An Adaptive Audio Watermarking System

Say Wei FOO, *Senior Member, IEEE*, Theng Hee YEO and Dong Yan HUANG

***Abstract*-- In this paper, an adaptive and content-based audio watermarking system based on echo hiding is presented. The proposed system aims to overcome the problems of audible echoes in simple echo hiding approach. Audibility of echoes is reduced especially for problematic signals by the applications of signal based attenuation, psycho-acoustic model and perceptual filter to adaptively modify the decay rate used in encoding the watermark.**

The proposed method is robust to common signal processing operations of noise addition, re-sampling, cropping, filtering and MPEG coding. If inaudibility and detection accuracy are of paramount importance, psycho-acoustic model coupled with single echo is the most appropriate encoding option to use when additional signal processing is imposed.

The application of multiple echo hiding is also investigated. It is found that although multiple echo hiding improves inaudibility in general, detection accuracy is compromised when additional signal processing is imposed.

Index Terms: **Echo-hiding method, Perceptual filter, Psycho-acoustic model, Watermarking.**

I. INTRODUCTION

DIGITAL watermarking, a method of data hiding, has been proposed as a means for owner identification of digital data. Unlike cryptographic techniques such as encryption, data hiding does not restrict or regulate access to the host signal, but ensures the embedded data remain inviolate and recoverable [1]. Some applications of watermarking are broadcast monitoring, owner identification, proof of ownership, authentication, transactional watermarks (fingerprinting), copy control, and covert communication [2].

To be effective in the protection of the ownership of intellectual property, the audio watermark should be inaudible even to golden ears. The watermark should be robust to manipulation and common signal processing operations such as filtering, re-sampling, adding noise, cropping, digital-to-analog and analog-to-digital conversions and lossless / lossy compression. It should be tamperproof — resistant to active, passive, collusion and forgery attacks.

Say Wei FOO and Theng Hee YEO are with the department of electrical and computer engineering, National University of Singapore.

Dong Yan HUANG is with the Institute of Microelectronics, Singapore.

On the other hand, watermark detection should unambiguously identify the owner.

Echo hiding is a method of audio watermarking for copyright protection. However, the simple echo hiding approach has several problems. One of which is the audibility of the echoes. For speech and musical signals of high dynamic range and when the watermarked signals are subject to further processing, the accuracy of detection of the embedded watermark is also less than perfect.

In this paper, an adaptive and content-based audio watermarking system based on single and multiple echo hiding are investigated. In the next two sections, the design of the encoder and decoder of the system are described. Experimental results are given in Section 4 in terms of audibility, computational efficiency and detection accuracy for a wide assortment of signal processing operations and distortions. The conclusion is presented in Section 5.

II.ENCODER DESIGN

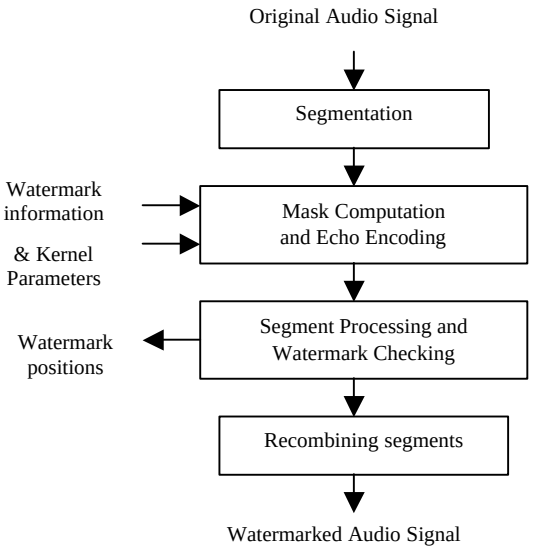


Fig.1. Encoder block diagram

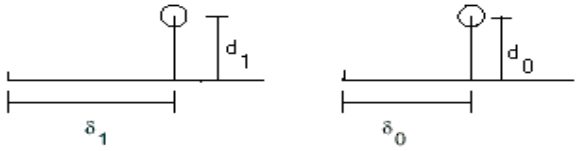
The block diagram of the encoder is shown in Fig.1. The operations of the various blocks are explained in detail in the following sub-sections.

A. Segmentation

The original audio is first divided into segments of 0.25s duration. The segment that satisfies the given criteria will be coded with one bit of the watermark.

IEEE Catalogue No. 01CH37239
0-7803-7101-1/ 01 / $10.00 © 2001 IEEE.

The single echo kernels and the multiple echo kernels are depicted in Fig. 2 and Fig.3 respectively.

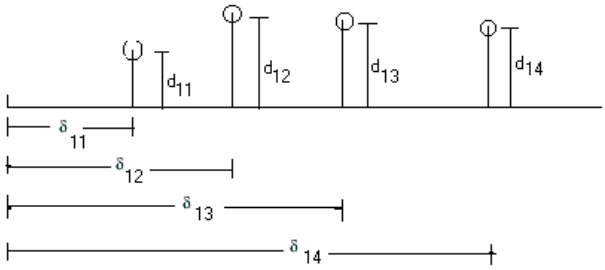


(a) "One" kernel (b) "Zero" kernel

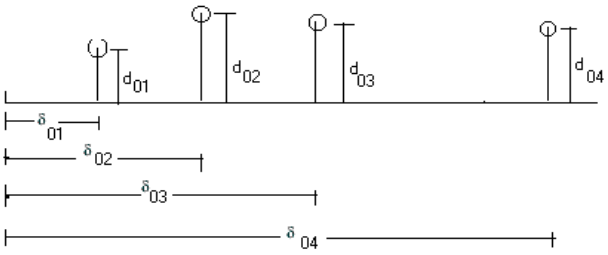
Fig. 2. Single echo kernels

The parameters used for the single echo kernels are given below:

- δ_1 = 0.001s (Delay for "one" kernel)
- δ_0 = 0.0013s (Delay for "zero" kernel)
- d_1 = 0.5 (Decay rate for "one" kernel to be adaptively modified by masking)d_0 = 0.5 (Decay rate for "zero" kernel to be adaptively modified by masking)



(a) "One" kernel



(b) "Zero" kernel

Fig 3. Multiple echo kernels

For multiple echo hiding investigated , 4 impulses are used to create 4 echoes as shown in Fig 3. The parameters used for the echo kernels are:

- δ_{11} = 0.00104s, δ_{12} = 0.00183s, δ_{13} = 0.00220s, δ_{14} = 0.00262s (Delays for "one" kernel)
- δ_{01} = 0.00127s, δ_{02} = 0.00133s, δ_{03} = 0.00136s, δ_{04} = 0.00238s (Delays for "zero" kernel)
- d_{11} = 0.15, d_{12} = 0.45, d_{13} = 0.40, d_{14} = 0.50 (Decay rates for "one" kernel to be adaptively modified by masking)d_{01} = 0.45, d_{02} = 0.35, d_{03} = 0.20, d_{04} = 0.35 (Decay rates for "zero" kernel to be adaptively modified by masking)

The parameters for echo kernels are determined using the Multi-objective Evolutionary Algorithm (MOEA) [7], an efficient search algorithm.

B. Mask Computation and Echo Encoding

Different segments have different levels of energy, the 'quiet' segments have extremely low level of energy. A constant decay rate for all audio segments will result in failure of detection for segments that consist of decaying signals or are simply problematic for one reason or another. As such, assessment is made of each segment to determine the maximum decay rates of the impulses in the kernel for the echoes.

The signal energy for a segment is computed using the formula below.

$$E_{rms} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} s_i^2} \qquad (1)$$

The value so obtained is compared with a threshold. Only segments with signal energy higher than the threshold are coded with the watermark bit.

Next, a kind of automatic gain control is carried out. The signal of selected segment is amplified to have at least desired amplitude of 0.95. After which, the decay rate is adaptively determined. That is, the echoes are set to values below a mask. One of the following methods is used to compute the mask: 1)Signal Dependent Attenuation, 2)Psycho-acoustic Model and 3)Perceptual Filter.

Signal based adaptation is the most primitive way of computing the mask, the mask being the original host signal. For each segment, the decay rates of the echo kernel (d_1 and d_0 for single echo kernels) are adjusted such that most of the echo components are below the original host wave components.

MPEG Psycho-acoustic Model 1 is another way of computing the mask. The mask is computed in the frequency domain for 26 critical bands. For each segment, decay rate is adjusted such that the frequency components of the echo from 1kHz to 5kHz are below the mask. An example is shown in Fig.4.

The principal steps of computing the mask are described for a signal with 32 kHz sampling rate in [5], but sampling rates of 44.1 kHz and 48 kHz are also supported.

Perceptual filter is another way of computing the mask. This is carried out in the time domain based on Linear Predictive Coding (LPC). The mask is computed by passing the original host signal through a perceptual weighting filter with the following transfer function

$$W_c(z) = C(z)/C(z/a), \qquad (2)$$

where a = 0.95 and $C(z) = 1 - \sum_{i=1}^{p} c_i z^{-i}$.

The optimal estimation of autoregressive (AR) model parameters, $\{c_i,\ 1 \le i \le p\}$, are calculated by solving the least squares Yule-Walker equation [6].

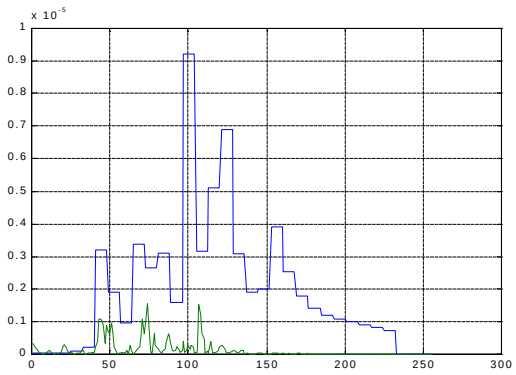


Fig. 4. Mask using psycho-acoustic model

C. Watermark checking

To ensure that the watermark can be decoded, decoding is carried out on the watermarked segment. If the peaks corresponding to the echo kernel being used are readily detected, the watermarked segment is accepted and the segment number is registered; otherwise, this segment is sipped and the watermark is used for the next selected segment. Information of the positions of the watermarked segments will be transmitted to the decoder. Thus the system belongs to the half-blind decoding scheme.

D. Recombining all segments

Finally, all segments, including the watermarked segments, are then combined in the temporal order of occurrence.

III.DECODER DESIGN

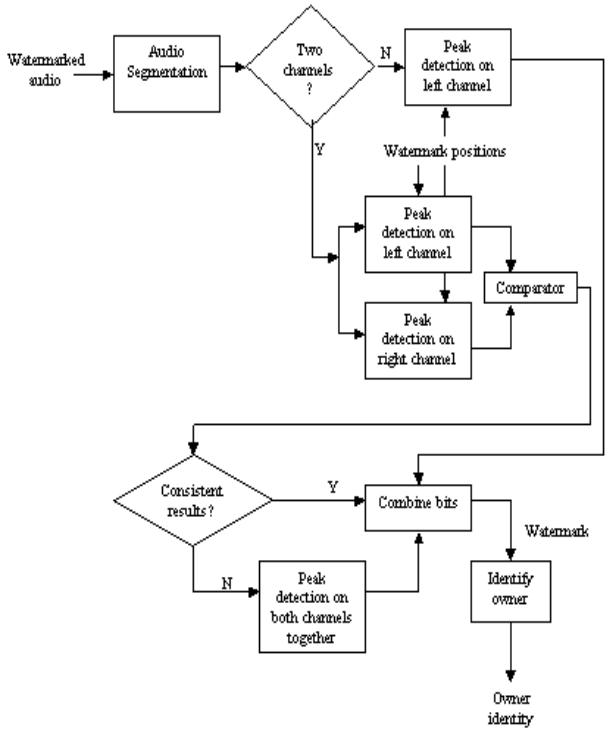


Fig. 5. Decoder block diagram

As illustrated in the decoder block diagram above, peak detection is performed on the watermarked segments using the information on watermark positions. The watermark bit stream is then recombined into a signature (watermark) and correlated with all owners in the database. The owner is identified as the one with the highest correlation.

Peak detection involves the computation of autocepstrum on the audio segments stored in the vector of watermark positions. For single echo hiding, the magnitude of the autocepstrum is examined at the two locations corresponding to the delays of the "one" and "zero" kernel respectively. If the autocepstrum is greater at d_1 than it is at d_0, it is decoded as "one". For multiple echo hiding, all peaks present in the autocepstrum are detected. The number of peaks corresponding to the delay locations of the "one" and "zero" kernels are then counted and compared. If there are more peaks at the delay locations for the "one" echo kernel, it is decoded as "one."

IV.EXPERIMENTS AND RESULTS

The audio quality and robustness of the various encoding options for echo hiding are tested. The following five audio pieces are used. 1) Pop song (PopSong, stereo), 2) Music played in saxophone (SaxoMusic, stereo), 3)Chinese Er Hu piece (ErHuMusic, stereo), 4)A capella song by Suzanne Vega (ACapSong, mono) and 5)Speech (SpeechClip, mono).

PopSong has very high signal energy in both channels while SaxoMusic and ErHuMusic have moderate signal energy. ACapSong has low signal energy and contains noticeable periods of silence. The SpeechClip, as in most speech, consists of a high percentage of silence intervals.

The audio quality of the watermarked signals is evaluated through listening tests. The listeners were presented with the original signals and the watermarked signals in random and asked to indicate their preference.

As for the detection accuracy, it is defined as

$$\text{detection accuracy} = \frac{(\text{no of bits correctly decoded}) \times 100}{\text{no of bits placed}}$$

The detection accuracies for various options are determined for the following situations:

1. *Immediately after encoding (closed-loop)*
2. *After adding noise*
3. *After re-sampling*: The watermarked audio was down-sampled, and then up-sampled.
4. *After cropping:* Five short pieces (each of duration 0.1s) are randomly selected from the watermarked signal and for each piece, colored noise is added, the noise-added pieces are then filtered, and placed back in the correct location in the audio signal.
5. *After filtering:* The watermarked signal is low-pass filtered using a Butterworth 15-tap low-pass filter, with a cutoff frequency set to 1/16 the sampling rate.

6. *6.After MPEG coding/decoding:* The coding/decoding was performed using L3ENC version 2.71, a command-line software implementation of ISO/MPEG-1 Audio Layer 3 with a bit rate of 112 kbits/s.

A. Results of Listening Tests

Listening tests reveal that the effect of watermarking is less audible in the adaptively modified watermarked signal than in the watermarked signal obtained using the simple echo-hiding approach. Adaptation using the psycho-acoustic model or perceptual filter gives better results than signal based adaptation. It is found that psycho-acoustic model yields the best result for PopSong, SaxoMusic, ErHuMusic and ACapSong (popular songs/music), while perceptual filter gives the best result for SpeechClip (speech). This may be related to the facts that psycho-acoustic model is a good model for the Human Auditory System (HAS) and perceptual filter is a good model for the vocal tract.

It is also observed that in terms of inaudibility of watermark, multiple echo hiding yields a better result than single echo hiding because the decay rate for each echo per segment is significantly reduced.

B. Detection Accuracy

The detection accuracy for the audio pieces PopSong, SaxoMusic and SpeechClip using different approaches and after different forms of processing is presented in Tables I, II and III respectively.

Table I. Detection accuracy for PopSong

	Closed-loop	Noise	Re-sample	Crop	Filter	MPEG
Simple (decay=0.8)	97.14%	100%	97.14%	97.14%	100%	91.43%
Simple (decay=0.2)	74.29%	71.43%	74.29%	74.29%	80%	74.29%
Signal based: (single)	100%	97.14%	100.00%	100%	91.43%	85.71%
Signal based: (multiple)	100%	91.43%	85.71%	100%	82.86%	62.86%
Psycho-acoustic (single)	100%	94.29%	91.43%	100%	88.57%	80.00%
Psycho-acoustic (multiple)	100%	97.14%	91.43%	100%	91.43%	74.29%
Perceptual (single)	100%	97.14%	97.14%	100%	85.71%	91.43%
Perceptual (multiple)	100%	88.57%	88.57%	100%	71.43%	62.71%

Table II. Detection accuracy for SaxoMusic

	Closed-loop	Noise	Re-sample	Crop	Filter	MPEG
Simple (decay=0.8)	94.29%	97.14%	91.43%	94.29%	91.43%	88.57%
Simple (decay=0.2)	77.14%	68.57%	57.14%	77.14%	62.86%	62.86%
Signal based: (single)	100%	91.43%	91.43%	100%	94.29%	77.14%
Signal based: (multiple)	100%	94.29%	80.00%	100%	62.86%	65.71%
Psycho-acoustic (single)	100%	94.29%	91.43%	100%	85.71%	91.43%
Psycho-acoustic (multiple)	100%	88.57%	80.00%	100%	74.29%	71.43%
Perceptual (single)	100%	91.43%	85.71%	100%	71.43%	77.14%
Perceptual (multiple)	100%	96.43%	75.00%	100%	60.71%	60.71%

Table III. Detection accuracy for SpeechClip

	Closed-loop	Noise	Re-sample	Crop	Filter	MPEG
Simple (decay=0.8)	97.14%	91.43%	85.71%	97.14%	88.57%	91.43%
Simple (decay=0.2)	80%	60%	68.57%	80%	65.71%	54.29%
Signal based: (single)	100%	71.43%	77.14%	97.14%	68.57%	60.00%
Signal based: (multiple)	100%	71.43%	65.71%	97.14%	60.00%	45.71%
Psycho-acoustic (single)	100%	68.57%	77.14%	94.29%	74.29%	68.57%
Psycho-acoustic (multiple)	97.14%	74.29%	80.00%	97.14%	68.57%	54.29%
Perceptual (single)	97.14%	62.86%	57.14%	91.43%	48.57%	42.86%
Perceptual (multiple)	94.29%	54.29%	62.86%	94.29%	54.29%	51.43%

It can be seen that the detection accuracy is dependent on additional signal processing performed. 100% accuracy can be achieved using any of the three adaptation methods if there is no additional signal processing. The most damaging additional operation is MPEG.

The average detection accuracy for the five audio pieces using different coding schemes is summarized in Table IV.

Table IV. Average detection accuracy

	Pop-Song	Saxo-Music	ErHu-Music	ACap-Song	Speech-Clip	Ave
Simple (decay=0.8)	97.14%	92.86%	98.57%	96.67%	91.90%	95.43%
Simple (decay=0.2)	74.76%	67.62%	69.05%	74.76%	68.10%	70.86%
Signal based: (single)	95.71%	92.38%	90.00%	80.95%	79.05%	87.62%
Signal based: (multiple)	87.14%	83.81%	87.62%	79.76%	73.33%	82.33%
Psycho-acoustic (single)	92.38%	93.81%	87.62%	86.90%	80.48%	88.24%
Psycho-acoustic (multiple)	92.38%	85.71%	85.24%	85.71%	78.57%	85.52%
Perceptual (single)	95.24%	87.62%	75.24%	79.76%	66.67%	80.90%
Perceptual (multiple)	85.21%	82.14%	84.13%	78.57%	68.57%	79.73%
Average	91.35%	87.58%	84.97%	81.94%	74.44%	84.06%

Note: The shaded boxes indicate the highest detection accuracy for enhanced echo hiding for that clip.

It shall be noted that on the average, the simple form of echo hiding is robust to the various forms of processing mentioned and the average detection accuracy is very high. However, the simple form of echo hiding loses out in audibility.

Signal characteristics and mask computation method become important consideration if adverse signal processing is imposed. PopSong, has very high signal energy with no gaps of silence in general. For this test clip, an excellent average detection accuracy of over 90% is achieved. The watermark is also easily recoverable even after common signal processing operations. On the other extreme, the average detection accuracy for SpeechClip, a short speech with many silence intervals, is only 70%.

In addition, psycho-acoustic model (single echo) and signal based adaptation (single echo) yield the highest accuracy among the clips. This is because in echo hiding, the inaudibility-robustness tradeoff means that using a high decay rate will yield high recovery accuracy at the expense of audio quality, while using a low decay rate will yield an inaudible result at the expense of recovery accuracy. Signal based adaptation and psycho-acoustic model allow a higher decay rate while maintaining inaudibility, making these two encoding options more robust to common signal processing operations.

Though multiple echo hiding introduces more echoes per segment, the decay rate for each echo is much lower compared to single echo hiding.

The proposed watermarking methods are more robust to cropping (average detection accuracy = 97.48%), noise addition (average detection accuracy = 84.82%) and re-sampling (average detection accuracy = 81.74%) than the other signal processing operations because filtering and MPEG coding/decoding remove much more signal information.

V.CONCLUSION

Echo hiding is one of the audio watermarking methods for copyright protection. However, simple echo hiding is problematic because the echoes are highly audible.

By adaptively modifying the decay rate in each audio segment, a significant improvement in audio quality is achieved.

The results show that detection accuracy is highly dependent on additional signal processing performed. 100% accuracy can be achieved if there is no additional signal processing. The method is more robust to cropping, noise addition and re-sampling than the other operations of filtering and MPEG coding/decoding.

In conclusion, if both inaudibility and detection accuracy are of paramount importance, psycho-acoustic model (single echo) is the most appropriate encoding option to use when additional signal processing is imposed.

REFERENCES

[1] W. Bender, D. Gruhl, N. Morimoto, A.Lu, “Techniques for data hiding”, IBM Systems Journal, Vol 35, Nos 3 & 4, pp. 313-336, 1996.

[2] I.J. Coz, M.L. Miller, J.A. Bloom, “Watermarking applications and their properties”, Int. Conf. on Information technology’2000, Las Vegas, 2000.

[3] C. Xu, J. Wu, Q. Sun, K. Xin, “Applications of Watermarking Technology in Audio Signals”, Journal Audio Engineering Society, Vol. 47, No. 10, 1999 October.

[4] D. Gruhl, A. Lu, W. Bender, “Echo Hiding”, in Proc. Information Hiding Workshop (University of Cambridge, U.K., 1996), pp. 295-315.

[5] “Information Technology – Coding of moving pictures and associated audio for digital storage up to about 1.5 Mbits/s”, ISO/IEC IS 11172, 1993.

[6] W.W. Chang, C.T. Wang, “Audio Coding Using masking-Threshold Adapted Perceptual Filter”,

[7] K.C. Tan, “Multi-objective Evolutionary Algorithm (MOEA) homepage,” http://vlab.ee.nus.edu.sg/~kctan