Classification of Depressed Speakers based on MFCC in Speech Samples

Thongluan Laosaphan and Thaweesak Yingthawornsuk

Abstract - This paper describes an approach of classification of speech sample using the Mel-Scale Frequency Cepstral Coefficients (MFCC) extracted from speech samples of depressed speakers and clinically approved normal speakers. Principal Component Analysis (PCA) is employed in feature dimensional reduction state, prior to training and testing MFCC extracts via Maximum Likelihood Classifier (ML). Based on experimental database of total 31 females categorized into 17 depressed (DP) and 14 remitted (RM) females, all speech samples were collected under acoustically control with same environmental condition. Based on results obtained from analysis and classification the sixteen-ordered MFCC extracts have shown the improvement in correct classification when training the ML with more MFCC samples by randomly selected from database.

Keywords - Classification, Speech, MFCC, PCA, ML.

I. INTRODUCTION

Speech Recognition is the process of automatically recognizing the speech samples of depressed and remitted 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 samples 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 speech 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 provide the details on database, processing methods of; voice-segment detection, MFCC feature extraction, principal component analysis and performance evaluation, finally results and discussion.

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 17 depressed “DP” females and another is a group of 14 remitted “RM” females. 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 samples of depressed (DP) in upper plot, remitted (RM) in lower plot and voiced segments of speech signal as followings.
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 $\delta_1 = 2^3$ and the highest energy level is $\delta_N = 2^6$. Segments of sound signal with its largest energy level estimated at scale $\delta_1 = 2^3$ are therefore identified as unvoiced segment, otherwise found to be voiced segments. The following equation is the energy threshold defining as unvoiced segment:

$$ n = (n|\delta_i = 2^3); \quad n = 1, ..., N $$

At which $\nu$ is the unvoiced segment of the $n$ segment with energy at scale $\delta_i$ maximized.

B. Acoustical Feature Extraction

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

- Segmenting all concatenated voiced speech signal into 256ms-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. All processes are implemented in Matlab 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:

$$ f_{\text{mel}} = 2595 \cdot \log_{10} \left[ 1 + \frac{f_{\text{lin}}}{700} \right] $$

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

Fig. 3 Logarithmic plot of the mapping frequencies between 0 and 10 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 $\left[ 0 - \frac{f_T}{2} \right]$. The center frequency of individual filter is defined:

$$ F_{C,i} = k_i \frac{f_s}{N}; \quad i = 1, 2, 3, ..., N_f $$

And its bandwidth is consequently computed by

$$ B_i = F_{C,i+1} - F_{C,i-1}; \quad i = 1, 2, 3, ..., N_f $$

Here $N'$ is the flt 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 $F_{C,N_f} < \frac{f_T}{2}$. 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:

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

Due to the computational simplicity, equation (5) without the superfluous factor $\frac{1}{2}$ 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

The classifiers, Maximum Likelihood (ML) 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 30%, 50% of the rest of dataset used later for testing the same classifier. Several trials on random selection of samples from our dataset with 20%, 30% 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];

\[
P(mfcc_i|\omega_2)P(\omega_2) > P(mfcc_i|\omega_1)P(\omega_1)
\]

(6)

This means \(mfcc_i\) is in class \(\omega_1\), otherwise \(mfcc_i\) is identified as class \(\omega_2\). Where \(P(\omega_i)\) 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 pro-bability for class \(i\). Furthermore, the inequality can be re-arranged to obtain another decision rule;

\[
L_R(mfcc_i) = \frac{P(mfcc_i|\omega_1)}{P(mfcc_i|\omega_2)} > \frac{P(\omega_2)}{P(\omega_1)} \tau_c = \frac{P(\omega_2)}{P(\omega_1)}
\]

(7)

Then \(x\) is in class \(\omega_1\). 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 \(\omega_1\). 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 \(\omega_1\), and if \(L_R < 1\), then we classify the observation as belonging to \(\omega_2\). We can also adjust this threshold to obtain a desired probability of false alarm.

IV. EXPERIMENTAL RESULTS

Results of the sixteen-ordered MFCC extracted from database of DP and RM speech samples are shown in Figure (4). The significant difference in quantity can be fairly identified between sample classes of two speech sample groups. The comparative performances obtained from several trials on sample selections in training and testing states are graphically plotted in box-and-whisker diagrams as convenient examination on statistical descriptives. Training results of classification shown in Figure (5), in case of ML classifier of samples from our dataset with 50%, provide much less error in term of median value with consistent scores from both training and testing MFCC samples as compared to dataset with 20%, 30%. In addition more decreasing change in maximum of samples from our dataset with 50% and minimum adjacent values depicted as top and bottom bars of individual box plots can be notified as well of ML samples from our dataset with 20%, 30%. These suggest that the ML classifier of samples from our dataset with 50% seems to provide more consistent and reliable performance on training sample state than ML of samples from our dataset with 20%, 30%, 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 ML classifying score from our dataset with 50% seem more expanded but consistent for both training and testing than those of ML of samples from our dataset with 20%, 30% for all percentages of dataset tested for classifications.
V. CONCLUSION

This paper has presented the principle of speech MFCC extraction for performing group classification between DP and RM speech samples. Details in technique are described and its efficiency performance on training scores confirmed improvement in correct classification when training ML classifier with more samples. Tendency of higher correct classification can be obtained with larger sample size than present study used.

ACKNOWLEDGEMENT

Author would like to gratefully thank Dr. Thaweesak Yingthawornsuk, Department of Media Technology, King Mongkuts University of Technology Thonburi - BangKhunthian Campus for valuable suggestions on experimental design and discussion on results. This research work has been financially granted by National Research Council of Thailand.

REFERENCES