S10.22

A STUDY OF LINE SPECTRUM PAIR FREQUENCIES FOR SPEECH RECOGNITION

K.K. Paliwal
Computer Systems and Communications Group
Tata Institute of Fundamental Research
Homi Bhabha Road, Bombay-400005, India

Abstract- The line spectrum pair (LSP) frequency representation has recently been proposed as an alternative linear prediction (LP) parametric representation. In the context of speech coding, this representation shows better quantization properties than the other LP parametric representations. In the present paper, the LSP representation is studied for speech recognition. Several distance measures based on this representation are investigated. The weighted LSP distance measure is found to result in the best performance. The performance of the weighted LSP distance measure is compared with that of the other popular LP distance measures (such as the Itakura, cepstral, weighted cepstral, root-power-sum, log area ratio and reflection coefficient distance measures). The weighted LSP distance measure is found to perform significantly better than these popular LP distance measures.

1. Introduction

Linear prediction (LP) analysis has been used extensively in the area of speech recognition over the last several years. It is possible to derive a number of parametric representations from the LP analysis of speech. Though each of these representations contain equivalent information about the LP spectral envelope, these parametric representation can, in general, lead to different recognition performances. In our earlier paper [1], we have studied different linear prediction (LP) parametric representations for speech recognition and found cepstral coefficient representation to be the best. Later on, we proposed [2] the following two distance measures based on the cepstral coefficient representation: 1) the quefrency-weighted cepstral distance measure (also known as the root-power-sum (RPS) distance measure) and 2) the statistically weighted cepstral distance measure (where the weights are inversely proportional to the standard deviations computed statistically from the speech data contained in the training set). We have found [2] that these weighted cepstral distance measures give better recognition results than the cepstral distance measure. Recently, these weighted cepstral distance measures have been studied by various researchers and a number of papers confirming these results have been reported in the literature [3-10].

In the present paper, we study the performance of the line spectrum pair (LSP) frequency representation for speech recognition. Application of this representation for speech recognition has not been reported so far in the literature. The LSP representation has recently been proposed by Itakura [11] as an alternative LP parametric representation. In the context of speech coding, a number of authors have shown [12,13] that this representation has better quantization properties than the other LP parametric representations (such as log area ratios and reflection coefficients). The LSP representation is capable of reducing the bit-rate for transmitting LP coefficients by 25-30% without degrading the coded speech quality [13]. Our interest in the present paper is to see whether we can get similar advantage from the LSP representation for speech recognition.

In the present paper, we study different distance measures based on LSP representation for speech recognition. For this, we use the single-frame vowel recognition system as a test-bed. We prefer this system over the more general word recognition system because of the following two reasons. Firstly, the purpose of a distance measure is to provide a quantitative measure of dissimilarity between the speech spectra of two different frames. If the word recognition system is used as a test-bed to evaluate the distance measure, the resulting recognition performance may not only be affected by the properties of the distance measure, but also by the other components of the recognition system (such as the dynamic time warping component) and the manner in which the distances from the individual frames are combined to compute the total distance between two words. Hence, it is not possible to say something conclusively about the distance measure on the basis of the results derived from the word recognition system. Secondly, most of the phonemes in conversational speech are

CH2561-9/88/0000-0485 $1.00 © 1988 IEEE

vowels. (For example, conversational English has about 38.2% vowels [14].) Because of these reasons, we use here the single-frame vowel recognition system for studying the LSP representation and the conclusions derived here are expected to be meaningful for more general speech recognition systems.

2. The LSP representation

In the LP analysis of speech, a short segment of speech signal is assumed to be generated as the output of a time-invariant linear all-pole filter $H(z)=1/A(z)$, where $A(z)$ is the inverse filter given by

$$A(z)=1+a_1z^{-1}+\ldots+a_Mz^{-M}.$$

Here M is the order of LP analysis and $\{a_i\}$ are the LP coefficients.

In order to define the LSP frequencies, the inverse filter polynomial is decomposed into two polynomials

$$P(z)=A(z)+z^{-(M+1)}A(z^{-1})$$

and

$$Q(z)=A(z)-z^{-(M+1)}A(z^{-1})$$

The roots of the polynomials $P(z)$ and $Q(z)$ are called the LSP frequencies. The LSP polynomials $P(z)$ and $Q(z)$ have the following two properties [12]: (1) All zeros of $P(z)$ and $Q(z)$ lie on the unit circle, and (2) Zeros of $P(z)$ and $Q(z)$ are interlaced with each other. These properties help in efficient numerical computation of the LSP frequencies from $P(z)$ and $Q(z)$. Using these two properties, it can be shown that (1) the minimum phase property of $A(z)$ can be easily preserved after quantization of LSP frequencies and (2) the LSP frequencies are amenable to interpolation.

3. Data acquisition and preprocessing

The speech data used in the present study consists of 900 utterances, having 30 repetitions of 10 different /b/-vowel-/b/ syllables, spoken by three speakers (2 male and one female). Recording of these utterances is done in an ordinary office room. The speech signal is digitized at a sampling rate of 10 kHz by means of 12-bit analog-to-digital converter. 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 the LSP frequencies are extracted from each of these 20 msec segments. The LP analysis is done here using the autocorrelation method (with 20 msec Hamming window and without preemphasis [15]).

4. Recognition procedure

The aim here is to classify the 10-dimensional vectors (each vector has 10 LSP frequencies 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 exhaustively treated in the literature [16]. In the present paper, we use the minimum distance classifier for vowel recognition. We study it for the following distance measures:

(1) Euclidean distance measure: It is given for the i-th class by

$$d_i=[(\mathbf{x}-\mathbf{m}_i)^t(\mathbf{x}-\mathbf{m}_i)]^{1/2},$$

where $\mathbf{x}$ is the test vector and $\mathbf{m}_i$ the mean vector of the i-th class. The mean vectors are computed from the data in the training set as follows:

$$\mathbf{m}_i=\frac{1}{N_i}\sum_{j=1}^{N_i}\mathbf{y}_{ij}, \quad 1\leqslant i\leqslant 10,$$

where N_i is the number of preclassified (training set) vectors in the i-th class and $\mathbf{y}_{ij}$ the j-th vector of the i-th class.

(2) Weighted Euclidean distance: It is given for the i-th class by

$$d_i=[\sum_{j=1}^{M}\{w_{ij}(x_j-m_{ij})\}^2]^{1/2},$$

where x_j and m_{ij} are the j-th components of the vectors $\mathbf{x}$ and $\mathbf{m}_i$, respectively, and w_{ij} is the weight associated with the j-th component for the i-th class and is computed statistically from the data in the training set by taking it to be inversely proportional to standard deviation; i.e.,

$$w_{ik}=K_i[\frac{1}{N_i}\sum_{j=1}^{N_i}(y_{ijk}-m_{ik})^2]^{-1/2},$$

where y_{ijk} is the k-th component of the vector $\mathbf{y}_{ij}$. The proportionality constants K_i for all the classes ($i=1,2,\ldots,10$) are determined from the constraints

$$\prod_{j=1}^{M}w_{ij}=1.$$

(3) Mahalanobis distance measure: It is given for the i-th class by

$$d_i=[(\mathbf{x}-\mathbf{m}_i)^t\mathbf{W}_i^{-1}(\mathbf{x}-\mathbf{m}_i)]^{1/2}$$

where $\mathbf{W}_i$ is the covariance matrix of i-th class. It is computed from the data in training set as follows:

$$\mathbf{W}_i=\frac{1}{N_i}\sum_{j=1}^{N_i}(\mathbf{y}_{ij}-\mathbf{m}_i)^t(\mathbf{y}_{ij}-\mathbf{m}_i).$$

5. Results

Different distance measures based on the LSP representation are studied here as to their vowel recognition performance for each of the three speakers. In the recognition experiment, speaker-specific training is used; i.e., both the training and the 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 in a number of ways to estimate the recognition performance [17]. 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 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 three distance measures based on LSP representation. Vowel recognition accuracies are listed here for each of the three speakers separately. It can be seen from this table that the weighted Euclidean distance measure results in the best performance. It might be noted that the Mahalanobis distance measure uses more information about the data and, hence, is more powerful than the weighted Euclidean distance measure. But, in the present experiment, it results in inferior performance than the weighted Euclidean distance measure. This happens here because the amount of data in training set is small and estimation of covariance matrix (specially its off-diagonal elements) requires relatively large amount of training data.

So far, we have studied different LSP-based distance measures and found the weighted LSP distance measure to be the best. It might be interesting to see how it compares with the other existing LP distance measures. Some of the LP distance measures which have been popular in the speech recognition literature are as follows: 1) Itakura (or, log-likelihood) distance measure [18], (2) Cepstral distance measure [1,2,18], (3) Weighted cepstral distance measure [2,4], (4) Root-power-sum distance measure [2,3], (5) Reflection coefficient distance measure [1], (6) Weighted reflection coefficient distance measure, (7) Log area ratio distance measure [1], and (8) Weighted log area ratio distance measure. The recognition performance of these distance measures along with that of the weighted LSP distance measure is listed in Table 2. We can see from this table that the weighted LSP distance measure performs consistently better than all the existing LP distance measures for each of the three speakers.

6. Conclusion

In this paper, the LSP representation is studied for speech recognition. Among the different LSP-based distance measures, the weighted LSP distance measure is found to result in the best performance. The weighted LSP representation is then compared as to its recognition performance with the existing LP distance measures. It is found to perform significantly better than these existing LP distance measures.

We have also compared the speech recognition performance of the LSP representation with that of the formant representation. The LSP representation has been found to result in better performance than the formant representation. However, these results are presented elsewhere [19]. It might be noted that the LSP representation not only gives good recognition performance, but it has also got nice interpolation properties which are useful in certain speech recognition systems [20].

References

[1] K.K. Paliwal and P.V.S. Rao, Signal Processing, pp. 323-327, 1982.

[2] K.K. Paliwal, Speech Communication, pp. 151-154, 1982.

[3] B.A. Hansen and H. Wakita, Proc. ICASSP, pp. 757-760, 1986.

[4] Y. Tohkura, Proc. ICASSP, pp. 761-764, 1986.

[5] H. Hermansky et al., Proc. ICASSP, pp. 1971-1974, 1986.

[6] B.H. Juang et al. Proc. ICASSP, pp.765-768, 1986.

[7] F. Itakura and T. Umezaki, Proc. ICASSP, pp. 1257-1260, 1987.

[8] D. Kahn, Proc. ICASSP, 1987.

[9] H. Hermansky, Proc. ICASSP, pp. 1159-1162, 1987.

[10] T. Applebaum et al., Proc. ICASSP, pp. 1155-1158, 1987.

[11] F. Itakura, J. Acoust. Soc. Amer., S 35(A), 1975.

[12] F.K. Soong and B.H. Juang, Proc. ICASSP, pp. 1.10.1-1.10.4, 1984.

[13] G.S. Kang and L.J. Fransen, Proc. ICASSP, pp. 244-247, 1985.

[14] M.A. Mines et al., Language and Speech, pp. 221-235, 1978.

[15] K.K. Paliwal, Speech Communication, pp. 101-106, 1984.
[16] R.O. Duda and P.E. Hart, Pattern Classification and Scene Analysis, Wiley, New York, 1973.
[17] G.T. Toussaint, IEEE Trans. IT-20, pp. 472-479, 1974.
[18] A.H. Gray and J.D. Markel, IEEE Trans. ASSP-24, pp. 380-391, 1976.
[19] K.K. Paliwal, under preparation.
[20] K.K. Paliwal and P.V.S. Rao, J. Acoust. Soc. Amer., pp. 1016-1024, 1982.

Table 1. Vowel recognition performance of different LSP-based distance measures.

Distance measure	Recognition accuracy (in %) for male speaker 1	male speaker 2	female speaker	average
Euclidean distance measure	92.3	94.7	82.3	89.8
Weighted Euclidean distance measure	98.7	99.0	92.3	96.7
Mahalanobis distance measure	97.7	98.7	89.3	95.2

Table 2. Comparision of the weighted LSP distance measure with the other LP distance measures.

Distance measure	Recognition accuracy (in %) for male speaker 1	male speaker 2	female speaker	average
Weighted LSP distance measure	98.7	99.0	92.3	96.7
Itakura distance measure	97.3	97.0	88.3	94.2
Cepstral distance measure	94.0	95.0	85.3	91.4
Weighted cepstral distance measure	95.7	95.0	87.3	92.7
Root-power-sum distance measure	95.0	96.0	86.3	92.4
Reflection coefficient distance measure	84.7	84.7	78.3	82.6
Weighted reflection coefficient distance measure	93.7	92.3	83.7	89.9
Log area ratio distance measure	89.7	86.0	79.7	85.1
Weighted log area ratio distance measure	93.7	92.0	87.3	91.0