Neural Network Models for Combining Evidence from Spectral and Suprasegmental Features for Text-Dependent Speaker Verification

S.R. Mahadeva Prasanna, Jinu Mariam Zachariah and B. Yegnanarayana
Speech and Vision Laboratory
Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai-600 036, India
{prasanna,jinu,yegna}@cs.iitm.ernet.in

Abstract

This paper proposes a method using neural network models for combining evidence from spectral and suprasegmental features for text-dependent speaker verification. Spectral features are extracted using the Dynamic Time Warping (DTW) technique. While extracting the spectral features, the DTW algorithm is used only to obtain a matching score and the information present in the warping path is ignored. In this work a method is discussed to extract suprasegmental features such as pitch and duration using the information in the warping path. Although the suprasegmental features may not yield good performance, combining the evidence from suprasegmental and spectral features improves the performance of the speaker verification system significantly.

1. INTRODUCTION

The advancement in technology has led to several applications like remote access to computers, voice dialling, banking transactions over telephone, voice mail and database access services. This technology advancement has also increased the threat to access information by unauthorized people. One way to provide security is to allow the persons to authenticate using biometrie features like speech, face, finger print and eyes. This is because the biometric features are known to be unique for each person. The advantage of using speech as biometric feature for person authentication is that it enables people to access information even from remote places. Speaker verification, a pattern recognition task, is the process of accepting or rejecting the identity claim of the speaker using the information obtained from the speech signal [1]. Text-dependent speaker verification system requires the speaker to provide speech for the same text in both training and testing. Speech signal contains information about the message, speaker, language and emotional status of the speaker. The speaker information manifested in the speech signal may be attributed to several factors such as, shape and size of the vocal track, rate of vibration of vocal folds, accent imposed by the speaker and speaking rate. The speaker information may be extracted by the analysis of speech at segmental or suprasegmental level. In segmental analysis speech is blocked into frames of 10-30 msec and spectral information is extracted from these frames. State-of-the-art speaker recognition systems are based on the spectral features [2, 3].

Humans use several features at the suprasegmental level like pitch, duration, idiolect (word usage), speaking rate and speaking style for recognizing speakers. Few attempts have been made to use the suprasegmental information for speaker recognition by machines. This is mainly due to the difficulty in extracting suprasegmental features from the speech signal. Some of the attempts made are as follows: Atal [4] proposed a speaker recognition method using pitch contours. Significance of long-term features like pitch and energy for speaker recognition is discussed in [5]. Statistical features of pitch, pitch tracks and local dynamics in pitch are also used in speaker verification [6]. A text-prompted speaker verification technique using pitch information in addition to spectral information is proposed in [7]. A novel speaker recognition technique which exploits the idiolectal differences or word usage among the speakers is also proposed recently [8]. The usefulness of prosody and lexical information for speaker identification is demonstrated in [9]. Recently a project titled *SuperSID* is undertaken for exploring the usefulness of high-level information for speaker recognition [10, 11]. In all these studies it is shown that the suprasegmental features provide significant improvement in performance when combined with spectral features for speaker verification.

In this paper we discuss methods for extracting pitch and duration information for text-dependent speaker verification task [12, 13]. These methods differ from the existing methods mainly in the way of deriving the pitch and duration information. This is achieved by exploiting the nature of warping path obtained by the Dynamic Time Warping (DTW) technique. The objective of this paper is to propose a method for combining the evidence from spectral features with suprasegmental features. The paper is organized as follows: Section 2 discusses about the text-dependent speaker verification system using spectral features. In Section 3 method for extracting suprasegmental features given in [12, 13] is discussed. A neural network based approach for combining evidence from spectral and suprasegmental features is proposed in Section 4. A summary of the work and scope for future study is given in Section 5.

2. SPEAKER VERIFICATION USING SPECTRAL FEATURES

Speech database for this study was collected from 30 cooperative speakers (21 male and 9 female) over microphone as well as telephone channels. A typical telephone channel has a passband of 300-3300 Hz. In addition to bandwidth

0-7803-8243-9/04/$17.00 © 2004 IEEE

limitation, telephone channels may introduce noise and distortion to the spectral characteristics of the speech signal. The speech data is collected for 10 sentences of Hindi. The number of words in these sentences vary from 5 to 7, and the durations of the sentences from 2 to 3 secs. Each of the 10 sentences was uttered 18 times by each speaker. The data was collected in a laboratory environment in different sessions for microphone and telephone cases.

A speaker verification system consists of four stages: preprocessing, feature extraction, pattern classification and decision making. Preprocessing involves mainly the detection of the end-points of a speech utterance. Correct detection of the end-points increases the accuracy of aligning the reference and test utterances. An algorithm based on the knowledge of Vowel Onset Point (VOP) is used for the detection of end-points [13]. VOP is the instant at which the onset of vowel takes place. As discussed in [13], this algorithm is found to be robust as compared to the existing end-points detection algorithms based on the amplitude of the speech signal.

Spectral information is extracted for each differenced and Hamming windowed frame of the speech signal using a 12^{th} order Linear Prediction (LP) analysis [14]. The spectral information is represented using Weighted Linear Prediction Cepstral Coefficients (WLPCC) and the corresponding delta cepstral coefficients [15]. Thus the feature vector for each frame of 20 msec consists of 25 components (20 WLPCCs and 5 delta cepstral coefficients). 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 [16]. The matching score is the minimum distance, which is obtained along the optimal warping path of the DTW algorithm.

For each speaker, out of the 18 utterances for each sentence, 3 utterances are used for creating reference templates. The remaining 15 utterances are used for conducting the genuine speaker tests. Thus there are 45 genuine trial scores (15 x 3) for each speaker. Hence the total number of genuine speaker tests per sentence for 30 speakers is 1350 (30 x 45). Imposter tests for each speaker are conducted by using the utterances of the remaining 29 speakers in the database. For each speaker, randomly chosen three utterances of the same sentence are taken for testing. Thus there are 261 impostor trial scores (87 x 3) for each speaker for each sentence. Hence, the total number of imposter speaker tests per sentence for 30 speakers is 2610 (30 × 87). Since there is data for ten sentences, the total number of genuine speaker trials are 13500 (1350 x 10), and the total number of impostor trials are 26100 (2610 x 10).

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 Table 1 (see Sl.No.3). The performance of the speaker verification system can be improved by incorporating additional information. In the next section we explore the use of suprasegmental information.

3. SPEAKER VERIFICATION USING SUPRASEGMENTAL FEATURES

3.1 Duration Information

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.

Figure 1 shows the warping path of an imposter speaker. The straight line is the regression line obtained by the least square fit of the points along the 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 $y'(k) = m\ x(k) + c$, 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

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

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 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 Eq.(1) 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 Table 1 (see Sl.No.2).

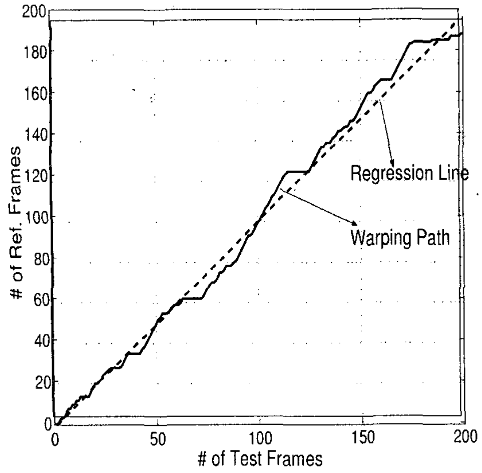

Figure 1. Regression line fit for the warping path of an impostor

3.2 Pitch information

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. 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

$$P_s = \sum_{i=1}^{L} |F_o(x(i)) - F_o(y(i))| \qquad (2)$$

where $F_o(x(i))$ = Pitch frequency of the frame $x(i)$ of the test utterance, $F_o(y(i))$ = Pitch frequency of the frame $y(i)$ of the reference utterance and 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 $F_o(x(i)) \neq 0$ and $F_o(y(i)) \neq 0$.

The speech database and the number of genuine and imposter speaker tests are same as described in the previous section. 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 Table 1 (see Sl.No.1).

These experiments show that the duration and the pitch contour gives useful speaker-specific information. Compared to the spectral features the relatively poor performance in case of suprasegmental features may be due to the large intra-speaker variability. Although the performance of the system is poor if duration or pitch alone is used for speaker verification, the performance improves significantly when combined with the evidence from spectral information. This is shown in the next section.

4. COMBINING EVIDENCE FROM SPECTRAL AND SUPRASEGMENTAL INFORMATION

Though the short-term spectral features give a strong evidence for speaker verification task, they are sensitive to the characteristics of the channel. The suprasegmental features like pitch and duration are known to be robust against channel distortion [4]. Studies have shown that the features and classifiers of different types may complement one another in giving better performance when used together [17]. Since the features derived separately from the spectral (segmental), duration and pitch (suprasegmental) information give nearly independent sources of evidence, they can be combined to improve the performance of a speaker verification system. In this section a method based on neural network models to combine the evidence obtained from different types of features is proposed.

The problem of accepting or rejecting the identity claim of the speaker may be viewed as a two-class pattern classification problem. A pattern classification problem involves determining the hypersurfaces separating the multidimensional patterns belonging to different classes [18, 19]. In case of a two-class problem, if the input dimension is one, then the hypersurface separating the two classes is a straight line. In case of 2-dimensional input pattern, the hypersurface is a plane. Similarly the hypersurface will be curved surface for any higher dimension. For instance, 2-D and 3-D plots showing the genuine and imposter scores obtained for a speaker using different features are shown in Figure 2(a) and 2(b). Before plotting all the scores are normalized in the range from -1 to +1. As it can be seen from the figure, there is clustering among the patterns belonging to each class.

The hypersurface separating the genuine and imposter scores may be captured using the Multilayer Perceptron (MLP) neural network [18, 19]. Multilayer perceptron, a feedforward neural network is known to be useful for pattern classification tasks [18, 19]. In particular a four layer network with the input layer consisting of linear units and the other three layers consisting of hard-limiting nonlinear units can perform any pattern classification task [18, 19].

There are 3 different features, namely, spectral, duration and pitch. We can obtain evidence from each of these features as described in the previous sections. There are 45 genuine trial scores and 261 imposter trial scores for each speaker per sentence. All the scores are normalized to the range -1 to +1. If we use 2 or 3 features, then correspondingly the genuine

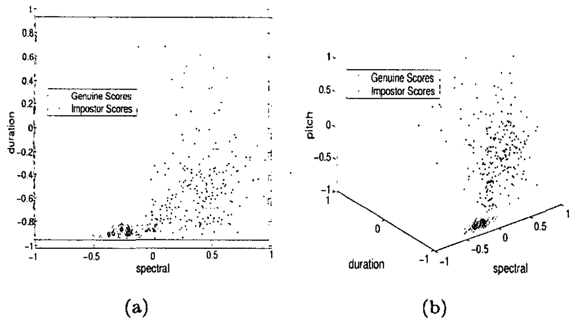

Figure 2. Distribution of the score vectors for genuine and impostor trials. (a) Spectral and duration information. (b) Spectral, duration and pitch information.

and imposter scores will be 2-D or 3-D respectively.

The objective is to develop a 2-class (genuine and imposter) classifier using the 2-D or 3-D score vectors, corresponding to the result of combining 2 or 3 features, respectively. In order to capture the dividing surface between the genuine and imposter classes, a MLP network is trained using 30 of the 45 genuine trail score vectors and 180 of the 261 imposter trial score vectors [18, 19]. The remaining 15 genuine trail score vectors and 81 imposter trial score vectors are used for evaluating the 2-class network classifier. The impostor examples used for testing are from different speakers than 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 2 and 3 features are respectively, $2L$ $4N$ $3N$ $1N$ and $3L$ $6N$ $3N$ $1N$, where L refers to linear units and N refers to nonlinear (activation function $tanh(.)$) unit, 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 imposter (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 10 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 2 and 3 features are combined are given in Table 1. We have obtained the EER values for the *small* test data using spectral information alone (see Sl.No.4). Note that these are not significantly different from the EER values obtained using the complete data (see Sl.No.1). But the EER values computed for spectral information from *small* test data 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 3 features are used. Another interesting point to be noted is that the degradation in performance is not significant due to channel variations, when evidence from multiple sources of information are combined.

Table 1. Performance of the text-dependent speaker verification system with microphone and telephone speech. The scores for telephone channel are shown inside parentheses (.).

Sl. No.	Feature Used	Equal Error Rate
1	pitch information	8.67 (8.44)
2	duration information	7.79 (7.50)
3	spectral information	2.54 (2.77)
4	spectral information (*small* test data)	2.60 (2.89)
6	spectral+duration	1.14 (1.28)
7	spectral+duration+pitch	0.75 (0.84)

5. CONCLUSIONS

Most present day systems for speaker verification use the information about the characteristics of the vocal tract, which are reflected in the spectral features. Features from spectral, duration and pitch provide evidence from independent sources of information. The evidence from the different sources were combined using a multilayer perceptron neural network. It was shown that not only that the performance of verification improved, but also the nonspectral features such as duration and pitch were found to be robust for variations due to channel. That is the reason why there is not much difference in the performance for telephone channel.

The excitation source of speech production is also known to contain information about the speaker. Hence methods may be developed to extract speaker-specific source information and use it as an additional evidence for further improving the performance of the speaker verification system.

REFERENCES

[1] D. O'Shaughnessy, Speech Communication: Human and Machine. New York: Addison-Wesley, 1987.

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

[3] N. Malayath, H. Hermansky, S. Kajarekar, and B. Yegnanarayana, "Data-driven temporal filters and alternatives to GMM in speaker verification," Digital Signal Processing, vol. 10, pp. 55–74, Jan. 2000.

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

[5] 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.

[6] 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, Australia), Nov.-Dec. 1998.

[7] 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. II, (Beijing, China), pp. 302–305, Oct. 2000.

[8] G. Doddington, "Speaker reognition based on idiolectal differences between speakers," in Proc. European Conf. Speech Processing, Technology, (Aalborg, Denmark), pp. 2521–2524, Sept. 2001.

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

[10] 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, Signal Processing, vol. IV, (Hong Kong), pp. 784–787, Apr. 2003.

[11] 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, Signal Processing, vol. IV, (Hong Kong), pp. 784–787, Apr. 2003.

[12] J. M. Zachariah, Text-dependent speaker verification using segmental, suprasegmental and source features. MS thesis, Indian Institute of Technology Madras, Department of Computer Science and Engg., Chennai, India, 2002.

[13] S. R. M. Prasanna, J. M. Zachariah, and B.Yegnanarayana, "Begin-end detection using vowel onset points," in Proc. Workshop on Spoken Language Processing, (Tata Institute of Fundamental Research, Mumbai, India), pp. 33–40, Jan. 2003.

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

[15] S. Furui, "Cepstral analysis technique for automatic speaker verification," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-29, pp. 254–272, Apr. 1981.

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

[17] J. J. Hull and S. N. Srihari, "Decision combination in multiple classifier systems," IEEE Trans. Pattern Analysis, Machine Intelligence, vol. 16, pp. 66–75, Jan. 1994.

[18] B. Yegnanarayana, Artificial Neural Networks. New Delhi: Prentice–Hall India, 1999.

[19] S. Haykin, Neural networks: A comprehensive foundation. New Jersey: Prentice-Hall Inc., 1999.