# Music retiler: Using NMF2D source separation for audio mosaicing

Hadrien Foroughmand, Geoffroy Peeters

# Music retiler: Using NMF2D source separation for audio mosaicing

Hadrien Foroughmand Aarabi
UMR STMS (IRCAM - CNRS - UPMC)
Paris, France
foroughmand@ircam.fr

Geoffroy Peeters
UMR STMS (IRCAM - CNRS - UPMC)
Paris, France
peeters@ircam.fr

## ABSTRACT

Musaicing (music mosaicing) aims at reconstructing a target music track by superimposing audio samples selected from a collection. This selection is based on their acoustic similarity to the target. The baseline technique to perform this is concatenative synthesis in which the superposition only occurs in time. Non-Negative Matrix Factorization has also been proposed for this task. In this, a target spectrogram is factorized into an activation matrix and a predefined basis matrix which represents the sample collection. The superposition therefore occurs in time and frequency. However, in both methods the samples used for the reconstruction represent isolated sources (such as bees) and remain unchanged during the musaicing (samples need to be pre-pitch-shifted). This reduces the applicability of these methods. We propose here a variation of the musaicing in which the samples used for the reconstruction are obtained by applying a NMF2D separation algorithm to a music collection (such as a collection of Reggae tracks). Using these separated samples, a second NMF2D algorithm is then used to automatically find the best transposition factors to represent the target. We performed an online perceptual experiment of our method which shows that it outperforms the NMF algorithm when the sources are polyphonic and multi-source.

## CCS CONCEPTS

• **Applied computing → Sound and music computing**;

## KEYWORDS

Music Information Retrieval (MIR), Mosaicing, NMF2D, Sound Synthesis

**ACM Reference Format:**
Hadrien Foroughmand Aarabi and Geoffroy Peeters. 2018. Music retiler: Using NMF2D source separation for audio mosaicing. In *Proceedings of Audio Mostly 2018: Sound in Immersion and Emotion (AM'18).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3243274.3243299


*AM'18, September 12–14, 2018, Wrexham, United Kingdom*

ACM ISBN 978-1-4503-6609-0/18/09…$15.00
https://doi.org/10.1145/3243274.3243299

## 1 INTRODUCTION

*Musaicing* (music mosaicing) aims at reproducing a target audio music track by using "elements" derived from a collection of source audio files.

Unlike other Music Information Retrieval (MIR) researches, the goal here is not to develop algorithms that estimate the most precise content description (to perform afterwards auto-tagging or playlist generation) but to develop algorithms based on content description that allow re-generation of music or the generation of new music. We believe such research fields will attract more and more interest in the MIR community.

Applications of such methods range from music creation to music repurposing (such as the automatic ring-tone generation proposed by [16]).

### 1.1 Related works

Over the years, different methods have been proposed to perform musaicing under various names (including corpus-based synthesis) [2, 3, 6, 7, 12–14, 16, 19]. We only review here the most important ones.

*1.1.1 Concatenative synthesis based musaicing.* The first method was proposed by Zils et al. [19] who actually invented the name "musaicing". In their method, a target sequence (a target music track) is reconstructed by selecting a sequence of audio samples from a predefined collection. The audio samples are selected to satisfy two constraints: 1) a segment constraint (which guarantees that the selected samples have the same loudness, pitch and timbre than the target segments), 2) a sequence constraint (which favours parameters continuity and recognition of the target temporal shape).

Therefore, the "tiles" used for the reconstruction are directly temporal audio slices taken from the collection of *source* music tracks. They are then temporally concatenated (using concatenative synthesis) without any superposition or modification to reproduce a *target* music track.

Such a method presents two main limitations.

**Limitation 1.** only a single "tile" is used for the reconstruction at each time,

**Limitation 2.** since the "tiles" are directly taken from the audio, they potentially encapsulate a polyphonic multi-source audio fragment (i.e. an instrument playing simultaneously several pitches and /or several instruments playing together). Better results are therefore obtained when the collection of samples used represent monophonic, single-source instruments.

*1.1.2 NMF-based musaicing.* To solve (Limitation 1), Driedger et al. proposed in [4] a method which relies on Non-Negative Matrix

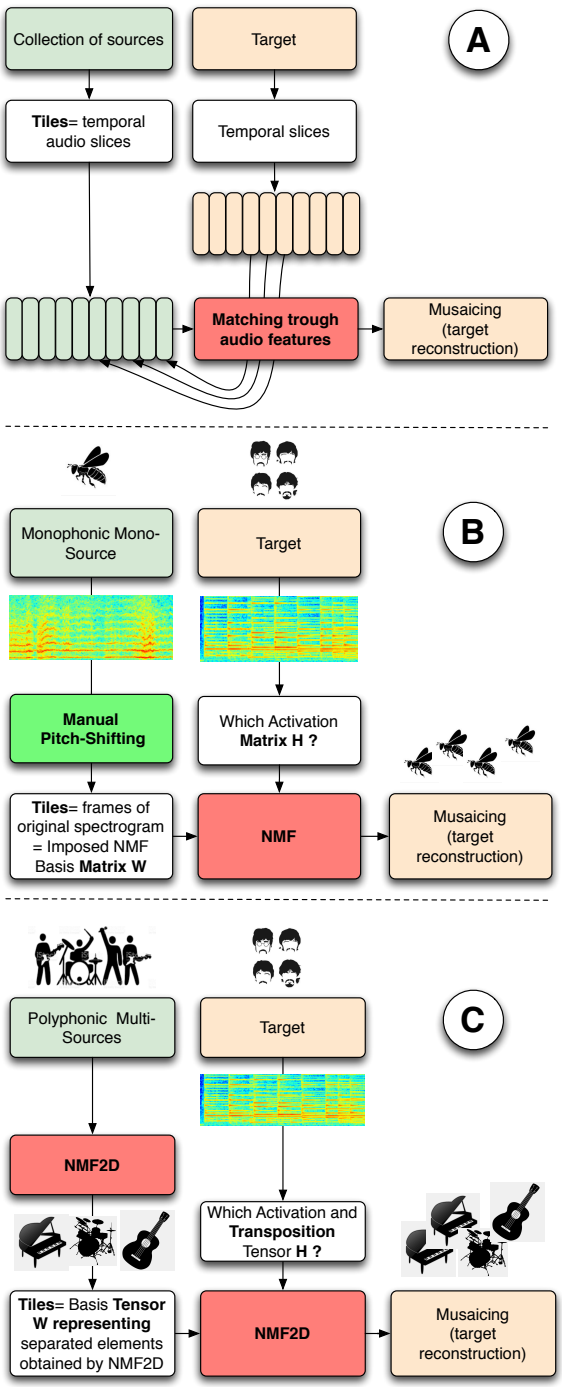


**Figure 1: Flowchart of Ⓐ a typical musaicing by concatenative synthesis system, Ⓑ Driedger musaicing method by NMF with imposed basis matrix W, Ⓒ Our proposed musaicing method in which sources are first decomposed into elements using NMF2D and then used for the reconstruction using NMF2D.**

Factorization (NMF)[1]. In [4], the "tiles" are defined as the successive frames of the Short Time Fourier Transform (STFT) of a source track. The target music track is also represented by the successive frames of its STFT. While in Zils method "tiles" are selected based on the similarity of their audio descriptors (loudness, pitch and timbre) to the target, in [4] the NMF algorithm is used to select those.

The NMF algorithm has been proposed by [8] to allow factorizing a given non-negative matrix $X \in \mathbb{R}_{\geq 0}^{N \times M}$ into two non-negative matrices: a *basis* (or atoms) matrix $W \in \mathbb{R}_{\geq 0}^{N \times K}$ and an *activation* matrix $H \in \mathbb{R}_{\geq 0}^{K \times M}$:

$$\underset{(n,m)}{X} \approx \underset{(n,m)}{\hat{X}} = \underset{(n,k)}{W} \cdot \underset{(k,m)}{H} \tag{1}$$

In audio, $X$ is usually chosen as the magnitude of the STFT and therefore $n$ denotes the number of frequencies, $m$ the number of time frames and $k$ the number of basis.

In Driedger et al. method, $X$ is the magnitude STFT of the target to be reconstructed. They impose the *basis* matrix $W$ to be the successive frames of the magnitude of the STFT of the source: $W = W_{tile} = |STFT[x_{source}]|$. The NMF algorithm is then only used to estimate the activation matrix $H$, i.e. the "tiles" which are necessary to reconstruct the magnitude STFT of the target $X_{target}$:

$$X_{target} \approx \hat{X}_{target} = W_{tile} \cdot H \tag{2}$$

In [4], the method is used to reconstruct $X_{target}$ = "Let it be" from The Beatles using $W_{tile}$ = pitched "bees buzzing" sounds. $\hat{X}_{target}$ then represents "Let it be" songs by "bees".

Compared to Zils method, the method of [4] makes it possible to use several "tiles" simultaneously for the reconstruction (Limitation 1 solved). Yet (Limitation 2) still applies since the "tiles" are directly taken as the STFT frames without any separation. The method thus works better when the source sounds represent a single instrument playing a single pitch at each time (such as "bees"). Also, in order to be able to reproduce the various pitches potentially existing in the target (such as the ones in "Let it be" from The Beatles), the source tracks should be manually pitch-shifted to all possible pitches.

**Details of the NMF method used in [4].** $H$ is estimated iteratively by minimizing a cost function between $X$ and $\hat{X}$. In [4], the Kullback-Leibler divergence is used for this since it is invariant to multiplicative factor of $X$:

$$D(X||WH) = \sum_{nm} X_{nm} \log\left(\frac{X_{nm}}{(WH)_{nm}}\right) - X_{nm} + (WH)_{nm} \tag{3}$$

Minimizing $D(X||WH)$ leads to the following multiplicative update of $H$ to be used at each iteration. For $X = X_{target}$ and $W = W_{tile}$ it is

$$H \leftarrow H \odot \frac{W_{tile}^{\intercal} \cdot \frac{X_{target}}{W_{tile} \cdot H}}{W_{tile}^{\intercal} \cdot J} \tag{4}$$

where $\odot$ is the Hadamard product (element-wise multiplication), $^{\intercal}$ denotes a matrix transposition and $J$ is a all-ones matrix of size $(n, m)$.

[1] It should be noted that the factorization using NMF for sound synthesis had been already proposed by Burred in [1] for cross-synthesis.

**Constraints added in [4].** To maintain the temporal and timbre properties of the "tile" sounds[2], Driedger et al. add three constraints which are applied sequentially at each iteration $\lambda_{it}$ of the NMF algorithm:

**Constraint 1.** To limit the successive activations over time of the same basis, [4] only keeps the highest activation in a fixed temporal neighborhood:

$$R_{k,m} = \begin{cases} H_{k,m} & \text{if } H_{k,m} = \mu^r_{k,m} \\ H_{k,m}(1-\lambda_{it}) & \text{otherwise} \end{cases} \quad (5)$$

where $\mu^r_{k,m}$ is the maximum value of $H_{k,m\in[m-r/2,m+r/2]}$ in a $r$-horizontal neighborhood.

**Constraint 2.** To limit the simultaneous activations of different basis at the same time, [4] only keeps the highest activation over the basis dimension $k$:

$$P_{k,m} = \begin{cases} R_{k,m} & \text{if } k \in \Omega^p_m \\ R_{k,m}(1-\lambda_{it}) & \text{otherwise} \end{cases} \quad (6)$$

where $\Omega^p_m$ is the $p$-maximum value of $R_{:,m}$.

**Constraint 3.** To favor ta temporal organisation of tiles similar to the one in the source [4] multiplies $P_{k,m}$ with a diagonal kernel of size (c,c):

$$C_{k,m} = \sum_{i=-c}^{c} P_{(k+i)(m+i)} \quad (7)$$

This lead to a sparse diagonal activation matrix.

The musaicing equation then becomes:

$$\hat{X}_{target} = W_{source} \cdot C \quad (8)$$

Given $\hat{X}_{target}$, the audio signal is reconstructed using the Griffin & Lim phase reconstruction iterative algorithm [5].

Such a method presents three main limitations.

**Limitation 3.** In [4], since the "tiles" (which are here the basis $W_{tile}$) are directly the magnitude of the STFT of the source, they potentially represent a polyphonic multi-source signal. Better results are therefore obtained when the sources are monophonic single-source instruments like [4] did with their "bees buzzing" sources. This is the same limitation as for the concatenative synthesis method

**Limitation 4.** In [4], each basis of $W_{tile}$ only represents a single STFT frame without any temporal evolution. This is the reason for (Constraint 3).

**Limitation 5.** In [4], since the basis $W_{tile}$ are directly (i.e. without any modification) used to regenerate the target, they should be able to represent the various potential pitches existing in the target. To guarantee this, they previously manually pitch-shifted the source spectrogram to any possible pitches.

## 1.2 Proposed musaicing method: music retiler

In this paper, we propose a new musaicing method named "music retiler" which allows us to solve the limitations (L1, L2, L3, L4, L5) mentioned above.

**Extracting the tiles.** In our method, we start from a collection of music tracks which can be polyphonic and multi-source. We first extract a set of time-frequency elements which become the tiles used to reconstruct the target music track.

For the estimation of these time and frequency constitutive elements we will use the NMF-2D Deconvolution (NMF2D) algorithm [10]. The NMF2D extends the NMF by adding a deconvolution over time (such as the NMFD) but also over frequency. The deconvolution over time allows to represent the temporal evolution of the timbre (like drums strokes with their resonance). The (Limitation 4) is therefore solved. The deconvolution over frequency allows to represent the spectral envelopes of pitched instruments independently of their pitch. The same basis tensor can therefore be transposed to represent the various notes of a given instrument.

As opposed to previous methods, our "tiles" therefore represent isolated notes and instruments (Limitation 3 solved). Our method can therefore be applied to source tracks with multiple simultaneous polyphonic instruments.

**Retiling the music** To reconstruct the target music track from these "tiles" we apply the same idea as Driedger et al. but extend it to the NMF2D case (deconvolution over time and frequency). We consider these "tiles" as the fixed basis tensor $W$ and estimate the activation $H$ of these "tiles" to be able to reconstruct the target track. Since several "tiles" can be used simultaneously (Limitation 1) is solved. In the case of the NMF2D, the activation tensor $H$ also represents the necessary pitch-shifting of the "tiles" to represent the target. Therefore, there is no need to "previously manually pitch-shift the source spectrogram" (as in Driedger et al. method) since these pitch-shiftings are automatically performed by the NMF2D algorithm. Therefore (Limitation 5) is also solved.

## 1.3 Paper organization

We present our method in section 2. We first review the NMFD (section 2.1), NMF2D algorithm (section 2.2) and explain how we use them to perform musaicing.

To obtain invariance over frequencies, we represent the audio signal in the Constant-Q-Transform domain (CQT). We therefore explain in section 2.3 how to reconstruct the audio signal in the time domain from the CQT by adapting the Griffin & Lim phase reconstruction iterative algorithm.

Finally, we evaluate the performances of our proposed NMF2D musaicing method in comparison to the NMF method of [4]. For this, we have set up an online perceptual experiment which results we discuss in section 3.

# 2 PROPOSED NMFD AND NMF2D SOURCE SEPARATION BASED MUSAICING

The method we propose first separates the source tracks into their constitutive elements (characteristic spectral patterns) then use those to reconstruct the target music track. In the following we propose to use the NMFD or the NMF2D algorithm to achieve this.

In a first step, the NMFD or the NMF2D algorithm is applied to a set of source music tracks in order to compute their constitutive elements which defines the set of "tiles".

In a second step, using these estimated "tiles" as basis (a basis tensor), we use the NMFD or NMF2D again to estimate the activations necessary to reconstruct the target.

[2]Those are named "source sounds" in [4]

The flowchart of NMF2D musaicing method is given in Figure 2 and detailed below.

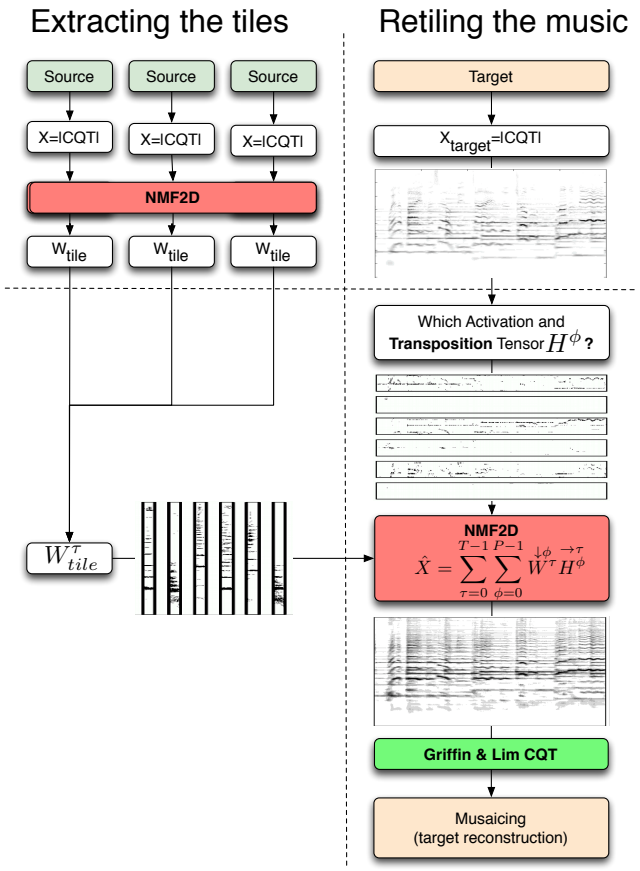


**Figure 2: Musaicing by NMF2D.**

## 2.1 NMFD

In [15], Smaragdis proposes the NMF-Deconvolution algorithm (NMFD). This method extends the NMF paradigm by making $W$ a basis tensor (instead of a basis matrix). Each basis in the $W^\tau$ tensor is a matrix representing the time evolution of the frequency spectrum. The basis of duration $T$ are convolved in time $\tau$ with the activations $H$.

$$X \approx \hat{X} = \sum_{\tau=0}^{T-1} W^\tau \overset{\tau\rightarrow}{H} \tag{9}$$

For our musaicing, we propose to apply twice the NMFD algorithm.

**Extracting the tiles.** First, we use the NMFD algorithm to estimate the set of "tiles" to be used for the reconstruction. For this we apply the NMFD algorithm to the set of original music tracks. In the NMFD, $W^\tau$ are updated iteratively using the followings:

$$W^\tau \leftarrow W^\tau \odot \frac{\frac{X}{\hat{X}} \cdot (\overset{\tau\rightarrow}{H})^\intercal}{J \cdot \overset{\tau\rightarrow}{H}} \tag{10}$$

$$H \leftarrow H \odot \frac{(W^\tau)^\intercal \cdot [\frac{\overset{\leftarrow\tau}{X}}{\hat{X}}]}{(W^\tau)^\intercal \cdot J} \tag{11}$$

The set of "tiles" to be used for the reconstruction $W^\tau_{tile}$ is then chosen as the resulting $W^\tau$.

**Retiling the music.** For the reconstruction of the target we use the NMFD algorithm again but we impose $W^\tau = W^\tau_{tile}$ and only update the activation $H$ using the same equation (9) but with $X = X_{target}$, $\hat{X} = \hat{X}_{target}$ and $W^\tau = W^\tau_{tile}$. At each iteration, the reconstructed target is given by:

$$\hat{X}_{target} = \sum_{\tau=0}^{T-1} W^\tau_{tile} \overset{\rightarrow\tau}{H} \tag{12}$$

In order to make the "tiles" more recognizable in the reconstructed target (we are of course not interested in a perfect reconstruction of the target since it would not allow us to recognize the "tiles"), we apply the constraints C1 and C2 (which favour sparsity) at each iteration of the update of H during the reconstruction.

**Pros and Cons.** The NMFD allows us to solve the above-mentioned limitations (L1), (L2) and (L3). Yet, limitation (L4) still applied and we still need to manually pitch-shift the sources. To solve this limitation we propose the use of the NMF2D algorithm described in the following.

## 2.2 NMF2D

In [9], Schmidt & Mørup propose the NMF2D-Deconvolution (NMF2D) algorithm. As in the NMFD algorithm, $W$ is a basis tensor $W_{tile}$ representing the time evolution of the frequency spectrum of each basis. In the NMF2D, $H$ is also a tensor. It represents the activation of each basis at every time step and at every frequency translation. For a given time, the same basis can therefore be activated at different frequency translations (hence at different pitches). A chord can therefore be represented by a single basis translated at difference frequencies. In the NMF2D, the basis are convolved in time but also in frequency $\phi \in [0, P-1]$ with the activations. The basis are thus both time and frequency-invariant.

$$X \approx \hat{X} = \sum_{\tau=0}^{T-1} \sum_{\phi=0}^{P-1} \overset{\downarrow\phi}{W^\tau} \overset{\rightarrow\tau}{H^\phi} \tag{13}$$

For our musaicing, the NMF2D algorithm is applied twice.

**Extracting the tiles.** We first use the NMF2D algorithm[3] to decompose a set of source music tracks into the set of "tiles": $W^\tau_{tile}$ (a set of time and frequency patterns). Using a KL-divergence as cost-function, the following update rules can be derived to update the tensors:

$$W^\tau \leftarrow W^\tau \odot \frac{\sum_\phi \overset{\uparrow\phi}{\frac{X}{\hat{X}}} \overset{\tau\rightarrow}{(H^\phi)^\intercal}}{\sum_\phi J \cdot \overset{\tau\rightarrow}{(H^\phi)^\intercal}} \tag{14}$$

$$H^\phi \leftarrow H^\phi \odot \frac{\sum_\tau \overset{\downarrow\phi}{(W^\tau)^\intercal} \overset{\leftarrow\tau}{\frac{X}{\hat{X}}}}{\sum_\tau \overset{\downarrow\phi}{(W^\tau)^\intercal} \cdot J} \tag{15}$$

[3] We use the SNMF2D Matlab Toolbox of Schmidt & Mørup [10]

The set of "tiles" to be used for the reconstruction $W^{\tau}_{tile}$ is then chosen as the resulting $W^{\tau}$.

**Retiling the music.** We then reconstruct the target music track $X_{target}$ imposing $W^{\tau}$ as $W_{tile}$. We therefore only estimate the activation tensor $H^{\phi}$, i.e. the weight and transposition factor of each $W^{\tau}_{tile}$ over time using eq. 14, 15. At each iteration, the reconstructed target is given by:

$$\hat{X}_{target} = \sum_{\tau=0}^{T-1} \sum_{\phi=0}^{P-1} W^{\tau}_{tile}{}^{\downarrow\phi} H^{\phi}{}^{\rightarrow\tau} \tag{16}$$

**Pros and Cons.** The NMF2D allows to solve all the above-mentioned limitations (L1), (L2), (L3) and (L4). However, we still use the two sparsity constrains (C1) and (C2). The reason for this being that the constraints favours the proper recognition of the sources in the reconstructed target (again, we are not interested in a perfect reconstruction of the target).

## 2.3 Griffin & Lim algorithm for the CQT

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 so that when transposed they represent the spectrum of a note played by an instrument.

In order to reconstruct the audio temporal signal from its CQT ($\hat{X}_{target}$), we need to reconstruct its phase. In the case of the STFT, the Griffin & Lim algorithm phase reconstruction iterative algorithm [5] is often used. In the case of the CQT, the representation is however bounded to a minimum and maximum frequency. We therefore apply a band-pass filter between those minimum and maximum frequencies to the random temporal signal used for the initialisation of the algorithm. We also replace the STFT and inverse STFT of the original algorithm [5] by the CQT and the inverse CQT. This is done using SchÃűrkhuber et al. implementation [11] which results in a perfect reconstruction of the CQT using non-stationary Gabor frames [17].

# 3 EXPERIMENT AND RESULTS

The objective of audio musaicing is to reproduce the harmonic and temporal structure of a target music track using the acoustic characteristics of a set of sources. It is therefore not possible to measure its quality using the usual source separation measures (as defined for example in the bsss_eval toolbox [18]). Instead, we have set up an online perceptual experiment asking people to rate the following criteria:

**Question 1** the audio quality of the produced signal,
**Question 2** whether or not the method allows the listener to recognize of the harmonic and temporal structure of the *target*,
**Question 3** whether or not the method allows the listener to recognize of the acoustic characteristics of the *sources*.
**Question 4** the creative aspect of the audio example

The rates are in the following range 1=bad, 2=poor, 3=fair, 4=good, 5=excellent.

With this, we asked people to rate Driedger et al. musaicing method (NMF) and our NMF2D proposal; each with or without the above mentioned-constraints (C1), (C2) and (C3).

**Parameters.** The audio result relatively depends on the parameters of the NMF2D. From several tests in the computation of audio examples, we have made the following observation. The number of basis $K$ chosen for extracting the "tiles" is linked to the number of instruments of each audio source (as the classic NMF algorithm). The frequency convolution factor $P$ is chosen to cover all the notes. The temporal convolution factor $T$ is more variable, by increasing $T$ the learned tiles will be longer ($T$ cannot be too long due to computation time).

We tested each method using two targets and two audio sources[4] (audio sources are used to extract "tiles"). The list of targets, source recordings and methods used for our experiment are listed in Table 1.

**Table 1: List of targets, source recordings and methods used for our experiment.**

| | |
|---|---|
| Targets | - Let It Be - The Beatles<br>- Funk jazz - Music Delta |
| Source | - Recordings of reggae music (two excerpts of Bob Marley songs)<br>- Recordings of five spoken vowels |
| Methods | - Driedger Musaicing (NMF) without/with constraints (C1,C2,C3)<br>- NMF2D Musaicing without/with constraints (C1,C2) |

20 people participated to the test (18 are from the audio signal processing field, 17 have already participated in a perceptual test, 5 women and 15 men, the average age being 30.1 years).

In Figures 3, 4, 5 and 6, we indicate the results in terms of mean rating and confidence interval ($\mu \pm 1.95\sigma$) for each question.

Looking at Figures 3, 4, 5 and 6, we first see that the results largely depend on the choice of sources (Reggae/Vowels). For example for Question 3 (Figure 5), the ratings are higher when the sources are "Vowels", whatever the choice of method or target. It is therefore difficult to compare the results from one pair of Target/Source to another.

For **Question 1 "Rate the audio quality of the produced signal"** (Fig. 3) we found better results – using Driedger w/o constraint for Beatles/Reggae, – using Driedger with constraint for Beatles/Vowels, – using our NMF2D w/o constraint for Funk/Reggae and – NMF2D w/o constraint for Funk/Vowels.

For **Question 2 "Rate whether or not the method allows you to recognize of the harmonic and temporal structure of the target"** (Fig. 4), the results strongly depend on the *source* used. When the *source* is polyphonic and multi-source (Reggae), the NMF2D w/o constraint is rated higher than Driedger method, this independently of the *target* (Beatles or Funk). When the source is monophonic and mono-source (Vowels), Driedger methods are rated better. Note that to apply the Driedger method to the Vowels,

[4]The generated audio examples are available here: https://www.dropbox.com/sh/9d8hw1bb9fnk7o7/AACyWYWIxXgxI6B1GbDJgfu-a?dl=0

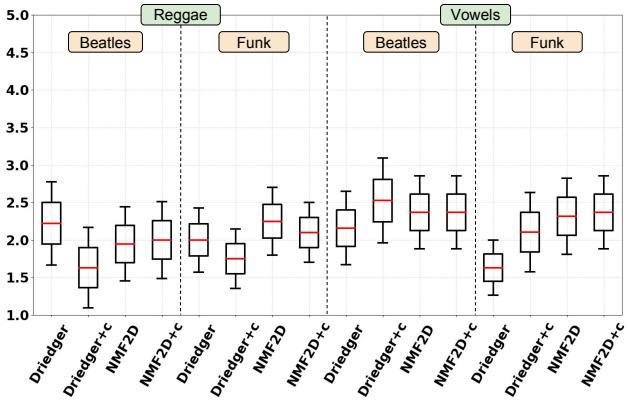


**Figure 3: Question 1: "Rate the audio quality of the produced signal".**

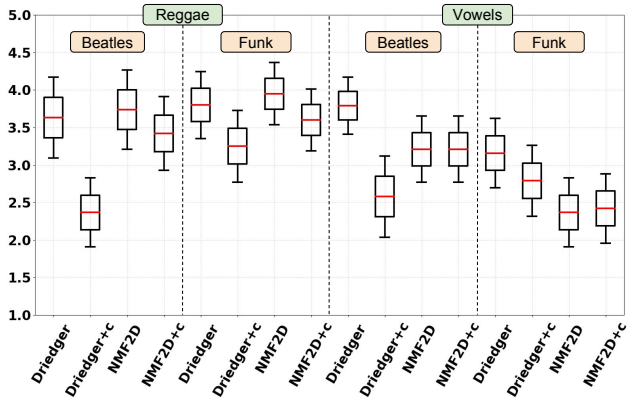


**Figure 4: Question 2: "Rate whether or not the method allows the recognition of the harmonic and temporal structure of the *target*".**

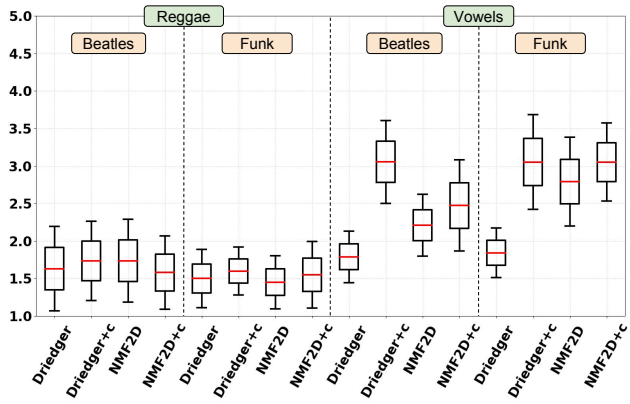


**Figure 5: Question 3: "Rate whether or not the method allows the recognition of the acoustic characteristics of the *sources*".**

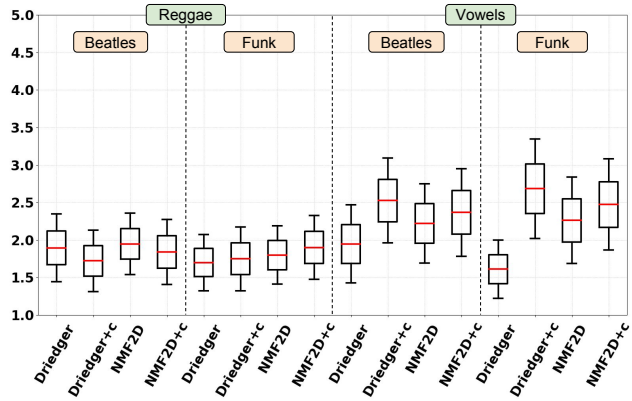


**Figure 6: Question 4: "Rate the creative aspect of the audio example".**

we had to manually pitch-shift the sources to all possible potential pitches. This is not the case for our NMF2D where the pitch-shifting is automatically done. For every methods, the non-constrained versions are rated better. This is explained by the fact that without any constraints, sources can be chosen more freely by the algorithm hence reproducing the target with more fidelity.

For **Question 3 "Rate whether or not the method allows you to recognize the acoustic characteristics of the sources"**, it is the opposite. Here, the constrained methods are rated better than the non-constrained ones. We also see that – when the *source* is Reggae, Driedger method and our NMF2D are judged equal, – when the *source* is Vowels, Driedger method with constraint is rated slightly better than our NMF2D with constraint (but the difference is not significant for Funk).

Finally, for **Question 4 "Rate the creative aspect of the audio example"**, – when the *source* is Reggae, both Driedger method and our NMF2D are again judged equal, – when the *source* is Vowels, Driedger method with constraint is again rated slightly better than the NMF2D with constraint (but the difference is not significant for Funk). Results are then very close between Question 3 and 4 which tends to show that people consider more "Creative" the "recognition of the acoustic characteristics of the sources" than the "recognition of the harmonic and temporal structure of the target". This also explains why the results are on average higher when using "Vowels" (which are more easely recognizable) than using "tiles" extracted from a set of Reggae tracks.

## 4 CONCLUSION

In this paper, we extended the possibilities of musaicing methods by proposing the use of the NMF2D. These algorithms are first used $n$-times to obtain "tiles" (a set of time/frequency patches) from a set of $n$ source music tracks. These "tiles" are then used to "retile" a target track using again the NMF2D algorithms but while imposing the basis tensors as "tiles". Using the NMF2D algorithm allows to solve the limitations we highlighted in previous methods: – the necessity to separate the sources so that the "tiles" represent monophonic single-source sounds, – the necessity to represent time evolution – the necessity to automatically transpose the "tiles" to the right

pitches. The signal is then re-synthesized with an adaptation we propose of the Griffin & Lim algorithm to the CQT case.

In order to assess the performances of our proposal, we have set up an online perceptual experiment comparing our NMF2D algorithm to the NMF algorithm proposed by [4]. We first showed that the results strongly depend on the choice of couple targets and "tiles" source sounds. We also showed that our NMF2D algorithm is more effective than the NMF algorithm when the "tiles" source sounds are polyphonic and multi-sources.

**Future works.** In our current system, "tiles" are extracted for each source of the collection separately. They are then concatenated to form $W_{tile}$. Therefore, some redundancy may exist in the learned "tiles". For future work, we would like to add a constraint in order to avoid this redundancy of "tiles". In our current system, all "tiles" have the same temporal duration. Another future work would be to automatically adapt this duration to the acoustic content of the "tiles" (drum sounds, voices, or sustained harmonic instruments).

## ACKNOWLEDGMENTS

This work has been partly founded by the EU Horizon 2020 research and innovation program under grant agreement no 761634 (Future Pulse project).

## REFERENCES

[1] Juan José Burred. A framework for music analysis/resynthesis based on matrix factorization. In *ICMC*, 2014.
[2] Graham Coleman, Esteban Maestre, and Jordi Bonada. Augmenting sound mosaicing with descriptor-driven transformation. In *Proceedings of DAFx*, 2010.
[3] Edward Costello, Victor Lazzarini, and Joseph Timoney. A streaming audio mosaicing vocoder implementation. 2013.
[4] Jonathan Driedger, Thomas Prätzlich, and Meinard Müller. Let it bee-towards nmf-inspired audio mosaicing. In *ISMIR*, pages 350–356, 2015.
[5] Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243, 1984.
[6] Jordi Janer and Maarten De Boer. Extending voice-driven synthesis to audio mosaicing. In *5th Sound and Music Computing Conference, Berlin*, volume 4, 2008.
[7] Ryoho Kobayashi. Sound clustering synthesis using spectral data. In *ICMC*, 2003.
[8] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.
[9] M. Mørup and M. N. Schmidt. Non-negative matrix factor 2-D deconvolution, apr 2006.
[10] Mikkel N Schmidt and Morten Mørup. Nonnegative matrix factor 2-d deconvolution for blind single channel source separation. In *International Conference on Independent Component Analysis and Signal Separation*, pages 700–707. Springer, 2006.
[11] Christian Schörkhuber, Anssi Klapuri, Nicki Holighaus, and Monika Dörfler. A matlab toolbox for efficient perfect reconstruction time-frequency transforms with log-frequency resolution. In *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*. Audio Engineering Society, 2014.
[12] Diemo Schwarz. A system for data-driven concatenative sound synthesis. In *Digital Audio Effects (DAFx)*, pages 97–102, 2000.
[13] Diemo Schwarz. Corpus-based concatenative synthesis. *IEEE signal processing magazine*, 24(2):92–104, 2007.
[14] Diemo Schwarz, Grégory Beller, Bruno Verbrugghe, and Sam Britton. Real-time corpus-based concatenative synthesis with catart. In *9th International Conference on Digital Audio Effects (DAFx)*, pages 279–282, 2006.
[15] Paris Smaragdis. Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In *International Conference on Independent Component Analysis and Signal Separation*, pages 494–499. Springer, 2004.
[16] Shih-Yang Su, Cheng-Kai Chiu, Li Su, and Yi-Hsuan Yang. Automatic conversion of pop music into chiptunes for 8-bit pixel art. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 411–415. IEEE, 2017.
[17] Gino Angelo Velasco, Nicki Holighaus, Monika Dörfler, and Thomas Grill. Constructing an invertible constant-q transform with non-stationary gabor frames. *Proceedings of DAFX11, Paris*, pages 93–99, 2011.
[18] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. 14(4):1462–1469, 2006.
[19] Aymeric Zils and François Pachet. Musical mosaicing. In *Digital Audio Effects (DAFx)*, volume 2, page 135, 2001.