Interspeech 2022
18-22 September 2022, Incheon, Korea

Practical Over-the-air Perceptual Acoustic Watermarking

Ameya Agaskar

Amazon Alexa
Cambridge, MA, USA
agaskar@amazon.com

Abstract

In this work, we demonstrate a novel technique for automatically scaling over-the-air acoustic watermarks to maximize amplitude while remaining imperceptible to human listeners. These watermarks have been demonstrated in prior work to be robust to the indoor acoustic channel. However, they require careful calibration to ensure that they are (a) detectable by the device and (b) imperceptible to humans. While previously this was done using listening tests, we show that psychoacoustic masking curves can be used to automatically scale each watermark frame's amplitude to be as high as possible while remaining below the masking level. This maximizes watermark detectability by the self-correlation decoder described in earlier work, while ensuring that the watermark is not heard.
Index Terms: audio watermarking, psychoacoustics, smart home

1. Introduction

The proliferation of smart speakers, networked sensors, and computing devices in modern homes has led to a vision of *ambient computing*, in which all of these devices will blend into the surroundings and improve consumers' lives while requiring less conscious input. This will require a broad range of devices from many different manufacturers to communicate and work together.

The indoor acoustic channel provides an underutilized communication mode between devices that may not be explicitly networked together. For example, a user watching a movie using a streaming app on their TV may want their smartphone, a "second screen", to show content related to the movie [1].

In [2], the authors described one of the first examples of an audio watermark that is robust to acoustic propagation over an indoor channel and time/frequency drift between the encoder and decoder. A watermark codec is the simplest form of communication, providing a binary message of presence or absence.

A challenge with over the air audio watermarking is that it requires the balancing of two imperatives: (1) it must be loud enough to be detectable by a smart device, and (2) it must be quiet enough to not be noticeable or distracting to users.

We present an extension of the watermark in [2] that automatically scales the amplitude over time in order to perceptually mask the watermark under the targeted audio's spectrum, while maintaining the detection performance resulting from the prior, manual strength selection.

1.1. Prior Work

This work is an extension of [2], which described an over-the-air watermarking system that was robust to the indoor acoustic channel and the lack of prior synchronization between the sound source and the detector. That paper cited several previous examples of watermarking systems that were robust to either indoor propagation [3]–[7] or de-synchronization [8, 9], but not both—it was the first practical example robust to both impediments.

There is one example in the literature of a watermark encoder that uses perceptual masking curves to determine its embedding strength [10]. That example, however, was robust to digital manipulations such as filtering and compression, but not transmission over an indoor acoustic channel.

Thus, this work is the first demonstration of a practical, perceptually-scaled watermark that is robust to the indoor acoustic channel and doesn't require prior synchronization.

1.2. Paper Outline and Summary of Contributions

In Section 2 we describe encoder and decoder of the watermarking system. In Section 3 we describe the perceptual normalization scheme that determines the amplitude of each watermark frame. In Section 4 we describe the results of listening tests. We offer concluding remarks in Section 5.

2. Watermarking System

In this section we provide a brief recapitulation of the spread spectrum watermarking encoder and the self-correlation decoder first presented in [2].

2.1. Spread-Spectrum Watermark Encoding

We denote the watermark frame length as $T = 10$ ms and the host audio's sampling frequency as f_s (the system works for any reasonable value of f_s). The watermark band is defined by the range $[f_L, f_H]$—in our case $f_L = 3$ kHz and $f_H = 4$ kHz. We define the watermark bandwidth $W = f_H - f_L$. We define the frame length in samples as $N = f_s T$, the number of frames in the watermark as M, and the number of DCT frequency bins in the watermark band as $L = 2WT$ (in our case, $L = 20$). The length N DCT can be identified with the matrix $\boldsymbol{F}$ (we assume a unitary normalization so that $\boldsymbol{F}\boldsymbol{F}^T = \boldsymbol{I}$.

We consider the NM samples of the host audio that need to be watermarked and divide them into frames, denoted by $x_1(1), \ldots, x_N(1), x_1(2), \ldots, x_N(2), \ldots, x_1(M), \ldots, x_N(M)$ and identified with the vectors

$$\boldsymbol{x}(t) = (x_1(t), \ldots, x_N(t))^T \tag{1}$$

We then compute the DCT of each frame as

$$\hat{\boldsymbol{x}}(t) = \boldsymbol{F}\boldsymbol{x}(t) \tag{2}$$

We will refer to the center frequencies of each DCT bin as

$$f_i = \frac{(i-1)f_s}{2N} \tag{3}$$

and the indices of the first and last DCT bin in the watermark

Copyright © 2022 ISCA

10.21437/Interspeech.2022-183

band as

$$i_L = \frac{2Nf_L}{f_s}$$
$$i_H = \frac{2Nf_H}{f_s}$$

The watermark signal is generated by repeating a "spreading sequence", which is a pre-defined wide-band signal, and modulating the sign and amplitude of the sequence with each repetition. In [2], the spreading sequence is chosen as an eigenvector of a random matrix; effectively it is a random Gaussian process bandlimited to the watermark band. For this reason, in that paper it is referred to as the "eigenvector based layer (EBL)".

In this work, we propose a new spreading code that is designed to have a flat spectrum within the watermark band. We avoid the EBL terminology because it is not chosen using the eigenvector method in [2]. We define the spreading sequence in the frequency domain by $\boldsymbol{s} = F^T\hat{\boldsymbol{s}}$, where $\hat{\boldsymbol{s}} = (\hat{s}_1, \ldots, \hat{s}_N)$ and

$$\hat{s}_i = \begin{cases} \pm 1 & i_L \le i \le i_H \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where the choice of sign can be made arbitrarily for each bin. In [2], multiple orthogonal watermarks could be superimposed; they chose the spreading sequences to be eigenvectors of a random matrix. For flat-spectrum spreading codes, they could be chosen as the rows of a Hadamard matrix, if one exists for order L. (For $L = 20$, such a matrix does exist [11]). In what follows we consider only a single watermark.

Given these definitions, each watermarked audio frame is given by

$$\boldsymbol{y}(t) = P_{\boldsymbol{s}}^{\perp}\boldsymbol{x}(t) + a(t)k(t)\boldsymbol{s}, \tag{5}$$

where $P_{\boldsymbol{s}}^{\perp} = I - \frac{\boldsymbol{s}\boldsymbol{s}^T}{||\boldsymbol{s}||^2}$ is the projection orthogonal to $\boldsymbol{s}$, $k(t) \in \{-1, +1\}$ is a key sequence (an arbitrary sequence of +1/-1 values that must be known by a detector) and $a(t)$ is the amplitude of the tth watermark frame.

The orthogonal projection ensures that the host audio is not interfering with the signal in the subspace defined by the spreading sequence. Because the decoder uses correlation between frames, the amplitude $a(t)$ of each frame can be chosen independently. We refer to this choice as "normalization". Larger amplitudes lead to a higher probability of detection by the decoder; however, this also increases the chances that a listener can perceive the watermark. In [2], $a(t)$ was chosen to be proportional to $\boldsymbol{x}(t)$, with the fixed proportionality constant chosen manually by an expert listener for each host audio file to be watermarked. This was a time-consuming process that prevented fully-automated watermark encoding. In Section 3, we describe the main contribution of this paper: a fully automated technique for normalizing the watermark frames using psychoacoustic masking curves.

2.2. Watermark Detection

This watermarking system can use the same detector as in [2]. Here, we provide a brief description of that detector. We consider a segment of audio received from a microphone that is the same length as the watermark. We again divide these NM samples into frames denoted by $z_1(1), \ldots, z_N(1), z_1(2), \ldots, z_N(2), \ldots, z_1(M), \ldots, z_N(M)$ and identified with the vectors

$$\boldsymbol{z}(t) = (z_1(t), \ldots, z_N(t))^T \tag{6}$$

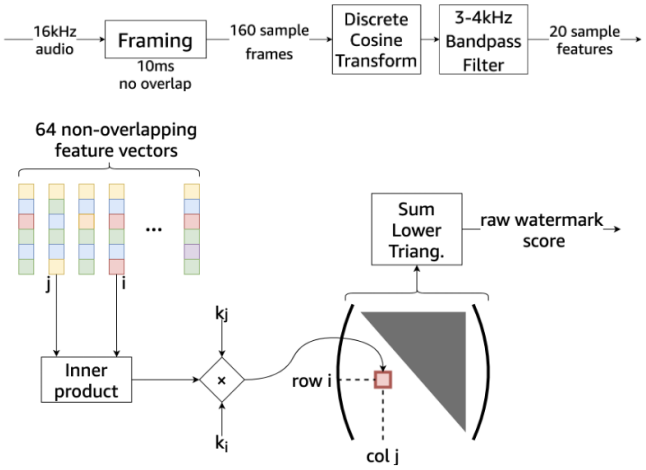

Figure 1: *Illustration of the computation of the watermark detector's features and raw score. In the matrix of correlation scores, index i corresponds to time s and index j corresponds to time t.*

We then compute the DCT of each frame and extract the bins in the watermark band using

$$\hat{\boldsymbol{y}}(t) = \boldsymbol{S}\boldsymbol{F}\boldsymbol{y}(t), \tag{7}$$

where S is a matrix consisting of the rows i_L through i_H of an identity matrix. Each feature vector $\hat{\boldsymbol{y}}$ is a vector of length L. The raw detection statistic is then given by the sum of the normalized, sign-corrected pairwise correlations between these feature vectors:

$$\rho(N) = \sum_{t=1}^{N-1}\sum_{s=t+1}^{N} k(t)k(s)\frac{\boldsymbol{y}(t)^T\boldsymbol{y}(s)}{||\boldsymbol{y}(t)||\,||\boldsymbol{y}(s)||} \tag{8}$$

As shown in [2], the mean and variance of the detection statistic under the null hypothesis in which no signal is present depends on the unknown acoustic channel. To correct with this, they proposed what is effectively a constant false alarm rate (CFAR) detector. A sequence of raw scores are computed over a sliding window that has the same length of the statistic. The CFAR score is computed by averaging the raw score over a short window and then normalizing it by the standard deviation of the score computed in a trailing window.

In particular, we define T_s, T_g, and T_n as the lengths of the signal window, gap, and noise window, respectively. The trailing mean of the score is computed as

$$\mu(N) = \frac{1}{T_n}\sum_{t=N-T_s-T_g-T_n+1}^{N-T_s-T_g}\rho(t) \tag{9}$$

and the trailing noise standard deviation as

$$\sigma(N) = \sqrt{\frac{1}{T_n}\sum_{t=N-T_s-T_g-T_n+1}^{N-T_s-T_g}|\rho(t) - \mu(N)|^2} \tag{10}$$

Finally, the CFAR statistic is computed as

$$\gamma(N) = \frac{1}{T_s}\sum_{t=N-T_s+1}^{N}\frac{\rho(t) - \mu(N)}{\sigma(N)} \tag{11}$$

.

The detection decision is then made by comparing the statistic to a threshold.

3. Perceptual Normalization

The main contribution of this paper is to describe a technique for automatically choosing the amplitude of each watermark frame to "hide" it under the host audio. One could imagine a naive approach of scaling the amplitude to be proportional to the amplitude of the host audio. However, there would be no universal scaling constant that would work for all host audio, or even for every frame within a single host audio clip. This is because the human auditory system (HAS) is a complicated, nonlinear system; the perceptibility of superimposed noise depends heavily on the spectral characteristics of the host audio and the noise.

Thus, in order to automatically choose the amplitude, we must incorporate knowledge about the HAS into our normalization scheme. To do this, we turn to the theory of psychoacoustic masking. We use a formula to compute a spectral mask from the host audio's spectrum. We then ensure that each watermark frame's amplitude is below the spectral mask so that it is inaudible. First, we convert the DCT bin frequencies to the Bark scale using the approximation [12]

$$z_i = 13\tan^{-1}(0.00076 f_i) + 3.5\tan^{-1}\left(\left(\frac{f_i}{7500}\right)^2\right) \quad (12)$$

Now we compute the *spreading function*, which determines how each frequency component of the host audio spreads to mask nearby frequencies. (This is unrelated to the *spreading code* described earlier; we simply have a collision of terminology between wireless communications and psychoacoustics.) Per [13]–[15], we compute the spreading coefficients in dB as

$$R_{ij}(t) = (z_i - z_j) \times \begin{cases} 31 & \text{if } i \le j \\ \min\left(-4, -24 - \frac{230}{f_j} + 2\log_{10}(\gamma|\hat{x}_j(t)|^2)\right) & \text{else.} \end{cases}$$

Here $\gamma = 10^{92/10}$ is a scaling factor for taking the digital audio samples, which we assume are scaled to the range $[-1, 1]$, and converting them to a sound pressure level (SPL) given a typical speaker volume and listening distance. This equation was developed by psychoacoustics experts to approximate experimental results.

This gives the relative masking level at frequency bin i caused by the energy in frequency bin j. We can then compute the masking level in dB for each frame as

$$m_i(t) = \max_j \left[10\log_{10}(|\hat{x}_j(t)|^2) + R_{ij}(t)\right]. \quad (13)$$

This gives the energy in dB that can be added to the ith frequency bin while remaining imperceptible. Since the spreading code has been chosen such that the energy in each bin is 0 dB, the desired power level of the spreading code in this frame is given (in dB) by

$$a_{\text{dB}}(t) = \min_{i_L \le i \le i_H} m_i(t). \quad (14)$$

We convert this to a linear scale to obtain

$$a(t) = 10^{a_{\text{dB}}(t)/20} \quad (15)$$

This level ensures that the level of the spreading code in each frequency bin is below the corresponding mask level. We cannot independently control the power in each bin without breaking the self-correlation.

The computation of the psychoacoustic mask requires computing the DCT of each frame, which is an $O(N \log N)$ operation (where N is the length of the frame), and the computation of the spreading function (which is $O(N^2)$). Without this perceptual computation, however, the only way to choose the appropriate scale to balance perceptibility and detectibility would be a manual listening test, a much more time-consuming process.

4. Experiments

4.1. Listening Tests

Thirteen participants in the listening tests were each given five examples to attempt to detect the watermark. For each example, the participants first listened to a clean example of the host audio. Then they listened to two other audio clips, one identical to the clean audio, the other watermarked, and were asked to choose which (if any) were different from the original (or to state that neither sounded different to them.). The results are shown in Table 1, along with p-values for a multinomial model where the null hypothesis is that the probability of correct and incorrect choices are identical. (In particular, p is the probability that a pool of random guessers would obtain at least as many correct answers and at most as many incorrect answers as our pool of respondents.) The listeners were unable to distinguish watermarked from unwatermarked for any of the samples at the 0.05 significance level. Furthermore, for each sample the vast majority of respondents said they could not hear the difference between either of the two test samples (one of which was watermarked) and the clean version. Of the 13 respondents, only four correctly identified any of the samples as being different from the clean version. However, in another test (not shown) we gave each respondent the same task, but with both samples *actually being identical*. Three of the four successful respondents incorrectly identified one of the identical, unwatermarked samples as being different from the clean one.

4.2. Detection Simulations

To evaluate the performance of such a system, we computed receiver operating characteristics (ROC) curves for the detector described in Section 2.2. To estimate false alarms, we used unwatermarked audio clips with the same length as the watermark. To estimate the detection rate, we used watermarked audio clips and used the peak watermark CFAR score. The audio clips are reverberated with room impulse responses (RIRs) output by a generative model that produces RIRs with a desired reverberation time. At any given threshold we can estimate the probability that the detection score crosses the threshold for the unwatermarked audio and for the watermarked audio, and using these values as the x- and y-coordinates, respectively. As the threshold is swept, curves are traced out. The results are shown in Figure 2 with ROC curves shown for RIRs with 60 dB reverberation time (RT60) of 1000, 2000, and 3000 ms. The results show that the watermark system is quite robust to significant amounts of reverberation.

Table 1: *Listening Results*

Sample ID	Correct	Incorrect	"Don't Know"	p
A	2	0	11	0.10
B	0	1	12	0.91
C	2	0	11	0.10
D	0	1	12	0.91
E	0	0	13	1.0

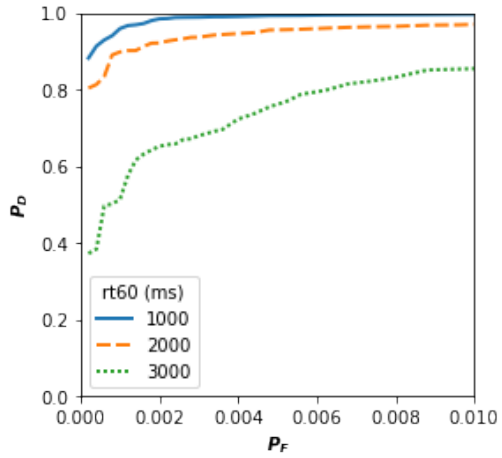

Figure 2: *Receiver operating characteristics (ROC) curves for the watermark detectors using simulations of reverberated audio with and without the watermark, with varying RT60 values for the reverberation shown to affect the results. The x-axis shows the probability of false alarm (P_F), i.e. the probability that a watermark is detected given unwatermarked audio; the y-axis shows the probability of detection (P_D). Up and left correspond to better performance, and it is evident from the plot that increased reverberation time results in poorer detection performance.*

5. Conclusions

In this paper, we demonstrated the use of psychoacoustic masking curves to optimally trade off an acoustic watermark's detectability by a device with its perceptibility to humans. We took an existing watermark encoder and normalized each frame to be proportional to the minimum level of the host audio's psychoacoustic mask within the watermark's band. Simulations showed that the same watermark detector could obtain reasonable performance, while a listening test showed that it was not perceptible to human listeners. This capability allows for the full automation of the watermark encoding process.

6. Acknowledgments

The author wishes to thank Sumit Garg, Chris Evans, Ahmed Abdelal, Nagaraj Mahajan, Joe Wang, Noah Stein, and Mike Rodehorst for helpful discussions.

7. References

[1] P. César, D. C. A. Bulterman, and J. Jansen, "Leveraging user impact: an architecture for secondary screens usage in interactive television," *Multim. Syst.*, vol. 15, no. 3, pp. 127–142, 2009.

[2] Y.-Y. Tai and M. F. Mansour, "Audio watermarking over the air with modulated self-correlation," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Brighton, United Kingdom: IEEE, May 2019, pp. 2452–2456. [Online]. Available:

[3] M. F. Mansour and A. H. Tewfik, "Time-scale invariant audio data embedding," *EURASIP J. Adv. Signal Process.*, vol. 2003, no. 10, pp. 1–8, Dec. 2003. [Online]. Available:

[4] C.-M. Pun and X.-C. Yuan, "Robust segments detector for de-synchronization resilient audio watermarking," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2412–2424, Nov. 2013, conference Name: IEEE Transactions on Audio, Speech, and Language Processing.

[5] X.-Y. Wang and H. Zhao, "A novel synchronization invariant audio watermarking scheme based on DWT and DCT," *IEEE Transactions on Signal Processing*, vol. 54, no. 12, pp. 4835–4840, Dec. 2006.

[6] Y. Xiang, I. Natgunanathan, S. Guo, W. Zhou, and S. Nahavandi, "Patchwork-based audio watermarking method robust to de-synchronization attacks," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, pp. 1413–1423, Sep. 2014.

[7] A. Nadeau and G. Sharma, "An audio watermark designed for efficient and robust resynchronization after analog playback," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 6, pp. 1393–1405, Jun. 2017.

[8] G. Del Galdo, J. Borsum, T. Bliem, A. Craciun, and S. Krägeloh, "Audio watermarking for acoustic propagation in reverberant en-

vironments," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 2364–2367, iSSN: 2379-190X.

[9] X. Zhang, D. Chang, W. Yang, Q. Huang, W. Guo, and Y. Zhao, "An audio digital watermarking algorithm transmitted via air channel in double DCT domain," in *2011 International Conference on Multimedia Technology*, Jul. 2011, pp. 2926–2930.

[10] M. D. Swanson, B. Zhu, A. H. Tewfik, and L. Boney, "Robust audio watermarking using perceptual masking," *Signal Processing*, vol. 66, no. 3, pp. 337–355, May 1998. [Online]. Available:

[11] K. J. Horadam, *Hadamard Matrices and Their Applications*. Princeton University Press, 2007.

[12] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models*. Springer Science & Business Media, Dec. 2006.

[13] P. Kabal, "An Examination and Interpretation of ITU-RBS.1387: Perceptual Evaluation of Audio Quality," Department of Electrical & Computer Engineering McGill University, Tech. Rep., 2002. [Online]. Available:

[14] T. Thiede, "PEAQ–The ITU standard for objective measurement of perceived audio quality," *J. Audio Eng. Soc.*, vol. 48, no. 1, p. 27, 2000.

[15] International Telecommunications Union, "RECOMMENDATION ITU-R BS.1387-1 - Method for objective measurements of perceived audio quality."