Presentation of Information Synchronized with the Audio Signal Reproduced by Loudspeakers Using an AM-based Watermark

Akira Nishimura
Department of Media and Cultural Studies, Faculty of Informatics,
Tokyo University of Information Sciences
1200–1, Yatoh-cho, Wakaba-ku, Chiba-city, Chiba, JAPAN
akira@rsch.tuis.ac.jp

Abstract

Reproducing stego audio signal via a loudspeaker and detecting embedded data from a recorded sound from a microphone are challenging with respect to the application of data hiding. A watermarking technique using subband amplitude modulation was applied to a system that displays text information synchronously with the watermarked audio signal transmitted in the air. The robustness of the system was evaluated by a computer simulation in terms of the correct rate of data transmission under reverberant and noisy conditions. The results showed that the performance of detection and the temporal precision of synchronization were sufficiently high. Objective measurement of the watermarked audio quality using the PEAQ method revealed that the mean objective difference grade obtained from 100 watermarked music samples exhibited an intermediate value between the mean ODGs of 96-kbps and 128-kbps MP3 encoded music samples.

1. Introduction

Applications of audio watermarking are not restricted to the area of copyright management. Another application of audio watermarking is the inclusion of augmentation data[2]. For example, if an announcement or music broadcasted from loudspeakers in a public area could contain the text form of the announcement or advertisement messages embedded in it, the embedded data could be extracted in real-time from the transmitted sound detected by a microphone. A receiver for the watermarked audio signal could be a personal digital assistant or a cellular phone.

Such an application requires the watermarking system to be robust against reverberations, reflections, and background noise. It also requires a greater data payload than copyright management systems.

The author has developed a new audio watermarking technique based on subband amplitude modulation. Evaluation of this watermarking system was conducted for data hiding in music [4] and in speech signals [6]. The system is robust against perceptual audio codings, additive Gaussian noise, spectral modifications, reflections, and reverberations. Deterioration of the sound quality resulting from embedding is relatively small because modulation masking and modulation detection interference occur in the human auditory system.

In the present paper, a technique for displaying embedded data synchronously with the watermarked sounds reproduced by loudspeakers is presented. The technique can be used to display Karaoke lyrics or captions of foreign films synchronously with the music or the sounds in the film. The robustness of the watermark with respect to background noise and reflections caused by air transmission is examined. The objective audio quality for the watermarked audio signals is also evaluated.

2 Audio watermarking based on amplitude modulation

2.1 Embedding process

At the beginning of the embedding process, a host signal $H(t)$, which is the length of one data frame period, is split into $2n$ subband signals $h_m(t)$ using an equal-bandwidth filter bank:

$$H(t) = \sum_{m=1}^{2n} h_m(t). \tag{1}$$

Sinusoidal amplitude modulations (SAMs) at a relatively low modulation frequency (f-Hz) are applied to the adjacent subband signals $h_{2m}(t)$ and $h_{2m+1}(t)$ in the opposite phase. An embedding key produced by a known pseudorandom number generator arbitrarily classifies the n subband pairs into k subband groups. It also defines the random initial phase angles $r(m)$ of the SAMs for each subband pair. The output of an amplitude modulated subband pair $x_m^i(t)$, which belongs to the i-th subband group, is given by

$$x_m^i(t)=h_{2m}(t)\Big(1+A(m)\sin\big(2\pi ft+r(m)+p(i)\big)\Big)+ h_{2m+1}(t)\Big(1-A(m)\sin\big(2\pi ft+r(m)+p(i)\big)\Big), \quad (2)$$

where $A(m)$ is the depth of the SAM of the m-th subband pair. Embedded information is encoded by phase shift keying (PSK), defined as the difference between the phase angles of the SAM of the first subband group and that of the i-th subband group, $p(1)$ and $p(i)$ $(i = 2, ..., k)$. Quarterly PSK encodes 2-bit information $(D_i = 0, 1, 2, 3)$ into every $\pi/2$ phase angle of $p(i)$.

$$p(i) = \begin{cases} 0 & i = 1; \\ \dfrac{\pi D_i}{2} & i = 2, ..., k. \end{cases} \quad (3)$$

As a result, $2(k - 1)$ bits of information are embedded per data frame period. Multiplex watermarking can be applied using different modulation frequencies simultaneously. Finally, a watermarked signal $X(t)$ is obtained by summing all of the amplitude-modulated signals $x_m(t)$.

$$X(t) = \sum_{m=1}^{n} x_m(t). \quad (4)$$

Synchronization of the data frames is achieved by inverting the relative phase of the SAMs between successive frames for the first subband group.

2.2 Extraction process

The amplitude envelopes of the subbands $E_m(\tau)(m = 1, 2, ..., 2n)$ of the watermarked frame signal can be derived from the amplitude spectrum of the half-overlapped running FFT, where τ is the period of time defined as half the FFT length. The embedded modulation waveform $G_m(\tau)$ is extracted by calculating the logarithmic ratio of the amplitude envelopes extracted from adjacent subband signals.

$$G_m(\tau) = \log \frac{E_{2m}(\tau)}{E_{2m+1}(\tau)}. \quad (5)$$

After compensation of the initial phase difference $r(m)$ for each $G_m(\tau)$, synchronized addition of $G_m(\tau)$, which belongs to the i-th subband group produces an accumulated AM waveform $G^i(\tau)$, which emphasizes the AM waveform. Consequently, the modulation depth $A(m)$ in the embedding process can be kept small. Initial phase differences between the first and the i-th subband group, that is, the embedded information, are obtained by comparing the phase angles of the FFT spectra calculated from $G^1(\tau)$ and $G^i(\tau)$.

The intensity of a watermark, which is defined as the depth of the SAM $A(m)$, was determined relative to the extracted modulation power of the subband signal of the host signal. This means that the $A(m)$ value that generates equal power of the effective value of $G_m(\tau)$ for the non-watermarked signal is set to 0 dB [5].

2.3 Detecting the starting point of a data frame

Before decoding the PSK data, the starting point of the embedded data frame must be detected. A rectangular temporal window of the data frame length T is iteratively applied to the modulation waveform $G^1(\tau)$ extracted from the first subband group. The starting point of the windowing is denoted by u in Eq. 6. Then, $F(u)$ is derived by subtracting the synchronized addition of the modulation waveforms in the odd-order windows from the synchronized addition of the modulation waveforms in the even-order windows.

$$R_u = \{G^1(u), G^1(u + 1), ..., G^1(u + T - 1)\}. \quad (6)$$

$$F(u) = \sum_{v=0} R_{u+2vT} - \sum_{v=0} R_{u+(2v+1)T}. \quad (7)$$

The Fourier amplitude of $F(u)$, which corresponds to the modulation frequency f, denoted by $\mathrm{AMP}_f\big(F(u)\big)$, exhibits a maximum when the position of the window overlaps completely the position of the frame. Consequently, the starting point of the data frame y is given by

$$y = \operatorname*{argmax}_u \mathrm{AMP}_f\big(F(u)\big). \quad (8)$$

The above-described method is applicable when the duration of the watermarked signal is at least $2T$. The detection performance is improved by using a longer watermarked signal, especially when a severe disturbance is applied to the watermarked signal. The present implementation utilizes a watermarked signal length of $6T$ to calculate $F(u)$.

3 Displaying information synchronously with watermarked sounds

Table 1 shows an example of the definition of the record of embedding data. The record consists of six data fields of 29 bits in total. BCH coding is used to encode 29 bits of the record into 127 bits with the error correction limit of 21 bits. Bit sequences in the data fields are interpreted as unsigned integer values. The length of each data field depends on the requirements of the application. The second field defines the position of the target frame to begin display relative to the embedded frame. The exact timing to begin display during the target frame is defined as the ratio from the beginning of the frame. The end timing of display is defined in the same manner as the start timing. The index data field points to the entity of the information for display, such as a song lyric or a movie caption that is already stored and indexed in the machine that runs the detection program.

Synchronization of displaying embedded data is based on real-time detection of the embedding frame segment from the watermarked signal. After frame detection, all decoded record data are stored in the buffer memory for the time being. Each time the incoming watermarked signal is transported from the AD converter, the timing data of the buffered records are scanned to display the indexed information.

Table 1. Data definition of the embedding record.

Auxiliary data	Relative starting frame	Starting time in the frame	Relative ending frame	Ending time in the frame	Index for information to display
2 bits	6 bits	4 bits	6 bits	4 bits	7 bits

4 Computer simulation of noisy and reverberant conditions

A computer simulation was performed in order to confirm that the data embedded in 100 pieces of popular music (RWC-MDB-P2001 [1]) could be successfully extracted under noisy and reverberant conditions. An AM- based watermark of the embedding condition shown in Table 2 is embedded in the initial 60-second left-channel signal of each piece of music.

The impulse response of the reverberation room having a reverberation time of 1.3 seconds (the ‘ir130.dat’ file in the RWCP Sound Scene Database in a real acoustic environment) was convolved with the watermarked music signals. One of the five types of background noise was added to the reverberant watermarked music. These background noises consisted of noises recorded in an airport lounge, a crowded intersection, an underground corridor, a station platform, and a lowpass noise that had a cutoff of 500 Hz and a slope of –9 dB/oct. The same noises were used in the previous research [6]. The overall signal-to-noise-ratio (SNR) was 15 dB. The role of the impulse response is to supply the band-limited characteristics of a loudspeaker and a microphone, not only to supply reflections and reverberations.

Table 2. Watermarking conditions.

Parameters	Values
embedding bit rate	43 bps
sampl. freq. [Hz]	44,100
embedding region	below 11,025 Hz
subband pairs (n)	64
subband groups (k)	17
frame period	3 s
mod. freq. [Hz]	1.67, 2, 2.33, 3
watermarking intensity	+12 dB

The watermarking performance was evaluated in terms of the number of data frames that achieved a bit error rate lower than the error correction limit of 21 bits. The accuracy of synchronization was evaluated in terms of the deviation of the detected boundary from the correct boundary of the data frame. Since these performance indices depend on the number of frames from the beginning of the watermarked music, the results are shown as a function of the frame number. Figure 1 shows the ratio of the correctly recovered data frames. Figure 2 shows the mean deviation from the correct boundary location. Each datum is collected from 500 simulated conditions (100 pieces of music versus five background noises).

The ratio of correctly detected data frames is insufficient in the initial part of the music. This is caused by the relatively large deviations of the detected frame boundaries from the correct timing and the small power level observed in the introduction of the music. However, after the fifth frame, the synchronization and detection performances are sufficient for the host signal of various types of popular music.

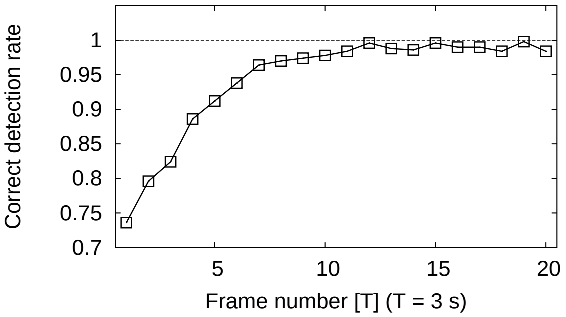

Figure 1. Correct rate of detected frames. Each datum was collected from 500 simulated conditions.

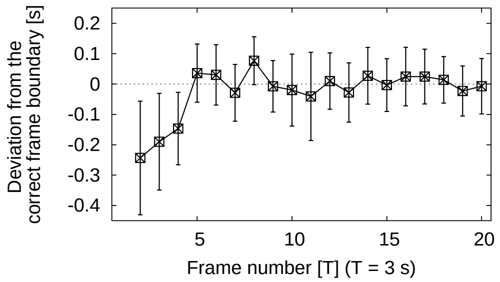

Figure 2. Deviation from the correct frame boundary. Error bars show ± 1 standard errors.

5 Objective audio quality measurement for watermarked music

A method for the objective measurement of perceived audio quality (ITU-R Recommendation BS.1387), referred to hereinafter as PEAQ, uses a number of psycho-acoustical measures that are combined to provide a measure of the quality difference between two instances of a signal (a reference and a test signal). The implementation of the basic version of PEAQ by Kabel [3] was used to assess the objective quality degradation of the watermarked music used in the previous section. The PEAQ measurement outputs the Objective Difference Grade (ODG), which corresponds to the Subjective Difference Grade obtained from the assessment procedure of the subjective quality degradation defined in ITU-R BS.1116-1.

Figure 3 shows the averaged result of PEAQ measurement for watermarked music, and 96-kbps and 128-kbps MP3 encoded music. The mean ODG obtained from the watermarked music is an intermediate value between the 96-kbps and 128-kbps MP3 encoded music. In other words, the degradation in quality of the music impaired by watermarking is slightly annoying but still tolerable.

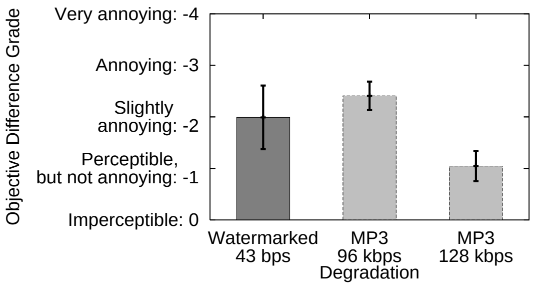

Figure 3. Mean ODGs for 100 pieces of music from RWC-MDB-P2001. Error bars show ± 1 standard errors.

6 Discussion

Practical implementation of the system requires embedding the same record iteratively. The limitation of this scheme is that recorded data must be embedded before being displayed. The optimal number of iterations for each embedding record can be determined automatically depending on the host signal.

In an application of displaying captions of a film, it is difficult to handle relatively long periods of silence in the soundtrack. A solution is to use a long synchronization frame divided into several data frames. The embedding parameters of the system are flexible, allowing the desired embedding bit rate to be to realized if the length of the frame is altered.

7 Summary

A watermarking technique using subband AM was applied to the synchronous display of text information with the audio signal. The robustness of the system was evaluated by a computer simulation in terms of the correct data transmission rate under reverberant and noisy conditions. The results showed that the performance of detection and temporal precision are sufficient. Objective measurement of the watermarked audio quality was conducted using the PEAQ method. The mean ODG obtained from the 100 watermarked music was an intermediate value between the mean ODGs of 96-kbps and 128-kbps MP3 encoded music.

References

[1] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. Rwc music database: Popular, classical, and jazz music databases. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, pages 287—288, 2002.

[2] D. Gruhl, A. Lu, and W. Bender. Echo hiding. In *Proceedings of the First International Workshop on Information Hiding LNCS 1174*, pages 295—315, 1996.

[3] P. Kabal. An examination and interpretation of ITU-R BS.1387: Perceptual evaluation of audio quality. *TSP Lab Technical Report, Dept. Electrical & Computer Engineering*, 2002.

[4] A. Nishimura. Audio watermarking based on sinusoidal amplitude modulation. In *Proceedings of ICASSP 2006, IV*, pages 797—800. IEEE, 2006.

[5] A. Nishimura. Audio watermarking based on subband amplitude modulation. In *Proceedings of the 2006 Symposium on Cryptography and Information Security*, number 3F4-2. IEICE, 2006.

[6] A. Nishimura. Data hiding for speech sounds using subband amplitude modulation robust against reverberations and background noise. In *Proceedings of IIH-MSP 2006*, pages 7–10. IEEE, 2006.