Proceedings of
ICISIP - 2005

Neural Network Models for Extracting Complementary Speaker-Specific Information from Residual Phase

Sri Rama Murty Kodukula [1], S. R. Mahadeva Prasanna [2], B. Yegnanarayana [1]
[1] *Speech and Vision Laboratory*
Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai-600036, India
{ksrm, yegna}@cs.iitm.ernet.in
[2] *Dept. of Electronics and Communication Engineering*
Indian Institute of Technology Guwahati, Guwahati-781039, India
prasanna@iitg.ernet.in

Abstract

In this paper using neural network models we demonstrate the presence of complementary speaker-specific information in the residual phase as compared to the conventional spectral features. The spectral features mainly represent the speaker-specific vocal tract system features. The proposed LP residual phase represents the speaker-specific excitation source information. Speaker recognition studies are conducted using NIST 2003 speaker recognition evaluation database. The speaker recognition system using only spectral features gives an Equal Error Rate (EER) of 15.5% and using only LP residual phase information gives an EER of 22.0%. However, combining the evidences from LP residual phase and spectral features increases the performance to an EER of 13.5%. This result clearly demonstrates the complementary nature of speaker-specific information present in the LP residual phase.

1. Introduction

Speaker recognition is the task of recognizing the speaker from his speech signal [1]. Speaker recognition can be either identification or verification. In speaker identification the goal is to identify the speaker of the speech signal from a given set of speakers. Speaker verification involves validating the identity claim of a given speaker. Further whether the speech for the same or different text is used during training and testing, we have text-dependent or text-independent categories. This study focuses on text-independent speaker verification task.

Speech is produced from a time-varying vocal tract system by a time-varying excitation source [2]. The vocal tract system and the excitation source contain speaker-specific information. Spectral features like Mel-Frequency Cepstral Coefficients (MFCC), Linear Prediction Cepstral Coefficients (LPCC), which mainly represent the vocal tract system features have been well exploited for speaker verification studies [3], [4]. Excitation source information like Linear Prediction (LP) residual and LP residual phase have been shown to contain speaker-specific information [5], [6]. Eventhough excitation source information contain significant speaker-specific information, the difficulty in extracting features make them less preferable compared to the existing spectral features. However, the fact that the evidences from excitation source and vocal tract system are from two independent sources, it will be interesting to verify further whether they contain some complementary speaker-specific information. This factor will be useful for improving the performance of state of the art speaker recognition systems using only spectral features [4] and hence the objective of this paper.

Linear Prediction (LP) analysis is performed on the speech signal to separate the vocal tract system (LP Coefficients) and excitation source information (LP residual). The LP residual is processed further using Hilbert transform relations to derive the phase information [7]. In LP analysis autocorrelation analysis is performed to estimate the Linear Prediction Coefficients (LPCs) and hence relations among the samples up to second order are removed in the LP residual. Therefore speaker-specific information in the LP residual phase is present in the higher order relations among the samples. Distribution of moments taken on these higher order relations may not give information about a speaker. Hence extraction of speaker-specific information from such relations involves nonlinear operation. In this work AutoAssociative Neural Network (AANN) models are used for extracting the speaker-specific information from the LP residual phase.

This paper is organized as follows: The extraction of residual phase information from the speech signal is discussed in Section 2. AANN models for extracting the speaker-specific information are discussed in Section 3. The speaker recognition studies are discussed in Section 4. The conclusions of this study and the scope for future work are given in Section 5.

0-7803-8840-2/05/$20.00 ©2005 IEEE

2. Speaker-Specific Information

The residual phase information can be extracted from the speech signal by LP analysis [8]. In LP analysis each sample is predicted as a linear combination of past p samples, where p is the order of prediction. The predicted sample $s(n)$ is given by

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

where $\{s(n)\}$ are the speech samples and $\{a_k\}$ are the LPCs. The LPCs are obtained as a process of minimizing the error between the actual and the predicted samples. This is achieved by solving the following set of normal equations:

$$\sum_{k=1}^{p} a_k R(n-k) = -R(n), \qquad n = 1,....p \quad (2)$$

where $R(m) = \sum_n s(n)s(n-m)$ is the autocorrelation function.

The error between the actual samples and their predicted versions is termed as LP residual and is given by

$$r(n) = s(n) - \hat{s}(n) \quad (3)$$

$$r(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k) \quad (4)$$

A 20 ms frame of speech signal, its LP spectrum obtained using the LPCs and the LP residual are shown in Fig. 1.

The features from the LP spectrum and the LP residual may be used independently for speaker recognition studies [5].

The Hilbert transform of the LP residual $r_h(n)$ is the 90^o phase shifted version of the LP residual and is obtained using the relations

$$r_h(n) = \left\{ \begin{array}{ll} IDFT[-jR(\omega)], & 0 < \omega < \pi \\ IDFT[jR(\omega)], & 0 > \omega > -\pi \\ 0 & \omega = 0, \pi \end{array} \right\} \quad (5)$$

where $R(\omega)$ is the Discrete Fourier Transform (DFT) of $r(n)$. The magnitude of the complex time signal constructed from the LP residual and its Hilbert transform is termed as Hilbert envelope and is given by

$$h_e(n) = \sqrt{r^2(n) + r_h^2(n)} \quad (6)$$

The residual phase information ($\sin\theta$) is obtained as the ratio of LP residual to the Hilbert envelope.

$$\sin\theta = r(n)/h_e(n) \quad (7)$$

A segment of speech signal, its LP residual, Hilbert transform of LP residual, Hilbert envelope of the LP residual and the residual phase information ($\sin\theta$) are shown in Fig. 2. As it can be observed, the phase information looks like a noise-like signal and hence it is difficult make out speaker information visually. For comparison of residual phase across different speakers, LP residual phase extracted for the same sound unit /a/ for five different male speakers are shown in Fig. 3

As it can be observed, it is difficult to visually make out any discriminatory features across different speakers. However, it

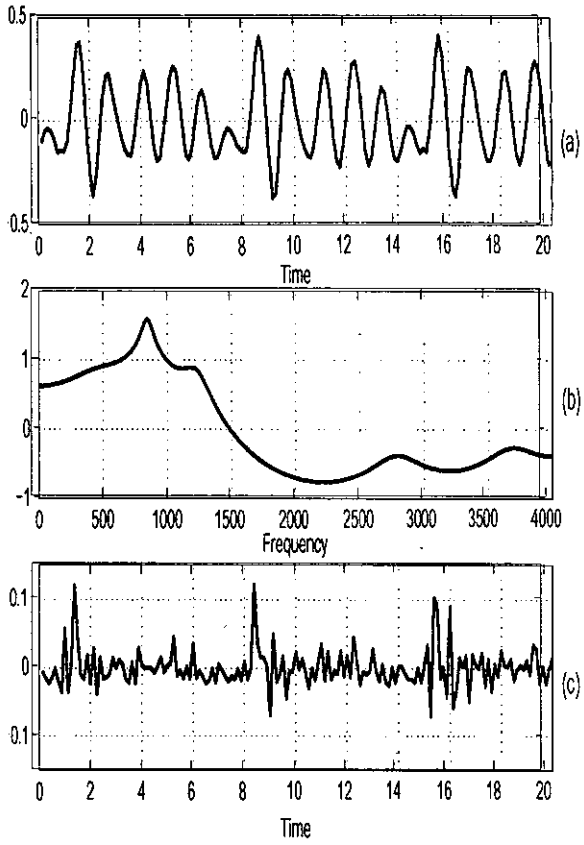

Fig. 1: (a)Speech signal, (b)LP spectrum, (c)Residual signal

is interesting to note that the speaker information is present in the higher order relations among the samples of the residual phase information [7], [9].

The LP residual phase signal mainly contains the information about the phase relations among the samples of the LP residual. In the LP residual, samples around the Glottal Closure (GC) events are high Signal-to-Noise Ratio (SNR) regions and are known to contain better speaker-specific information [10]. In a similar way, the phase information around the samples of the GC events in the LP residual phase contains better speaker-specific information compared to other places [9]. Hence the knowledge of GC events is used for extracting the phase relations for speaker recognition studies. The difference between the LP residual and LP residual phase information is that the strength of the excitation around the GC event present in the LP residual is eliminated in the LP residual phase information. Thus in the LP residual phase, speaker information is present only in the sequence of the samples.

3. Neural Network Models for Speaker Verification

Since LP analysis extracts the second order statistical features through the autocorrelation matrix, the LP residual phase

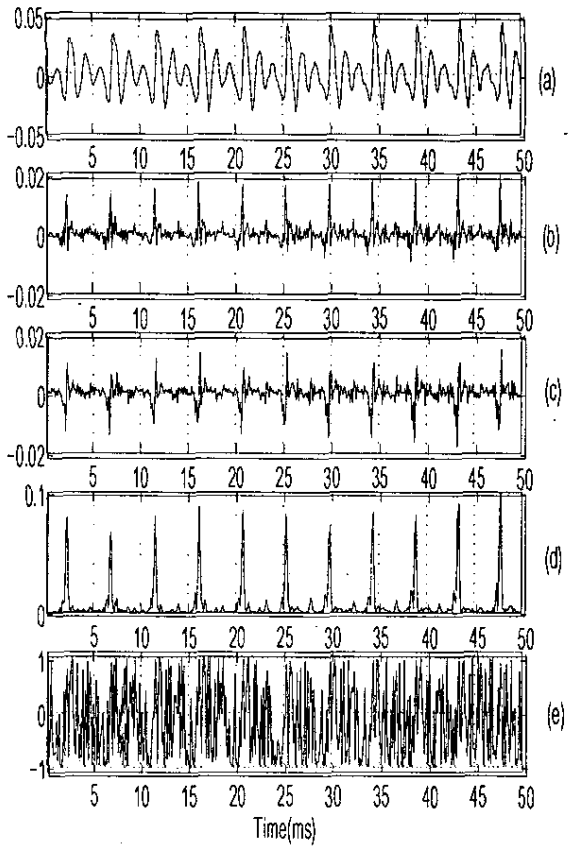

Fig. 2: (a)Speech signal, (b)LP Residual, (c)Hilbert Transform, (d)Hilbert envelope, (e)Residual phase ($sin(\theta)$)

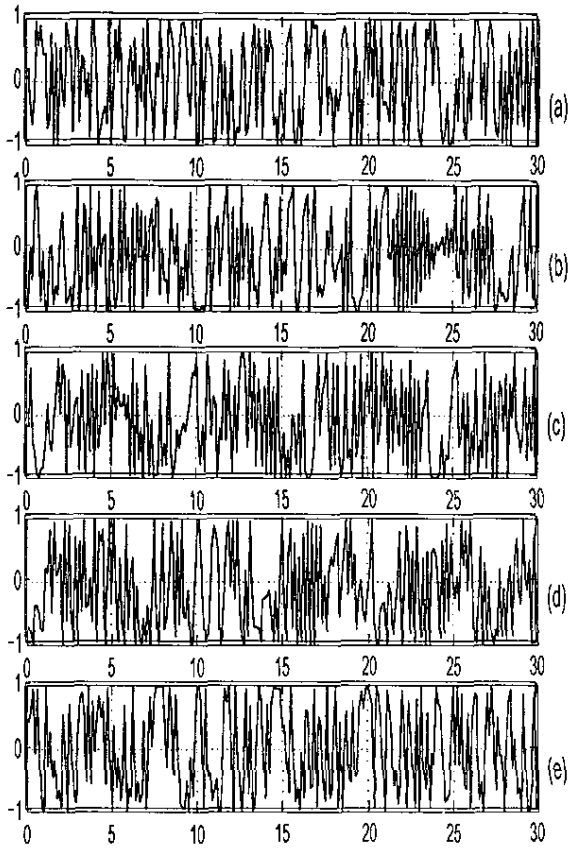

Fig. 3: Residual phase for five different speakers in the steady region of vowel /a/

does not contain any significant second order features corresponding to the shape of the vocal tract. That is why the autocorrelation function of the LP residual phase signal has low correlation values for nonzero time lags, like for a whitening process. We conjecture that the speaker-specific information may be present in some higher order relations among the samples of the residual phase signal. It is not clear how this information can be extracted from the residual phase signal. Statistical features like higher order moments of the distribution of the samples of residual phase do not seem to capture the desired speaker-specific information. It is conjectured that the extraction of such an information may involve nonlinear processing. Neural network models can be trained to capture the nonlinear information present in the signal. We explore these models. In particular, we propose AANN models to extract the desired information from the residual samples.

AANN models are basically Feed Forward Neural Network (FFNN) models which try to map an input vector onto itself, and hence the name autoassociation or identity mapping. It consists of an input layer, an output layer and one or more hidden layers. The number of units in the input and output layers are equal to the size of the input vectors. The number of nodes in the middle hidden layer is less than the number of units in the input or output layers. The middle layer is also the dimension compression hidden layer. The activation function of the units in the input and output layers are linear, whereas the activation function of the units in hidden layer can be either linear or nonlinear.

AANN models can capture the distribution of the input data, if the data is a set of feature vectors in the feature space [5]. However, when an AANN is presented with raw PCM signal samples, such as samples of speech or LP residual or LP residual phase signal, the AANN captures the implicit nonlinear (higher order) relations among the samples. Therefore the behavior of the AANN depends on the type of input given to the network.

The speech signal contains both the second (autocorrelation) and higher order relations among the samples. If the speech signal itself is given to the AANN, then the dominant second order correlations among the samples will be captured in the training of the network. When the second order correlations are removed from the speech signal through the LP analysis, and the resulting LP residual phase is used as input to the AANN, then the implicit higher order relations in the LP residual

phase signal are captured. Later we show experimentally that these relations do correspond to the desired speaker-specific information in the excitation component.

When the input to an AANN consists of samples of random noise, then the network weights will not converge. On the other hand if blocks of speech samples or LP residual phase samples are given as input, the error between the input (also the desired output) and the actual output is reduced during training, indicating that there is some relation among the samples. As the number of LP residual phase samples per block is increased, then the relations over longer length of the block are capture ed. But, if the length of the block exceeds a pitch period, then the effect of pitch period also influences the training of the network. Therefore in this study the number of samples per block are limited to less than a pitch period. If the number of units in the dimension compression layer is large, then too many details in the input data may be captured, and these details may not be consistent across several blocks. If the number of units in the compression layer is very small (4 or 5), then important speaker-specific information may be missing. The training error is an indication of the minimum number of units required in the compression layer. Typically the training error reaches a low value when the number of units in the compression layer are increased to about 12, and thereafter the error does not significantly reduce even if the number of units are increased. Note that a lower number is preferable as it reduces the size (in terms of the weights) of the network.

4. Speaker Recognition Studies

A. Database for the Study

All the speaker recognition experiments of this study are conducted on NIST 2003 speaker recognition evaluation database of male speakers [11]. Our objective is only to demonstrate the complementary nature of the speaker-specific information in the residual phase information and hence only male speakers part of the database was chosen for the study. There are 149 male speakers, and the duration of training data for each speaker is about 2 minutes. There are 1343 test utterances, each having a duration of 15-45 sec duration. Each test utterance has 11 claimants, where genuine speaker may or may not be present. All speech signals were sampled at 8 kHz.

B. Studies using LP Residual Phase Information

The LP residual phase information ($\sin\theta$) is extracted from the speech signals as explained earlier. Only voiced frames are selected for the study using a method based on the autocorrelation of the Hilbert envelope. GC events are also detected using a method based on the Hilbert envelope. LP residual phase information around the GC events is considered in blocks of 40 samples for the study. The structure of the network used for the study is shown in Fig. 4 The structure of the network is 40*L* 48*N* 12*N* 48*N* 40*L*, where, *L* represents linear, *N* represents nonlinear and the numerals represent number of units in the layer. The structure of the network was

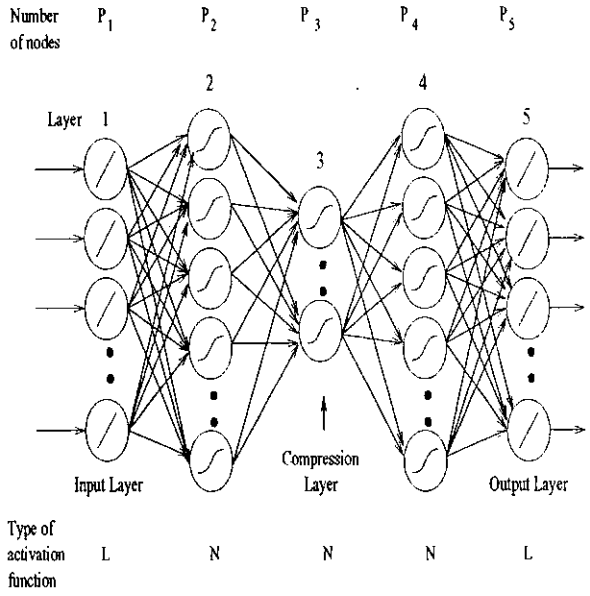

Fig. 4: Structure of the AANN model

determined experimentally. The performance does not depend critically on the structure of the network.

During training phase, 6 blocks of 40 samples around each GC event are considered in shifts of one sample. Each block is applied to input as well as output of the AANN model and was allowed to learn the higher order relations among the samples in the block. One AANN model was trained for each speaker for 500 epochs.

During testing phase, 6 blocks of 40 samples around each GC event are considered in shifts of one sample. Each block is applied as input to the AANN and the output of the AANN is noted. The error between the input block and the output of the AANN is computed. The error is converted into a confidence value using the relation $c_i = exp(-\lambda e_i)$ where e_i and c_i are error and confidence of block i, respectively, and $\lambda = 1$ throughout the study. The average confidence of all the blocks in the test utterance gives the score of the speaker for the given test utterance. The performance of the speaker recognition system is given as Detection Estimation (DET) curve and is shown in Fig. 5 From the DET curve the EER is found to be 22%.

C. Studies using Spectral Features

The Linear Prediction Cepstral Coefficients (LPCCs) derived from the LPCs are used as the spectral features for the speaker verification study. Voiced frames are detected as mentioned earlier. 19 dimension LPCCs are computed for each frame of 20 ms with a shift of 5 ms. The structure of AANN for capturing distribution of LPCCs of each speaker is 19*L* 38*N* 8*N* 38*N* 19*L*, where *L* refers to linear nodes and *N* refers to nonlinear nodes.

During training, the LPCCs are fed in random order to the AANN and one AANN is trained for each speaker for 60 epochs. We found the performance obtained for AANNs

trained with 60 epochs is almost equal to the AANNs trained with 500 epochs, but with significant reduction in time. Hence we used only 60 epochs for training AANN models using LPCCs.

During testing, LPCCs extracted from the test utterance for every block of 20 ms with a shift of 5 ms are applied to the AANN models. For each block of 20 ms the error between the LPCCs and the output of AANN are computed. The error is converted into confidence and the average confidence across all frames represent the score of the model for the given test utterance. The performance of the speaker recognition system using spectral features shown in the form of DET curve in Fig. 5. From the DET curve the EER for spectral features is 15.5 %.

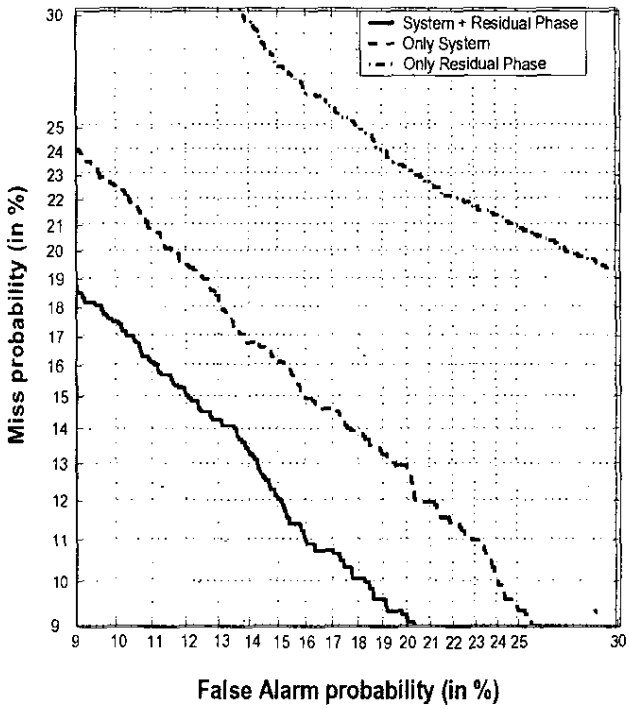

Fig. 5: DET curves for the systems built using residual phase, spectral features and the combined system

D. Combining evidence from Spectral and LP residual phase

The scores obtained for each speaker using LPCCs and the LP residual phase are combined by simple addition of the two scores. The performance of the combined system is plotted as DET curve in Fig. 5. The EER for the same is 13.5%. This study infers that the speaker-specific information in the LP residual phase is complementary to existing spectral features and hence the improvement in the performance of the combined system.

5. Conclusions

The objective of this paper was to demonstrate the complementary nature of speaker-specific information present in the residual phase information. This was was demonstrated by conducting speaker verification experiments on NIST 2003 Speaker Recognition evaluation database. The speaker recognition system using only residual phase information gives an EER of 22%, using only spectral features gives an EER of 15.5% and the combined system gives an EER of 13.5%.

In this work it was experimentally demonstrated that AANN models indeed capture speaker-specific information. However, efforts are needed to give analytical framework for the functioning of AANN models in the two cases.

References

[1] D. O'Shaughnessy, "Speaker recognition," *IEEE ASSP Magazine*, vol. 3, pp. 4–17, Oct. 1986.

[2] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, New Jersy: Prentice-Hall, 1993.

[3] 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.

[4] 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.

[5] K. S. Reddy, *Source and system features for speaker recognition*. MS thesis, Indian Institute of Technology Madras, Chennai-600 036, India, Department of Computer Science and Engg., 2001.

[6] C. S. Gupta, S. R. M. Prasanna, and B. Yegnanarayana, "Autoassociative neural network models for online speaker verification using source features from vowels," in *Proc. Int. Joint Conf. Neural Networks*, (Honolulu, Hawii, USA), May 2002.

[7] L. Mary, K. S. R. Murty, S. R. M. Prasanna, and B. Yegnanarayana, "Features for speaker and language identification," in *Proc. ODYSSEY 2004: the speaker and language recognition workshop*, (Toledo, Spain), May-June 2004.

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

[9] K. S. R. Murty, S. R. M. Prasanna, and B. Yegnanarayana, "Speaker-specific information from residual phase," in *SPCOM 2004 (IISc. Bangalore)*, Dec 2004.

[10] C. S. Gupta, *Significance of source features for speaker recognition*. MS thesis, Indian Institute of Technology Madras, Chennai-600 036, India, Department of Computer Science and Engg., 2003.

[11] "NIST speaker recognition evaluation plan," in *Proc. NIST speaker recognition workshop*, (University of Maryland, College Park, MD, USA), 2003.