Autoassociative Neural Network Models for Online Speaker Verification using Source Features from Vowels

Cheedella S. Gupta, S.R. Mahadeva Prasanna, and B. Yegnanarayana
Speech and Vision Laboratory
Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai - 600 036 India
Email: {gupta,prasanna,yegna}@speech.cs.iitm.ernet.in

Abstract - *In this paper we demonstrate the usefulness of excitation source information for text-dependent speaker verification. The nature of vibration of vocal folds may be unique for a given speaker. This can be studied by considering vowels, since the excitation in this case is only due to glottal vibration. Linear prediction (LP) residual contains mostly source information. We propose autoassociative neural network models for capturing speaker-specific source information present in the LP residual. Speaker models are built for each vowel to study the extent of speaker information in each vowel. Using this knowledge an online speaker verification system is developed. This study demonstrates that excitation source indeed contains significant speaker information, which can be exploited for speaker recognition tasks.*

I. Introduction

Speaker verification involves accepting or rejecting the claim of a speaker [1–4]. Speech signal carries information related to not only the message to be conveyed, but also about speaker, language, emotional status of the speaker, environment and so on. In a speaker recognition task the speech signal is processed to extract the speaker-specific information. Speech is the result of exciting a time-varying vocal tract system with time-varying excitation. The interplay/coupling between the source and system for producing speech is likely to have unique speaker characteristics. Since it is not known how to characterize this coupling, an alternative approach is to extract the features characterizing source and system separately, and use them for speaker recognition task.

Most of the existing speaker recognition systems use spectral features, which characterizes the vocal tract system of the given speaker [5–9]. It is interesting to note that human beings recognize speakers mostly from the source characteristics such as glottal vibrations, and prosodic features such as intonation and duration [10,11]. Due to variability and also due to difficulty involved in extracting these features, not much effort has gone in using these features for speaker recognition. In this paper we explore the usefulness of the source of excitation for speaker verification. We use autoassociative neural network models for capturing the source information.

The sources of excitation in speech are plosive, fricative and glottal vibration. Plosive excitation is due to the total closure and sudden release of the vocal tract system, and it results in the production of stop consonants. Fricative excitation is due to narrow constriction some where along the length of vocal tract system, which results in the production of fricative sounds. Glottal vibration produces voiced sounds like vowels, nasals and semivowels. Glottal vibration is the major excitation of speech, as more than 70% of the speech is voiced. Moreover, if voicing is replaced by random noise excitation to produce whispered type of speech, one notices that most of the speaker's identity is lost. Thus it appears that significant speaker-specific information may be present in the nature of vibration of the vocal folds. Among the voiced sounds, speaker information may be significant in the case of vowels. Hence we consider the source information for the case of five vowels */a/*, */i/*, */u/*, */e/* and */o/* in this study.

This paper is organized as follows: In Section II we discuss the significance of source of excitation for speaker recognition. The extraction of speaker-specific source information using autoassociative neural network models is described in Section III. Section IV discusses the extent of speaker information in different vowels, by conducting speaker recognition experiments for each of the vowels. An online speaker verification system using the source features is described in Section V. Some conclusions form this study, as well as some issues to be addressed further, are discussed in the last section.

0-7803-7278-6/02/$10.00 ©2002 IEEE

II. Significance of Source of Excitation for Speaker Recognition

The first step in using the source information for speaker recognition studies is to separate the source of excitation from the speech signal. This can be done conveniently by using linear prediction (LP) analysis [12]. In the linear prediction analysis of speech, each sample is predicted as a linear weighted sum of the past p samples, where p represents the order for prediction.

If $s(n)$ is the present sample, then it is predicted by the past p samples as,

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

The difference between the actual and predicted sample value is termed as prediction error or residual, which is given by

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

The linear prediction coefficients $\{a_k\}$ are determined by minimizing the mean squared error over an analysis frame.

It has been shown that the LP order used for extracting the residual plays a crucial role on the performance of speaker recognition systems [13,14]. The study shows that the optimal range of the LP order for speaker recognition is in the range 8-16 for speech signal sampled at 8 kHz [13,14].

The five vowels considered in the present study may be grouped into three categories depending on the position of tongue hump as, front vowels (/i/), middle vowels (/a/ and /e/) and back vowels (/u/ and /o/). The vowels are also classified depending on the lip rounding as rounded (/u/ and /o/) or unrounded (/a/, /i/ and /e/). Even though the source of excitation is glottal vibration in all the cases, the characteristics of the excitation source will be different for different vowels due to the position of tongue hump and lips. This can be seen in Figure 1, where segments of LP residuals for the five vowels are given for a speaker. As can be seen in the figure, the excitation in case of vowels /u/ and /o/ are not as sharp as for the other vowels. Perceptually also speaker characteristics seem to be manifested well for unrounded vowels compared to rounded vowels. Thus the extent of speaker information manifested in the excitation source may be different for different vowels. This is also confirmed by the experimental studies to be discussed later.

The excitation source characteristics are also different among different speakers. This is illustrated in Figure 2, where the LP residuals for segments of vowel /a/ are shown for five different speakers. As shown in the figure, the rate of vibration of the vocal folds and the strength of excitation are different for different speakers. In the next section we discuss methods to capture the speaker-specific source information from the LP residual.

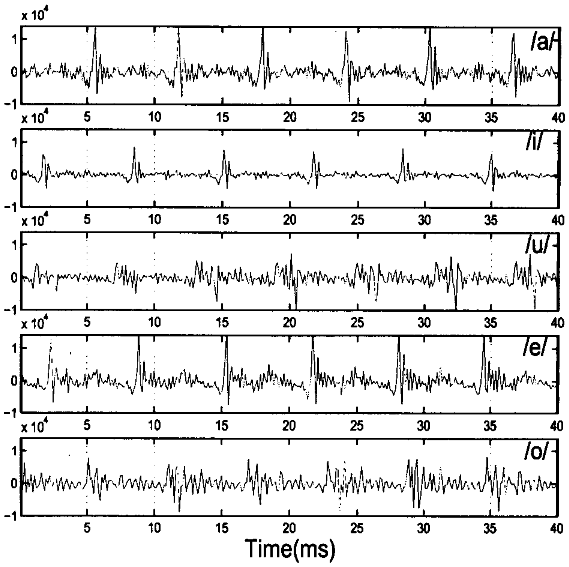

Fig. 1. LP residuals for the same speaker for segments of five different vowels.

III. AANN models for Capturing Source Information

Since LP analysis extracts the second order statistical features through the autocorrelation matrix, the LP residual does not contain any significant second order statistics corresponding to the vocal tract system. But the source characteristics are present in the LP residual. We conjecture that the source features may be present in the higher order statistics in the residual signal. Since it is not clear what specific set of parameters are to be extracted to represent the source information, and also since the extraction of such an information may involve nonlinear processing, we propose neural network models to capture the source information from the residual [15].

Autoassociative neural network models are feedforward neural networks performing an identity mapping of the input space [16]. AANN models were shown to capture the source features [17]. For capturing the speaker-specific source information present in the LP residual signal, a five layer AANN model with the structure shown in Figure 3 is used. The structure of the network used

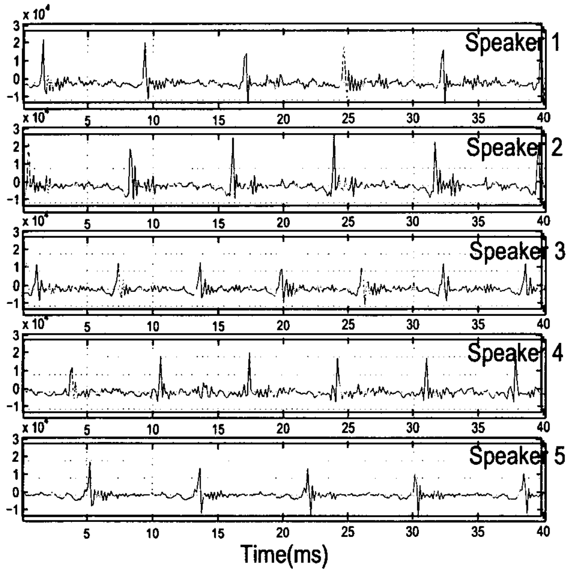

Fig. 2. LP residuals for the segments of the same vowel $/a/$ for five different speakers.

in our study is 40L 48N 12N 48N 40L, where L refers to linear units and N to nonlinear units. A *tanhx* is used for the nonlinear activation function. The performance of the network does not depend critically on the structure of the network [18].

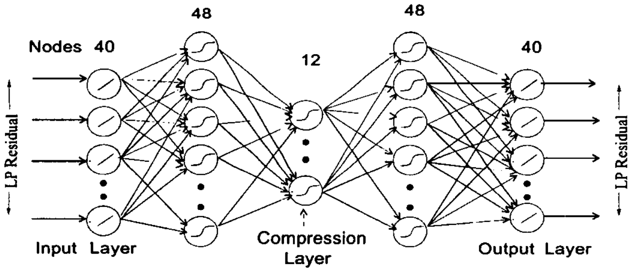

Fig. 3. Structure of AANN Model used for capturing speaker-specific source information

IV. Speaker Recognition Studies using Vowels

To study the effectiveness of the speaker-specific source information for each of the vowels, we conducted recognition experiments separately for each vowel. The data for the recognition experiments is collected from 20 cooperative speakers. For building speaker models, we collected vowels of duration 1-3 sec. The speech signal is collected by a microphone in the laboratory environment. The signal is sampled at 8 kHz, and is stored as 16 bit integers. LP residual is extracted from the speech signal using a 12^{th} order LP analysis, and the residual is normalized to unit magnitude before feeding it to the AANN models. Residual samples are given in blocks of 40 samples with every sample shift. The speaker models are trained for 60 epochs using backpropagation learning algorithm [15]. The training error curves for all the five vowels of a speaker are given in Figure 4. The low training error values for vowels $/a/$, $/i/$ and $/e/$ shows that speaker-specific information may be represented better in these cases. Higher training error values for vowels $/u/$ and $/o/$ may be attributed to poor representation of speaker-specific information. One model is built for each vowel for each speaker.

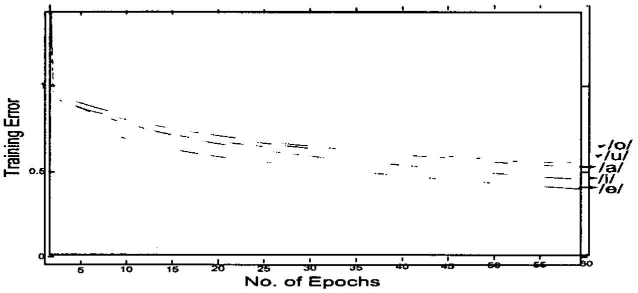

Fig. 4. Training error curves for the five vowels $/a/$, $/i/$, $/u/$, $/e/$, and $/o/$ for a speaker.

During verification, a test utterance of typically 0.5 sec duration is used. The LP residual is computed using a 12^{th} order LP analysis, and is normalized to unit magnitude for each block of 40 samples. The blocks are presented with one sample shift to all the models. The output of each model is compared with its input to compute the squared error for each block. The error (E_i) for the i^{th} block is transformed into a confidence value using $C_i = exp(-\lambda E_i)$, where the constant $\lambda = 1$. The frame confidences for a segment of all the vowels, for both the genuine as well as an impostor speaker are shown in Figure 5 to Figure 9. As shown in the figures, the confidence values for genuine speakers in case of vowels $/a/$, $/i/$ and $/e/$ have high discrimination compared to the confidences of the impostors. The average confidence for the genuine speaker is around 0.6, whereas that for the impostor speaker it is around 0.5. For $/u/$ and $/o/$, the discrimination between the confidences of genuine and impostor speakers is very low. As shown in the figure, the average confidences for both the cases are around 0.5. The average of all the frame confidences for a given test utterance is given by $C = (1/N)\sum_{i=1}^{N} C_i$, where N is number of blocks in the test utterance. The average confidence value is used to evaluate the performance of the test utterance with respect to the given model.

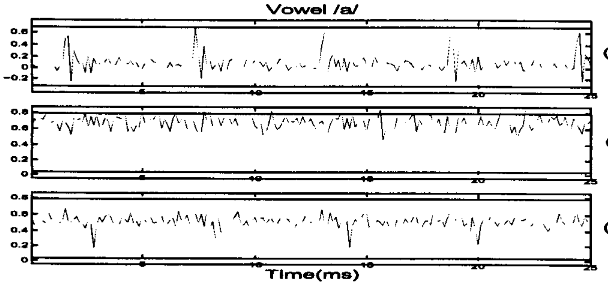

Fig. 5. For a segment of vowel /a/, (a) Normalized LP residual, (b) Frame confidences for genuine speaker, and (c) Frame confidences for an impostor speaker.

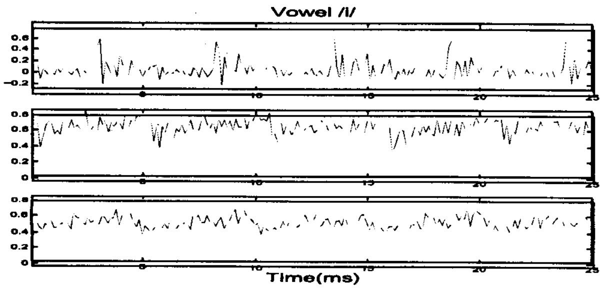

Fig. 6. For a segment of vowel /i/, (a) Normalized LP residual, (b) Frame confidences for genuine speaker, and (c) Frame confidences for an impostor speaker.

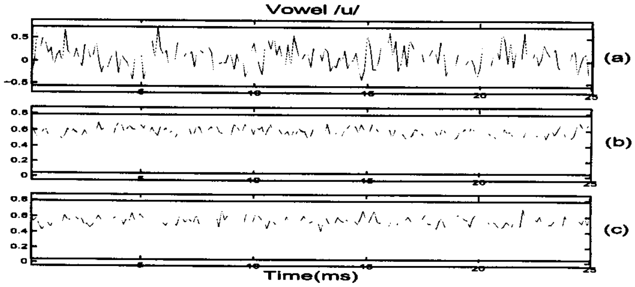

Fig. 7. For a segment of vowel /u/, (a) Normalized LP residual, (b) Frame confidences for genuine speaker, and (c) Frame confidences for an impostor speaker.

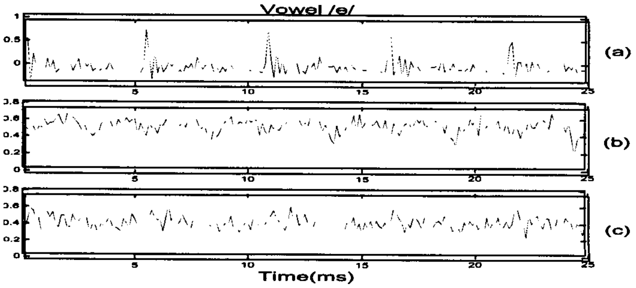

Fig. 8. For a segment of vowel /e/, (a) Normalized LP residual, (b) Frame confidences for genuine speaker, and (c) Frame confidences for an impostor speaker.

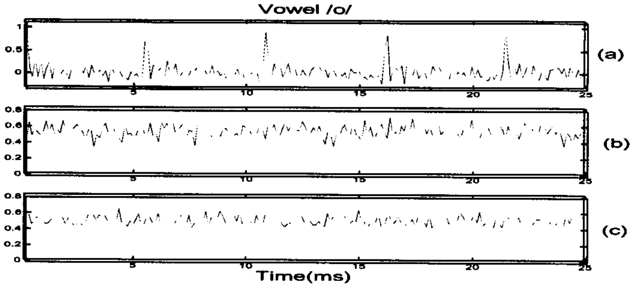

Fig. 9. For a segment of vowel /o/, (a) Normalized LP residual, (b) Frame confidences for genuine speaker, and (c) Frame confidences for an impostor speaker.

To evaluate the performance we conducted 50 genuine trials and 50 impostor trials for each of the vowels. The performance is expressed in terms of False Acceptance (FA) and False Rejection (FR), and are expressed in percentage. The result of the testing for all the vowels is given in Table I. The high false rejection in case of vowels $/u/$ and $/o/$ indicates the poor presence of speaker-specific information in these vowels. Even though the FR in other cases is considerably high, treating the output of speaker verification system for each vowel as independent evidence, and combining all these evidences, improves the performance significantly. This feature is exploited in building an online speaker verification system, which is explained in the next section.

TABLE I

Performance of speaker verification system using each of the five vowels. False acceptance and false rejection are expressed in percentage out of total 50 trials conducted for each case.

Vowel	FA in %	FR in %
$/a/$	2	40
$/i/$	4	36
$/u/$	0	62
$/e/$	2	18
$/o/$	4	60

V. Online Speaker Verification System using source Features

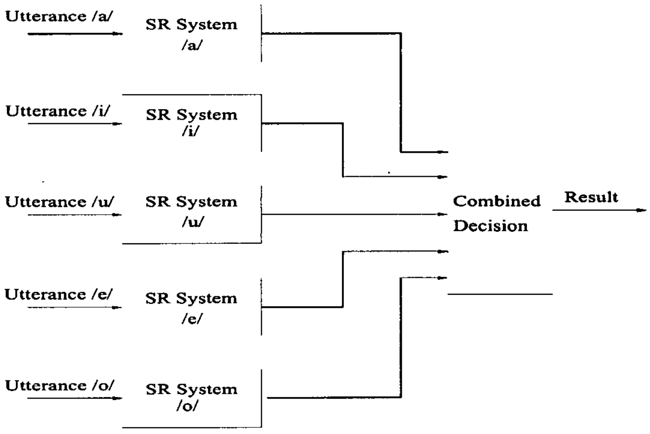

Fig. 10. Block diagram of online speaker verification system.

The block diagram of the proposed online speaker verification system is shown in Figure 10. As shown in the figure, the system uses speaker recognition systems built for each vowel, followed by a decision logic for combining the evidences from each of these systems, to come up with a decision for accepting or rejecting the claim of a speaker.

The online speaker verification system consists of two phases: (1) Enrollment phase and (2) verification phase. During enrollment, the speaker utters five vowels $/a/$, $/i/$, $/u/$, $/e/$ and $/o/$ in isolation. The speakers are instructed to speak these vowels for a duration of about 1-3 sec as naturally as possible. One model is generated for each vowel and hence we have five models for each speaker. The time taken to train one model is approximately 15 minutes (on Pentium III processor with Linux 6.0). During verification, the claimant has to speak each of the five vowels for about 0.5 sec. The test utterance of each vowel is given to all the enrolled speaker models of the same vowel. The average confidence values for all the vowels and their ranks are used to come up with the decision to accept or reject the claim of speaker.

The logic for combining the evidences from different verification systems is as follows:

The claimant model is accepted as genuine, if for the five vowel test utterances, any of the following conditions are satisfied, otherwise it is rejected.

- The claimant model has the majority compared to all other models
- The claimant model comes as *Rank*1 in exactly two cases and another speaker also comes as *Rank*1 in two other cases, but the average confidence for all the five vowels of claimant speaker is more than the average confidence of the other speaker

Performance of the online speaker verification system is evaluated for 20 cooperative speakers in terms of False Acceptance (FA) and False Rejection (FR). In this evaluation, we have considered 19 models of other speakers as background models for computing the ranks.

TABLE II

Performance of online speaker verification system.

Claimant	No. of Tests	FA in %	FR in %
Genuine	50	-	10
Impostor	50	0	-

Table II shows the performance of the online speaker verification system. Comparing the results in Table I and Table II, we can infer that by suitably combining the evidence from different vowels, we can come up with a

0-7803-7278-6/02/$10.00 ©2002 IEEE

decision for accepting or rejecting with minimum FA and FR values. Table II also demonstrates the usefulness of source information for speaker verification. The results also demonstrate that the AANN models have indeed captured the speaker-specific source information present in the LP residual.

VI. Conclusions

In this paper our objective is to demonstrate the feasibility of using source information for speaker recognition task. The recognition studies shows that source of excitation contains significant speaker-specific information. One more important point to be noted is that, these models do not require large amount of data as in the case of systems based on spectral features.

For any online system it is necessary to reduce the time for enrollment and verification. At present, for testing a given claimant model, we are considering the confidences values of all the frames. It is possible to evolve a frame selection criterion, which may improve the performance significantly.

References

[1] S. Furui, "Recent advances in speaker recognition," *Pattern Recognition Lett.*, vol. 18, pp. 859–872, 1997.

[2] A. E. Rosenberg, "Automatic speaker verification: A review," *Proc. IEEE*, vol. 64, pp. 475–487, Apr 1976.

[3] D. O'Shaughnessy, "Speaker recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, pp. 4–17, Oct. 1986.

[4] Joseph P. Campbell, "Speaker recognition : A tutorial," *Proc. IEEE*, vol. 85, pp. 1436–1462, Sept 1997.

[5] D. A. Reynolds, "Speaker identification and verification using gaussian mixture models," *Speech Communication*, vol. 17, pp. 91–108, Aug. 1995.

[6] NIST, "Speaker recognition workshop notebook," in *Proc. NIST Speaker Recognition Workshop*, Jun. 1999.

[7] B. Yegnanarayana, S. P. Kishore, and A. V. N. S. Anjani, "Neural network models for capturing probability distribution of training data," in *Proc. Int. Conf. Cognitive and Neural Systems*, Boston, 2000.

[8] B. Yegnanarayana and S. P. Kishore, "AANN-An alternative to GMM for pattern recognition," *Accepted for publication in Neural Netwroks.*

[9] S. P. Kishore, *Speaker verification using autoassociative neural netwrok models*, MS Thesis, Indian Institute of Technology Madras, Department of Computer Science and Engg., Chennai, 2000.

[10] G. R. Doddington, "Speaker recognition-Identifying people by their voices," *Proc. IEEE*, vol. 73, pp. 1651–1664, Nov. 1985.

[11] B. Yegnanarayana, A.S. Madhukumar and V.R. Ramachandran, "Robust features for applications in speech and speaker recognition," in *Proc. ESCA workshop speech processing in adverse conditions, Cannes Mandelieu, France*, Nov. 1992.

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

[13] S. R. Mahadeva Prasanna, Cheedella. S. Gupta, and B. Yegnanarayana, "Autoassociative neural network models for speaker verification using source features," in *Int. Conf. Cognitive and Neural Systems (Communicated)*, Boston, 2002.

[14] S. R. Mahadeva Prasanna, Cheedella. S. Gupta, and B. Yegnanarayana, "Source information from linear prediction residual for speaker recognition," *Communicated to the Journal of Acoustical Society of America.*

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

[16] M. A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," vol. 37, pp. 233–243, Feb. 1991.

[17] B. Yegnanarayana, K. Sharat Reddy, and S. P. Kishore, "Source and system features for speaker recognition using AANN models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2001.

[18] K. Sharat Reddy, *Source and system features for speaker recognition*, MS Thesis, Indian Institute of Technology Madras, Department of Computer Science and Engg., Chennai (submitted), Sept. 2001.

0-7803-7278-6/02/$10.00 ©2002 IEEE