MULTI-SOURCE MUSAICING USING NON-NEGATIVE MATRIX FACTOR 2-D DECONVOLUTION

Hadrien Foroughmand
UMR STMS (IRCAM - CNRS - UPMC)
hadrien.foroughmand@ircam.fr

Geoffroy Peeters
UMR STMS (IRCAM - CNRS - UPMC)
geoffroy.peeters@ircam.fr

ABSTRACT

A recent trend is to use Music Information Retrieval algorithms for *creativity*. When considering the audio signal as observation, a well-known method of data-driven synthesis is the "concatenative synthesis" also named *musaicing* (audio mosaicing) [7]. We propose here to extend the musaicing framework by including recent advances in audio source separation. We propose a method which allows reproducing a target music track by superimposing in time and frequency constitutive elements obtained using source separation methods. [1]

1. INTRODUCTION

Musaicing based on concatenative synthesis: Some musical practices consist in temporally concatenating slices of audio from a set of *source* music tracks to reproduce a *target* music track. In order to automate this technique, the slices are specifically chosen based on their similarity (through audio features) with the ones of the target such as to best reproduce the temporal evolution of the acoustic properties of the target. This method is known as "concatenative synthesis" [5] or *musaicing* (audio mosaicing) [7]. A limitation of this method is that only a single audio slice of the source (representing the whole frequency spectrum) is used at each time.

Musaicing based on NMF: Recently Driedger has proposed in [1] to extend this paradigm by allowing the use of multiple slices at each time. For this, a Non-Negative Matrix Factorization (NMF) paradigm is used.

The NMF algorithm [3] allows to factorize a non-negative matrix X into two non-negative matrices: a basis (or atoms) matrix W and an activation matrix H: $X \approx \hat{X} = W \cdot H$. When applied to audio, X is often the magnitude spectrogram of the signal. While in usual NMF, the basis matrix W is estimated, in [1] it is imposed as directly the spectrogram of a source sound. We denote it by W_{source}. Only the activation matrix H is estimated such that the superposition of the basis slices W_{source} best reproduce the target spectrogram X_{target}:

$$X_{target} \approx \hat{X}_{target} = W_{source} \cdot H \quad (1)$$

[1] **Acknowledgement**: Work partly founded by the EU Horizon 2020 research and innovation program under grant agreement no 761634 (Future Pulse project).

As an example, X_{target} can be chosen as the spectrogram of "Let it be" by The Beatles, W_{source} can be imposed as the spectrogram of "bee buzzing". H is then estimated to best reproduce X_{target}. $\hat{X}_{target}$ is then the spectrogram of "Let it be" sung by bees. The audio signal is then reconstructed using Griffin & Lim iterative algorithm [2] to reconstruct the phase of $\hat{X}_{target}$.

Constraints: To maintain the temporality and the timbre properties of the sources, [1] proposes three constraints which are applied during the iterations of the algorithms:
1. Limiting successive activations of the same basis
2. Limiting simultaneous activations of different basis
3. Favoring the temporal order of the slices as present in the source spectrogram

Limitations of Driedger method:

1) In [1], the basis W_{source} are directly the spectrogram slices. If the source track is made of simultaneously playing instruments, each basis therefore represents a mixture. The "bees buzzing" are hopefully a single monophonic "instrument". $\Rightarrow$ In our proposal, in order to be able to use source tracks with multiple simultaneous polyphonic instruments, we decompose those into their constitutive elements. We do this using the recently proposed NMF 2-D Deconvolution (NMF2D) algorithm [4] described below.

2) In [1], the basis W_{source} provided by the NMF algorithm only represent slices without any temporal evolution. Because of this, Driedger added the "favoring the temporal order" constraint. $\Rightarrow$ In our proposal, the basis provided by the NMF2D directly represent time evolution (as with NMF-Deconvolution algorithm [6]).

3) In [1], since the basis W_{source} are directly used to regenerate the target, those should be able to represent the various potential pitches existing in it. To guarantee this, he previously manually pitch-shift the source spectrogram to any possible pitches. $\Rightarrow$ In our proposal, the basis provided by the NMF2D are shift-invariant over frequency and the NMF2D algorithm automatically estimate the best pitch-shifting of those to reproduce the target.

We illustrate in Figure 1 the main differences between Driedger method and our proposal.

2. PROPOSED MUSAICING METHOD

The method we propose here - first separate the source tracks into their constitutive elements (their characteristic spectral patterns) - then use those to regenerate the target. Both steps are achieved using the NMF2D algorithm :

NMF2D: In the NMF, W is a basis matrix (each basis is a frequency slice). In the NMFD, W becomes a basis tensor (each basis is a time and frequency pattern). The basis

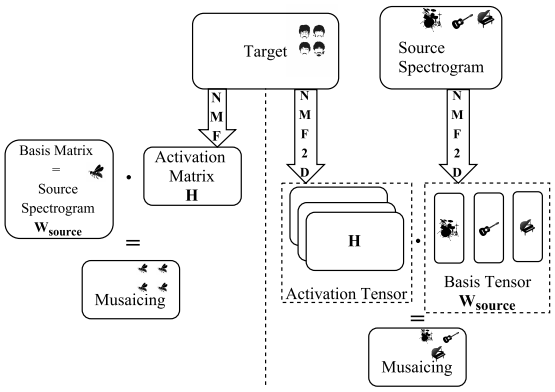

Figure 1. [Left] Driedger musaicing method using NMF and no source separation [Right] our proposal using NMF2D and source separation.

of duration T are convolved in time τ with the activations. The NMF2D (Schmidt & Mørup in [4]) extends the NMFD by making H also a tensor. H now represents not only the activation over time of each basis but also its frequency transposition factor. The basis are now convolved in time but also in frequency $\phi \in [0, P-1]$ with H.

$$X \approx \hat{X} = \sum_{\tau=0}^{T-1} \sum_{\phi=0}^{P-1} \overset{\downarrow\phi}{W^{\tau}} \overset{\rightarrow\tau}{H^{\phi}} \tag{2}$$

To achieve the invariance over frequency, the method is applied to a log-frequency representation: the constant-Q transform (CQT). We expect that the basis will represent the prototype spectral envelope of the various instruments such that when transposed they represent the spectrum of a note played by a given instrument.

In a **first step**, the NMF2D is used to decompose the source tracks into their constitutive elements. We therefore estimate both W and H. Only W is used for the remaining; it is the set of basis (time and frequency patterns) used for the musacing. We denote it by W_{source}.

In a **second step**, the NMF2D is applied to the target spectrogram and only the activation tensor H is estimated; the basis tensor is imposed as $W = W_{source}$.

$$X_{target} \approx \hat{X}_{target} = \sum_{\tau=0}^{T-1} \sum_{\phi=0}^{P-1} \overset{\downarrow\phi}{W^{\tau}_{source}} \overset{\rightarrow\tau}{H^{\phi}} \tag{3}$$

The audio signal is then obtained from $\hat{X}_{target}$ using an adaptation of the Griffin & Lim algorithm to the CQT case. Our method is summarized in Figure 2. Compare

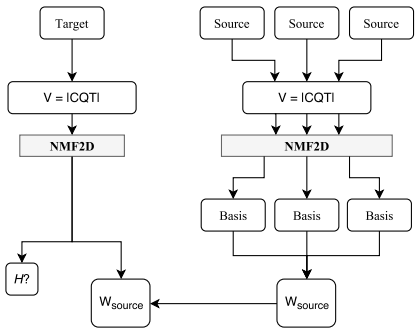

Figure 2. Musaicing by NMF2D.

to Driedger, our method allows to avoid the necessity to previously manually pitch-shift the sources. It also allows to avoid the use of the "maintaining temporal continuity" constraint. However, we still use the two constrains "limiting successive" and "simultaneous activations". The reason is that those two constraints allow to favor the recognition of the sources in the reconstructed target (a perfect reconstruction will not allow that).

3. PERCEPTUAL EXPERIMENT

In order to assess the performances of our method we set up an online perceptual experiment. In this, people were asked to compare various musaicing methods including Driedger one (NMF), another based on NMFD and our NMF2D proposal; each with or without constraints. Two targets and four sources were used for each method. 30 people participated to the test. People were asked to rate the following criteria: 1) the audio quality of the produced signal, 2) whether or not the method allows the recognition of the harmonic and temporal structure of the *target*, 3) of the acoustic characteristics of the *sources*, 4) the creative interest of the method. We analyzed the results in terms of mean and confidence interval for each question: The results strongly depend on the choice of the target and sources. When the sources are mono and stationary (such as bee buzzing), Driedger method (which does not separate the sources into constitutive elements) actually works very well. In the other case, the NMFD method allows a better conservation of the acoustic properties of the *sources* (this is due to the temporal deconvolution). The NMF2D musaicing allows to retain the structure of the *target* (this is due to the automatic transposition of the basis) especially in a multi-source case. The use of the constraints, allows to better maintain the characteristics of the *sources* but it is then more difficult to recognize the *target*. Finally, the methods that allow for better recognition of sources are deemed more creative. Numerical results : `http://recherche.ircam.fr/anasyn/expe/musaicing/`.

4. REFERENCES

[1] J. Driedger, T. Prätzlich, and M. Müller. Let it beetowards nmf-inspired audio mosaicing. In *ISMIR*, 2015.

[2] D. Griffin and J. Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1984.

[3] D. D Lee and H S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 1999.

[4] M. Schmidt and M. Mørup. Nonnegative matrix factor 2-d deconvolution for blind single channel source separation. In ICA, 2006.

[5] D. Schwarz, G. Beller, B. Verbrugghe, and S. Britton. Real-time corpus-based concatenative synthesis with catart. In DAFx, 2006.

[6] P. Smaragdis. Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In ICA, 2004.

[7] A. Zils and F. Pachet. Musical mosaicing. In *Digital Audio Effects (DAFx)*, 2001.