One-Sample Speech Recognition of Mandarin Monosyllables using Unsupervised Learning

By

Tze Fen Li

Institute of Management, Ming Dao University, Chang-Hua, Taiwan, ROC

and

Shui-Ching Chang

Department of Information Management, The Overseas Institute of Technology, Taichung, Taiwan, ROC

Abstract

In the speech recognition, a mandarin syllable wave is compressed into a matrix of linear predict coding cepstra (LPCC), i.e., a matrix of LPCC represents a mandarin syllable. We use the Bayes decision rule on the matrix to identify a mandarin syllable. Suppose that there are K different mandarin syllables, i.e., K classes. In the pattern classification problem, it is known that the Bayes decision rule, which separates K classes, gives a minimum probability of misclassification. In this study, a set of unknown syllables is used to learn all unknown parameters (means and variances) for each class. At the same time, in each class, we need one known sample (syllable) to identify its own means and variances among K classes. Finally, the Bayes decision rule classifies the set of unknown syllables and input unknown syllables. It is an one-sample speech recognition. This classifier can adapt itself to a better decision rule by making use of new unknown input syllables while the recognition system is put in use. In the speech experiment using unsupervised learning to find the unknown parameters, the digit recognition rate is improved by 22%.

Key words and phrases: classification, dynamic processing algorithm, EM (estimate maximize) algorithm, empirical Bayes, maximum likelihood estimation, speech recognition.

Corresponding author address: Tze Fen Li, Institute of Management, Ming Dao University, 369 Wen-Hua Road, Pee-Tow, Chang-Hua (52345), Taiwan, ROC.

email address(Tze Fen Li): tfli@mdu.edu.tw

1. Introduction

A speech recognition system in general consists of feature extractor and classification of an utterance [1-5]. The function of feature extractor is to extract the important features from the speech waveform of an input speech syllable. Let x denote the measurement of the significant, characterizing features. This x will be called a feature value. The function performed by a classifier is to assign each input syllable to one of several possible syllable classes. The decision is made on the basis of feature measurements supplied by the feature extractor in a recognition system. Since the measurement x of a pattern may have a variation or noise, a classifier may classify an input syllable to a wrong class. The classification criterion is usually the minimum probability of misclassification [1].

In this study, a statistical classifier, called an empirical Bayes (EB) decision rule, is applied to solving K-class pattern problems: all parameters of the conditional density function $f(x \mid \omega)$ are unknown, where ω denotes one of K classes, and the prior probability of each class is unknown. A set of n unidentified input mandarin monosyllables is used to establish the decision rule, which is used to separate K classes. After learning the unknown parameters, the EB decision rule will make the probability of misclassification arbitrarily close to that of the Bayes rule when the number of unidentified patterns increases. The problem of learning from unidentified samples (called unsupervised learning or learning without a teacher) presents both theoretical and practical problems [6-8]. In fact, without any prior assumption, successful unsupervised learning is indeed unlikely.

In our speech recognition using unsupervised learning, a syllable is denoted by a matrix of features. Since the matrix has 8x12 feature values, we use a dynamic processing algorithm to estimate the 96 feature parameters (means and variances). Our EB classifier, after unsupervised learning of the unknown parameters, can adapt itself to a better and more accurate decision rule by making use of the unidentified input syllables after the speech system is put in use. The results of a digit speech experiment are given to show the recognition rates provided by the decision rule.

2. Empirical Bayes Decision Rules for Classification

Let X be the present observation which belongs to one of K classes $c_i, i = 1, 2, \cdots, K$. Consider the decision problem consisting of determining whether X belongs to c_i. Let $f(x \mid \omega)$ be the conditional density function of X given ω, where ω denotes one of K classes and let $\theta_i, i = 1, 2, \cdots, K$, be the prior probability of c_i with $\sum_{i=1}^{K} \theta_i = 1$. In this study, both the parameters of $f(x \mid \omega)$ and the θ_i are unknown. Let d be a decision rule. A simple loss model is used such that the loss is 1 when d makes a wrong decision and the loss is 0 when d makes a correct decision. Let $\theta = \{(\theta_1, \theta_2, \cdots, \theta_K); \theta_i > 0, \sum_{i=1}^{K} \theta_i = 1\}$ be the prior probabilities. Let $R(\theta, d)$ denote the risk function (the probability of misclassification) of d. Let $\Gamma_i, i = 1, 2, \cdots, K$, be K

regions separated by d in the domain of X, i.e., d decides c_i when $X \in \Gamma_i$. Let ξ_i denote all parameters of the conditional density function in class c_i, $i = 1, ..., K$. Then

$$R(\theta, d) = \sum_{i=1}^{K} \int_{\Gamma_i^c} \theta_i \, f(x \mid \xi_i) dx \tag{1}$$

where Γ_i^c is the complement of Γ_i . Let D be the family of all decision rules which separate K pattern classes. For θ fixed, let the minimum probability of misclassification be denoted by

$$R(\theta) = \inf_{d \in D} R(\theta, d). \tag{2}$$

A decision rule d_θ which satisfies (2) is called the Bayes decision rule with respect to the prior probability vector $\theta = (\theta_1, \theta_2, \cdots, \theta_K)$ and given by Ref.[1]

$$d_\theta(x) = c_i \quad \text{if} \quad \theta_i \, f(x \mid \xi_i) > \theta_j \, f(x \mid \xi_j) \quad \textit{for all} \;\; j \neq i. \tag{3}$$

In the empirical Bayes (EB) decision problem [9], the past observations (ω_m, X_m), $m = 1, 2, \cdots, n$, and the present observation (ω, X) are i.i.d., and all X_m are drawn from the same conditional densities, i.e., $f(x_m \mid \omega_m)$ with $p(\omega_m = c_i) = \theta_i$. The EB decision problem is to establish a decision rule based on the set of past observations $\mathbf{X}_n = (X_1, X_2, \cdots, X_n)$. In a pattern recognition system with unsupervised learning, $\mathbf{X}_n$ is a set of unidentified input patterns. The decision rule can be constructed using $\mathbf{X}_n$ to select a decision rule $t_n(\mathbf{X}_n)$ which determines whether the present observation X belongs to c_i. Let $\xi = (\xi_1, ..., \xi_K)$. Then the risk of t_n, conditioned on $\mathbf{X}_n = \mathbf{x}_n$, is $R(\theta, t_n(\mathbf{x}_n)) \geq R(\theta)$ and the overall risk of t_n is

$$R_n(\theta, t_n) = \int R(\theta, t_n(\mathbf{x}_n)) \prod_{m=1}^{n} p(x_m \mid \theta, \xi) \, dx_1 \cdots dx_n \tag{4}$$

where $p(x_m \mid \theta, \xi)$ is the marginal density of X_m with respect to the prior distribution of classes, i.e., $p(x_m \mid \theta, \xi) = \sum_{i=1}^{K} \theta_i f(x_m \mid \xi_i)$. The EB approach has been recently used in many areas including classification [10,11], sequential estimation [12], reliability [13-15], multivariate analysis [16,17], linear models [18,19], nonparametric estimation [20,21] and some other estimation problems [22,23]. Let

$$S = \{(\theta, \xi); \theta = (\theta_1, ..., \theta_K), \; \xi = (\xi_1, ..., \xi_K)\} \tag{5}$$

define a parameter space of prior probabilities θ_i and parameters ξ_i representing the i-th class, $i = 1, ..., K$. Let P be a probability distribution on the parameter space S. In this study, we want to find an EB decision rule which minimizes

$$\hat{R}_n(P, t_n) = \int R_n(\theta, t_n) dP(\theta, \xi). \tag{6}$$

Similar approaches to constructing EB decision rules can be found in the recent literature [11,15,24]. From (1) and (4), (6) can be written as

$$\hat{R}_n(P,t_n) = \int \sum_{i=1}^{K} \int_{\Gamma_{i,n}^c} \big[\int f(x \mid \xi_i)\theta_i \prod_{m=1}^{n} p(x_m \mid \theta,\xi) dP(\theta,\xi)\big] dx\, dx_1 \cdots dx_n \tag{7}$$

where, in the domain of X, $\Gamma_{i,n},\ i = 1, 2, \cdots, K$, are K regions, separated by $t_n(\mathbf{X}_n)$, i.e., $t_n(\mathbf{X}_n)$ decides c_i when $X \in \Gamma_{i,n}$ and hence they depend on the past observations $\mathbf{X}_n$. The EB decision rule which minimizes (7) can be found in Ref[24]. Since the unsupervised learning in this study is based on the following two theorems given in Ref[24], both theorems and their simple proofs are provided in this paper.

Theorem 1 [24]. The EB decision rule $\hat{t}_n$ with respect to P which minimizes the overall risk function (7) is given by

$$\begin{aligned} \hat{t}_n(\mathbf{x}_n)(x) = c_i \qquad \text{if} \qquad & \int f(x \mid \xi_i)\, \theta_i \prod_{m=1}^{n} p(x_m \mid \theta,\xi) dP(\theta,\xi) > \\ & \int f(x \mid \xi_j)\, \theta_j \prod_{m=1}^{n} p(x_m \mid \theta,\xi) dP(\theta,\xi) \end{aligned} \tag{8}$$

for all $j \neq i$, i.e., $\Gamma_{i,n}$ is defined by the definition of the inequality in (8).

Proof. To minimize the overall risk (7) is to minimize the integrand

$$\sum_{i=1}^{K} \int_{\Gamma_{i,n}^c} \big[\int f(x|\xi_i)\theta_i \prod_{m=1}^{n} p(x_m|\theta,\xi) dP(\theta,\xi)\big] dx$$

of (7) for each past observations $\mathbf{x}_n$. Let the past obervations $\mathbf{x}_n$ be fixed and let i be fixed for $i = 1, ..., k$. Let

$$g_i(x) = \int f(x|\xi_i)\theta_i \prod_{m=1}^{n} p(x_m|\theta,\xi) dP(\theta,\xi).$$

Then the integrand of (7) can be written as

$$\sum_{i=1}^{K} \int_{\Gamma_{i,n}^c} g_i(x)dx = \int_{\Gamma_{i,n}^c} g_i(x)dx + \sum_{j\neq i} [\int g_j(x)dx - \int_{\Gamma_{j,n}} g_j(x)dx]$$

$$= \sum_{j\neq i} \int g_j(x)dx + \sum_{j\neq i} \int_{\Gamma_{j,n}} [g_i(x) - g_j(x)]dx \quad (\Gamma_{i,n}^c = \sum_{j\neq i} \Gamma_{j,n})$$

which is minimum since $\Gamma_{j,n} \subset \{x|g_j(x) > g_i(x)\}$ for all $j \neq i$ by the definition of $\Gamma_{j,n}$.

In applications, we let the parameters $\xi_i,\ i = 1, ..., K$, be bounded by a finite numbers M_i. Let $\rho > 0$ and $\delta > 0$. Consider the subset S_1 of the parameter space S defined by

$$S_1 = \{(n_1\rho, n_2\rho, ..., n_K\rho, n_{K+1}\delta, n_{K+2}\delta, ..., n_{2K}\delta); \quad integer\ n_i > 0,\ i = 1, ..., K,$$
$$\sum_{i=1}^{K} n_i\rho = 1, |n_i\delta| \le M_i,\ integer\ n_i,\ i = K+1, ..., 2K\} \tag{9}$$

where $(n_1\rho, ..., n_K\rho)$ are prior probabilities and $(n_{K+1}\delta, ..., n_{2K}\delta)$ are the parameters of K classes. In order to simplify the conditional density of (θ, ξ), let P be a uniform distribution on S_1 so that the conditional density can later be written as a recursive formula. The boundary for class i relative to another class j as separated by (8) can be represented by the equation

$$E[f(x \mid \xi_i)\theta_i \mid \mathbf{x}_n] = E[f(x \mid \xi_j)\theta_j \mid \mathbf{x}_n] \tag{10}$$

where $E[f(x \mid \xi_i)\theta_i \mid \mathbf{x}_n]$ is the conditional expectation of $f(x \mid \xi_i)\theta_i$ given $\mathbf{X}_n = \mathbf{x}_n$ with the conditional probability function of (θ, ξ) given $\mathbf{X}_n = \mathbf{x}_n$ equal to

$$h(\theta, \xi \mid \mathbf{x}_n) = \frac{\prod_{m=1}^{n} p(x_m \mid \theta, \xi)}{\sum_{(\theta'\xi')\in S_1} \prod_{m=1}^{n} p(x_m \mid \theta', \xi')} \tag{11}$$

The actual region for class i as determined by (8) is the intersection of the regions whose borders are given by (10), relative to all other classes.

The main result in Ref[24] is that the estimates $E[\theta_i \mid \mathbf{X}_n]$ converge almost sure (a.s.) to a point arbitrarily close to the true prior probability and $E[\xi_i|\mathbf{X}_n]$ will converge to a point arbitrarily close to the true parameter in the conditional density for the i-th class. Let $\lambda = (\theta_1, ..., \theta_K, \xi_1, ..., \xi_K)$ in the parameter space S. Let λ^o be the true parameter of λ.

Lamma 1 (Kullback, 1973 [25]). Let

$$H(\lambda^o, \lambda) = \int ln\, p(x|\lambda) p(x|\lambda^o) dx.$$

Then the Kullback-Leibler information number $H(\lambda^o, \lambda^o) - H(\lambda^o, \lambda) \ge 0$ with equality if and only if $p(x|\lambda) = p(x|\lambda^o)$ for all x, i.e., $H(\lambda^o, \lambda)$ has an absolutely maximum value at $\lambda = \lambda^o$.

Let $\lambda' = (\theta', \xi') \in S_1$ such that $H(\lambda^o, \lambda') = max_{\lambda \in S_1} H(\lambda^o, \lambda)$. Since S_1 has a finite number of points, $H(\lambda^o, \lambda') - H(\lambda^o, \lambda) \ge \epsilon$ for some $\epsilon > 0$ and for all $\lambda \in S_1$. Since $H(\lambda^o, \lambda)$ is a smooth (differentiable) function of $\lambda \in S$, the maximum point λ' in S_1 is arbitrarily close to the true parameter λ^o in S if the increments δ and ρ are small.

Theorem 2 [24]. Let λ^o be the true parameter of λ. Let $\lambda = (\theta, \xi)$ in S. The conditional probability function $h(\lambda|\mathbf{x}_n)$ given $\mathbf{X}_n = \mathbf{x}_n$ in (11) has the following property: for each $\lambda \in S_1$,

$$\lim_{n\to\infty} h(\lambda \mid \mathbf{x}_n) = 0 \qquad \text{if } \lambda \ne \lambda'$$
$$= 1 \qquad \text{if } \lambda = \lambda' \tag{12}$$

and hence $E[\lambda \mid \mathbf{X}_n]$ converges to λ' with probability 1.

Proof. $H(\lambda^o, \lambda)$ has an absolutely maximum value at $\lambda = \lambda'$ on S_1. Let $\lambda \in S_1$ and $\lambda \neq \lambda'$. Consider

$$\frac{1}{n} ln \frac{\prod_{m=1}^{n} p(X_m|\lambda)}{\prod_{m=1}^{n} p(X_m|\lambda')} = \frac{1}{n} \sum_{m=1}^{n} ln\, p(X_m|\lambda) - \frac{1}{n} \sum_{m=1}^{n} ln\, p(X_m|\lambda')$$

which converges almost sure to $H(\lambda^o, \lambda) - H(\lambda^o, \lambda') < -\epsilon$ by a theorem (the strong law of large numbers, Wilks, (1962) [26]), i.e., there exists a $N > 0$ such that for all $n > N$,

$$\frac{1}{n} ln \frac{\prod_{m=1}^{n} p(X_m|\lambda)}{\prod_{m=1}^{n} p(X_m|\lambda')} < -\frac{\epsilon}{2}.$$

Hence, for all $n > N$, $\frac{1}{n} ln\, h(\lambda|\mathbf{X}_n) < -\frac{\epsilon}{2}$, i.e., for all $n > N$, $ln\, h(\lambda|\mathbf{X}_n) < -n\frac{\epsilon}{2}$. This implies that $lim_{n\to\infty} ln\, h(\lambda|\mathbf{X}_n) = -\infty$ and $lim_{n\to\infty} h(\lambda|\mathbf{X}_n) = 0$ for $\lambda \neq \lambda'$ almost sure. Obviousy, $\sum_{\lambda\in S_1} h(\lambda|\mathbf{X}_n) = 1$ implies $lim_{n\to\infty} h(\lambda'|\mathbf{X}_n) = 1$ almost sure.

3. Feature Extraction

The measurements of features made on the speech waveform include energy, zero crossings. extrema count, formants, LPC cepstrum (LPCC) and the Mel frequency cepstrum coefficient (MFCC). The LPCC and MFCC are most commonly used for the features to represent a syllable. 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 wave. 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.

3.1. Preprocessing Speech Signal

In the real world, all signals contain noise. In our speech recognition system, the speech data must contain noise. We propose two simple methods to eliminate noise. One way is to use the sample variance of a fixed number of sequential sampled points of a syllable wave to detect the real speech signal, i.e., the sampled points with small variance does not contain real speech signal. Another way is to compute the sum of the absolute values of differences of two consecutive sampled points in a fixed number of sequential speech sampled points, i.e., the speech data with small absolute value does not contain real speech signal. In our speech recognition experiments, the latter provides slightly faster and more accurate speech recognition.

3.2. Linear Predict Coding Cepstrum (LPCC)

For speech recognition, the most common features to be extracted from a speech signal are Mel-frequency cepstrum coefficient (MFCC) and linear predict coding cepstrum (LPCC). The MFCC was proved to be

better than the LPCC for recognition [27], but we have shown [28] that the LPCC has a slightly higher recognition rate. Since the MFCC has to compute the DFT and inverse DFT of a speech wave, the computational complexity is much heavier than that of the LPCC. The LPC coefficients can be easily obtained by Durbin's recursive procedure [2,29,30] and their cepstra can be quickly found by another recursive equations [2,29,30]. The LPCC can provide a robust, reliable and accurate method for estimating the parameters that characterize the linear and time-varying system like speech signal [2,4,29-30]. Therefore, in this study, we use the LPCC as the feature of a mandarin syllable. The following is a brief discussion on the LPC method:

It is assumed [2-4] that the sampled speech wave $s(n)$ can be linearly predicted from the past p samples of $s(n)$. Let

$$\hat{s}(n) \ = \ \sum_{k=1}^{p} a_k s(n-k) \tag{13}$$

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

$$E \ = \ \sum_{n=0}^{N-1} [s(n) - \hat{s}(n)]^2. \tag{14}$$

The unknown $a_k, \ k = 1, ...p$, 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 [31]. Here in our speech experiment, $p = 12$, because the cepstra in the last few elements are almost zero.

3.3. Feature Extraction

Our feature extraction from LPCC is quite simple. Let $x(k) = (x(k)_1,...,x(k)_p)$, $k = 1,..,n$, be the LPCC vector for the k-th frame of a speech wave in the sequence of n vectors. Normally, if a speaker does not intentionally elongate pronunciation, a mandarin syllable has 30-70 vectors of LPCC. After 50 vectors of LPCC, the sequence does not contain significant features.

Since an utterance of a syllable is composed two parts: stable part and feature part. In the feature part, the LPCC vectors have a dramatic change between two consecutive vectors, representing the unique characteristics of syllable utterance and in the stable part, the LPCC vectors do not change much and stay about the same. Even if the same speaker utters the same syllable, the duration of the stable part of the sequence of LPCC vectors changes every time with nonlinear expansion and contraction and hence the duration of the stable parts and the duration of the whole sequence of LPCC vectors are different every time. Therefore, the duration of stable parts is contracted such that the compressed speech waveforms have about the same length of the sequence of LPCC vectors. Li [32] proposed several simple techniques to contract the stable parts of the sequence of vectors. We state one simple technique for contraction as follows:

Let $x(k) = (x(k)_1, ..., x(k)_p),\ k = 1, ..., n$, be the k-th vector of a LPCC sequence with n vectors, which represents a mandarin syllable. Let the difference of two consecutive vectors be denoted by

$$D(k) = \sum_{i=1}^{p} |x(k)_i - x(k-1)_i|, \quad k = 2, ..., n. \tag{15}$$

In order to accurately identify the syllable utterance, a compression process must first be performed to remove the stable and flat portion in the sequence of vectors. A LPCC vector $x(k)$ is removed if its difference $D(k)$ from the previous vector $x(k-1)$ is too small. Let $x'(k),\ k = 1, ..., m(< n)$, be the new sequence of LPCC vectors after deletion. We think that the first part (about 40 vectors or less) of an utterance of a mandarin syllable contains main features which can most represent the syllable and the rest of the sequence contains the "tail" sound, which has a variable length. If a speaker intentionally elongates pronunciation of a syllable, the speaker only increases the tail part of the sequence and the length of the feature part stays about the same. We partition the feature part (the first 40 vectors of the new sequence) into 6 equal segments since the feature part of LPCC vectors has a dramatic change and partition the tail part into 2 equal segments. If the whole length of the new sequence is less than 40, we neglect the tail sound and partition the new sequence into 8 equal segments. The average value of the LPCC in each segment is used as a feature value. Note that the average values of samples tend to have a normal distribution [26]. This compression produces 12x8 feature values for each mandarin syllable.

4. Stochastic Approximation

Stochastic approximation [1,2,33,34] is an iterative algorithm for random environments, which is used for parameter estimation in pattern recognition. Its convergence is guaranteed under very general circumstances. Essentially, a stochastic approximation procedure [1,2,33,34] should satisfy: (1) the successive expression of the estimate of a parameter can be written as an estimate calculated from the old n patterns and the contribution of the new $(n+1)$-st pattern and (2) the effect of the new pattern may diminish by using a decreasing sequence of coefficients. The best known of the stochastic approximation procedures are the Robbins-Monro procedure [1,33,34] and the Kiefer-Wolfowitz procedure [1,34].

For the unsupervised learning, (11) can be written in the recursive form

$$h(\lambda|\mathbf{x}_{n+1}) = \frac{p(x_{n+1}|\lambda)h(\lambda|\mathbf{x}_n)}{\sum_{\lambda' \in S_1} p(x_{n+1}|\lambda')h(\lambda'|\mathbf{x}_n)} \quad for\ n = 0, 1, 2, ... \tag{16}$$

where $h(\lambda|\mathbf{x}_n) = 1$, if $n = 0$. Equ. (16) is different from the above two types of procedures. It does not have a regression function or an obvious decreasing sequence of coefficients, but it appears to be a weighted product of the estimates calculated from the old patterns and the contribution of the new pattern. In each

step of evaluation, (16) multiplies a new probability factor with the old conditional probability $h(\lambda|\mathbf{x}_n)$ based on the new pattern x_{n+1}. The convergence of (16) is guaranteed by Theorem 2.

5. A Dynamic Processing Algorithm

As in Section 3, a mandarin syllable is represented by a 12x8 matrix of feature values, which tend to be normally distributed. Let $\mathbf{x}_n = (x_1, ..., x_n)$ denote n unidentified syllables, where each x_m, $m = 1, ..., n$, denotes a 12x8 matrix of feature values, which are used to learn the means μ_{kij}, variances σ^2_{kij}, $i = 1, ..., 12$, $j = 1, ..., 8$, $k = 1, ..., K$, of normal distributions of 12x8 feature values and the prior probabilities θ_k (the probability for a syllable to appear) for K classes of syllables. For large number of classes, the stochastic approximation procedure in Section 4 is not able to estimate the means and variances, because the recursive procedure (16) needs tremendous size of computer memory. For simplicity, we let $\theta_k = 1/K$, i.e., each syllable has an equal chance to be pronounced. Let λ denote all parameters, i.e., Kx12x8 means and variances for K classes of syllables. Let λ^o be the true parameters. From Theorem 2 in Section 2, the conditional probability $h(\lambda|\mathbf{x}_n)$ has the maximum probability at $\lambda = \lambda^o$ for large n, i.e., the numerator

$$F(\mathbf{x}_n|\lambda) = \prod_{m=1}^{n} p(x_m|\lambda) \tag{17}$$

is maximum at $\lambda = \lambda^o$ for large n, where x_m, $m = 1, ..., n$, is the 12x8 matrix. Therefore, to search the true parameter λ^o by the recursive equation (16) is to find the MLE of λ.

To find the MLE of unknown parameters is a complicated multi-parameter optimization problem. First one has to evaluate the likelihood function F on a coarse grid to locate roughly the global maximum and then apply a numerical method (Gauss method, Newton-Raphson or some gradient-search iterative algorithm). Hence the direct approach tends to be computationally complex and time consuming. Here, we use a simple dynamic processing algorithm to find the MLE, which is similar to an EM [35,36] algorithm.

5.1. The Log Likelihood Function

A syllable is denoted by a matrix of feature values X_{ij}, $i = 1, ..., 12$, $j = 1, ..., 8$. For simplicity, we assume that the 12x8 random variables X_{ij} are stochastically independent (as a matter of fact, they are not independent). The marginal density function of an unidentified syllable X_m with its matrix denoted by $x_m = (x^m_{ij})$ in (17) can be written as

$$p(x_m|\lambda) = \sum_{k=1}^{K} \theta_k \prod_{ij} f(x^m_{ij}|\mu_{ijk}, \sigma_{ijk}) \tag{18}$$

where $f(x_{ij}^m|\mu_{ijk}, \sigma_{ijk})$ is the conditional normal density of the feature value X_{ij}^m in the matrix if the syllable $X_m = (X_{ij}^m)$ belongs to the k-th class. The log likelihood function can be written as

$$ln\, F(\mathbf{x}_n|\lambda) = \sum_{m=1}^{n} ln\,\{\sum_{k=1}^{K}\theta_k \prod_{i=1}^{12}\prod_{j=1}^{8}\frac{1}{\sqrt{2\pi}\sigma_{kij}}e^{-\frac{1}{2}(\frac{x_{ij}^m-\mu_{kij}}{\sigma_{kij}})^2}\}. \tag{19}$$

5.2. A Dynamic Processing Algorithm

From the log likelihood function (19), we present a simple dynamic processing algorithm to find the MLE of unknown parameters μ_{ijk} and σ_{ijk}. Our algorithm is an EM algorithm [35,36], more and less like the Viterbi algorithm [2-4]. We state the our dynamic processing algorithm as follows:

1. In the matrix, pick up an initial value of $(\mu_{kij}, \sigma_{kij})$, $k = 1, ..., K$, for K classes.

2. For $k = 1$ and for each $i = 1, ..., 12$ and $j = 1, ..., 8$, pick up a point $(\hat{\mu}_{1ij}, \hat{\sigma}_{1ij})$ such that $ln\, F$ in (19) is maximum.

3. Continue step 2 for $k = 2, ..., K$.

4. If (19) continues increasing, go to step 2, otherwise, stop the dynamic processing and the final estimates $(\hat{\mu}_{kij}, \hat{\sigma}_{kij})$ are the MLE of $(\mu_{kij}, \sigma_{kij})$ for all K classes and are saved in a database.

5.3. Finding the Means and Variances for each Syllable by a Known Sample

For each element (i, j) in the matrix, we have found the MLE $(\hat{\mu}_{kij}, \hat{\sigma}_{kij})$ for each syllable. There are totally K matrices of MLE representing K different syllables, but we do not know which matrix of MLE belongs to the syllable $c_i,\ i = 1, ..., K$. We have to use one known sample from each syllable to identify its own matrix of MLE. In this paper, we simply use the distance to select a matrix of MLE among K matrices for the known sample.

5.4. Classification by the Bayes Decision Rule

After each syllable obtains its means and variances which are identified by a known sample of the syllable, the Bayes decision rule (3) with the estimated means and variances (MLE) classifies the set of all unidentified syllables. After simplification [32], the Bayes decision rule (3) can be reduced to

$$l(c_k) = \sum_{ij} ln(\hat{\sigma}_{kij}) + \frac{1}{2}\sum_{ij}(\frac{x_{ij}-\hat{\mu}_{kij}}{\hat{\sigma}_{kij}})^2 \tag{20}$$

where $\{x_{ij}\}$ denotes the matrix of LPCC of an input unknown syllable. The matrix of LPCC of an unknown syllable is compared with each known syllable c_k represented by $(\hat{\mu}_{kij}, \hat{\sigma}_{kij})$. The Bayes rule (20) selects a syllable c_k with the least value of $l(c_k)$ from K known syllables to be the input unknown syllable.

Note that new input unidentified syllables can update the estimated means and variances (MLE) which are closer to the true unknown means and variances, and hence the Bayes decision rule will become a more accurate classifier.

6. Speech Experiment on Classification of Digits

Our speech recognition is implemented in a classroom. The data of 10 mandarin digits are created by 10 different male and female students, each pronouncing 10 digits (0-9) once. The mandarin pronunciation for 1 and 7 is almost the same. It is hard to classify these two syllables.

6.1. Speech Signal Processing.

The speech signal of a mandarin monosyllable is sampled at 10k Hz. A Hamming window with a width of 25.6 ms is applied every 12.8 ms for our study. A Hamming window with 256 points is used to select the data points to be analyzed. In this study, the 12x8 unknown parameters of features representing a digit are estimated by unsupervised learning. After learning the parameters, there are 10 12x8 matrices of estimates representing 10 digits. For each digit, use one known sample to identify a 12x8 matrix of estimates to represent the digit.

In our speech experiments, we use this database to produce the LPCC and obtain a 12x8 matrix of feature values for each syllable. There are totally 100 matrices of feature values.

6.2. To Learn Means and Variances using Unsupervised Learning

The simple dynamic processing algorithm in Section 5 produces 10 matrices of MLE (estimated means and variances). After a known sample of each digit (0,1,...,9) picks up its own matrix of MLE, the 10 matrices are ranked in order from 0 to 9 as fellows: $(\hat{\mu}_{kij}, \hat{\sigma}_{kij})$, $i = 1, ..., 12$, $j = 1, ..., 8$, for $k = 0, ..., 9$. One of 10 students pronounces 10 digits which are considered as 10 known samples (each for one digit) and the other 9 students pronounce 10 digits (90 samples), which are considered as unknown samples. The total 100 samples (10 known samples and 90 unknown samples) are used for finding the matrices of MLE of the means and variances for 10 digits. This experiment is implemented five times, each time for one of five different students whose 10 digit pronunciations are considered as known samples. Note that the only training samples are the only one sample for each digit pronounced by a student and note that the testing samples are the mixed 90 unknown samples of 10 digits pronounced by the other 9 students. Actually, the experiment is a speaker-independent speech recognition. The 10 training samples and the 90 testing samples (90 mixed unknown samples also used for unsupervised learning of parameters) are totally separated.

6.3. Speech Classification on the Mixed Samples

In this study, two different classifiers are used to classify 90 unknown mixed digital samples since 10 digital samples pronounced by one student are already known.

(a). Bayes Decision Rule.

The estimated means and variances of each digit obtained in (6.2) are placed into the Bayes decision rule (20). The Bayes decision rule classifies 90 mixed samples (except 10 known samples for 10 digits (0-9)). The recognition rates are listed in Table 1.

(b). Distance Measure from 10 Known Samples

The known sample of a digit (0-9) identifies 90 other mixed unknown samples using distance measure from the known sample, i.e., to classify an unknown sample, we select a known sample from 10 known samples which is the closest to the unknown sample to be the unknown sample. Its recognition rates are also listed in Table 1. From Table 1, the Bayes decision rule using unsupervised learning gives the higher recognition rate 79%, 22% more than the rate 57% given by the distance measure using one known sample.

Table 1. Recognition rates for 10 digits given by the Bayes decision rule with unsupervised learning to classify 90 unknown samples as compared with the distance measure without unsupervised learning.

	student 1	student 2	student 3	student 4	student 5	average
Bayes rule with	72	70	69	68	76	71.0
unsupervised learning	.80	.78	.77	.76	.84	.79
distance measure	55	51	39	54	58	54.4
	.61	.57	.43	.60	.64	.57

Discussions and Conclusion

This paper is the first attempt to use an unsupervised learning for speech recognition. Actually, this paper presents an one-sample speech recognition. An unsupervised learning needs a trmendous amount of unknown samples to learn the unkown parameters of syllables. From Theorem 2, the estimates using unsupervised learning will converge the true parameters and hence, our classifier can adapt itself to a better decision rule by making the use of unknown input syllables for unsupervised learning and will become more and more accurate after the system is put in use. Theoretically, from Theorem 2, our one-sample speech

recognition rate will approach to the rate given by supervised learning classifiers if a syllable does not have too many unknown parameters. In our experiments, we only have 9 samples for each syllable (a total of 90 unknown samples after 90 samples are mixed) for unsupervised learning of 96 parameters for each syllable and hence we only obtain 79% accuracy, 22% more than the rate without unsupervised learning.

Acknowledgments

The authors are grateful to the editor and the referees for their valuable suggestions to improve this paper.

References

[1]. K. Fukunaga, Introduction to Statistical Pattern Recognition, New York: Academic Press, 1990.

[2]. Sadaoki Furui, Digital Speech Processing, Synthesis and Recognition, Marcel Dekker, Inc., New York and Basel, 1989.

[3]. L. Rabiner and B. H. Juang, Fundamentals of Speech Recognition, Prentice Hall, PTR, Englewood Cliffs, New Jersey, 1993.

[4]. X. D. Huang, A. Acero, and H. W. Hon, Spoken Language Processing - A guide to theory, algorithm, and system development, Prentice Hall, PTR, Upper Saddle River, New Jersey, USA, 2001.

[5]. L. Deroye, L. Gyorfi, and G. Lugosi, A Probabilistic Theory of Pattern Recognition, Elsevier, New York, 1996.

[6]. R. L. Kasyap, C. C. Blayton, and K. S. Fu, Stochastic Approximation in Adaptation, Learning and Pattern Recognition Systems: Theory and Applications, J. M. Mendel and K. S. Fu. Eds., New York, Academic, 1970.

[7]. T. Y. Young and T. W. Calvert, Classification, Estimation and Pattern Recognition, New York: Elsevier, 1974.

[8]. A. G. Barto and P. Anandan, Pattern recognizing stochastic learning automata, IEEE Trans. Syst., Man, Cybern., Vol. SMC-15(May 1985) 360-375.

[9]. H. Robbins, An empirical Bayes approach to statistics, Proc. Third Berkeley Symp. Math. Statist. Prob., Vol. 1, University of California Press, (1956), 157-163.

[10]. Y. Lin, A note on margin-based loss function in classification, Statist. and Pro. Letters, 68(1)(2004), 73-81.

[11]. T.F. Li and S.C. Chang, Classification on defective items using unidentified samples, Pattern Recognition, 38(2005), 51-58.

[12]. R. J. Karunamuni, Empirical Bayes sequential estimation of the means, Sequential Anal., 11(1)(1992),

37-53.

[13]. A. Sarhan, Non-parametric empirical Bayes procedure, Reliability Engineering and System, 80(2)(2003), 115-122.

[14]. A. Sarhan, Empirical Bayes estimation in exponential reliability model, Applied Math. and Computation, 135(2)(2003), 319-332.

[15]. T. F. Li, Bayes empirical Bayes approach to estimation of the failure rate in exponential distribution, Commu.-Stat. Meth., 31(9)(2002), 1457-1465.

[16]. M. Ghosh, Empirical Bayes minimax estimators of matrix normal means, J. Multivariate Anal., 38(2)(1991), 306-318.

[17]. S. D. Oman, Minimax hierarchical empirical Bayes estimation in multivariate regression, J. Multivariate Anal., 80(2)(2002), 285-301.

[18]. R. Basu, J. K. Ghosh, and R. Mukerjee, Empirical Bayes prediction intervals in a normal regression model: higher order asymptotics, Statist. and Pro. Letters, 63(2)(2003), 197-203.

[19]. L. Wei and J. Chen, Empirical Bayes estimation and its superiority for two-way classification model, Statist. and Prob. Letters, 63(2)(2003), 165-175.

[20]. M. Pensky, Nonparametric empirical Bayes estimation of the matrix parameter of the Wishart distribution, J. Multivariate Anal., 69(2)(1999), 242-260.

[21]. M. Pensky, A general approach to nonparametric empirical Bayes estimation, Statistics, 29(1)(1997), 61-80.

[22]. S. Majumder, D. Gilliland, and J. Hannan, Bounds for robust maximum likelihood and posterior consistency in compound mixture state experiments, Statist. and Prob. Letters, 41(3)(1999), 215-227.

[23]. Y. Ma, Empirical Bayes estimation for truncation parameters, J. Statistical Planning and Inference, 84(1)(2000), 111-120.

[24]. T. F. Li and T. C. Yen, A Bayes Empirical Bayes decision rule for classification, Communications in Statistics-Theory and Methods, 34(2005), 1137-1149.

[25]. S. Kullback, Information Theory and Statistics, Gloucester, MA: Peter Smith, 1973.

[26]. S.S. Wilks, Mathematical Statistics, New York: John Wiley and Son, 1962.

[27]. S. B. Davis and P. Mermelstein, Comparison of parametric representation for monosyllabic word recognition in continously spoken sentences, IEEE. Trans. Acoust., Speech, Signal Processing, 28(4)(1980), 357-366.

[28]. T. F. Li, A note on Mel frequency cepstra in speech recognition, Department of Applied Mathematics, Chung Hsing University, Taichung, Taiwan, (2006).

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

[30]. J. Makhoul, Linear prediction: a tutorial review, Proc. IEEE, 63(4)(1975), 561-580.

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

[32]. T. F. Li, Speech recognition of mandarin monosyllables, Pattern Recognition, 36(2003), 2712-2721.

[33]. H. Robbins and S. Monro, A stochastic approximation method, Ann. Math. Statist., 22(1951), 400-407.

[34]. A. Abert and L. Gardner, Stochastic Approximation and Nonlinear Regression, Cambridge, MA, M.I.T., 1967.

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

[36]. C. F. J. Wu, On the convergence properties of the EM algorithm, Ann. Stat., 11(1983), 95-103.