Speech Recognition using MFCC

Chadawan Ittichaichareon, Siwat Suksri and Thaweesak Yingthawornsuk

***Abstract*—** This paper describes an approach of speech recognition by using the Mel-Scale Frequency Cepstral Coefficients (MFCC) extracted from speech signal of spoken words. Principal Component Analysis is employed as the supplement in feature dimensional reduction state, prior to training and testing speech samples via Maximum Likelihood Classifier (ML) and Support Vector Machine (SVM). Based on experimental database of total 40 times of speaking words collected under acoustically controlled room, the sixteen-ordered MFCC extracts have shown the improvement in recognition rates significantly when training the SVM with more MFCC samples by randomly selected from database, compared with the ML.

***Keywords*—**Speech Signal, MFCC, SVM, ML

I. Introduction

SPEECH recognition is the process of automatically recognizing the spoken words of person based on information in speech signal. Recognition technique makes it possible to the speaker's voice to be used in verifying their identity and control access to services such as voice dialing, banking by telephone, telephone shopping, database access services, information service, voice mail, security control for the confidential information areas, and remote access to computers. The acoustical parameters of spoken signal used in recognition tasks have been popularly studied and investigated, and being able to be categorized into two types of processing domain: First group is spectral based parameters and another is dynamic time series. The most popular spectral based parameter used in recognition approach is the Mel Frequency Cepstral Coefficients called MFCC [2,3]. Due to its advantage of less complexity in implementation of feature extraction algorithm, only sixteen coefficients of MFCC corresponding to the Mel scale frequencies of speech Cepstrum are extracted from spoken word samples in database. All extracted MFCC samples are then statistically analyzed for principal components, at least two dimensions minimally required in further recognition performance evaluation.

The following sections give details on database, processing methods of; voice-segment detection, MFCC feature extraction, principal component analysis and performance evaluation, finally results and discussion.

Chadawan Ittichaichareon is with Department of Media Technology, King Mongkut's University of Technology Thonburi - BangKhuntian Campus.

Siwat Suksir is with Department of Media Technology, King Mongkut's University of Technology Thonburi, KMUTT – BangKhuntian Campus.

Thaweesak Yingthawornsuk is with Department of Media Technology, KMUTT – BangKhuntian Campus.

II. Database

Database consists of two groups of speech samples recorded in an environmentally controlled recording room to have all possibly less acoustical interferes to the quality of sound sample during the recording time. The first group comprises of thirty spoken sound samples of a word "MFCC" and another is a group of thirty sound samples of a word "PCA". All sound signals are recorded under most similar setting condition such as the same length of recording time, and the level of sound amplitude. The sampling frequency is originally set at 44.1 KHz for making all sound records in order to preserve acoustical quality of sound signals. Prior to detect for voiced segments in speech sounds, signals are digitized offline via a 16-bit A/D converter. Thereafter, signals are monitored and edited for all possible sound artifacts that could affect in further processing phases. Furthermore, the longer silences than a half second are manually removed as well in the Goldwave sound editor program.

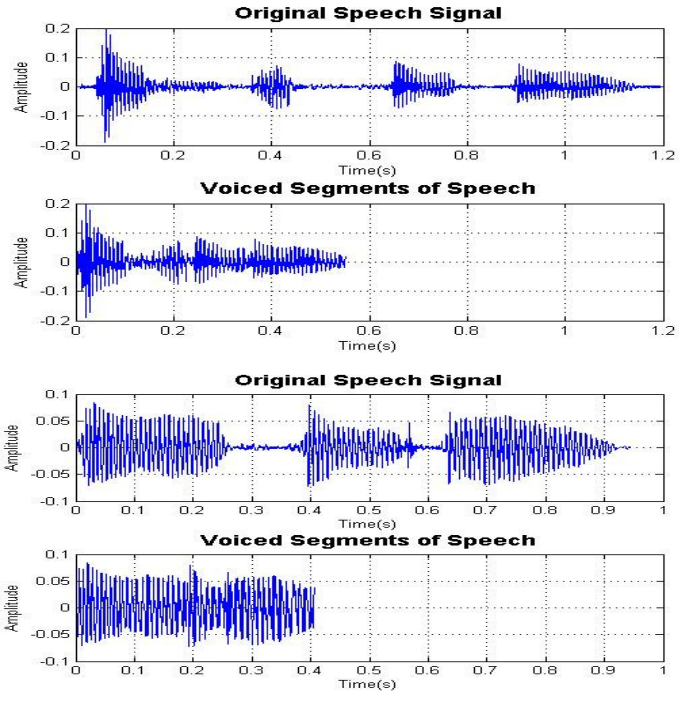

Fig.1 Speech signals of spoken words "MFCC" in upper plot, "PCA" in lower plot, and the detected segments of voiced speech in following plots.

III. Methodology

A. Voiced/Unvoiced Detection

Pre-processed signals are estimated for their energy and then weighted using the Dyadic Wavelet Transform (DTW) on each 256 samples/frame. The lowest energy level is at scale $\delta_1 = 2^1$and the highest energy level is $\delta_5 = 2^5$. Segments of sound signal with its largest energy level estimated at scale $\delta_1 = 2^1$ are therefore identified as unvoiced segment, otherwise found to be voiced segments. The following equation is the energy threshold defining as unvoiced segment;

$$uv = (n|\delta_i = 2^1); \qquad n = 1, \dots, N \tag{1}$$

At which uv is the unvoiced segment of the n segment with energy at scale δ_1 maximized.

B. Acoustic Feature Extraction

Only voiced segments of speech signal are processed for MFCC extraction. The procedure to determine MFCC [1] is described as follows:

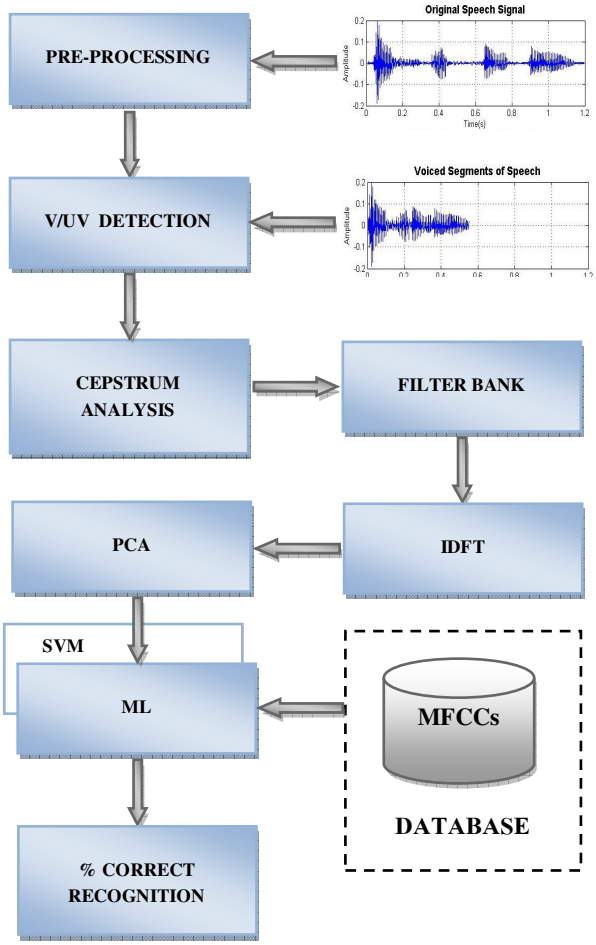

Fig.2 Workflow for the MFCC based speech classification.

- Segmenting all concatenated voiced speech signal into 25.6ms-length frames.
- Estimating the logarithm of the magnitude of the discrete Fourier Transform (DFT) for all signal frames.
- Filtering out the center frequencies of the sixteen triangle band-pass filters corresponding to the mel frequency scale of individual segments.
- Estimating inversely the IDFT to get all 16-order MFCC coefficients.
- Analyzing all extracted MFCC dataset for two dimension principal components and then used as an input vector for testing and training with ML and SVM. All processes are implemented in Maltab program.

The mel-scale used in this work is to map between linear frequency scale of speech signal to logarithmic scale for frequencies higher than 1 kHz. This makes the spectral frequency characteristics of signal closely corresponding to the human auditory perception [5]. The mel-scale frequency mapping is formulated as:

$$f_{mel} = 2595 * \mathrm{LOG}_{10}\left[1 + \frac{f_{lin}}{700}\right] \tag{2}$$

in which f_{mel} is the perceived frequency and f_{lin} is the real linear frequency in speech signal.

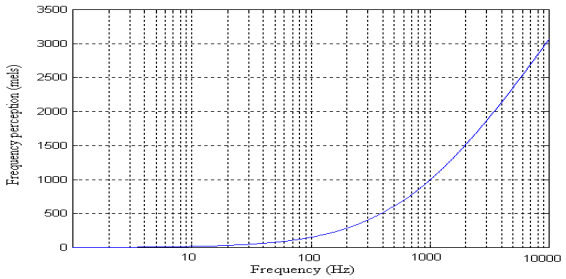

Fig.3 Logarithmic plot of the mapping frequencies between 0 and10 kHz.

In filtering phase, a series of the 16 triangular band-pass filters, $N_f = 16$ is used for a filter bank whose center frequencies and bandwidths are selected according to the mel-scale. They span the entire signal bandwidth for $\left[0 - \frac{f_s}{2}\right]$. The center frequency of individual filter is defined;

$$F_{C,i} = \mathrm{k}_i \frac{f_s}{N'}; \qquad i = 1, 2, 3, \dots, N_f \tag{3}$$

And its bandwidth is consequently computed by

$$B_i = F_{C,i+1} - F_{C,i-1}; \qquad i = 1, 2, 3, \dots, N_f \tag{4}$$

Here N' is the fft bin equal to 256, k_i is the DFT index of the center frequency of filter i, B_i and $F_{C,i+1}$ are the bandwidth and the center frequency of filter i, respectively. It is also important to see that $F_{C,0} = 0$ and $F_{C,N_f} < \frac{f_s}{2}$. Once the center

frequencies and bandwidth of the filters are obtained, the log-energy output of each filter *i* is computed and encoded to the MFCC by performing a Discrete Cosine Transform (DCT) defined as follows:

$$C_n = \frac{2}{N'}\sum_{i=1}^{N_f} X_k \cos\left(k_i \frac{2\pi}{N'} n\right);\ 1 \le n \le p \qquad (5)$$

Due to the computational simplicity, equation (5) without the superfluous factor $\frac{2}{N}$ is employed in our algorithm for the computation of mel-cepstral filter bank coefficients.

C. Principal Component Analysis

In this paper, we have applied the PCA technique [4] to MFCC features to extract the most significant components of feature. The main concept of PCA is to project the original feature vector onto principal component axis. These axes are orthogonal and correspond to the directions of greatest variance in the original corresponding to the directions of greatest variance in the original feature space. Projecting input vectors onto the principal subspace helps reducing the redundancy in original feature space and dimension as well. The analyzed MFCC features are projected onto a two dimensional space which is adequate for data training and testing in next classification state.

D. Feature Classification

Two classifiers, Maximum Likelihood (ML) and Support Vector Machine (SVM) are selected to train and test on two-dimensional MFCC dataset and then compared to each other for performances on correct classification. Firstly, samples are randomly selected for 20% of sample dataset, and then used to train a classifier, and another 80% of the rest of dataset used later for testing the same classifier. Several trials on random selection of samples from our dataset with 20%, 35%, and 50% for training and the rest for testing are further proceeded to find more on the performance of classifier which may be affected by sample sizes. In case of ML classification the Bay's Decision Rule has been approximated from our samples by following the denoted equation [4];

$$P(mfcc_i|\omega_1)\mathrm{P}(\omega_1) > P(mfcc_i|\omega_2)\mathrm{P}(\omega_2) \qquad (6)$$

This means $mfcc_i$ is in classω_1, otherwise $mfcc_i$ is identified as class ω_2. Where $\mathrm{P}(\omega_i)$ is known as the prior probability that it would be in class *i* and $P(mfcc_i|\omega_i)$ is known as the state conditional probability for class *i*. Furthermore, the inequality can be re-arranged to obtain another decision rule;

$$L_R(mfcc_i) = \frac{P(mfcc_i|\omega_1)}{P(mfcc_i|\omega_2)} > P(\omega_2)\frac{P(\omega_2)}{P(\omega_1)}$$
$$\tau_c = P(\omega_2)\frac{P(\omega_2)}{P(\omega_1)} \qquad (7)$$

Then x is in class ω_1. The ratio on the left of equation 7 is called the likelihood ratio and quantity on the right is the threshold. If $L_R > \tau_c$ then we decide that the case belongs to class ω_1. If $L_R < \tau_c$, then the threshold is one ($\tau_{c\ =\ 1}$). Thus, when $L_R > 1$, we assign the observation or pattern to ω_1, and if $L_R < 1$,then we classify the observation as belonging to ω_2. We can also adjust this threshold to obtain a desired probability of false alarm.

Another SVM classifier [6, 7] is also used for performance validation. This type of classifier achieves relatively robust pattern recognition performance using well established concepts in optimization theory. SVM separates an input $\boldsymbol{x} \in \boldsymbol{R}^d$ into two classes. A decision function of SVM separates two classes by $f(\boldsymbol{x}) > 0$ or $f(\boldsymbol{x}) < 0$. The training data which is used in training phase is $\{\boldsymbol{x}_i, y_i\}$, for $i = 1,\ldots,l$ where $\boldsymbol{x}_i \in \boldsymbol{R}^d$ is the input pattern for the *i*th sample and $y_i \in \{-1, +1\}$ is the class label. Support Vector Classifier maps $\boldsymbol{x}_i$ into some new space of higher dimensionality which depends on a nonlinear function $\phi(\boldsymbol{x})$ and looks for a hyperplane in that new space. The separating hyperplane is optimized by maximization of the margin. Therefore, SVM can be solved as the following quadratic programming problem,

$$\max_{\alpha_i}\{\sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum\sum\alpha_i\alpha_j\, y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j)\} \qquad (8)$$

Subject to $0 \le \alpha_i \le C$ and $\sum_{i=1}^{l}\alpha_i \boldsymbol{y}_i = 0$

where *C* is a parameter to be chosen by user, a larger *C* corresponding to assigning a higher penalty to errors and $\alpha \ge 0$ are Lagrange multipliers. When the optimization problem has solved, system provides many $\alpha_i > 0$ which are the required support vector. Note that the Kernel function $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi^T(\boldsymbol{x}_i)\phi(\boldsymbol{x}_j)$ where $\phi(\cdot)$ is a nonlinear operator mapping input vector $\boldsymbol{x} \in \boldsymbol{R}^d$ to a higher dimensional space. In this work, we choose the polynomial kernel $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = <\boldsymbol{x}_i, \boldsymbol{x}_j>^d$ as the kernel function, where $d \in N$. In addition, other kernels can also be applied. Classification consists of two steps: training and testing. In the training phase, SVM receives some feature patterns as input. These patterns are the extracted speech features represented by N feature parameters that can be seen as points in N-dimensional space. In this study sixteen MFCC features are formed for input feature matrix which is only two dimensional. Then the classifying machine becomes able to find the labels of new vectors by comparing them with those used in the training phase. For every training classifier, the cross validations have been completed for approximately 100 times. The tendencies of performance evaluated from both SVM and ML classifiers are provided in next section.

IV. Experimental Results

Results of the sixteen-order MFCC extracted from database of spoken words are shown in Figure (4). The significant difference in quantity can be clearly identified between sample classes of different words. The comparative performances obtained from several trials on sample selections in training and testing states are graphically plotted in box-and-whisker diagrams for convenient examination on statistical descriptive. Training results of classification shown in Figure (5), in case of SVM classifier, provide much compact distribution with consistent training scores as compared to ML classification. In addition more decreasing change in maximum and minimum adjacent values depicted as top and bottom bars of individual box plots can be notified as well for SVM. These suggest that the SVM classifier seems to give more consistent and reliable performance on training sample state than ML does. The testing results shown in Figure (6) consistently reveal the similar tendency of improving recognition on larger size of samples used in testing state. The distributions of SVM scores seem more tense and consistent than those of ML for all percentages of dataset tested for recognitions.

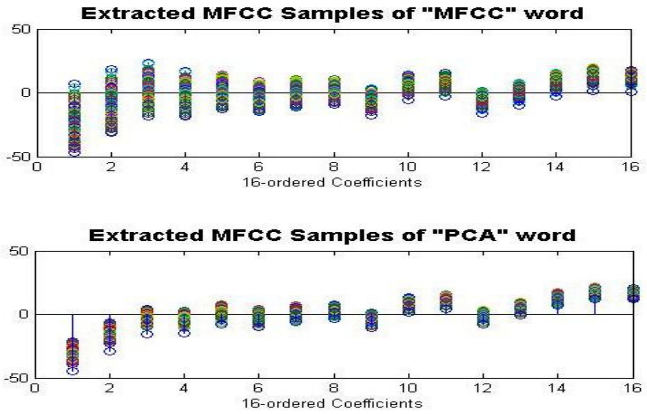

Fig. 4 The 16-ordered MFCC extracted from voiced speech samples

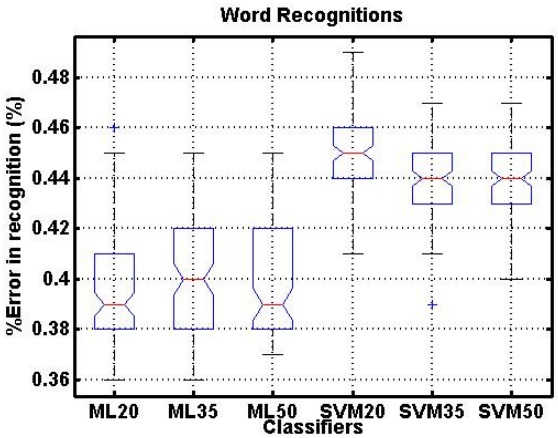

Fig. 5 Comparative training results

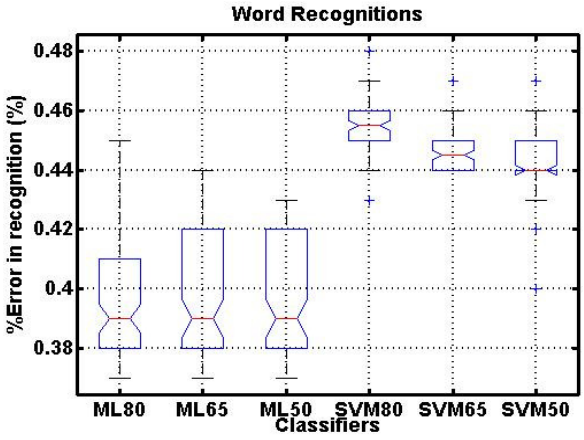

Fig. 6 Comparative testing results

V. Conclusion

This paper addressed the principle of speech MFCC extraction for performing word recognition. Details in technique are described and its efficiency performance on training scores agree with improvement in recognition rates when training words with support vector machine.

References

[1] Ozdas, A., Shiavi, R.G., Wilkes, D.M., Silverman, M., Silverman, S.,"Analysis of Vocal Tract Characteristics for Near-term Suicidal Risk Assessment", Meth. Info. Med., vol. 43, pp 36-38, 2004.

[2] Godino-Llorente J.I., Gomez-Vilda P., and Blanco-Velasco M., "Dimensionality Reduction of a pathological Voice Quality Assessment System Based on Gaussian Mixture Models and Short Term Cepstral Parameters", IEEE Transaction on Biomedical Engineering, 53(10):1943-1953, 2006.

[3] Lu-Shih Alex Low, et al., "Content Based Clinical Depression Detection in Adolescents", 17th EUSIPCO 2009, Scotland Aug. 24-28, 2009.

[4] A.J. Richard, *Applied Multivariate Statistical Analysis.* 3th ed., Prentice hall, New Jersey, 1992.

[5] Koeing, W., "A new frequency scale for acoustic measurements", Bell Telephone Laboratory Record", Vol. 27, pp. 299-301, 1949

[6] C. Thanawattano and S. Tan-a-ram, "Cardiac arrhythmia detection based on signal variation characteristic", BMEI2008, Hainan, China, 2008.

[7] C. Cortes and V.N. Vapnik , "Support vector networks", *Machine Learning*, vol.20, pp. 1-25, 1995.