Journal of Zhejiang University SCIENCE
ISSN 1009-3095
http://www.zju.edu.cn/jzus
E-mail: jzus@zju.edu.cn

Adaptive audio watermarking based on SNR in localized regions*

WU Guo-min (吴国民)†, ZHUANG Yue-ting (庄越挺)†‡, WU Fei (吴 飞)†, PAN Yun-he (潘云鹤)

(*School of Computer Science, Zhejiang University, Hangzhou 310027, China*)

†E-mail: wuguomin@263.net; yzhuang@cs.zju.edu.cn; wufei@cs.zju.edu.cn

Received Nov. 12, 2004; revision accepted Jan. 12, 2005

Abstract: In this paper, a novel localized audio watermarking scheme based on signal to noise ratio (SNR) to determine a scaling parameter α is proposed. The basic idea is to embed watermark in selected high inflexion regions, and the intensity of embedded watermarks are modified by adaptively adjusting α. As these high inflexion local regions usually correspond to music edges like sound of percussion instruments, explosion or transition of mixed music, which represent the music rhythm or tempo and are very important to human auditory perception, the embedded watermark is especially expected to escape the distortions caused by time domain synchronization attacks. Taking advantage of localization and SNR, the method shows strong robustness against common problems in audio signal processing, random cropping, time scale modification, etc.

Key words: Copyright protection, Audio watermarking, Time scale modification, MDCT
doi:10.1631/jzus.2005.AS0053 **Document code:** A **CLC number:** TP309.7

INTRODUCTION

Digital watermarking is a process by which a user-specified signal (watermark) is hidden or embedded into another signal (cover data) (Jin *et al*., 2004). The highly successful MPEG-2 layer 3 (MP3) audio coding and the rapid growth of the Internet enable the promising marketing of on-line music worldwide distribution. Consequently, it is necessary to protect the copyright of digital audio information. Digital watermarking or data hiding is one solution for this kind of protection.

At present, there are few algorithms that can effectively resist the time scale synchronization attack, which is a serious problem especially for audio watermarking. Mansour and Tewfik (2001a; 2001b) proposed to embed watermark by changing the relative length of the middle segment between two successive maximum and minimum of the smoothed waveform or by changing the interval lengths between salient points in the signal. Both performances depend highly on the selection of appropriate threshold, which is difficult work. Tachibana *et al.*(2001) proposed a method to calculate and manipulate the magnitudes of segmented areas in the time-frequency plane, which is robust against random stretching up to ±4%. Tachibana (2002) further improved the performance up to ±8% by using multiple pseudo-random arrays. The above mentioned methods can be made robust to time scaling modification, but they have the same limitation in that they all heavily rely on adjusting threshold or some assumed coefficients, which make them difficult to apply to different kinds of music.

In this paper, we present a novel localized robust audio watermarking algorithm based on the SNR. The main idea is to embed watermark in MDCT coefficients of high inflexion regions, and the intensity of embedded watermark is determined by the SNR. High inflexion regions, which generally represent music transition like a piano entrance after the orchestra in a

‡Corresponding author

*Project supported by the National Natural Science Foundation of China (No. 60272031), the National Basic Research Program (973) of China (No. 2002CB312101), Zhejiang Provincial Natural Science Foundation (No. ZD0212), Doctorate Research Foundation of the State Education Commission of China (No. 20010335049), and the State Education Commission Foundation for Scholars Returned from Overseas, China

concerto, a rock guitar solo, a change of speaker, etc., or sound of percussion instruments like drum, bell and xylophone or explosion, are closely related to music rhythm or tempo and are very important to human auditory perception. To maintain high auditory quality, such regions should be left unchanged or altered very little under different kinds of modification including random cropping, which often takes place in less important parts outside of these important regions. So watermark embedded in these regions will show natural resistance to many audio distortions, especially those time domain synchronization attacks. Usually, a scaling parameter α is proposed to determine watermark embedded intensity. This scheme has two disadvantages. First, to ascertain a proper α is inefficient and to get the best α is almost impossible. Second, for an ascertained α, the intensity is invariable, which can cause excessive watermark energy at some places and weak watermark energy at other places. Therefore we propose to use SNR method to determine the scaling parameter α on different audio signal segments. Compared with the method proposed by Wang *et al.*(2004), which also embeds watermark in modified discrete cosine transform (MDCT) without using localization and SNR, our algorithm can be expected to perform more robustly and inaudibly.

The main procedure of our algorithm is to produce watermarked MP3 files during the process of MP3 compression, which is separate from the MP3 compression implemented for the robustness experiment. The embedding result is a slight modification of MDCT coefficients in a way that does not produce any perceived effect. The algorithm first calculates the audio features during MP3 compression, after which a scheme is designed to select appropriate regions for watermark embedding according to the extracted features. Then the embedding intensity is determined by SNR theory; for a predefined threshold R_{SNR}, the scaling parameter α is adaptively adjusted by the audio signals. Finally, the watermark w is embedded into the region according to the dynamic scaling parameter α given in Eq.(1), where x is the original signal, y is the watermarked signal.

$$y=x+\alpha w \tag{1}$$

EMBEDDING REGIONS SELECTION

Feature extraction

During MP3 compression, layer 3 filterbank is used to translate time domain into frequency domain, after which, the original 576 music signals will be translated into 576 MDCT coefficients, the value of which represents the bank energy for 576 subbands. The analysis is performed on blocks of 576 samples (about 25 ms at 22050 Hz, mono) that correspond to one MPEG audio frame. For each frame, a root mean squared (*RMS*) subband vector is extracted as:

$$M[i]=\sqrt{\frac{\sum_{t=1}^{18}(S_t[i]^2)}{18}}, \quad i=1,...,32.$$

S_t is the 32-demensional subband vector. M is a 32-dimensional vector describing the spectral content of sound for that frame. Then we can calculate the following four features (Tzanetakis and Cook, 2000).

(1) *Centroid* is the balancing point of the vector. It can be calculated using:

$$C=\sum_{i=1}^{32} iM[i] \Big/ \sum_{i=1}^{32} M[i].$$

(2) *Rolloff* is the value R, the rate at which a frequency response curve decreases by 3 dB, such that:

$$R=\arg\left(\sum_{i=1}^{R} M[i]=0.85\sum_{i=1}^{32} M[i]\right).$$

(3) *Spectral Flux* is the 2-norm of the difference between normalized M vectors evaluated at two successive frames.

(4) *RMS* is a measure of the loudness of the frame. As the high inflexion region often occurs at the edge of a sudden loud sound, it is a very important feature in region selection, and is denoted as:

$$RMS=\sqrt{\frac{\sum_{i=1}^{32}(M[i])^2}{32}}.$$

The above four features reflect the static and dynamic characteristics of the frame, and are used for describing the audio frame.

Embedding regions selection

(1) The feature vector f_t calculated for each frame includes *Centriod*, *Rolloff*, *Spectral Flux*, and *RMS*.

(2) A Euclidean distance d_t is calculated between successive frames of sound. In order to prevent the resulting distance being too larger, we use logarithm to zoom the distance. The resulting formula is:

$$d_t = \log[(f_t - f_{t-1})^{\mathrm{T}}(f_t - f_{t-1}) + 1] \qquad (2)$$

(3) The mean difference $\mathrm{d}f_t$ is calculated between the prior k+1 length window and the next k+1 length window.

$$\mathrm{d}f_t = \frac{1}{k+1}\left|\sum_{i=t-k}^{t} d_i - \sum_{i=t}^{t+k} d_i\right| \qquad (3)$$

(4) The $\mathrm{d}f_t$ will be low for smoothly changing textures and high during sudden transitions. The peaks roughly correspond to the 'texture' changes of sound. And the larger the peak value, the more suitable is the region for embedding and detecting watermark, so in later embedding procedure, we always selected the 15 largest peaks for embedding regardless of the total length of the original audio, if there exist 15 peaks.

THEORY OF USING SNR TO DETERMINE EMBEDDING INTENSITY

Generally, the human auditory system is more sensitive to distortions than the visual system thus making the generation of imperceptible audio watermarks a challenging task. Embedding watermark into the audio signal corresponds to adding inaudible noise to the audio signal, so SNR method can be adopted to determine watermark embedding intensity.

Suppose $X(i)$ is MDCT coefficient of a given audio frame, i is the number of coefficients, $Y(i)$ is the watermarked signal, w and α are watermark and embedding intensity. We can calculate SNR by

$$R_{SNR} = 10\log\frac{\sum_i X^2(i)}{\sum_i [X(i)-Y(i)]^2} \qquad (4)$$

Based on Eqs.(1), (4) and the watermark bit $w\in\{1,-1\}(0\rightarrow-1,1\rightarrow+1)$ compute the scaling α:

$$\alpha = \sqrt{\left(\sum_i X^2(i)\right)10^{-R_{SNR}/10}} \qquad (5)$$

For a predefined threshold R_{SNR}, the scaling parameter α can be adaptively adjusted by audio signals. A similar approach can be widely used in any other watermarking algorithm.

WATERMARK EMBEDDING

First, according to Eq.(3), all magnitude peaks of the original audio are calculated. Let *nPeakNum* be the number of all detected peaks, if *nPeakNum* exceeds 15, then we always select the 15 largest peaks for embedding; if *nPeakNum* is less than 15, then all the peak regions are selected for embedding.

Next, the watermark adopted in out experiment is a 384-bit pseudorandom sequence, $W=\{w(i)|w(i)\in\{-1,1\},1\leq i\leq 384\}$. Experimental results showed that a 384-bit watermark maintains high inaudibility, while a 512-bit or bigger watermark may introduce some distortions, that is, exceeding the nonzero 576-coefficient of embedding regions.

Finally, the watermark sequence W is repeatedly embedded into all the selected high inflexion regions according to Eq.(6) (Cox *et al.*, 1997). In the same way, for large quantity watermark, the watermark can be divided into sub watermark in length of 384-bit and embedded into the successive selected regions repeatedly, corresponding to a large *nPeakNum.*

$$Y_k(i) = X_k(i) + \alpha(k)w(i) \qquad (6)$$

where $\alpha(k)$ is the kth high inflexion region scaling, which is calculated by Eq.(5); $X_k(i)$ is the original MDCT coefficient; $Y_k(i)$ is the watermarked signal.

WATERMARK DETECTION

This detection scheme needs the original audio signal. The detailed detection steps are described as follows:

Step 1: The same method with embedding is used to calculate all magnitude peaks of the original audio signal. Let *omaxnum* be the number of all detected peaks; if *omaxnum*>15 then *omaxnum*=15. And the top *omaxnum* large peak regions are selected in descending order from *oregion*[1], *oregion*[2]... to *oregion*[*omaxnum*].

Step 2: The same method used in step 1 is applied to the watermarked and may be attacked audio signals. Let *wmaxnum* be the number of all detected peaks; if *wmaxnum*>15 then *wmaxnum*=15. And the top *wmaxnum* large peak regions are selected for detection in descending order from *wregion*[1], *wregion*[2]... to *wregion*[*wmaxnum*].

Step 3:

```
for $i$=1 to $wmaxnum$        // (matching $wregion[i]$ with an
                              // original $oregion[j]$) loop 0
  for $j$=1 to $omaxnum$      // loop 1
    {
    if ($oregion(j)$ has matched with some $wregion$)
      continue;
    // (select some unmatched original $oregion$ to match
    // with the detection region $wregion[i]$, if $oregion[j]$
    // has matched earlier with some $wregion$, then
    // ignore the $oregion$ and select the next $oregion$
    // for matching)
    Eq.(5), with a predefined threshold $R_{SNR}$, is used to
    calculate the scaling parameter $\alpha$ of $oregion[j]$;
    for $k$=1 to 384    // get the watermark in the detection
                        // region $wregion[i]$
      {
      Together with the original signal $X(k)$ in $oregion[j]$,
      the watermarked signal $Y[k]$ in $wregion[i]$,
      and the scaling parameter $\alpha$, the watermark is
      retrieved from IMDCT coefficients as Eq.(7);

      $w'(k)=(Y(k)-X(k))/\alpha$                          (7)
      }
    Put the extracted sequence into pseudo-random inversed order,
    and obtain the final watermark $W^*$ in
    the detection region $wregion[i]$;
    Do similar test between the detected watermark $W^*$
    and the original watermark $W$ by $Sim(W^*,W)$ (Cox et
    al., 1997. Threshold=6);
    $result[i]=Sim(W^*,W)$;   //save the detection result in
                              // result list $result[i]$
    if (the watermark exists in the detection region)
      // which means $Sim$>6
      {
      Mark the $oregion[j]$ that has matched with some
      $wregion$;
      break;    // exit the loop 1, go to next detection
                // $wregion$
      }
    }               // end of loop 1
return $result$;    // return the full detection $result$.
```

EXPERIMENTAL RESULTS

The algorithm was applied to a set of audio signals including piano, rock, pop and electronic organ or a mixed music of them (20 s, mono, 16 bits/sample 22050 Hz). We selected audio signal quality threshold R_{SNR}=32 to calculate the scaling α. The waveform of the original and watermarked music is shown in Fig.1, SNR=39.8 dB, which means almost no distortions have been introduced and there is no obvious difference by using listening test.

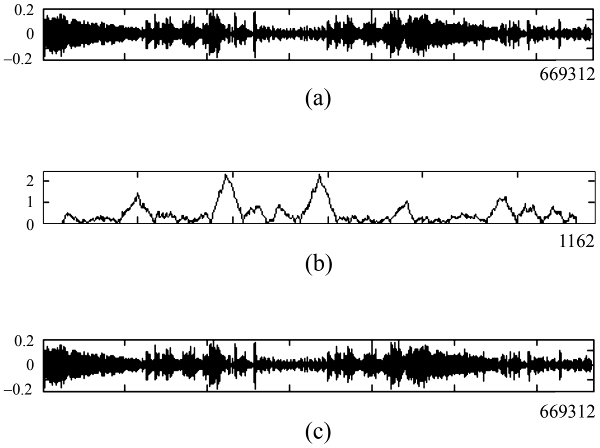

Fig.1 (a) The original waveform of mixed music; (b) The result of d*f* in time domain. The 15 largest peaks are selected out for embedding watermark; (c) The waveform of watermarked music

We tested the robustness of the algorithm by a series of audio distortions including common audio signal processing, random cropping and time scale modification. The experimental conditions are: (1) Downsampling is decreasing the rate from 22.05 kHz to 11.025 kHz, and increasing it to 22.05 kHz by interpolation technique. Upsampling is from 22.05 kHz to 44.1 kHz and return to 22.05 kHz by decimation technique; (2) Echo addition with a delay of 100 ms and decay of 50%, respectively; (3) Low pass filtering (LPF) using a second-order Butterworth filter, cutoff frequency of 6 kHz; (4) MP3 compression using 32, 48, 64, and 96 kbit/s, with the procedures being separated from the watermark embedding compression; (5) White noise with 1% of the power of the audio signal was added; (6) Colored noise was added by a Gaussian white noise shaped by low pass filtering; (7) Denoise by using DWT to carry out global threshold noise removing; (8) Random cropping 20% of the audio signal at each of 16 samples; (9) Time scale modification from –18% to 18% of the total audio excerpt length. Pitch-invariant time scale modification is a challenging problem in audio watermarking; it can be viewed as a special form of random cropping, removing or adding some parts of audio signal while preserving the pitch.

Watermark detection results after the attacks described above are shown in Table 1. The algorithm showed very high detection performance in audio signal processing, random cropping and time scale modification. Especially for time scale modification, the algorithm shows strong robustness to this attack up to at least ±12%, which is mainly due to the properties of high inflexion regions that are naturally strongly resistant to such attacks. In the case of detection failure or the result cannot be accepted, the auditory quality is also distorted severely.

Table 1 RCDR (Ratio of Correctly Detected Regions), Sim (Cox *et al.*, 1997. Threshold=6), BER of watermarked audio signal under series audio distortions (–: all region detections failed)

Attack type	RCDR	*Sim*	*BER*
Unattacked	15/15	19.5959	0
Downsampling	11/15	19.5959	0
Upsampling	13/15	19.5959	0
Echo, 100ms, 50%	10/15	19.4939	2.6×10^{-3}
LPF, 6 kHz	11/15	19.5959	0
MP3 comp, 32 kbit/s	4/15	19.0856	1.3×10^{-2}
MP3 comp, 48 kbit/s	8/15	19.2897	7.81×10^{-3}
MP3 comp, 64 kbit/s	13/15	19.4939	2.6×10^{-3}
MP3 comp, 96 kbit/s	13/15	19.5959	0
White noise, 1%	6/15	19.4939	2.6×10^{-3}
Colored noise, audible	2/15	19.3918	5.21×10^{-3}
Denoise, DWT	1/15	16.1258	8.85×10^{-2}
Crop, 20%, 6 samples	11/15	19.5959	0
TSM–2%	13/15	19.5959	0
TSM–6%	9/15	19.5959	0
TSM–10%	7/15	19.5959	0
TSM–12%	6/15	19.5959	0
TSM–14%	3/15	18.9835	1.56×10^{-2}
TSM–18%	0/15	–	–
TSM+2%	12/15	19.5959	0
TSM+6%	10/15	19.5959	0
TSM+10%	7/15	19.5959	0
TSM+12%	5/15	19.5959	0
TSM+14%	4/15	19.1877	1.04×10^{-2}
TSM +18%	1/15	12.5536	1.8×10^{-1}

CONCLUSION

A novel audio watermarking algorithm that embeds a watermark in MDCT coefficients of high inflexion regions using SNR method is presented. The selection of high inflexion regions and the exact region matching in detection are the most crucial steps in this algorithm. Detection results showed high robustness against common audio signal processing, random cropping and time scale modification. Our future work is to find more suitable regions to further improve the robustness against more audio signal processing attacks and exert more effort in blind watermark detection.

References

Cox, I.J., Kilian, J., Leighton, T., Shamoon, T., 1997. Secure spread spectrum watermarking for multimedia. *IEEE Transactions on Image Processing*, **6**(12):1673-1687.

Jin, J.Q., Dai, M.Y., Bao, H.J., Peng, Q.S., 2004. Watermarking on 3D mesh based on spherical wavelet transform. *Journal of Zhejiang University SCIENCE*, **5**(3): 251-258.

Mansour, M., Tewfik, A., 2001a. Time-Scale Invariant Audio Data Embedding. IEEE International Conference on Multimedia and Expo, Tokyo, Japan.

Mansour, M., Tewfik, A., 2001b. Audio Watermarking by Time-Scale Modification. IEEE International Conference on Acoustics, Speech and Signal Processing, Tokyo, Japan.

Tachibana, R., 2002. Improving Audio Watermarking Robustness Using Stretched Patterns against Geometric Distortion. 3rd IEEE Pacific-Rim Conference on Multimedia (PCM2002), (LNCS), **2532**:647-654.

Tachibana, R., Shimizu, S., Nakamura, T., Kobayashi, S., 2001. An Audio Watermarking Method Robust against Time and Frequency Fluctuation. SPIE Conference on Security and Watermarking of Multimedia Contents 3, **4314**: 104-115.

Tzanetakis, G., Cook, P., 2000. Sound Analysis Using MPEG Compressed Audio. Icassp, Istanbul, p.2-761.

Wang C.T., Chen, T.S., Chao, W.H., 2004. A New Audio Watermarking Based on Modified Discrete Cosine Transform of MPEG/Audio Layer 3. IEEE International Conference on Networking Sensing & Control, Taiwan, China, p.21-23.