Available online at www.sciencedirect.com

Applied Mathematics and Computation 177 (2006) 644–651

www.elsevier.com/locate/amc

A simple statistical speech recognition of mandarin monosyllables

Tze Fen Li [a], Shui-Ching Chang [b], Chung-Bow Lee [a,*]

[a] *Department of Applied Mathematics, National Chung Hsing University, Kuo-Kuang Road, Taichung 40227, Taiwan, ROC*
[b] *Department of Business Administration, The Overseas Chinese Institute of Technology, Taichung, Taiwan, ROC*

Abstract

Each mandarin syllable is represented by a sequence of vectors of linear predict coding cepstra (LPCC). Since all syllables have a simple phonetic structure, in our speech recognition, we partition the sequence of LPCC vectors of all syllables into equal segments and average the LPCC vectors in each segment. The mean vector of LPCC is used as the feature of a syllable. Our simple feature does not need any time consuming and complicated nonlinear contraction and expansion as adopted by the dynamic time-warping. We propose several probability distributions for the feature values. A simplified Bayes decision rule is used for classification of mandarin syllables. For the speaker-independent mandarin digits, the recognition rate is 98.6% if a normal distribution is used for feature values and the rate is 98.1% if an exponential distribution is used for the absolute values of the features. The feature proposed in this paper to represent a syllable is the simplest one, much easier to be extracted than any other known features. The computation for feature extraction and classification is much faster and more accurate than using the HMM method or any other known techniques.

Keywords: Bayes decision rule; Linear predict coding; Speech recognition

1. Introduction

The rapid spread of computer usage by the human being has stimulated the need for automatic speech recognition (ASR). The ultimate goal of research on ASR is the construction of machines that are indistinguishable from humans in their ability to communicate in natural spoken language. The speech recognition has been investigated about 40 years. The early research in ASR can be found in [1–4]. In the recent years, a lot of speech recognition devices with limited capabilities are now available commercially. These devices are usually able to deal only with a small number of acoustically distinct words. The ability to converse freely with a machine still represents the most challenging topic in speech recognition research. A speech recognition system basically contains extraction of features and classification of an utterance. The measurements made on

[*] Corresponding author.
E-mail address: cblee@amath.nchu.edu.tw (C.-B. Lee).

0096-3003/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.amc.2005.09.094

the speech waveform include energy, zero crossings, extrema count, formants, LPC cepstrum [5–10] and the Mel frequency cepstrum coefficient (MFCC) [11]. The LPC method provides a robust, reliable and accurate method for estimating the parameters that characterize the linear, time-varying, system which is recently used to approximate the nonlinear, time-varying system of the speech waveform. The MFCC method uses the bank of filters scaled according to the Mel scale to smooth the spectrum, performing a processing that is similar to that executed by the human ear. For recognition, the performance of the MFCC was better than the LPC cepstrum [11]. As to classification of an input utterance, the most successful speech recognition methods are the pattern matching using dynamic time-warping (DTW) [12–22], vector quantization (VQ) [23–30] and hidden Markov model (HMM) [31–49]. Since the same word uttered by the same speaker may have different duration of the same phoneme, the DTW process nonlinearly expands or contracts the time axis to match the same phoneme or landmark positions between the input speech and reference templates. This process can usually be accomplished by using the dynamic programming (DP) technique as the method for comparing patterns. This approach has been proved successful, but the computational cost is extremely high. The VQ is an information theoretic data compression principle introduced by Shannon [23]. When it is applied to speech compression, a training sequence is used to generate a set of reproduction vectors (codeword), called the codebook of the speech. In general, the selection of a perceptually meaningful distortion measure in clustering and the construction of an optimal codebook are difficult. It is also difficult to apply the VQ to a large vocabulary because the computational cost is still high in clustering. The theory of HMM was published by Baum et al. [31], but widespread understanding and applications of the theory of HMMs to speech processing have occurred only within the past 10 years. The HMM technique has significantly reduced the computational cost and has been used for large vocabulary connected and continuous speech recognition applications. The mandarin syllable recognition was recently studied by Wagner et al. [41], Chen et al. [42] and Lee et al. [47–49] and the digit recognition can be found in [1,3,10,40,50].

It seems to us that all existing speech recognition methods are computationally complex and time consuming. In this paper, we present the simplest, fastest and the most accurate speech recognition method for mandarin syllables. For feature extraction of a syllable, we just partition the whole sequence of LPCC into equal segments and use the average of LPCC in each segment to represent the syllable, which does not need any time consuming and complicated nonlinear compression and expansion as processed in the DTW and for speech recognition, we simply use a simplified Bayes decision rule where each step is a simple calculation and which has the minimum probability of misclassification. The recognition results are excellent and the computational cost is very low.

2. Bayes decision rules

Let $X = (X_1, \ldots, X_k)$ be the input feature vector of a speech data, which belongs to one of m categories (syllables) c_i, $i = 1, \ldots, m$. Consider the decision problem consisting of determining whether X belongs to c_i. Let $f(x|c_i)$ be the conditional density function of X given category c_i. Let θ_i be the prior probability of c_i such that $\sum_{i=1}^{m}\theta_i = 1$, i.e., the θ_i is the probability for the category c_i to occur. Let d be a decision rule. A simple loss function $L(c_i, d(x))$, $i = 1, \ldots, m$, is used such that the loss $L(c_i, d(x)) = 1$ when $d(x) \neq c_i$ makes a wrong decision and the loss $L(c_i, d(x)) = 0$ when $d(x) = c_i$ makes a right decision. Let $R(\tau, d)$ denote the risk function (the probability of misclassification) of d. Let Γ_i, $i = 1, \ldots, m$, be m regions separated by d in the k-dimensional domain of X, i.e., d decides c_i when $X \in \Gamma_i$. Let $\tau = (\theta_1, \ldots, \theta_m)$. Then

$$R(\tau, d) = \sum_{i=1}^{m} \theta_i \int L(c_i, d(x)) f(x|c_i)\,\mathrm{d}x = \sum_{i=1}^{m} \theta_i \int_{\Gamma_i^c} f(x|c_i)\,\mathrm{d}x, \tag{2.1}$$

where Γ_i^c is the complement of Γ_i. Let D be the family of all decision rules which separate m categories. Let the minimum probability of misclassification be denoted by

$$R(\tau) = \inf_{d \in D} R(\tau, d). \tag{2.2}$$

A decision rule d_τ which satisfies (2.2) is called the Bayes decision rule with respect to the prior distribution τ and is given in (2.3) [51]. An easy proof is given in Theorem 2.1.

Theorem 2.1 [51]. *The Bayes decision rule with respect to τ is defined by*

$$d_\tau(x) = c_i \quad \text{if} \ \ \theta_i f(x \mid c_i) > \theta_j f(x \mid c_j) \tag{2.3}$$

for all $j \neq i$, i.e., $\Gamma_i = \{x|\theta_i f(x|c_i) > \theta_j f(x|c_j)\}$ for all $j \neq i$.

Proof. The probability of misclassification can be written as

$$\begin{aligned} R(\tau, d) &= \sum_{j=1}^{m} \theta_j \int_{\Gamma_j^c} f(x \mid c_j)\,\mathrm{d}x \\ &= \theta_i \int_{\Gamma_i^c} f(x \mid c_i)\,\mathrm{d}x + \sum_{j\neq i} \theta_j \int_{\Gamma_j^c} f(x \mid c_j)\,\mathrm{d}x \quad \left(\Gamma_i^c = \sum_{j\neq i} \Gamma_j\right) \\ &= \theta_i \sum_{j\neq i} \int_{\Gamma_j} f(x \mid c_i)\,\mathrm{d}x + \sum_{j\neq i} \theta_j \left[\int f(x \mid c_j)\,\mathrm{d}x - \int_{\Gamma_j} f(x \mid c_i)\,\mathrm{d}x\right] \\ &= \sum_{j\neq i} \theta_j + \sum_{j\neq i} \int_{\Gamma_j} \left[\theta_i f(x \mid c_i) - \theta_j f(x \mid c_j)\right] \mathrm{d}x \end{aligned}$$

which is minimum since $\Gamma_j \subset \{x|\theta_i f(x|c_i) < \theta_j f(x|c_j)\}$ for $i \neq j$ by the definition of Γ_j. □

Note that if $\theta_i = 1/m$, $i = 1, \ldots, m$, the Bayes decision rule (2.3) becomes a ML classifier. In speech recognition, a possible probability distribution for the feature X_l, $l = 1, \ldots, k$, is the normal or gamma distribution. Hence, in our experiments, the conditional density $f(x|c_i)$ is assumed to have normal distributions with weighted variances or gamma distribution (on the absolute values of X_l) and in order to reduce computational time, the components of the feature vector $X = (X_1, \ldots, X_k)$ are assumed to be stochastically independent.

3. Feature extraction

3.1. Linear predict coding cepstrum (LPCC)

The acoustic wave produced in human speech is represented by a continuously varying (or analog) time waveform and then is digitized by a A/D converter into a sequence of sampled speech signal. The MFCC was proved to be better than the LPC cepstrum for recognition by using the DTW method [11], but the computational complexity for the MFCC is much heavier than that of the LPC cepstrum. First of all, for the MFCC, one has to obtain the DFT of all frames of the signal and after the Mel filter banks smooth the spectrum, performs the inverse DFT on the logarithm of the magnitude of filter bank output. Our goal in this study is to find a simple, fast and accurate classification method to identify all mandarin syllables. Therefore, in this study we use LPC cepstrum in stead of MFCC for recognition. The LPC coefficients can be easily obtained by Durbin's recursive procedure [52] and their cepstra can be quickly found by another recursive equations [52] without computing the DFT and the inverse DFT, which are computationally complex and time consuming.

Since the LPC method can provide a robust, reliable and accurate method for estimating the parameters that characterize the linear, time-varying, system, we transform the speech data into the LPC coefficients. The following is a brief discussion of LPC method. It is assumed [52] that the sampled speech waveform $s(n)$ can be linearly predicted from the past p samples of $s(n)$. Let

$$\hat{s}(n) = \sum_{k=1}^{p} a_k s(n-k), \tag{3.1}$$

and let E be the squared difference between $s(n)$ and $\hat{s}(n)$ over N samples of $s(n)$, i.e.,

$$E = \sum_{n=0}^{N-1} [s(n) - \hat{s}(n)]^2. \tag{3.2}$$

The unknown a_k, $k = 1, \ldots, p$, are called the LPC coefficients and can be solved by the least square method. The most efficient method known for obtaining the LPC coefficients is Durbin's recursive procedure [7,52]. Here in our experiments, $p = 16$ and $N = 200$ (the frame size of the Hamming window). In order to apply distance measures in classification, LPC coefficients are transformed into LPC cepstra (LPCC), which for each mandarin syllable are represented in a sequence of vectors, each having 16 LPCC. In extracting features, we only use the first 10 elements of a vector of LPCC because the cepstra in the last few elements are almost zeros. Here we let $p = 10$.

3.2. Feature extraction from LPCC

Our method to extract the feature from LPCC is quite simple. Let $x(k) = (x(k)_1, \ldots, x(k)_p)$, $k = 1, \ldots, n$, be the LPCC vector for the kth frame of a speech waveform in the sequence. Since all mandarin syllables have a simple phonetic structure (i.e., they need about the same length of time to pronounce), they have about the same length n of LPCC sequence. We partition the whole sequence of LPCC into r equal segments. Note the number r of segments is fixed for all mandarin monosyllables. The average value of the LPCC in each segment is used as the feature value of a speech wave. Since the first 10 elements of a LPCC vector are used, we obtain $r \times 10$ feature values. Hence the $r \times 10$ matrix is the feature of a syllable. This feature extraction is the simplest among the existing speech methods.

4. Experimental results

4.1. Speech processing

The database of 10 mandarin digits was created by 21 different persons who pronounced 10 digits (0–9) once. The speech signal of a mandarin monosyllable is sampled at 8 kHz, and pre-emphasized using a transfer $1 - 0.95z^{-1}$. A Hamming window with a width of 25 ms is applied every 12.5 ms for our study.

In our experiments, we use this database to produce the LPCC and obtain a $r \times 10$ matrix for each digit sample. Among 21 samples of each mandarin digit, pick up any one sample for recognition and the rest of 20 samples is used for training, i.e., the rest of 20 samples is used to estimate the parameters which represent the mandarin digit. We assume that each element of the $r \times 10$ matrix has (1) normal distribution with weighted variance c and (2) exponential distribution (the simplest distribution of the gamma case) on the absolute values of each element in the matrix. Since the magnitude of the element in the matrix representing a syllable identifies the syllable itself, we can use the exponential distribution to classify the syllables. We also assume that $k = r \times 10$ elements in the matrix are stochastically independent in order to reduce the computational time. In the experiments, the number of segments is $r = 2, \ldots, 21$.

(1) *Normal distributions with weighted variances.* Since the average value of a random sample tends to have a normal distribution, let us assume that each element of the feature matrix has the normal distribution. The conditional normal density with weighted variance for class c_i can be represented as

$$f(x_1, \ldots, x_k \mid c_i) = \left[\prod_{l=1}^{k} \frac{1}{\sqrt{2\pi} c\sigma_{il}}\right] \mathrm{e}^{-\frac{1}{2}\sum_{l=1}^{k}\left(\frac{x_l-\mu_{il}}{c\sigma_{il}}\right)^2}, \tag{4.1}$$

where $i = 1, \ldots, m$, $k = r \times 10$ and c is a weighted factor for the variance. Taking logarithm on both sides of (4.1), the Bayes decision rule (2.3) with equal prior on each syllable becomes

$$l(c_i) = -\sum_{l=1}^{k} \log(c\sigma_{il}) - \frac{1}{2}\sum_{l=1}^{k}\left(\frac{x_l - \mu_{il}}{c\sigma_{il}}\right)^2, \quad i = 1, \ldots, m. \tag{4.2}$$

The Bayes decision rule (4.2) decides a syllable c_i with the largest $l(c_i)$ to which the feature matrix $x = (x_1, \ldots, x_k)$ belongs. For the Bayes decision rule, 20 samples of the syllable c_i are used for estimating its mean μ_{il} and variance σ_{il}^2. The weighted factor c is selected from 0.5 to 2.5. The recognition rates are in

Table 1
Digit recognition rates using normal distributions with weighted variance and exponential distribution

Segmental numbers	Normal										Exponential
r	c										
	0.50	0.90	1.10	1.25	1.37	1.50	1.75	2.00	2.50	Average	
2	.929	.957	.971	.971	.967	.967	.967	.957	.914	.956	.900
3	.957	.971	.971	.974	.976	.976	.976	.967	.933	.967	.938
4	.943	.976	.976	.978	.977	.981	.981	.967	.948	.970	.952
5	.957	.981	.981	.986	.986	.981	.976	.967	.948	.974	.971
6	.952	.976	.976	.976	.976	.976	.971	.967	.943	.968	.974
7	.957	.981	.986	.981	.981	.981	.976	.971	.948	.974	.976
8	.962	.976	.976	.976	.981	.986	.976	.967	.948	.972	.981
9	.957	.967	.971	.976	.981	.981	.976	.962	.943	.968	.961
10	.967	.971	.981	.981	.986	.986	.981	.967	.952	.975	.967
11	.967	.981	.981	.986	.986	.981	.976	.967	.943	.974	.967
12	.957	.976	.976	.976	.976	.986	.981	.971	.952	.972	.971
13	.962	.976	.981	.981	.976	.976	.971	.971	.948	.971	.962
14	.957	.981	.986	.986	.986	.986	.981	.976	.948	.976	.967
15	.962	.976	.976	.981	.981	.986	.981	.971	.957	.975	.967
16	.962	.976	.976	.976	.976	.976	.981	.957	.948	.970	.971
17	.962	.971	.971	.981	.981	.976	.971	.967	.948	.970	.971
18	.957	.967	.976	.976	.976	.981	.967	.967	.943	.968	.967
19	.952	.971	.976	.981	.981	.981	.971	.967	.952	.970	.972
20	.967	.976	.981	.981	.981	.981	.976	.971	.943	.973	.971
21	.967	.976	.976	.981	.981	.976	.971	.957	.938	.969	.971
Average	.956	.974	.977	.979	.980	.980	.975	.971	.944	.970	.965

c is a weighted factor for the variance and r is the number of segments.

Table 1. From Table 1, the best recognition rate is 98.6% and the higher rates scatter in weighted factors $c = 1.1$–1.5 and in the segment numbers $r = 5$–14.

(2) *Exponential distributions.* Since the magnitude of the LPCC of a syllable identifies the syllable itself, we let the absolute value of each element in the matrix representing a syllable to have an exponential distribution, which is the simplest probability distribution for a positive random variable. Let $Y_l = |X_l|$, $l = 1, \ldots, k$. The conditional density of $Y = (Y_1, \ldots, Y_k)$, where $k = r \times 10$, of an exponential distribution can be written as

$$f(y_1, \ldots, y_k \mid c_i) = \left(\prod_{l=1}^{k} \frac{1}{\lambda_{il}} \right) \exp\left(-\sum_{l=1}^{k} \frac{y_l}{\lambda_{il}} \right), \tag{4.3}$$

where $\lambda_{il} = E[Y_l \mid c_i]$, i.e., the mean of the absolute value of feature value in the matrix. Taking logarithm on both sides of (4.3), the Bayes decision rule (2.3) with equal prior on each syllable can simplify to

$$l(c_i) = -\sum_{l=1}^{k} \log \lambda_{il} - \sum_{l=1}^{k} \frac{y_l}{\lambda_{il}}, \tag{4.4}$$

where c_i, $i = 1, \ldots, m$, is the ith syllable. The Bayes decision rule (4.4) decides a syllable c_i with the largest $l(c_i)$ to which the feature matrix $x = (x_1, \ldots, x_k)$ belongs. For the Bayes decision rule, 20 samples of the syllable c_i are used to estimate the meanλ_{il}. The recognition rates are listed in the last column of Table 1. From Table 1, the best recognition rate is 98.1% which occurs at the segment number $r = 8$. Although the rate obtained by using exponential distributions is not quite different from the rate obtained by using normal distributions, we think that the negative values from normal distributions still contribute a little recognition rate. However, the computational time from exponential distribution is much less than that from normal distributions. Our speech recognition method using exponential distributions on feature values is the simplest and fastest classifier. Note that the calculation of $\log \lambda_{il}$ in (4.4) is part of training time, not recognition time.

5. Discussions and conclusion

In this paper, we have presented the simplest speech recognition method with low computational cost. In our method, we partition a speech data (a sequence of LPCC vectors) into equal segments and take the average of the LPCC in each segment as a feature. Our feature extraction does not use any nonlinear compression and expansion as processed by the DTW. In order to reduce the computation for classification, we use a simplified Bayes decision rule which is simpler than the simplest distance measure.

Among the existing speech recognition methods, the HMM system is the most efficient method for classification. As compared with the HMM system, our recognition system is, more or less, similar to a simplified HMM method (with left–right state model and without state durations). In our method, we partition a sequence of LPCC vectors into r equal segments. Each segment corresponds to a state in a degenerate HMM in that every state $j, j = 1, \ldots, r-1$, has stationary transition probability $\pi_{j,j+1} = 1$ and that the initial distribution on state 1 is $\pi(1) = 1$ Hence, the states represented by segments in our method are not hidden and have known transition probabilities. The average value of the LPCC in each segment is the observation which is assumed to have a normal distribution $N(\mu_{il}, \sigma_{il}^2)$ since the observation (average value) tends to have a simple normal distribution or which is assumed to have a much simpler exponential distribution. We only use the parameters $(\mu_{il}, \sigma_{il}^2)$ or only one parameter $\lambda_{\ il}$ to represent one syllable and the parameters can be easily estimated by their sample means and sample variances. To represent a syllable, the HMM system uses the parameters, which include the initial distribution, the transition probabilities and the probability of the observation O_t in each state, denoted by $(\pi, \pi_{ij}, b_j(O_t))$. The number of parameters in the HMM is much more than that of our system. The estimation of these parameters is not easy. First, one has to choose initial estimates of the HMM parameters so that the forward-backward reestimation algorithm [31,53,54], which is special case of the EM algorithm [53–55], modifies the parameters to increase the probability of the training data until a local maximum has been reached. The key question is that the local maximum may not be the global maximum of the likelihood function [54,55]. Furthermore, the probability $b_j(O_t)$ of an observation O_t in each state is difficult to estimate, since it does not have a simple normal distribution like a sample mean. For a syllable, the HMM method uses various probability distributions including Gaussian mixtures to find the probability $b_j(O_t)$ of the observation O_t. The computational cost for the convergence of the iterative reestimation algorithm and the clustering of training data into states for Gaussian mixtures is extremely high. In our method, we only estimate the individual mean in each equal segment for a syllable. In this study, we use a simplified Bayes rule for classification, which is the best classifier ((4.2) and (4.4)) with the minimum probability of misclassification and which is fast in computation. Every step in the simplified Bayes decision rule is a simple calculation except the logarithm of sample variance in (4.2) or the logarithm of sample mean in (4.4), which is part of training time, not recognition time. If each mandarin syllable has an equal probability to occur, then the Bayes rule becomes a ML classifier. However, each syllable does not occur equally likely. The classification in the HMM is to compute the probability of a sequence of observations $(O_1, \ldots, O_T)$ which is the sum of the probabilities of the sequence of observations in all possible state sequences. For the large vocabulary classification, it is time consuming. Hence, our speech recognition system for feature extraction (mean of LPCC) and classification (simplified Bayes rule) is much faster and more accurate than any other known techniques. Since our recognition method has a low computation cost, it can be widely used for speech recognition in the large vocabulary.

It is possible to extend our method to polysyllable or continuous speech. A Mandarin sentence is partitioned into a set of monosyllables which can be recognized by our method. Since a sentence can provide more information than a monosyllable, it may raise its recognition rate. In the English language, a word is partitioned into a set of basic phonemes which can be also recognized by our method.

References

[1] K.H. Davis, R. Biddulph, S. Balashek, Automatic recognition of spoken digits, J. Acoust. Soc. Amer. 24 (1952) 637–642.

[2] H. Dudley, S. Balashek, Automatic recognition of phonetic patterns in speech, J. Acoust. Soc. Amer. 30 (1958) 721–739.

[3] P.B. Denes, M.V. Mathews, Spoken digit recognition using time frequency pattern matching, J. Acoust. Soc. Amer. 32 (1960) 1450–1455.

[4] V.W. Zue, The use of speech knowledge in automatic speech recognition, Proc. IEEE 73 (11) (1985) 1602–1615.

[5] S.S. McCandless, An algorithm for automatic format extraction using linear prediction spectra, IEEE Trans. Acoust. Speech Signal Process. ASSP-22 (2) (1974) 135–141.

[6] B.S. Atal, S.L. Hanauer, Speech analysis and synthesis by linear prediction of the speech wave, J. Acoust. Soc. Amer. 50 (1971) 637–655.

[7] J. Makhoul, J. Wolf, Linear Prediction and the Spectral Analysis of Speech, Bolt, Baranek, and Newman, Inc., Cambridge, MA, 1972, Rep. 2304.

[8] F. Itakura, Minimum prediction residual principle applied to speech recognition, IEEE Trans. Acoust. Speech Signal Process. 23 (1) (1975) 67–72.

[9] J. Tierney, A study of LPC analysis of speech in additive noise, IEEE Trans. Acoust. Speech, Signal Process. 28 (4) (1980) 389–397.

[10] M.R. Sambur, L.R. Rabiner, A speaker-independent digit recognition system, B.S.T.J. 54 (1) (1975) 84–102.

[11] S.B. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Trans. Acoust. Speech Signal Process. 28 (4) (1980) 357–366.

[12] Y. Tohkura, A weighted cepstral distance measure for speech recognition, in: IEEE ICASSP 86, Tokyo, 1986, pp. 761–764.

[13] S. Morishima, H. Harashima, H. Miyakawa, A proposal of a knowledge based isolated word recognition, in: IEEE ICASSP 86, Tokyo, 1986, pp. 713–716.

[14] S. Furui, Speaker-independent isolated word recognition using dynamic features of speech spectrum, IEEE Trans. Acoust. Speech Signal Process. ASSP-34 (1) (1986) 52–59.

[15] M. Kuhn, H. Tomaschewski, H. Ney, Fast nonlinear time alignment for isolated word recognition, in: Proc. 1981 ICASSP, May 1981, pp. 736–740.

[16] J.L. Gauvain, J. Mariani, J.S. Lienard, On the use of time compression for word-based recognition, in: Proc. 1983 ICASSP, April 1983, pp. 1029–1032.

[17] J.L. Gauvain, J. Marini, Evaluation of time compressing for connected word recognition, in: Proc. 1984 ICASSP, Boston, MA, pp. 391–394.

[18] B.P. Landell, R.E. Wohlford, L.G. Bahler, Improved speech recognition in noise, in: IEEE ICASSP 86, Tokyo, 1986, pp. 749–751.

[19] S.K. Das, Some experiments in discrete utterance recognition, IEEE Trans. Acoust. Speech Signal Process. 30 (5) (1982) 766–770.

[20] A. Aktas, B. Kammerer, W. Kupper, H. Lagger, Large-vocabulary isolated word recognition with past coarse time alignment, in: IEEE ICASSP 86, Tokyo, 1986, pp. 709–712.

[21] S.L. Banner, Simulating an acoustic recognizer, IEEE ICASSP 86 (1986) 725–728.

[22] L.R. Rabiner, S.E. Levinson, A.E. Rosenberg, J.G. Wilson, Speaker independent recognition of isolated words using clustering techniques, IEEE Trans. Acoust. Speech Signal Process. 27 (1979) 336–349.

[23] C.E. Shannon, Coding theorems for a discrete source with a fidelity criterion, in: R.E. Machol (Ed.), Information and Decision Processes, McGraw-Hill, New York, 1960, pp. 93–126.

[24] Y. Linde, A. Buzo, R.M. Gray, An algorithm for vector quantizer design, IEEE Trans. Communicat. COM-28 (1) (1980) 84–95.

[25] R. Gray, Vecter quantization, IEEE ASSP Mag. (April) (1984) 4–29.

[26] J. Makhoul, S. Roucos, H. Gish, Vector quantization in speech coding, Proc. IEEE 73 (11) (1985) 1551–1588.

[27] A. Buzo, A. Gray, R. Gray, J. Markel, Speech coding based upon vector quantization, IEEE Trans. Acoust. Speech Signal Process. ASSP-28 (5) (1980) 562–573.

[28] J.E. Shore, D.K. Burton, Discrete utterance speech recognition without time alignment, IEEE Trans. Inform. Theory 29 (1983) 473–491.

[29] D. Burton, J. Shore, J. Buck, Isolated-word speech recognition using multisection vector quantization codebooks, IEEE Trans. Acoust. Speech Signal Process. ASSP-33 (4) (1985) 837–849.

[30] B.H. Juang, D.Y. Wong, A.H. Gray Jr., Distortion performance of vector quantization for LPC voice coding, IEEE Trans. Acoust. Speech Signal Process. 30 (2) (1982) 294–303.

[31] L.E. Baum, T. Petrie, G.R. Soules, N. Weiss, A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, Ann. Math. Statist. 41 (1) (1970) 164–171.

[32] L.R. Rabiner, S.E. Levinson, M.M. Sondhi, On the application of vector quantization and hidden Markov models to speaker-independent, isolated word recognition, B.S.T.J. 62 (6) (1983) 1075–1105.

[33] A. Poritz, A. Richter, On hidden Markov models in isolated word recognition, in: IEEE ICASSP 86, Tokyo, 1986, pp. 705–708.

[34] L. Bahl, P.F. Brown, P.V. de Souza, R.L. Mercer, Maximum mutual information estimation of hidden Markov model parameters, in: IEEE ICASSP 86, Tokyo, 1986, pp. 49–52.

[35] S. Soudoplatoff, Markov modeling of continuous parameters in speech recognition, in: IEEE ICASSP 86, Tokyo, 1986, pp. 45–48.

[36] Y. Kamp, State reduction in hidden Markov chain used for speech recognition, IEEE Trans. Acoust. Speech Signal Process. ASSP-33 (4) (1985) 1138–1145.

[37] L.R. Bahl, F. Jelinek, R.L. Mercer, A maximum likelihood approach to continuous speech recognition, IEEE Trans. Pattern Anal. Mach. Intell. PAMI-5 (2) (1983) 179–190.

[38] L.R. Rabiner, S. Levinson, A speaker-independent, syntax-directed, connected word recognition system based on hidden Markov models and level building, IEEE Trans. Acoust. Speech Signal Process. ASSP-33 (3) (1985) 561–672.

[39] K.F. Lee, Context-dependent phonetic hidden Markov models for speaker independent continuous speech recognition, IEEE Trans. Acoust. Speech Signal Process. ASSP-38 (4) (1990) 599–609.

[40] L.R. Rabiner, R.H. Juang, S.E. Levinson, M.M. Sondhi, Recognition of isolated digit using hidden Markov models with continuous mixture densities, AT&T Tech. J. 64 (6) (1985) 1211–1233.

[41] M. Wagner, W. Wang, H. Ho, M. O'Kane, Isolated-word recognition of the complete vocabulary of spoken Chinese, in: Proc. ICASSP, Tokyo, 1986, pp. 701–706.

[42] Y.K. Chen, C.Y. Liu, G.H. Chiang, M.T. Lin, The recognition of mandarin monosyllables based on the discrete hidden Markov model, in: Proc. Telecommun. Symp., Taiwan, ROC, 1990, pp. 133–137.

[43] L.S. Lee, C.Y. Tseng, F.H. Liu, C.H. Chang, H.Y. Gu, S.H. Hsieh, C.H. Chen, Special speech recognition approaches for the highly confusing mandarin syllables based on hidden Markov models, Comput. Speech Language 5 (2) (1991) 191–201.

[44] H.Y. Gu, C.Y. Tseng, L.S. Lee, Markov modeling of mandarin Chinese for decoding the phonetic sequence into Chinese characters, Comput. Speech Language 5 (4) (1991) 363–377.

[45] L.S. Lee, C.Y. Tseng, H.Y. Gu, K.J. Chen, F.H. Liu, C.H. Chang, S.H. Hsieh, C.H. Chen, A real-time mandarin dictation machine for Chinese language with unlimited tests and very large vocabulary, in: Proc. 1990 ICASSP, Albuquerque, NM, USA, 1990, pp. 65–68.

[46] L.S. Lee, C.Y. Tseng, K.J. Chen, I.J. Hung, M.Y. Lee, L.F. Chien, Y.M. Lee, R.Y. Lyu, H.M. Wang, Y.C. Chang, T.S. Lin, H.Y. Gu, C.P. Nee, C.Y. Liao, Y.J. Yang, Y.C. Chang, R.C. Yang, Golden mandarin (II)—an improved single-chip real-time mandarin dictation machine for Chinese language with vary large vocabulary, in: Proc. 1993 ICASSP, pp. 503–506.

[47] L.S. Lee, J.T. Chen, An initial study on speaker adaptation techniques for isolated mandarin syllable recognition, in: Proc. Telecommun. Symp., Taiwan, 1990, pp. 115–121.

[48] Y.Q. Gao, T.Y. Huang, Z.W. Lin, B. Xu, D.X. Xu, A real-time Chinese speech recognition system with unlimited vocabulary, in: Proc. ICASSP 1991, pp. 257–260.

[49] H.W. Hon, B.S. Yuan, Y.L. Chow, S. Naryan, K.F. Lee, Toward large vocabulary mandarin Chinese speech recognition, in: Proc. ICASSP 1994, pp. 545–548.

[50] M. Elghonemy, M. Fikri, M. Hashish, E. Talkhan, Speaker independent isolated Arabic word recognition system, in: IEEE ICASSP 1986, Tokyo, pp. 697–699.

[51] K. Fukunage, Introduction to Statistical Pattern Recognition, Academic Press, New York, 1990.

[52] L. Rabiner, B.H. Juang, Fundamentals of Speech Recognition, Prentice-Hall PTR, Englewood Cliffs, NJ, 1993.

[53] L.E. Baum, An inequality and associated maximization technique in statistical estimation of probabilistic functions of Markov processes, Inequalities 3 (1972) 1–8.

[54] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Ann. R. Statist. Soc. 39 (1977) 1–35.

[55] C.F.J. Wu, On the convergence properties of the EM algorithm, Ann. Statist. 11 (1983) 95–103.