Speech Communication 3 (1984) 101-106
North-Holland

EFFECT OF PREEMPHASIS ON VOWEL RECOGNITION PERFORMANCE

K.K. PALIWAL *

Division of Telecommunications, University of Trondheim, Trondheim-NTH, Norway

Abstract. Preemphasis of the speech signal at higher frequencies is a preprocessing step employed in various speech processing applications. In the present paper, the effect of preemphasis on vowel recognition performance is studied. Preemphasis of the speech signal is achieved by the first-order differencing of the speech signal. Cepstral coefficients derived through linear prediction analysis are used as recognition parameters. A minimum distance classifier is used for vowel recognition and the recognition performance is studied for four different distance measures: Euclidean distance measure, correlation distance measure, Mahalanobis distance measure and Itakura distance measure. It is shown that preemphasis of the speech signal brings about a deterioration in the vowel recognition performance. Implications of this result for isolated word recognition are also discussed.

Zusammenfassung. Bei mannigfachen Sprachverarbeitungsanwendungen wird von Höhenanhebung des Sprachsignals Gebrauch gemacht. Im vorliegenden Aufsatz wird der Einfluss der Höhenanhebung auf die Vokalerkennung untersucht. Die Höhenanhebung wird durch ein Differenzierglied erster Ordnung realisiert. Als Erkennungsparameter dienen Cepstralkoeffizienten, welche durch LPC-Analyse gewonnen wurden. Für die Vokalerkennung wird ein Minimaldistanzklassifikator verwendet, und die Erkennungsleistung wird für vier verschiedene Distanzmasse untersucht: Euklidisches, Korrelations-, Mahalanobis- und Itakura- Distanzmass. Es wird nachgewiesen, dass die Höhenanhebung des Sprachsignals die Vokalerkennungsleistung herabsetzt. Die Auswirkungen dieses Ergebnisses auf die Erkennung isolierter Worte werden ebenfalls diskutiert.

Résumé. Diverses applications du traitement de la parole utilisent la préaccentuation du signal aux fréquences élevées. Dans cet article, nous étudions l'effet de la préaccentuation sur la performance de reconnaissance de voyelles. La préaccentuation est réalisée par différentiation du premier ordre. Les coefficients cepstraux, obtenus à partir d'une analyse par prédiction linéaire, sont utilisés comme paramètres de reconnaissance. Il est fait usage d'un classificateur par minimum de distance, et la performance de reconnaissance est étudiée pour quatre mesures de distance différentes: la distance euclidienne, la distance par corrélation, la distance de Mahalanobis et la distance d'Itakura. Nous montrons que la préaccentuation conduit à une détérioration du taux de reconnaissance des voyelles. Des implications de ce résultat pour la reconnaissance de mots isolés sont également discutées.

Keywords. Speech recognition, linear prediction, cepstral coefficients, preemphasis

1. Introduction

Preemphasis of the speech signal at higher frequencies has become a standard preprocessing step in many speech processing applications such as linear prediction (LP) analysis-synthesis [1,2] and speech recognition [3-7]. For LP analysis-synthesis systems, preemphasis serves a useful purpose because, at the analysis stage, it reduces the dynamic range of the speech spectrum and this helps in estimating the LP parameters more accurately [8] while, at the synthesis stage, speech synthesised from the LP parameters representing the preemphasised speech is deemphasised. But, it is not clear how preemphasis helps in speech recognition systems. On the contrary, there is some evidence that preemphasis of the speech signal at higher frequencies may deteriorate the speech recognition performance. For example, Davis and Mermelstein [9] showed that the mel-frequency cepstra which are computed by giving less weight to the

* On leave from Speech and Digital Systems Group, Tata Institute of Fundamental Research, Bombay-400005, India.

higher frequency region of the spectrum give a better recognition performance than the linear frequency cepstra. Also, Paliwal et al. [10] showed that the zero crossing rate of the raw speech signal is a better parameter for speech recognition than the zero crossing rate of the preemphasised speech signal. So, the aim of the present study is to investigate the effect of preemphasis on speech recognition performance. Differencing of the speech signal is used here for achieving preemphasis at higher frequencies.

Since most of the phonemes (about 38.7% [11]) occurring in conversational English are vowels, we will investigate here the effect of preemphasis on the performance of a vowel recognition system. Hopefully, the results obtained from this vowel recognition experiment will be useful for a general acoustic-phenomic recognition of continuous speech. Since LP analysis is commonly employed in speech recognition systems [3–7] and the cepstral coefficient representation is known to be the best LP parametric representation for speech recognition [12–15], in the present study we will use the cepstral coefficients derived through LP analysis as recognition parameters. We will study here the effect of preemphasis for four different distance measures: Euclidean, correlation, Mahalanobis and Itakura. [1]

2. Data acquisition and preprocessing

For studying the effect of preemphasis on vowel recognition performance, the speech data base was prepared as follows. We constructed thirty lists of all the ten Hindi vowels (/i/, /I/, /e/, /æ/, /ʌ/, /a/, /ɔ/, /o/, /U/ and /u/). Each list had a different random ordering of the ten vowels. The vowels from these lists were spoken in a /b/-vowel-/b/ context by three speakers (two male and one female). The recording of these 900 uttrances was done in an ordinary office room. The speech signal was digitised at a sampling rate of 10 kHz by using a 12-bit analog-to-digital converter. To avoid aliasing, a 6-th order Butterworth-type lowpass filter with a cutoff frequency of 4 kHz was used prior to digitisation. The steady-state part of the vowel segment was manually located for each of the 900 utterances and a 20 ms segment was excised from its centre. These 900 steady-state vowel segments (10 vowels × 30 repetitions × 3 speakers) formed the data base used in the present study. A 10-th order LP analysis was performed for each of the 20 ms segments. The speech signal was weighted by a 20 ms Hamming window and the autocorrelation method of linear prediction was used for analysis. The ten linear predictor coefficients $\{a_k\}$, $1 \leqslant k \leqslant 10$, were computed from the Hamming-windowed speech signal $\{x_n\}$, $0 \leqslant n \leqslant N-1$, (where N (= 200 samples) is the duration of speech segment) by solving the following set of linear equations [17]:

$$\sum_{k=1}^{10} a_k R(|i-k|) = -R(i), \quad 1 \leqslant i \leqslant 10, \tag{1}$$

where

$$R(i) = \frac{1}{N} \sum_{n=0}^{N-1-i} x_n x_{n+i}. \tag{2}$$

The cepstral coefficients $\{c_k\}$, $1 \leqslant k \leqslant 10$, were computed recursively from the linear predictor coefficients by using the following relations [18]:

$$c_k = a_k - \frac{1}{k} \sum_{n=1}^{k-1} (k-n) c_{k-n} a_n, \quad 1 \leqslant k \leqslant 10. \tag{3}$$

All the 900 segments were then preemphasised by taking a first-order difference of successive speech samples and the cepstral coefficients were computed again for each of the preemphasised 20 ms segments.

3. Recognition procedure

Each of the 900 segments represented by ten cepstral coefficients can be considered as a vector c in a 10-dimensional space, where $c = (c_1, c_2, \ldots, c_{10})^t$ and the superscript t denotes the transpose of the vector. The aim here is to classify each of the 10-dimensional vectors into ten vowel classes: /i/, /I/, /e/, /æ/, /ʌ/, /a/, /ɔ/, /o/, /U/ and /u/. This is a standard problem in

[1] Preliminary results using *only* Euclidean distance measure have already been reported at a conference [16].

statistical pattern recognition and has bee treated exhaustively in the literature [19]. In the present study, a minimum distance classifier is used for vector classification. The input vector $\boldsymbol{c}$ is classified here into the i-th class if $d_i < d_j$ for all $j \neq i$, where d_i is the distance of the vector $\boldsymbol{c}$ from the i-th class in the 10-dimensional space.

The performance of the vowel recognition system is studied for four different distance measures: Euclidean distance measure, correlation distance measure, Mahalanobis distance measure and Itakura distance measure. These distance measures are briefly described below.

1) *Euclidean distance measure*

Euclidean distance d_i for the i-th class is given by

$$d_i^2 = (\boldsymbol{c} - \bar{\boldsymbol{c}}_i)^{\mathrm{t}} (\boldsymbol{c} - \bar{\boldsymbol{c}}_i), \tag{4}$$

where $\bar{\boldsymbol{c}}_i\ (= (\bar{c}_{i1}, \bar{c}_{i2}, \ldots, \bar{c}_{i10})^{\mathrm{t}})$ is the mean vector of cepstral coefficients for the i-th class.

2) *Correlation distance measure*

Correlation distance d_i for the i-th class is defined here by [20]

$$d_i^2 = 1 - \frac{\boldsymbol{c}^{\mathrm{t}} \bar{\boldsymbol{c}}_i}{(\boldsymbol{c}^{\mathrm{t}}\boldsymbol{c})^{1/2} (\bar{\boldsymbol{c}}_i^{\mathrm{t}} \bar{\boldsymbol{c}}_i)^{1/2}}. \tag{5}$$

3) *Mahalanobis distance measure*

Mahalanobis distance d_i for the i-th class is given by [19]

$$d_i^2 = (\boldsymbol{c} - \bar{\boldsymbol{c}}_i)^{\mathrm{t}} \boldsymbol{W}_i^{-1} (\boldsymbol{c} - \bar{\boldsymbol{c}}_i), \tag{6}$$

where $\boldsymbol{W}_i$ is the 10×10 covariance matrix of cepstral coefficients for the i-th class.

4) *Itakura distance measure*

Itakura distance d_i for the i-th class is given by [3]

$$d_i = \frac{\bar{\boldsymbol{a}}_i^{\mathrm{t}} \boldsymbol{R} \bar{\boldsymbol{a}}_i}{\boldsymbol{a}^{\mathrm{t}} \boldsymbol{R} \boldsymbol{a}}, \tag{7}$$

where

$$\boldsymbol{a} = (1, a_1, a_2, \ldots, a_{10})^{\mathrm{t}},$$

$$\bar{\boldsymbol{a}}_i = (1, \bar{a}_{i1}, \bar{a}_{i2}, \ldots, \bar{a}_{i10})^{\mathrm{t}},$$

and $\boldsymbol{R}$ is a 11×11 matrix of autocorrelation coefficients with its element $R_{ij} = R(|i-j|)$. The elements of the vector $\bar{\boldsymbol{a}}_i$ are computed recursively from the average cepstral coefficients $\{\bar{c}_{in}\}$, $1 \leqslant n \leqslant 10$, for the i-th class by using Eq. (3). The elements of the vector $\boldsymbol{a}$ and the matrix $\boldsymbol{R}$ are computed from the speech segment to be recognised by using Eqs. (1) and (2).

For training the classifier, the mean vectors and class-conditional covariance matrices of cepstral coefficients are required for all the 10 vowel classes. These are computed from the data in the training set by using the following relations:

$$\bar{\boldsymbol{c}}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \boldsymbol{c}_{ij}, \quad 1 \leqslant i \leqslant 10, \tag{8}$$

and

$$\boldsymbol{W}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} (\boldsymbol{c}_{ij} - \bar{\boldsymbol{c}}_i)^{\mathrm{t}} (\boldsymbol{c}_{ij} - \bar{\boldsymbol{c}}_i), \quad 1 \leqslant i \leqslant 10, \tag{9}$$

where N_i is the number of preclassified vectors for the i-th class in the training-set data and $\boldsymbol{c}_{ij}$ the j-th preclassified vector of the i-th class in the training-set data.

4. Results

In order to study the effect of preemphasis at higher frequencies, the performance of the vowel recognition system is estimated separately for both raw speech and preemphasised speech, for each of the three speakers, and for each of the four distance measures described in Section 3. The vowel recognition system is trained by speakers' own utterances; i.e., speaker specific training is used.

For evaluating the performance of the vowel recognition system, we have a fixed sample of 300 preclassified vectors (obtained from 30 repetitions for each of the ten vowel classes). We are interested in getting an unbiased estimate of vowel recognition performance. For this, it is necessary to divide the fixed sample of preclassified vectors into two independent sets: a training set and a test set. This division can be done in a number of ways, as suggested by Toussaint [21]. We employ here a procedure which is described below. For each vowel class, one repetition is used for testing the classifier and the remaining twenty-nine repetitions for training the classifier. This step is repeated 30 times; each time, a new repetition is taken for testing and the remaining twenty-nine

Table 1
Vowel recognition performance *without* preemphasis

Distance measure	Vowel recognition rate (in %)			
	for first speaker	for second speaker	for third speaker	average
Euclidean	94.0	95.0	85.3	91.4
Correlation	94.0	93.7	84.7	90.8
Mahalanobis	97.3	97.7	86.7	93.9
Itakura	97.0	96.3	88.0	94.0

Table 2
Vowel recognition performance *with* preemphasis

Distance measure	Vowel recognition rate (in %)			
	for first speaker	for second speaker	for third speaker	average
Euclidean	93.3	94.0	84.3	90.6
Correlation	93.0	93.0	84.3	90.1
Mahalanobis	96.3	96.3	84.7	92.4
Itakura	95.0	94.0	87.3	92.1

repetitions for training. In this manner, a total of 300 decisions are made by the classifier for each speaker.

The results of the vowel recognition experiment are given in Table 1 for the case where the speech signal is not preemphasised at higher frequencies. Here, the vowel recognition rates are listed for all the four distance measures and for each of the three speakers. Table 2 lists the performance of the vowel recognition system for the case where the speech is preemphasised. The effect of preemphasis can be seen by comparing the vowel recognition rates from these two tables. For example, for the Euclidean distance measure it can be seen that when the speech signal is not preemphasised, the vowel recognition system gives a recognition accuracy of 94% of the first speaker, 95% for the second speaker and 85.3% for the third speaker. For preemphasised speech, this recognition accuracy is 93.3% for the first speaker, 94% for the second speaker and 84.3% for the third speaker. Thus, we see that, for Euclidean distance measure if speech is preemphasised at higher frequencies the recognition accuracy of the system goes down consistently for each of the three speakers. Similar observations can be made from Tables 1 and 2 for the correlation, and the Mahalanobis and Itakura distance measures. Thus, with preemphasis, the vowel recognition performance consistently deteriorates for each of the three speakers and for each of the four distance measures.

It is not difficult to find the reason for this deterioration in vowel recognition performance due to preemphasis at higher frequencies. It is well known that, for vowel sounds, most of the acoustic cues lie in the lower frequency region. Preemphasis puts undue weight on high frequency components and thus causes a deterioration in vowel recognition performance. If this reasoning is extended to consonant sounds, we expect the consonant recognition performance to improve with preemphasis because most of the acoustic cues are known to lie in the higher frequency region for most of the consonants, particularly fricatives. An experiment to verify this conjecture for consonant sounds will be conducted at a further date. However, Dautrich et al. [22] of Bell Laboratories have recently reported an experiment which was carried out on isolated word recognition and which indirectly supports this conjecture. In their experiment, they used the vocabulary of English alpha-digits which, for successful word recognition, requires more discrimination between consonants. (For example, consonants /s/ and /z/ have to be discriminated for the recognition of alphabets 'C' and 'G'.) They showed that, for this vocabulary, the isolated word recognition performance improved with preemphasis. In the isolated word recognition experiment reported by Paliwal et al. [10], a vocabulary of Hindi digits was used. Since this vocabulary required more discrimination between vowels, the isolated word recognition performance in their experiment deteriorated with preemphasis.

Thus, in an isolated word recognition system the use of preemphasis as a preprocessing step depends on the vocabulary of the system. If the vocabulary requires more discrimination between consonants, preemphasis should be used. On the contrary, if the vocabulary requires more discrimination between vowels, preemphasis should not be used. However if the vocabulary is such that both types of discriminations (i.e., between consonants and between vowels) are important, a trade-off in the amount of preemphasis at higher frequencies is necessary. This can be done by choosing an appropriate value for the parameter, a, in the transfer function $H(z) = 1 - az^{-1}$ representing the preemphasis operation.

It should be noted here that, in the present study, we have investigated the effect of preemphasis on the recognition performance of vowel sounds only. The effect of preemphasis on consonant recognition performance will be studied at a future date.

5. Conclusion

The effect of preemphasising the speech signal at higher frequencies on vowel recognition performance has been studied. Ten cepstral coefficients derived through LP analysis have been used as recognition parameters. A minimum distance classifier with four different distance measures (Euclidean, correlation, Mahalanobis and Itakura) have been used for vowel classification. It has been found that preemphasis of the speech signal at higher frequencies causes a deterioration in vowel recognition performance.

References

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

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

[3] 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.

[4] L.R. Rabiner, "On creating reference templates for speaker-independent recognition of isolated words", *IEEE Trans. Acoust. Speech signal Process.*, Vol. ASSP-26, No. 1, Feb. 1978, pp. 34–42.

[5] L.R. Rabiner, S.E. Levinson, A.E. Rosenberg and J.G. Wilpon, "Speaker-independent recognition of isolated words using clustering techniques", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-27, No. 4, Aug. 1979, pp. 336–349.

[6] L.R. Rabiner and J.G. Wilpon, "Speaker-independent, isolated word recognition for a moderate size (54 word) vocabulary", *IEEE Trans. Acoust. speech Signal Process.*, Vo. ASSP-27, No. 6, Dec. 1979, pp. 583–587.

[7] L.R. Rabiner and C.E. Schmidt, "Application of dynamic time warping to connected digit recognition", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-28, No. 4, Aug. 1980, pp. 377–388.

[8] A.H. Gray, Jr. and J.D. Markel, "A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-22, No. 3, June 1974, pp. 207–217.

[9] S.B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-28, No. 4, Aug. 1980, pp. 357–366.

[10] K.K. Paliwal, S.S. Sinha and A. Agarwal, "An isolated word recognition system for Hindi digits using linear time normalisation", *Journal of the Institution of Electronics and Telecom. Engrs.*, Vol. 29, No. 1, Jan. 1983, pp. 18–22.

[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] 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.

[13] M. Stella, "Comparison de différents coefficients de prédiction linéaire pour la reconnaissance des mots isolés, *Proc. Speech Symposium*, Budapest, 1980, pp. 129–134.

[14] 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.

[15] K.K. Paliwal, "On the performance of the quefrency-weighted cepstral coefficients in vowel recognition", *Speech Communication*, Vol. 1, No. 2, Aug. 1982, pp. 151–154.

[16] K.K. Paliwal, "Effect of preemphasis on speech recognition performance", *Proc. 11th Int. Cong. Acoustics*, Paris, France, July 19–27, 1983, Vol. 4, pp. 211–214.

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

[18] 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.

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

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

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

[22] B.A. Dautrich, L.R. Rabiner and T.B. Martin, "The effects of selected signal processing techniques on the performance of a filter-bank-based isolated word recognizer", *Bell Syst. Tech. Jour.*, Vol. 62, No. 5, May-June 1983, pp. 1311–1336.