Speech Communication 1 (1982) 151–154
North-Holland Publishing Company

ON THE PERFORMANCE OF THE QUEFRENCY-WEIGHTED CEPSTRAL COEFFICIENTS IN VOWEL RECOGNITION

K.K. PALIWAL

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

Received 8 March 1982
Revised 18 May 1982

Abstract. From a number of linear prediction parametric representations each of which furnishes equivalent information about the linear predictor, the cepstral coefficients representation is known to provide the best speech recognition performance. Since the cepstral coefficients of a stable all-pole filter are inversely proportional to their quefrencies, these coefficients are multipled by their respective quefrencies. The quefrency-weighted cepstral coefficients (also known as the root-power sums) are studied as to their effectiveness in a vowel recognition experiment and found to perform better than the cepstral coefficients with a Euclidean distance measure.

Zusammenfassung. Von mehreren Parameterdarstellungen der linearen Prädiktion, die alle jeweils äquivalente Information über den linearen Prädiktor liefern, gibt bekanntlich die Darstellung durch cepstrale Koeffizienten die besten Ergebnisse bei der Spracherkennung. Da die Cepstrum-Koeffizienten eines stabilen Nur-Pole-Filters umgekehrt proportional zu ihren Quefrenzen sind, werden diese Koeffizienten mit ihren jeweiligen Quefrenzen multipliziert. Die quefrenzgewichteten Cepstrum-Koeffizienten (auch als Wurzel-Potenz-Summen bekannt) werden auf ihre Effektivität in einem Vokalerkennungsexperiment untersucht. Es zeigt sich, dass sie mit einem euklidischen Abstandsmass bessere Ergebnisse liefern als die Cepstrum-Koeffizienten.

Résumé. Parmi un certain nombre de représentations paramétriques équivalentes de la prédection linéaire, la représentation par coefficients du cepstre est reconnue comme fournissant les meilleures performances en reconaissance de la parole. Comme les coefficients cepstraux d'un filtre tout pôle sont inversement proportionnels à leurs quéfréquences, ces coefficients sont multipliés par leur quéfréquence respective. Les coefficients cepstraux pondérés en quéfréquence sont étudiés en fonction de leur efficacité dans une expérience de reconnaissance de voyelles. Il est montré que leurs performances sont supérieures à celles des coefficients cepstraux non pondérés.

Keywords. Cepstral coefficients, quefrency, speech recognition, linear prediction analysis.

1. Introduction

In an earlier paper [1], we experimentally compared a number of linear prediction parametric representations as to their vowel recognition performance. Though each of these parametric representation provided equivalent information about the linear predictor, the performance of the cepstral coefficients representation was found to be the best in the vowel recognition task. Similar results were obtained by Pfeifer [2] and Atal [3] in a speaker recognition task and by Ichikawa et al. [4] and Stella [5] in a word recognition task.

Since the cepstral coefficient of a stable all-pole filter are inversely proportional to their quefrencies [6], these coefficients are multiplied in the present paper by their respective quefrencies. We shall study the performance of the quefrency-weighted cepstral coefficients (also known as the root-power sums [7]) in a vowel recognition task and compare their performance with that of the cepstral coefficients.

2. Data acquisition and preprocessing

The speech data consists of 900 utterances containing 30 repetitions of 10 different /b/-vowel-

0167-6393/82/0000-0000/$02.75

/b/ syllables spoken by 3 speakers (2 male and 1 female). Recording of these utterances was carried out in an ordinary office room. The speech signal was 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 6th order Butterworth type lowpass filter with a cutoff frequency of 4 kHz was used to avoid aliasing.

The steady-state part of the vowel segment was manually localized for each of the 900 utterances and a 20 ms segment was excised from its centre. A 10th order linear prediction analysis was performed and then cepstral coefficients were derived from each of these 20 ms segments. The autocorrelation method of linear prediction ensured the stability of the estimated all-pole filter and hence, was used here for analysis (with 20 ms Hamming window and without pre-emphasis [1]).

The linear predictor coefficients $\{a_k\}$, $1 \leq k \leq M$, of the Mth order inverse filter

$$A(z) = 1 + \sum_{k=1}^{M} a_k z^{-k}$$

were computed from the Hamming-windowed speech signal $\{x_n\}$, $0 \leq n \leq N-1$, (where N is the duration of the speech segment) by solving the following set of equations [8]:

$$\sum_{k=1}^{M} a_k R(|i-k|) = -R(i), \quad 1 \leq i \leq M,$$

where

$$R(i) = \sum_{n=0}^{N-1-i} x_n x_{n+i}.$$

The cepstral coefficients $\{c_k\}$, $1 \leq k \leq M$, were computed recursively from the linear predictor coefficients by using the following relations [3]:

$$c_k = -a_k - \frac{1}{k} \sum_{n=1}^{k-1} (k-n) c_{k-n} a_n, \quad 1 \leq k \leq M.$$

Schroeder [7] has recently given direct (nonrecursive) relations for computing the cepstral coefficients from the linear predictor coefficients. The quefrency-weighted cepstral coefficients $\{c'_k\}$, $1 \leq k \leq M$, are obtained from the cepstral coefficients by using the following relations:

$$c'_k = k c_k, \quad 1 \leq k \leq M.$$

3. Recognition procedure

The aim here is to classify the M-dimensional vectors (each vector has M (= 10) cepstral coefficients 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 [9]. In the present paper, the classification scheme used is the forced decision pattern matching method and is studied using the Euclidean distance measure which, for the ith class, is given by

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

where $\boldsymbol{x}$ is the test vector and $\boldsymbol{m}_i$ the mean vector of the ith class. The superscript t denotes here the transpose of the vector. The test vector $\boldsymbol{x}$ 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 are computed from the data in the training set by using the following relations:

$$\boldsymbol{m}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \boldsymbol{y}_{ij}, \qquad 1 \leq i \leq 10,$$

where N_i is the number of preclassified vectors in the ith class and $\boldsymbol{y}_{ij}$ the jth vector of the ith class.

4. Results

Vowel recognition performance of both the cepstral coefficients and the quefrency-weighted cepstral coefficients is studied here separately for each of the three speakers. In the present experiment, a speaker specific training was used; i.e., both the training and test data were derived from the same speaker.

[1] The linear prediction analysis was carried out here without pre-emphasis because it has been found that, for the data used in the present experiment, the pre-emphasis of the speech signal deteriorates the vowel recognition performance.

In order to estimate the vowel recognition performance we had a fixed sample of 300 preclassified vectors for each speaker (obtained from 30 repetitions for each of the 10 vowel classes). This fixed sample can be used, as suggested by Toussaint [10], in a number of ways to estimate the recognition performance. The procedure adopted in the present paper for estimating recognition performance employs different training and test data sets and is described below. For each vowel class, twenty-nine repetitions were used for training the classifier and the thirtieth repetition for testing. This procedure of training and testing was repeated 30 times and a total of 300 decisions were made by the classifier for each speaker.

In Table 1, we show the vowel recognition performance (measured in terms of recognition accuracy) of the cepstral coefficients and the quefrency-weighted cepstral coefficients using a Euclidean distance measure for each of the three speakers. It can be seen from this table that the quefrency-weighted cepstral coefficients give consistently better recognition acccuracy than the cepstral coefficients for each of the three speakers. This indicates the importance of the quefrency-weighted cepstral coefficients in a vowel recognition task.

It may be noted that the Euclidean distance measure employed with quefrency-weighted cepstral coefficients is equivalent to the weighted Euclidean distance measure used with the cepstral coefficient where the weigths associated with different coefficients are fixed a priori and are independent of the data in the training set. Now we will study the vowel recognition performance of cepstral coefficients with the weighted Euclidean distance measure where the weights are computed by statistical methods from the data contained in the training set. The weighted Euclidean distance measure for the ith class is defined by [11];

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

where x_j and m_{ij} are the jth components of the vectors $\boldsymbol{x}$ and $\boldsymbol{m}_i$, respectively, and w_{ij} is the weight associated with the jth component (or the cepstral coefficient) for the ith class and is computed statistically from the data in the training set by taking it to be inversely proportional to standard deviation; i.e.,

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

where y_{ijk} is the kth component of the vector $\boldsymbol{y}_{ij}$. The proportionality constants K_i are determined from the constraints

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

for all the classes (i.e., $i = 1, 2, \ldots, 10$).

The vowel recognition scores obtained by means of this statistically weighted Euclidean distance measure are 95.7%, 95.0% and 87.3% for the first, second and third speakers, respectively[2]. By comparing these scores with those listed in Table 1, we see that the weighted Euclidean distance measure with statistical weight performs slightly better than that with prefixed quefrency weights. But, the former has the disadvantage that additional computation is required to estimate the weights from the data in the training set. Also, these training set

Table 1
Vowel recognition performance for three speakers using the cepstral coefficients and the quefrency-weighted cepstral coefficients

	Recogntion performance (in %) using	
	cepstral coefficients	quefrency weighted cepstral coefficients
First male speaker	94.0	95.0
Second male speaker	95.0	96.0
Female speaker	85.3	86.3

[2] The Mahalanobis distance measure with its class-conditional covariance matrices computed from the data in the training set yielded the recognition scores of 97.3%, 97.7% and 86.7% for the first, second and third speakers, respectively.

data have to be quite large in order to obtain reliable esimates of the weights. On the other hand, the weighted Euclidean distance measure with quefrency weights uses the property of cepstral coefficients and has its weights fixed a priori. These weights are not to be computed from the data in the training set and are thus data-independent. The data in the training set in this case need not be large.

5. Conclusion

The quefrency-weighted cepstral coefficients (also known as the root-power sums) of an all-pole filter were studied as to their effectiveness in a vowel recognition experiment. These coefficients were found to give better recognition performance than cepstral coefficients with a Euclidean distance measure.

References

[1] K.K. Paliwal and P.V.S. Rao, "Evaluation of various linear prediction parametric representations in vowel recognition", *Signal Processing*, Vol. 4, No. 4 July 1982, pp. 323-327.

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

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

[4] 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.

[5] M. Stella "Comparaison de différents coefficients de prédiction linéaire pour la reconnaissance des mots isolés", *Proc. Speech Symposium*, Budapest, Sept, 30-Oct. 2, 1980, pp. 129-134.

[6] A.V. Oppenheim and R.W. Schafer, "Homomorphic analysis of speech", *IEEE Trans. Audio Electroacoust.*, Vol. AU-16, No. 2, June 1968, pp. 221-226.

[7] M.R. Schroeder, "Direct (nonrecursive) relations between cepstrum and predictor coefficients", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-29, No. 2, Apr. 1981, pp. 297-301.

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

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

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

[11] G.S. Sebestyen, *Decision-Making Processes in Pattern Recognition*, Macmillan, New York, 1962, pp. 17-23.