LET IT BEE – TOWARDS NMF-INSPIRED AUDIO MOSAICING

Jonathan Driedger, Thomas Prätzlich, Meinard Müller
International Audio Laboratories Erlangen
{jonathan.driedger,thomas.praetzlich,meinard.mueller}@audiolabs-erlangen.de

ABSTRACT

A swarm of bees buzzing "Let it be" by the Beatles or the wind gently howling the romantic "Gute Nacht" by Schubert – these are examples of *audio mosaics* as we want to create them. Given a *target* and a *source* recording, the goal of audio mosaicing is to generate a *mosaic* recording that conveys musical aspects (like melody and rhythm) of the target, using sound components taken from the source. In this work, we propose a novel approach for automatically generating audio mosaics with the objective to preserve the source's timbre in the mosaic. Inspired by algorithms for *non-negative matrix factorization* (NMF), our idea is to use update rules to learn an activation matrix that, when multiplied with the spectrogram of the source recording, resembles the spectrogram of the target recording. However, when applying the original NMF procedure, the resulting mosaic does not adequately reflect the source's timbre. As our main technical contribution, we propose an extended set of update rules for the iterative learning procedure that supports the development of sparse diagonal structures in the activation matrix. We show how these structures better retain the source's timbral characteristics in the resulting mosaic.

1. INTRODUCTION

Using the sounds in a recording of buzzing bees to recreate a recording of the song "Let it be" by the Beatles is a typical example of an audio mosaic. In this example, the recording of the bees serves as *source*, while the Beatles recording is called the *target*. Ultimately, one should be able to identify the target recording when listening to the mosaic, but at the same time perceive the timbre of the source sounds. Therefore, the audio mosaic of "Let it be" with the bee recording could give the impression of bees being musicians, buzzing the song's tune.

Audio mosaicing is an interesting audio effect which has found its way into both artistic work as well as academic research. Artists like John Oswald used thousands of manually selected source audio snippets to create new musical compositions [1] and real-time audio mosaicing has been used by musicians as an instrument in live performances [4, 22]. Over the years, many different systems for audio mosaicing were proposed [1, 3, 5, 11, 13, 17, 18]. The core idea of most automated systems is to split the source into short audio segments, which are suitably concatenated afterwards to match spectral and temporal characteristics of the target [19].

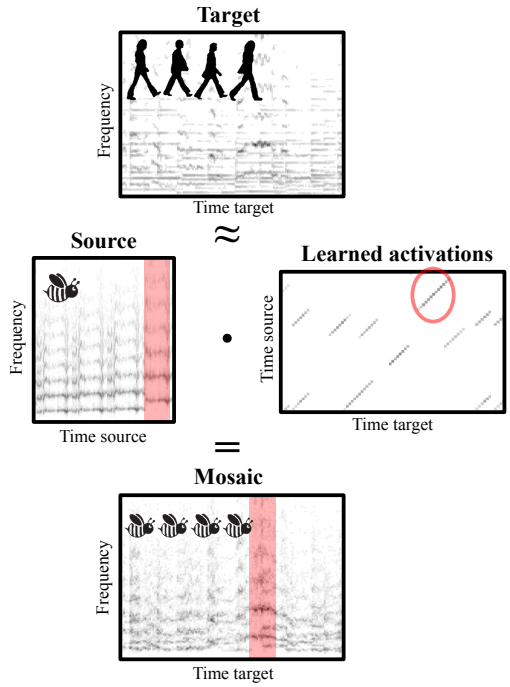

Figure 1. Schematic overview of our proposed audio mosaicing method. The sparse diagonal structures in the activation matrix are important in order to preserve the timbre of the source in the mosaic.

In this work, we propose a novel way to create audio mosaics. Our idea is to learn an *activation matrix* that, when multiplied with the spectrogram of the source recording, approximates the spectrogram of the target recording (see Figure 1). The source spectrogram hereby serves as a *template matrix* which is fixed throughout the learning process. This way, as opposed to many previous automated mosaicing approaches, a frame of the target can be resynthesized as the superposition of several spectral frames of the source , thus allowing "polyphony" of the source sounds.

[1] Especially on his album *Plexure* [16].

As a first contribution, we propose an audio mosaicing procedure which is inspired by well-known algorithms for *non-negative matrix factorization* (NMF) [14]. Keeping the template matrix fixed (the source's magnitude spectrogram), this basic procedure learns an activation matrix by iteratively applying a standard NMF update rule to a randomly initialized matrix. Experiments show that in case the source recording offers an appropriate amount of different sounds, this procedure can closely approximate the spectrogram of the target recording. However, the source's timbre is often barely recognizable in the resulting mosaics. The reason is that the procedure recreates every target frame independently, thus destroying temporal characteristics of the source in the final audio mosaic. Furthermore, the method can superimpose an arbitrary number of spectral frames from the source to construct a good numerical approximation of a single target frame. A superposition of a large number of source sounds may however result in a timbre that is no longer similar to the actual timbre of the source. Therefore, an exact approximation of the target's spectrogram cannot be our procedure's sole goal.

As our main technical contribution, we therefore propose an extended set of update rules that supports the development of sparse diagonal structures in the activation matrix during the learning process (see the activation matrix in Figure 1). Rather than single frames, diagonal structures activate whole frame sequences in their original order. This preserves the source's temporal characteristics in the resulting mosaic. Furthermore, the extended set of update rules also limits the number of simultaneous activations, making the learned activation matrix sparse and reducing the problem of too many source sounds being audible simultaneously. This way, we trade some approximation quality for a better preservation of the source's timbre.

The idea of activating sequences of frames is inspired by methods like *non-negative matrix factor deconvolution* (NMFD) and related formulations [20, 21], where template sequences of frames from a dictionary are activated by single activation values. However, our approach is conceptually different. Instead of changing the NMF problem formulation, our approach stays in the standard NMF setting, supporting the activation of whole frame sequences directly in the activation matrix with additional update rules. Besides being computationally very efficient and easy to implement, this also has the advantage that we do not need to choose a maximal length of the sequences as in NMFD. Similarly, the sparseness constraint imposed by our procedure is not enforced by penalty terms in the problem formulation (as for example in [8, 10, 12, 23]), but also by additional update rules.

The remainder of this paper is structured as follows. In Section 2 we introduce the basic concept of using NMF-inspired update rules for the task of audio mosaicing. In Section 3 we present the extended set of update rules that supports the development of sparse diagonal structures in a learned activation matrix. The effects of these update rules on the audio mosaics are discussed and demonstrated in Section 4.

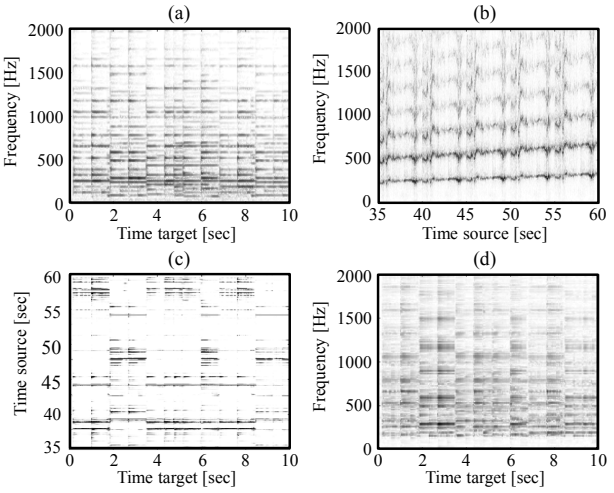

Figure 2. Basic NMF-inspired audio mosaicing. **(a):** Magnitude spectrogram of "Let it be" V (target). **(b):** Magnitude spectrogram of a recording of bees W (source). **(c):** Activation matrix H. **(d):** The product WH (mosaic).

2. BASIC NMF-INSPIRED AUDIO MOSAICING

Non-negative matrix factorization (NMF) has been applied very successfully in a large variety of music processing tasks and beyond. Given a non-negative matrix $V \in \mathbb{R}_{\geq 0}^{N \times M}$, the goal of NMF is to decompose this matrix into two factors $W \in \mathbb{R}_{\geq 0}^{N \times K}$ and $H \in \mathbb{R}_{\geq 0}^{K \times M}$, where $N, M, K \in \mathbb{N}$. The distance between the product WH and the matrix V is minimized with respect to some distance measure, for example the Kullback-Leibler divergence

$$(V||\,WH) = \sum_{nm} V_{nm} \log \frac{V_{nm}}{(WH)_{nm}} - V_{nm} + (WH)_{nm}. \quad (1)$$

In the context of music processing, the matrix V is usually a magnitude spectrogram of a music recording, the matrix W is interpreted as a set of spectral templates, and the matrix H constitutes an activation matrix. Non-zero values in a row of H activate the associated template in W at the respective time instance. The two factors W and H are usually learned by iteratively applying multiplicative update rules to two suitably initialized matrices [14].

Fixing the template matrix W to be the magnitude spectrogram of the source recording, the basic idea of our proposed audio mosaicing approach is to learn only the activation matrix H. More precisely, we proceed as follows. Given the target recording x_{tar} and the source recording x_{src}, we first compute the complex valued spectrograms X_{tar} and X_{src} by applying the short-time Fourier transform (STFT) to both recordings. Afterwards, we set $V := |X_{tar}|$, $W := |X_{src}|$, and randomly initialize $H^{(1)} \in (0, 1]^{K \times M}$. Fixing a number of iterations L, we then iteratively update H with

$$H_{km}^{(\ell+1)} = H_{km}^{(\ell)} \frac{\sum_n W_{nk} V_{nm} / (W H^{(\ell)})_{nm}}{\sum_n W_{nk}}, \quad (2)$$

for $k \in [1:K]$, $m \in [1:M]$, and the iteration index $\ell \in [1:L-1]$. Finally, we set $H := H^{(L)}$. The learned activation matrix H is then multiplied with the complex valued X_{src}, yielding the complex valued spectrogram of the audio mosaic $X_{mos} := X_{src}H$. To compute the audio mosaic x_{mos}, we apply an "inverse" STFT to the spectrogram X_{mos} which also adjusts the phases such that artifacts from phase discontinuities are reduced [9].

Figure 2 shows this basic procedure applied to our running example. In Figure 2a we see an excerpt of the magnitude spectrogram of the song "Let it be". Our goal is to create an audio mosaic of this song, using the recording of buzzing bees, which can be seen in Figure 2b. To increase the range of different pitches occurring in our source, we used a pitch-shifting algorithm [6] to create differently pitched versions of the bee recording and concatenated them. Figure 2c shows an excerpt of the activation matrix H, derived by applying the basic procedure described above. A first observation about H is the predominance of horizontal activation structures. These patterns correspond to single spectral frames in the source which are activated repeatedly to mimic the stable spectral structures in the target. Although the resulting mosaic, shown in Figure 2d, closely resembles these spectral structures, one can hear a "stuttering" effect when listening to the reconstructed audio recording. This stuttering originates from the same frame of the source being repeated over and over again. In Section 3.1, we aim to prevent the learning process from activating the same frame in fast repetition with an additional update rule.

A second observation is that the matrix H usually activates many source frames simultaneously. The learning process can thus closely approximate the spectral shapes of the target frames. However, in the context of audio mosaicing, this has several drawbacks. Since H is multiplied with the complex spectrogram X_{src}, phase cancellation artifacts may arise when superimposing many complex spectral frames. This way, especially low pitched sounds tend to cancel each other out and are not audible in the final audio mosaic. Furthermore, since a sound's timbre is also closely related to the energy distribution in its frequency spectrum, adapting the spectral shapes may change the timbre of the source. An update rule which sets a limit on the maximal number of simultaneous activations is presented in Section 3.2.

A third problem connected with the activation matrix shown in Figure 2c is the loss of temporal characteristics of the source. The typical "buzzing sound" of the bees, which results from pitch modulations (see Figure 2b), is lost in the mosaic (see Figure 2d). This is the case since the spectral frames of the source are activated independently of their order in the source spectrogram. To preserve some temporal characteristics, the update rule presented in Section 3.3 supports the development of diagonal structures in the activation matrix.

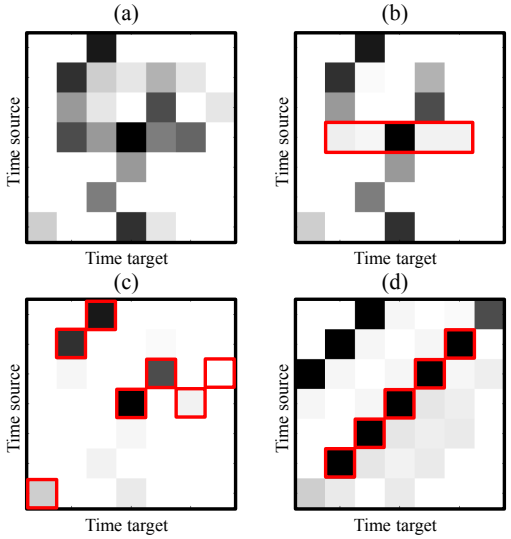

Figure 3. **(a):** Activation matrix $H^{(\ell)}$. **(b):** Repetition restricted activation matrix $R^{(\ell)}$. The horizontal neighborhood is indicated in red. **(c):** Polyphony restricted activation matrix $P^{(\ell)}$. For each column, the highest value is indicated in red. **(d):** Continuity enhancing activation matrix $C^{(\ell)}$. The diagonal kernel is indicated in red.

3. LEARNING SPARSE DIAGONAL ACTIVATIONS

The core idea to overcome the issues of the basic NMF-inspired audio mosaicing procedure is to impose specific constraints on the learned activation matrices by adapting the iterative update process. As discussed in the previous section, we identified three main problems of the mosaics generated by the basic procedure, all related to properties of the the derived activation matrices. First, horizontal activation patterns cause stuttering artifacts in the mosaics. Second, too many simultaneous activations lead to phase cancellations and overfitting of the spectral shapes. Third, the source's temporal characteristics are destroyed by activating source frames independently of each other. We therefore introduce additional update rules to approach these issues, see also Figure 3.

3.1 Avoiding repeated activations

To avoid activating the same spectral frame of the source in subsequent time-instances, the idea is to only keep the highest activations in a horizontal neighborhood of the matrix H, suppressing the remaining values. However, we do not want to interfere too much with the actual learning process in the first few update iterations. The amount of suppression applied to the smaller values is therefore dependent on the iteration index ℓ. Given the activation matrix $H^{(\ell)}$, the size of a horizontal neighborhood r, and the number of iterations L, we compute a *repetition restricted* activation matrix $R^{(\ell)}$ by

$$R^{(\ell)}_{km} = \begin{cases} H^{(\ell)}_{km} & \text{if } H^{(\ell)}_{km} = \mu^{r,(\ell)}_{km} \\ H^{(\ell)}_{km}(1 - \frac{(\ell+1)}{L}) & \text{otherwise} \end{cases}, \quad (3)$$

with $\ell \in [1 : L-1]$ and $\mu_{km}^{r,(\ell)}$ being the maximum value of $H^{(\ell)}$ in a horizontal neighborhood

$$\mu_{km}^{r,(\ell)} = \max(H_{k(m-r)}^{(\ell)}, \ldots, H_{k(m+r)}^{(\ell)}) \,. \quad (4)$$

Note that the suppression of smaller values becomes strict in the last update iteration for $\ell = L-1$. Intuitively, the parameter r defines the minimal horizontal distance (and therefore the minimal time interval) between two activations of the same source frame. Figure 3b shows the repetition restricted activation matrix $R^{(\ell)}$ derived from the toy example activation matrix shown in Figure 3a, using $r = 2$, $\ell = 8$, and $L = 10$. As opposed to $H^{(\ell)}$, there are no two dominant values next to each other in $R^{(\ell)}$.

3.2 Restricting the number of simultaneous activations

Next, we address the problem of too many simultaneous activations. Setting a limit $p \in \mathbb{N}$ on the number of activations in one column of the activation matrix, we compute a *polyphony restricted* activation matrix $P^{(\ell)}$ in a similar manner as $R^{(\ell)}$ by

$$P_{km}^{(\ell)} = \begin{cases} R_{km}^{(\ell)} & \text{if } k \in \Omega_m^{p,(\ell)} \\ R_{km}^{(\ell)}(1 - \frac{(\ell+1)}{L}) & \text{otherwise} \end{cases} , \quad (5)$$

where $\Omega_m^{p,(\ell)}$ contains the indices of the p highest values in the m^{th} column of $R^{(\ell)}$. The parameter p can be directly interpreted as the desired degree of polyphony in the mosaic. For example, setting $p = 1$ results in a mosaic where the source sounds are not heavily superimposed but mainly concatenated to mimic the most dominant features of the target. In Figure 3c, we see the polyphony restricted activation matrix $P^{(\ell)}$ derived from $R^{(\ell)}$, using $p = 1$. One can see that in $P^{(\ell)}$ there is (at most) one single dominant value left in every column.

3.3 Supporting time-continuous activations

To support the development of diagonal structures that activate successive frames of the source, we now compute a *continuity enhancing* activation matrix $C^{(\ell)}$. The idea here is to convolve the matrix P with a diagonal kernel. Choosing $c \in \mathbb{N}$, which defines the length of the kernel, we compute

$$C_{km}^{(\ell)} = \sum_{i=-c}^{c} P_{(k+i)(m+i)}^{(\ell)} \,. \quad (6)$$

Intuitively, the length $2c+1$ of the kernel defines the minimal number of source frames that we would like to successively activate. Figure 3d shows the matrix $C^{(\ell)}$ for our toy example, computed with $c = 2$. Note that in $C^{(\ell)}$ the number of simultaneous dominant activations may locally exceed the limit which was imposed in the computation of the polyphony restricted activation matrix $P^{(\ell)}$. In practice, this is however not a problem and even desirable since this way, the diagonal structures can overlap with each other to some degree. Therefore, the corresponding audio segments of the source are overlapped in the final mosaic as well, leading to smooth transitions between them.

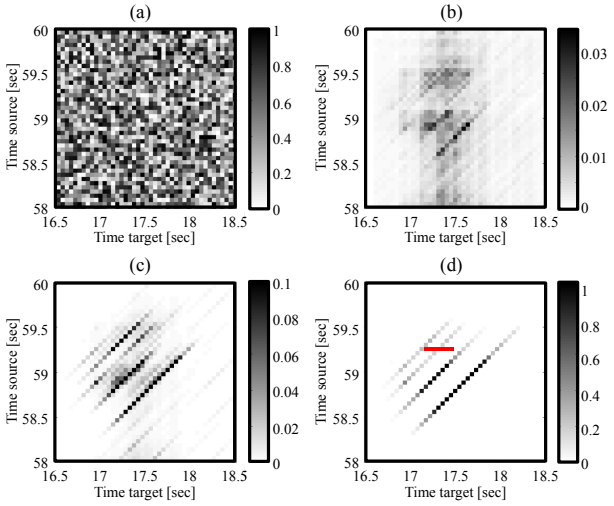

Figure 4. The activation matrix H for the mosaic of "Let it bee" with a recording of bees in different states. **(a):** $H^{(1)}$. **(b):** $H^{(3)}$. **(c):** $H^{(6)}$. **(d):** $H^{(10)}$. The repetition restricting neighborhood is indicated in red.

3.4 Adapting the activations to fit the target

Finally, we perform the standard NMF update step to let the mosaic adapt to the target again. Similarly to Equation (2), we compute the activation matrix for the next iteration by

$$H_{km}^{(\ell+1)} = C_{km}^{(\ell)} \frac{\sum_n W_{nk} V_{nm} / (W C^{(\ell)})_{nm}}{\sum_n W_{nk}} \,. \quad (7)$$

In summary, a single update step of the activation matrix H is computed by applying Equations (3), (5), (6), and (7) sequentially.

Note that in one update iteration, the three intermediate update rules (3), (5), and (6) are insensitive to the target and therefore may increase the distance measure of Equation (1). However, as already discussed in Section 1, we are not interested in minimizing this measure, but trade some approximation accuracy for a better preservation of the source's timbre. In practice, our procedure usually yields an activation matrix that, when multiplied with the source spectrogram, approximates the target spectrogram to a sufficient degree, while preserving the source's timbre in the mosaic much better than the basic procedure described in Section 2.

Figure 4 shows an excerpt of the activation matrix H of our running example "Let it be" for several iteration indices ℓ. Here, we set the repetition restriction parameter to $r = 3$, the limit of simultaneous activations to $p = 10$, the kernel parameter to $c = 3$ (resulting in a diagonal kernel of length 7), and the number of update iterations to $L = 10$. Figure 4a shows the random initialization of the activation matrix $H^{(1)}$. After two iterations, one can already notice diagonal patterns in $H^{(3)}$, see Figure 4b. Figure 4c shows the activations after another three update iterations. The diagonal patterns in $H^{(6)}$ are even more prominent and one can observe that separate diagonal structures start to

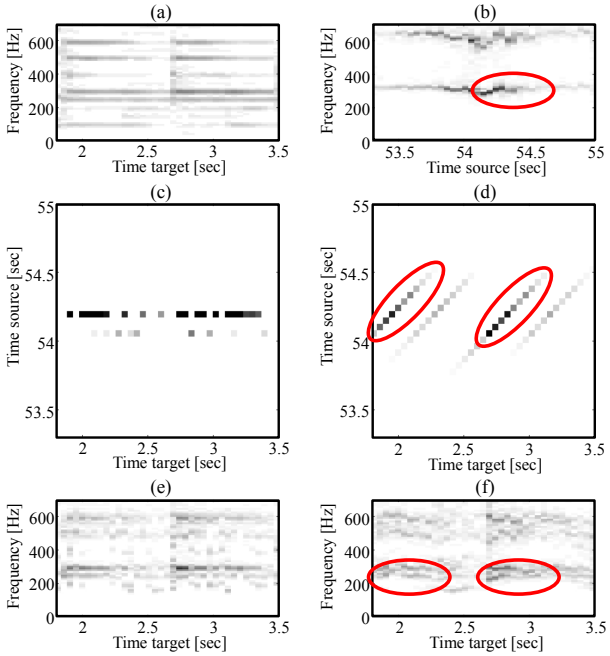

Figure 5. The effect of diagonal activation patterns. **(a):** Spectrogram of the target recording "Let it be". **(b):** Spectrogram of the source recording of buzzing bees. **(c):** Activation matrix H derived with the basic approach. **(d):** Activation matrix H derived with the extended set of update rules. **(e):** Spectrogram of the audio mosaic resulting from the basic approach. **(f):** Spectrogram of the audio mosaic resulting from the extended procedure.

emerge, leaving regions of lower values inbetween them. In Figure 4d, the activation matrix $H^{(10)}$ is shown. In this final activation matrix, four clear diagonal structures have emerged. The remaining activations are outside the visible range. Looking at the two upper diagonals, one can see that although they seem to be rather close together, they obey the repetition restricting horizontal neighborhood indicated in red. Furthermore, it is noteworthy that the length of the diagonals greatly exceeds the length of the diagonal kernel. For example, while we used a diagonal kernel of length 7, the lowest diagonal has a length of 25 non-zero activations, corresponding to an audio segment in the source of roughly one second. This means that the procedure uses a whole one-second patch of source audio material to recreate the target between second 17 and 18.

4. EXPERIMENTS AND EXAMPLES

In this section, we both visually and acoustically demonstrate the effectiveness of our proposed method. As discussed in previous sections, the main drawbacks of the basic audio mosaicing approach described in Section 2 were both the loss of temporal characteristics and spectral shapes of the source sounds in the resulting audio mosaics. The idea was to approach these problems by supporting the development of sparse diagonal structures in the activation matrix with an extended set of update rules. In the following, we exemplify how these structures can preserve the source's desired characteristics in the audio mosaic.

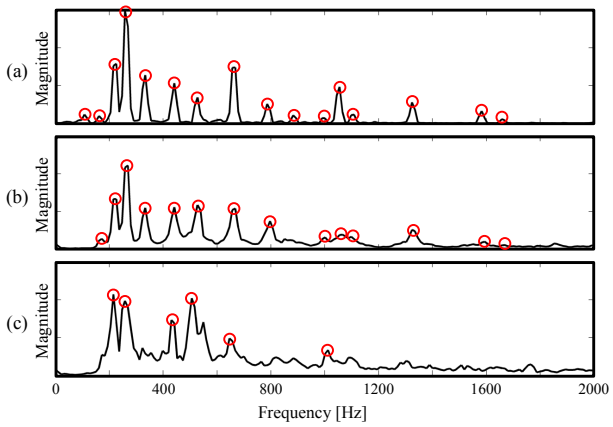

Figure 6. Comparison of spectral shapes. **(a):** A single spectral frame of the target recording ("Let it be"). Harmonics are indicated by red circles. **(b):** The spectral frame of the mosaic computed with the basic procedure at the same temporal position. Harmonics which are present in both the original frame as well as in the mosaic are indicated by red circles. **(c):** The spectral frame of the mosaic computed by using the extended set of update rules.

4.1 Preserving temporal characteristics of the source

In Figure 5, we once again revert to our running example. Here, spectrogram excerpts of the target recording "Let it be" as well as the source recording of buzzing bees are shown in Figures 5a and 5b, respectively. The spectrogram of the target recording exhibits sounds with very stable pitches, resulting from the solo piano at the beginning of the song. In contrast, the buzzing of the bees leads to rather strong amplitude modulations that are characteristic for the sound. Figure 5c shows an excerpt of the activation matrix H as derived by the basic NMF-inspired audio mosaicing procedure. In this excerpt of H, only two different spectral frames of the source are activated repeatedly by the procedure to mimic the stable pitch of the piano sound. The resulting spectrogram of the audio mosaic, shown in Figure 5e, approximates the target's spectrogram quite precisely. However, the characteristic pitch modulations of the buzzing bee sound are lost almost completely. Looking at Figure 5d, one can see the activation matrix H derived by our proposed procedure based on the extended set of update rules. The diagonal patterns shown activate segments of the source that have a duration of roughly half a second. As can be seen by comparing the regions marked in red in the source (Figure 5b) and the mosaic spectrogram (Figure 5f), the temporal structures of these segments are preserved in the mosaic. While the mosaic computed with the extended set of update rules exhibits a lot of pitch modulations, which reflect the preserved timbre of the buzzing bee sound, the tonal content as well as rhythmic structures of the target are still maintained. For example, the two strong partials of the target recording at around 270 Hz and

Name of the target	Description of the target	Name of the source	Description of the source
LetItBe	An excerpt of the song "Let it be" by the Beatles (piano & singing).	Bees	Recording of a buzzing swarm of bees.
GuteNacht	An excerpt of "Gute Nacht" by Franz Schubert which is part of the romantic *Winterreise* song cycle, taken from [15].	Wind	Recording of howling wind.
FunkJazz	An excerpt from a jazz piece performed by the band "Music Delta" (saxophone, synthesizer, bass, and drums), taken from [2].	Whales	Recording of whale songs and whale sounds.
Stepdad	Excerpt from the song "My leather, my fur, my nails" by the pop band Stepdad (synthesizers, drums, and singing).	Chainsaw	Recording of a chainsaw's sawing and engine sounds.
Freischütz	Excerpt from the opera "Der Freischütz" by Carl Maria von Weber (full orchestra, applause at the end).	AirRaid	Recording of an air raid siren.
Vermont	An excerpt of the song "Vermont" by the band "The Districts" (singing, guitar, bass, and drums), taken from [2].	RaceCars	Recording of engine sounds of starting race cars.

Table 1. List of target and source recordings used in our experiments.

300 Hz in Figure 5a are also visible in the audio mosaic in Figure 5f, only this time pitch modulated. Similarly, the onset in the target at second 2.6 is present in the mosaic as well.

4.2 Preserving spectral shapes of the source

In Figure 6, we investigate typical spectral shapes of the target as well as the mosaic for our running example. Figure 6a shows the spectral frame of the target's spectrogram at second 4.6 as a frequency-magnitude plot. One can see the harmonic structure with several clear partials in this frame, resulting from the piano sound in the target. The corresponding spectral frame of the mosaic computed by the basic procedure shown in Figure 6b shows a very similar spectral structure. Most of the harmonics visible in the target are also present in this frame (indicated by the red circles) and even the relations between peak heights are often preserved. In contrast, the spectral frame of the mosaic computed with the extended set of update rules only roughly corresponds to the spectral shape of the target frame, see Figure 6c. However, some of the dominant peaks in the target frame are still present in the mosaic, leading to a sound that captures only the dominant tonal characteristics of the target. The noisy timbre of the buzzing bees, visible by the increased noise level in the frame, is therefore preserved.

4.3 Audio examples

In order to also give an auditory demonstration of our method, we set up an accompanying website for this paper at [7]. On this website, one finds the target recordings as well as source recordings listed in Table 1. To ensure that each source recording offers an adequate pitch range, we computed several pitch-shifted versions of it (using a pitch-shifting algorithm from [6]) and concatenated them. For each pair of target and source, we then generated an audio mosaic using both the basic mosaicing procedure described in Section 2 as well as the procedure based on the extended set of update rules proposed in Section 3. For these experiments, we used music recordings sampled at 22050 Hz, an STFT frame length of 2048 samples and a hop size of 1024 samples to compute the spectrograms. In order to derive the activation matrices for both procedures, we performed $L = 20$ iterations of the respective update steps. For the extended set of update rules, we set the repetition restriction parameter to $r = 3$, the limit of simultaneous activations to $p = 10$, and the kernel parameter to $c = 3$. To reconstruct time-domain signals from the derived complex valued mosaic spectrograms, we finally performed 20 iterations of the STFT inversion procedure proposed in [9].

5. CONCLUSION AND FUTURE WORK

In this work we presented a novel approach for automatically generating an audio mosaic of a target recording using the sounds from a source recording. The core idea of this NMF-inspired procedure was to learn an activation matrix that, when multiplied with the spectrogram of the source recording, yields the spectrogram of the mosaic recording. As our main technical contribution, we proposed an extended set of update rules that supports the development of sparse diagonal structures in the activation matrix during the learning process. Our experiments showed that these diagonal activation structures correspond to the activation of whole sequences of spectral frames and help to preserve timbral characteristics of the source in the mosaic.

In future work we want to investigate if our proposed procedure can also be applied in scenarios beyond audio mosaicing. One possibility is to examine whether supporting the development of diagonal structures in the activation matrix can also be beneficial when learning not only the activation matrix, but also the template matrix. Such an NMF procedure could be applied for learning and identifying repeating patterns in feature sequences, similar to [24] who used techniques based on NMFD for this task. In this context, we hope that our approach may yield a simpler implementation as well as more flexibility since the maximal length of sequences does not need to be fixed.

Acknowledgments:
This work has been supported by the German Research Foundation (DFG MU 2686/6-1). The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institut für Integrierte Schaltungen. Furthermore, we would like to thank Colin Raffel and the other organizers of the *HAMR* Hack Day at ISMIR 2014, where the core ideas of the presented work were born.

6. REFERENCES

[1] G. Bernardes. *Composing Music by Selection: Content-Based Algorithmic-Assisted Audio Composition*. PhD thesis, Faculty of Engineering, University of Porto, 2014.

[2] R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello. Medleydb: A multitrack dataset for annotation-intensive MIR research. In *Proc. of the 15th International Society for Music Information Retrieval Conference ISMIR*, pages 155–160, Taipei, Taiwan, October 2014.

[3] G. Coleman, E. Maestre, and J. Bonada. Augmenting sound mosaicing with descriptor-driven transformation. In *Proc. of the International Conference on Digital Audio Effects (DAFx)*, Graz, Austria, 2010.

[4] J. M. Comajuncosas, A. Barrachina, J. O'Connell, and E. Guaus. Nuvolet: 3D gesture-driven collaborative audio mosaicing. In *Proc. of the International Conference on New Interfaces for Musical Expression*, pages 252–255, Oslo, Norway, 2011.

[5] E. Costello, V. Lazzarini, and J. Timoney. A streaming audio mosaicing vocoder implementation. In *Proc. of the 16th International Conference on Digital Audio Effects (DAFx)*, Maynooth, Ireland, September 2013.

[6] J. Driedger and M. Müller. TSM Toolbox: MATLAB implementations of time-scale modification algorithms. In *Proc. of the International Conference on Digital Audio Effects (DAFx)*, pages 249–256, Erlangen, Germany, 2014.

[7] J. Driedger, T. Prätzlich, and M. Müller. Accompanying website: Let it bee – towards NMF-inspired audio mosaicing. `http://www.audiolabs-erlangen.de/resources/MIR/2015-ISMIR-LetItBee/`.

[8] J. Eggert and E. Körner. Sparse coding and NMF. In *Proc. of the IEEE International Joint Conference on Neural Networks*, volume 4, pages 2529–2533, July 2004.

[9] D. W. Griffin and J. S. Lim. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 32(2):236–243, 1984.

[10] P. O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004.

[11] J. Janer and M. de Boer. Extending voice-driven synthesis to audio mosaicing. In *5th Sound and Music Computing Conference*, Berlin, Germany, July 2008.

[12] J. Kim and H. Park. Toward faster nonnegative matrix factorization: A new algorithm and comparisons. In *Proc. of the IEEE International Conference on Data Mining (ICDM)*, pages 353–362, Pisa, Italy, 2008.

[13] R. Kobayashi. Sound clustering synthesis using spectral data. In *Proc. of the International Computer Music Conference (ICMC)*, Singapore, 2003.

[14] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Proc. of the Neural Information Processing Systems (NIPS)*, pages 556–562, Denver, USA, 2000.

[15] M. Müller, V. Konz, W. Bogler, and V. Arifi-Müller. Saarland music data (SMD). In *Proc. of the International Society for Music Information Retrieval Conference (ISMIR): Late Breaking session*, 2011.

[16] J. Oswald. Plexure. CD, 1993. `http://www.allmusic.com/album/plexure-mw0000621108`.

[17] N. Schnell, M. A. S. Cifuentes, and J.-P. Lambert. First steps in relaxed real-time typo-morphological audio analysis/synthesis. In *Sound and Music Computing*, Barcelona, Spain, 2010.

[18] D. Schwarz. A system for data-driven concatenative sound synthesis. In *Proc. of the International Conference on Digital Audio Effects (DAFx)*, Verona, Italy, July 2000.

[19] D. Schwarz. Concatenative sound synthesis: The early years. *Journal of New Music Reaserch*, 35(1), March 2006.

[20] P. Smaragdis. Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In *Independent Component Analysis and Blind Signal Separation*, volume 3195 of *Lecture Notes in Computer Science*, pages 494–499. Springer Berlin Heidelberg, 2004.

[21] P. Smaragdis, B. Raj, and M. Shashanka. Sparse and shift-invariant feature extraction from non-negative data. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP*, pages 2069–2072, Las Vegas, Nevada, USA, 2008.

[22] P. A. Tremblay and D. Schwarz. Surfing the waves : Live audio mosaicing of an electric bass performance as a corpus browsing interface. In *Proc. of the International Conference on New Interfaces for Musical Expression*, pages 447–450, Sydney, Australia, September 2010.

[23] T. Virtanen. Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech and Language Processing*, 15(3):1066–1074, 2007.

[24] R. J. Weiss and J. P. Bello. Unsupervised discovery of temporal structure in music. *IEEE Journal of Selected Topics in Signal Processing*, 5:1240–1251, 2011.