NEW ECHO EMBEDDING TECHNIQUE FOR ROBUST AND IMPERCEPTIBLE AUDIO WATERMARKING

Hyen O Oh[†], *Jong Won Seok*[††], *Jin Woo Hong*[††], *Dae Hee Youn*[†]

[†] ASSP Lab., Dept. of Electrical and Electronic Eng., Yonsei University, Seoul, KOREA
[††]Radio & Broadcasting Tech. Lab., Electronics and Telecommunications Research Institute, Daejon, KOREA
E-mail: oho@assp.yonsei.ac.kr

ABSTRACT

Conventional echo watermarking techniques often exhibit inherent trade-offs between imperceptibility and robustness. In this paper, a new echo embedding technique is proposed. The proposed method enables one to embed high energy echoes while the host audio quality is not deteriorated, so that it is robust to common signal processing modifications and resistant to tampering. It is possible due to echo kernels that are designed based on psychoacoustic analyses. Subjective and objective evaluations confirmed that the proposed method could improve the robustness without perceptible distortion.

1. INTRODUCTION

Over the last few years, audio watermarking has become an issue of significant interest. This is primarily motivated by a need to provide copyright protection to digital audio content. Digital watermarking is a technique to embed copyright or other information into the underlying data. The embedded data should be perceptually inaudible to maintain the quality of the host signal.

There are a number of desirable characteristics that a watermark should exhibit. These include that it should be difficult to notice, robust to common signal processing, resistant to malicious attack of third parties, efficient to implement the system and detectable without original signal [1]. Among them the first and most important problem that all watermarking schemes need to address is that of inserting data in the audio signal without deteriorating its perceptual quality. The early emphasis was on this requirement, since the applications were not concerned with signal distortions or intentional tampering that might remove a watermark. However, as watermarks are increasingly used for purposes of copyright control, robustness to common signal processing and resistance to tampering have become important considerations. To improve robustness the embedded watermarks should have larger energies. The importance of perceptual modeling and the need to embed a signal in perceptually significant regions of an audio have been recognized. However, this requirement conflicts with the need for the watermark to be imperceptible.

Several possible audio watermarking techniques have been developed including phase coding, spread spectrum and echo watermarking [1][2]. In particular, Boney *et al* [3] present robust spread spectrum watermarking algorithm that exploits temporal and frequency masking. But since high quality audio compression algorithms also exploits these characteristics, it is not enough to exploit them further by inserting marks that are just above the truncation threshold of the compression but still below the threshold of perception. If perfect compression scheme exists in future, embedding of pseudo random sequence may be trivial or impossible. Moreover, listeners do not prefer noise addition itself to the host audio signal although it may be inaudible.

Echo watermarking marks audio signals by adding an echo [2]. In this case, the embedded signal is not additive noise but audio signal having the same statistical and perceptual characteristics. And the echo is even perceived as added resonance. Nevertheless, embedding a large echo to increase robustness of a watermark is subject to arise audible distortion. Inherent trade-offs still exist in the conventional echo watermarking.

In this paper, we propose a new echo embedding technique to improve robustness to signal processing attacks without deteriorating the host audio quality. Proposed watermarking system can embed relatively large echo without compromising with the quality, so that it can increase the reliability of the system and thus robustness. In order to make it possible, the echo kernel is designed based on psychoacoustic analysis. The proposed echo watermarking is evaluated through subjective quality tests and robustness tests, and results are compared with conventional echo embedding methods.

2. ECHO WATERMARKING

Echo watermarking embeds data into a host audio signal by introducing an echo [2]. The data can be hidden by varying the offset (time delay) of the echo. To embed binary data the coder uses two different offsets. Both offsets are below the threshold at which the human ear can resolve the echo. The encoding process can be represented as a system that has one of two possible system functions. In the time domain, the system functions are impulse response functions (denoted as kernels) with only several discrete impulses as shown in Fig. 1. In priori method [2], only two impulses (one to copy the original signal and one to create an echo) are chosen for the kernels. In order to encode more than one bit, the original signal is divided into smaller segments. Each individual segment can then be echoed with the desired bit by considering each as an independent signal. The final encoded signal is the recombination of all independently encoded signal segments. To prevent abrupt changes in the resonance of the encoded signal, smooth transition mechanism between segments should be considered.

Extraction of the embedded information involves the detection of spacing between the echoes. And it is considered as time delay estimation problem with so small delay that cannot be resolved by autocorrelation method. The cepstrum can be a convenient

solution to this problem. The magnitude of the autocorrelation of the encoded signal's cepstrum is used for the detection. Considering the relationship between the autocorrelation and the power spectrum, the autocorrelation of the cepstrum, denoted as *autocepstrum*, about the input signal, $x_m(n)$ can be represented as

$$\tilde{x}_m(n) = F^{-1}\left(\left(\log(F(x_m(n))\right)^2\right)$$

where F represents short-time Fourier Transform, F^{-1} represents its inverse transform and subscript m is a segment index. In each segment, the embedded binary data can be decoded by detecting the peak of autocepstrum.

3. NEW ECHO KERNEL DESIGN BASED ON PSYCHOACOUSTIC ANALYSIS

3.1 Perception of echo

In spite of the same physical origin, perception of echo by human ear breaks into two different way. One is echo and the other is coloration. This phenomenon can be explained by considering the time and frequency response of a single echo together with ear model. A single echo takes place when there exists only single wall between the sound source and receiver as shown in Fig. 2. Both the impulse response and frequency response of the single echo can be simplified as shown in Fig. 3. Since sound reflected from a single wall has two paths to the receiving point, the impulse response consists of two pulses and its frequency response varies over frequency as shown in the figure. Suppose the wall is distant from the source producing approximately a 50 ms delay in the reflected pulse. In this case, a clear single echo is heard. When multiple reflections of this type occur, as in a large hard-walled room, one hears the resulting reverberation as a clutter of echoes. On the contrary, with a signal having short echo delay (approximately a 2 ms delay), the perception is one of a change in timbre of the sound, usually called coloration.

In the ear model theory, it is well known that the incoming signal is broken into frequency bands, corresponding to the ear's critical bands [4]. These critical bands are the result of the cochlear filters, which have a memory, or duration, that is roughly inversely proportional to their bandwidth and is about 5 to 20 ms, depending on the critical band [5]. The long-delay pulse produces no variation in amplitude across frequency because the filters of the ear have shorter memory than the distance between the pulses, and the time or echo nature of the signal is well preserved. In the short-delay case, frequency content of the pulse pair is well-preserved in the output, but all time information is lost because the memory of the cochlear filters is greater than the distance between pulses, causing the pulses to interact within the cochlea.

In temporal masking, post-masking effects render weaker signals inaudible (during 50-200 ms, exponentially decaying) after the stronger masker is turned off. When temporal masking is considered, it is preferred that the offsets of the echo kernels are chosen to be small values and, thus, echo kernels are placed in coloration region. But too small offsets may distort the host signal as dull sound and may not be decoded by the cepstrum method. It is time consuming process to select adequate offsets.

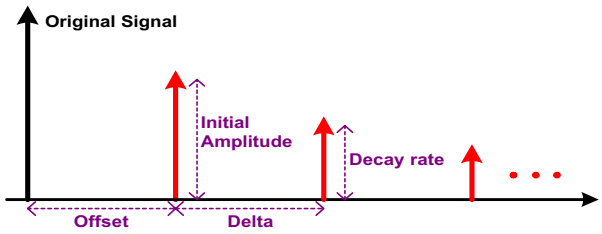

Fig. 1. An echo kernel.

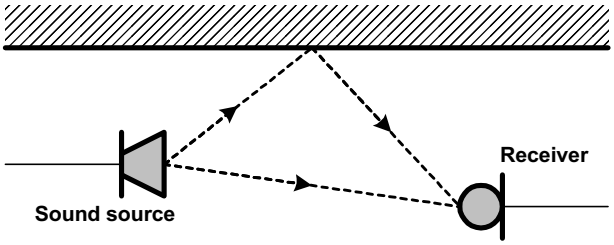

Fig. 2. Single echo conditions in real world.

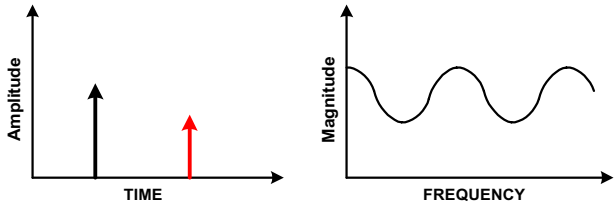

Fig. 3. Impulse response and frequency response of single echo.

3.2 Positive single and multiple echo

Since echo kernels used in echo watermarking are placed in coloration region, perceptual analysis of echo should be considered in the frequency domain.

Fig. 4 shows the frequency response of a conventional positive single echo where positive refers to the polarity of echo pulse. We chose critical-band rate scale as the horizontal axis in the figure [4]. This scale more precisely describes the human perception and provides a good approximation of frequency resolution of human auditory system. Ripple shape across critical bands comes from the combination of pulses in the kernel. The varying rate of ripple corresponds to the offset of echo and the amplitude of ripple is determined by the initial amplitude of kernel. For the watermark to be ideally transparent, the plain response over the entire critical band would be desirable. But this implies a trivial kernel that has only one impulse in the time domain. No signal can be embedded. Nonetheless, to make the critical band response as plain as possible is an important issue in designing echo kernels.

We have run numerous tests to recognize more detailed relationship between the response shape and echoed signal quality. Based on the test results, we could know that, (1) overall spectral envelope is more important than the fine spectral shape in the critical-band rate scale, (2) lower bands response (0-10 barks) plays a key role to determine the sound quality, and (3) espe-

cially, the shape at a few lowest critical bands determine the timbre of echoed audio signal. In the positive single echo case, the lowest three bands are amplified as shown in Fig. 4, which acted as a kind of bass booster and produced slightly richer sound.

Previously, Xu *et al* [6] proposed a multiple echo technique. Instead of embedding one large echo into an audio segment, four smaller echoes with different offsets were chosen. By using the multiple echo kernel, the amplitude of echoes can be reduced with the same detection rate. This may reduce the possibility of echo detection and removal by third parties because they do not know the parameters. Fig. 5 shows the frequency response of a positive multiple echo example in which two echo pulses are present. For comparison, the sum of the initial amplitudes is chosen to be equal to the initial amplitude used in Fig. 4. The positive multiple echo kernel produced similar change of timbre to the positive single echo, because the response in low frequency bands up to 12 barks is almost the same shape as in Fig. 4. But the high frequency bands envelop generated the annoying distortion when larger echoes were embedded. As increasing the number of pulses, the spectral shape was changed to the direction of having more complicated ripples. Multiple echoes may reduce the possibility of echo detection by third parties but cannot increase the robustness because the change of timbre increases in proportion to the sum of pulse amplitudes.

3.3 Proposed echo kernel

Instead of positive echo, we first introduced a negative echo which had negative initial amplitude and, thus, didn't add the copied signal but subtracted it from host signal. In Fig. 6, the frequency response of a negative single echo is presented. In this case, inversed response shape having similar ripples as in the positive single echo case can be noticed. The lowest bands' attenuation made the sound slightly sharper one. In the decoding stage, both the positive and negative echo could be detected by the cepstrum method with the same detection rate.

Frequency response of a new echo kernel proposed in this paper is depicted in Fig. 7. The proposed echo kernel comprises multiple echoes by both positive and negative pulses but with different offsets in the kernel. By combining closely located positive and negative echoes, more transparent response can be obtainable than the conventional positive single or multiple echo methods. It is obvious from Fig. 7 that much more plain response can be acquired in lower bands than the conventional case shown in Figs. 4 and 5. Large ripples are located in the higher bands that are perceptually less important, and their amplitudes are still no more than positive single case in Fig. 4.

4. EVALUATION

To evaluate the performance of the proposed echo kernel, subjective quality tests and robustness tests were conducted. Test results are compared with conventional methods. Echo embedding methods considered for the experiments are summarized in Tab. 1. Scalefactor in the table corresponds to total amplitude of echo pulses. α was experimentally selected at the level that can make the echo embedded sound by Method I transparent.

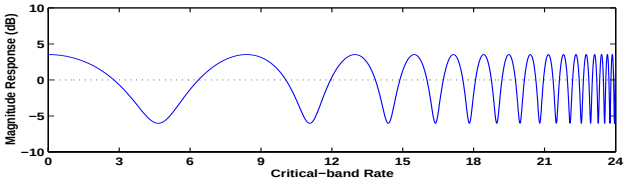

Fig. 4. Frequency response of a positive single echo.

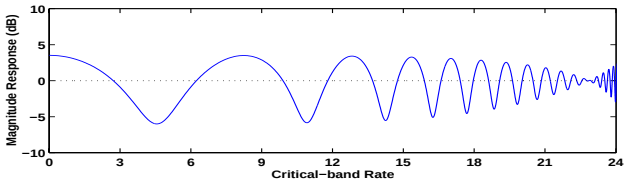

Fig. 5. Frequency response of a positive multiple echo.

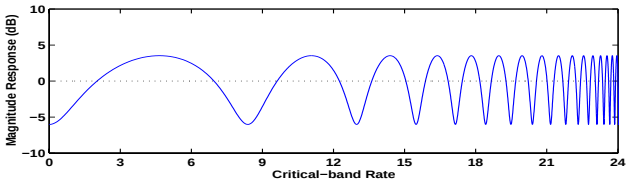

Fig. 6. Frequency response of a negative single echo.

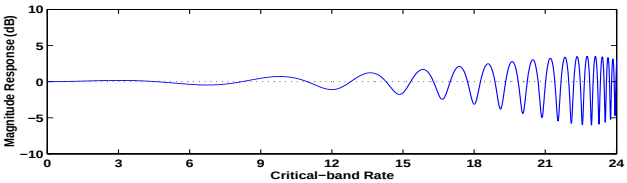

Fig. 7. Frequency response of the proposed echo.

4.1 Subjective Quality Test

Subjective evaluation of watermarked material was done via *Double blind triple stimulus with hidden reference* listening test (ITU-R recommendations BS.1116 [7]). 15 listeners were involved in the experiments. Some of them were experienced and familiar with the test material. 6 different music signals including classical, jazz, pop, and a cappella were selected.

Listening test results are summarized in Tab. 2. Diffgrade and '# of Transparent Items' results are presented. Diffgrade is equal to the subject rating given to the watermarked test item minus the rating given to the hidden reference. Therefore, a diffgrade near 0.00 indicates a high level of quality. Diffgrade can even be positive which means incorrect identification of the watermarked item, *i.e.*, the quality is transparent. The diffgrade scale is partitioned into five ranges: "imperceptible (>0.00)," "not annoying (0.00~ -1.00)," "slightly annoying (-1.00~ -2.00)," "annoying (-2.00~ -3.00)," and "very annoying (-3.00~ -4.00)." '# of Transparent Items' in the table stands for the number of incorrectly identified items over 90 test items.

It is apparent from the results that only Method I and Method V (proposed method) show imperceptible quality and Method IV

shows the worst quality corresponding to "slightly annoying." The '# of Transparent Items' results also show similar pattern. From the subjective evaluation, it can be said that the proposed echo kernel (Method V) provides a comparable quality to the single echo kernels (Method I and II).

4.2 Robustness Evaluation

In the decoding process, the peak of the autocepstrum in each segment is detected. Since the possible peak positions, *i.e.*, offsets were presumed to be known, detection problem becomes decision problem. Many techniques including simple maximum-likelihood decision criterion can be adopted [8].

To evaluate robustness, we first investigate the empirical pdf of the autocepstrum at pre-known offsets. Fig. 8 shows the results for 5 embedding methods analyzed over 1289 segments. The pdf which has larger mean and smaller variance is better one for detection performance. In particular, common signal processing operations such as compression, filtering and noise addition are subject to blurring the pdf, *i.e.,* increasing the variance. So it is highly encouraged to make the mean value as large as possible to preserve reliable decoding after signal processing attack. It is apparent from Fig. 8 that the proposed method(Method V) together with Method IV represents the most large mean. Empirical means and variances are summarized in Tab. 3. The detection results expressed by bit error rate (BER) after embedding binary watermark bitstream are also shown in Tab. 3. The BER was measured for both unprocessed signals and 56kbps MPEG-Audio Layer-III codec passed signals [9].

Tab. 1. Echo embedding methods for evaluation

Label	Type	Scalefactor
Method I	Positive Single	α
Method II	Negative Single	α
Method III	Positive Single	1.7α
Method IV	Positive Multiple	2α
Method V	Proposed	2α

Tab. 2. Results of subjective quality test.

Embedding Method	Diffgrade	# of Transparent Items
Method I	0.15	47
Method II	-0.08	39
Method III	-0.53	4
Method IV	-1.41	0
Method V	0.04	42

Tab. 3. Results of robustness evaluation.

Embedding Method	Mean	Var.	BER	
			Not processed	MP3 coded
Method I	0.288	0.0019	0.16%	1.78%
Method II	0.311	0.0019	0.16%	1.40%
Method III	0.488	0.0019	0.00%	0.47%
Method IV	0.584	0.0036	0.00%	0.39%
Method V	0.592	0.0043	0.00%	0.39%

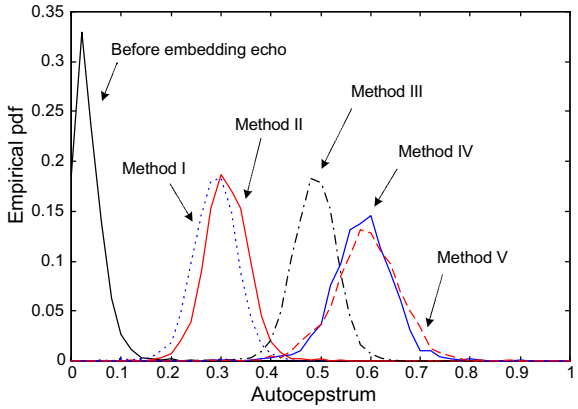

Fig. 8. Empirical pdf of autocepstrum in each method.

5. CONCLUSION

In this paper, we proposed a new echo embedding technique for robust and imperceptible audio watermarking. Psychoacoustic analysis in frequency domain based on perception of echo has been conducted to derive the direction of echo kernel design. Subjective evaluation and robustness tests verified that using the proposed method, it was possible to embed echoes whose energies are two times larger than conventional approaches and, thus, to acquire corresponding robustness without degrading the host signal quality.

6. REFERENCES

[1] M. Swanson, M. Kobayashi, and A. Tewfik, "Multimedia Data-Embedding and Watermarking Technologies (invited paper)," *Proc. of the IEEE, Vol. 86, No. 6*, 1998 Jun.

[2] W. Bender, D. Gruhl, N. Morimoto, A. Lu , "Techniques for data hiding", *IBM System Journal*, Vol 35, Nos 3&4, 1996.

[3] L. Boney, A. Tewfik and K. Hamdy, "Digital Watermarks for Audio Signals," *IEEE Int. Conference on Multimedia Computing and Systems*, pp 473-480, 1996.

[4] E. Zwicker, *Psychoacoustics.* Springer-Verlag, New York, 1982.

[5] G. Studebaker and I. Hochberg, *Acoustical Factors Affecting Hearing Aid Performance*, Allyn and Bacon, 1993.

[6] C. Xu, and *et al.*, "Applications of Digital Watermarking Technology in Audio Signals," *J. Audio Eng. Soc.*, Vol. 47, No. 10, 1999 Oct.

[7] ITU-R Rec. BS.1116, "Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems," International Telecommunication Union, Geneva, Switzerland, 1994.

[8] J. Melsa and D. Cohn, Decision and estimation theory, McGraw-Hill, New York, 1978.

[9] http://www.iis.fhg.de/amm/techinf/layer3/index.html