Digital Watermark for Real Musical Instrument Sounds Using Non-negative Matrix Factorization

Harumi Murata[1] and Akio Ogihara[2]
[1]Department of Information Engineering, School of Engineering, Chukyo University,
101 Tokodachi, Kaizu-cho, Toyota, 470-0393, Japan
[2]Department of Informatics, Faculty of Engineering, Kindai University,
1 Takaya Umenobe, Higashi-Hiroshima City, Hiroshima, 739-2116, Japan

Abstract: In this paper, we propose an audio watermarking method using nonnegative matrix factorization (NMF). The amplitude spectrogram of the observed signal is decomposed the basis matrix and the activation matrix, which are nonnegative matrices, by NMF. For embedding watermarks, we use the activation matrix. Onset time and offset time are estimated from the coefficients of activation matrix and this interval is defined as duration. The estimated notes are regarded as root notes and watermarks are embedded by operating the activation coefficients of dominant notes.

Keywords—**Audio watermarking, Nonnegative matrix factorization, Music theory**

1. Introduction

Digital watermarking is a technique to embed another digital data into digital contents such as music, images, and videos. For audio signals, the sound quality of the stego signal should not deteriorate as much as host signal. With current methods [1]–[7], high sound quality means that the difference between the host and stego signals is small. In these methods, watermarks are embedded by operating the components of the host signal. If noise resulting from embedding watermarks is perceived, it tends to be heard as an annoying sound.

However, host signal are not always shown to users in the actual systems that apply information hiding technology; rather, it is believed that host signals are not shown in many cases. Therefore, there is no problem even if another sound, except for the host signal, is perceived when the sound quality of the stego signal is maintained as music. In this case, we regard the sound quality of the stego signal as high in this paper.

Accordingly, we focus on a chord of music theory and embed the watermarks in consonance with host signal. A consonance is defined as the frequency ratio between notes consisting of a chord and is represented as simple whole numbers or its approximate value. In this paper, we use a diatonic chord as consonance. A diatonic chord is the most basic chord and it is composed of root, mediant, and dominant notes. A mediant is the third note from the root note and a dominant is the fifth note from the root note.

Moreover, we use nonnegative matrix factorization (NMF) [8], [9] for embedding watermarks. NMF is an algorithm that decomposes a nonnegative matrix that is an amplitude spectrogram of a target signal into two nonnegative matrices that corresponds to the spectral patterns of the target signal and the activation of each spectrum. A group of spectral patterns of the target signal is represented with the basis matrix and the intensity variation of each spectral pattern is represented with the activation matrix. In this paper, the watermarks are embedded into the activation matrix which is obtained by NMF.

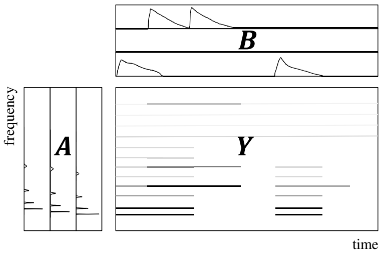

Figure 1. An example of applying NMF in the amplitude spectrogram.

The root notes are estimated from the activation matrix coefficients and the key of host signal are identified. The watermarks are embedded by operating the activation coefficients of dominant notes corresponding to the estimated root notes. A dominant is the fifth note from the root note and the root and dominant notes are relationship of consonance. Furthermore, the watermark signal becomes instrumental sound in the proposed method because the basis matrix is composed by spectral patterns of instrumental sound. Hence, even if the notes which are not included in host signal are perceived, we consider that there is no problem in case that these notes are arranged based on music theory.

2. Nonnegative Matrix Factorization (NMF)

2.1 Outline of NMF

NMF is one of the techniques used for separation of an audio mixture that consists of multiple instrumental sources. The following equation represents the decomposition of a simple NMF,

$$\boldsymbol{Y} \simeq \boldsymbol{A}\boldsymbol{B}, \tag{1}$$

where $\boldsymbol{Y}$ is an observed nonnegative matrix, which represents the time-frequency amplitude spectral components obtained via short-time Fourier transform (STFT), and $\boldsymbol{A}$ and $\boldsymbol{B}$ are nonnegative matrices. In addition, the matrix $\boldsymbol{A}$ is called basis matrix which represents spectral patterns of observed spectrogram $\boldsymbol{Y}$, and $\boldsymbol{B}$ is called activation matrix which involve activation information for $\boldsymbol{A}$. Figure 1 shows an example of applying NMF in the amplitude spectrogram.

Moreover, the multiplicative update algorithms of standard NMF based on Euclidean distance are shown in Eqs.(2) and (3).

$$a_{m,k} = \frac{[\boldsymbol{Y}\boldsymbol{B}^{\mathrm{T}}]_{m,k}}{[\boldsymbol{A}\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}]_{m,k}} a_{m,k}, \tag{2}$$

$$b_{k,n} = \frac{[\boldsymbol{A}^{\mathrm{T}}\boldsymbol{Y}]_{k,n}}{[\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{B}]_{k,n}} b_{k,n}, \tag{3}$$

where $a_{m,k}$ and $b_{k,n}$ are the elements of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$.

2.2 constrained supervised NMF

In the standard NMF, it is difficult to gather factorized spectral bases as spectral components of one musical instrument. Moreover, there is possibility of including spectral components of multiple instruments in one basis. Hence, it is extremely difficult to extract only specific instrumental information. Therefore, NMF is applied to spectrogram of specific instrumental sound as prior learning.

A specific instrumental sound of observed signal is regarded as a supervised signal. This supervised signal is transformed to the STFT domain and the amplitude spectrogram $\boldsymbol{Y}_{\mathrm{target}}$ is obtained. NMF is applied to the obtained spectrogram as shown in Eq.(4), and it is factorized into two non-negative matrices.

$$\boldsymbol{Y}_{\mathrm{target}} \simeq \boldsymbol{F}\boldsymbol{Q}. \tag{4}$$

The basis matrix $\boldsymbol{F}$ is regarded as teaching information of a specific instrumental sound.

Using the basis matrix $\boldsymbol{F}$ which is obtained by prior learning, the spectrogram of observed signal is factorized as follows:

$$\boldsymbol{Y} \simeq \boldsymbol{F}\boldsymbol{G} + \boldsymbol{H}\boldsymbol{U}. \tag{5}$$

Here, $\boldsymbol{G}$ is the activation matrix corresponding to the basis matrix $\boldsymbol{F}$. $\boldsymbol{H}$ is the basis matrix except for $\boldsymbol{F}$ and ideally represents the elements except for the specific instrumental sound.

3. Proposed Audio Watermarking Method Using NMF

In this section, we explain about proposed audio watermarking method using NMF.

The host signal is divided into frames of L samples and is transformed to the STFT domain with 50% overlap between successive frames. The amplitude spectrogram $\boldsymbol{Y}$ is decomposed into the basis matrix $\boldsymbol{A}$ and the activation matrix $\boldsymbol{B}$. Here, objective function is obtained by the square of the Euclidean distance between each column of $\boldsymbol{Y}$ and its approximation $\boldsymbol{A}\boldsymbol{B}$.

In the standard NMF, it is difficult to gather factorized spectral bases as spectral components of one musical instrument. Moreover, there is possibility of including spectral components of multiple instruments in one basis. Hence, it is extremely difficult to extract specific instrumental information. Moreover, the basis matrix $\boldsymbol{A}$ would like to be fixed to prevent the extreme change of the decomposition results of embedding and extracting process. Therefore, NMF is applied to spectrogram of specific instrumental sound as prior learning, and the factorized spectral basis matrix is used as teaching information. The teaching information is regarded as the basis matrix $\boldsymbol{A}$, and NMF is applied to the amplitude spectrogram $\boldsymbol{Y}$. The activation matrix $\boldsymbol{B}$ is obtained by NMF and it is used for embedding watermarks.

Next, onset time and offset time are estimated from the obtained activation matrix $\boldsymbol{B}$. For the activation matrix coefficients corresponding to a basis of $\boldsymbol{A}$, the duration between the estimated onset time and offset time is defined as one note. One-bit watermark is embedded in one note. The estimated notes are regarded as root notes, and we operate the activation matrix coefficients corresponding to notes which become consonance with root notes. Here, consonance is defined as the frequency ratio between notes consisting of a chord and is represented as simple whole numbers or its approximate value. In this paper, we use a diatonic chord as consonance. A diatonic chord is the most basic chord and it is composed of root, mediant, and dominant notes. A mediant is the third note from the root note and a dominant is the fifth note from the root note. Moreover, the root note is set to seven notes include in the key of music and has seven kinds of triads and tetrads that are composed by only note included in the scale of that key. Hence, a key of the host signal needs to be estimated, and watermarks should be embedded using suitable consonances based on the estimated key.

3.1 Key detection

The key of the host signal is detected after note estimation. The key has major scale and minor scale, but a chord is uniquely determined if seven notes which is required for key detection are estimated. Hence, in this paper, we estimate seven notes, but we have no discussion about identification of major scale or minor scale. The process of key detection is described below.

The pitch names are identified from the basis matrix corresponding to the activation coefficients which are estimated as notes. If seven pitch names are identified, a key which includes the identified pitch names are regarded as the key of host signal. Moreover, if the number of the estimated pitch names is less than seven, we calculate the correlation the estimated pitch names and the pitch names of each key. The key which has the maximum correlation value is regarded as the key of host signal. On the other hand, if the number of the estimated pitch names is more than seven, the number of notes corresponding to each pitch name is counted. Seven pitch names are identified in descending order of number of notes, and the key including them are regarded as the key of host signal.

However, the key detection procedure operates to identify the key for embedding watermarks, and there is no problem even if the estimated key is different from the real key of host signal in this paper.

3.2 Embedding watermarks

One-bit watermark is embedded into each estimated note. Before embedding watermarks, the activation coefficients are quantized in order to improve tolerance against attacks as shown in Fig.2. Next, the dominant note is specified from the root note. The root and dominant notes are relationship of consonance, and the watermarks are embedded by operating the activation coefficients of dominant notes. Hence, even if the notes which are not included in host signal are perceived,

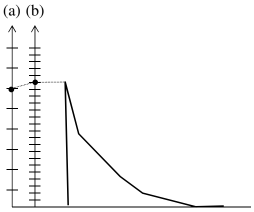

Figure 2. The activation coefficients: (a) after quantization, (b) before quantization.

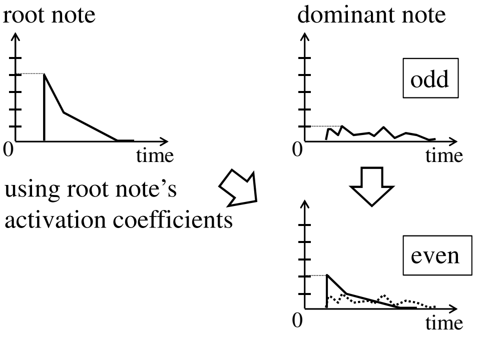

Figure 3. An example of embedding watermark bit '0'.

we consider that there is no problem in case that these notes are arranged based on music theory.

Next, the maximum value of activation coefficients of dominant note is calculated and the watermarks are embedded by even or odd of this value. If the maximum value of activation coefficients of dominant note is even, this condition is defined as the watermark bit '0'. On the other hand, the maximum value of activation coefficients of dominant note is odd, this condition is defined as the watermark bit '1'. If an embedding watermark bit and the condition of a frame are different, the activation coefficients are amplified by adding its root note's activation coefficients which is changed the magnification.

An example of embedding watermark bit '0' is shown in Fig.3. The condition of dominant note and an embedding watermark bit are different, and the activation coefficients of dominant note are amplified using its root note's activation coefficients.

After embedding watermarks in all notes, the amplitude spectrogram $\boldsymbol{Y'}$ are obtained by Eq.(6).

$$\boldsymbol{Y'} = \boldsymbol{AB'}, \tag{6}$$

where $\boldsymbol{B'}$ is the activation matrix after embedding watermarks. The amplitude spectrogram $\boldsymbol{Y'}$ is inverse short-time Fourier transformed and the stego signal is obtained. In addition, the phase information uses that of the host signal.

3.3 Extracting watermarks

In the same manner as the embedding process, the stego signal is divided into frames of L samples and is transformed to the STFT domain with 50% overlap between successive frames. The amplitude spectrogram $\boldsymbol{Y'}$ is decomposed the basis matrix $\boldsymbol{A}$ and the activation matrix $\boldsymbol{B'}$. Here, the basis matrix $\boldsymbol{A}$ uses the same matrix with embedding process. The root notes are estimated from the coefficients of activation matrix $\boldsymbol{B'}$ and the dominant notes are specified. If the maximum value of activation coefficients of dominant note is even, the watermark bit '0' is extracted. Otherwise, the watermark bit '1' is extracted.

4. Experiments

To confirm the validity of the proposed method, we conducted an evaluation of its tolerance against various attacks based on the evaluation criteria for audio information hiding technologies [10] and a subjective evaluation of sound quality. For testing, we used 10 pieces of music selected from [11] of 60 seconds duration at a 44.1-kHz sampling rate and with the stereo channel. These are simple piano etude for beginner and number of notes is small. Moreover, the watermarks were embedded in left channel and the embedding capacity of watermarks was 1.72 bps on average.

The frame length L of STFT was 8192 sample. The supervised signal of basis matrix $\boldsymbol{A}$ was used 36 notes of a single piano sound from octave 3 to octave 5 made by Cubase Artist7. As prior learning, NMF was applied for a supervised signal and the basis matrix was obtained. This basis matrix was used as $\boldsymbol{A}$.

4.1 Tolerance against attacks

We examined the tolerance against following attacks.

- MP3 128 kbps joint stereo
- MP3 128 kbps (joint stereo) tandem coding
- MPEG4 HE-AAC 96 kbps
- Gaussian noise addition (overall SNR 36 dB)
- Bandpass filtering 100 Hz – 6 kHz, -12 dB/oct.

The bit error rate (BER) of the watermarks is expressed as

$$\mathrm{BER} = \frac{\text{number of error bits}}{\text{embeddingbits}} \cdot 100\ [\%]. \tag{7}$$

The BER must be less than 10% according to the criteria [10].

Table 1 lists the BER results. The BERs were less than 10% for all attacks and it was confirmed that the tolerance to these attacks are high.

4.2 Sound quality

As a simple trial listening experiment, we conducted mean opinion score (MOS) for the subjective evaluation of sound quality involving two test subjects. Test subject scored the sound on a scaled from 1 to 5, a score of 1 being lowest quality and a score of 5 being highest quality. The average score of all stego signals was 3.3. From this result, sound quality of stego signal needs to be more improved.

However, the timbre of watermark signal was similar to the instrumental sound because the basis matrix $\boldsymbol{A}$ was represented by the spectral patterns of piano sound in prior learning. Therefore, it is considered that the sound quality of stego signal is not necessarily low.

Table 1. BER [%]: (A) MP3, (B) MP3 tandem coding, (C) MPEG4 HE-AAC, (D) Gaussian noise addition, (E) Bandpass filtering.

music number	BER [%]				
	(A)	(B)	(C)	(D)	(E)
1	7.50	6.25	2.50	1.25	6.25
2	6.89	5.75	3.45	0	2.30
3	6.35	7.94	9.52	0	3.17
4	7.43	6.08	7.43	1.35	3.38
5	9.62	3.85	3.85	1.92	1.92
6	7.81	7.81	1.04	0.52	1.56
7	7.37	7.37	3.16	0	5.26
8	8.47	8.47	3.39	0	5.08
9	4.67	6.54	0.93	0	3.74
10	6.67	6.67	6.00	0	2.67
average	7.28	6.67	4.13	0.50	3.53

5. Conclusions

We proposed an embedding method using NMF for real instrumental sound and studied the validity of this proposed method. From the experimental results, the proposed method was tolerant against attacks and the timbre of watermark signal was similar to the real instrumental sound. However, one-bit of watermark was embedded in one chord, and it is difficult to say that the embedding capacity of watermarks is high. Hence, we will improve the embedding capacity of watermarks for future work.

Acknowledgment

This work was supported by Grants-in-Aid for Scientific Research (KAKENHI) 26870681, and 26330214. proceedings.

References

[1] W.N. Lie and L.C. Chang, “Robust and high-quality time-domain audio watermarking based on low-frequency amplitude modification,” IEEE Trans. on Multimedia, vol.8, no.1, pp.46-59, 2006.

[2] L. Boney, A.H. Tewfik, and K.N. Hamdy, “Digital watermarks for audio signals,” Proc. IEEE Intl. Conf. on Multimedia Computing and Systems, vol.17, no.02, pp.473-480, 1996.

[3] P. Bassia, I. Pitas and N. Nikolaidis, “Robust audio watermarking in the time domain,” IEEE Trans. Multimedia, vol.3, no.2, pp.232-241, 2001.

[4] S. Shin, O. Kim, J. Kim and J. Choil, “A robust audio watermarking algorithm using pitch scaling,” Proc. IEEE 14th Int. Conf. Digital Signal Processing, pp.701-704, 2002.

[5] X. Li and H.H. Yu, “Transparent and robust audio data hiding in subband domain,” Proc. IEEE Int. Conf. Information Technology: Coding and Computing, pp.74-79, 2000.

[6] S.-S. Kuo, J.D. Johnston, W. Turin and S.R. Quackenbush, “Covert audio watermarking using perceptually tuned signal independent multi-band phase modulation,” Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, vol.II, pp.1753-1756, 2002.

[7] H. Murata, A. Ogihara and M. Uesaka, “Sound quality evaluation for audio watermarking based on phase shift keying using BCH code,” IEICE Trans. Inf. & Syst., vol.E98-D, no.1, pp.89-94, 2015.

[8] D.D. Lee, and H.S. Seung, “Algorithms for non-negative matrix factorization,” Neural Inf. Process. Syst., vol.13, pp.556-562, 2001.

[9] K. Yagi, Y. Takahashi, H. Saruwatari, K. Shikano. K. Kondo, “Music signal separation by orthogonality and maximum-distance constrained nonnegative matrix factorization with target signal information,” Proc. Audio Engineering Society 45th International Conference, pp.142-147, 2012.

[10] IHC Evaluation Criteria and Competition, `http://www.ieice.org/iss/emm/ihc/IHC_criteriaVer4.pdf`, Accessed December 21, 2015.

[11] “Classic anthology 100 played easily for adults,” Rittor-Music, 2009.