Improving Instrumentality of Sound Collage Using CNMF Constraint Model

Sora Miyaguchi[1], Naotoshi Osaka[1], Yusuke Ikeda[1]

[1]Tokyo Denki University
23fmi32@ms.dendai.ac.jp, {osaka,yusuke.ikeda}@mail.dendai.ac.jp

Abstract. In this study, the improvement in a new audio effect called sound collage, whereby one sound waveform (target sound) is synthesized using another sound waveform (element sound), is investigated. We propose a new model of convolutional NMF (CNMF) with constraints. And we compared the performance of three methods: the original CNMF, the new CNMF constraint model, and modified of Driedger's NMF (non-negative matrix factorization method). Sound collage sounds are synthesized using a combination of animal calls as the target sound and several instrumental sounds as the element sounds. Psychological experiments are conducted to evaluate the extent to which the target sound and instrumental character, namely reproducibility and instrumentality, are demonstrated. The results confirm that the instrumental nature of the synthesized sounds for both models improve compared with CNMF.

Keywords: CNMF, NMF, Audio mosaicking, sound collage, instrumentality

1 Introduction

Audio effects have applications in various domains, such as game music and animation; furthermore, new effects are desired to achieve richer expression. Previously, we have analyzed an effect called “sound collage” or “audio mosaicking” whereby one sound waveform (target sound) is synthesized using another sound waveform (element sound). Our interest here is a case of an environmental sound as a target sound and instrumental sound as an element sound. Furthermore, we have studied several methods to improve the performance of this effect. Additionally, we have defined two indices for evaluating this effect: (1) reproducibility, which is the degree to which the target sound is represented, and (2) instrumentality, which is the degree to which the sound is perceived to be instrumental.

Previously, we proposed a method for sound collage based on nonnegative matrix factorization (NMF) for sound source separation [1]. This method reproduces the sound by fitting a very short frame of the element sounds, and it has very high reproducibility;

however, it has low instrumentality owing to the destruction of the temporal structure of the element sounds. To overcome this limitation, Ikeda et al. proposed a method that improved on Driedger's NMF method with three constraints [2] by adding one more constraint (NMF_DM) [3] and a method using convolutive NMF [4]. Although NMF_DM improved the instrumentality, the convolutional NMF (CNMF) method at that time was not guaranteed to be an optimal solution and was impractical.
Later, a new optimal solution was reported [5], and a revised method based the new CNMF was proposed by the authors [6]. As the CNMF method can treat all part of element sound as a single basis, the temporal structure of element sounds is preserved; however, the same sound is repeated multiple times in a short period of time, thus rendering difficulty in perceiving instrumentality. In this study, a horizontal proximity restriction was added to the temporal activation of the CNMF method to create a CNMF constraint model (CNMF_C), and a sound collage was synthesized.

Herein, we compared the three methods, including the new method, and conducted psychological experiments to improve both the instrumentality and reproducibility to the greatest extent possible.

2 Sound collage

2.1 Sound collage with NMF

The NMF algorithm decomposes matrix *V* into the product of matrices *W* and *H*, with error matrix *C*, as follows.

$$V = W \times H + C \qquad (1)$$

To estimate *W* and *H*, *C* is minimized with various criteria, such as by using Frobenius norm. *W* and *H* are not estimated by an analytical method but rather as an optimization problem, wherein the error *C* with the original data is reduced through iterative computation.

In audio signal processing, *V* is a spectrogram. Therefore, *W* consists of spectra of the target sound (basis matrix) and *H* is the temporal activation corresponding to the basis matrix. The original NMF is a supervised algorithm, which simultaneously estimates *W* and *H*. However, in sound collage, we adopted unsupervised algorithm, where the spectrogram is synthesized, considering the target sound to *V* and spectra of element sounds as the basis matrix to *W*, and only the time-axis activation *H* is estimated.

This method can represent the target sound with significantly high reproducibility because it fits a very short frame of the element sound; however, the temporal structure of the element sound is destroyed, which renders difficulty in perceiving the instrumentality of the element sound.

2.2 Sound collage with NMF modified model

To improve the instrumentality, a model which preserves the temporal structure is necessary. Driedger proposed an improved NMF, NMF_D [2], which imposes constraints on the estimated activation matrix with respect to

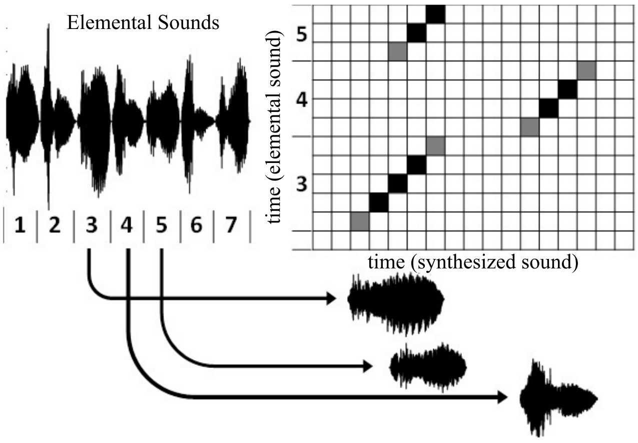
1. Horizontal Repetition Restriction: This constraint limits the repetition of spectra within a certain interval along the horizontal direction of the activation matrix.
2. Polyphony Inhibition in Activation Matrix: The proposed polyphony-restricted activation matrix suppresses the presence of multiple sounds within a single frame.
3. Enhanced Element Sound Continuity: Another constraint aims at enhancing the continuity of element sounds, which is manifested as diagonal patterns in the activation matrix.

However, the NMF_D is insufficient to improve instrumentality as it synthesizes only a portion of the element sound, rather than the whole. Moreover, this modification results in degradation caused by the synthesis of sound solely from the power spectrum, devoid of phase information.

To improve the instrumentality, we modify Driedger's third proposal with the addition of the following constraint referred to as NMF-DM:

1. Instead of utilizing a portion of an element sound, the entire sound is employed from start to finish.
2. To prevent any compromise in sound quality, the original waveform is retained, while the amplitude is drawn from the activation result, which is different from normal Griffin-Lim method [7].

Fig.1 depicts the correspondence between element sounds as basis and synthesized signal using modified activation (inverted upside down). Both are drawn in wave instead of spectrogram in convenience.



Synthesized Sound

Fig. 1. Modification of activation and preservation of temporal structure in NMF-DM.

2.3 Sound collage with CNMF

The CNMF algorithm does not differ from NMF in its basic structure of decomposition in the form of a product of matrices; however, the basis matrix is decomposed into a third-order tensor. For the length of the sound, the prescribed matrix W is provided, thus allowing for the preservation of temporal ordering. The structural schematic is shown in Fig. 2 and is expressed as follows.

$$V \approx \sum_{t=0}^{T} W(t) \times H^{t\rightarrow} \tag{2}$$

where the right arrow (→) indicates that the matrix is shifted t to the right and 0 is assigned to the vacant space.

Previously, we studied sound collage using CNMF [8]; in this CNMF version, the value of the evaluation function did not decrease monotonously, and the result could not be guaranteed as an optimal solution. However, in 2019, Dylan Fagot et al. proposed a new method to explore the optimal solution of the evaluation function [5]. We applied this method to implement a new sound collage synthesis method. Despite the mathematical optimization, this method has a problem in that the same sound is played multiple times in a short period of time, which causes stuttering and degrades instrumentality.

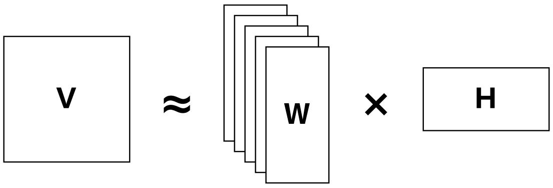


Fig. 2. Structural schematic of CNMF.

2.4 Sound collage CNMF with constraint model (CNMF_C)

We proposed a new model CNMF_C that prevents temporal proximity caused by an estimated activation. The algorithm that imposes a constraint to the activation is shown in Fig 3. It modifies the activation as a post-processing of CNMF such that only a dominant (local maximum) value survives in a fixed interval and the rest approaches zero as the iteration continues, as shown in the most outside loop.

Fig. 3. Constraint algorithm for *H* in CNMF_C (constraint part only).

```
w = appropriate frame length
for i = 1 → N_iteration
  for k_number = 1 → number of element sounds
    R = frame length of the element sound
    for j = 1 → R
     j0 = argmax_[j-w,j+w](H(j))
      if j == j0
        H(j) remains unchanged.
      else
         % Damping H(j)
        H(j)=H(j)*(1-(i + 1/N_iteration))
      end
    end
  end
end
```

3 Experiments

We considered three models: CNMF (as baseline), CNMF_C, and NMF_DM, and executed psychological evaluation test. In the experiment, sound sources were played back randomly; furthermore, six male and four female experimental collaborators in their 20s were asked to rate the reproducibility and instrumentality of the two items in an opinion test (five-category test).

3.1 Experiment details

For the experiment, animal calls were used as the target sound and instrumental sounds were used as the element sounds. The experimental parameters are listed in Table 1. Furthermore, element sounds of each number are presented in Table 2. For the experiment participants, the target and instrument of the synthesized sound to be heard were written in advance on an evaluation sheet. In addition, each original sound was also demonstrated in advance. The synthesized sound is also available at the following website.

https://acl.im.dendai.ac.jp/index.php/team/sora-miyaguchi/

Regarding the instrumentality, the participants were asked to evaluate the degree to which the synthesized sound resembled an instrument sound on a 5-point scale (1~5). They were instructed to give a score of 5 if it felt very similar. For reproducibility, the participants were asked to evaluate how close the synthesized sound was to the target sound on a 5-point scale (1~5). They were instructed to give a score of 5 if it felt very close.

Table 1. Experimental parameters.

Target sound	Frog, cicada, horse, elephant
Element sound	Marimba, Accordion, Metallophone, violin (single note, glissando, trill, pizzicato)
Evaluation method	MOS
Test participants	6 men and 4 women

Table 2. Element sound of each number.

	Target Sound	Element Sound
1	Cicada	Violin (Glissando), Metallophone
2	Frog	Marimba
3	Frog	Violin
4	Frog	Accordion
5	Elephant	Marimba
6	Elephant	Violin (Glissando), Accordion
7	Elephant	Violin (Single note)
8	Elephant	Violin (Glissando, Trill), Accordion
9	Horse	Marimba
10	Horse	Accordion
11	Horse	Violin

3.2 Comparison of CNMF and CNMF_C with CNMF

The instrumentality and reproducibility results of the experiment are shown in Figs. 4 and 5, respectively. The 95% confidence intervals are indicated on the bar graph. Compared with the CNMF, both CNMF_C and NMF_D exhibited higher instrumentality, as shown in Fig. 4; thus, the instrumentality improved. By contrast, for reproducibility, the evaluation changes significantly depended on the target sound, as shown in Fig. 5. In particular, when the elephant was used as the target sound, the results for CNMF_C were significantly lower.

Comparing CNMF and CNMF_C, the evaluation was higher for instrumentality except for conditions #6 and #7. On the other hand, for reproducibility, the evaluations were low except for conditions #3 and #11.

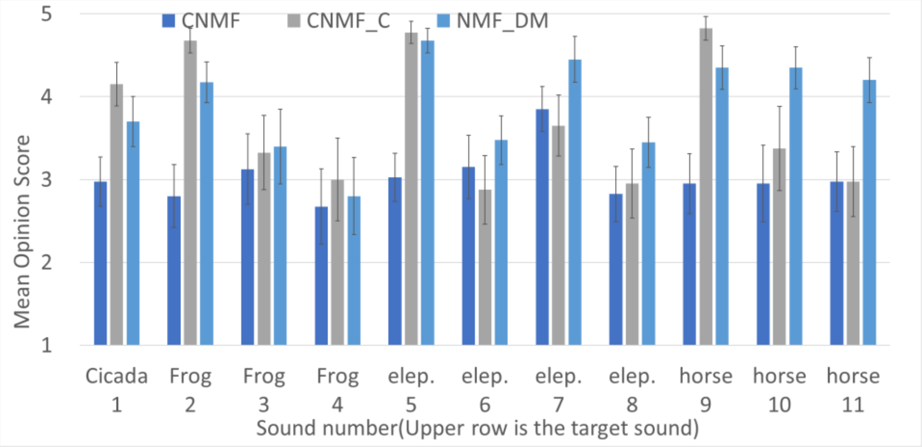


Fig. 4. Evaluation results of the three models for instrumentality.

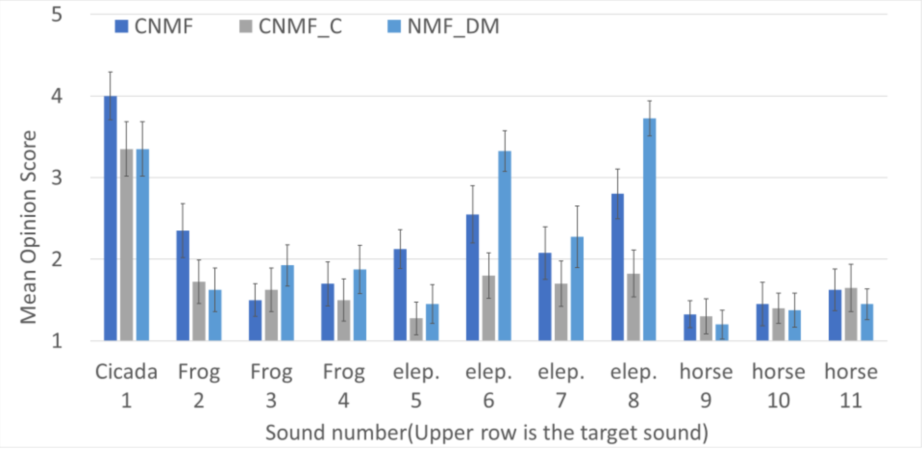


Fig. 5. Evaluations results of the three models for reproducibility.

3.3 Discussion

Compared to CNMF, CNMF_C showed improved instrumentality in almost all conditions, except when the target sound is an elephant. On the other hand, the reproducibility results were lower under almost all conditions. This demonstrates that limiting the temporal proximity of element sounds affects instrumentality. Additionally, combining CNMF with CNMF_C makes it possible to control the trade-off between instrumentality and reproducibility. Instrumentality and reproducibility are a trade-off: when one rises, the other falls.

Certainly, NMF_DM showed higher instrumentality than CNMF_C depending on the combination of element sounds and target sounds. This indicates that NMF_DM has the potential to show higher instrumentality than CNMF_C with the appropriate combination of timbres. However, In terms of controllability, CNMF_C, a generalization of CNMF, is higher because of its wider control over reproducibility and instrumentality. Therefore, in designing sound collages, CNMF_C, which allows moderate control of the two, is desirable and will better meet the user's needs.

4 Conclusions

In this paper, we compared three methods: CNMF_C, CNMF, and NMF_DM. Through experimentation, it was demonstrated that, by adding constraints in CNMF_C, the instrumental quality could be improved in most cases compared to CNMF, although the reproducibility decreased. This suggests that when CNMF_C is defined as a generalization of CNMF, there is a trade-off between instrumental quality and reproducibility, and that this trade-off can be controlled. Since users of Sound collage should be able to synthesize at their preferred level of reproducibility, we can say that our research was successful in improving the performance as Sound collage. Certainly, in some conditions, NMF_DM showed higher instrumental score than CNMF_C. However, controllability is important in sound collage, so CNMF_C can be considered more suitable in this study than NMF_DM.

In the future, we plan to explore ways to improve the performance of CNMF_C, such as finding more appropriate combinations of sounds, and examining finer control over instrumental quality and reproducibility. Furthermore, we will attempt to define a comprehensive evaluation in scalar values, incorporating both subjective evaluation and yet to be defined physical evaluation.

References

1. Tanaka, M., Osaka, N.: Sound quality evaluation of sound collage using NMF 2022 Autumn meeting of the Acoustical Society of Japan, 1-1-19, (2022). (In Japanese)
2. Driedger, J et al.: LET IT BEE-Towards NMF-Inspired audio mosaicking, Proc. of the 16th ISMIR, Malaga, Spain (2005).
3. Masaya, I., Osaka, N.: Synthesis of sound collage using NMF, IPSJ, Vol. 2020-MUS-126, No. 8, 1–6 (2020). (In Japanese)
4. Paris Smaragdis, Non-negative Matrix Factor Deconvolution; Extraction of Multiple Sound Sources from Monophonic Inputs, ICA 2004: Independent Component Analysis and Blind Signal Separation, pp 494-499 (2004).
5. Fagot, D. et al.: Majorization-minimization Algorithms for Convolutive NMF with the Beta-divergence, ICASSP 2019, Brighton, UK.
6. Miyaguchi, S., Osaka, N.: Improvement of instrumentality for sound collage using CNMF, 2023 Spring meeting of the Acoustical Society of Japan, 1-9-20, (2023). (In Japanese)
7. D. W. Griffin and J. S Lin, Signal Estimation from modified Short-Time Fourier Transform, ASSP-32, April 1984, pp. 236-242 (1984)
8. “nmf - toolbox”, https://github.com/colinvaz/nmf-toolbox, last accessed 2023/05/03.