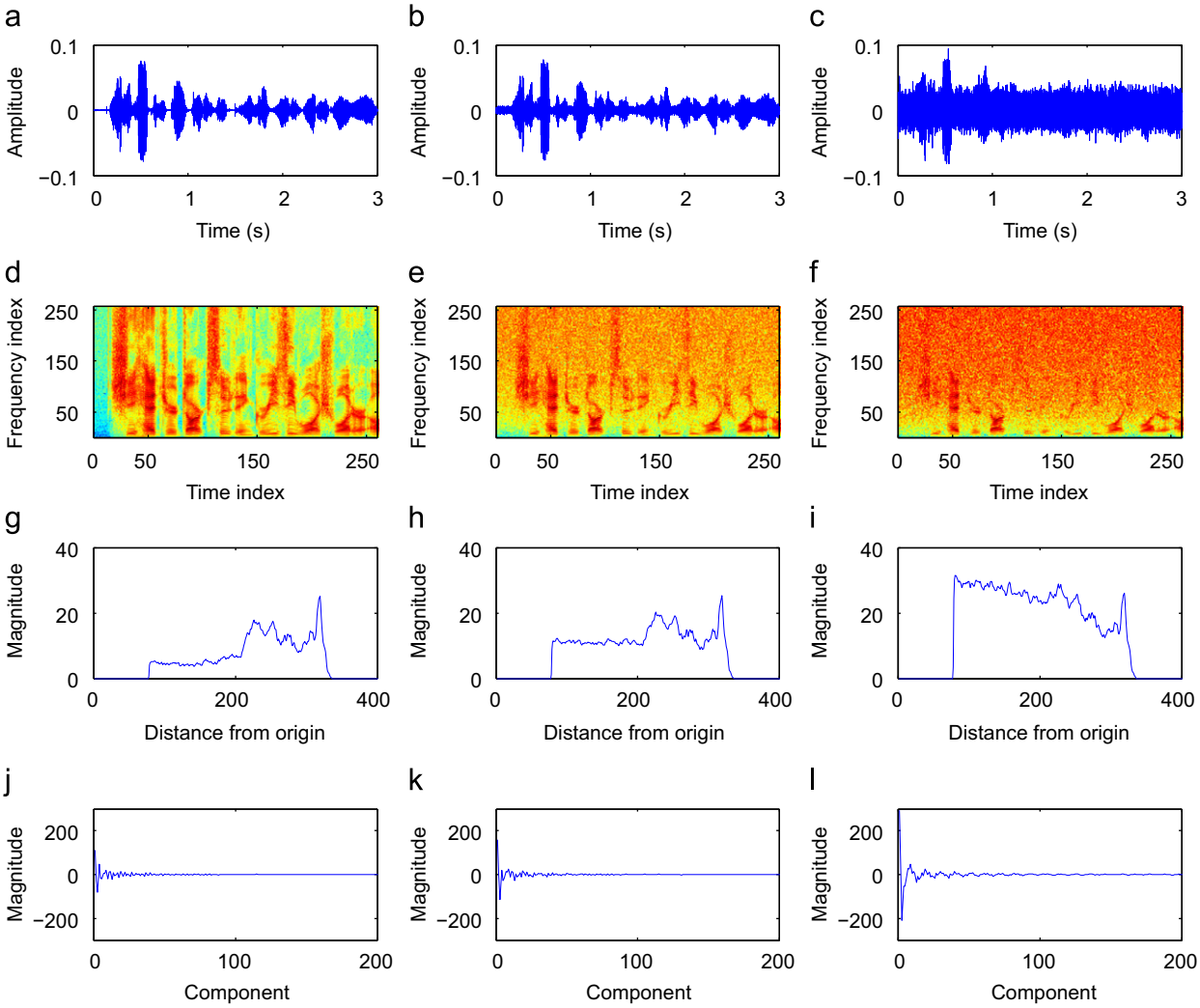

Fig. 10. Typical speech signals, spectrograms, Radon projections and DCT speech for 'no noise', 15 dB and 5 dB SNR: (a)–(c) speech waveforms; (d)–(f) corresponding spectrograms; (g)–(i) Radon projections of respective spectrograms at an angle of 90° and (j)–(l) DCT coefficients of Radon projections.

Table 5
Comparison of computational time for different algorithms (for 100 speakers from TIMIT database).

Algorithm	Dimension of feature vector	Computational time		% Recognition rate
		Training time (s)	Testing time (s)	
MFCC	9984	6341.840	0.5551	98
Modulation spectrogram	3200	4975.341	0.6541	90
TDCT	29133	9012.044	1.0799	92
Proposed approach	772	21.266000	0.2218	98

databases. The approach is also insensitive to additive noise and robust to channel and session variations.

Acknowledgments

The authors are very much thankful to the anonymous reviewers for their constructive comments and valuable suggestions, which helped to improve the quality of this manuscript significantly.

References

[1] M. He, S.J. Horng, P. Fan, R.S. Run, R.J. Chen, J.L. Lai, M.K. Khan, K.O. Sentosa, Performance evaluation of score level fusion in multimodal biometric systems, Pattern Recognition 43 (5) (2010) 1789–1800.
[2] S.J. Horng, D. Mulyono, A study of finger vein biometric for personal identification, in: Proceedings of the IEEE International Symposium on Biometrics and Security Technologies, Islamabad, 2008, pp. 22–23.
[3] S.J. Horng, Y.H. Chen, R.S. Run, R.J. Chen, J.L. Lai, K.O. Sentosal, An improved score level fusion in multimodal biometric systems, in: Proceedings of the International Conference on Parallel and Distributed Computing, Applications and Technologies, Japan, 2009, pp. 239–246.
[4] T. Kinnunen, H. Li, An overview of text-independent speaker recognition: from features to supervectors, Speech Communication 52 (1) (2010) 12–40.
[5] R.J. Mammone, X. Zhang, R.P. Ramachandran, Robust speaker recognition: a feature-based approach, IEEE Signal Processing Magazine 13 (5) (1996) 58–71.
[6] J. Pelecanos, S. Sridharan, Feature warping for robust speaker verification, in: Proceedings of the International Speech Communication Association a Speaker Odyssey—The Speaker Recognition Workshop, Greece, 2001, pp. 213–218.
[7] P. Kenny, G. Boulianne, P. Ouellet, P. Dumouchel, Speaker and session variability in GMM-based speaker verification, IEEE Transactions on Audio Speech and Language Processing 15 (4) (2007) 1448–1460.
[8] R. Vogt, S. Sridharan, Explicit modeling of session variability for speaker verification, Computer Speech and Language 22 (1) (2008) 17–38.
[9] R. Auckenthaler, M. Carey, H. Lloyd-Thomas, Score normalization for text independent speaker verification systems, Digital Signal Processing 10 (1–3) (2000) 42–54.
[10] D. Reynolds, T. Quatieri, R. Dunn, Speaker verification using adapted Gaussian mixture models, Digital Signal Processing 10 (1) (2000) 19–41.

[11] C.W. Seo, K.Y. Lee, J. Lee, GMM based on local PCA for speaker identification, Electronics Letter 37 (24) (2001) 1486–1488.
[12] W. Zhang, Y. Yang, Z. Wu, L. Sang, Experimental evaluation of a new speaker identification framework using PCA, in: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Washington, 2003, pp. 4147–4152.
[13] S.B. Davis, P. Mermelstain, Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences, IEEE Transactions on Acoustic, Speech and Signal Processing 28 (4) (1980) 357–366.
[14] D.A. Reynolds, Experimental evaluation of features for robust speaker identification, IEEE Transactions on Speech and Audio Processing 2 (4) (1994) 639–643.
[15] A. Ljolje, The importance of cepstral parameter correlations in speech recognition, Computer Speech and Language 8 (1994) 223–232.
[16] K. You, H. Wang, Joint estimation of feature transformation parameters and Gaussian mixture model for speaker identification, Speech Communication 28 (3) (1999) 227–241.
[17] L. Liu, J. He, On the use of orthogonal GMM in speaker recognition, in: Proceedings of the IEEE International Conference on Acoustic, Speech, and Signal Processing, Arizona, 1999, pp. 845–848.
[18] M.J.F. Gales, Semi-tied covariance matrices for hidden Markov model, IEEE Transactions on Speech and Audio Processing 7 (3) (1999) 272–281.
[19] D.A. Reynolds, R.C. Rose, Robust text-independent speaker identification using Gaussian mixture speaker models, IEEE Transactions on Speech and Audio Processing 3 (1) (1995) 72–82.
[20] R.A. Cole, A.I. Rudnicky, V.M. Zue, Performance of an expert spectrogram reader, Journal of Acoustic Society of America 65 (1979) 81–87.
[21] T.F. Quatieri, Discrete-Time Speech Signal Processing: Principles and Practice, Prentice Hall, Massachusetts, 2002.
[22] V.W. Zue, An expert spectrogram reader: a knowledge-based approach to speech recognition, in: Proceedings of International Conference on Acoustic Speech and Signal Processing, Japan, 1986, pp. 1197–1200.
[23] R.A. Cole, A.I. Rudnicky, V.W. Zue, D.R. Reddy, in: R.A. Cole (Ed.), Speech as patterns on paper, perception and production of fluent speech, Erlbaum, , 1980.
[24] L. He, M. Lech, N. Maddage, N. Allen, Emotion recognition in speech of parents of depressed adolescents, in: Proceedings of the Third International Conference on Bioinformatics and Biomedical Engineering, China, 2009, pp. 5–8.
[25] L. He, M. Lech, N. Maddage, N. Allen, Emotion recognition in spontaneous speech within work and family environments, in: Proceedings of the Third International Conference on Bioinformatics and Biomedical Engineering, China, 2009, pp. 1–4.
[26] L. He, M. Lech, S. Memon, N. Allen, Recognition of stress in speech using wavelet analysis and teager energy operator, in: Proceedings of the Ninth Annual Conference of the International Speech Communication Association, Australia, 2008, pp. 605–608.
[27] M. Kleinschmidt, V. Hohmann, Sub-band SNR estimation using auditory feature processing, Speech Communication 39 (1–2) (2003) 47–63.
[28] M. Kleinschmidt, Methods for capturing spectro-temporal modulations in automatic speech recognition, Acta Acustica 8 (2001) 1–6.
[29] T. Chih, P. Ru, S. Shamma, Multi resolution spectro-temporal analysis of complex sounds, Journal of Acoustic Society of America 118 (2005) 887–906.
[30] J. Bouvrie, T. Ezzat, T. Poggio, Localized spectro-temporal cepstral analysis of speech, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Las Vegas, 2008, pp. 4733–4736.
[31] T. Ezzat, J. Bouvrie, T. Poggio, Spectro-temporal analysis of speech using 2-D Gabor filters, in: Proceedings of the Tenth International Conference on Spoken Language Processing, Belgium, 2007, pp. 506–509.
[32] T. Ezzat, T. Poggio, Discriminative word-spotting using ordered spectro-temporal patch features, in: Proceedings of the Workshop on Statistical and Perceptual Audition, Australia, 2008, pp. 35–40.
[33] K. Saeed, M.K. Nammous., A speech-and-speaker identification system: feature extraction, description, and classification of speech-signal image, IEEE Transactions on Industrial Electronics 54 (2) (2007) 887–897.
[34] L. Rabiner, B.H. Juang, Fundamentals of Speech Recognition, Prentice Hall, Englewood Cliffs, NJ, 1993.
[35] G. Beylkin, Discrete radon transform, IEEE Transactions on Acoustics Speech and Signal Processing 35 (2) (1987) 162–172.
[36] Kourosh Jafari-Khouzani, Humid Soltanian-Zadeh, Rotation invariant multi-resolution texture analysis using radon and wavelet transform, IEEE Transactions on Image Processing 14 (6) (2005) 783–794.
[37] W. Xuan, X. Bin, M. JianFeng, B. Xiu-Li, Scaling and rotation invariant approach to object recognition based on Radon and Fourier–Mellin transforms, Pattern Recognition 40 (12) (2007) 3503–3508.
[38] W. Chen, M.J. Er, S. Wu, PCA and LDA in DCT domain, Pattern Recognition Letter 26 (15) (2005) 2474–2482.
[39] T. Kinnunen, C. Koh, L. Wang, H. Li, E. Chng, Temporal discrete cosine transform: towards longer term temporal features for speaker verification, in: Proceedings of the Fifth International Symposium on Chinese Spoken Language Processing, Singapore, 2006, pp. 547–558.

Pawan K. Ajmera received the M.E. degree from SGGS Institute of Engineering and Technology, Nanded, India. He is presently a research scholar in Instrumentation Engineering in SGGS Institute of Engineering and Technology, Nanded (India). The areas of his research interest are digital signal processing, image processing and biometrics.

Dr. Dattatray V. Jadhav received the B.E. degree in Electronics Engineering from Marathwada University in 1991 and M. Tech. degree in 1997 from Dr. BA Marathwada University, India. He received the Ph.D. degree from SGGS Institute of Engineering and Technology, Nanded, India in 2009. He is associated with Bhivarabai Swant College of Engineering and Research, Pune. His areas of interests include computer vision, biometrics and image processing.

Dr. Raghunath S. Holambe received the Ph.D. degree from Indian Institute of Technology, Kharagpur, in India and he is presently a professor in Instrumentation Engineering in SGGS Institute of Engineering and Technology, Nanded (India). The areas of his research interest are digital signal processing, image processing, applications of wavelet transform, biometrics, and real time signal processing using DSP processors.