Dynamic time warping for speech recognition with training part to reduce the computation

Sun Xihao * Yoshikazi Miyanaga †
Graduate School of Information Science and Technology Hokkaido University
E-mail: sonkikou@icn.ist.hokudai.ac.jp* miya@icn.hokudai.ac.jp†

***Abstract*—Dynamic time warping (DTW) is a popular automatic speech recognition (ASR) method based on template matching[1], [2]. DTW algorithm compares the parameters of an unknown word with the parameters of one reference template. But the recognition rate is limited. To increase the number of reference templates for the same word will improve the recognition rate, but it will lead to spend a lot of computing time and memory resource. In this paper we proposed a method to reduce the number of reference templates, thus reduces the computing time and memory resource and also keep the high recognition rate.**

I. Introduction

There are two main techniques in speech recognition. One is hidden markov model (HMM) , the other is DTW. Although HMM is a very popular technique in speech recognition, DTW is still used in the small-scale embedded systems (e.g. cell phones, mobile applications) because of simplicity of its hardware implementation, straightforwardness and speed of the training procedure[3],[4]. The Fig. 1 shows a simple speech recognition system using DTW.

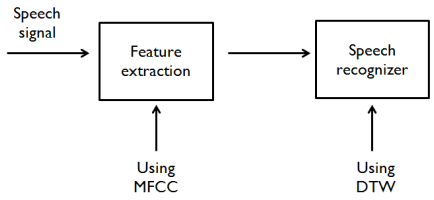

Fig. 1. Example of speech recognition using DTW

Conventional DTW has fast search and low complexity, but it has poor speech recognition rate. Therefore, DTW has mostly been used for speech recognition in clear speech environments. In order to improve the recognition rate in the noisy environments using DTW, a better way is to increase the number of templates for the same word. In this paper we will give a way to find a appropriate reference template to replace the increasing templates.

II. Parameters Extraction and DTW

A. Mel-Frequency Cepstral Coefficients (MFCC)

A well known parameter extraction method is Mel-frequency cepstral coefficients(MFCC)[5].That is because MFCC can better describe the nonlinear relation that humans ear feels the frequency of speech signal.By analyzing the spectrum of speeches, we can obtain the better accuracy and robust. Fig.2 shows the structure of MFCC to get the feature extraction.

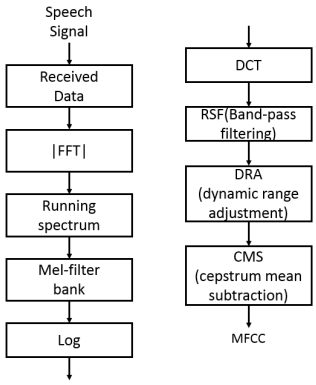

Fig. 2. Feature Extraction

B. DTW

The objective of DTW is to warp two speech templates $P = (p_1, p_2, \cdots, p_I)$ and $Q = (q_1, q_2, \cdots, q_J)$ in the time dimension as represented in Fig. 3. Each p_i and q_j is a vector of parameters (MFCC).

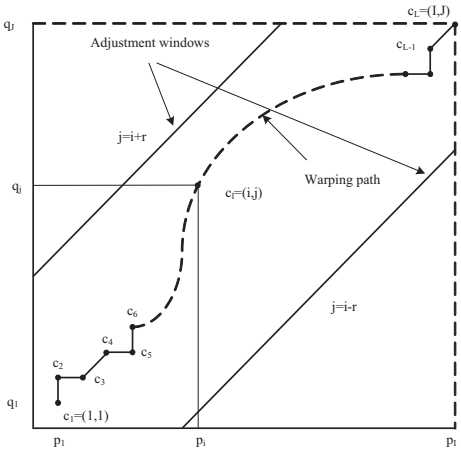

Fig. 3. Example of speech templates

These two speech templates are of the same category, the timing differences between them can be depicted by a

978-1-4673-6143-9/13/$31.00 ©2013 IEEE

sequence of points $c = (i, j)$:

$$C = c(1), c(2), \cdots, c(L) \tag{1}$$

where

$$c(l) = (i(l), j(l)) \tag{2}$$

This sequence can be considered to represent a warping path which approximately realizes a mapping from the time axis of template P onto that of template Q. As a measure the difference between two speech vectors p_i and q_j, a distance $d(i, j)$ is defined.

$$d(c) = d(i, j) = \|a_i - b_j\| \tag{3}$$

We will compute the distance between the starting point (1, 1) and the end point (I, J) from left to right $D(I, J)$.

$$D(C) = \sum_{l=1}^{L} d(c(l)) \tag{4}$$

Since there are X possible paths from (1, 1) to (I, J), We will identify the smallest accumulated distances from (1, 1) to (I, J) among all possible, and the path which has the minimum $D(I, J)$ is the optimal path between P and Q.

III. PROBLEM AND SOLUTION

As our said the more templates we used for the same word, the more memory resources and computing time we need to pay. So the problem become to how to find a best reference template to replace the reference templates. Actually, the DTW algorithm provides the optimal path which is the key to find a best reference template. We will detailed explain in the training part.

A. training part

1) One pair of vectors: for simplicity we will assume the case of a one dimensional feature.Then, the two reference templates will be represented by two vectors

$$P = (p_1, p_2, \cdots, p_i) \tag{5}$$

$$Q = (q_1, q_2, \cdots, q_j) \tag{6}$$

where p_i and q_j are scalars.

The DTW algorithm provides the optimal path C the one that minimizes the cumulative error. That also means the optimal path is the most similar vector between P and Q, so new vector can be used to replace the P and Q vectors (Our experiment use 100 reference words for each recognized word, after one time training the reference words become to 50).

For getting the new vector C, we observed the optimal path, and consider continuous three cells $c(x-1)$, $c(x)$, $c(x+1)$ (Although consider three cells, but every time we just compute the cell $c(x)$). There are 9 kind of paths, and we divide them into two classes as represented in Fig. 4.

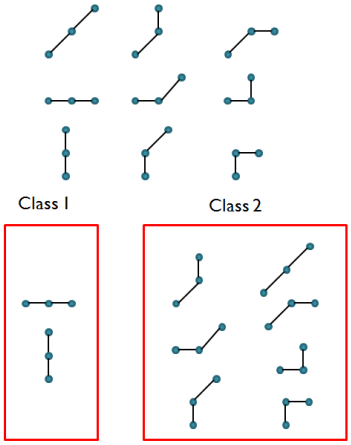

Fig. 4. Class of optimal path

Class 1 is the cells of optimal path on the same row or the same column. We will consider cell $c(x) = c(i, j)$ from the optimal path and calculate its as:

$$c(i, j)_m = (p_i + q_j) \tag{7}$$

Class 2 is the cells of optimal path not on the same row and the same column. We will consider cell $c(x) = c(i, j)$ from the optimal path and calculate its centroid as:(If the cells $c(x-1)$, $c(x)$ in the same row the p_i will be reset to zero, if the cells $c(x-1)$, $c(x)$ in the same column the q_j will be reset to zero in equation 8)

$$c(i, j) = \frac{(p_i + q_j + c(i, j)_m)}{N_m} \tag{8}$$

where N_m is the number of merge cell.

We define the new vector $C = (c_1, c_2, \cdots, c_N)$, as a set of centroid calculated using equation 8.

This new vector can be used to replace the P and Q vectors.

2) Pairs of MFCC vectors: for simplicity, we assume that only one scalar is associated with each window. In reality for recognition we use K-MFCC parameters.(Our experiment use 36-dimensions MFCC.)

The first reference template $P = [P(1), P(2), \cdots, P(I)]$ is defined by the following equation:

$$P = \begin{bmatrix} p(1)(1) & p(2)(1) & \ldots & p(I)(1) \\ p(1)(2) & p(2)(2) & \ldots & p(I)(2) \\ \ldots & \ldots & \ldots & \ldots \\ p(1)(K) & p(2)(K) & \ldots & p(I)(K) \end{bmatrix} \tag{9}$$

The second reference template (for the same word) is $Q = [Q(1), Q(2), \cdots, Q(J)]$ and it is defined by the following equation:

$$Q = \begin{bmatrix} q(1)(1) & q(2)(1) & \ldots & q(J)(1) \\ q(1)(2) & q(2)(2) & \ldots & q(J)(2) \\ \ldots & \ldots & \ldots & \ldots \\ q(1)(K) & q(2)(K) & \ldots & q(J)(K) \end{bmatrix} \tag{10}$$

Similarly, we will compute the centroid for each class. In this case each centroid for each class is also a vector. The resulted vector C defined by the optimal path based on the defined classes.

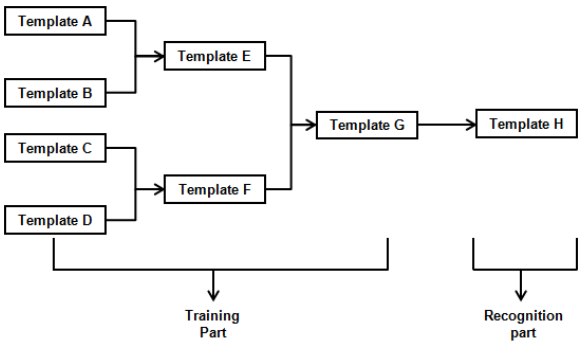

Fig. 5. Example of system structure

The detail of the proposed method is shown below:

1) In the training part of DTW, for each same word $N = 2p$ waves will be recorded which the MFCC parameters will be computed.

2) The resulted reference templates (N) will be divided into two subsets: A and B (same cardinality p). Each reference template A_p will be randomly paired with only one B_P.

3) For each pair A_p and B_p the optimal path will be computed (using the conventional DTW algorithm).

4)Using the DTW algorithm we can get the optimal path between A_p and B_p. According to the equations 7 and 8, we create the new vector C_i.

5) The new vector $C_i = (c_1, c_2, \ldots, c_p)$ will replace the template A_p and B_p).

6) Repeat the process starting with step 4,considering the pair (C_i, $C_i + 1$) as a new A_i , B_i pair.

7) In the recognition part of DTW, the unknown word (X) will be compared to each reference model(C_i).

B. recognition part

In the recognition part we use a nonlinear median filter(NMF)[6]. We assume there are M reference words. For each reference word there are N available reference speech utterances from different speakers. The distance computed between the unknown speech waveform and the n^{th} utterance of the m^{th} reference word is denoted as d_{mn}, $1 \leqslant m \leqslant M$, $1 \leqslant n \leqslant N$. The distances computed between the unknown speech waveform and all utterances of the m^{th} reference word are collected in vector $\mathbf{d}_m = [d_{m1}\, d_{m2} \ldots d_{mn} \ldots d_{mN}]^{\mathrm{T}}$. Then, all distances between the unknown speech waveform and all reference utterances can be represented in matrix form as

$$\mathbf{D} = \begin{bmatrix} \mathbf{d}_1^{\mathrm{T}} \\ \mathbf{d}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{d}_M^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} d_{1,1} & d_{1,2} & \ldots & d_{1,N} \\ d_{2,1} & d_{2,2} & \ldots & d_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ d_{M,1} & d_{M,2} & \ldots & d_{M,N} \end{bmatrix} \tag{11}$$

Sorting the distances for every reference word into ascending order yields $\mathbf{d}'_m$

$$\mathbf{d}'_m = \begin{bmatrix} d'_{m,1} & d'_{m,2} & \ldots & d'_{m,N} \end{bmatrix} \tag{12}$$

i.e., $d'_{m,1}$ and $d'_{m,N}$ are the minimum and maximum distances, respectively. We thus obtain the ordered distance matrix $\mathbf{D}'$.

Now, employing an NMF of window length k, with $1 \leqslant k \leqslant N$ (According to the experiment's result, when $K = 4$ we will get the best speech recognition rate), we extract from $\mathbf{d}'_m$, $1 \leqslant m \leqslant M$, the median distance

$$a_m = Med(\mathbf{d}'_m) = \begin{cases} d'_{m,\frac{k+1}{2}} & , \text{ if } k \text{ is odd} \\ \frac{1}{2}\left[d'_{m,\frac{k}{2}} + d'_{m,\frac{k}{2}+1}\right] & , \text{ if } k \text{ is even} \end{cases} \tag{13}$$

We thus obtain the vector with all minimum distances from the reference waveforms $\mathbf{a} = [a_1\, a_2 \ldots a_M]^{\mathrm{T}}$.

Whereas in the conventional DTW approaches, the recognized word corresponds to

$$arg \min_{m=1:M} d'_{m,1} \tag{14}$$

in the new approach we propose that the recognized word corresponds to

$$arg \min_{m=1:M} a_m \tag{15}$$

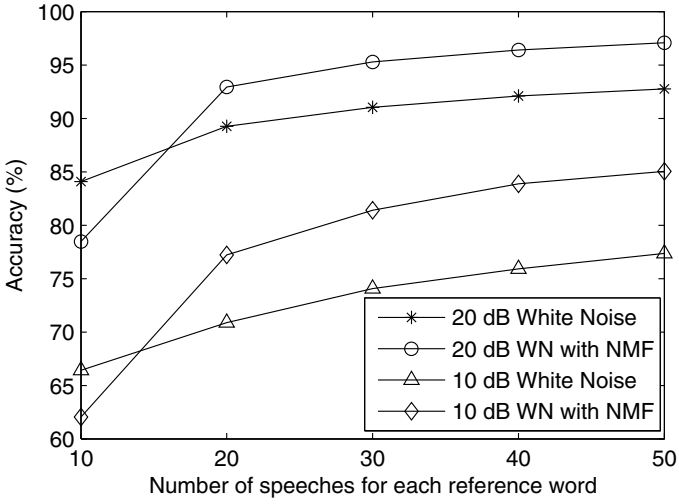

Fig. 6. the recognition accuracy of DTW with NMF K = 4

Fig.6 shows the recognition accuracy of DTW with NMF compare with conventional DTW.

IV. EXPERIMENT AND RESULTS

Conventional recognition systems consist of ordinary feature extraction based on MFCC. The entire recognition system is implemented using MATLAB. The reference database consists of 100 isolated Japanese words, and every word has 100 waveforms spoken by 50 persons. Test words are 50 isolated Japanese words and every word has 100 waveforms

spoken by the other 50 persons. MFCC feature vectors are extracted. These vectors comprise 36-dimensions : 12 cepstral coefficients($s_i(k), i = 1, 2, \ldots, 12$, k : time index), 12 delta cepstral coefficients($\Delta s_i(k) = s_i(k) - s_i(k-1)$), 12 delta-delta cepstral coefficients ($\Delta^2 s_i(k) = \Delta s_i(k) - \Delta s_i(k-1)$). The reference 100 isolated Japanese words have been trained three times, that means the reference 100 isolate Japanese words become to 13 new reference words, other conditions are described in Table 1.

TABLE I
EXPERIMENT SETTINGS AND PARAMETERS

Recognition task	Isolated 100 words
Speech data	100 Japanese region names from JEIDA
Sampling	11.025 kHz, 16 bits
Window length	23.2 ms (256 samples)
Frame length	11.6 ms (128 samples)
Bandwidth of bandpass filter	1-16 Hz
NMF order	4
Feature vector	36-dimensional MFCC
Noise type	white noise 20dB

Let us see the computing time(most of the computing time is spent on DTW). We assume UW is number of unknown word, every unknown word have UWS templates, RW is number of reference word,every reference word have RWS templates. The conventional DTW will calculate $UW \times UWS \times RW \times RWS$ times DTW. According to the proposed method, after one time training we calculate $\frac{\frac{1}{2}RWS \times RW \times UW \times UWS + \frac{1}{2}RWS \times RW}{UW \times UWS \times RW \times RWS}$ times DTW, that means almost 50% computing time have been reduced. Similarly, after three times training almost 85% computing time have been reduced.

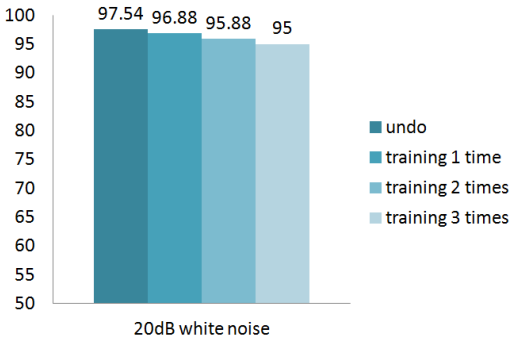

Fig. 7. experiment result

Fig. 7 shows the DTW recognition accuracy in 20dB white noise. We can see after three times training that means 85% computing time have been reduced, meanwhile the recognition accuracy is 95.00% in 20dB white noise.

V. CONCLUSION AND FURTHER WORK

In this paper presents a new training technique to prepare the reference templates for DTW-based speech recognition systems. Significant improvements have been obtained with this training technique. The computation have been reduced almost 85%, meanwhile the recognition accuracy also keep in a high lever.

Although we proposed method has improved the efficiency of automatic speech recognition system with DTW algorithm, the recognition accuracy also keep in a high lever, but there also have some issues.

1) After three times training, the recognition accuracy will significantly reduce.

2) Compare with HMM, the benefit of using DTW is which we do not need to training the data. But for keeping the recognition accuracy and reducing the computation, we add the training part.

In the further, we will improve the algorithm, make sure the recognition accuracy will keep in a high lever even after three times training.

REFERENCES

[1] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics,Speech and Signal Processing,* vol. 26, no. 1, pp. 43 C 49, feb 1978.

[2] G. Kang and S. Guo, "Variable sliding window dtw speech identification algorithm," *in Ninth International Conference on Hybrid Intelligent Systems,* vol. 1, 12-14 2009, pp. 304 C307.

[3] J. Di Martin. "Dynamic Time Warping Algorithms For Isolated And Connected Word Recognition" *in Nato Asi Series,* Vol. F16. R. De Mori and C.Y. Suen, Ed. Berlin: Springer-Verlag, 1985.

[4] T. Zaharia, S. Segarceanu, M. Cotescu, and A. Spataru, "Quantized dynamic time warping (DTW) algorithm," *in The 8th International Conference on Communications (COMM),* Jun. 2010, pp. 91C94.

[5] S. B. Davis. and P. Mermelstein,"Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,"*IEEE Transactions on acoustics, Speech and Signal Processing,* vol.28,no.4,pp.357-366, Aug 1980.

[6] Z. Yuxin, Y. Miyanaga, and C. Siriteanu,"An improved dynamic time warping algorithm employing nonlinear median filtering" *Journal of signal processing,* vol.16, no.2,pp. 147-157, Mar.2012.