MFCC와 DTW에 알고리즘을 기반으로 한 디지털 고립단어 인식 시스템

장 한, 정 길 도
전북대학교 전자정보공학부

Digital Isolated Word Recognition System based on MFCC and DTW Algorithm

Xian Zang, Kil To Chong
Department of Electronic Information Engineering, Chonbuk National University, Jeonju, Republic of Korea

Abstract - The most popular speech feature used in speech recognition today is the Mel-Frequency Cepstral Coefficients (MFCC) algorithm, which could reflect the perception characteristics of the human ear more accurately than other parameters. This paper adopts MFCC and its first order difference, which could reflect the dynamic character of speech signal, as synthetical parametric representation. Furthermore, we quote Dynamic Time Warping (DTW) algorithm to search match paths in the pattern recognition process. We use the software "GoldWave" to record English digitals in the lab environments and the simulation results indicate the algorithm has higher recognition accuracy than others using LPCC, etc. as character parameters in the experiment for Digital Isolated Word Recognition (DIWR) system.

Key Words -Digital Isolated Word Recognition, Mel-Frequency Cepstral Coefficients, Dynamic Time Warping

1. INTRODUCTION

The goal in research of speech recognition is to make machine understand spoken language from human. "Understanding" contains two implications, one is the reliable transcription of spoken language into written words, and another is the ability to make correct response to the request or command included in the spoken language. This technique is an important embranchment of pattern recognition. In speech recognition systems, the extraction of features in the acoustic signal is an important task. These character parameters should be useful representation of the information carried by the speech signal. There are several kinds of parametric representations for the acoustic signals, such as energy, pitch, formants, Linear Predictive Cepstral Coefficients (LPCC) which are based on the model of vocal tract, and Mel-Frequency Cepstral Coefficients (MFCC) which are based on auditory response. Among them MFCC is the most widely used [1-3]. As we know, the human ear does not show a linear frequency resolution but builds several groups of frequencies and integrates the spectral energies within a given group. Furthermore, the mid-frequency and bandwidth of these groups are non-linearly distributed. The non-linear warping of the frequency axis can be modeled by the so-called mel-scale. The frequency groups are assumed to be linearly distributed along the mel-scale. The MFCC, which are the result of a cosine transform of the real logarithm of the short-term energy spectrum expressed on a mel-frequency scale. The superiority of MFCC consists in producing higher recognition accuracy in presence of channel noise and spectral distortion.

2. DIGITAL PROCESSING OF SPEECH SIGNAL

Fig.1 is the block diagram of the MFCC extraction algorithm.

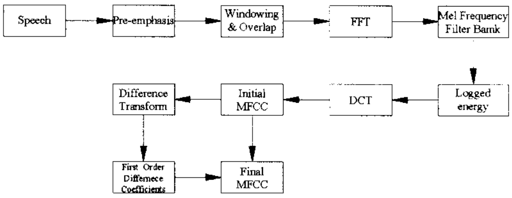

Fig. 1 Block diagram of the MFCC extraction algorithm

2.1 Pre-processing

Pre-processing include pre-filtering, sampling, A/D conversion, pre-emphasis and windowing.

The speech is first pre-filtered for two purposes: (1) the low-pass filter is used to restrain the high-frequency parts that over (is sampling frequency) in order to prevent aliasing interference in the time domain (2) the high-pass filter is used to restrain the 50Hz power interference.

Then according to Nyquist sampling theorem and depending on the implementation, a sampling frequency and an 8 bit quantization of the signal amplitude is used. After digitizing the analog speech signal, we get a series of speech samples .

Next we use a pre-emphasis filter to sharpen the amplitude of high-frequency parts for flattening the signal spectrum since there is decrease of -6dB per octave. The relationship between the output and the input of the pre-emphasis block is shown in (1).

(1)

where, $\alpha$is between 0.9 and 1usually. The default value of $\alpha$is 0.97.

Then we separate the pre-emphasized speech into short segments called frame. The frame length is set to32ms (256samples) to guarantee stationarity inside the frame and there is a 16ms (128 samples) overlap between two adjacent frames to ensure stationary between frames.

A frame can be seen as the result of the speech waveform multiplies a rectangular pulse whose width is equal to the frame length. This will introduce significant high frequency noise at the beginning and end points of the frame because of the sudden changes from zero to signal and from signal to zero. To reduce this edge effect, a 256-points Hamming window is applied to each frame. The mathematical expression of the Hamming window is shown in equation (2),

(2)

where N is equal to 256, the number of points in one frame. By multiplying our speech signal with the time window, we get a short speech segment .

(3)

2.2 MFCC Extraction Process

After the FFT block, the spectrum of each frame is filtered by s set of filters and the power of each band is calculated. To obtain a good frequency resolution, a 256-point FFT is used [4]. Because of the symmetry property of FFT, we only need to calculate the first 128 coefficients.

The so-called mel-frequency can be computed from the linear frequency f by using Fant's expression (4):

(4)

Here, we computed the maximum frequency of the sampled signal:

(5)

where is sampling frequency, .

The filter bank consists of 24 triangular shaped band-pass filters, which are centered on equally spaced frequencies in the Mel domain between 0Hz and 4kHz.

We can calculate the 24 values of Mel-Frequency filter bank energies by integrating the spectral energy within triangular frequency bins, and then perform the discrete cosine transform (DCT) on the natural logarithm of the filter-bank energies using Eq. (6),

(6)

where is the output power of the filter of the filter bank, and n is from 1 to 12. Thus we get 12 cepstrum coefficients. However, they couldn't reflect the dynamic character of speech signal, so we add the first order difference to the basic static parameters to enhance the performance of the speech recognition system. The first order difference coefficients are obtained from the following formula:

(7)

After combining static MFCC and its first order difference, the total number of MFCC for each frame is 24.

2.4 Dynamic Time Warping (DTW) Algorithm

The dynamic time warping algorithm used in the pattern recognition process serves to estimate the similarity between an unknown token and a reference template. This algorithm is simple and effective therefore it's very applicable for the small-vocabulary isolated word recognition system.

We use T and R to denote test template and reference template respectively, then compute the distance $D[T,R]$ to estimate the similarity, where $D[T,R]$ is the difference between character parameters of corresponding frames from two templates. The distance smaller, the similarity will be higher. We assume n and m are the arbitrary frame in T and R respectively, and $d[T_n, R_m]$ are the distance only between the two frames. If we mark the frame sequence of test template $n = 1 \sim N$ as horizontal axis of planar rectangular coordinate system, and take the each frame of reference template $m = 1 \sim M$ as vertical axis, we can establish the grid shown in Fig.2. Each crossing (n, m) denotes the intersection of some frame from test and reference template. DTW algorithm can be concluded to search an optimal path, in which the crossings are the frame sequence for distance calculation.

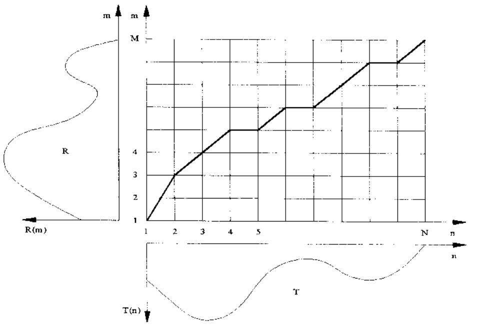

Fig. 2 Search path of DTW algorithm

The slope of the path is limit between 0 to 2, thus the lattice point before next reachable one only could be (n_{i-1}, m_i), $(n_{i-1}, m_i - 1)$ and $(n_{i-1}, m_i - 2)$, the cumulative distance to the three points are $D[(n_{i-1}, m_i)]$, $D[(n_{i-1}, m_i - 1)]$ and $D[(n_{i-1}, m_i - 2)]$. Then (n_i, m_i) must select one of them as the next point, which should be closest to (n_i, m_i), to ensure the cumulative distance is minimal and get the optimal path sequentially. That is,

$$D[(n_i, m_i)] = d[T_{n_i}, R_{m_i}] + D[(n_{i-1}, m_{i-1})] \tag{8}$$

where $n_{i-1} = n_i - 1$, m_{i-1} will be determined by the following formula:

$$D[(n_{i-1}, m_{i-1})] = \min\{D[(n_{i-1}, m_i)], D[(n_{i-1}, m_i - 1)], D[(n_{i-1}, m_i - 2)]\} \tag{9}$$

Thus, we could search from $(n_0, m_0) = (1, 1)$, each (n_{i-1}, m_{i-1}) saves the anterior lattice point (n_{i-1}, m_{i-1}) and corresponding distance $D[(n_i, m_i)]$. We only reserve an optimal path when searching to the end, then search backwards point-by-point to get the whole path.

3. SIMULATION RESULTS

We use the software "GoldWave" to record English digitals 0-9 in the lab environments in our experiments, each acoustic signal was low-pass filtered at 5 kHz and sampled at 8 kHz, also an 8 bit quantization of the signal amplitude was used. Then each speech signal was pre-emphasized first. Fig. 3 showed the frequency spectrum of speech "9" before and after pre-emphasis. The frame length was set to 32ms, and there was a 16ms overlap between two adjacent frames. Then a 250-points Hamming window was applied to each frame to select the data points to be analyzed. Fig. 4 showed the comparison of speech "2" before and after Hamming window. For the MFCC computations, 24 triangular bandpass filters were simulated. We calculated the 24 values of Mel-Frequency filter bank energies and performed the discrete cosine transform (DCT) on the natural logarithm of them to get 12 cepstrum coefficients. In order to enhance the performance of the speech recognition system, we added the first order difference to the basic static parameters to form 24-order MFCC parameters for each frame. The last step was the pattern recognition process using dynamic time warping algorithm.

We performed several simulations using different test templates of 0-9, the recognition results indicated the algorithm based on MFCC and DTW has good recognition accuracy about 95% for Digital Isolated Word Recognition (DIWR) system, which is higher than others using LPCC, etc.

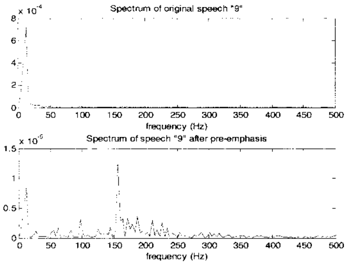

Fig. 3 The frequency spectrum of speech "9" before and after pre-emphasis.

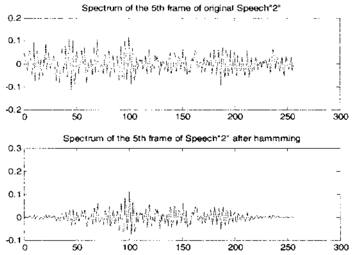

Fig. 4 The spectrum of speech "2" before and after Hamming window.

4. CONCLUSION

This paper introduced the MFCC and DTW algorithm for Digital Isolated Word Recognition (DIWR) system; also we demonstrated the detail process for MFCC extraction and pattern recognition. From the experimental results, we can tell that why the MFCC is superior to other parameters consists in that MFCC imitates the auditory characteristics of the human ear, while other feature parameters such as LPCC, etc. which are based on the model of vocal tract, maybe could result in low recognition accuracy in presence of noise. Further more, we used DTW to solve the problem of each speech section interval varying in different in different situations. Based on the combination of MFCC and DTW, the DIWR system performed recognition well.

[REFERENCES]

[1] L. Rabiner and Biing-Hwang Juang, Fundamentals of Speech Recognition, Prentice Hall PTR, c1993.

[2] Joseph W. Picone, "Signal Modeling Techniques in Speech Recognition", Proceeding of the IEEE, vol.81, No.9, pages 1215-1247, 1993.

[3] Steven B. Davis and Paul Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-28, No.4, August 1980.

[4] Steven W. Smith, The Scientist and Engineer's Guide to Digital Signal Processing, California Technical Publishing, 1997, pages 169-174.