Journal of Signal Processing, Vol.18, No.2, pp.89-96, March 2014

PAPER

Dynamic Time Warping for Speech Recognition with Training Part to Reduce the Computation

Xihao Sun, Yoshikazu Miyanaga and Baiko Sai

Graduate School of Information Science and Technology, Hokkaido University
Kita 14 Nishi 9 Kita-ku, Sapporo 060-0814, Japan
E-mail: sonkikou@icn.ist.hokudai.ac.jp, miya@ist.hokudai.ac.jp, baiko.sai@ist.hokudai.ac.jp

Abstract **In this paper, we proposed a dynamic time warping (DTW) method with a training part. DTW is a popular automatic speech recognition (ASR) method based on template matching. Conventional DTW is fast and of low complexity, however its recognition accuracy is limited. Recently, a DTW with multireferences (mDTW) algorithm has also been developed to improve the recognition accuracy to be comparable to that of the hidden Markov model (HMM) algorithm under noisy conditions. However the mDTW algorithm increases the calculation cost. Therefore, in order to reduce the calculation cost, in this paper, a training part will be added to the DTW-based ASR system, unlike the mDTW, which tries to find appropriate reference utterances to replace the increasing utterances. The results show that the average recognition accuracy of the proposed method is similar to that of the mDTW, and the calculation cost was reduced by 41.6%.**

Keywords: dynamic time warping (DTW), robust speech recognition system, calculation cost, memory resource

1. Introduction

There are two main techniques in speech recognition. One is the hidden Markov model (HMM) and the other is dynamic time warping (DTW). Although DTW has low recognition accuracy, it is still used in small-scale embedded systems (e.g., cell phones and mobile applications) because of the simplicity of its hardware implementation, straightforwardness, and speed of the training procedure[1,2]. Recently, a multireference DTW (mDTW) has been developed to increase the recognition accuracy. Even if the mDTW improves the recognition accuracy to be similar to that of the HMM algorithm however, the cost of mDTW increases dramatically. Therefore, in this study, we attempt to reduce the total cost of the mDTW-based speech recognition approach and we call the modified method training dynamic time warping (tDTW).

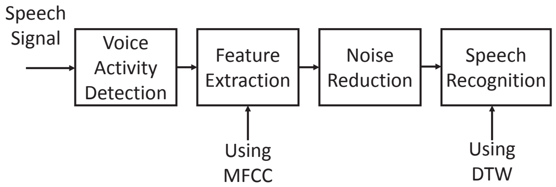

Fig. 1 Example of speech recognition system using DTW

Figure 1 shows a simple ASR system using DTW. It comprises modules for voice activity detection (VAD), feature extraction, noise reduction, and speech recognition. These modules will be discussed below.

The VAD block is used to detect the beginning and end of speech waveforms and to exclude nonspeech segments [3-7]. This technique is used in this work because nonspeech segments can degrade recognition performance, especially at a low signal-to-noise ratio (SNR)[3-5]. In noisy environments, the noise will smear speech waveforms, thus a robust speech recognition algorithm, such as cepstrum mean subtraction (CMS) [10], running spectrum filtering (RSF), and dynamic range adjustment (DRA)[8-11], is required. These techniques help yield high recognition accuracy, even at a low SNR. However, these techniques do not reduce the calculation cost.

Therefore, in this paper, we propose a tDTW-based ASR system. First, we deploy VAD to prepare the waveform as well as to reduce the response time and computational cost. Then, we employ the CMS, RSF, and DRA noise rejection techniques filter. Finally, we use the tDTW-based ASR system to determine the recognition accuracy. Our experiments indicate that DTW recognition accuracy is similar to that of the mDTW for the same noise reduction method, but the calculation cost is reduced by 41.6%.

The paper is organized as follows In Sect. 2, we describe the VAD method based on a short-time energy method and give the details of the RSF and DRA noise-reduction methods. In Sect. 3, we discuss conventional DTW and mDTW. In Sect. 4, we describe the proposed method that can reduce the computing time and memory resource. In Sect. 5, we present our experimental result. Finally, we draw conclusions in Sect. 6.

2. Conventional Methods

2.1 VAD

The VAD algorithm typically relies on the short-time energy and zero-pass ratio. In this work, we employ the short-time energy method for VAD. The samples of a waveform of the input signal is defined as $x(m)$, where m is the sample index. The short-time square energy of the speech signal, $E_{sqr}(i)$, is defined as

$$E_{sqr}(i) = \sum_{m=-\infty}^{+\infty} [x(m)\omega(m-i)]^2 \tag{1}$$

where $\omega(i)$ is the a window function whose small width in samples represents the frame size. The Hamming window function is used in speech signal processing. The Hamming window function is defined as

$$\omega(i) = \begin{cases} 0.54 - 0.64\cos(\frac{2i\pi}{N-1}) & 0 \leqslant i \leqslant N-1 \\ 0 & other \end{cases} \tag{2}$$

where N is the window length. Then, if $E_{sqr}(i)$ exceeds a certain threshold, frame i is classified as a speech frame, otherwise it is classified as a nonspeech frame.

This threshold must be adjusted to the level of the input signal as follows:

$$\Upsilon = \frac{1}{I}\sum_{i=1}^{I} E_{sqr}(i) \tag{3}$$

We assume that the first five frames are nonspeech data; therefore, their average energy equals the average energy of nonspeech data. Thus, the maximum energy of a nonspeech frame is defined as the threshold Υ . If the energy of some later frame is larger than Υ, then it is considered to be a speech frame.

2.2 Mel-frequency cepstral coefficients (MFCC)

A well-known parameter extraction method is that of Mel-frequency cepstral coefficients (MFCCs)[12]. MFCC can better describe the nonlinear relation. By analyzing the spectrum of speeches, we can obtain better accuracy and robustness. Figure 2 shows the structure of the MFCC method for the feature extraction.

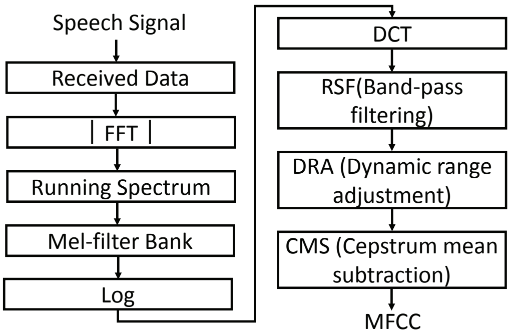

Fig. 2 Feature extraction

2.3 RSF

RSF is a noise reduction method that exploits the difference in temporal variability between the spectra of speech and noise signals to remove the noise [10]. Thus, we can evaluate the different characteristics of speech and noise signals. In the modulation spectrum, we have found that the noise spectrum is concentrated in the direct component (DC). Most of the noise energy is distributed in the low-frequency band of the modulation spectrum. Thus, removing low-frequency components with a high-pass filter can reduce the noise. Therefore, we can use a band-pass filter to separate speech from noise.

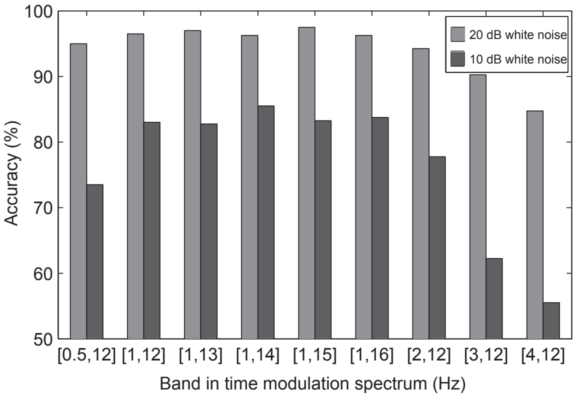

Fig. 3 DTW recognition accuracy vs band for RSF

Figure 3 depicts the recognition accuracies of different bands filtered by RSF. It is shown that the best performance in 10 dB white noise is obtained for the band from 1 to 14 Hz and that in 20 dB white noise is for the band form 1 to 15 Hz. Note that the 1 to 15 Hz band-pass filter is selected in this paper, since the accuracy for the 1 to 15 Hz band is the highest among the 10 and 20 dB SNR conditions.

2.4 DRA

One of the major causes of noise corruption is derived from the differences in the dynamic ranges of the cepstrum. The dynamic range of the cepstrum indicates the difference between the maximum and minimum cepstral values in each order. However, the amplitude difference between clean and noisy speech deteriorates the recognition accuracy. DRA can be used to compensate for this difference using the following normalization:

$$x_i'(t) = \frac{x_i(t)}{\max\limits_{j=1,\cdots,h} |x_j(t)|} \tag{4}$$

2.5 CMS

CMS is a channel normalization approach to compensate for the acoustic channel. The time invariant channel parameters in a recording system and convolutional disturbance noise are evaluated by CMS, and these removal of such noise results in the observed speech waveform. The working of CMS is simple. After feature extraction, the MFCC feature vectors are obtained in the cepstral domain. In a longtime range, almost all speech features are changed with time. On the other hand, the time-invariant noise features in such a range are considered to be almost constant. The subtraction of the time-invariant features from noisy speech features results in the reduction of noise components. Noise reduction is then executed as

$$c_t' = c_t - \frac{1}{T}\sum_{j=1}^{T} c_j \tag{5}$$

where t is the frame time index, T is the total number of frames, and $\frac{1}{T}\sum_{j=1}^{T} c_j$ is the mean MFCC vector from each MFCC feature vector c_t.

3. DTW

3.1 Conventional DTW

The DTW algorithm is based on dynamic programming and provides a means of template matching for different lengths of pronunciation [13,14]. It is a nonlinear warping technique where time series are stretched and compressed to match the reference template. In the field of speech recognition, the objective of DTW is to warp two sequences of speech feature vectors until an optimal match is found.

The sequence of feature vectors of test speech P is $[p(1), p(2), \cdots, p(i), \cdots, p(I)]$. The sequence of feature vectors of the reference speech Q is $[q(1), q(2), \cdots, q(j), \cdots, q(J)]$.

In order to clarify the nature of the differences, let us consider an $i - j$ plane, as shown in Fig. 4, where speeches P and Q are developed along the i-axis and j-axis, respectively. The differences between them can be depicted by a sequence of points $c = (i, j)$:

$$C = [c_1, c_2, \cdots, c_l, \cdots, c_L] \tag{6}$$

where

$$c_l = (i(l), j(l)) \tag{7}$$

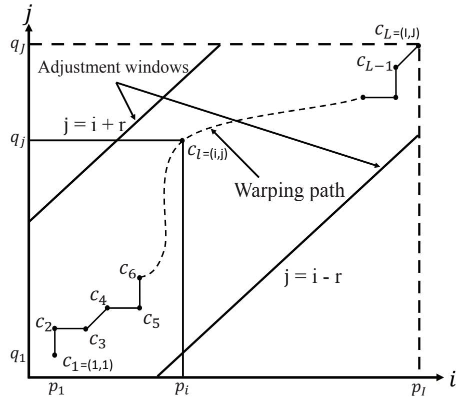

Fig. 4 DTW

As a measure of the difference between two speech vectors p_i and q_j, the warping path distance from $(1, 1)$ to (i, j) is defined as $d(i, j)$. We will compute the distance between the starting point $(1, 1)$ and the end point (I, J) from left to right $D(I, J)$.

$$D(C) = \sum_{l=1}^{L} d(c_l) = \sum_{i=1}^{I}\sum_{j=1}^{J} \|p(i) - q(j)\| \tag{8}$$

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 paths, and the path with the minimum $D(I, J)$ is the optimal path between P and Q. To determine the optimal warping path, the following conditions can be realized as the following restrictions on the warping path and are shown in Fig. 4.

1) Monotonic conditions:

$$\begin{aligned} i(l-1) &\leqslant i(l) \\ j(l-1) &\leqslant j(l) \end{aligned} \tag{9}$$

The monotonic conditions express the characteristics of the time sequence of speech signals. The precedence order cannot be changed after a warped sequence.

2) Continuous conditions:

$$\begin{aligned} i(l) - i(l-1) &\leqslant 1 \\ j(l) - j(l-1) &\leqslant 1 \end{aligned} \tag{10}$$

The continuous conditions express how to choose the adjacent frame.

3) Boundary conditions:

$$\begin{aligned} i(1) &= 1, j(1) = 1 \\ i(L) &= I, j(L) = J \end{aligned} \tag{11}$$

The boundary conditions define the beginning point as point $c_1 = (1, 1)$ and the end point as $c_L = (I, J)$ for all paths. These also express the characteristics of the time sequence of speech signals. The two endpoints of two patterns first must be aligned.

4) Adjustment window condition:

$$\left|i(l) - j(l)\right| \leqslant r \tag{12}$$

In fact, all warping paths from $(1, 1)$ to (I, J) may not cross all points. Thus, adjustment windows defined the computation area for the warping function. The points out side of the adjustment windows are excluded from the calculation. In other words, the paths that cross the points out side of the adjustment windows are not the optimal path. However, the calculation cost of the DTW algorithm can be reduced much more efficiently.

According to the above restrictions, the optimal warping path distance can be expressed as

$$D(i, j) = \min \begin{pmatrix} D(i-1, j) + d(i, j) \\ D(i-1, j-1) + 2d(i, j) \\ D(i, j-1) + d(i, j) \end{pmatrix} \tag{13}$$

3.2 DTW with multireferences (mDTW)

Conventional DTW is capable of fast search and low complexity, but it has poor speech recognition accuracy. In order to improve the recognition accuracy in noisy environments using DTW, a better way is to increase the number of utterances for the same word.

mDTW [15] has been developed. First, we assume there are M reference words, and each word has N speech utterances from difference 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

$$\mathbf{D} = \begin{bmatrix} \mathbf{d}_1^{\mathrm{T}} \\ \mathbf{d}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{d}_M^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} d_{1,1} & d_{1,2} & \ldots & d_{1,N} \\ d_{2,1} & d_{2,2} & \ldots & d_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ d_{M,1} & d_{M,2} & \ldots & d_{M,N} \end{bmatrix} \tag{14}$$

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

$$\mathbf{d}'_m = \begin{bmatrix} d'_{m,1} & d'_{m,2} & \cdots & d'_{m,N} \end{bmatrix} \tag{15}$$

That is, $d'_{m,1}$ and $d'_{m,N}$ are the minimum and maximum distances, respectively.

In contrast, in the mDTW approaches, the recognized word corresponds to

$$\underset{m=1:M}{\operatorname{argmin}}\, d'_{m,1} \tag{16}$$

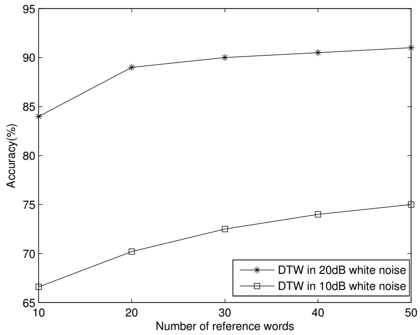

Fig. 5 Recognition accuracy of mDTW

Figure 5 shows the recognition accuracy of the mDTW algorithm for different numbers of reference utterances for each word. For this implementation, the reference database consists of 100 isolated Japanese words, and every word has 10 to 50 waveforms spoken by different persons, and the test words are 50 isolated Japanese words. Other conditions are described in Table 1. Note that although accuracy continues to improve with a higher number of reference utterances for each word, calculation complexity also increases substantially because of the increasingly large reference database. In the following section, we present a way of finding an appropriate reference utterance to replace the increasing number of utterances, thus reducing the calculation cost while maintaining the high recognition accuracy.

4. Proposed tDTW

As stated above, the more utterances we used for the same word, the more memory resources and computing time we need to pay. Therefore, the problem becomes how to find the best reference utterance to replace the large number of reference utterances. Actually, the DTW algorithm provides the optimal path for finding the best reference template. We give a detailed explanation in the following part.

4.1 One pair of vectors

For simplicity, first, we assume one pair of speech feature for the same word, $P = [p(1), p(2), \cdots, p(i), \cdots, p(I)]$ and $Q = [q(1), q(2), \cdots, q(j), \cdots, q(J)]$, as mentioned in Section 3. Then, by using the DTW algorithm, the optimal path between P and Q is defined as

$$C_{opt} = [c_1, c_2, \cdots, c_l, \cdots, c_L] \tag{17}$$

where c_l is a point on the $i - j$ plane, the coordinates of which are $(i(l), j(l))$, with the value $(p(i(l)), q(j(l)))$.

The optimal path C_{opt} is the one that minimizes the cumulative error path between P and Q. In other words, the value of each optimal path point is the cloest value between P and Q. Therefore, let us consider defining a new vector C' to replace P and Q on the basis of the optimal path.

First, we consider the optimal path to represent a function that approximately realizes mapping from the axis of speech feature P onto that of speech feature Q. The slope of every two points in this function is calculated by

$$S = \frac{i(l+1) - i(l)}{j(l+1) - j(l)} \tag{18}$$

where S is the slope of two points. Actually, there are only three kinds of slope, as represented in Fig. 6.

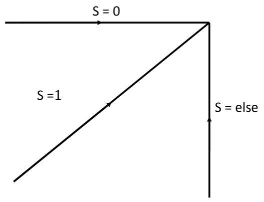

Fig. 6 Types of slope

Then, every two points other than the starting point and end point will be merged into a new point with the value expressed as

$$c'(l) = \begin{cases} \frac{p(i(l))+p(i(l+1))+q(j(l))}{3} & \text{if S = 0} \\ \frac{p(i(l))+q(j(l))}{2} & \text{if S = 1} \\ \frac{p(i(l))+q(j(l))+q(j(l+1))}{3} & \text{else} \end{cases} \tag{19}$$

The starting point and end point remain the original values.

Finally, we define the new vector C' as a set of centroids calculated using Eq. (19).

$$C' = [c'_1, c'_2, \cdots, c'_N] \tag{20}$$

Figure 7 shows the details of the merging rule.

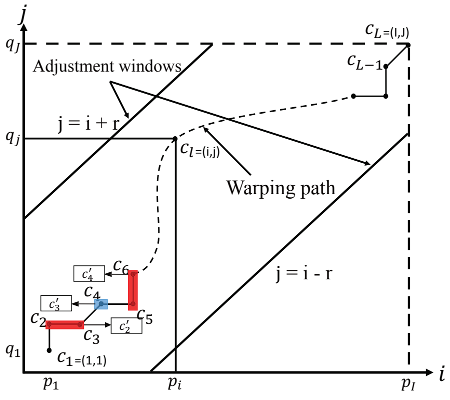

Fig. 7 Merging rule

4.2 Pairs of vectors

On the basis of the above explanation, we solve the condition in which there are pairs of vectors. The proposed method proceeds in the following steps.

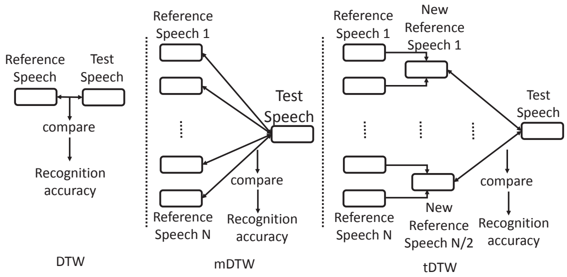

Fig. 8 Basic algorithms of different DTW methods

1. We assume there are M reference words, where each word has N speech utterances from difference speakers. For each reference word, N speech utterances will be divided into two subsets.

2. For each pair of subsets, the optimal path will be computed. According to Eq. (19), the new vector will replace the pair of subsets. The number of speech utterances will be reduced to $N' = N/2$.

3. If we repeat step 2, the number of speech utterances will be further reduced. In other words, if we repeat step 2 t times (we call it training t times), then the number of speech utterances will be reduced to $\frac{1}{2^t}N$.

4. The distances computed between the unknown speech waveform and all utterances of reference word M are collected in a matrix as

$$\mathbf{D} = \begin{bmatrix} \mathbf{d}_1^{\mathrm{T}} \\ \mathbf{d}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{d}_M^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} d_{1,1} & d_{1,2} & \dots & d_{1,N'} \\ d_{2,1} & d_{2,2} & \dots & d_{2,N'} \\ \vdots & \vdots & \ddots & \vdots \\ d_{M,1} & d_{M,2} & \dots & d_{M,N'} \end{bmatrix} \tag{21}$$

5. As in the mDTW algorithm, sort the distances for each reference word. The recognized word corresponds to

$$\underset{m=1:M}{\operatorname{argmin}}\, d'_{m,1} \tag{22}$$

6. Finally, in the recognition part, the recognition accuracy will be calculated.

Figure 8 shows the basic algorithm of three kinds of DTW. The conventional DTW uses the reference speech compared with the test speech. Its algorithm is simple and fast, but the recognition accuracy is low. mDTW uses N reference speeches compared with the test speech. Although the algorithm increases the robustness of the reference speeches and the recognition accuracy is very high, the computation cost is significantly increased. The proposed method not only reduces the computation cost but maintains a high recognition accuracy (Case of training once).

5. Simulation Experiments

5.1 Evaluation measure 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 each word has 100 waveforms spoken by 50 persons. The test words are 50 isolated Japanese words, and each word has 100 waveforms spoken by another 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 ($\Delta s_i(k) = s_i(k) - s_i(k-1)$), and 12 delta-delta cepstral coefficients ($\Delta^2 s_i(k) = \Delta s_i(k) - \Delta s_i(k-1)$). Other conditions are described in Table 1.

Table 1 Experimental settings and parameters

Recognition task	Isolated 100 words
Speech data	100 Japanese region names
Sampling	11.025 kHz, 16 bits
Window length	23.2 ms (256 samples)
Frame length	11.6 ms (128 samples)
Band of bandpass filter	1-15 Hz
Feature vector	36-dimensional MFCC
Noise type	white noise and babble noise

In this study, we have two main goals. One is to reduce the calculation cost. In the following part, we will show the calculation costs of mDTW and tDTW.

To obtain the calculation cost of mDTW, we must evaluate the following cost:

$$C_{T,i}^{ID}(\mathbb{A}) = MNC_D(H_i, \mathbb{A}) + C_R(\mathbb{A}) \qquad (23)$$

where $C_{T,i}^{ID}$ is the total calculation cost of mDTW, C_D is the calculation cost of DTW, and C_R is the calculation cost of noise reduction. M is the total number of target words, and N is the total number of speeches for each speech word (in the experiment, M is 100 and N is 100). We define $\mathbb{A}$ as a feature vector of speech and H_i as the i^{th} reference feature vector.

In the case of tDTW-based ASR, the total calculation cost is

$$C_{T,i}^{TD}(\mathbb{A}) = C_T(H_i, \mathbb{A}) + \frac{1}{2t} MNC_D(H_i, \mathbb{A}) + C_R(\mathbb{A}) \quad (24)$$

where $C_{T,i}^{TD}$ is the total calculation cost of tDTW, C_T is the calculation cost of the training part, and t is number of training repetitions. We assume training of only once, then, $C_T(H_i, \mathbb{A})$ can be expressed as

$$C_T(H_i, \mathbb{A}) = \frac{1}{2} NC_D(H_i, \mathbb{A}) \qquad (25)$$

Since $\frac{1}{2}MNC_D(H_i, \mathbb{A})$ is M times $\frac{1}{2}NC_D(H_i, \mathbb{A})$, in other words, $C_T(H_i, \mathbb{A}) \ll MNC_D(H_i, \mathbb{A})$, then $C_{T,i}^{TD}(\mathbb{A}) \approx \frac{1}{2}C_{T,i}^{ID}(\mathbb{A})$. Apparently, the calculation cost of mDTW has been reduced; after training only once, the calculation cost has been reduced by almost 50%.

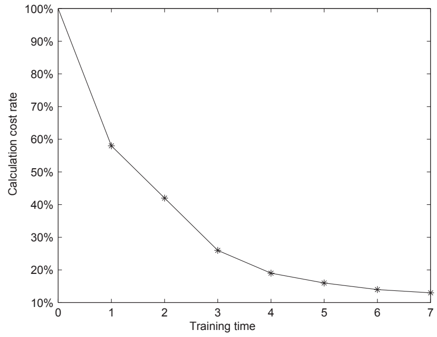

Fig. 9 Computing time of proposed DTW

Figure 9 shows the practical calculation cost rate. We used the Epson Pro7500 computer with the Core(TM) i7-3820 CPU @ 3.6 GHz. Note that zero training time represents the mDTW calculation cost, and all the calculation cost rate were compared with the mDTW calculation cost. Apparently, after training once, computing time has been reduced 41.6%. On the other hand, when the numbers of reference words becomes half, the computing time is significantly reduced.

Our other goal is to maintain a high recognition accuracy. Figure 10 shows the recognition accuracy of the two DTW algorithms with 10 dB and 20 dB white and babble noise. Our approach yields 96.94% accuracy compared with the 97.54% accuracy of mDTW in 20 dB white noise and 84.4% accuracy compared with 86.44% accuracy of mDTW in 10 dB white noise. Our approach yields 94.12% accuracy compared with 94.14% accuracy of mDTW in 20 dB babble noise and 80.82% accuracy compared with 81.64% accuracy of in 10 dB babble noise (case of training once).

Furthermore, Fig. 11 shows the tDTW recognition accuracy when the reference utterances have been trained more than once in 10 dB and 20 dB white and babble noise.

5.2 Discussion

Compared with HMM-based ASR, the merit of DTW-based ASR is less or no complicated training procedure in these system. In other words, only when speech sound can be obtained does conventional DTW start recognition immediately without training. If its merit is applied during the recognition stage, the DTW-based ASR can recognize any speech, and at the same time, always learn new speech data.

As mentioned in the previous section, the proposed tDTW-based ASR can reduce the total calculation cost compared with mDTW-based ASR, while its recognition accuracy can still be kept sufficiently high. However, un-

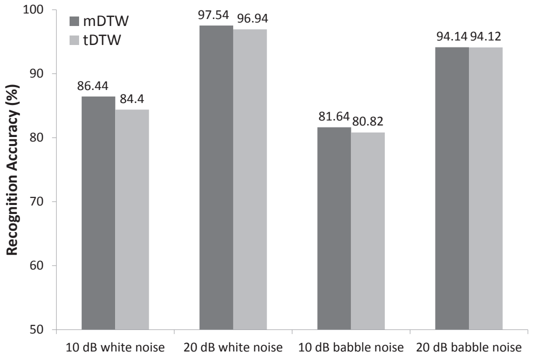

Fig. 10 Recognition accuracy of tDTW algorithms with 10 dB and 20 dB white and babble noise

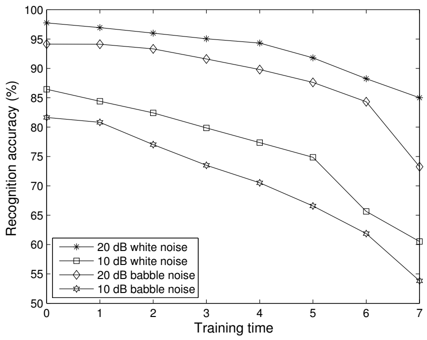

Fig. 11 Recognition accuracy of proposed DTW

fortunately, the proposed ASR requires a training procedure. This means the tDTW-based ASR cannot be applied as an on-line training ASR system, whereas any conventional DTW can be used as an on-line training ASR, i.e., speech recognition by increasing the number of references.

When the complexity of the training procedure in tDTW-based ASR is explored, as mentioned in the previous section, it is found that new references can be obtained by merging original references. On the other hand, the training of HMM-based ASR needs the parameter estimation of HMM models with convergence calculation and usually requires a high calculation cost in the training stage.

When the tDTW-based ASR is implemented into an autonomous ASR or an on-line training ASR, its training stage should also be implemented within the recognition stage in order to overcome its demerit. This will be one of the future issues.

6. Conclusion

In this paper, we proposed a robust ASR technique that includes VAD, noise-reduction, and tDTW-based recognition. It was shown in this paper that the tDTW method can reduce the computing time and memory resources. The proposed tDTW can also provide high recognition accuracy even when a training procedure is additionally required.

As mentioned in this paper, the DTW-based ASR system is valuable for small-scale systems, considering memory resources and calculation cost. The proposed system also provides a similar solution for speech recognition applications. However, the proposed system requires a training stage before it can be applied for recognition. If the target ASR includes on-line training, its training procedure should be implemented into its ASR system. This will be one of the future issues.

Acknowledgments

This study is supported in part by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (A1) (24240007), the Japan Science and Technology Agency for A-Step Program (AS2416901H), and KDDI Laboratories.

References

[1] J. Di Martin: Dynamic time warping algorithms for isolated and connected word recognition, Nato Asi Series, Vol. F16, R. De Mori and C.Y. Suen, Ed, Springer-Verlag, 1985.

[2] T. Zaharia, S. Segarceanu, M. Cotescu and A. Spataru: Quantized dynamic time warping (DTW) algorithm, 8th International Conference on Communications (COMM), pp. 91-94, June 2010.

[3] K. Yamamoto, F. Jabloun, K. Reinhard and A. Kawamura: Robust endpoint detection for speech recogntion based on discriminative feature extraction, IEEE Int. Conf. Acount. Speech and Signal Process, Vol. 1, pp. 805-808, May 2006.

[4] S. Al-Haddad, S. Samad, A. Hussain, K. Ishak and H. Mirvaziri: Decision fusion for isolated Malay digit recognition using dynamic time warping (DTW) and hidden Markov model (HMM), 5th Student Conf.on Research and Development, pp. 1-6, December 2007.

[5] G. Hongbin, P. Weiyi, H. Chunru and Z. Yongqiang: A speech endpoint detection based on dynamically updated threshold of box-counting dimension, Int. Forum on Information Technoloy and Applications, Vol. 2. pp. 397-401, May 2009.

[6] Z. Lu, B. Liu and L. Shen: Speech endpoint detection in strong noisy environmnet based on the Hilbert-Huang transform, Int. Conf. Mechatronics and Automation, pp. 4322-4326, August 2009.

[7] G. Xu, B. Tong and X. He: Robust endpoint detection in Mandarin based on MFCC and short-time correlation coefficient, Int. Conf. Intelligent Computation Technology and Automation, Vol. 2. pp. 336-339, October 2009.

[8] Q. Zhu, N. Ohtsuki, Y. Miyanaga and N. Yoshida: Robust speech analysis in noisy environment using running spectrum filtering, Int. Sym. Communications and Information Technologies, Vol. 2, pp. 995-1000, October 2004.

[9] N. Wada, N. Hayasaka, S. Yoshizawa and Y. Miyanaga: Robust speech recognition with feature extraction using combined method of RSF and DRA, Int. Sym. Communications and Information Technologies, Vol. 2, pp. 1001-1004, October 2004.

[10] N. Hayasaka, K. Khankhavivone, Y. Miyanaga and K. Songwatana: New robust speech recognition by using nonlinear running spectrum filter, Int. Sym. Communications and Information Technologies, pp. 133-136, October 2006.

[11] N. Hayasaka and Y. Miyanaga: Spectrum filtering with FRM for robust speech recognition, IEEE Int. Sym. Circuits and System, pp. 3285-3288, November 2006.

[12] 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, August 1980.

[13] 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. 43-49, February 1978.

[14] G. Kang and S. Guo: Variable sliding window dtw speech identification algorithm, Ninth International Conference on Hybrid Intelligent Systems, Vol. 1, pp. 304-307, December 2009.

[15] 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, March 2012.

Xihao Sun received his M.S. degree in computer engineering from Hokkaido University, Japan, in 2011. Currently, he is working toward a Ph.D. degree at Hokkudai University, Japan. His main research interest is robust speech recognition in noisy environments.

Yoshikazu Miyanaga received his B.S., M.S., and Dr. Eng. degrees from Hokkaido University, Sapporo, Japan, in 1979, 1981, and 1986, respectively. Since 1983 he has been with Hokkaido University. He is now Professor in the Division of Information Communication Systems, Graduate School of Information Science and Technology, Hokkaido University. From 1984 to 1985, he was a visiting researcher at the Department of Computer Science, University of Illinois, USA. He served as an Associate Editor of IEICE Transactions on Fundamentals of Electronics, Communications and Computer Science from 1996 to 1999, and Editor of IEICE Transactions on Fundamentals, Special Issues. He is also an Associate Editor of the Journal of Signal Processing, RISP Japan (2005-present). He was a delegate of IEICE, Engineering Sciences Society Steering Committee, i.e., IEICE ESS Officers from 2004 to 2006. He was a Chair of the Technical Group on Smart Info-Media System, IEICE (IEICE TG-SIS) during the same period and is now a member of the advisory committee, IEICE TG-SIS. He served as a member of the board of directors, IEEE Japan Council, as a Chair of the student activity committee, from 2002 to 2004. He is a Chair of student activity committee in IEEE Sapporo Section (1998-present). His research interests are in the areas of speech signal processing, wireless communication signal processing, and low-power VLSI system design.

Baiko Sai received his B.E. degree from Shanghai University of Science and Technology, China in 1985. He received his M.E. degree from Yokohama National University, Yokohama, Japan, in 1990, his Ph.D. degree from Hokkaido University in 2011. He worked at the Space Science and Technology, Institute of China from 1985 to 1987. He worked at Pioneer Electronic Corporation, Tokyo, Japan from 1990 to 1996. He worked at Philips Japan Ltd., Tokyo, Japan, from 1996 to 2001. He is also a Guest Professor in the Department of Computer Science and Electronics, Kyushu Institute of Technology. Since 2013, he has been with the Graduate School of Information Science and Technology of Hokkaido University as an Associate Professor. His research interests include wireless communication, digital signal processing, digital broadcasting system, high-speed interface, and information security technologies.

(Received September 18, 2013; revised January 10, 2014)