Speech recognition of mandarin syllables using both linear predict coding cepstra and Mel frequency cepstra

黎自奮　Tze Fen Li
明道大學管理研究所
Institute of Management
Ming Dao University
tfli@mdu.edu.tw

張水清　Shui-Ching Chang
僑光技術學院資訊管理系
Department of Information Management
The Overseas Chinese Institute of Technology
monet@ocit.edu.tw

Abstract

This paper is to compare two most common features representing a speech word for speech recognition on the basis of accuracy, computation time, complexity and cost. The two features to represent a speech word are the linear predict coding cepstra (LPCC) and the Mel-frequency cepstrum coefficient (MFCC). The MFCC was shown to be more accurate than the LPCC in speech recognition using the dynamic time warping method. In this paper, the LPCC gives a recognition rate about 10% higher than the MFCC using the Bayes decision rule for classification and needs much less computational time to be extracted from speech signal waveform, i.e., the MFCC needs computational time 5.5 time as much as the LPCC does. The algorithm to compute a LPCC from a speech signal much simpler than a MFCC, which has many parameters to be adjusted to smooth the spectrum, performing a processing that is similar to be adjusted to smooth the spectrum, performing a processing that is similar to that executed by the human ear, but the LPCC is easily obtained by the least squares method using a set of recursive formula.

Key words: Bayes decision rule, linear predict coding, Mel-frequency cepstrum coefficient, signal processing, speech recognition.

1. Introduction

A speech recognition system basically contains extraction of features and classification of an utterance of an acoustical word. The measurements made on the speech waveform include energy, zero crossings, extrema count, formants, LPC cepstrum (LPCC) [1-4] and the Mel frequency cepstrum coefficient (MFCC) [5-8]. 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. The filters with Mel scales spaced linearly at low frequencies and logarithmically at high frequencies are used to capture phonetically the characteristics of speech [8]. For recognition, Davis and Mermelstein [5] used the dynamic time warping algorithm to show that the performance of the MFCC was better than the LPCC.

In this paper, we use a simple technique [9] for speech data compression of the sequence of MFCC vectors and the sequence of LPCC vectors to obtain a matrix of feature values respectively. For speech recognition, we simply use a simplified Bayes decision rule with weighted variance, where each step is a simple calculation and which has the minimum probability of misclassification. In our study, there are two speech recognition experiments. In the first experiment, since both LPCC and MFCC are said to be robust and reliable to noise and estimation errors, our speech experiment is implemented in a noisy environment to test which feature is better on speech recognition. Pick up 9 female and 10 male students and each pronounces 10 digits once using a common (not high-quality) microphone. Some students pronounce mandarin syllables not very clearly, since we have several types of accents to pronounce the same mandarin syllables. In the second experiment, there are 87 students to pronounce the mandarin syllables in a quiet classroom, which are most commonly used in usual conversations. Our speech experiment is done like natural talking. Hence our speech system can be commonly used for all peoples and in all environments. The recognition rate using LPCC is significantly better than the rate using MFCC and the LPCC needs much less computational time to be extracted from speech signal waveform.

2. Bayes Decision Rules

Let $X = (X_1, ..., X_k)$ be the input feature vector of a speech data, which belongs to one of m categories (syllables) c_i, $i = 1, ..., m$. Consider the decision problem consisting of determining whether X belongs to c_i. Let $f(x \mid c_i)$ be the conditional density function of X given category c_i. Let θ_i be the prior probability of c_i such that $\Sigma_{i=1}^{m} \theta_i = 1$, 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 $L(c_i, d(x)), i = 1, ..., m,$ 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 $\tau = (\theta_1, ..., \theta_{\mathrm{m}})$ and let $R(\tau, d)$ denote the risk function (the probability of misclassification) of d. Let $\Gamma_i, i = 1, ..., m$, be m regions separated by d in the k-dimensional domain of X, i.e., d decides c_i when $X \in \Gamma_i$. Then

$$\begin{aligned} R(\tau, d) &= \sum_{i=1}^{m} \theta_i \int L(c_i, d(x)) f(x \mid c_i) dx \\ &= \sum_{i=1}^{m} \theta_i \int_{\Gamma_i^c} f(x \mid c_i) dx \end{aligned} \tag{2.1}$$

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) [10]. We state the Bayes decision rule in the following theorem.

Theorem 2.1. [10] The Bayes decision rule with respect to τ is defined by

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

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

Note that if $\theta_i = 1/m$, $i = 1, ..., m$, the Bayes decision rule (2.3) become a ML classifier.

3. Feature Extraction

3.1 Preprocessing Speech Signal

Since our speech recognition experiment is implemented in a noisy environment, 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 samples to detect the real speech signal, i.e., the samples with small variance does not contain speech signal. Another way is to compute the sum of the absolute values of difference of two consecutive samples in a fixed number of sequential speech samples, i.e., the speech data with small absolute value do not contain real speech signal. In our speech recognition experiment, the latter provides slightly faster and more accurate speech recognition.

3.2 Mel-Frequency Cepstrum Coefficient (MFCC)

The MFCC is a representation defined as the real cepstrum of a windowed short-time signal derived from the fast Fourier transform of the speech signal. In the MFCC, a nonlinear frequency scale is used, which approximates the behavior of the auditory system. The discrete cosine transform of the real logarithm of the short-time energy spectrum expressed on this nonlinear frequency scale is called the MFCC. Davis and Mermelstein [5] showed the MFCC representation to be beneficial for speech recognition. We detail the MFCC as follows [8]:

Let $s[n]$ denote the N samples of a speech waveform. The discrete Fourier transform (DFT) $X[k]$ of the speech signal is defined by

$$X[k] = \sum_{n=0}^{N-1} s[n] e^{-j2\pi nk/N}, \qquad 0 \le k < N \tag{3.1}$$

We define a filterbank with M filters $(m = 1, ..., M)$, where filter m is a triangular filter given

$$\begin{aligned}
H[m,k] &= 0 && if \quad k < f[m-1] \\
H[m,k] &= (k - f[m-1])/(f[m] - f[m-1]) && if \quad f[m-1] \le k \le f[m] \\
H[m,k] &= (f[m+1] - k)/(f[m+1] - f[m]) && if \quad f[m] \le k \le f[m+1] \\
H[m,k] &= 0 && if \quad k > f[m+1]
\end{aligned} \tag{3.2}$$

which satisfies $\Sigma_{m=1}^{M} H[m,k] = 1,\ k = 0, 1, ..., N-1$.

Let f_l and f_h be the lowest and highest frequencies of the filterbank in H_z and let F_s be the sampling frequency in H_z. The boundary points $f[m]$ are uniformly spaced in the mel-scale:

$$f[m] = (\frac{N}{F_s}) B^{-1}(B(f_l) + m\frac{B(f_h) - B(f_l)}{M+1}) \tag{3.3}$$

where $B(f) = 1125 ln(1 + f/700)$ and $B^{-1}(b) = 700(e^{(b/1125)} - 1)$. The log-energy is computed by

$$S[m] = ln\{\sum_{k=0}^{N-1} |X[k]|^2 H[m,k]\}, \qquad 0 < m \le M. \tag{3.4}$$

The MFCC is then the discrete cosine transform of the M filters outputs:

$$c(n) = \sum_{m=0}^{M-1} S[m] cos(\pi n(m - 0.5)/M) \qquad 0 \le n < M \tag{3.5}$$

For speech recognition, normally, the number M of filters is from 10 to 20 and the MFCC produced from the first few filters are the most effective in recognition. In our experiment, we use $M = 12$

3.3 Linear Predict Coding Cepstrum (LPCC)

The MFCC was proved to be better than the LPC cepstrum for recognition by using the dynamic time warping (DTW) method [5], but the computational complexity for the MFCC is much heavier than that of the LPC cepstrum. The LPC coefficients can be easily obtained by Durbin's recursive procedure [11-13] and their cepstra can be quickly

found by another recursive equations [11-13] without computing the discrete Fourier transform (DFT) and the inverse DFT, which are computationally complex and time consuming.

The LPC method can also provide a robust, reliable and accurate method for estimating the parameters that characterize the linear and time-varying system [3, 11-13]. The following is a brief discussion of LPC method. It is assumed [13] that the sampled speech waveform $\hat{s}(n)$ can be linearly predicted from the past p samples of $s(n)$. Let

$$\hat{s}(n) = \sum_{k=1}^{p} a_k s(n-k) \tag{3.6}$$

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

$$E = \sum_{n=0}^{N-1} [s(n) - \hat{s}(n)]^2. \tag{3.7}$$

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 [3, 11-13]. Here in our experiments, $p = 12$, because the cepstra in the last few elements are almost zeros.

Both LPCC and MFCC are the method to compress or simplify the huge speech data $s(n)$ of a syllable into a simple data without loss of speech information. The LPCC is more or less like the sufficient statistics of a random samples in statistics [14]. The LPC coefficients a_k, $k = 1, ..., p$, are actually the least squares estimators of the regression coefficients, i.e., the minimum variance linear estimators a_k of the regression coefficients [14]. The huge data of a frame are well-represented by the LPC coefficients unless LPC coefficients are too small, i.e., the estimates a_k of the regression coefficients are not significant as compared with noise. On the other hand, to produce a MFCC, one has to obtain the DFT of a frame of the huge data and after the Mel filter banks smooth the spectrum, performs the inverse DFT on the logarithm of the magnitude of filter bank output. It seems to us that the formula in (3.1)-(3.5) to produce a MFCC are a little arbitrarily or artificially or experimentally adjusted for human ears. There is no theoretical theory to support the MFCC to well represent a syllable without loss of information. Hence in this paper, we create a huge database from common mandarin sentences to obtain the recognition rates using the LPCC and MFCC respectively.

3.4 Feature Extraction [9]

Our method to extract the feature from LPCC (MFCC) is quite simple. Let $x(k) = (x(k)_1, ..., x(k)_p)$, $k = 1, ..., n$, be the LPCC (MFCC) vector of size $p = 12$ for the k-th frame of a speech waveform, where n is the length of the LPCC (MFCC) sequence and p is the number of LPC coefficients in each frame. Normally, if a speaker does not intentionally elongate pronunciation, a mandarin syllable has 30-70 vectors of LPCC (MFCC).

Since an utterance of a syllable is composed of two basic parts: stable part and feature part. In the feature parts, the vectors have a dramatic change between two consecutive vectors, representing the unique characteristics of the syllable utterance and in the stable parts, the vectors stay about the same. Even if the same speaker utters the same syllable, the duration of stable parts of the sequence of LPCC (MFCC) vectors changes every time with nonlinear expansion and contraction and hence the duration of the portion of feature vectors and duration of stable parts are different every time. Therefore, the duration of stable parts is contracted such that the compressed speech waveform has about the same length of the sequence of the vectors. Li [9] proposed several simple compression techniques to contract the stable parts of the sequence of vectors. We state a simple one with good recognition rate as follows:

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

$$D(k) = \sum_{i=1}^{p} | x(k)_i - x(k-1)_i |, \quad k = 2, ..., n. \tag{3.8}$$

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 (MFCC) vector is removed if its absolute difference $D(k)$ from the previous vector $x(k-1)$ is too small. In this study, a squared difference criterion is also used to remove the stable and flat portion of the sequence. The criterion is expressed as follows:

$$D(k) = \sum_{i=1}^{p} [x(k)_i - x(k-1)_i]^2, \quad k = 2, ..., n. \tag{3.9}$$

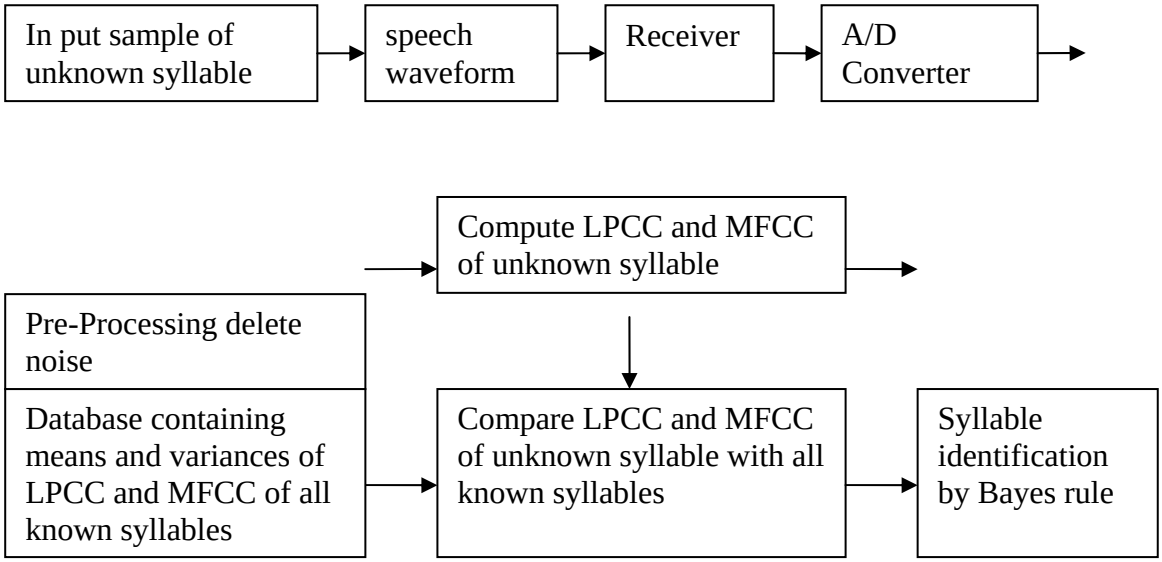
Let $x'(k),\ k = 1, ..., m(< n)$, be the new sequence of LPCC (MFCC) vectors after deletion. We think that the first part (about first 40 vectors) of an utterance of a mandarin syllable contains main features which can most represent the syllable and the rest of the sequence contains the "tail" sound, which has a variable length. If a speaker intentionally elongates pronunciation of a syllable, the speaker only increases the tail part of the sequence. The length of the feature part stays about the same. As in [9], we partition the feature part (first 40 vectors of the new sequence) into 8 equal segments and partition the tail part with variable length into two equal segments. If the length of the new sequence of vectors representing a syllable is less than 40, we neglect the tail sound and partition the new sequence into 10 equal segments. The average value of the LPCC (MFCC) in each segment is used as a feature value. Note that the average values of samples tend to have a normal distribution. This compression produces 12×10 feature values for each mandarin syllable.

4. Experimental Results

There are two speech recognitions implemented in our study. One is the digit recognition in a noisy environment and the other is the speech recognition on the mandarin monosyllables which are most commonly used in general conversations.

The following is a flow chart to show the speech recognition on a syllable.

Figure 1. Flowchart of a syllable recognition



4.1 The Digit Recognition

The digit recognition is implemented in a noisy environment, a classroom with windows open, which has noise from students inside classroom and from students and autos on the street outside classroom. The database of 10 mandarin digits is created by 19 persons (9 female and 10 male students) who pronounce 10 digits (0-9) once. The speech signal of a mandarin monosyllable is sampled at $10\,kHz$. A Hamming window with a width of 25.6 *ms* is applied every 12.8 *ms* for our study. A 256 point Hamming window is used to select the data points to be analyzed.

In our experiments, we use this database to produce the LPCC (MFCC) and obtain a 12×10 matrix for each digit sample. On the average, the time to produce a MFCC using DFT and formula in Section 3.2 is 5.5 times as much as to produce a LPCC. Among 19 samples (pronounced by 19 students) of each mandarin digit, pick up one sample (from one student) for recognition and the rest of 18 samples (from the other 18 students) of the digit is used for training, i.e., the rest of 18 samples of this digit is used to estimate the parameters which represent the digit. Hence each of 19 students has to be tested, i.e., there are 19 testing samples for each digit.

Since the average value of samples tends to be normally distributed. In order to reduce computation for classification, we assume that all elements in the 12×10 matrix of feature values are stochastically independent. It was proved [15] that using weighted variance in the Bayes decision rule for each class may increase the recognition rate. Hence, the conditional normal density given syllable c_i with weighted variance c can be represented as

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

where $i = 1, ..., m = 10$, k=12×10 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

$$l(c_i) = \sum_{l=1}^{k} \ln(c\sigma_{il}) + \frac{1}{2} \sum_{l=1}^{k} (\frac{x_l - \mu_{il}}{c\sigma_{il}})^2, \ i = 1, ..., m. \quad (4.2)$$

The Bayes decision rule (4.2) decides a syllable c_i with the least $l(c_i)$ to which the feature matrix $x = (x_1, ..., x_k)$ belongs. For the Bayes decision rule, 18 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 0.8 to 1.3.

Note that in the Bayes decision rule, a matrix of feature values representing the testing digit pronounced by one student is compared with 10 matrices of means representing 10 digits' parameters. The means are computed from the feature values pronounced by the rest of 18 students. Hence the feature values of the digits pronounced by the student to be tested are independent of the feature values of the digits pronounced by the other 18 students and in the training data to produce 10 matrices of means (each matrix represents one digit's parameters μ_{il}, $l = 1, ..., k = 12 \times 10$), the feature values of 10 digits between any two persons of the other 18 students are mutually independent. Therefore, the Bayes rule uses simple normal distributions. Table 4.1 shows that the number of correct digits of 190 testing samples and the recognition rates are obtained using LPCC and MFCC features with absolute difference and squared difference criteria.

Table 4.1 Correct digit recognition rates

	absolute difference criterion		squared difference criterion	
	LPCC	MFCC	LPCC	MFCC
total testing samples = 190				
19 students	181	178	182	179
	(95.3%)	(93.7%)	(95.8%)	(94.2%)
total testing samples = 100				
10 students (pronounce most clearly)	100	96	100	96
	(100%)	(96%)	(100%)	(96%)

Table 4.1 also shows the misclassified digits pronounced by the 10 students who

pronounce most clearly and distinctly. Since all mandarin syllables are monosyllables, the speech wave for each monosyllable is short. If the monosyllables are not pronounced clearly, it is difficult to recognize by the human ear. Hence, to test the recognition ability of the Bayes decision rule, which should not be damaged by the ambiguous pronunciation, we select 10 students (4 female and 6 male) among 19 students, who pronounce most clearly and distinctly. As in the first speech experiment, 10 digits pronounced by each student are used for testing and 90 samples (9 samples for each digit) from the other 9 students are used for training the means and the variances of each digit. There are 10 testing samples for each digit. From the classification in digits, the LPCC for speech recognition is lightly better than the MFCC for two criteria (absolute and squared differences). After compression of a sequence of LPCC and MFCC vectors, the two compression criteria give about the same recognition rates, but the squared difference criterion takes less time to compute. The same speech recognition experiment was implemented in a quiet environment [15] and gave the correct digit recognition rate 98.6%. For the robustness to the noise, our results show that the LPCC gives a recognition rate no less than the MFCC. This contradicts to the results obtained by Davis and Mermelstein [5] in a quiet environment.

4.2 The Speech Recognition

In this speech recognition experiment, 87 students participate in the experiment. Each pronounces loudly and clearly several sentences of mandarin syllables, which are commonly used in the usual conversation in our life. We cut these sentences into single words (syllables). We select the syllables which have at least 9 samples, i.e., each syllable as a candidate for speech recognition should appear in the sentences at least 9 times. Hence there are 102 different syllables to be classified. The 102 syllables appear in the sentences from 9 to 45 times. There are totally 1644 samples for 102 syllables to be tested. This experiment is designed as in the digit recognition in the first experiment. Each of 1644 samples is tested and the other 1643 samples are used for training, i.e., the syllables with 9 samples have 8 samples for training and the syllable with 45 samples has 44 samples for training. To compress the speech wave of a syllable into a 12×10 matrix of feature values, we only use the absolute difference criterion, since in the first experiment on digit recognition, there are no difference on recognition rates between the absolute difference and the squared difference criteria. The simplified Bayes decision rule with $m = 102$ and the weighted factor $c = 1.2$ in (4.1) and (4.2) is used to classify 102 different mandarin syllables. Table 4.2 shows the results. The table shows that the LPCC feature has the recognition rate 0.9057 better than the rate 0.8102 obtained by the MFCC feature. The total time needed to compute the MFCC of 1644 samples based on the formula in Section 3 is 5.5 times as much as that needed to compute the LPCC of the same 1644 samples.

Both recognition rates are not high enough, since some syllables having only 8 samples for training have poor rates. Hence we increase the minimum number of samples for training to 10, i.e., we select the syllables from the sentences, which should have at least 11 samples (to appear at least 11 times in the sentences) as candidates for speech recognition. This restriction results in 91 different mandarin syllables with a total 1523 samples to be tested in speech recognition, i.e., each of 1523 samples is used for testing and the remain of 1522 samples are used to train 91 different syllables. The recognition rates are increased to 0.9140 for the LPCC and 0.8188 for the MFCC. The recognition results for 91 syllables are shown in Table 4.2.

The above recognition rates all show that the LPCC features a little higher than the MFCC. Hence we make a statistical hypothesis testing in our study. We adopt two nonparametric methods (McNemar test and Cochran Q-test) [16] to test if the LPCC is better than the MFCC. The McNemar test is to compare two rates provided by the LPCC and MFCC individually. We obtain the approximate standard normal *z*-value 7.4593 for 102 syllables and 7.3899 for 91 syllables. Both are strongly significant at the level $\alpha = 0.0001$.

The Cochran Q-test is to compare two features (MFCC and LPCC) if they are equally effective in classification. We obtain the approximate Chi-square (df =1) Q value 55.6411 for 102 syllables and 54.6104 for 91 syllables, which are both strongly significant at the level $\alpha = 0.0001$. Both tests show in Table 4.3. Obviously, the two nonparametric tests make a decision to favor the LPCC.

Table 4.2 Correct syllable recognition rates pronounced by 87 students

features	LPCC	MFCC
total samples=1644 for 102 different syllables		
correct samples	1489	1332
correct rates	90.57%	81.02%
total samples=1523 for 91 different syllables		
correct samples	1392	1247
correct rates	91.40%	81.88%

Tables 4.3 Statistical testing hypotheses

Mc Nemar test :

H_0 : two recognition rates are equal

$z - \text{value} = 7.4593$ for 102 syllables. $= 7.3899$ for 91 syllables

$p - values < 0.0001$ for both 102 syllables and 91 syllables

decision : reject H_0 for both tests at the level $\alpha = 0.0001$

Cochran's Q – test :

H_0 : LPCC and MFCC are equally effective in classifiction

Q – value $= 55.64$ for 102 syllables. $= 54.61$ for 91 syllables

$p - values < 0.0001$ for both 102 syllables and 91 syllables

decision : reject H_0 for both tests at the level $\alpha = 0.0001$

Discussions and Conclusion

In this paper, we have used two speech recognition experiments to test if the LPCC feature has a higher ability in classification of the mandarin monosyllables than the MFCC. The speech waveform of a mandarin syllable is extracted into a sequence of LPCC (MFCC) vectors and the sequence of vectors is then compressed into a matrix of LPCC (MFCC) values, which tend to have a normal distribution. Using the Bayes decision rule, we have found that in the first digit experiment, the mandarin digit recognition rate using LPCC feature is no less than the rate using the MFCC feature. In the second speech recognition experiment, we build a large amount of mandarin syllables, which are the most commonly used in usual conversations. From the nonparametric statistical analysis, the LPCC has a significant higher ability in classification than the MFCC. Furthermore, the LPCC feature needs much less computational time to be extracted from speech signal waveform than the MFCC.

References

[1]. S. S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra", IEEE Trans. Acoust., Speech, Signal Processing, ASSP-22(2), 135-141, 1974.

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

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

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

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

[6]. W. Q. Zhang, L. He, Y. L. Chow, R. Z. Yang, and Y. P. Su, "The study on distributed speech recognition system", IEEE 2000 ICASSP, 1431-1434.

[7]. T. Fukuda, M. Takigawa, and T. Nitta, "Peripheral feature for HMM-based speech recognition", IEEE 2001 ICASSP, 129-132.

[8]. 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.

[9]. T. F. Li, "Speech recognition of mandarin monosyllables", Pattern Recognition, 36, 2713-2721, 2003.
[10]. K. Fukunage, Introduction to Statistical Pattern Recognition, New York: Academic Press, 1990.
[11]. Sadaoki Furui, Digital Speech Processing, Synthesis and Recognition, Marcel Dekker, Inc., New York and Basel, 1989.
[12]. J. Durbin, The fitting of time-series models, Rev. Inst. Int. Statist., 28(3) (1960) 233-243.
[13]. L. Rabiner and B. H. Juang, Fundamentals of Speech Recognition, Prentice Hall PTR, Englewood Cliffs, New Jersey, 1993.
[14]. S. S. Wilks, Mathematical Statistics, New York: J. Wiley and Sons, 1962.
[15]. T. F. Li and T. F. Lin, On probability distribution of feature values for speech digit recognition, Technique Report, Department of Applied Mathematics, Feng Chia University, Taichung, Taiwan, 1994.
[16]. W. W. Daniel, Applied Nonparametric Statistics, Georgia State University, 1979.