Combining Evidence From Source, Suprasegmental and Spectral Features for a Fixed-Text Speaker Verification System

B. Yegnanarayana, *Senior Member, IEEE*, S. R. Mahadeva Prasanna, Jinu Mariam Zachariah, and Cheedella S. Gupta

***Abstract*—This paper proposes a text-dependent (fixed-text) speaker verification system which uses different types of information for making a decision regarding the identity claim of a speaker. The baseline system uses the dynamic time warping (DTW) technique for matching. Detection of the end-points of an utterance is crucial for the performance of the DTW-based template matching. A method based on the vowel onset point (VOP) is proposed for locating the end-points of an utterance. The proposed method for speaker verification uses the suprasegmental and source features, besides spectral features. The suprasegmental features such as pitch and duration are extracted using the warping path information in the DTW algorithm. Features of the excitation source, extracted using the neural network models, are also used in the text-dependent speaker verification system. Although the suprasegmental and source features individually may not yield good performance, combining the evidence from these features seem to improve the performance of the system significantly. Neural network models are used to combine the evidence from multiple sources of information.**

***Index Terms*—Dynamic time warping, speaker verification, source features, spectral features, suprasegmental features, vowel onset point.**

I. Introduction

SPEAKER recognition by machine is the task of recognizing a person based on the information obtained from his speech signal [1]. Speaker recognition may be divided into speaker identification and speaker verification. Speaker identification is the process of determining to which of the registered speakers a given utterance belongs [2]. Speaker verification is the process of accepting or rejecting the identity claim of the speaker [3]. Based on the text to be spoken, speaker recognition methods can also be grouped into text-dependent and text-independent cases [2]. Text-dependent speaker recognition systems require the speaker to produce speech for the same text in both training and testing, whereas text-independent speaker recognition systems do not depend on the text being spoken. The focus of this paper is on the text-dependent speaker verification system using fixed-text.

Speech signal contains information about the message to be conveyed, identity of the speaker, and language used for communication [1]. Speaker recognition involves extraction of the speaker-specific information from the speech signal. The uniqueness of the speaker-specific information may be attributed to several factors such as the shape and size of the vocal tract, dynamics of the articulators, rate of vibration of the vocal folds, accent imposed by the speaker and speaking rate. All these factors are reflected in the speech signal, and hence are useful for speaker recognition.

The variation in the size and shape of the vocal tract from one speaker to another is reflected as the differences in the resonance frequencies of the short-time spectrum envelope of the speech signal. Present day systems mostly use the spectral information for speaker recognition. The spectral information is extracted from segments of 10–30 ms of the speech signal. These systems ignore other speaker-specific information such as suprasegmental information like pitch and duration, and the information regarding the characteristics of the excitation source. This is mainly due to the difficulty involved in utilizing this information for text-independent speaker verification. However, features from suprasegmental and source information are relatively easier to extract in the case of a text-dependent speaker verification system, when compared to a text-independent case. These features not only provide additional evidence for a speaker verification task, but they are also robust to channel or handset variations.

The present work is an effort to investigate the effectiveness of the suprasegmental and source information for text-dependent speaker verification. An attempt is made to incorporate these features into a baseline text-dependent speaker verification system which uses only the spectral information. The paper is organized as follows: In Section II, the database used for the study is discussed. Section III gives a brief description of the baseline system, and discusses an improved method for end-points detection. Section IV describes the proposed method for extracting pitch and duration information. Section V gives the description of the text-dependent speaker verification system using the characteristics of the excitation source. Section VI describes the proposed method for combining the evidence from different sources of information for a text-dependent speaker verification system. Section VII concludes with a summary of the key ideas proposed in this work.

Manuscript received July 15, 2002; revised April 1, 2004. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ramesh A. Gopinath.

B. Yegnanarayana, S. R. Mahadeva Prasanna, and J. M. Zachariah are with the Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai 600036, India (e-mail: yegna@cs.iitm.ernet.in; prasanna@cs.iitm.ernet.in; jmz00@hotmail.com).

C. S. Gupta is with the Xalted Information Systems Pvt., Ltd, Bangalore, India (e-mail: guptacheedella@yahoo.com).

Digital Object Identifier 10.1109/TSA.2005.848892

II. Speech Database for the Study

The speech database for this study was collected from 30 cooperative speakers (21 male and nine female) over microphone as well as telephone channels. A typical telephone channel has a passband of 300–3300 Hz. In addition to bandwidth limitation, telephone channels may introduce noise and distortion to the spectral characteristics of the speech signal. The speech data was collected for ten sentences of Hindi (an Indian language). The number of words in these sentences vary from five to seven, and the durations of the sentences from 2 to 3 s. Each of the ten sentences was uttered 18 times by each speaker. The data was collected in a laboratory environment in different sessions for microphone and telephone cases. However, one set of all the 18 utterances for each sentence by a speaker was collected in a single session. Thus with this data it is not possible to obtain inter-session variability for the same channel. However, the effect of inter-session variation can be studied along with the effect of inter-channel variation by matching the microphone data with the reference templates for the telephone data or vice versa. The speech data was sampled at 8000 Hz and stored as 8 bit samples.

For each speaker, out of the 18 utterances for each sentence, three utterances are used for creating reference templates. The remaining 15 utterances are used for conducting the genuine speaker tests. Thus there are $15 \times 3 = 45$ genuine trial scores for each speaker. The total number of genuine speaker tests per sentence for 30 speakers is $30 \times 45 = 1350$. Impostor tests for each speaker are conducted by using the utterances of the remaining 29 speakers in the database. For each speaker, three utterances of the same sentence are taken for testing. Thus there are $87 \times 3 = 261$ impostor trial scores for each sentence. Hence, the total number of impostor speaker tests per sentence for 30 speakers is $30 \times 87 = 2610$. Since there is data for ten sentences, the total number of genuine speaker trials are $1350 \times 10 = 13\,500$, and the total number of impostor trials are $2610 \times 10 = 26\,100$.

III. Speaker Verification Using Spectral Features

A speaker verification system consists of four stages: preprocessing, feature extraction, pattern classification and decision making. Preprocessing involves mainly the detection of end-points of a speech utterance. Correct detection of the end-points increases the accuracy of aligning the reference and test utterances [4]. An algorithm based on the amplitude of the speech signal is normally used for detection of the end-points in the baseline system [5]. In this paper we propose an end-point detection algorithm based on Vowel Onset Point (VOP) [6]. A VOP is the instant at which the onset of vowel takes place. The VOPs are obtained using the Hilbert envelope of the linear prediction (LP) residual [6], [7]. The LP residual signal $r(n)$ is obtained by passing the speech signal samples through the inverse filter derived from the LP analysis of speech. The Hilbert envelope $h(n)$ of the residual signal $r(n)$ is given by [6]

$$h(n) = \sqrt{r^2(n) + r_h^2(n)} \tag{1}$$

where $r_h(n)$ is the Hilbert transform of $r(n)$. The Hilbert transform of a signal $r(n)$ is obtained by exchanging the real and imaginary parts of the Discrete Fourier Transform (DFT) of $r(n)$, and then computing the inverse DFT.

TABLE I
Algorithm for Automatic Detection of the Vowel Onset Points (VOP)

1. Preemphasize the input speech by differencing (Fig. 1(a)).
2. Low pass (cut off freq 2.5 kHz) filter the speech signal.
3. Compute the LP residual using 12^{th} order LP analysis, using a frame size of 20 ms and a frame shift of 5 ms.
4. Compute the Hilbert envelope of the LP residual (Fig. 1(b)).
5. Obtain the *VOP evidence plot* (Fig. 1(c)) from the Hilbert envelope by passing the signal through a filter given by
   $g(n) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{n^2}{2\sigma^2}} cos(\omega n)$
   where σ is the spatial spread and ω is the modulating frequency.
   (In this work $\sigma = 100$, $\omega = 0.0114$ and analysis window size = 100 ms)
6. Find the maximum in the *VOP evidence plot*, and identify the peaks candidates for VOPs (Fig. 1(d)).
7. Eliminate the spurious peaks by checking for the presence of the vowel region between two peaks, which is indicated by a negative region in the *VOP evidence plot.*
8. Hypothesize the remaining peaks as VOPs (Fig. 1(e)).

Table I gives the algorithm used for locating the VOPs. A speech utterance from the database and its detected VOPs are shown in Fig. 1. As shown in the figure the speech utterance has 8 VOPs. The proposed algorithm hypothesized all the 8 VOPs correctly. Additionally one spurious VOP is also hypothesized. To evaluate the performance of the VOP detection algorithm, VOPs for 60 randomly chosen utterances from the database are manually marked using the knowledge of the Hilbert envelope of the LP residual. There are totally 480 VOPs. For the same 60 utterances, the VOPs are detected automatically using the proposed algorithm. It was found that 95.6% of the VOPs are correctly detected within a deviation of ± 30 ms [6]. In the chosen 60 utterances, among the total 480 VOPs, 459 are correctly hypothesized, 21 are missing, and 26 are spurious. Among the missing VOPs, very few of them correspond to either the first or the last VOP of the utterance. These missing VOPs are the cases when the strengths of the first vowel and the last vowel are comparable to that of the noise level. These failures can be attributed to the VOP detection algorithm which presently uses only the strength of the LP residual. The first and the last VOPs are used to locate the end-points. The point 300 ms before the first VOP is marked as the begin point of the speech utterance. Similarly, the point 300 ms after the last VOP is marked as the end point of the utterance.

Spectral information is extracted for each differenced and Hamming windowed frame of the speech signal using linear prediction (LP) analysis [7]. The spectral information is represented using weighted linear prediction cepstral coefficients

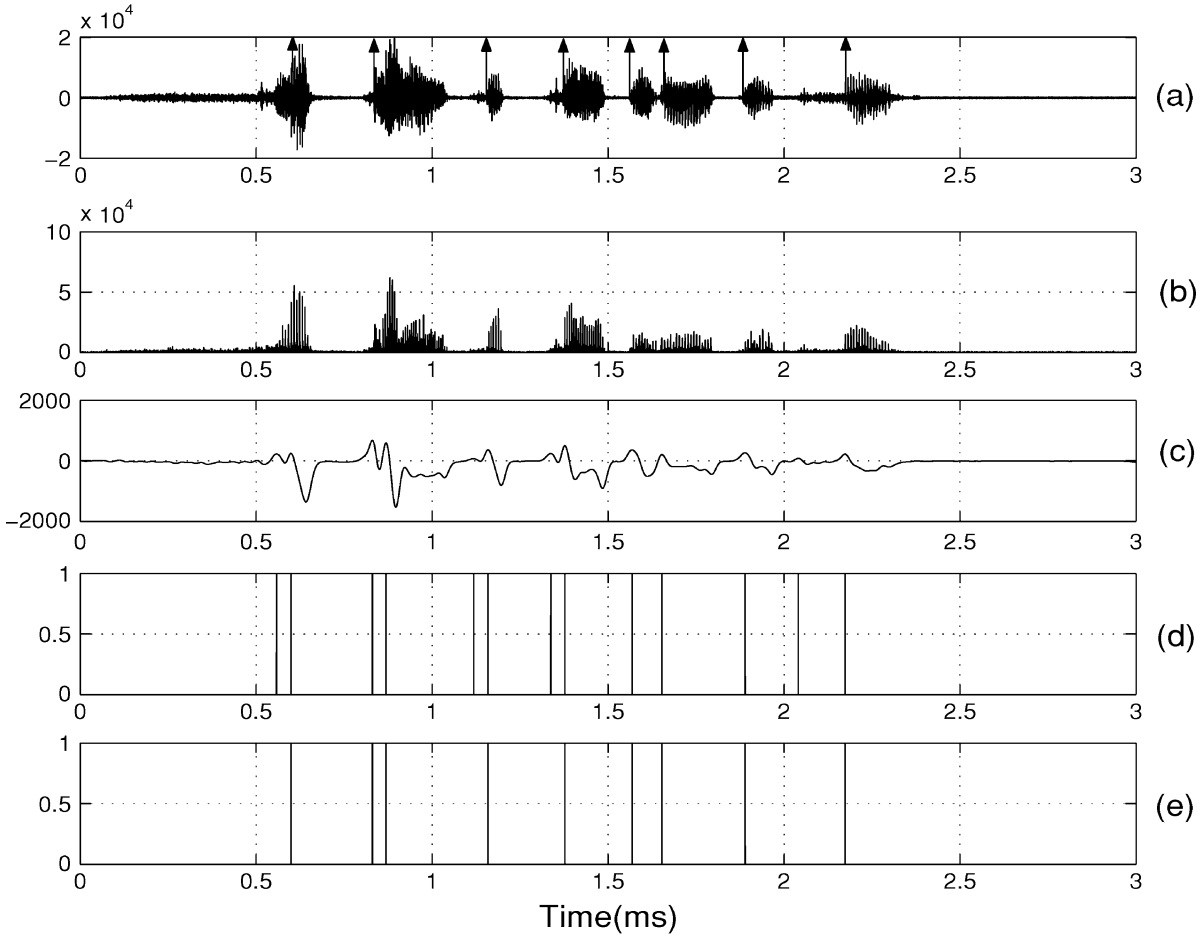

Fig. 1. Steps in the detection of VOPs. (a) Speech signal with manually marked VOPs. (b) Hilbert envelope of the LP residual. (c) *VOP evidence plot*. (d) Peaks as candidates of VOPs. (e) Hypothesized VOPs.

(WLPCC) and the corresponding delta cepstral coefficients [3]. A 12th order LP analysis is used to derive the 20 weighted linear prediction cepstral coefficients for each 20 ms frame. The delta cepstral coefficients are obtained by deriving the average slope of the contour for each of the WLPCCs from seven successive frames [3]. Only the first five delta cepstral coefficients are considered, as it was experimentally found that the other delta coefficients did not contribute much to the performance of the speaker verification system [8]. Thus the feature vector for each frame consists of 25 components (20 WLPCCs and 5 delta cepstral coefficients). We use this 25 dimension vector to represent segmental features of speech.

Both the reference and test utterances are represented by a sequence of 25 dimension feature vectors. The reference and test utterances are matched using the Dynamic Time Warping (DTW) algorithm [9]. The matching score is the minimum distance, which is obtained along the optimal warping path of the DTW algorithm.

The performance of the speaker verification system is evaluated as follows: For each speaker for each sentence, the genuine and the impostor scores are normalized to the range from -1 to 1. The threshold is linearly varied from -1 to 1, and at each threshold the fraction of the False Acceptance (FA) and the fraction of the False Rejection (FR) are noted. The point at which the FA and FR curves as a function of the threshold meet is the Equal Error Rate (EER) for that speaker. The average value of the EER for all the speakers and for all the sentences is given in the first row of Table II. We found that the results obtained using the proposed VOP-based end-points detection are significantly better (reduction in EER by more than half) than the results obtained using simple amplitude-based approach for end-points detection [6].

Comparing the performances of the system for microphone and telephone speech data, it can be seen that the performance of the system degrades for the telephone speech as well as for inter-session/inter-channel case. The performance of the speaker verification system can be improved by incorporating additional information. In the following sections we explore the use of pitch and duration information, as well as the information in the excitation component of speech.

IV. Duration and Pitch Information for Speaker Verification

State-of-the-art automatic speaker recognition systems are based on the spectral features extracted over short segments (10–30 ms) of speech [10]. Humans use several features at suprasegmental level like pitch, duration, idiolect (or word usage), speaking rate and speaking style for recognizing speakers. Atal [11] proposed a speaker recognition method

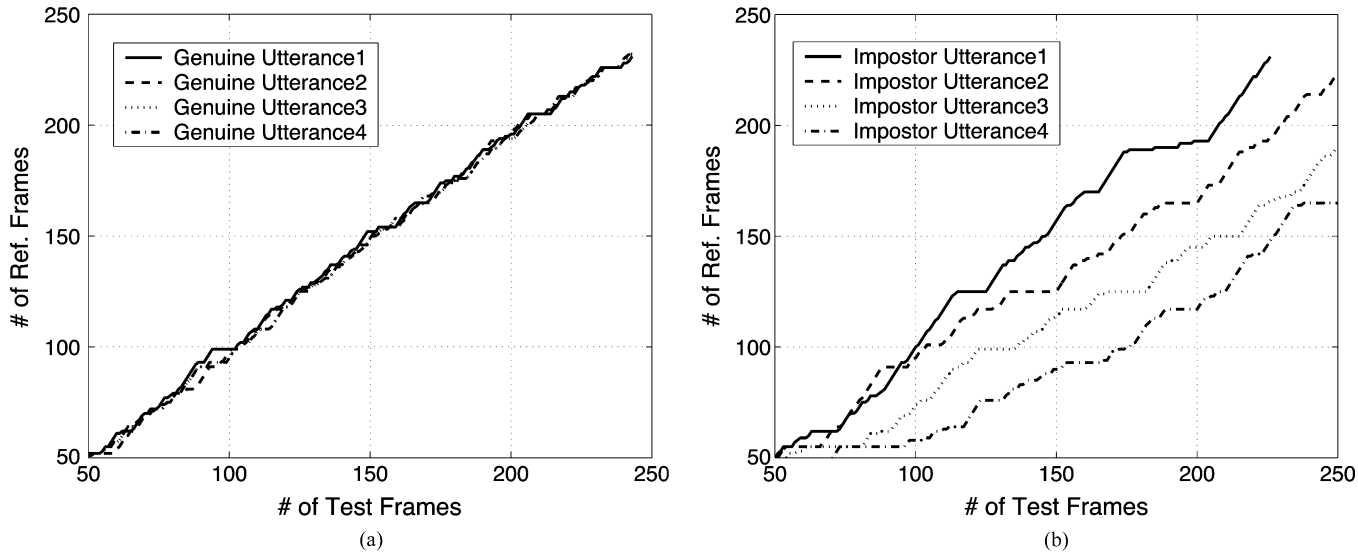

Fig. 2. Optimal warping paths of four different test utterances matched against the same reference template. (a) Genuine case. (b) Impostor case.

using pitch contours. Significance of long-term features like pitch and energy for speaker recognition is discussed in [12]. Statistical features of pitch, pitch tracks and local dynamics in pitch are also used in speaker verification [13]. In these studies ([12], [13]) it is shown that addition of the suprasegmental features improves the performance of spectral-based speaker verification system. A text-prompted speaker verification technique using pitch information in addition to spectral information is proposed in [14]. The system is found to be rejecting impostor testing using synthetic speech better compared to the speaker verification based only on spectral features.

The usefulness of prosody and lexical information for speaker identification is demonstrated in [15]. Recently a project titled **SuperSID** is undertaken for exploring the usefulness of high-level information for speaker recognition [16]. The objective of this project is to analyze, characterize, extract and apply high-level information to the speaker recognition task. Some of the results of this project are published in [17], [18]. Approaches for modeling the dynamics in the pitch and energy trajectories is proposed in [18]. The study of suprasegmental features for speaker recognition is also reported in [17]. It is shown that the suprasegmental features provide significant improvement when added with the spectral-based speaker verification systems.

In this work we propose methods to extract duration and pitch information for the text-dependent speaker verification task. The novelty of the proposed methods lies in the way of exploiting the nature of warping path to derive duration and pitch information.

A. Duration Information

Interest in the use of duration information for speaker recognition appears to be less because of the difficulty in locating the boundaries, and measuring the duration of the units such as syllable, word or phrase. It is useful to extract the duration information without explicitly locating the boundary of any unit. This is accomplished in this work by using the nature of the optimal warping path obtained in the DTW algorithm. The nature of the DTW path indicates the extent of mismatch between the relative durations of the units in the reference and the test utterances.

The baseline system uses the DTW algorithm only for obtaining the matching score. It ignores the information present in the resulting warping path. The DTW path is represented by a sequence of K points $C(1), C(2), \ldots, C(k) \ldots, C(K)$, where $C(k) = (x(k), y(k)), x(k)$ is the frame index of the test utterance, and $y(k)$ is the frame index of the reference utterance. An analysis was carried out to study the nature of the warping path by matching the reference and test utterances of genuine and impostor speakers. It was observed that the nature of the warping path that joins the points $C(k), k = 1, \ldots, K$ follows closely the diagonal line in the x-y plane for genuine speakers, whereas it deviates significantly from the diagonal line for impostor speakers. Fig. 2 illustrates the behavior of the warping paths for genuine and impostor speaker test utterances with a reference utterance of a target speaker. The significance of this behavior of the warping path can be explained as follows.

The duration of the test utterance by a genuine or an impostor speaker for a given text may or may not match with the duration of the reference utterance of the target speaker. As a result, it is not possible to arrive at a conclusion based on matching the durations (amount of time taken) of the entire utterances, or even by matching the durations of each of the corresponding units (such as syllable or word or phrase) in the utterances. But it is interesting to note that, although the total duration of the utterance of the same text may vary from that of the reference utterance for the genuine speaker, the relative durations or the percentage durations of the units in the utterance are usually consistent. This consistency in the relative durations of the units in the reference and test utterances results in a warping path which is nearly straight. If a mismatch occurs between the relative durations of the units of the reference and test utterances, then the nature of the warping path will be highly irregular. In other words, the extent of mismatch between the relative durations of the units of the reference and test utterances is related to the deviation of the warping path from a straight line. The straight line is the regression line obtained by the least square fit of the points along the

TABLE II
PERFORMANCE OF THE TEXT-DEPENDENT SPEAKER VERIFICATION SYSTEM WITH SPEECH FROM MICROPHONE, TELEPHONE, AND INTER-SESSION. THE ABBREVIATIONS MIC, TEL, AND INT INDICATE MICROPHONE, TELEPHONE, AND INTER-SESSION, RESPECTIVELY

Type of system	Row No.	Feature	Equal Error Rate		
			MIC	TEL	INT
Basic systems	1	Spectral	2.54	2.77	3.73
	2	Duration	7.79	7.50	7.54
	3	Pitch	8.67	8.44	8.70
	4	Source	13.57	14.54	14.24
Combination systems	5	Spectral (using *small* test data)	2.60	2.89	3.94
	6	Spectral+Duration	1.14	1.28	1.99
	7	Spectral+Duration+Pitch	0.75	0.84	1.12
	8	Spectral+Duration+Pitch+Source	0.33	0.70	0.74

warping path. The deviation of each point $y(k)$ of the warping path from its regression line is an indication of the mismatch in the relative durations between the units of the reference and test utterances. The regression line of the warping path is given by $y'(k) = mx(k) + c$, where $y'(k)$ is the point on the regression line corresponding to the frame $x(k)$ on the x-axis, m is the slope of the regression line, and c is the intercept of the regression line. The slope and the intercept of the regression line are computed by means of the least squares method. The deviation of the actual warping path from the regression line is indicated by the average sum of the squared error (E_d), given by

$$E_d = \frac{\sum_{k=1}^{K}(y'(k) - y(k))^2}{K} \tag{2}$$

where $K =$ Number of points along the warping path.

In order to have a comparative study of the effectiveness of the duration information and the spectral information for speaker verification, the same speech database described in the previous section is used. The test utterances are matched against the three reference utterances of the target speaker using the DTW algorithm. The error as given by (2) is computed for each of the comparisons. These error values are then normalized to the range -1 to $+1$. As the threshold is varied linearly from -1 to $+1$, the fraction of FA and FR are noted. The point at which the FA and FR curves as function of the threshold meet is noted as the EER for that speaker. The average value of the EER for all the speakers and for all the sentences is given in the second row of Table II.

It can be observed from the table that the duration information does not critically depend on the channel characteristics. The performance of the system will be poor when the duration information alone is used. But when combined with the evidence from the spectral information, the performance of the speaker verification improves significantly as discussed in Section VI.

B. Pitch Information

The pitch contour of an utterance also contributes to the speaker-specific information. Pitch is the acoustic correlate of the rate of vibration of the vocal folds. The uniqueness of the rate of vibration of the vocal folds is due to differences in the size of the vocal folds, and also due to the accent imposed by the speaker (which is a learnt attribute). The physiological constraints determine the average pitch of the speaker. The speaking style determines the pitch pattern or the variation of the pitch frequency as a function of time. The local variations of the pitch contour is more representative of the speaker than the average pitch.

In this work an attempt is made to incorporate the pitch pattern information in a text-dependent speaker verification system. The evidence obtained from the pitch information is likely to be complementary to that obtained from the spectral information of the speech signal. It is also known that the pitch features are not sensitive to channel variations [11].

The similarity of the pitch contours of the reference and test utterances can be captured by using the optimal warping path obtained in the DTW algorithm. The pitch contour of an utterance is computed using the Simple Inverse Filtering Technique (SIFT) algorithm [19]. The absolute difference of the pitch frequencies for a few selected matching frames in the reference and test utterances are summed up to get the pitch score (P_s). Twenty pairs of matching frames are selected such that the Euclidean distance between the spectral feature vectors of these frame pairs are the lowest among all the points in the warping path. Also it should be ensured that none of these pairs have a zero pitch frequency in the reference and test frames. In this way we ensure that the sound units are similar in both the reference and test utterances, and that those units are voiced. The pitch score is computed as

$$P_s = \sum_{i=1}^{L} |F_0(x(i)) - F_0(y(i))| \tag{3}$$

where

$F_0(x(i))$ pitch frequency of the frame $x(i)$ of the test utterance;

$F_0(y(i))$ pitch frequency of the frame $y(i)$ of the reference utterance;

L number of points (20) chosen for computing the pitch score. These points correspond to the least distant pairs, and which also satisfy the condition $F_0(x(i)) \neq 0$ and $F_0(y(i)) \neq 0$.

The speech database and the number of genuine and impostor speaker tests are the same as described in the previous sections. The pitch scores are normalized to the range -1 to $+1$. As the threshold is varied linearly from the -1 to $+1$, the fraction of FA and FR are noted. The point at which the FA and FR curves as function of the threshold meet is noted as the EER for that speaker. The results of the text-dependent speaker verification system using the pitch information is given in the third row of Table II.

The experiments show that the pitch contour gives useful speaker-specific information. Although the performance of the system is poor if pitch alone is used for speaker verification, the performance improves significantly when combined with the evidence from spectral and duration information. This is discussed in Section VI.

V. Source Information for Speaker Verification

The excitation source in the production of speech seems to provide some cues for automatic speaker verification [20]. Information about the excitation source is present in the linear prediction (LP) residual of the speech signal. It was shown that auto-associative neural network (AANN) models can be used to capture the speaker-specific information in the residual signal [20]. This section describes studies made on the source characteristics of the speech signal for a text-dependent speaker verification [8].

In the LP analysis the signal s_n is predicted using a linear weighted sum of the past p samples. The error between the actual value s_n and the predicted value $\hat{s_n}$ is given by

$$e_n = s_n - \hat{s_n} \tag{4}$$

where

$$\hat{s_n} = -\sum_{k=1}^{p} a_k s_{n-k}. \tag{5}$$

The error e_n is the LP residual of the speech signal, when the optimal values of $\{a_k\}$ are used in (5). The optimal values of $\{a_k\}$ are obtained by using the least square error formulation of the linear prediction [7]. This involves solution of the normal equations given by

$$\sum_{k=1}^{p} a_k R(n-k) = -R(n), \quad n = 1, \ldots p \tag{6}$$

where $R(n) = \sum_m s(m)s(m-n)$ is the autocorrelation function. The autocorrelation function is a second order statistic. Although the second order correlation is removed by means of the LP analysis, most of the higher order relations among the samples may still remain in the LP residual. These relations in the residual may contain speaker-specific information, and can be captured using AANN models [20]. An AANN consists of one input layer, one output layer, and one or more hidden layers. The units in the input and output layers are linear units, and the units in the hidden layers are nonlinear. The number of units in the input and output layers are equal to the dimension of the input data. The middle hidden layer may have fewer units than in the input and output layers.

The AANN model used in this study has five layers with the structure shown in Fig. 3. The structure of the network is $40L\,48N\,12N\,48N\,40L$, where L denotes linear units and N denotes nonlinear units. The structure is based on the extensive studies made on these models for extracting speaker-specific information [20]. The AANN is trained using the LP residual signal. Blocks of 40 samples of the LP residual are considered with a shift of 1 sample. Each block is normalized to unit magnitude by dividing each sample with the magnitude of the vector of samples in the block. The target output is same as the input vector. The weights of the network are adjusted using the backpropagation algorithm [21]. The network weights are initialized to random values, and the network is trained for 60 epochs. One epoch consists of giving all the frames of the residual signal in succession. The number of frames is almost equal to the number of the samples, except those belonging to the silence and the low energy regions. The final weights define the model for the speaker. Three models are created for the three reference utterances of a given speaker.

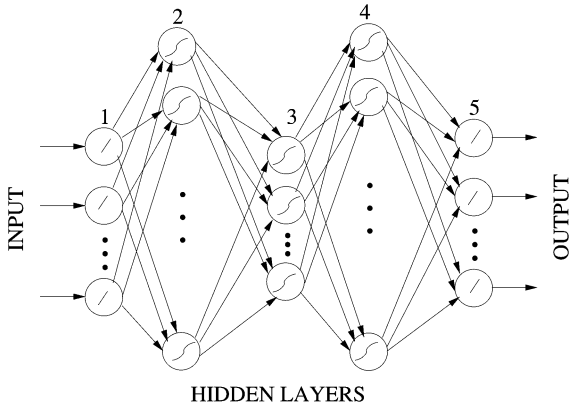

Fig. 3. Structure of AANN.

Blocks of the normalized residual samples from the test utterance are given as input to the AANN model of the target speaker. The squared error (E_i) between the actual output and the target output (which is same as the input vector) is obtained for each frame. The average of the error values over all the frames is computed as $c = (1/N)\sum_{i=1}^{N} E_i$, where N is the number of the frames used in testing. The average error value per frame is the score obtained by comparing the test utterance with the target model.

In order to evaluate the effectiveness of the source features, both genuine and impostor speaker tests were conducted. For consistency and comparison of the performance of the system using different features, the database is same as used in the previous sections. The results of the speaker verification system using the source information is given in the fourth row of Table II. Although the performance of the speaker verification system using the source information alone is poor compared to the performance obtained using the spectral, duration and pitch information, the performance improves significantly when this evidence is combined with the evidence from other sources of information as discussed in the next section.

VI. Combining Evidence From Segmental, Suprasegmental and Source Information

Studies have shown that features and classifiers of different types may complement each other, and thus give an improvement in the classification performance when they are used together [22]. Since the features derived separately from the spectral (segmental), duration and pitch (suprasegmental) and excitation (source) information give nearly independent sources of evidence, they can be combined to improve the performance of a speaker verification system. In this section we discuss methods based on neural network models to combine the evidence obtained from different types of features.

There are four different features, namely, spectral, duration, pitch and excitation source. We can obtain evidence from each of these features as described in the previous sections. There are 45 genuine trial scores and 261 impostor trial scores for each speaker per sentence. All the scores are normalized to the range

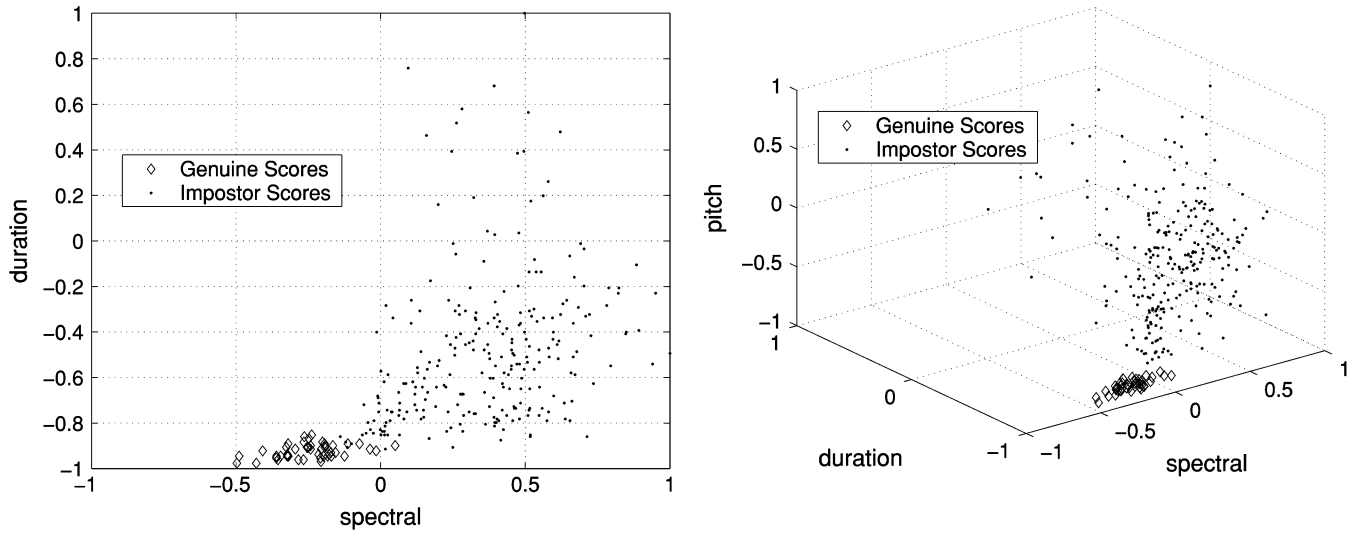

Fig. 4. Distribution of the score vectors for genuine and impostor trials. (a) Spectral and duration information. (b) Spectral, duration, and pitch information.

-1 to $+1$. Since there are 4 different features, we get a 4-dimensional vector for each of the genuine trial scores and impostor trial scores. If we use 2 or 3 features only out of the 4, then correspondingly the genuine and impostor scores will be 2-D and 3-D, respectively.

The objective is to develop a 2-class (genuine and impostor) classifier using the 2-D or 3-D or 4-D score vectors, corresponding to the result of combining two, three, or four features, respectively. Fig. 4(a) and (b) illustrate the distribution of the score vectors for a given speaker for one sentence. In order to capture the dividing surface between the genuine and impostor classes, a multilayer perceptron (MLP) network is trained using 30 of the 45 genuine trail score vectors and 180 of the 261 impostor trial score vectors. The remaining 15 genuine trail score vectors and 81 impostor trial score vectors are used for evaluating the 2-class network classifier. The impostor examples used for testing are obtained from speakers different from those used for training the MLP. The structure of the MLP depends on the number of features used to combine the scores. The 4-layer structures of the MLP for combining two, three, and four features are respectively, $2L4N3N1N, 3L6N3N1N$ and $4N8N3N1N$, where L refers to linear units, N refers to nonlinear with activation function $\tanh(\cdot)$, and the numbers refer to the number of units in that layer. These structures were determined based on some preliminary experiments. However, the number of layers and the number of units in the layers can vary significantly without affecting the performance of classification.

For each classifier, the genuine (15) and impostor (81) trial score vectors of the test data for each speaker and for all the sentences are used to evaluate the performance of the combination feature. There are 4500 genuine trials and 8700 impostor trials. The threshold at the output layer is varied from -1 to $+1$, and the fraction of FA and FR utterances are obtained. The value of the FA/FR at which the FA and FR curves as a function of the threshold intersect will give the EER. The EERs are obtained for all the ten sentences and for all the 30 speakers, and the average of these EERs gives an indication of the performance of the verification system. The average EER values, when two, three, and four features are combined, are given in rows 6, 7, and 8, respectively, in Table II. We have obtained the EER values for the *small* test data using spectral information alone, and the results are given in the fifth row of the table. Note that these are not significantly different from the EER values obtained using the complete data (see first row). But the values shown in the fifth row are useful for comparing the EER values of the combined systems.

As expected, the performance of the speaker verification system improves as more features from independent sources of information are used in combination. For example, the performance of the system is better when the duration information is combined with the spectral information, compared to the performance with the spectral information alone. Likewise, the performance obtained is best when all the four features are used. The overall performance is good for both telephone and also for inter-channel tests. The inter-channel tests also correspond to inter-session tests. Hence the degradation in performance is not significant due to channel and session variations, when evidence from multiple sources of information are combined.

VII. Summary

Most present day systems for speaker verification use the information about the characteristics of the vocal tract, which are reflected in the short-time spectral features. These systems ignore the information present at the suprasegmental level such as duration and pitch. They also ignore the speaker characteristics present in the excitation source during speech production. We have proposed a method to extract duration and pitch information from the warping path of the dynamic time warping algorithm. The speaker-specific characteristics of the excitation source are extracted from the LP residual using an AANN model. Features from spectral, duration, pitch and excitation components of speech provide evidence from independent sources of information. Hence the performance of the speaker verification system could be improved by combining the evidence from these multiple sources of information. The evidence from the different sources were combined using a MLP network. It was shown that not only that the performance of verification improved, but also the nonspectral features such as duration, pitch and excitation

source were found to be robust for variations due to channel and session. That is the reason why there is not much difference in the performance for telephone channel and for inter-channel comparison. The slightly poorer performance for the telephone channel compared to microphone channel is due to generally low SNR of the telephone speech data.

We have also proposed a robust method for end-points detection, which is crucial for a text-dependent speaker verification system based on template matching. Therefore with a good end-points detection algorithm, coupled with combining evidence from multiple sources of information, it is possible to build robust text-dependent speaker verification systems.

REFERENCES

[1] D. O'Shaughnessy, "Speaker recognition," in *IEEE Acoust., Speech, Signal Process. Mag.*, vol. 3, Oct. 1986, pp. 4–17.

[2] H. Gish and M. Schmidt, "Text—independent speaker identification," *IEEE Signal Process. Mag.*, vol. 11, pp. 18–32, Oct. 1994.

[3] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 29, pp. 254–272, Apr. 1981.

[4] L. Rabiner and M. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell Syst. Tech. J.*, vol. 54, pp. 297–315, 1975.

[5] M. Mathew, "Combining evidences from multiple classifiers for text-dependent speaker verification," M.S. thesis, Indian Inst. Technol., Madras, Chennai, 1999.

[6] S. R. M. Prasanna, J. M. Zachariah, and B. Yegnanarayana, "Begin-end detection using vowel onset points," in *Proc. Workshop on Spoken Language Processing*. Mumbai, India, Jan. 2003, pp. 33–40.

[7] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, Apr. 1975.

[8] J. M. Zachariah, "Text-Dependent Speaker Verification Using Segmental, Suprasegmental and Source Features," M.S. thesis, Indian Inst. Technol. Madras, Chennai, 2002.

[9] L. Rabiner, A. Rosenberg, and S. Levinson, "Considerations in dynamic time warping algorithms for discrete word recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-26, pp. 575–582, Dec. 1978.

[10] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted mixture models," *Digital Signal Process.*, vol. 10, pp. 181–202, 2000.

[11] B. S. Atal, "Automatic speaker recognition based on pitch contours," *J. Acoust. Soc. Amer.*, vol. 52, no. 6, pp. 1687–1697, 1972.

[12] M. J. Carey, E. S. Parris, H. Lloyd-Thomas, and S. Bennett, "Robust prosodic features for speaker identification," in *Proc. Int. Conf. Spoken Language Processing*, Philadelphia, PA, USA, Oct. 1996.

[13] M. K. Sonmez, E. Shriberg, L. Heck, and M. Weintraub, "Modeling dynamic prosodic variation for speaker verification," in *Proc. Int. Conf. Spoken Language Processing*, Sydney, NSW, Australia, Nov.–Dec. 1998.

[14] T. Masuko, K. Tokudo, and T. Kobayashi, "Imposture using synthetic speech against speaker verification based on spectrum and pitch," in *Proc. Int. Conf. Spoken Language Processing*, vol. 2, Beijing, China, Oct. 2000, pp. 302–305.

[15] F. Weber, L. Manganaro, B. Peskin, and E. Shriberg, "Using prosodic and lexical information for speaker identification," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Orlando, FL, May 2002, pp. 141–144.

[16] D. Reynolds, W. Andrews, J. Campbell, J. Navartil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Codfrey, D. Jones, and B. Xiang, "The SuperSID project: Exploiting hig-level information for high-accuracy speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 4, Hong Kong, Apr. 2003, pp. 784–787.

[17] B. Peskin, J. Navratil, J. Abramson, D. Jones, D. Klusacek, D. A. Reynolds, and B. Xiang, "Using prosodic and conversational features for high-performance speaker recognition: Report from JHU WS'02," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. IV, Hong Kong, Apr. 2003, pp. 784–787.

[18] A. G. Adami, R. Mihaescu, D. A. Reynolds, and J. J. Godfrey, "Modeling prosodic dynamics for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. IV, Hong Kong, Apr. 2003, pp. 784–787.

[19] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 367–377, Dec. 1972.

[20] B. Yegnanarayana, K. S. Reddy, and S. P. Kishore, "Source and system features for speaker recognition using AANN models," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Salt Lake City, UT, May 2001.

[21] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, J. L. McClelland and D. E. Rumelhart, Eds. Cambridge, MA: MIT Press, 1986, vol. 1, ch. 8. PDP Research Group.

[22] J. J. Hull and S. N. Srihari, "Decision combination in multiple classifier systems," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 16, pp. 66–75, Jan. 1994.

B. Yegnanarayana (M'78–SM'84) was born in India in 1944. He received the B.E., M.E., and the Ph.D. degrees in electrical communication engineering from the Indian Institute of Science, Bangalore, in 1964, 1966, and 1974, respectively.

He was a Lecturer from 1966 to 1974 and an Assistant Professor from 1974 to 1978, in the Department of Electrical Communication Engineering at the Indian Institute of Science. From 1978 to 1980, he was a Visiting Associate Professor of computer science at Carnegie Mellon University, Pittsburgh, PA. Since 1980, he has been a Professor in the Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai. He was the chairman of the department from 1985 to 1989. His current research interests are in signal processing, speech, vision, neural networks, and man-machine interfaces. He has published papers in reviewed journals in these areas. He is also the author of the book *Artificial Neural Networks* (India: Prentice-Hall, 1999).

Dr. Yegnanarayana is a Fellow of Indian National Science Academy, Indian National Academy of Engineering, and Indian Academy of Sciences. He is an Associate Editor of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING.

S. R. Mahadeva Prasanna was born in India in 1971. He received the B.E. degree in electronics engineering from Sri Siddartha Institute of Technology, Bangalore University, India, in 1994, and the M.Tech. degree in industrial electronics from National Institute of Technology, Surathkal, India, in 1997. He received the Ph.D. degree in computer science and engineering from the Indian Institute of Technology Madras, Chennai, in 2004.

He is currently an Assistant Professor at IIT, Guwahati. His research interests are in speech signal processing and neural networks.

Jinu Mariam Zachariah was born in India in 1977. She received the B.Tech. degree in electrical and electronics engineering from M.A. College of Engineering Kothamangalam, Kerala, in 1998 and the M.S. degree in computer science and engineering, from the Indian Institute of Technology Madras, Chennai, in 2002.

Her research interests are in speaker recognition and neural networks.

Cheedella S. Gupta was born in India in 1975. He received the B.Tech. degree in electronics and communication engineering from Jawaharlal Nehru Technological University, India, in 2000, and the M.S. degree in computer science and engineering from the Indian Institute of Technology Madras, Chennai, in 2003.

He is currently a Member of Technical Staff with Xalted Information Systems Pvt. Ltd., Bangalore, India. His research interests are in speech signal processing and VLSI design.