AANN Models for Speaker Recognition Based on Difference Cepstrals

Guruprasad. S, Dhananjaya. N and B. Yegnanarayana
Speech and Vision Laboratory
Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai - 600 036 India
Email: {guru, dhanu, yegna}@cs.iitm.ernet.in

Abstract - *This paper presents a novel method for representing speaker characteristics present in the speech signal, by the way of deemphasizing the linguistic content of the signal. Cepstral coefficients that are widely employed as features for automatic speaker recognition task, contain considerable speech information in addition to the speaker information, and hence do not highlight the latter. The proposed method is based on using the difference between all-pole spectra due to higher order and lower order of linear prediction analysis. Distribution of the feature vectors in the multi-dimensional feature space is captured by employing autoassociative neural network models. A speaker recognition system is developed using the proposed method of feature extraction, whose performance is evaluated against that of the system based on cepstral coefficients. The complementary nature of evidence due to the proposed feature is also examined, so as to improve the overall system performance.*

I. Introduction

The objective of automatic speaker recognition task is to recognize a person solely based on his/her voice, by a machine. Speech is a composite signal that primarily carries with it the message to be conveyed and speaker information. Speech signal conveys the message through a sequence of sound units which are produced by exciting the time varying vocal tract system with a time varying input. Speaker-specific variations in speech signal are partly due to the anatomical differences in speech-producing organs, and partly due to idiosyncrasies of the speaker, such as speaking habits and emotional state. Both forms of variation are of importance to automatic speaker recognition task, which entails the following three steps in the main, when approached from a pattern recognition perspective:

1. Representation of speaker-specific properties of speech signal and their efficient measurement, collectively known as feature extraction
2. Modeling the speaker characteristics represented by the feature vectors
3. Decision mechanism for verification/identification, based on the response of the speaker model for features in the test utterance

Speaker characteristics are manifested in both the components of speech production mechanism, namely, the excitation source and the vocal tract system. Variations in shape of the vocal tract are captured in the form of resonances, antiresonances and spectral roll-off characteristics, while the excitation source is characterized by voice pitch, glottal vibrations and suprasegmental features such as intonation, duration, stress and coarticulation. Features extracted from short-time analysis techniques provide reasonable approximation to both the components of speech production mechanism and are relatively easy to extract. Hence, they are widely used for speaker recognition, unlike suprasegmental features that are difficult to characterize and represent with existing techniques.

Short-time (10-30 ms) analysis of quasi-stationary regions of speech is performed to approximate the spectral response of the vocal tract, either by directly computing the Discrete Fourier Transform (DFT) spectrum, or using linear prediction (LP) analysis which provides an all-pole approximation of the spectral envelope. Cepstral coefficients derived from the DFT spectrum or the all-pole LP spectrum are widely used as features for speaker recognition [1] [2]. However, these features contain information about the speaker as well as the linguistic message, and hence do not highlight speaker-dependent properties.

Methods based on speaker-specific mapping of features have been attempted to capture speaker-specific information, by mapping a set of features that characterize linguistic content, to a set of features having both linguistic and speaker-specific information. A mapping from a low (7) order perceptual linear predictive (PLP) model to a high (14) order PLP model is discussed in [3]. Another experiment employed feedforward neural networks to capture the mapping between cepstrals derived from low (6) order LP analysis, and those derived from high (12) order LP analysis [4]. The objective of the present study is to propose a feature that highlights the speaker-dependent information contained in the signal by deemphasizing information specific to the speech sound. The feature is based on the difference between LP spectra due to higher and lower order analysis.

The paper is organized as follows. Section II provides an overview of LP analysis of speech, while Section III proposes a feature for highlighting the speaker characteristics present in the speech signal. Section IV describes autoassociative neural network (AANN) models for capturing the distribution of feature vectors. Section V outlines a speaker verification system based on the proposed feature, while Section VI discusses the performance of the system. The study is summarized in Section VII.

II. Linear Prediction Analysis of Speech

Linear prediction analysis of speech signal predicts a given speech sample at time n as a linear weighted sum of the previous p samples,

$$\hat{s}(n) = \sum_{k=1}^{p} a_k s(n-k)$$

where $\hat{s}(n)$ is the predicted sample at time at time n, $s(n)$ is the speech sample at time n, and $a_k, k = 1, 2, ...p$ are the predictor coefficients [5].
The prediction error $e(n)$ is defined as

$$e(n) = s(n) - \hat{s}(n).$$

The mean square of the prediction error over an analysis frame of N samples is given by

$$E = \sum_{n=0}^{N-1} e^2(n)$$

Minimizing E with respect to the set of predictor coefficients $\{a_k\}$ results in a set of p normal equations. The predictor coefficients $\{a_k\}$ are obtained by solving this set of p normal equations.

The vocal tract system can be modeled as an all-pole filter whose spectral response is described by the set of predictor coefficients $\{a_k\}$. The prediction order p has significant bearing on the ability of the all-pole filter to closely approximate the short-time spectrum of the speech segment. For larger values of p, from 16 to 30, the LP model tries to match spurious spectral peaks of the speech signal, like the individual pitch harmonics.

An important parameter set that can be derived from the LP coefficients is the set of cepstral coefficients [6]. Cepstral coefficients provide a compact representation of the resonances and the spectral roll-off characteristics of the vocal tract system. The set of cepstral coefficients $\{c_k\}$ is obtained from the set of predictor coefficients $\{a_k\}$, using the following recursive relation:

$$\begin{aligned} c_0 &= logE_{min} \\ c_k &= -a_k + \sum_{j=1}^{k-1} \frac{j}{k} c_j a_{k-j} && 1 \le k \le p \\ c_k &= \sum_{j=k-p}^{k-1} \frac{j}{k} c_j a_{k-j} && p < k < l \end{aligned}$$

where l is the number of cepstral coefficients, and E_{min} is minimum mean squared prediction error.

III. A Feature Based on the Difference Spectrum

The short-time spectrum of speech for a voiced speech sound has two components: harmonic peaks due to periodicity of voiced speech, and gross envelope of the spectrum that reflects the vocal tract response and glottal-pulse shape [7]. The periodicity of voiced speech is due to the excitation source, that is a characteristic of the speaker. The spectral envelope is shaped by formants, that reflect the resonances of the vocal tract. Formant locations and bandwidths show considerable variation for different speakers, for a given category of sound unit [8]. This is due to the varying vocal tract shapes and lengths for different speakers. This variation may be more pronounced in the finer fluctutations in the spectral envelope, indicating that speaker-dependent properties are relatively better manifested in these variations. Also, a smoothed spectral envelope is still intelligible, implying that the smoothed spectral envelope is more representative of linguistic information than the speaker information.

The order of LP analysis determines how closely the resultant all-pole spectrum matches the short-time spectrum of the signal. The all-pole spectrum obtained by a lower order of LP analysis such as 6, approximates only the prominent formants, while that due to a higher order LP analysis such as 12 or 14, matches all the formants and other spectral envelope information. This can be observed from Fig. 1 which shows the LP log-spectra for two different orders of LP analysis along with the short-time spectrum, for a voiced region of speech signal. A difference of the above two all-pole spectra can be expected to deemphasize the linguistic information of the sound unit, and yet preserve the variations that contain speaker-dependent information.

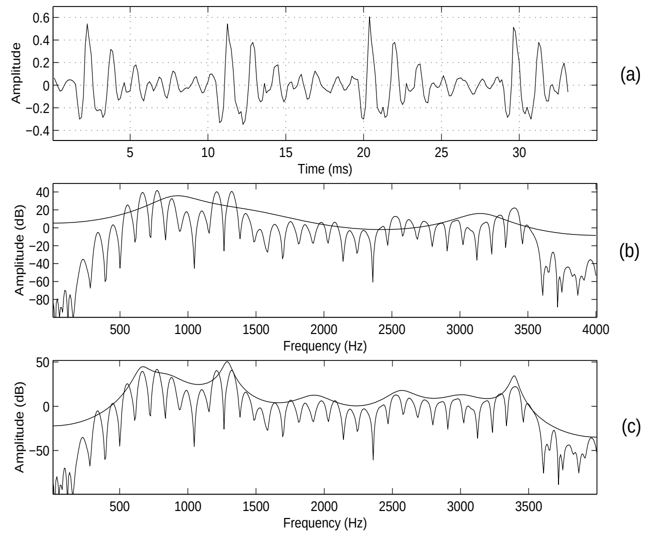

Fig. 1. LP log-spectra for two different orders of prediction. (a) A voiced region of speech. (b) Short-time spectrum and LP log-spectrum for LP order $p = 6$. (c) Short-time spectrum and LP log-spectrum for $p = 14$.

The weighted difference cepstral coefficients d_k can be expressed as

$$d_k \quad = \quad k(c_k^h - c_k^l) \qquad\qquad 0 < k < m$$

where c_k^h are the cepstral coefficients due to a higher order of LP analysis, c_k^l are the coefficients due to a lower order of LP analysis, and m is the number of cepstral coefficients.

The differencing of the cepstra also reduces the influence of the transmission channel characteristics on the speech signal. This obviates the need for cepstral mean subtraction, that is normally employed to remove the mean of the time trajectory of each cepstral coefficient [1] [2]. However, the comparable range of amplitudes of the cepstral coefficients of the two spectra increases noise content in the difference cepstrum. Hence, the weighted difference cepstral coefficients d_k are averaged over a window of w contiguous frames of a region of voiced speech, as follows:

$$\hat{d}_{k,j} \quad = \quad \frac{1}{w}\sum_{i=j-\frac{w}{2}}^{j+\frac{w}{2}} d_{k,i} \qquad\qquad 0 < k < l,$$

where $\hat{d}_{k,j}$ are the averaged weighted difference cepstral coefficients for segment j of the region of voiced speech, and $d_{k,i}$ are the weighted difference cepstral coefficients for frame i.

IV. AANN Models for Capturing the Distribution of Feature Vectors

Autoassociative neural network models are feedforward neural networks that perform an identity mapping of the input space [9]. The desired output for AANNs is same as the input vector, and hence the input and output layers have the same number of units. The number of hidden layers and the number of units in each hidden layer depend on the problem. Typically, any hidden layer with number of input units less than the dimension of the input vector results in compression of the input vector to a lower dimension. For instance, a three-layer AANN model with linear units can capture the principal components of a feature set in the feature space. For such a network, it can be shown that the mean square of the error between the input and output patterns is minimized by choosing those weights that correspond to principal vectors of covariance matrix of the input data [9].

Fig. 2 shows a five-layer AANN model that performs nonlinear principal component analysis.

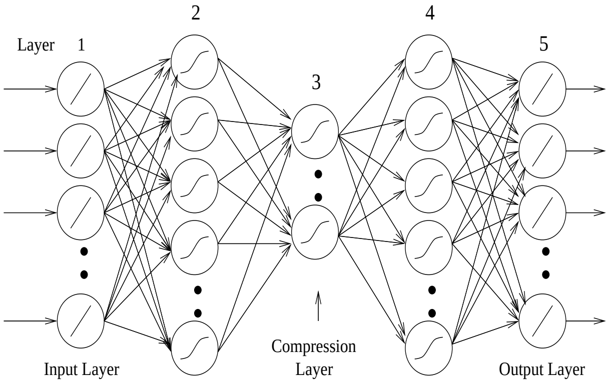

Fig. 2. A five-layer AANN Model.

The second and the fourth layer have nonlinear elements, and these layers have more elements than the first and the fifth layer. The third layer, with P units, serves as the compression layer. The first three layers form a mapping network, that projects the input space R^M on to a subspace R^P, where M is the dimension of the input vector, and $P < M$. The mapping is nonlinear, and a nonlinear subspace is formed at the third layer. The last three layers form a demapping network that projects the nonlinear subspace R^P back on to the input space R^M. The nonlinear subspaces captured by the network remain similar for different training sessions [10].

The ability of AANN models to capture nonlinear subspaces has been demonstrated in [10]. The authors examine the possibility of designing an AANN to capture the characteristics of distribution of the data. They study the importance of error surface of the training data in the feature space. It was observed that the average error was lower for the most frequently occurring input vectors than for the less frequently occurring ones. It was demonstrated experimentally that a network can be designed such that the training error surface relates to the distribution of the given data, depending on the constraints imposed by the structure of the network. When the surface representing the distribution of features is highly non-linear in the multi-dimensional space, AANN models have a distinct advantage over Gaussian mixture models (GMMs), which are constrained by the fact that the shape of the components of the distribution is assumed to be Gaussian, and that the number of mixtures are generally fixed a priori.

V. A Speaker Verification System Based on the Difference Cepstrum

The verification system involves extraction of features from training and testing data, building AANN models for speakers and testing each utterance against a certain number of claimant models to detect the identity of the speaker of that utterance from among the claimants.

Speech corpus used in this study is a subset of the cellular database of NIST 2002 speaker recognition evaluation [11]. There are 139 male speakers, and the duration of training data for each speaker is about 2 minutes. The present study uses 300 test utterances, each having a duration of about half a minute. Each test utterance has 11 claimants, where the genuine speaker may or may not be one of the claimants. All speech was sampled at 8 kHz.

Speech signal is pre-emphasized using a difference operator. Hamming window is applied on frames of 20 ms of differenced speech, with a shift of 5 ms. An amplitude threshold is used to mark silence frames. LP analysis is performed on the non-silence frames, using lower and higher orders of analysis. Weighted difference cepstral coefficients are then obtained and smoothed over 5 consecutive frames within the given non-silence region.

For each speaker, two sets of AANN models are trained, one using difference cepstrals, and another using weighted LP cepstrals. Difference cepstrals of three different dimensions, namely, 10, 14 and 19 are used to train three AANN models, whose structures are, *10L 20N 3N 20N 10L*, *14L 30N 4N 30N 14L* and *19L 38N 4N 38N 19L*, respectively. Here L refers to a linear unit, and N refers to a nonlinear unit. The second set of AANN models is trained on 19-dimensional weighted LP cepstrals, derived from LP analysis of orders 12 and 14. The structure of these AANN models is *19L 38N 4N 38N 19L*. The choice of above network structures is based on a systematic study reported in [12]. Error backpropagation algorithm is used to update the weights of the network [13].

Features extracted from each test utterance are fed to the corresponding 11 claimant models. The confidence score of a model is defined as

$$C = \frac{1}{N}\sum_{i=1}^{N} \exp(-D_i), \qquad D_i = \|\mathbf{x}_i - \mathbf{y}_i\|^2$$

where $\mathbf{x}_i$ is the input vector to the model, $\mathbf{y}_i$ is the output of the model, and N is the number of feature vectors of the test utterance. Speaker model normalization (ZNorm) and test utterance normalization (TNorm) are performed on the raw confidence scores. For TNorm, 20 speakers are chosen for developing background models from NIST 2001 cellular development data set.

VI. Results and Discussion

The performance of speaker recognition systems is commonly evaluated in terms of equal error rate (EER), which can be measured from the Detection Error Trade-off (DET) curves [11]. Fig. 3 compares the performance of speaker recognition systems based on cepstrals with that of systems based on difference cepstrals. The EER of systems based on difference cepstrals is 19.1% and 19.5%, for the cases (12,6) and (14,6) respectively, as compared to 15.1% and 15.9% of the systems based on cepstrals derived from LP orders of 14 and 12, respectively. Here, the ordered pair indicates the high and low orders of LP analysis used to obtain the difference cepstrals. The lower performance of the difference cepstrals can be attributed to their high noise content resulting from differencing. The performance was also analyzed by examining whether the highest scorer (winner) among the 11 claimants for each test utterance was indeed the actual speaker (genuine winner) of that test utterance. The analysis was done on two systems: system1 based on weighted cepstrals derived from the LP analysis of order 14, and system2, based the on difference cepstrals for the case (14,6). Of the 300 test utterances, the actual speaker is one of the claimants in only 251 cases. It was observed that 139 genuine winners were common to both the systems, while 28 genuine winners were specific to system1, and 30 to system2. This clearly indicates some complementary evidence represented by the difference cepstrals, an observation that is not evident from the DET curves of Fig. 3.

Fig. 4 shows the DET curves for systems based on difference cepstrals, for the cases (14,6), (14,8) and (14,10). As the lower order of LP analysis increases, more speaker characteristics are captured by the cepstrals of

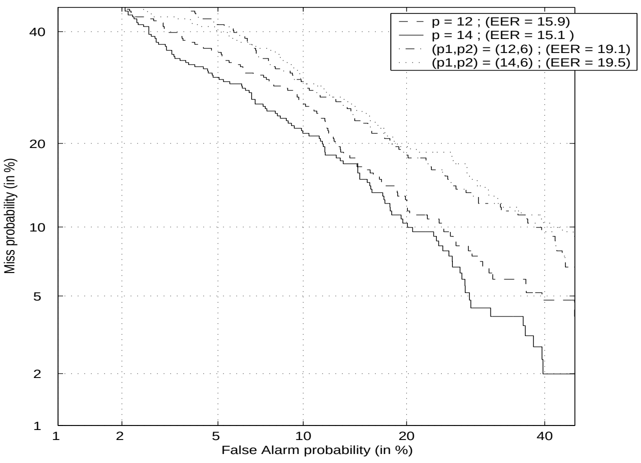

Fig. 3. DET curves to compare systems based on cepstrals and difference cepstrals. p_1, p_2 denote higher and lower orders of LP analysis, respectively.

lower LP order in addition to speech information, leading to a decline in the performance of difference cepstrals.

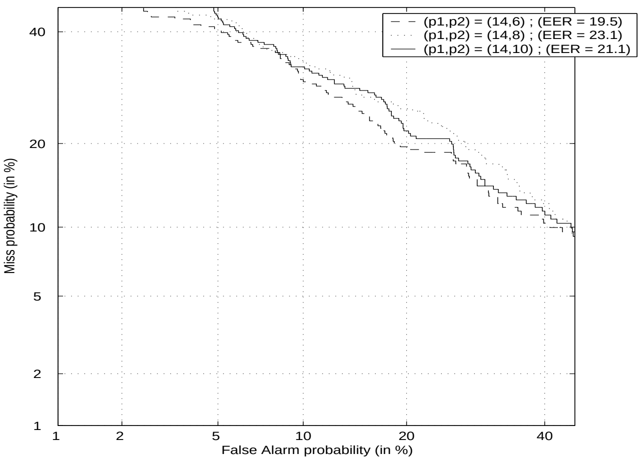

Fig. 4. DET curves for difference cepstrals for varying lower LP orders and a fixed higher LP order of 14.

The DET curves for a system based on difference cepstrals for the case (14,6) are plotted in Fig. 5, for varying dimensions of the feature vector. The 14 and 10 dimensional feature vectors are obtained by truncating the 19 dimensional feature vectors. Evidently, the decline in performance is marginal even when only 10 dimensions of the difference cepstrals are used. This result is important from a dimension reduction perspective, that simplifies the structure of the AANN model, and effectively reduces the number of parameters for characterizing a speaker.

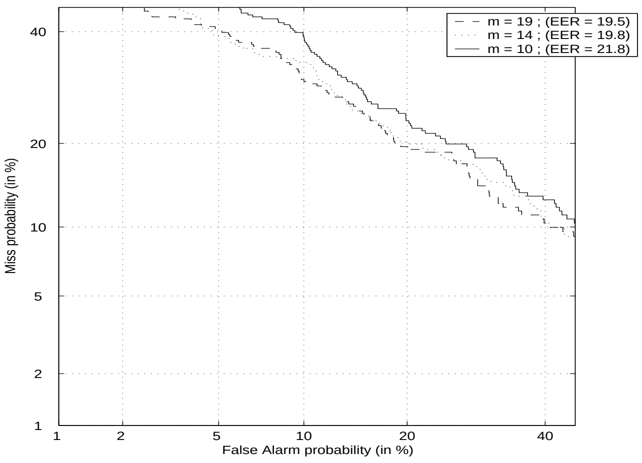

Fig. 5. DET curves for difference cepstrals for varying dimension m of the feature vector, for the case (14,6).

VII. Conclusions

This paper proposes a feature based on the difference between higher order and lower order LP spectra, with the objective of deemphasizing speech information, and thereby highlighting speaker-dependent information. The difference cepstrals represent speaker-information that may provide some complementary information to that represented by the weighted cepstrals. Thus the evidences may be combined to enhance the system performance. Also, a reduction in the dimension of the input vectors is possible without significantly affecting the recognition performance of the system. Better scoring techniques for decision making and algorithms for combining evidences need to be explored.

References

[1] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 29, no. 2, pp. 254–272, Apr. 1981.

[2] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *Journal of the Acoustical Society of America*, vol. 55, pp. 1304–1312, Jun. 1974.

[3] H. Hermansky and N. Malaynath, "Speaker verification using speaker-specific mapping," in *Speaker Recognition and its Commercial and Forensic Applications*, France, May 1998.

[4] M. Shajith Ikbal H. Misra and B. Yegnanarayana, "Speaker-specific mapping for text-independent speaker recognition," *Speech Communication*, vol. 39, no. 3-4, pp. 301–310, Feb. 2003.

[5] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, pp. 561–580, April 1975.

[6] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs and N.J., 1993.

[7] B. S. Atal, "Automatic recognition of speakers from their

voices," *Proceedings of the IEEE*, vol. 64, pp. 460–475, April 1976.

[8] G. E. Peterson and H. L. Barney, "Control methods used in a study of vowels," *Journal of the Acoustical Society of America*, vol. 24, no. 2, pp. 175–194, Mar. 1952.

[9] B.Yegnanarayana, *Artificial Neural Networks*, Prentice-Hall of India, New Delhi, 1999.

[10] B. Yegnanarayana and S. P. Kishore, "AANN: an alternative to GMM for pattern recognition," *Neural Networks*, vol. 15, no. 3, pp. 459–469, May 2002.

[11] "Speaker recognition workshop notebook," in *Speaker Recognition Workshop*, Vienna, Virginia USA, May 2002.

[12] S. P. Kishore, *Speaker Verification Using Autoassociative Neural Network Models*, MS thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras, Dec. 2000.

[13] Simon Haykin, *Neural networks: A Comprehensive Foundation*, Prentice-Hall International, New Jersey, 1999.