ICICS-PCM 2003
15-18 December 2003
Singapore

Improved Linear Predictive Coding Method for Speech Recognition

Jiang Hai; Er Meng Joo

School of Electrical and Electronic Engineering
Nanyang Technological University
Singapore 639798

Abstract

In this paper, improved Linear Predictive Coding (LPC) coefficients of the frame are employed in the feature extraction method. In the proposed speech recognition system, the static LPC coefficients + dynamic LPC coefficients of the frame were employed as a basic feature. The framework of Linear Discriminant Analysis (LDA) is used to derive an efficient and reduced-dimension speech parametric speech vector space for the speech recognition system. Using the continuous Hidden Markov Model (HMM) as the speech recognition model, the speech recognition system was successfully constructed. Experiments are performed on the isolated-word speech recognition task. It is found that the improved LPC feature extraction method is quite efficient.

1. Introduction

Several adverse conditions are also often present during operation, such as ambient and transmission noise, distortions due to room acoustics and transducers, and even changes in speech characteristics due to psychological awareness of talking to a machine. These conditions need to be dealt with in order for the recognizer to be able to deliver reliable results. Thus, the issue of robustness is one of the important problems in automatic speech recognition.

One method that achieves robust results is to extract the most representative feature from the speech utterance in bad environment conditions. Feature extraction of speech is one of the most import issues in the field of speech recognition. In order to achieve high recognition accuracy, the feature extractor is required to discover salient characteristics suited for classifications.

There are two dominant acoustic measurements of speech signals. One, which occurs in the temporal domain, is the parametric modeling approach, which is developed to match closely the resonant structure of the human vocal tract that produces the corresponding sound. It is mainly derived from Linear Predictive analysis, such as the Linear Predictive Coding (LPC) method. The other, which occurs in the frequency domain, is the nonparametric modeling method originated from the human auditory perception system. FFT-based Mel Frequency Cepstral Coefficients (MFCC) are utilized for this purpose. The MFCC feature extraction method was widely adopted in many popular speech recognition systems by many researchers. [1], [2] , [3].

To increase the speech recognition accuracy, some modern speech recognizers were employed such as Discrete Cosine Transform (DCT) [4],[5] and Principal Component Analysis (PCA) method to increase the discriminant of feature vectors and reduce the feature dimension [6] , [7]. Some researchers have designed new optimal feature extraction methods, for example, M.Chetouani uses Neural Predictive Coding method which is an extension of LPC method by modeling nonlinear speech signals [8], Satya Dharanipragada uses Minimum Variance Distortionless Response spectrum based modeling of speech [9] and Ralf Schluter attempts to use Linear Discriminant Analysis method to optimize the Mel Frequency Cepstral Coefficients (MFCC) [10].

In spite of these developments, effective feature extraction is certainly far from being a solved problem. The LPC method is not optimal because the underlying speech production model is nonlinear. LPC starts with the assumption that the speech signal is produced by a buzzer at the end of a tube. For ordinary vowels, the vocal tract is well represented by a single tube. However, for nasal sounds, the nose cavity forms a side branch. In practice, this difference is partly ignored and partly dealt with during the encoding of the residue.

Most of the speech recognition systems use MFCC for phoneme recognition. Essentially, in all these computation methods, Fourier Transform (FT) is used. It is a well-known fact that the windowed FT or the Short Time Fourier Transform (STFT) has uniform resolution over the time-frequency plane. Because of this, it is difficult to detect sudden burst in a slowly varying signal by using STFT. This phenomenon is observed in phoneme recognition when "stops" are encountered.

In this paper, we attempt to use Linear Discriminant Analysis (LDA) to increase the discriminant characteristic of conventional LPC coefficients and increase the classifier design accuracy. Moreover, LDA is employed to decrease the dimension of feature vectors and to combine the two procedures, i.e. feature extraction and classification.

0-7803-8185-8/03/$17.00 © 2003 IEEE

In the proposed system, the static LPC coefficients + dynamic LPC coefficients of the frame vector space were compressed by using the LDA and the reduced dimension feature space can be applied in our speech recognition system.

2. Linear Predictive Coding

Parametric representation of a spectrum using linear prediction is a powerful technique in speech processing. For speech coding applications, LPC was introduced about 30 years ago [11] and it is still the main and effective tool in that field. In speech applications, the main advantage is usually attributed to the all-pole characteristics of vowel spectra. However, because the human ear is more sensitive to spectral poles than zeros [12], LPC also has advantages in terms of human hearing. In comparison with nonparametric spectral modeling techniques, LPC is also more powerful in compressing the spectral information into few filter coefficients which can be more efficiently quantized.

Because speech signals are only stable in a short time, LPC is also a short-term estimation method as other speech signal analysis methods. There are two-ways processing short-term analysis methods. First, each speech frame is multiplied by the window function $w(n)$ so as to obtain the windowed speech frame $s_w(n)$. For each frame, a vector of LPC coefficients is computed from the autocorrelation vector using a Levinson or a Durbin recursion method. Second, since the speech frame is not windowed, we use the covariance method analysis to get the LPC coefficients.

The LPC method considers a speech sample at time n, $s(n)$ and approximates it by a linear combination of the past p speech samples in the following way:

$$s(n) \approx a_1 s(n-1) + a_2 s(n-2) + \ldots + a_p s(n-p)$$

where $a_1, \ldots, a_p$ are constant coefficients. The above equation can be transformed, by including an excitation term $Gu(n)$, to:

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + Gu(n) \qquad (1)$$

where G is the gain and $u(n)$ is the normalized excitation. Transforming equation (1) to z-domain, we obtain:

$$S(z) = \sum_{i=1}^{p} a_i z^{-i} S(z) + GU(z)$$

and consequently, the transfer function $H(z)$ becomes

$$H(z) = \frac{S(z)}{GU(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-1}} = \frac{1}{A(z)}$$

This corresponds to the transfer function of a digital time-varying filter. The main parameters that can be obtained with the LPC model are the classification voiced/unvoiced, the pitch period, the gain and the coefficients a_i. It is important to note that the higher the order of the model is, the better the all-pole model can represent spoken sounds. A linear predictor with coefficients α_k is defined as follows:

$$P(z) = \sum_{k=1}^{p} \alpha_k z^{-k}$$

whose output is

$$\tilde{s}(n) = \sum_{k=1}^{p} \alpha_k s(n-k)$$

The prediction error $e(n)$ is defined as:

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} \alpha_k s(n-k)$$

which is the output of the system $A(z) = 1 - \sum k = 1^p \alpha_k z^{-k}$. If $a_k = \alpha_k$, we have $H(z) = \frac{G}{A(z)}$. The main goal now is to obtain the set of coefficients α_k that minimizes the square of the prediction error in a short segment of speech (typically 10-20ms frame). The mean short-time prediction error per frame is defined as:

$$E_n = \sum_m e_n^2(m) = [s_n(m) - \sum_{k=1}^{p} \alpha_k s_n(m-k)]^2$$

where $s_n(m)$ is a segment of speech selected in the neighborhood of sample n: $s_n(m) = s(m+n)$. The values of the coefficients α_k that minimize the error E_n can be obtained from $\frac{dE_n}{d\alpha_i} = 0, \quad i = 1,2,\ldots,p$. This results in the next equation:

$$\sum_m s_n(m-i)s_n(m-k) = \sum_{k=1}^{p} \alpha_k' \sum_m s_n(m-i)s_n(m-k) \quad 1 \le i \le p \quad (2)$$

where α_k' are the values of α_k that minimize E_n. Defining $\Phi_n(i,k) = \sum_m s_n(m-i)s_n(m-k)$, equation (2) can be written as:

$$\sum_{k=1}^{p} \alpha_k \Phi_n(i,k) = \Phi_n(i,0) \quad i = 1,2,\ldots,p \quad (3)$$

This is a system of p equations with p variables that can be solved to find the α_k coefficients for the segment $s_n(m)$. It can be demonstrated that:

$$E_n = \sum s_n^2(m) - \sum_{k=1}^{p} \alpha_k \sum_m s_n(m)s_n(m-k)$$

and in compact form:

$$E_n = \Phi_n(0,0) - \sum_{k=1}^{p} \alpha_k \Phi_n(0,k) \qquad (4)$$

Now, the values $\Phi_n(i,k)$ have to be obtained for $1 \le i \le p$ and $0 \le k \le p$, and the α_k coefficients are obtained by solving equation (3). The system given by equation (4) can be solved using the autocorrelation method.

The autocorrelation method considers the segments $s_n(m)=0$ outside the interval $0 \le m \le N-1$ and $s_n(m)=s(m+n)w(m)$ in the interval (where $w(m)$ is a finite-length window). If $s_n(m)$ differs from zero for $0 \le m \le N-1$, the correspondent prediction error $e_n(m)$ for a linear predictor of order p will be different from zero in the interval $0 \le m \le N-1+p$. Hence, $E_n = \sum_{m=0}^{N-1+p} e_n^2(m)$, Using this method, the prediction error is large at the beginning and at the end of the interval due to the prediction of null samples in the extremes. For this reason, every segment should be applied a windowing process (for example, a Hamming window) for reducing the border values. Considering that $s_n(m)$ is null outside the interval $0 \le m \le N-1$, it can be demonstrated that:

$$\Phi_n(i,k) = \sum_{m=0}^{N-1+p} s_n(m-i)s_n(m-k) \quad 0 \le k \le p \quad 1 \le i \le p$$

which can be rewritten as:

$$\Phi_n(i,k) = \sum_{m=0}^{N-1-(i-k)} s_n(m)s_n(m+i-k) \quad 1 \le i \le p \quad 0 \le k \le p$$

In this case, $\Phi_n(i,k)$ is related to the short-time autocorrelation function valued for $i-k$:

$$\Phi_n(i,k) = R_n(i,k)$$

where $R_n(k) = \sum_{m=0}^{N-1-k} s_n(m)s_n(m+k)$ is a pair function, so:

$$\Phi_n(i,k) = R_n(|i-k|) \quad i = 1,2,...,p \quad k = 0,1,...,p$$

Therefore,

$$\sum_{k=1}^{p} \alpha_k R_n(|i-k|) = R_n(i) \qquad 1 \le i \le p$$

In an analogous way, the square prediction error is:

$$E_n = R_n(0) - \sum_{k=1}^{p} \alpha_k R_n(k)$$

The equation system can be expressed in the following matrix-vector form:

$$\begin{bmatrix} R_n(0) & R_n(1) & R_n(2) & \dots & R_n(p-1) \\ R_n(1) & R_n(0) & R_n(1) & \dots & R_n(p-2) \\ R_n(2) & R_n(1) & R_n(0) & \dots & R_n(p-3) \\ \dots & \dots & \dots & \dots & \dots \\ R_n(p-1) & R_n(p-2) & R_n(p-3) & \dots & R_n(0) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \dots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} R_n(1) \\ R_n(2) \\ R_n(3) \\ \dots \\ R_n(p) \end{bmatrix}$$

The previous matrix equation can be solved using the more efficient Durbin algorithm.

The Durbin's algorithm is used to solve equation systems where the elements across the diagonal are identical and the matrix of coefficients is symmetric (Teoplitz matrix). The complexity of this method, consists of solving $p^2 + o(p)$ operations and the memory required is only 2p locations. The equations to solve are in the form: $\sum_{k=1}^{p} \alpha_k R_n(|i-k|) = R_n(i)$ with $1 \le i \le p$. The complete Durbin algorithm is:

$$E^{(0)} = R(0)$$

$$k_i = \left[R(i) - \sum_{j=1}^{i-1} \alpha_j^{i-1} R(i-j) \right] \Big/ E^{i-1}$$

$$\alpha_i^i = k_i$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} = k_i \alpha_{i-j}^{(i-1)}$$

$$E^{(i)} = (1-k_i^2)E^{(i-1)}$$

These equations are solved recursively for $i = 1,2,...,p$ and the final solution is given by $\alpha_j = \alpha_j^{(p)}$ where $1 \le j \le p$.

3. Linear Discriminant Analysis

Pattern recognition problems often include two parts, namely feature extraction and classification. In the literature, many classification methods are developed in two different ways. The first type of methods considers the raw pattern matrix without any concern about properties in the initial data set. The objective is to find significant clusters using all the available measurements, i.e. all the parameters defining the objects or events. Then, each cluster is represented by a model, for instance, by the centroid and the within class covariance, and additional objects are classified using these characteristics.

The second type of classification methods uses the pattern class information. A transformation can be applied on this pattern matrix in order to find the optimal discriminant directions. Next, additional objects are classified in this optimal subspace. The principal subject of this method is either for discriminating various categories of observed objects or for finding a relevant subspace of vectors through transformation of a high-dimensional pattern matrix.

The first case corresponds to finding an optimal set of discriminant vectors, while the second permits to reduce the dimension of high-dimensional data without losing significant scatters.

LDA is one of the common tools for multigroup data classification and dimensionality reduction. It is a technique based on the transform and attempts to minimize the ration of the within-class scatter to the between-class scatter thereby guaranteeing the maximal separability [13].

This key idea of this method is to find the best set of discriminant vectors in order to separate predefined classes of objects and reduce the dimension of high-dimensional data. Each object is represented by a set of frequently quite large raw measurements. LDA aims at finding a linear transform from a N-dimensional vector space to n-dimensional vector space. We can perform the dimensionality reduction of the vector space ($N \ge n$). Let X be a N-dimensional vector, and U a $N \times n$ transformation matrix. The n-dimensional transformed vector is then expressed as $U^t X$. The transformation matrix U consists of the n columns of d_n. The commonly

used method is to obtain n discriminant vectors d_n such that the ratio of the between-class variance to the within-class is maximized [14], [15] and [16]. Such a criterion can be expressed as:

$$C = \frac{d_n^t B d_n}{d_n^t W d_n}$$

$$B = \frac{1}{N}\sum_{k=1}^{K} n_k (v_k - v)(v_k - v)^t$$

$$W = \frac{1}{N}\sum_{k=1}^{K}\sum_{n=1}^{n_k} (x_{kn} - v_k)(x_{kn} - v_k)^t$$

$$v_k = \frac{1}{n_k}\sum_{n=1}^{n_k} x_{kn}$$

$$v = \frac{1}{N}\sum_{k=1}^{K} n_k v_k$$

where B is the between-class covariance matrix and W is the within-class covariance matrix with the following constraints:

$$d_1^t d_n = d_2^t d_n = ... = d_{n-1}^t d_n = 0$$

An additional constraint on the norm of d_n can be chosen as

$$d_n^t W d_n = 1$$

It is always possible to restore the final norm to

$$d_n^t d_n = 1$$

after computation of the vector direction in order to obtain an orthonormal set of vectors.

The first solution is the Fisher linear discriminant d_1 which can be obtain from

$$W_{-1} B d_1 = \lambda_1 d_1$$

Let us use the method of Lagrange multipliers to transform the C criterion including all the constraints in order that we can computer the n^{th} discriminant vector:

$$C_L = d_n^t B d_n - \lambda[(d_n^t W d_n) - 1] - \mu_1 d_n^t d_1 - ... - \mu_{n-1} d_n^t d_{n-1}$$

The optimization is performed by setting the partial derivative of C_L with respect to d_n equal to zero:

$$\frac{\partial C_L}{\partial d_n} = 0 = 2Bd_n - 2\lambda W d_n - \mu_1 d_1 ... - u_{n-1} d_{n-1} = 0 \quad (5)$$

Multiplying the left side of (5) by d_n^t, we obtain

$$2d_n^t B d_n - 2\lambda d_n^t W d_n = 0 \Rightarrow \lambda = \frac{d_n^t B d_n}{d_n^t W d_n}$$

Thus, we use λ to represent maximizing the expression.

Multiplying the left side of (5) successively by $d_1^t W^{-1}, ..., d_{n-1}^t W^{-1}$, we get a set of $n-1$ expressions as follows:

$$\mu_1 d_1^t W^{-1} d_1 + \mu_2 d_1^t W^{-1} d_2 + ... + \mu_{n-1} d_1^t W^{-1} d_{n-1} = 2 d_1^t W^{-1} d_n$$

$$\mu_1 d_2^t W^{-1} d_1 + \mu_2 d_2^t W^{-1} d_2 + ... + \mu_{n-1} d_2^t W^{-1} d_{n-1} = 2 d_2^t W^{-1} d_n ...$$

$$\mu_1 d_{n-1}^t W^{-1} d_1 + \mu_2 d_{n-1}^t W^{-1} d_2 + ... + \mu_{n-1} d_{n-1}^t W^{-1} d_{n-1} = 2 d_{n-1}^t W^{-1} d_n$$

Using the following matrix notation:

$$\mu^{(n-1)} = \begin{bmatrix} u_1 \\ ... \\ u_{n-1} \end{bmatrix}, \quad D^{(n-1)} = \begin{bmatrix} d_1^t \\ ... \\ d_{n-1}^t \end{bmatrix}$$

$$S^{(n-1)} = [S_{ij}^{(n-1)}], \quad S_{ij}^{n-1} = d_i^t W^{-1} d_j .$$

the previous (n-1) equations can be described in the following matrix equation:

$$S^{(n-1)} u^{(n-1)} = 2 D^{(n-1)} W^{-1} B d_n$$

or equivalently

$$u^{(n-1)} = 2[S^{(n-1)}]^{-1} D^{(n-1)} W^{-1} B d_n \quad (6)$$

Multiplying the left side of (5) by W^{-1} yields

$$2W^{-1} B d_n - 2\lambda d_n - \mu_1 W^{-1} d_1 ... - u_{n-1} W^{-1} d_{n-1} = 0$$

Using the matrix notation, we obtain

$$2W^{-1} B d_n - 2\lambda d_n - W^{-1}[D^{(n-1)}]^t u^{(n-1)} = 0$$

Using (6), we obtain

$$(I - W^{-1}[D^{(n-1)}]^t [S^{(n-1)}]^{-1} D^{(n-1)}) W^{-1} B d_n = \lambda d_n$$

where d_n is the eigenvector of

$$M = (I - W^{-1}[D^{(n-1)}]^t [S^{(n-1)}]^{-1} D^{(n-1)}) W^{-1} B$$

associated with the largest eigenvalue of M.

Computing the n set of d_n eigenvectors, we get the unique transformation matrix U.

In our system, the training data sets were transformed using the algorithm as above and the unique transformation matrix U was obtained at the same time. The test speech feature data sets would be transformed by multiplying the transformation matrix U.

4. Experimental Results

In the proposed system, a more robust acoustic feature effective for automatic speech recognition, i.e. the classical static + dynamic LPC feature vector space is given by LPC + ΔLPC coefficients where LPC coefficients are the 12th-order LPC applied on windowed speech frame and ΔLPC coefficients computed using a classical regression were added to the LPC. We projected this 24-dimensional vector space into a 12-dimensional vector space obtained from the LDA transformation. A high performance speech recognition system utilizing the simple five states from left to right continuous Hidden Markov Model is implemented.

A speech independent English digits recognition system is built in the experiment. The TI46 corpus of isolated words which was designed and collected at Texas Instruments (TI) is used in the proposed system. The TI46 corpus contains 16 speakers: 8 males labeled and 8 females. There are 15 utterances of each English digit (0~9) from each speaker: 10 designated as training tokens and 5 designed as testing tokens in the proposed system. To compare the recognition result, a speech recognition system based on the conventional LPC method is built at the same time. The test results at different Signal-to-Noise Ratio (SNR) are displayed as follows.

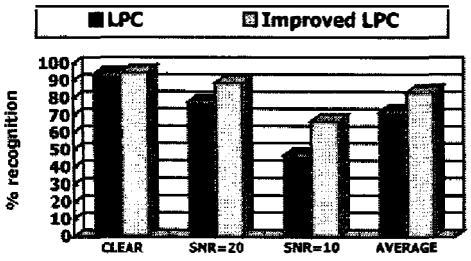

Fig. 1. Speech recognition results

5. Conclusions

In this paper, we have presented a new strategy for feature extraction. Using LPC+△LPC coefficients feature vectors and transforming them into another kind of simple representation with LDA, we are able to absorb the dynamic information and result in static size.

By applying the improved LPC feature attraction method, a high performance speech recognition system utilizing the simple five states from left to right continuous Hidden Markov Model is implemented successfully. The speaker-independent experiments to recognize ten English digits are carried out successfully.

References

[1] M.N.Stuttle and M.J.F.Gales. "A Mixture of Gaussians Front End for Speech Recognition", Eurospeech 2001-Scandinavia.

[2] Potamisfis,I. et al . "Improving the robustness of noisy MFCC features using minimal recurrent neural networks", Neural Networks, 2000.IJCNN2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on, Volume:5, 2000, pp. 271-276.

[3] Wei-Wen Hung and Hsiao-chuan Wang. "On the use of weighted filter bank analysis for the derivation of robust MFCCs". IEEE Signal Processing Letters, Volume:8 Issue:3, March 2001, pp. 70-73

[4] Jong-Hwan Lee, et al. "Speech Feature Extraction Using Independent Component Analysis". 0-7803-6293-4/00/2000,IEEE.

[5] Oh-Wook Kwon, et al. "Application of Variational Bayesian PCA For Speech Feature Extraction". 0-7803-7402-9/02,2002,IEEE.

[6] M.E.Tipping and C.M.Bishop. "Mixtures of probabilistic principal component analysers". Neural computation, 1998.

[7] C.M.Bishop. "Variational principal components". Proc. ICANN, 1999.

[8] M.Chetouani, et al. "Discriminative Training For Neural Predictive Coding Applied to Speech Features Extraction". 0-7803-7278-6/02/2002, IEEE.

[9] Satya Dharanipragada. "Feature Extraction for Robust Speech Recognition". 0-7803-7448-7/02/2002, IEEE.

[10] Ralf Schluter and Hermann Ney. "Using Phase Spectrum Information for Improved Speech Recognition Performance". 0-7803-7041-4/01 2001 IEEE.

[11] B. S. Atal and M. R. Schroeder. "Predictive coding of speech signals", Proceedings of 1967 IEEE Conf, Communication Processing, pp. 360–361.

[12] M. R. Schroeder. "Linear prediction, extremal entropy and prior information in speech signal analysis and synthesis", Speech Commun., vol., no. 1, pp. 9–20.

[13] Suresh Balakrishnama, et al; "Linear Discriminant Analysis For Signal Processing Problems", 0-7803-5237-8/99 IEEE.

[14] Duchene and S. Leclercq. "An Optimal Transformation for Discriminant Principal Component Analysis", IEEE Trans. On Pattern Analysis and Machine Intelligence,Vol. 10, No 6.

[15] J. W. Sammon. "An optimal discriminant plane". IEEE Trans. Comput.,vol.C-19,pp.826-829.

[16] D. H. Foley and J. W. Asmmon. "An optimal set of discriminant vectors". IEEE Trans. Comput., vol. C-24, pp.281-289.