Recovering Modified Watermarked Audio Based on Dynamic Time-Warping Technique

* **Changsheng Xu, **Yusuk Lim, **David Dagan Feng
*Kent Ridge Digital Labs
21 Heng Mui Keng Terrace
Singapore 119613
**Department of Computer Science
The University of Sudney
NSW 2006 Australia

Abstract

An audio registration method based on dynamic time-warping (DTW) technique was described. DTW technique can be used to register two audio signals whose type of misalignment in time domain is unknown. By measuring the frame dissimilarities in two audio signals, the best alignment between a pair of audio signals can be obtained to register the two audio signals. This method can solve the registration problem for audio signal which is processed by re-scaling in the time domain. It is useful in many audio applications and has been applied in digital audio watermarking detection.

1. Introduction

Registration is a fundamental task in image processing used to match two or more pictures taken, for examples, at different times, from different sensors, or from different viewpoints. These images need to be aligned with one another so that differences can be detected. The key step of image registration is to find an optimal transformation to match one image with another image. The determination of the optimal transformation depends on the types of variations between the images. Currently applications of image registration mainly focus on three areas: (1) medical image analysis[1]-including diagnostic medical imaging, such as tumour detection and disease localization, and biomedical research including classification of microscopic images of blood cells, cervical smears, and chromosomes; (2) computer vision and pattern recognition[2]-for numerous different tasks such as segmentation, object recognition, shape recognition, motion tracking, stereo mapping and character recognition; (3) remotely sensed data processing[3]-for civilian and military applications in agriculture, geology, oceanography, oil and mineral exploration, pollution and urban studies, forestry and target location and identification. So far, a broad range of techniques in image registration has been developed for various types of data and problems. These techniques include correlation and sequential method[4], Fourier method[5], point mapping method[6], and elastic model-based matching method[7].

Comparing with image processing, it is difficult to find a transformation between the original audio and the processed audio suffered intentional attacks or common signal manipulations[8] since there are so fewer invariant features in an audio signal. However, audio registration is required in many audio processing applications, especially in digital audio watermarking technology[9]. For a watermarked audio signal, it will be difficult to detect the watermark if it suffers attacks such as randomly remove or add a frame and/or re-scale in time domain. In order to avoid such problems, audio registration must be performed before watermark detection.

In this paper, we use Dynamic Time-Warping (DTW) technique to make audio registration and it has been embedded in our digital audio watermarking system. The DTW technique is the primary approach taken to register two audio signals whose type of misalignment in time domain is unknown. By measuring the frame dissimilarities in two audio signals, the best alignment between a pair of audio signals can be obtained to register the two audio signals. This is functionally equivalent to finding an optimal path through a grid mapping the features of one audio frame to the features of the other audio frame. For the audio signal under consideration and the reference audio signal, they are first divided into fixed-length frames, and then the power spectral parameters in each frame are calculated using non-linear frequency scale method. An optimal path will be generated by calculating the minimum dissimilarity of relevant frames between reference audio and considered audio. The registration is performed according to this optimal path. By doing so, any possible shifting, scale, or other non-linear time domain distortion between two audio signals will be detected and relevant operations will be done to register the two audio signals.

2. DTW Technique

DTW technique uses dynamic programming to solve the time alignment and normalization between two audio signals.

Consider two audio signals F_x and F_y, represented by the frames $(x_1, x_2, \cdots, x_{T_x})$ and $(y_1, y_2, \cdots, y_{T_y})$ respectively. We use i_x and i_y to denote the time indices of F_x and F_y. T_x and T_y are the duration of two audio signals. The dissimilarity between F_x and F_y is defined by considering some function of the spectral distortion $d(x_{i_x}, y_{i_y})$, which will be denoted for simplicity of notation as $d(i_x, i_y)$ where $i_x = 1,2,\cdots,T_x$ and $i_y = 1,2,\cdots,T_y$ without ambiguity. A more general time alignment and normalization scheme involves the use of two warping functions, ϕ_x and ϕ_y, which relate the indices of the two audio signals, i_x and i_y respectively, to a common "normal" time axis k, i.e.,

$$i_x = \phi_x(k) \quad k = 1,2,\cdots,T \tag{1}$$

$$i_y = \phi_y(k) \quad k = 1,2,\cdots,T \tag{2}$$

A global frame dissimilarity measure $d_\phi(F_x, F_y)$ can be defined based on the warping function pair $\phi = (\phi_x, \phi_y)$ as the accumulated distortion over the entire audio signals, namely,

$$d_\phi(F_x, F_y) = \sum_{k=1}^{T} d(\phi_x(k), \phi_y(k)) m(k) \tag{3}$$

where $d(\phi_x(k), \phi_y(k))$ is a spectral distortion defined for $x_{\phi_x(k)}$ and $y_{\phi_y(k)}$, and $m(k)$ is a nonnegative path weighting coefficient.

To complete the definition of a dissimilarity measure for the (F_x, F_y) pair of signals, we need to specify the path $\phi = (\phi_x, \phi_y)$ as indicated in Eq.(3). There is obviously an extremely large number of possible warping function pairs. The key issue then is which path should be chosen such that the overall path dissimilarity can be measured with consistency. We define the dissimilarity $D(F_x, F_y)$ as the minimum of $d_\phi(F_x, F_y)$, over all possible paths, such that

$$D(F_x, F_y) = \min_{\phi} d_\phi(F_x, F_y) = \min_{\phi_x, \phi_y} \sum_{k=1}^{T} d(\phi_x(k), \phi_y(k)) m(k) \tag{4}$$

Similarly, the minimum partial accumulated distortion along a path connecting (1,1) and (i_x, i_y) is

$$D(i_x, i_y) = \min_{\phi_x, \phi_y, T'} \sum_{k=1}^{T'} d(\phi_x(k), \phi_y(k)) m(k) \tag{5}$$

where $\phi_x(T') = i_x$ and $\phi_y(T') = i_y$ are implied.

The dynamic programming recursion with constrains thus becomes

$$D(i_x, i_y) = \min_{(i'_x, i'_y)} [D(i'_x, i'_y) + \xi((i'_x, i'_y), (i_x, i_y))] \tag{6}$$

where ξ is the weighting accumulated distortion (local distance) between point (i'_x, i'_y) and point (i_x, i_y).

$$\xi((i'_x, i'_y), (i_x, i_y)) = \sum_{l=0}^{L_s} d(\phi_x(T'-l), \phi_y(T'-l)) m(T'-l) \tag{7}$$

with L_s being the number of moves in the path from (i'_x, i'_y) to (i_x, i_y) according to ϕ_x and ϕ_y. Again

$$\phi_x(T'-L_s) = i'_x \ , \ \phi_y(T'-L_s) = i'_y \tag{8}$$

Therefore, the DTW implementation for finding the best path through a T_x by T_y grid, beginning at (1,1) and ending at (T_x, T_y) can be summarized as follows.

(1) Initialization

$$D(1,1) = d(1,1) m(1) \tag{9}$$

(2) Recursion

For $1 \le i_x \le T_x$, $1 \le i_y \le T_y$ such that i_x and i_y stay within the allowable grid, compute

$$D(i_x, i_y) = \min_{(i'_x, i'_y)} [D(i'_x, i'_y) + \xi((i'_x, i'_y), (i_x, i_y))] \tag{10}$$

where $\xi((i'_x, i'_y), (i_x, i_y))$ is defined by Eq.(7).

(3) Termination

$$d(F_x, F_y) = D(T_x, T_y) \tag{11}$$

1. Audio Registration Algorithm

Audio registration is a fundamental task in audio processing used to match the original audio signal with the processed audio signal suffered intentional attacks such as randomly remove or add a segment in the audio signal or common signal manipulations such as re-scale audio signal in time domain.

In the registration process, DTW technique is used to register two audio signals whose type of misalignment in time domain is unknown. By measuring the frame dissimilarities in two audio signals, the best alignment between a pair of audio signals can be obtained to register the two audio signals. Fig.1 shows the registration process of two audio signals.

Feature extraction is used to calculate the feature vectors of each frame. Based on the feature measures, the frame dissimilarities between two audio signals can be obtained. In our method we use mel scale method to calculate the feature vectors of each frame. The mel scale has a simple analytical form:

$$m = 1125 \ln(0.0016 f + 1) \qquad f > 1000Hz \tag{12}$$

where f is the frequency in Hz and m is the mel scaled frequency. For $f \le 1000Hz$, the scale is linear.

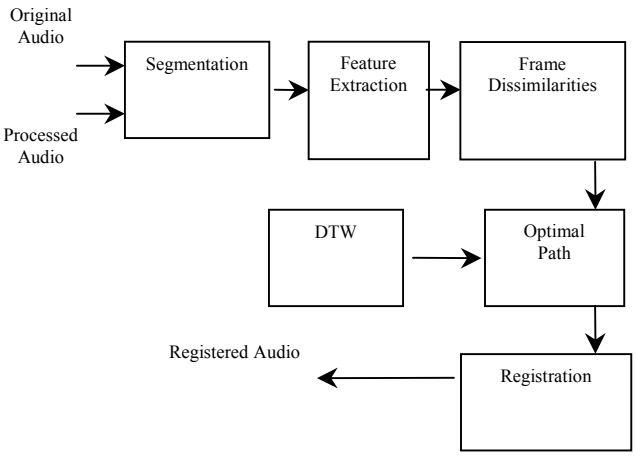

Figure 1. Audio Registration

The implementing procedure of audio registration is described as follows:

(1) For the original audio s and the processed audio s', segment them with the same fixed-length. Frame of s and s' can be expressed as $s_i\ (i=1,\cdots,m)$ and $s'_j\ (j=1,\cdots,n)$;

(2) Calculate the feature vectors of s_i and s'_j using mel scales:

$$V_i=\{v_{i1},v_{i2},\cdots,v_{il}\} \tag{13}$$

$$V'_j=\{v'_{j1},v'_{j2},\cdots,v'_{jl}\} \tag{14}$$

where l is the channel number of mel scales;

(3) Find an optimal path from $m\times n$ grid by mapping the feature vectors of s_i to the power vectors of s'_j;

(a) Initialisation:
Define local constraints and global path constraints;

(b) Recursion:
For $1\le i\le m, 1\le j\le n$ such that i and j stay within the allowable grid, calculate

$$D_{ij}=\min_{(i',j')}[D_{i'j'}+\zeta((i',j'),(i,j))] \tag{15}$$

where

$$\zeta((i',j'),(i,j))=\sum_{l=0}^{L_s} d_{i-l,j-l} \tag{16}$$

with L_s being the number of moves in the path from (i',j') to (i,j).

$$i-L_s=i',\ j-L_s=j' \tag{17}$$

$$d_{ij}=\sqrt{\sum_{k=1}^{l}(v_{ik}-v'_{jk})^2} \tag{18}$$

(c) Termination:
D_{mn}

(d) Form an optimal path from (1,1) to (m,n) according to D_{mn}:

$$P=\{p_{ij}\mid i\in[1,\cdots,m], j\in[1,\cdots,n]\} \tag{19}$$

(4) Register the watermarked audio with the original audio according to the optimal path:
For $p_{ij}\in P$
If $i<j$, add the ith frame of s to s';
If $i>j$, remove the jth frame from s'.

The feature extraction process can be described as follows:

(1) For each audio frame s_i, transform it to frequency domain using FFT;

$$S_i(j\omega)=F(s_i) \tag{20}$$

(2) Search the maximum and minimum frequency in the frequency spectrum;
$f_{\max}$, $f_{\min}$

(3) Determine the channel number n_1 and n_2, where n_1 for $f\le 1kHz$ and n_2 for $f>1kHz$;

(4) For $f\le 1kHz$, calculate the bandwidth of each band:

$$b=\frac{1000-f_{\min}}{n_1} \tag{21}$$

(5) For $f\le 1kHz$, calculate the center frequency of each band:

$$f_i=ib+f_{\min} \tag{22}$$

(6) For $f>1kHz$, calculate the maximum and minimum mel scale frequency:

$$\begin{aligned} m_{\max}&=1125\ln(0.0016f_{\max}+1)\\ m_{\min}&=1125\ln(0.0016\times 1000+1)\end{aligned} \tag{23}$$

(7) For $f>1kHz$, calculate the mel scale frequency interval of each band:

$$\Delta m=\frac{m_{\max}-m_{\min}}{n_2} \tag{24}$$

(8) For $f>1kHz$, calculate the center frequency of each band:

$$f_i=(\exp((i\Delta m+1000)/1125)-1)/0.0016 \tag{25}$$

(9) For $f>1kHz$, calculate the bandwidth of each band:

$$b_i=f_{i+1}-f_i \tag{26}$$

(10) For each center frequency and bandwidth, determine a triangle window function shown in Fig.2:

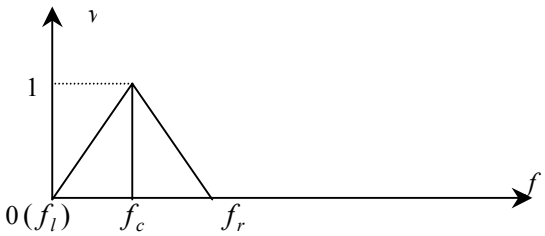

Figure 2. Triangle Window Function

$$w=\begin{cases}\dfrac{1}{f_c-f_l}f-\dfrac{f_l}{f_c-f_l} & f_l\le f\le f_c\\ \dfrac{1}{f_c-f_r}f-\dfrac{f_r}{f_c-f_r} & f_c\le f\le f_r\end{cases} \quad (27)$$

Where f_c, f_l, f_r are the center frequency, minimum frequency and maximum frequency of each band.

(11) For each band, calculate its spectral power and generate the feature vector:

$$\vec{V}=\{P_1,P_2,\cdots,P_{n_1+n_2}\} \qquad P_i=\sum_{j=f_l}^{f_r} w_j s_j \quad (28)$$

Where s_j is the spectrum of each frequency band.

2. Experiments and Application

This proposed audio registration method has been applied to our digital audio watermarking system. To prevent from attackers who may re-scale audio signal in time domain, and/or randomly remove or add a frame of audio signal, audio registration must be performed before watermark extraction. The watermark embedding and extracting scheme is shown in Fig.3.

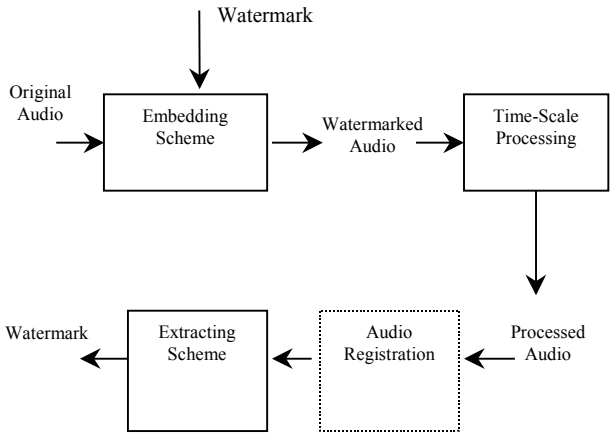

Figure 3. Watermark Embedding and Extracting Scheme

In order to further illustrate the registration algorithm, a watermarked audio and its processed version are applied to make a test. Fig.4 shows the watermarked audio with 22.05kHz sampling rate and 8-bit quantization. For the watermarked audio, we randomly remove a frame from it, and then randomly add a frame into it. The processed audio is shown in Fig.5. On the basis of the watermarked audio and the processed audio, we can obtain the registered audio according to our registration algorithm. Fig.6 shows the registered audio. It can be seen from the Fig.4 and Fig.6 that the result is quite good.

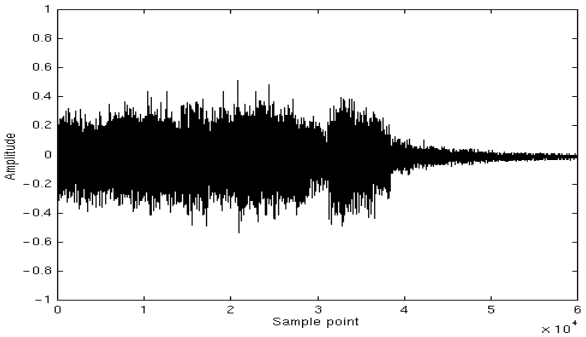

Figure 4. Watermarked Audio Signal

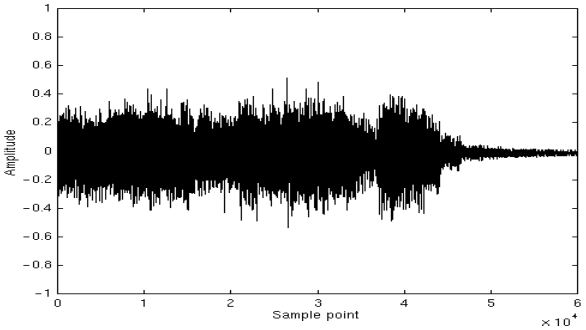

Figure 5. Processed Audio Signal

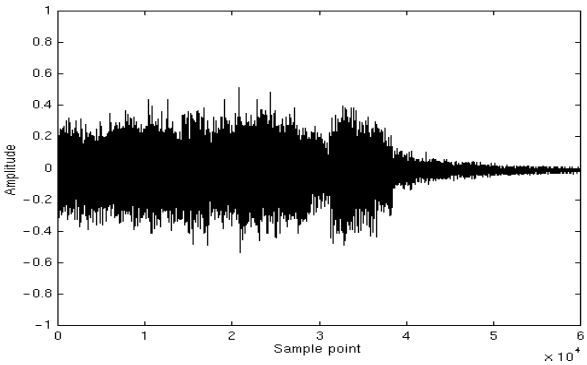

Figure 6. Registered Audio Signal

3. Conclusion

Audio registration is a fundamental task in audio processing used to match two audio signals whose type of misalignment in time domain is unknown. In order to implement it, we proposed a dynamic time-warping registration method and embed it in our digital audio

watermarking system. According to this method, an optimal path will be generated by calculating the minimum dissimilarity of relevant frames between reference audio and considered audio. And according to this optimal path, any possible shifting, scale, or other non-linear time domain distortion between two audio signals will be detected and relevant operations will be done to register the two audio signals.

References

[1] M.Herbin, A.Venot, Y.Devaux, E.Walter, F.Lebruchec, L.Dubertret, and C.Roucayrol, Automated registration of dissimilar images: Application to medical imagery, Comput. Vision Graph. Image Process., 47(1):77-88, 1989.

[2] Y.Bresler, and J.Merhav,, Recursive image registration with application to motion estimation, IEEE Trans. Acoust. Speech Signal Proc., ASSP-35(1):70-85, 1987.

[3] M.Haralick, Automatic remote sensor image registration, Topic in Applied Physics, Vol.11, Digital Picture Analysis, A. Rosenfeld, Ed. Springer-verlag, New York, pp.5-63, 1989.

[4] A.Rosenfeld, and C.Kak, Digital Picture Processing, Academic Press, Oriando, Fla., 1982.

[5] E. De Castro, and C.Morandi, Registration of translated and rotated images using finite Fourier Transforms, IEEE Trans. Patt. Anal. Machine Intell. PAMI-9(5):700-703, 1987.

[6] J.Ton, and K.Jain, Registering Landsat images by point matching, IEEE Trans. Geosci. Remote Sensing, 27(6): 642-651, 1989.

[7] M.Mehran, R.Surendra and N.Ken, Three-dimensional elastic matching of volumes, IEEE Trans. Image Processing, 3(2):128-138, 1994.

[8] A.Fabien, R.Anderson and M.Kuhn, Attacks on copyright marking systems, in vol. 1525 of Lecture Notes in Computer Science Portland, Oregon, USA, 14--17 April, 1998, pp. 218--238.

[9] M.Swanson, B.Zhu and A.Tewfik, Current State of the art, challenges and future directions for audio watermarking, IEEE International Conference on Multimedia Computing and Systems, Florence, Italy, 7-11 June, 1999, pp.19-24.

[10]J.Flusser, An adaptive method for image registration, Pattern Recognition, 25(1): 45-54, 1992.

[11]Y.Ohta, K.Takano and K.Ikeda, A high-speed stereo matching system based on dynamic programming, Proceedings of the International Conference in Computer Vision, London, England, 1987, pp.335-342.

[12]G.Medioni and R.Nevatia, Matching images using linear features. IEEE Trans. Patt. Anal. Machine Intell. PAMI-6, pp.675-685, 1984.