Data hiding in speech sounds using subband amplitude modulation robust against reverberations and background noise

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

Data hiding in audio signals can be used for transmitting auxiliary information related to the content of the audio signal. Such an application requires a greater data payload than music copyright management systems. Also, it is difficult to extract the embedded data from sounds played through a loudspeaker and detected by a microphone because of additive background noise, reflections and reverberations and the band-limited characteristics of loudspeakers and microphones. In this study, a watermarking technique using subband amplitude modulation was evaluated by computer simulation in terms of robustness against background noises and reverberations. The effects of amplitude modulation on the articulation scores of 125 vowel-consonant-vowel (VCV) syllables were also investigated. The results showed that reverberant speech signals with various background noises having a SNR of 10 dB can transmit more than 90% of embedded data at 48 bps, with only a small deterioration in the syllable identification scores.

1. Introduction

Audio watermarking techniques are usually associated with content protection and digital rights management of music. In such applications, the embedded data is generally the copy control code and the content identification code. Therefore, the robustness of watermarks to perceptual codecs, additive noise and lowpass filtering is critical when watermarked music is distributed by computer networks or broadcasting.

Another application of audio watermarks is the inclusion of augmentation data[1]. For example, if an announcement broadcast from loudspeakers in a public area could have a text form of the announcement embedded in it, the embedded data could be extracted in real-time from the speech signal that is detected by the microphone. Such an application would be invaluable for hearing-impaired people. A receiver for the watermarked audio signal could be a portable device such as 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, while speech has a narrower frequency band compared with that for music, which will result in a low embedding data rate. Moreover, a certain degree of quality degradation is acceptable in the case when the cover data is speech, while clarity is essential for watermarked speech.

However, very few attempts have been made to produce watermarking systems for speech or to verify the robustness of watermarks against reverberations. Generally speaking, most audio watermarking techniques that adopt relatively short embedding time frames, say below 1 s, cannot be used in reverberant spaces, since the delayed signal frame masks and interferes with the following frame. Such watermarking techniques include spread-spectrum, echo-hiding and phase-manipulations technologies.

The author has developed a new audio watermarking technique based on subband amplitude modulation. This system is robust against perceptual audio codings, additive Gaussian noise and spectral modifications [2]. It is also robust against reflections and reverberations, since it applies relatively slow amplitude modulation in a long embedding frame of several seconds.

Evaluation of this watermarking system was initially conducted for data hiding in music[2]. However, the watermarking system can also be used for data hiding in speech signals.

In this paper, the effectiveness of watermarking based on amplitude modulation in applying speech signals is examined in terms of the embedding data rate, the detection rate and the correct identification score of listeners.

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 filterbank:

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

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 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

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

where $A(m)$ is the depth of the SAM of the m-th subband pair. Embedded information is encoded by phase shift keying, defined as the differences between the phase angles of the SAM of the first subband group and that of the i-th subband group, $p(1)$ and $p(i)$ $(i = 2, ..., k)$. 4-phase shift keying encodes 2-bit information $(D_i = 0, 1, 2, 3)$ to 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 up all 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 is defined as the depth of the SAM $A(m)$. In a previous study, $A(m)$ was determined relative to the weighted modulation power [2] or the extracted modulation power of the subband signal of the host signal [3]. In the present study, the values of $A(m)$ for all m are equal and they are simply determined as parameter values, because the calculation cost should be minimized for real-time embedding.

2.3 Finding a starting point of the data frame

Before decoding the phase shift keying 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.

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

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

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 with the position of the frame. Consequently, the starting point of the data frame y is given by

$$y = \underset{u}{\mathrm{argmax}}\, \mathrm{AMP}_f\big(F(u)\big). \qquad (8)$$

3 Computer simulation of noisy and reverberant conditions

A computer simulation was performed to confirm that the embedded data in speech signals can be successfully extracted in noisy and reverberant conditions. The watermarking conditions used in this simulation are shown in Table 1.

Table 1. Watermarking conditions.

Parameters	Values	
embedding bit rate	64 bps	48 bps
sampl. freq. [Hz]	22050	←
embedding region	below 6034 Hz	←
subband pairs (n)	140	←
subband groups (k)	25	←
frame period	3 s	4 s
mod. freq. [Hz]	1, 1.67, 2.33, 3	1, 1.5, 2, 2.5

Speech segments having a duration of 36 seconds and spoken by 12 female and 10 male speakers were selected from the Continuous Speech Database for Research (Vol. 1) published by the Acoustical Society of Japan. These speech segments were used as the cover audio data. Their sampling frequencies were converted from 16 kHz to 22.05 kHz.

In the simulation, a room impulse response recorded in a variable 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 speech signals. One of the five types of background noise was added to the reverberant watermarked speech. These background noises consisted of noise recorded in an airport lounge, a crowded intersection, an underground corridor, a station platform and a lowpass noise which had a cutoff of 500 Hz and a slope of –9 dB/oct. (this is representative of a typical average spectrum of background noise). The averaged spectra of the background noises are shown in Fig. 1. The overall signal-to-noise-ratio (SNR) was 10 dB and 20 dB. The watermarking performance was evaluated in terms of the correct bit rate of the extracted data when compared with that of the embedded data. No error correction was performed on the embedded data.

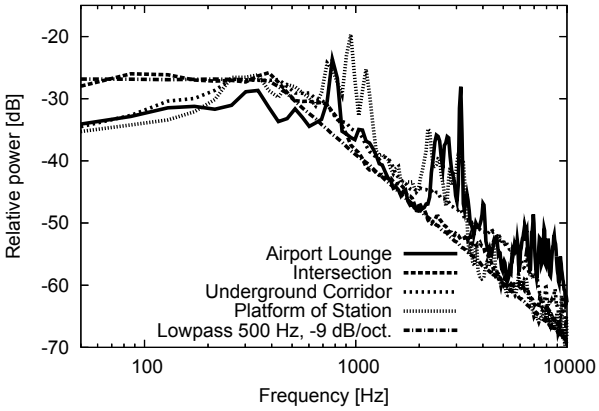

Figure 1. Averaged power spectra of the additive background noises.

The impulse response was recorded in a variable reverberation room, which is a different environment from those of the background noises. However, 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.

Figure 2 shows the medians of the bit detection rate for 110 conditions (22 speakers by five background noises) as a function of the watermark intensity $A(m)$. The error bars denote the 10th to 90th percentile of the detection rate obtained in the 110 conditions. The results of the detection rates demonstrate that if data is embedded at 48 bps and an AM depth of more than 0.5, more than 88% data bits survive for more than 90% conditions of the speaker and background noise combinations at a SNR of 10 dB. They also reveal that embedding at 64 bps is effective for high SNR conditions.

4 Identification test for watermarked VCV syllables

Identification tests for watermarked VCV syllables were conducted to investigate the effect of watermark on articulation scores.

125 watermarked VCV syllables, consisting of five first vowels (a, i, u, e, o), 25 Japanese consonants and the second vowel 'a', were presented via a headphone (Sennheiser HDA-200) diotically in a soundproof room. The conditions of watermarking were the same as those

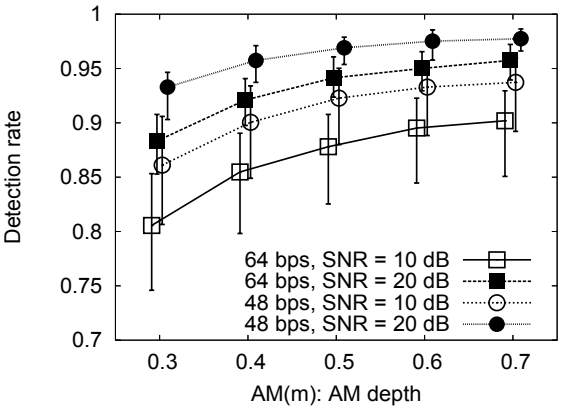

Figure 2. Medians of bit detection rate. Error bars denote 10th to 90th percentile of the detection rate obtained in 110 conditions.

given in Table.1 except that the embedding frequency region was under 8 kHz. The subjects were asked to type what they heard using the keyboard of a computer. The watermark intensity ($A(m)$ in Eq.2) was set to 0.4, 0.6 and 0 (no watermarking). The VCV syllables were presented in random order with or without a low-pass noise which had a cutoff frequency of 500 Hz and a slope of –9 dB/oct. at a SNR 10 dB. Overall SPL of the VCV syllables was 72 dB. Five subjects with normal hearing participated in the same experiment which was conducted twice on separate days.

The mean articulation scores of the 125 VCV syllables are shown in Fig. 3. The error bars denote the minimum and maximum scores of the subjects. Since speech intelligibility is generally higher than the articulation scores of syllables, no serious deterioration in the intelligibility of the watermarked speech is expected.

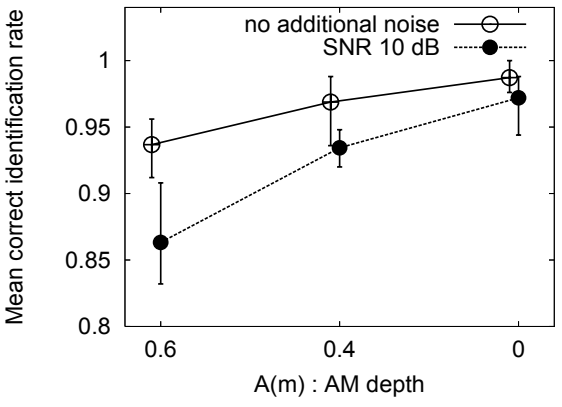

Figure 3. Mean articulation scores of 125 VCV syllables. Error bars denote minimum and maximum among the five subjects.

5 Discussion

The advantage of the identification test of the VCV syllables is availability of confusion matrix analysis. Analysis of confusion matrix shows that the misidentification rates of the syllables which were misidentified when there was no watermark but background noise was present increase when the intensity of watermarking increases and background noise is added. The finding of occasional misidentification of the first and second vowels in all conditions and the dependence of misidentification of the consonants on the first vowel, indicates that amplitude modulation impairs temporal changes which result from coarticulation between the first vowels and the consonants. Detailed analysis of the confusion matrix will enable improvements on the embedding algorithm.

In the case of an audio watermark that carries text data of the speech, the payloads of embedded data, including a practical error correction scheme, should be in the range of 200 to 300 bps in order to achieve a real data rate of approximately 100 bps, which represents the average text data rate transmitted by speech. Since the present data rate is four or five times smaller than the above requirement, effective data compression for embedding and improved watermark detection is required.

6 Summary

A watermarking technique using subband amplitude modulation was evaluated by a computer simulation in terms of robustness against background noises, reflections and reverberations. The effects of amplitude modulations on the articulation scores of 125 VCV syllables were also investigated. The results showed that reverberant speech signals with various background noises having a SNR of 10 dB can transmit more than 90% of embedded data at 48 bps, with only a small deterioration in syllable identification scores.

References

[1] D. Gruhl, A. Lu, and W. Bender. Echo hiding 1996. In *Information Hiding*, pages 295—315, 1996.

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

[3] 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.