Combining Evidence From Residual Phase and MFCC Features for Speaker Recognition

K. Sri Rama Murty and B. Yegnanarayana, *Senior Member, IEEE*

***Abstract*—The objective of this letter is to demonstrate the complementary nature of speaker-specific information present in the residual phase in comparison with the information present in the conventional mel-frequency cepstral coefficients (MFCCs). The residual phase is derived from speech signal by linear prediction analysis. Speaker recognition studies are conducted on the NIST-2003 database using the proposed residual phase and the existing MFCC features. The speaker recognition system based on the residual phase gives an equal error rate (EER) of 22%, and the system using the MFCC features gives an EER of 14%. By combining the evidence from both the residual phase and the MFCC features, an EER of 10.5% is obtained, indicating that speaker-specific excitation information is present in the residual phase. This information is useful since it is complementary to that of MFCCs.**

***Index Terms*—Autoassociative neural network, glottal closure instant, linear prediction (LP) residual, residual phase, speaker verification.**

I. Introduction

THE OBJECTIVE of automatic speaker recognition is to recognize a person from a spoken utterance [1]. A speaker recognition system can be operated in either identification mode or verification mode. In speaker identification, the goal is to identify the speaker of an utterance from a given population, whereas speaker verification involves validating the identity claim of a person. Speaker recognition systems can be classified into text-dependent systems and text-independent systems. Text-dependent systems require the recitation of a predetermined text, whereas text-independent systems accept speech utterances of unrestricted text. This letter deals with text-independent speaker verification.

Speech is a composite signal that mainly carries information about the message to be conveyed, speaker characteristics, and the language. Speaker characteristics in the speech signal can be attributed to the dimensions of the vocal tract system, characteristics of excitation, and the learning habits of the speakers. The speaker-specific vocal tract information is mainly represented by spectral features like mel-frequency cepstral coefficients (MFCCs) and linear prediction (LP) cepstral coefficients [2]. Efforts are being made to exploit the usefulness of features extracted from excitation source characteristics and suprasegmental characteristics for speaker recognition [3]–[6]. The goal of the search for new features is to improve the performance of the existing speaker recognition systems, which are based mainly on the characteristics of the vocal tract system. One of the desirable properties of the new features is that they should provide speaker-specific information complementary to the spectrum-based features like MFCC. Then by combining the evidence from the new features with the evidence from the existing features, the performance of the speaker recognition system can be improved.

In this letter, we demonstrate that the residual phase signal contains speaker-specific information that is complementary to the MFCC features. The residual phase is defined as the cosine of the phase function of the analytic signal derived from the LP residual of a speech signal. The speaker-specific information from the residual phase is captured using an autoassociative neural network (AANN) model. The speaker-specific information from the MFCC features is also captured using AANN models [7]. The evidence from both the models is used to validate the claim.

The letter is organized as follows. In Section II, a brief overview of spectral features and the need for complementary features is presented. Computation of the residual phase feature is described in Section III. Speaker recognition studies based on spectral features (MFCCs), excitation source features (residual phase), and combining evidence from both the systems are discussed in Section IV.

II. Spectral Features for Speaker Recognition

Most of the present-day systems use the characteristics of vocal tract system for speaker recognition. This information is extracted using short time spectrum analysis of segments of 20–30 ms (3–5 pitch periods) of speech signal. Cepstral coefficients are derived from the short time spectrum of speech signal. The cepstrum is the inverse Fourier transform of log-magnitude spectrum of speech signal. Since speech production is usually modeled as a convolution of the impulse response of the vocal tract filter with an excitation source, the cepstrum effectively deconvolves these two parts, resulting in a low-time component corresponding to the vocal tract system and a high-time component corresponding to the excitation source [8]. In speaker recognition systems based on spectral features, the initial 15–20 cepstral coefficients are used for a speech signal sampled at a frequency of 8 kHz. The high-time component of the cepstrum is mostly ignored, which means that the cepstral features usually ignore the information present in the excitation source.

MFCCs are widely used spectral features for speaker recognition. Computation of the MFCCs differs from the basic procedure described earlier, where the log-magnitude spectrum

Manuscript received June 13, 2005; revised August 18, 2005. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Frederic Bimbot.

The authors are with the Department of Computer Science and Engineering, Indian Institute of Technology–Madras, Chennai 600 036, India (e-mail: ksrm@cs.iitm.ernet.in; yegna@cs.iitm.ernet.in).

Digital Object Identifier 10.1109/LSP.2005.860538

is replaced with the logarithm of the mel-scale warped spectrum, prior to the inverse Fourier transform operation. Hence, the MFCCs represent only the gross characteristics of the vocal tract system. In this letter, we show that the characteristics of the time-varying excitation component of the speech signal are also useful for automatic speaker recognition. We also show that the speaker-specific information present in the excitation source complements the information present in the features representing the vocal tract system.

III. Computation of Residual Phase Through LP Analysis

In LP analysis, the sample $s(n)$ is estimated as a linear weighted sum of the past p samples. The predicted sample $\hat{s}(n)$ is given by

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

where p is the order of prediction, and $\{a_k\}$, $k = 1, 2, \ldots, p$ is the set of linear prediction coefficients (LPCs). The LPCs are obtained by minimizing the mean-squared error between the predicted sample value and the actual sample value over the analysis frame. The error between the actual value $s(n)$ and the predicted value $\hat{s}(n)$ is given by

$$r(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k). \tag{2}$$

This error $r(n)$ is called the LP residual of the speech signal. The LP residual contains mostly information about the excitation source. Values of LP residuals are large around the instants of glottal closure for voiced speech. Due to large fluctuations in amplitude, it is difficult to derive information from short segments of LP residual. Hence, we propose to use the phase of the analytic signal derived from the LP residual [9]. The analytic signal $r_a(n)$ corresponding to $r(n)$ is given by [10]

$$r_a(n) = r(n) + jr_h(n) \tag{3}$$

where $r_h(n)$ is the Hilbert transform of $r(n)$ and is given by

$$r_h(n) = \text{IFT}\left[R_h(\omega)\right] \tag{4}$$

where

$$R_h(\omega) = \begin{cases} -jR(\omega), & 0 \leq \omega < \pi \\ jR(\omega), & 0 > \omega \geq -\pi. \end{cases} \tag{5}$$

Here $R(\omega)$ is the Fourier transform of $r(n)$, and IFT denotes the inverse Fourier transform. The magnitude of the analytic signal $r_a(n)$ [Hilbert envelope $h_e(n)$] is given by

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

and the cosine of the phase of the analytic signal $r_a(n)$ is given by

$$\cos\left(\theta(n)\right) = \frac{\text{Re}\left(r_a(n)\right)}{|r_a(n)|} = \frac{r(n)}{h_e(n)}. \tag{7}$$

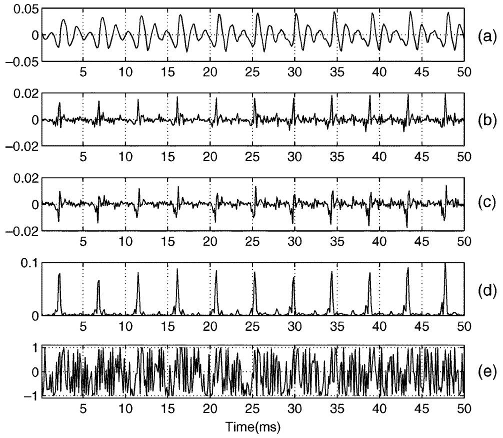

Fig. 1. (a) Speech signal. (b) LP residual. (c) Hilbert transform of residual. (d) Hilbert envelope of residual. (e) Residual phase.

A segment of voiced speech, its LP residual, the Hilbert transform of the LP residual, the Hilbert envelope, and residual phase $(\cos(\theta(n)))$ are shown in Fig. 1. It is difficult to see any feature from the plot of the residual phase. However, since during LP analysis, only the second-order relations are removed, the higher order relations among the samples of the speech signal are retained in the samples of the residual phase. Since speech production mechanism is not a Gaussian process, it is reasonable to expect that the speaker-specific information is present in the higher order relations among the samples of the residual phase. In this letter, residual phase is used to characterize the speaker-specific information in the excitation source. Note that in the residual phase, the amplitude information of the LP residual samples is not preserved. In the LP residual, the region around the glottal closure (GC) instant within each pitch period corresponds to high signal-to-noise ratio (SNR) region due to impulse-like excitation. These regions are known to contain speaker-specific information [6]. We also assume that the phase information around the GC instants in the residual contains better speaker-specific information compared to other regions [11]. The knowledge of the GC instants is used for selecting the residual phase segments for extracting the relations among the samples. The difference between the LP residual and the residual phase information is that the strength of the excitation around the GC instant present in the LP residual is eliminated in the residual phase information. Thus, in the residual phase, speaker-specific information is expected to be present only in the sequence of the samples.

IV. Speaker Verification Studies

A. *Database for the Study*

The speaker verification experiments presented in this letter are conducted using the NIST-2003 speaker recognition corpus of male speakers [12]. There are 149 male speakers, and the duration of training data for each speaker is about 2 min. There are 1343 test utterances, each having a duration of 15–45 s. Each test utterance has 11 claimants, where the genuine speaker may or may not be present. All speech signals were sampled at 8 kHz.

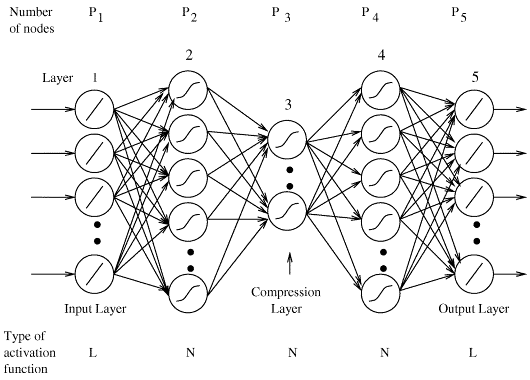

Fig. 2. Structure of the AANN model.

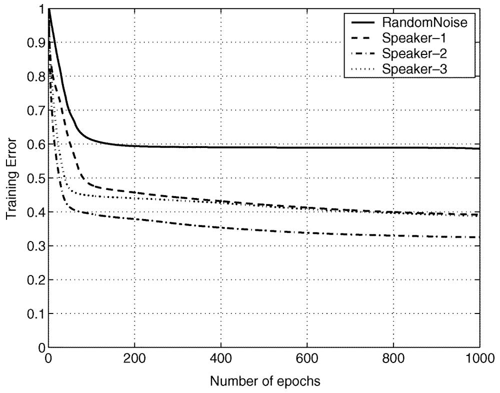

Fig. 3. Training error curves for random noise and residual phase.

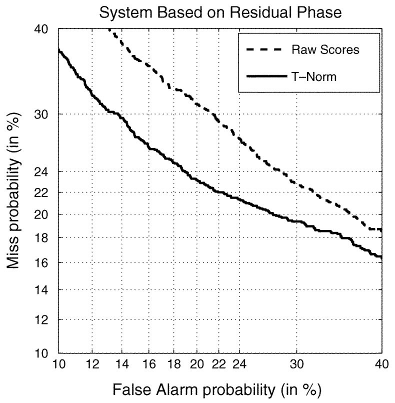

Fig. 4. DET curves for the system based on residual phase.

B. Speaker Verification Using Residual Phase

LP residual is obtained from the speech signal using a tenth-order LP analysis. The residual phase feature is computed from the LP residual using the method described in Section III. Voiced regions are identified using a method based on autocorrelation analysis of the Hilbert envelope of the LP residual. The GC instants are also detected using the Hilbert envelope [9]. The residual phase values around the GC instants in the voiced segments are used to extract the speaker-specific information. The ability of AANNs to capture the nonlinear relations is exploited for developing speaker-specific models.

AANN is a feedforward neural network model that performs identity mapping [13]. After training, the AANN model should be able to reproduce the input vector at the output with minimum error, if the input is from the same system. The AANN model consists of one input layer, one output layer, and one or more hidden layers. The nodes in the input and output layers are linear, whereas the nodes in the hidden layers are nonlinear. The neural network is expected to capture the speaker-specific information present in the higher order relations among the samples of residual phase. A five-layer neural network architecture (see Fig. 2) is considered for the letter. The structure of the network is $40L$ $48N$ $12N$ $48N$ $40L$, where L represents linear nodes, N represents nonlinear nodes, and numerals represent number of nodes in the layer. The structure of the network was arrived at empirically after some preliminary studies. During the training phase, six blocks of 40 samples around each GC event are considered with a shift of one sample. Each block is presented as input to the network, and the output is computed. The error vector between input and output is used to update the weights of the network using the back propagation algorithm [14], [15]. The AANN model is trained for 500 epochs. The network indeed captures the higher order relations present among the samples of the residual phase, as can be seen by the reduction in the error in the training error curve (see Fig. 3).

During the testing phase, six blocks of 40 samples around each GC instant are considered with a shift of one sample. Each block is applied as input to the AANN, and the output is computed. The squared error (e_i^2) between the input and the output of the AANN is converted into a confidence score using the relation $c_i = \exp(-\lambda e_i^2)$, where i refers to the block index. In this letter, we have chosen $\lambda = 1$. The average confidence score for a given test utterance is computed as $C = (1/N) \sum_{i=1}^{N} c_i$, where N is the total number of blocks in the test utterance. The performance of the speaker recognition system is given by the detection error tradeoff (DET) curve [16] shown in Fig. 4. From the DET curve, the EER is found to be 26%. Test utterance normalization (TNorm) is performed on the raw confidence scores in order to transform the scores into a similar range. After TNorm, the performance has improved to 22%.

C. Speaker Verification Using MFCCs

The MFCC features are extracted from voiced segments of the speech signal. The first 19 cepstral coefficients, other than the zeroth value (average of the log-spectral values), are used. Cepstral mean subtraction is performed to reduce the channel effects. The structure of the AANN for capturing the distribution of the MFCCs of each speaker is $19L$ $38N$ $8N$ $38N$ $19L$, as described in [17]. During training, the feature vectors are presented in a random order to the AANN. One model is trained for each speaker for 60 epochs. The performance of the AANNs did not improve, even if the number of epochs was increased to 500. Hence, all the AANN models were trained for only 60 epochs.

During testing, the MFCCs are extracted from the test utterance. The MFCC feature vectors are applied to the AANN models. For each frame of 20 ms, the squared error between the MFCCs and the output of AANN is computed. The squared error is converted into a confidence value, and the average confidence across all frames is used to score the test utterance against a given model. The performance of the speaker

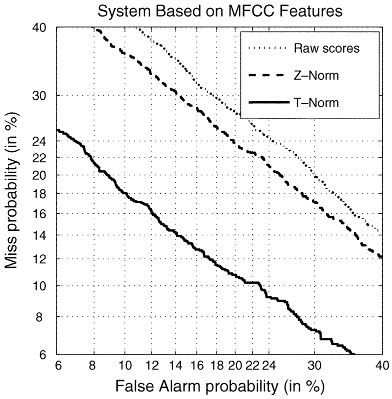

Fig. 5. DET curves for the system based on MFCC features.

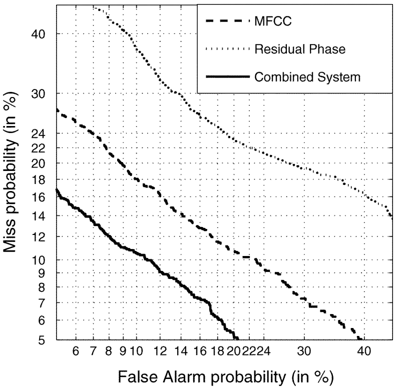

Fig. 6. DET curves for the combined system. DET curves for system based on MFCCs and residual phase are also given for comparison.

recognition system using the MFCC-based spectral features is shown in the form of the DET curve in Fig. 5. From the DET curve, the EER for the raw scores obtained from the spectral features is 24%. The performance has improved to 14% when normalization techniques [model normalization (ZNorm) and TNorm] are employed [18].

D. Combination of Speaker Recognition Systems

The confidence scores C_s and C_p obtained using MFCCs and the residual phase, respectively, for each speaker are combined using the linear weighted sum, given by $C_c = \alpha C_s + (1-\alpha) C_p$. The performance of the combined system is plotted as the DET curve in Fig. 6. For $\alpha = 0.5$, the EER of the combined system is 10.5%. This shows that, due to complementary speaker-specific information present in the residual phase, the performance of the system based on the spectral features can be improved by combining the scores.

V. Conclusion

The objective of this letter was to demonstrate the complementary nature of the residual phase and to show that this information indeed helps in improving the performance of the conventional systems based on spectral features such as MFCC. It was demonstrated by conducting speaker verification experiments on the NIST-2003 speaker recognition evaluation database. The speaker recognition system using only the residual phase information resulted in an EER of 22%, and that using only the MFCC features resulted in an EER of 14%. However, the combined system led to an EER of 10.5%, which is significantly better than both of the individual systems. The AANN models used in this letter were not optimized in terms of the network structure and training. Also the scores may be combined in a better way to improve the performance of the combined system.

References

[1] D. O'Shaughnessy, "Speaker recognition," *IEEE ASSP Mag.*, vol. 3, no. 4, pp. 4–17, Oct. 1986.

[2] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-29, no. 2, pp. 254–272, Apr. 1981.

[3] 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, Signal Processing*, Salt Lake City, UT, May 2001, pp. 409–412.

[4] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 5, pp. 569–586, Sep. 1999.

[5] G. Doddington, "Speaker recognition based on idiolectal differences between speakers," in *Proc. Eur. Conf. Speech Processing, Technology*, Aalborg, Denmark, Sep. 2001, pp. 2521–2524.

[6] C. S. Gupta, "Significance of source features for speaker recognition," M.S. thesis, Dept. Comput. Sci. Eng., Indian Inst.Technol.–Madras, Chennai, India, 2003.

[7] B. Yegnanarayana and S. P. Kishore, "AANN: An alternative to GMM for pattern recognition," *Neural Netw.*, vol. 15, pp. 459–469, Apr. 2002.

[8] D. O'Shaughnessy, *Speech Communication: Human and Machine*. Reading, MA: Addison-Wesley, 1987.

[9] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 4, pp. 309–319, Aug. 1979.

[10] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1975.

[11] K. S. R. Murty, S. R. M. Prasanna, and B. Yegnanarayana, "Speaker-specific information from residual phase," in *Proc. Int. Conf. Signal Processing Communications*, Bangalore, India, Dec. 2004.

[12] "NIST speaker recognition evaluation plan," in *Proc. NIST Speaker Recognition Workshop*, College Park, MD, 2003.

[13] M. S. Ikbal, H. Misra, and B. Yegnanarayana, "Analysis of autoassociative mapping neural networks," in *Proc. Int. Joint Conf. Neural Networks*, Washington, DC, Jul. 1999, pp. 854–858.

[14] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Englewood Cliffs, NJ: Prentice-Hall., 1999.

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

[16] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. Eur. Conf. Speech Processing Technology*, Rhodes, Greece, Sep. 1997, pp. 1895–1898.

[17] S. P. Kishore, "Speaker verification using autoassociative neural network models," M.S. thesis, Dept. Comput. Sci. Eng., Indian Inst.Technol.–Madras, Chennai, India, 2001.

[18] R. Aukenthaler, M. Carey, and H. L. Thomas, "Score normalization for text-independent speaker verification systems," *Dig. Signal Process.*, vol. 10, pp. 42–54, Jan. 2000.