CASCADE REALIZATION OF DIGITAL INVERSE FILTER FOR EXTRACTING SPEAKER DEPENDENT FEATURES

V.V.S. Sarma and B. Yegnanarayana
Department of Electrical Communication Engineering
Indian Institute of Science, Bangalore-560012, INDIA

Summary

Quest for new speaker dependent features is a constant problem in the design of automatic speaker recognition systems. In speech, information about the speaker usually arises along with the semantic information which makes its independent use difficult. In this paper, a method based on linear prediction (LP) analysis is described which yields features that are more speaker dependent than the usual linear predictor coefficients (LPC). In this method the LPC contours are obtained through cascade realization of digital inverse filtering (DIF) for speech signals. A low order (2-4) DIF removes the gross spectral characteristics such as the large dynamic range and some significant peaks which tend to mask the weaker formants. Visual comparison of the contours and a preliminary statistical analysis indicate that the LPC contours obtained by processing the output signal of the first stage contain better features for speaker dependency than the direct LPC contours.

Introduction

Linear Prediction (LP) analysis provides an efficient means of representing speech[1]. The parameters of the all-pole model of speech production assumed in LP analysis are obtained by minimizing the total squared error over a chosen length of speech segment. The LP formulation can be viewed as designing an optimum digital inverse filter (DIF) (an all-zero filter of order M) such that the output energy for a given input speech segment is minimized. In spectral domain this may be interpreted as minimizing the integrated ratio of the actual power spectrum, $P(\omega)$ of the original signal to the modelled spectrum, $\hat{P}(\omega)$. The inverse spectrum of the all-zero filter for a given order M, represents the best approximation to the envelope of the short time spectrum of the speech segment. The mathematical steps involved in the autocorrelation formulation of LP analysis are summarized below.

Speech samples: $\{s_n\}$ $n = 0,1,.., N-1$

Linearly predicted samples: $\hat{s}_n = \sum_{k=1}^{M} a_k s_{n-k}$

Total squared error: $E = \sum_{n=0}^{N-1} (s_n - \hat{s}_n)^2$

Normal equations, obtained by setting $\frac{\partial E}{\partial a_k} = 0$.

$$\sum_{k=1}^{M} a_k R_{|i-k|} = -R_i \quad , \quad i=1,2,.., M$$

where

$$R_k = \sum_{n=0}^{N-1-k} s_n s_{n+k}$$

Normalized error: $\eta = \frac{E}{R_0} = 1 + \sum_{k=1}^{M} a_k r_k$

where

$$r_k = R_k/R_0$$

Although the LP coefficients $\{a_k\}$ have been originally suggested to represent speech information their effectiveness in automatic speaker recognition has also been explored in recent studies[2,3,4]. The advantages of using $\{a_k\}$ in an automatic speaker recognition system are: (1) they are easily determined from speech and (2) they represent combined information about the formants, their bandwidths and the glottal waveform. Sambur[4] also observes that $\{a_k\}$ are slightly superior to formants for speaker recognition. As adjacent a_k contours are highly correlated, Sambur suggests that only LPC contour sets which are widely spaced in number (such as a_2-a_4-a_7) need be used to obtain good speaker recognition scores. Also the recognition potential is not reduced if the order of linear prediction is reduced from 12 to 8.

To exploit the LP analysis technique for extracting better speaker dependent features, it is worthwhile to note the nature of spectral envelope produced by the DIF. Since the coefficients are obtained by minimizing $P(\omega)/\hat{P}(\omega)$, the approximation would be better at the peaks of the $\hat{P}(\omega)$[1]. If there is a sharp formant peak the approximation at the peak would improve as M is increased. It is quite possible that increasing the order may not provide good approximation to higher formant peaks which were found to be good speaker dependent features[5]. As a result $\{a_k\}$ may not possess all the significant speaker information. This also explains Sambur's observation that recognition accuracy is not reduced by reducing the order of the DIF from 12 to 8.

Authorized licensed use limited to: INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY. Downloaded on August 26, 2009 at 01:56 from IEEE Xplore. Restrictions apply.

The high correlation between adjacent LPC contours is also not entirely unexpected if the η vs. M curve[6] is observed. For the addition of every odd coefficient only a real zero is added, which approximates the slope of the spectrum but not the formant peaks. This is because two coefficients are required to specify a complex pair of roots which produce a formant peak.

Another point which is worth noting is that one obtains a surprisingly high intelligibility of speech synthesized using only a predictor of order 2, which was utilized by Makhoul in speech recognition[7,8]. This clearly shows that most of the semantic information in speech can be obtained in the temporal variation of the coefficients of a low order predictor, which approximate only the gross spectral behaviour. Hopefully, the speaker dependency is present to a significant extent in the output of an inverse filter of a very low order.

A quick survey of speaker recognition literature indicates that most of the 'systems' considered so far are exploratory studies establishing the feasibility of automatic speaker recognition[9]. Most of these works concentrate on search for efficient features providing a speaker's identity such as spectral features, pitch, formants, nasals and those derived from LPC's. Wolf[10] and Sambur[4] recently stressed the importance of feature selection and evaluation. A feature could be a scalar feature or a vector feature describing the contour of a particular parameter in an utterance. The objective of this paper is to investigate the speaker dependency of the cascaded DIF coefficient contours.

Cascade Realization

The output of a low order DIF for an input speech signal contains information about pitch, formants and glottal waveform since the first stage merely removes the gross spectral features. To extract the information about formants an 8th or 10th order (corresponding to 4 or 5 formants) DIF should be derived from the output of the first stage. This method of cascading two stages of DIF has been suggested by Markel[6] to extract the weaker and closely spaced formants and also by Makhoul[1] to reduce the dynamic range of the speech spectrum thus reducing the ill-conditioning of the autocorrelation matrix.

Fig. 1 shows the reciprocal of the DIF spectrum when the direct form (M = 12) and a cascade form of two stages (M_1 = 2 and M_2 = 8) are used. It is very clear that the third formant which does not appear in Fig. 1a shows up well in Fig. 1b. This formant information is reflected in the second stage DIF coefficients which provide useful feature for speaker recognition. Fig. 2 describes both the direct and cascaded forms of DIF. If we assume that the total mean squared error is same in both cases, (it will not be, in general) we may write

$$1 + \sum_{k=1}^{M} a_k z^{-k} = \left(1 + \sum_{k=1}^{M_1} b_k z^{-k}\right)\left(1 + \sum_{k=1}^{M_2} c_k z^{-k}\right) \quad (1)$$

where

$$M = M_1 + M_2$$

It follows that

$$a_1 = b_1 + c_1$$

$$a_2 = b_2 + b_1 c_1 + c_2 \quad (2)$$

and so on.

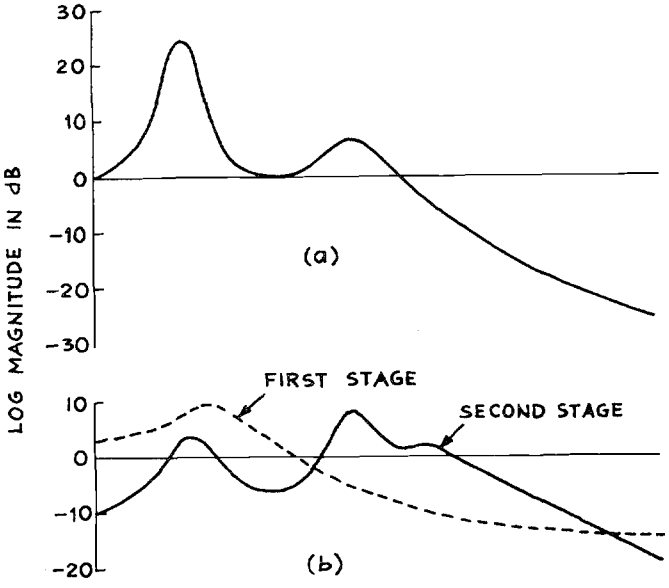

Fig. 1: Reciprocal of DIF spectrum for a voiced speech signal (a) Direct form(M=12) (b) cascaded form (M_1=2, M_2=8)

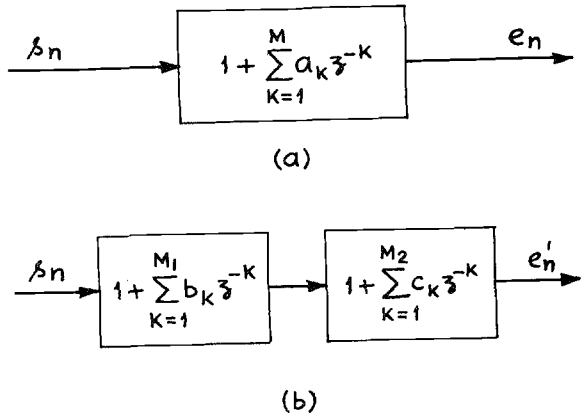

Fig.2: Digital inverse filtering (a)Direct form (b) cascaded form

Authorized licensed use limited to: INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY. Downloaded on August 26, 2009 at 01:56 from IEEE Xplore. Restrictions apply.

Equations (1) and (2) give the relation between processing of speech by direct DIF and cascaded DIF and these provide the basis on which the features from the two approaches can be compared.

In Fig. 3 the contours of a_1 for two speakers showing the inter and intra speaker variation are given. The remarkable similarity of the shape of the contours may be observed. The coefficient contours of first stage of cascaded DIF (b_1 contours) also are strikingly similar. On the other hand the coefficient contours of second stage (c_1 contours) bring out the inter-speaker variation in the features. A detailed statistical analysis is being performed on these features to support our hypothesis.

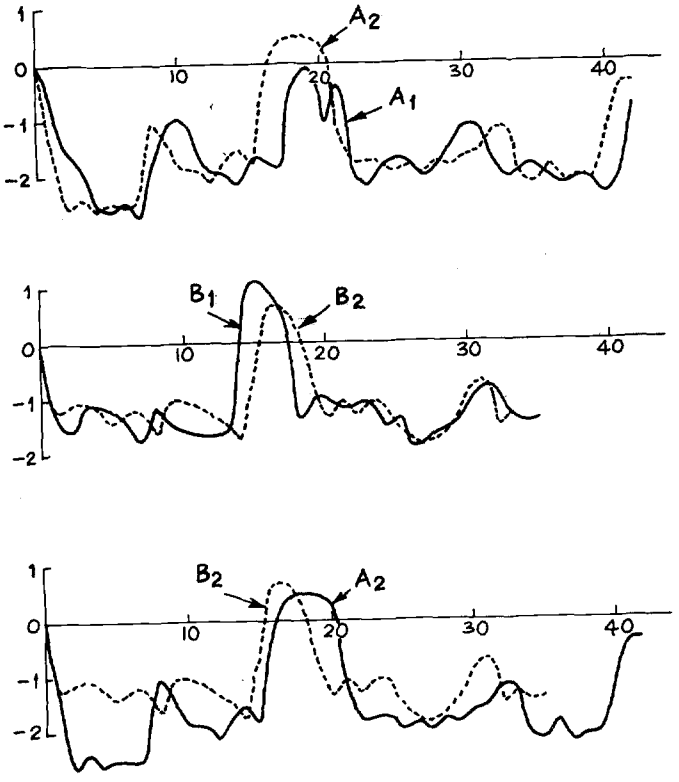

Fig. 3: a_1 contour for the utterance 'Have you seen Bill?' showing inter and intra speaker variation

A_i : utterance i(i=1,2) of speaker A

B_i : utterance i(i=1,2) of speaker B

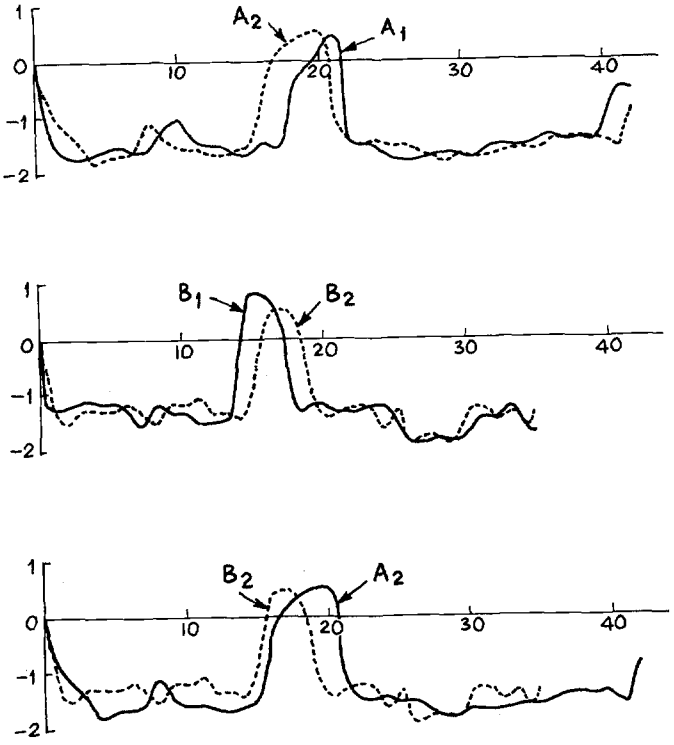

Fig. 4: b_1 contour for the utterance 'Have you seen Bill?' showing inter and intra speaker variation

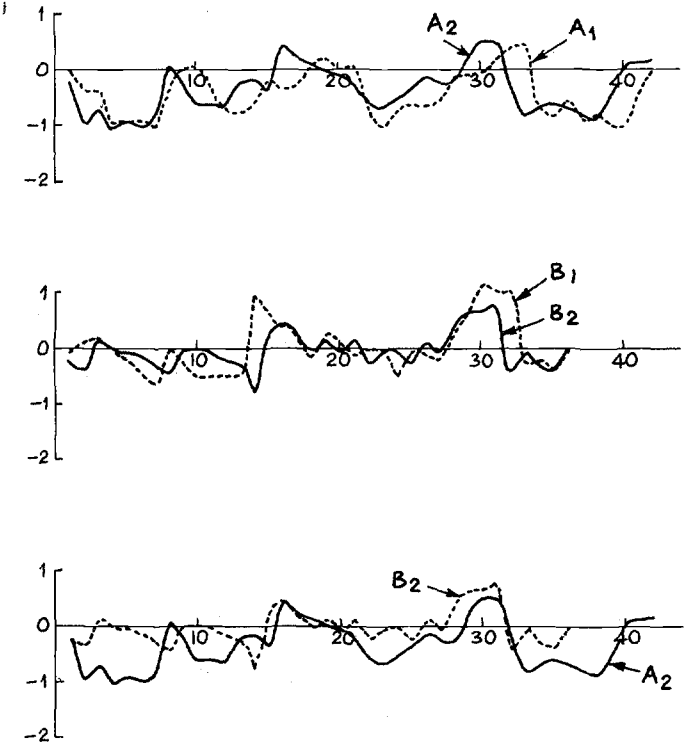

Fig.5: c_1 contour for the utterance 'Have you seen Bill?' showing inter and intra speaker variation

Authorized licensed use limited to: INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY. Downloaded on August 26, 2009 at 01:56 from IEEE Xplore. Restrictions apply.

Feature Evaluation

This section describes the data base and procedures being followed to demonstrate the use of cascaded DIF coefficients as features in an automatic speaker recognition scheme.

Data Base

Examples of speech from 11 adult Indian speakers are used in these studies. The speakers are all staff of Indian Institute of Science with 9 men and 2 women. The recording was done in a single session. The speakers were not given any special training and were asked to speak normally and they were informed of the nature of the experiment. Each speaker uttered his/her name followed by 10 utterances of each of the two sentences

(i) Have you seen Bill?

(ii) We were away a year ago.

The speakers belong to different linguistic groups of India (5 Telugu, 4 Kannada, 1 Malayalam, 1 Marathi) and English was not the mother tongue of any of them. The English pronunciation of an average Indian is influenced by his native language and by itself may provide a clue to speaker recognition but this aspect is not explored in the present study.

The analog recording of these utterances was done in an anechoic chamber with a BEL studio type single channel tape recorder and a Shure microphone. The speech was sampled at 20 KHz without prefiltering and quantized uniformly to 12 bit accuracy and stored on digital magnetic tape.

Evaluation Procedure

The feature that was considered for preliminary analysis was c_1 contour. Each utterance has a duration of about 0.83 to 1 sec. Only the utterance 'Have you seen Bill?' of 5 speakers has been used in this analysis. Each contour spanned 35-42 frames of 476 samples each and thus each is a vector of 35-42 components. The average value of c_1 over the speech segment 'Have' is considered as a scalar feature and compared with the corresponding value of a_1 using F-ratio[10]. The c_1 and a_1 contours are also compared as vector features by means of Wilk's lamba statistic.[11]

Discussion

The F-ratio obtained for the particular feature selected from c_1 is 100 as compared to 84 for the corresponding feature from a_1. Similarly the Wilk's lambda statistic has also shown that c_1 is superior to a_1. While this is yet an inadequate proof for establishing the superiority of c_1 contour over a_1 contour as a speaker dependent feature, it supports the hypothesis proposed in the paper. However, a detailed analysis of all the coefficient contours $\{c_n\}$ over the complete data set can confirm the validity of this conclusion. Statistical analysis by itself does not solve the feature selection problem. Performance error in a simulated recognition scheme utilizing this particular feature provides the final answer towards its usefulness.

References

1. J. Makhoul, 'Linear Prediction: A Tutorial Review', Proc. IEEE, Vol. 63, pp.561-580, April 1975.

2. B.S. Atal, 'Effectiveness of Linear Prediction Characteristics of Speech for Automatic Speaker Identification and Verification', J.Acoust.Soc.Amer., Vol.55, pp. 1304-1312, 1974.

3. A.E. Rosenberg and M.R. Sambur, 'New Techniques for Automatic Speaker Verification', IEEE Trans. Acoust., Speech, and Sig.Processing, Vol. ASSP-23, pp.169-176, April 1975.

4. M.R. Sambur, 'Selection of Acoustic Features for Speaker Identification', IEEE Trans.Acoust., Speech, and Sig.Processing Vol. ASSP-23, pp. 176-182, April 1975.

5. D. Lewis and C. Tuthill, "Resonant Frequencies and Damping Constants of Resonators Involved in the Production of Sustained Vowels'O' and 'Ah'", J.Acoust.Soc. Amer., Vol. 11, pp. 451-456, 1940.

6. J.D. Markel, 'Formant Trajectory Estimation from a Linear Least-Squares Inverse Filter Formulation', SCRL Monograph 7, Speech Communication Research Laboratory, Santa Barbara, California, October 1971.

7. B.S. Atal and S.L. Hanauer, 'Speech Analysis and Synthesis by Linear Prediction of the Speech Wave', J.Acoust.Soc.Amer., Vol. 50, pp 637-655, August 1971.

8. J. Makhoul, 'The use of a two-pole linear prediction model in speech recognition', BBN Rep. 2537, 1973.

9. V.V.S. Sarma and B. Yegnanarayana, 'Automatic Voice Recognition: Problems and Prospects', Rept.49, Dept. ECE, IISc, Bangalore, Dec. 1975.

10. J.J. Wolf, 'Efficient Acoustic Parameters for Speaker Recognition', J.Acoust.Soc. Amer., Vol. 51, pp. 2044-2056, 1972.

11. W.W. Cooley and P.R. Lohnes, 'Multivariate Data Analysis' John Wiley, New York, 1971.

Authorized licensed use limited to: INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY. Downloaded on August 26, 2009 at 01:56 from IEEE Xplore. Restrictions apply.