Signal Processing 4 (1982) 323–327
North-Holland Publishing Company

SHORT COMMUNICATION

EVALUATION OF VARIOUS LINEAR PREDICTION PARAMETRIC REPRESENTATIONS IN VOWEL RECOGNITION

K.K. PALIWAL and P.V.S. RAO

Speech and Digital Systems Group, Tata Institute of Fundamental Research, Homi Bhabha Road, Bombay 400005, India

Received 22 December 1980
Revised 13 July 1981

Abstract. Several alternate linear prediction parametric representations are experimentally compared as to their vowel recognition performance. The speech data used for this purpose consist of 900 utterances of 10 different vowels spoken by 3 speakers in a /b/-vowel-/b/ context. The cepstral coefficients representation is found to be the best linear prediction parametric representation.

Zusammenfassung. In diesem Beitrag werden verschiedene parametrische Darstellungen von Sprachsignalen, die auf dem Prinzip der linearen Prädiktion (LPC) basieren, auf ihre Verwendbarkeit in einem Spracherkennungssystem für Vokale experimentell untersucht. Das verwendete Sprachmaterial besteht hierbei aus 900 Beispielen von 10 verschiedenen Vokalen in einem /b/-Vokal-/b/-Kontext, gesprochen von 3 Sprechern. Die Darstellung der LPC-Parameter mit Hilfe der Cepstrumkeoffizienten erwies sich als die günstigste.

Résumé. Plusieurs représentations paramétriques déduites de la prédiction linéaire sont comparées expérimentalement en ce qui concerne leurs performances en reconnaissance des voyelles. Les données de parole utilisées pour cela sont 900 échantillons de 10 voyelles différentes prononcées par 3 locuteurs dans le contexte /b/ voyelle /b/. Les coefficients cepstraux se révèlent les meilleurs paramètres, suivis par les coefficients de la réponse impulsionnelle du filtre auto-régressif.

Keywords. Parametric representation of speech, linear prediction analysis, vowel recognition, distance measure.

1. Introduction

Linear prediction (LP) analysis has been used extensively over the last several years for speech processing applications such as speech analysis-synthesis [1, 2], speech recognition [3–8] and speaker recognition [9, 10]. LP analysis of speech can lead to a number of parametric representations, all of which provide equivalent information about the linear predictor. These parametric representations have been compared in a speaker recognition task by Pfeifer [9] and Atal [10] and in a word recognition task by Ichikawa *et al.* [3] and Stella [8]. However, there is no report in the literature of any comparative study of their suitability in an acoustic-phonemic recognition system. Since most of the phonemes (about 38.2% [11]) occurring in conversational English are vowels, we will compare here several alternate LP parametric representations as to their performance in the recognition of ten different vowels. Hopefully, the results obtained from this vowel recognition experiment will be useful in the design of a general acoustic-phonemic recognition system for continuous speech.

2. Alternate LP parametric representations

The main assumption on which LP analysis is based is that speech can be modelled as the output

0165-1684/82/0000–0000/$02.75

of an Mth order all-pole filter of the form

$$H(z) = G/(1 + \sum_{n=1}^{M} a_n z^{-n})$$

where a_n, $1 \leq n \leq M$, are the coefficients of the inverse filter $A(z) = 1 + \sum_{n=1}^{M} a_n z^{-n}$ and G is the gain of the filter.

Various LP parametric representations which uniquely define the inverse filter $A(z)$ and are used in the present investigation are listed below. (For details about these LP parametric representations and their estimation procedure, see [12] and [13].)

(1) Impulse response of the inverse filter $A(z)$ (or the predictor coefficients), a_n, $1 \leq n \leq M$.

(2) Impulse response of the all-pole filter $H(z)$, h_n, $0 \leq n \leq M$.

(3) Autocorrelation coefficients of $\{a_n\}$,

$$b_n = \sum_{k=0}^{M-n} a_k a_{k+n}, \qquad a_0 = 1, \quad 0 \leq n \leq M.$$

(4) Autocorrelation coefficients of $\{h_n\}$[1],

$$R_n = \sum_{k=0}^{\infty} h_k h_{k+n}, \quad 0 \leq n \leq M.$$

(5) Cepstral coefficients of $A(z)$,

$$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log A(e^{j\omega})\, e^{jn\omega}\, d\omega, \quad 1 \leq n \leq M.$$

(6) Area coefficients A_n, $1 \leq n \leq M$.

(7) Reflection coefficients k_n, $1 \leq n \leq M$.

(8) Poles of the all-pole filter $H(z)$ (or zeros of $A(z)$).

The poles of $H(z)$ can not be used directly as recognition parameters because they are not naturally ordered (i.e., interchanging the values of two poles does not change the filter $H(z)$). We therefore order these poles artificially by using their resonance frequencies. The real and imaginary parts of these ordered poles in the upper half of the z-plane are used as recognition parameters. Also, some of the parametric representations listed above have M parameters while others have $M+1$ parameters. For the latter representations, the M parameters (from the first to the Mth) are normalised by dividing them by the zeroth parameter, thus keeping the number of parameters the same (i.e., M) for all the parametric representations.

[1] In the autocorrelation method of linear prediction, the first $M+1$ autocorrelation coefficients of the impulse response of $H(z)$ are identical to the corresponding autocorrelation coefficients of the speech signal [14]. These coefficients are the starting point of LP analysis and can be computed from the speech signal s_n, $n = 0, 1, \ldots, N-1$ as follows: $R_n = \sum_{k=0}^{N-1-n} s_k s_{k+n}$, $0 \leq n \leq M$.

3. Data acquisition and preprocessing

The speech data consist of 900 utterances, having 30 repetitions of 10 different /b/-vowel-/b/ syllables, spoken by 3 speakers (2 male and 1 female). Recording of these utterances is done in an ordinary office room. The speech signal is digitised at a sampling rate of 10 kHz by means of a 12-bit analog-to-digital converter and stored on magnetic tape for further processing. A lowpass filter with a cutoff frequency of 4 kHz is used as a dealiasing filter.

The steady-state part of the vowel segment is manually located for each of the 900 utterances and a 20 msec segment is excised from its centre. A 10-th order LP analysis is performed and various LP parametric representations are derived from each of these 20 msec segments. The autocorrelation method of linear prediction (with 20 msec Hamming window and without pre-emphasis) is used here for analysis. LP analysis of the speech signal is done on the general purpose computer, DEC System 10, using the floating point arithmetic.

4. Recognition procedure

The aim here is to classify the 10-dimensional vectors (each vector has 10 LP parameters as its components) representing the vowel segments into ten vowel classes: /i/, /I/, /e/, /æ/, /ʌ/, /a/, /ɔ/,

/o/, /U/ and /u/. This is a standard problem in statistical pattern recognition and has been treated exhaustively in the literature [15]. In the present paper, the classification scheme used is the forced decision pattern matching method and is studied using three different distance measures:

(1) *Correlation distance measure*

The correlation distance measure d_j for the jth class is defined here as

$$d_j^2 = 1 - C_j$$

where C_j is the normalised correlation between the test vector X and the mean vector M_j of the jth class and is given by

$$C_j = \frac{X^t M_j}{(X^t X)^{1/2}(M_j^t M_j)^{1/2}}.$$

The superscript t denotes here the transpose of the vector.

(2) *Euclidean distance measure*

The distance measure d_j for the jth class is given here by

$$d_j^2 = (X - M_j)^t (X - M_j).$$

(3) *Mahalanobis distance measure*

The distance measure d_j for the jth class is given by

$$d_j^2 = (X - M_j)^t W^{-1} (X - M_j)$$

where W is the pooled intraclass covariance matrix.

The test vector X is classified here into the ith class if $d_i < d_j$ for all $j \neq i$.

The mean vectors for all the ten vowel classes and the pooled intraclass covariance matrix are computed from the data in the training set by using the following relations:

$$M_j = \frac{1}{N_j} \sum_{i=1}^{N_j} X_{ji}, \quad 1 \leq j \leq 10,$$

and

$$W = \sum_{j=1}^{10} \frac{1}{N_j} \sum_{i=1}^{N_j} (X_{ji} - M_j)^t (X_{ji} - M_j)$$

where N_j is the number of preclassified vectors in the jth class and X_{ji} the ith vector of the jth class.

5. Results and discussion

Vowel recognition performance of all the LP parametric representations is studied here separately for all the three speakers. In the present experiment, speaker specific training is used; i.e., both the training and test data are derived from the same speaker.

In order to estimate the vowel recognition performance, we have, for each speaker, a fixed sample of 300 preclassified vectors (obtained from 30 repetitions for each of the 10 vowel classes). This fixed sample can be used, as suggested by Toussaint [16], in a number of ways to estimate the recognition performance. We use here the following procedure for estimating the recognition performance. For each vowel class, twenty-nine repetitions are used as the training set and the thirtieth repetition is used as the test set. Each of the 30 repetitions is used in turn as the test set. All the 300 vectors of a given speaker are thus classified into 10 vowel classes.

In Table 1, we show the vowel recognition performance of the eight LP parametric representations using the correlation distance measure.

Table 1

Vowel recognition performance for 3 speakers using the correlation distance measure

LP parametric representation	Recognition performance (in %) for		
	First male speaker	Second male speaker	Female speaker
1. Predictor coefficients $\{a_n\}$	74.7	82.7	75.7
2. Impulse response $\{h_n\}$ of the all-pole filter	91.3	86.0	78.3
3. Autocorrelation coefficients of $\{a_n\}$	54.7	67.7	66.3
4. Autocorrelation coefficients of $\{h_n\}$	80.7	79.3	73.7
5. Cepstral coefficients	94.0	93.7	84.7
6. Area coefficients	74.0	61.7	75.3
7. Reflection coefficients	84.0	83.7	75.3
8. Poles of the all-pole filter	61.3	69.7	51.3

Vowel recognition scores are listed here separately for each of the three speakers. It can be seen from this table that the cepstral coefficients representation gives consistently better recognition results than the other LP parametric representations for all the three speakers.

In Table 2, we pool the recognition scores of the three speakers and show the recognition performance of the eight LP parametric representations for three different distance measures (namely, the correlation, Euclidean and Mahalanobis distance measures). Even here, we see that the cepstral coefficients representation is the best LP parametric representation for all the three distance measures[2].

Thus, the cepstral coefficients representation consistently ranks first among the eight LP parametric representations for different speakers and for different distance measures used. This signifies the importance of this LP parametric representation in a vowel recognition task. It is interesting to note here that the cepstral coefficients representation was found to be the best LP parametric representation in a speaker recognition task by Atal [10] and in a word recognition task by Ichikawa *et al.* [3] and Stella [8]. (Atal [10], Ichikawa *et al.* [3] and Stella [8] have studied only some of the LP parametric representations investigated in this paper.)

Table 2

Vowel recognition performance using three different distance measures

LP parametric representation	Recognition performance (in %) using		
	Correlation dist. measure	Euclidian dist. measure	Mahalanobis dist. measure
1. Predictor coefficients $\{a_n\}$	77.7	70.3	91.2
2. Impulse response $\{h_n\}$ of the all-pole filter	85.2	87.4	92.0
3. Autocorrelation coefficients of $\{a_n\}$	62.9	60.7	79.7
4. Autocorrelation coefficients of $\{h_n\}$	77.9	80.0	87.1
5. Cepstral coefficients	90.8	91.4	96.0
6. Area coefficients	70.3	55.4	80.9
7. Reflection coefficients	81.0	82.6	89.2
8. Poles of the all-pole filter	60.8	60.6	83.7

[2] Itakura distance measure is specially suited for comparing two speech segments represented in terms of the LP coefficients [4, 17–19]. This measure is under investigation and the detailed results will be presented in a later paper. However, the vowel recognition performance using this measure is found to be 94.2% for the three speakers.

We can make a few other observations from Tables 1 and 2. These observations are listed below.

(1) The impulse responses $\{a_n\}$ and $\{h_n\}$ are better parametric representations for vowel recognition than their respective autocorrelation coefficients $\{b_n\}$ and $\{R_n\}$.

(2) The poles of the all-pole filter are very poor recognition parameters (yielding a recognition score of 60.6% when used with the Euclidean distance measure). If, on the other hand, the first three formant frequencies extracted from these poles are used as the recognition parameters, vowel recognition performance improves significantly (i.e., by 24%). (The first three formant frequencies are extracted from the poles of $H(z)$ by making the decisions about spurious poles manually.)

(3) Both the correlation and Euclidean distance measures use first order statistics. But the correlation distance measure has an important property that it remains unchanged even if either the test vector X or the mean vector M_j or both are multiplied by a constant [20], [21]. This property is not always desirable but can be of considerable advantage in some situations where parameter variations due to some random scale factor are to be ignored. This can be seen from Table 2 where the correlation distance measure gives much better recognition results than the Euclidean distance measure for area coefficients representation.

(4) Mahalanobis distance measure uses second order statistics and thus is more complex than the

correlation and Euclidean distance measures. This measure gives better recognition results than the correlation and Euclidean distance measures for all the eight LP parametric representations (see Table 2).

6. Conclusion

Eight different LP parametric representations (namely, predictor coefficients $\{a_n\}$, impulse response $\{h_n\}$ of the all-pole filter, autocorrelation coefficients of $\{a_n\}$, autocorrelation coefficients of $\{h_n\}$, cepstral coefficients, area coefficients, reflection coefficients and poles of the all-pole filter) are experimentally compared with respect to their vowel recognition performance. Although all these parametric representations provide equivalent information about the linear predictor, their vowel recognition capabilities are shown to be different. The cepstral coefficients representation is shown to be the best LP parametric representation.

Acknowledgement

The authors are thankful to the referees for their constructive comments, and to Mr. Dinesh Sharma, Mr. Akhil Ranjan and Mrs. Rukmani Paliwal for providing their voices for experimentation.

References

[1] B.S. Atal and S.L. Hanauer, "Speech analysis and synthesis by linear prediction", *J. Acoust. Soc. Amer.*, Vol. 50, No. 2, Aug. 1971, pp. 637–655.

[2] J.D. Markel and A.H. Gray, Jr., "A linear prediction vocoder simulation based upon the autocorrelation method", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-22, No. 2, Apr. 1974, pp. 124–134.

[3] A. Ichikawa, Y. Nakano and K. Nakata, "Evaluation of various parameter sets in spoken digits recognition", *IEEE Trans. Audio Electroacoust.*, Vol. AU-21, No. 3, June 1973, pp. 202–209.

[4] F. Itakura, "Minimum prediction residual principle applied to speech recognition", *IEEE Trans. Acoust., Speech Signal Process.*, Vol. ASSP-23, No. 1, Feb. 1975, pp. 67–72.

[5] M.R. Sambur and L.R. Rabiner, "A statistical decision approach to the recognition of connected digits", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-24, No. 6, Dec. 1976, pp. 550–558.

[6] H. Wakita, "Normalization of vowels by vocal-tract length and its application to vowel identification", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-25, No. 2, Apr. 1977, pp. 183–192.

[7] L.R. Rabiner and M.R. Sambur, "Application of an LPC distance measure to the voiced-unvoiced-silence detection problem", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-25, No. 4, Aug. 1977, pp. 338–343.

[8] M. Stella, "Comparaison de differents coefficients de prediction lineaire pour la reconnaissance des mots isoles", *Proc. Speech Symposium*, Budapest, Sept. 30–Oct. 2, 1980, pp. 129–134.

[9] L.L. Pfeifer, "Inverse filter for speaker identification", RADC-TR-74-214, Final Report, Speech Communications Research Laboratory, Santa Barbara, CA, 1974.

[10] B.S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", *J. Acoust. Soc. Amer.*, Vol. 55, No. 6, June 1974, pp. 1304–1312.

[11] M.A. Mines, F. Hanson and J.E. Shoup, "Frequency of occurrence of phonemes in conversational English", *Language and Speech*, Vol. 21, Part 3, July–Sept. 1978, pp. 221–235.

[12] J.D. Markel and A.H. Gray, Jr., *Linear Prediction of Speech*, Springer-Verlag, Berlin, 1976.

[13] R. Viswanathan and J. Makhoul, "Quantization properties of transmission parameters in linear predictive systems", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-23, No. 3, June 1975, pp. 309–321.

[14] J. Makhoul, "Linear prediction: A tutorial review", *Proc. IEEE*, Vol. 63, No. 4, Apr. 1975, pp. 561–580.

[15] R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.

[16] G.T. Toussaint, "Bibliography on estimation of misclassification", *IEEE Trans. Information Theory*, Vol. IT-20, No. 4, July 1974, pp. 472–479.

[17] A.H. Gray, Jr. and J.D. Markel, "Distance measures for speech processing", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-24, No. 5, Oct. 1976, pp. 380–391.

[18] R.L. Kashyap, "Optimal feature selection and decision rules in classification problems with time series", *IEEE Trans. Information Theory*, Vol. IT-24, No. 3, May 1978, pp. 281–288.

[19] J.M. Tribolet, L.R. Rabiner and M.M. Sondhi, "Statistical properties of an LPC distance measure", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-27, No. 5, Oct. 1979, pp. 550–558.

[20] B.S. Atal, "Automatic recognition of speakers from their voices", *Proc. IEEE*, Vol. 64, No. 4, Apr. 1976, pp. 460–475.

[21] H.F. Silverman and N.R. Dixon, "A comparison of several speech-spectra classification methods", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-24, No. 4, Aug. 1976, pp. 289–295.