SPEAKER-SPECIFIC INFORMATION FROM RESIDUAL PHASE

K. Sri Rama Murty [1], *S. R. Mahadeva Prasanna* [2] *and B. Yegnanarayana* [1]

[1] Speech and Vision Laboratory
Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai - 600 036, INDIA
Email: {ksrm,yegna}@cs.iitm.ernet.in

[2] Department of Electronics and Communication Engineering
Indian Institute of Technology Guwahati, Guwahati-781039, India
Email: prasanna@iitg.ernet.in

ABSTRACT

This paper demonstrates the presence of speaker-specific information in the residual phase using Autoassociative Neural Network (AANN) models. The residual phase is extracted from the speech signal after eliminating the vocal tract information by the Linear Prediction (LP) analysis. AANN models are used for capturing the speaker-specific information present in the residual phase. The speaker recognition studies infer that the residual phase contains significant speaker-specific information and it is indeed captured by the AANN models. In this study we also demonstrate that in voiced speech segments, regions around the instants of glottal closure are more speaker-specific compared to other regions.

1. INTRODUCTION

Speaker recognition is the task of recognizing speakers using the information from their speech signals [1]. Depending on the objective, the speaker recognition task involves either identification or verification. In speaker identification the objective is to identify the speaker of the test speech signal from a given set of speakers. Alternatively, speaker verification involves verifying the identity claim of the speaker to accept or reject the claim. In this work the presence of speaker-specific information in the residual phase is demonstrated through the speaker identification studies.

The speaker-specific information in the speech signal may be attributed to the dimensions of the vocal tract, characteristics of the excitation source and learning habits of the speaker. The vocal tract information is represented by the spectral features. The distribution of the spectral features is assumed to be unique for each speaker and is captured by either Gaussian Mixture Models (GMM) or Autoassociative Neural Network (AANN) models [2–4]. In the excitation source signal the speaker-specific information is assumed to be present among some higher order relations of the samples and is captured by the AANN models [4]. The learning habits of the speaker like idiolect is also known to contain speaker-specific information and has been exploited for the speaker recognition studies [5].

The speaker-specific vocal tract and excitation source information may be extracted from the speech signal by the Linear Prediction (LP) analysis [6]. The LP residual obtained by the LP analysis mostly contains excitation source information. The ability to recognize speakers by informal listening to the LP residual infers that the LP residual indeed contains significant speaker-specific information. However, the speaker-specific information from the magnitude of the LP residual may dominate while listening directly to the LP residual. The presence of speaker-specific information in the LP residual phase may be demonstrated as follows: As will be explained in the next section the phase information ($sin\theta(n)$) may be extracted from the LP residual using the Hilbert envelope derived from it [7, 8]. The speech signal is synthesized by exciting the time-varying filter obtained by the LP analysis using only $sin\theta(n)$ as excitation. Similarly, the speech signal is also synthesized by exciting the time-varying filter using only random noise as the excitation. By informal listening it was found that the presence of speaker information is more in the speech synthesized by $sin\theta(n)$, compared to that using random noise. This experiment motivated us to explore the usefulness of the LP residual phase information for speaker recognition studies.

This paper is organized as follows: In Section 2 we discuss about deriving the LP residual phase information. The speaker identification studies from the phase information using AANN models is described in Section 3. In Section 4 the significance of the regions around the Glottal Closure (GC) instants for speaker identification is demonstrated. GC is the instant at which closure of vocal folds occurs in a pitch

0-7803-8674-4/04/$20.00 ©2004 IEEE

period. The summary of various issues discussed and the scope for future work is given Section 5.

2. RESIDUAL PHASE INFORMATION

In LP analysis each sample is predicted as a linear combination of past p samples [6]. The predicted sample ($\hat{s}(n)$) is given by

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

where p is the order of prediction and $\{a_k\}$ are the Linear Prediction Coefficients (LPCs) obtained by LP analysis.

The LP residual is the difference between the original and the predicted samples and is given by

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

The Hilbert envelope of the LP residual is computed as follows [7, 8]:

$$h_e(n) = \sqrt{r^2(n) + r_h^2(n)} \quad (3)$$

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

$$r_h(n) = \left\{ \begin{array}{rl} IDFT[-jR(\omega)], & 0 < \omega < \pi \\ IDFT[jR(\omega)], & 0 > \omega > -\pi \\ 0 & \omega = 0, \pi \end{array} \right\} \quad (4)$$

where IDFT is the Inverse Discrete Fourier Transform, and $R(\omega)$ is the discrete Fourier transform of $r(n)$.

The Hilbert envelope represents the magnitude of the analytic signal whose real part is the LP residual and imaginary part is the Hilbert transform of the LP residual. Hence the phase information $sin\theta(n)$ may be extracted from the LP residual as

$$sin\theta(n) = r(n)/h_e(n) \quad (5)$$

A segment of voiced speech and its LP residual, Hilbert transform of the LP residual, Hilbert envelope and $sin\theta(n)$ are shown in Figure 1. Since $sin\theta(n)$ is derived using the analytic signal concept, it does not suffer from the phase warping problem [7]. Hence in this work $sin\theta(n)$ is used as residual phase information for the speaker identification studies. It is difficult to make out any speaker-specific information from the plots of $sin\theta(n)$. However, during LP analysis since the second order correlations are removed, we conjecture that the speaker-specific information may be present among some higher order relations of the samples of $sin\theta(n)$. As it is not clear how to extract the speaker-specific information from these higher order relations, we propose to use the AANN models.

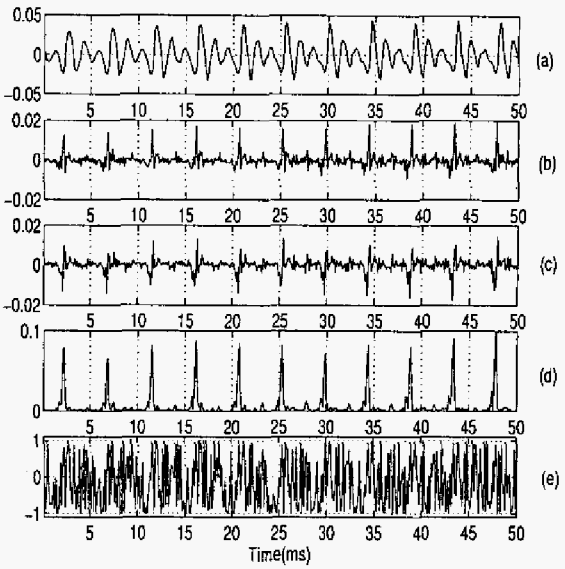

Fig. 1. (a) A segment of voiced speech and its (b) LP residual, (c) Hilbert transform of the LP residual, (d) Hilbert envelope of the LP residual and (e) $sin\theta(n)$ derived from the LP residual.

3. SPEAKER IDENTIFICATION USING RESIDUAL PHASE

The speech data for this study was taken from two dialect regions namely, **dr1** and **dr2** of TIMIT database. The **dr1** consists of 38 speakers (14 female and 28 male) and **dr2** consists of 76 speakers (23 female and 53 male). The speech signals originally sampled at 16 kHz were downsampled to 8 kHz. For each speaker there are ten speech utterances and among these seven are used for training and remaining three for testing. The two training sets are **MSET1** for **dr1** and **MSET2** for **dr2**. One test utterance for each speaker is chosen to form a test set. Hence there are six test sets namely, **TSET11, TSET12, TSET13** for **dr1** and **TSET21, TSET22, TSET23** for **dr2**.

The LP residual is extracted from the speech signal by 10^{th} order LP analysis. The $sin\theta(n)$ is computed from the LP residual using the Hilbert envelope of the LP residual. The voiced segments are identified from the speech signal by the autocorrelation analysis on the Hilbert envelope of the LP residual and setting a threshold on the normalized peak strength (first major peak after center peak) in the autocorrelation sequence. The $sin\theta(n)$ values in the voiced segments are considered for the study.

AANN is a neural network model which performs identity mapping [9, 10]. In other words, the samples in a block are mapped on to themselves to learn the relationship among the samples. It consists of one input layer, one output layer and one or more hidden layers. The elements in the input

and output layers are linear, whereas the elements in the hidden layer are nonlinear. The nonlinear processing employed during learning (training) of the neural network may help in capturing the speaker specific information that may be present in some higher-order relations among the samples of $sin\theta(n)$. A five layer AANN model with structure $40L\ 48N\ 12N\ 48N\ 40L$ is used in the present study.

During training, $sin\theta(n)$ in the voiced segments considered in blocks of 40 samples with a shift of one sample are applied to the AANN models in a sequence. One AANN model is trained per speaker for 1000 epochs. The training error curves of the AANN models for three speakers is shown in Figure 2. The decrease in the error values from one epoch to the next indicates that there is some information among the samples of $sin\theta(n)$ and is being learnt by the AANN model. The training error curve for a random noise sequence is also shown in the Figure 2 and large error with nearly straight line indicates that there is no relation among the samples in the random noise.

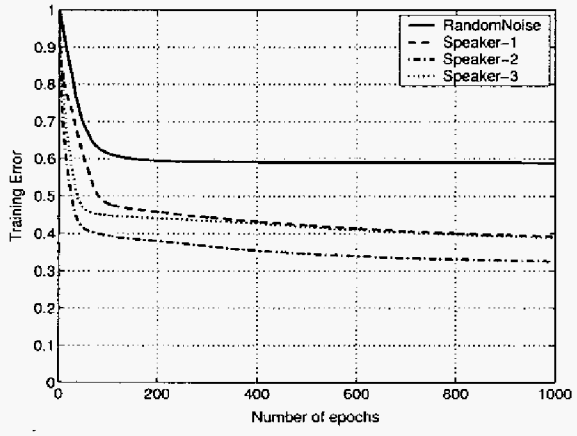

Fig. 2. Training error curves for the speech data from three speakers and also from random noise.

During testing, $sin\theta(n)$ in the voiced segments of the test speech signal considered in blocks of 40 samples with a shift of one sample are applied to the AANN models. For each block, the error (e_i) between the input and the output values of AANN models is computed. The error is converted into confidence value as $c_i = exp(-e_i)$ and the average confidence across all the blocks is computed as $C = \frac{1}{N}\sum_{i=1}^{N} c_i$ where N is the number of blocks.

The performance of the speaker identification system measured in terms of the ratio of number of genuine models with first rank to the total number of models expressed as percentage is given in Table 1. The consistent performance across different test sets indicates that speaker information is present in the residual phase and is captured by the AANN models. The relatively poor performance in **MSET2** compared to **MSET1** is mainly due to the increase in number of speakers.

All the blocks in the voiced segments may not contain significant information about the speaker. The performance of the speaker identification system may be improved by selecting blocks which contain more speaker-specific information. For instance, by nature of speech production regions around the GC instants are high Signal to Noise Ratio (SNR) regions and may contain more speaker-specific information and will be discussed in the next section.

4. SIGNIFICANCE OF REGIONS AROUND THE GC INSTANTS

Since the significant excitation of vocal tract occurs at the GC instant, a large error in the LP residual should indicate the location of the GC instant. The GC instant may be identified better in the Hilbert envelope of the LP residual (see Figure 1), due to its unipolar nature. The peaks in the Hilbert envelope of the LP residual are identified. Though most of the peaks coincide with GC instants, some spurious peaks do exist. Most of the spurious peaks are eliminated based on the hypothesis that the time gap between the two successive GC instants is not likely to vary much in the adjacent pitch periods.

For comparing the performance same speech data used in the previous section is considered. A 10^{th} order LP analysis is performed for computing the $sin\theta(n)$ values. The voiced segments and the GC instants are detected. In each voiced segment five blocks of 40 samples each around the GC instants with a shift of one sample are considered for training and testing the AANN models. One AANN model is trained per speaker for 1000 epochs. The training error curve for *Speaker-1* shown in Figure 3 and also the training error curve for the same speaker by considering only the blocks around the GC instants are shown in Figure 3. The lower values in the training error curve using the knowledge of the GC events indicates that the AANN models are learning better using the information around the GC instants. During testing, the average confidence values are computed. The performance of the speaker identification system using the knowledge of the GC instants is shown in Table 1. The significantly improved performance by considering the blocks around the GC instants indicates that regions around the GC instants contain more speaker-specific information. To explain this fact, the confidence scores of different speaker models for a test utterance of *Speaker-23* of **dr2** with and without using the knowledge of the GC events are plotted in Figure 4. The confidence score of the model of *Speaker-23* is significantly higher when information about the GC events is used. Another interesting observation is that the scores of most of other speaker models which were higher than the score of *Speaker-23* model have been reduced significantly. This indicates that the models

have learnt their respective speaker characteristics better.

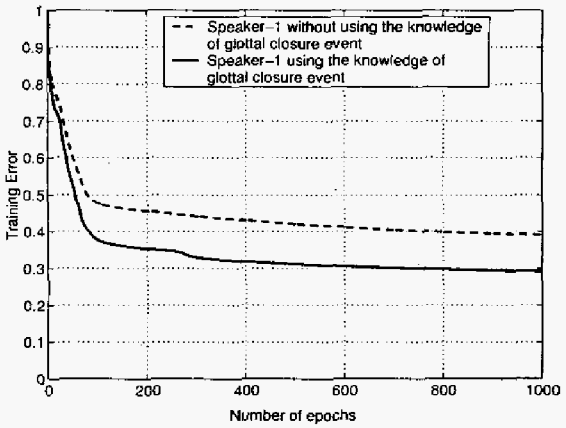

Fig. 3. Training error curves for the speech data from the same speaker with and without using the knowledge of the GC events.

Table 1. Performance of speaker identification system in terms of the ratio of number of models with first rank to the total number of models expressed as percentage.

Speakers	without GC instants		
	TSET11	TSET12	TSET13
MSET1	50 %	47 %	50 %
	TSET21	TSET22	TSET23
MSET2	45 %	43 %	48 %

Speakers	with GC instants		
	TSET11	TSET12	TSET13
MSET1	84 %	87 %	84 %
	TSET21	TSET22	TSET23
MSET2	76 %	74 %	75 %

5. SUMMARY AND CONCLUSIONS

In this work we have demonstrated the presence of speaker-specific information in the residual phase sequence using the AANN models. The significantly improved performance by using the GC instants indicates that selection of the regions containing more speaker-specific information helps in improving the performance.

In the present study clean speech collected over microphones is used. Studies need to extended for degraded speech case in which the speech is collected over either telephone channel or cell phone. The phase information is one component of speech and efforts needs to be made to combine the speaker-specific information from phase with other components like speaker-specific information from spectral and excitation source features.

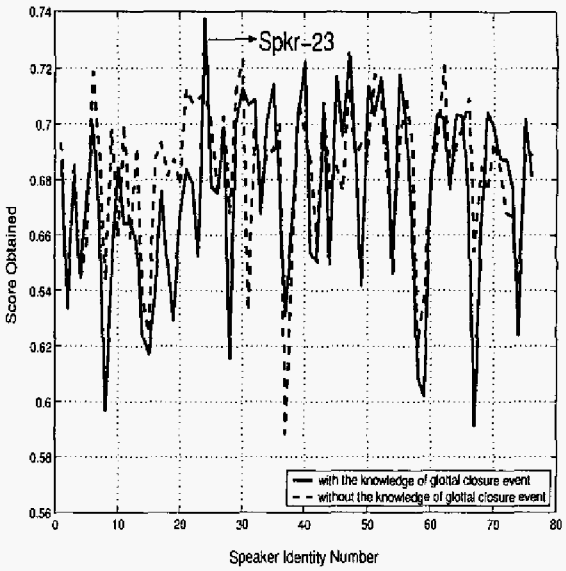

Fig. 4. The confidence scores of the model of **spkr-23** of **dr2** with and without using the information about the GC events.

6. REFERENCES

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

[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. 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, Utah, USA), pp. 409–412, May 2001.

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

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

[7] A. V. Oppenheim and R. W. Schafer, *Digital signal processing*. Englewood Cliffs, New Jersey: Prentice Hall, 1975.

[8] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 309–319, Aug. 1979.

[9] S. Haykin, *Adaptive Filter Theory*. New Jersey: Prentice Hall, second ed., 1991.

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