ACCOUNTING FOR PHASE CANCELLATIONS IN NON-NEGATIVE MATRIX FACTORIZATION USING WEIGHTED DISTANCES

Sebastian Ewert *Mark D. Plumbley* *Mark Sandler*

Queen Mary University of London, London, United Kingdom

ABSTRACT

Techniques based on non-negative matrix factorization (NMF) have been successfully used to decompose a spectrogram of a music recording into a dictionary of templates and activations. While advanced NMF variants often yield robust signal models, there are usually some inaccuracies in the factorization since the underlying methods are not prepared for phase cancellations that occur when sounds with similar frequency are mixed. In this paper, we present a novel method that takes phase cancellations into account to refine dictionaries learned by NMF-based methods. Our approach exploits the fact that advanced NMF methods are often robust enough to provide information about how sound sources interact in a spectrogram, where they overlap, and thus where phase cancellations could occur. Using this information, the distances used in NMF are weighted entry-wise to attenuate the influence of regions with phase cancellations. Experiments on full-length, polyphonic piano recordings indicate that our method can be successfully used to refine NMF-based dictionaries.

Index Terms— Weighted Distances, NMF, Phase Cancellation, Source Separation.

1. INTRODUCTION

Non-negative matrix factorization (NMF) and its variants have been widely adopted in music signal processing, with applications in music transcription, source separation and remixing, or pre-processing in music information retrieval [1–4]. A central idea in NMF is to learn a dictionary of spectral templates, which can be used as building blocks to approximate a time-frequency representation of a given signal [1]. In a musical context, each template typically reflects the spectral envelope associated with a single musical pitch played on an specific instrument. By reconstructing a signal based on these learned dictionary elements, NMF is often used to explain the structure of a signal and to disclose its constituent parts.

To increase the robustness of the dictionary learning process, one makes simplifying assumptions in NMF. For example, an implicit assumption in NMF is that the time-frequency representation of a mix of sound sources is equal to the sum of the individual time-frequency representations of the sources [5]. While this is true for complex spectrograms, it is only approximately correct for magnitude or power spectrograms, which are commonly employed in NMF. This way, the NMF model does not account for so called phase cancellations: the destructive interference between two sounds with similar frequency during the mixing process. In particular, at positions in a magnitude spectrogram where phase cancellations occur, less energy is available than expected without a cancellation. As the NMF model does not account for such effects, cancellations can have a negative influence on the NMF learning process.

Various extensions to NMF have been proposed which increase the robustness of the dictionary learning process in general, and which often also mitigate negative effects resulting from phase cancellations. For example, enforcing a specific structure in the templates imposes constraints over how sounds can be represented in NMF [4], which typically stabilizes the learning process. Furthermore, sparsity constraints [6, 7] can be used to penalize the use of a high number of templates to represent sounds, which typically leads to more meaningful templates and thus also stabilizes the learning process. However, whether such extensions have a significant effect on the dictionary learning process often depends on the particular recording being processed.

The effect of phase cancellations on the learning process has also been directly addressed by some NMF extensions. For example, one can show that if a given spectrogram can be assumed to be the result of a random process, and if this random process has certain properties (e.g. entries in the phase field are i.i.d. random variables), then using a modified dictionary learning process (Itakura-Saito (IS)-NMF) on a power spectrogram leads to a dictionary which is invariant against phase cancellations in the signal [5, 8]. While this approach provides the means to interpret NMF as a probabilistic process, some of these assumptions are often not met using real-world recordings [9]. Furthermore, a variant referred to as complex-NMF was introduced, which operates on complex-valued spectrograms [10]. While the relaxation of the strict non-negativity constraints used in standard NMF leads to more freedom in the signal model, it can also lower the robustness of the learning process in some cases. High-resolution (HR)-NMF was recently proposed as an extension to IS-NMF, which temporally couples the random variables in IS-NMF using an auto-regressive model [11]. This way, for each frequency band, energy patterns typical for a source can be captured (such as beating in piano sounds) and can be used to model overlapping frequency components with high accuracy, without explicitly modeling the phase field.

In this paper, we present a novel method that refines learned dictionaries by taking possible phase cancellations into account during a re-training step. Instead of modeling the phase field and phase cancellations directly, we exploit that NMF (or one of its variants) is often robust enough to obtain an initial model of the signal that can be used to indicate positions in the spectrogram where phase cancellations might have occurred. More precisely, after using an initial NMF, we look for positions in a spectrogram, where energy is explained by more than one template. Such positions often indicate where sources might overlap, and hence where phase cancellation could have occurred. To refine the dictionary, we then attenuate the influence of these positions in the spectrogram on the learning process. To this end, we create a matrix that contains a weight for each entry in the spectrogram and incorporate this matrix into typical spectral distance measures used in NMF, which enables us to control how much influence on the distance each entry should have. Finally, the NMF dictionary is refined during a retraining step using new update rules that take the modified distance measures into account. Since this general strategy does not rely on a specific NMF

978-1-4799-2893-4/14/$31.00 ©2014 Crown

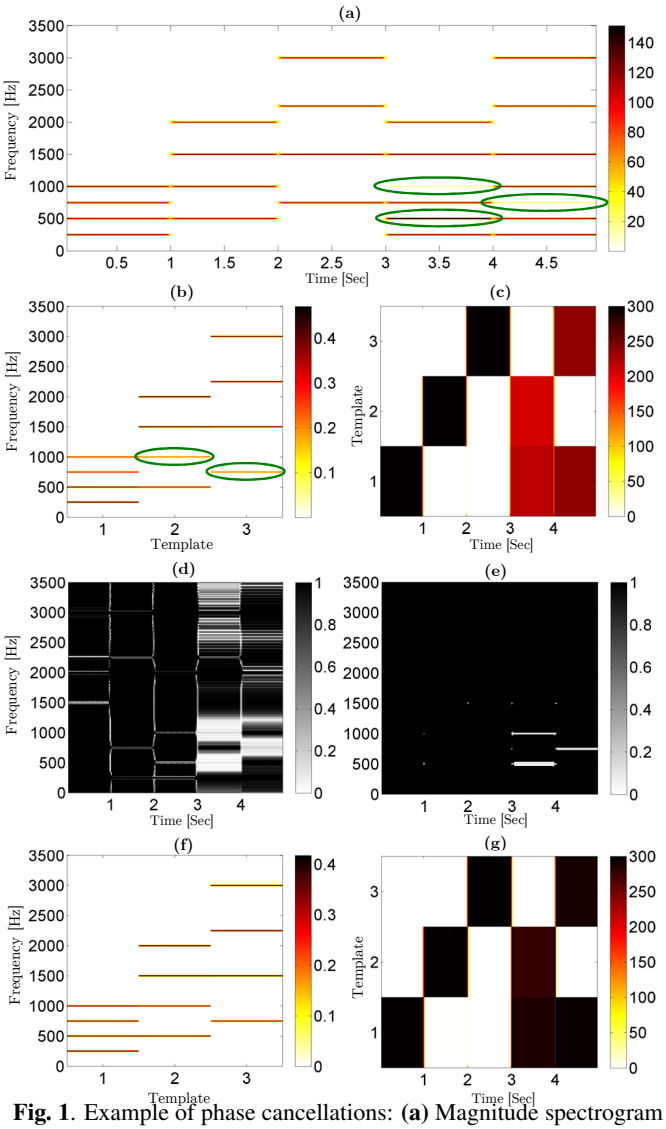

Fig. 1. Example of phase cancellations: **(a)** Magnitude spectrogram of the recording. **(b)/(c)** Dictionary and activation matrix learned via classic NMF. **(d)/(e)** Weighting matrices $\widetilde{\mathcal{W}}$ and $\mathcal{W}$. **(f)/(g)** Refined dictionary and activation matrix learned via proposed method.

variant, it can be combined with various NMF extensions to increase the robustness of the learning process.

In the following, we briefly introduce NMF discussing examples of shortcomings in the dictionary learning process related to phase cancellations. Then, we describe our method based on weighted spectral distances to refine NMF-based dictionaries and illustrate its effects on the learning process. After presenting the results of an experiment using full-length piano recordings, we conclude with a prospect on future extensions.

2. NON-NEGATIVE MATRIX FACTORIZATION

In classic non-negative matrix factorization, one approximates a spectral representation of a given recording by a product of two non-negative matrices. More exactly, given a magnitude spectrogram $V \in \mathbb{R}_+^{M\times N}$ of a music recording, NMF seeks to find non-negative matrices $T \in \mathbb{R}_+^{M\times K}$ and $A \in \mathbb{R}_+^{K\times N}$ such that $V \approx T \cdot A$. In this context, the matrix T is referred to as *dictionary*, its columns are referred to as *templates* and the rows of A as the corresponding *activations*. To compute such a factorization, one seeks to minimize a distance measure $D(V, T \cdot A)$ with respect to T and A. In the following, we use the modified *Kullback-Leibler (KL) divergence*, which is for $V, \widetilde{V} \in \mathbb{R}_+^{M\times N}$ defined as:

$$D(V,\widetilde{V}) := \sum_{m,n}\Big(V_{m,n}\ln\frac{V_{m,n}}{\widetilde{V}_{m,n}} - V_{m,n} + \widetilde{V}_{m,n}\Big).$$

To find a local minimum of D with respect to T and A, Lee and Seung proposed multiplicative update rules [12]:

$$T \leftarrow T \odot \frac{(\frac{V}{T\cdot A})\cdot A^\top}{J\cdot A^\top} \quad \text{and} \quad A \leftarrow A \odot \frac{T^\top\cdot(\frac{V}{T\cdot A})}{T^\top\cdot J},$$

where the $\cdot$ operator denotes the usual matrix product, the $\odot$ operator denotes the Hadamard product (point-wise multiplication), $J \in \mathbb{R}^{M\times N}$ denotes the matrix of ones, and the division is understood point-wise. After initializing T and A with non-negative random values, these rules monotonically decrease $D(V, T \cdot A)$.

To study the effect of phase cancellations on the NMF learning process, we consider in the following a simple synthetic example shown in Fig. 1(a). For this recording, we synthesized three harmonic sounds, having a fundamental frequency of 250 Hz, 500 Hz, and 750 Hz, respectively. Each one second long sound is played once at the beginning of the recording, compare the first three seconds in Fig. 1(a). Between seconds 3 and 4 we mixed the 250 Hz and 500 Hz sounds, and between seconds 4 and 5 we mixed the 250 Hz and 750 Hz sounds. Since the fundamental frequencies are integer multiples of 250, some of the partials overlap in the mixing section, see green circles in Fig. 1(a). Here, we can observe constructive interference (green circle around 500 Hz between seconds 3 and 4), as well as destructive interference (the other two green circles).

We now investigate how NMF behaves on this recording. To this end, we set the number of templates K=3 and apply NMF as described above with 100 update iterations[1]. The result is illustrated in Fig. 1(b) and Fig. 1(c), which show the learned dictionary T and the corresponding activation A, respectively. In this example, NMF identifies the three harmonic sounds approximately correctly, see Fig. 1(b). However, there are some inaccuracies. Looking at the three original sounds shown in Fig. 1(a), we see that all four partials have the same intensity in each sound. In Fig. 1(b), however, the second partial in the second template, and the first partial in the third template are weaker compared to the others. Furthermore, the loudness of individual sounds was not changed during the mixing process, but still the lower activity values in Fig. 1(c) (between seconds 3 and 5) incorrectly indicate that the volume of the individual sounds in the mixing section is much lower compared to the original sounds. In both cases, this behavior is caused by the unexpected local energy loss resulting from the two phase cancellations.

3. WEIGHTED-DISTANCE NMF

Modeling the phase field and the cancellations resulting from it can be very complex, and would introduce additional degrees of freedom in the signal model which could lead to numerical instabilities in the learning process. Instead, we exploit the fact that NMF (and in particular its more advanced variants) often yield quite robust signal models. In particular, if the initial signal model indeed explains the

[1] In total, we used 20 different random initializations for NMF and kept the factorization result having the lowest KL-divergence.

inner structure of a recording approximately correctly, we can use it to estimate where in a spectrogram sounds might be overlapping and hence where phase cancellations are likely to occur. Then we can pay less attention to these entries when we learn (or re-learn) our dictionary.

To this end, we now define a weighting matrix $\mathcal{W} \in [0,1]^{M \times N}$, which captures how much attention each spectrogram entry should receive in the distance measure D. We start with the definition of a helper matrix $\widetilde{\mathcal{W}} \in [0,1]^{M \times N}$:

$$\widetilde{\mathcal{W}}_{m,n} := \max_{k \in [1:K]} \max \left[\frac{(T^k \cdot A^k)_{m,n}}{(T \cdot A)_{m,n}} \cdot 2 - 1, \epsilon \right]$$

where T^k denotes the k-th column of T and A^k the k-th row of A, and $\epsilon \in \mathbb{R}_+$ is a small positive number. $\widetilde{\mathcal{W}}_{m,n}$ is equal to 1 if the entire energy in the model at the (m,n)-th position $(T \cdot A)_{m,n}$ is explained by a single template $(T^k \cdot A^k)_{m,n}$, and close to zero if the energy is equally explained by two or more templates, see Fig. 1(d) for an example. In particular, $\widetilde{\mathcal{W}}_{m,n} \approx 0$ if two or more sources are overlapping according to the model, and $\widetilde{\mathcal{W}}_{m,n} \approx 1$ if no overlap was detected. Next, we integrate $\widetilde{\mathcal{W}}$ into our distance measure D to reduce the influence of spectrogram entries with potential phase cancellations:

$$D_{\widetilde{\mathcal{W}}}(V,W) := \sum_{m,n} \widetilde{\mathcal{W}}_{m,n} \cdot \left(V_{m,n} \ln \frac{V_{m,n}}{W_{m,n}} - V_{m,n} + W_{m,n} \right).$$

However, integrating $\widetilde{\mathcal{W}}$ as defined above would have some drawbacks. Most importantly, ignoring entries in the spectrogram essentially creates wildcards where the model behaviour is not significantly penalized anymore. Using $\widetilde{\mathcal{W}}$, however, often leads to a high number of wildcards (Fig. 1(d)), which results in additional degrees of freedom in the model and lowers the overall robustness of the NMF learning process. To limit the number of entries being ignored we incorporate two additional requirements into $\widetilde{\mathcal{W}}$:

$$\mathcal{W}_{m,n} := \begin{cases} (\widetilde{\mathcal{W}}_{m,n})^C, & (T \cdot A)_{m,n} - V_{m,n} \geq b_1 \\ & \quad \wedge V_{m,n} \geq b_2 \\ 1, & \text{otherwise,} \end{cases}$$

where $b_1, b_2, C \in \mathbb{R}_+$. The first requirements means that the model expects more energy in the spectrogram than actually exists, which is a basic requirement for a phase cancellation. The second requirement means that only entries having a minimum amount of energy can be ignored. Here, the idea is that, in practice, a phase cancellation is not perfect but instead there is almost always some residual energy (b_2) left. If there is almost no energy at all, there probably was not a phase cancellation. Furthermore, a parameter C is introduced to non-linearly drive small entries in $\widetilde{\mathcal{W}}$ more towards zero. Fig. 1(e) shows an example for $\mathcal{W}$, with $b_1 = 0$, b_2 =[-40dB below $\max_{m,n} V_{m,n}$], and $C = 1.5$. As we can see, the matrix captures the three positions, where phase cancellations occur, quite accurately, except for some spurious entries.

To actually employ $\mathcal{W}$ to refine the dictionary learned using classic NMF, we need NMF update rules that minimize the distance $D_{\mathcal{W}}(V, T \cdot A)$. Such update rules have been introduced by Blondel et al. in [13].

$$T \leftarrow T \odot \frac{(\frac{\mathcal{W} \odot V}{T \cdot A}) \cdot A^\top}{\mathcal{W} \cdot A^\top} \quad \text{and} \quad A \leftarrow A \odot \frac{T^\top \cdot (\frac{\mathcal{W} \odot V}{T \cdot A})}{T^\top \cdot \mathcal{W}}.$$

An example is given in Fig. 1(f) and Fig. 1(g), which show the refined dictionary T and activations A for the example shown in Fig. 1(a), respectively. All partials in the refined templates as well as all refined activations correctly have the same intensity.

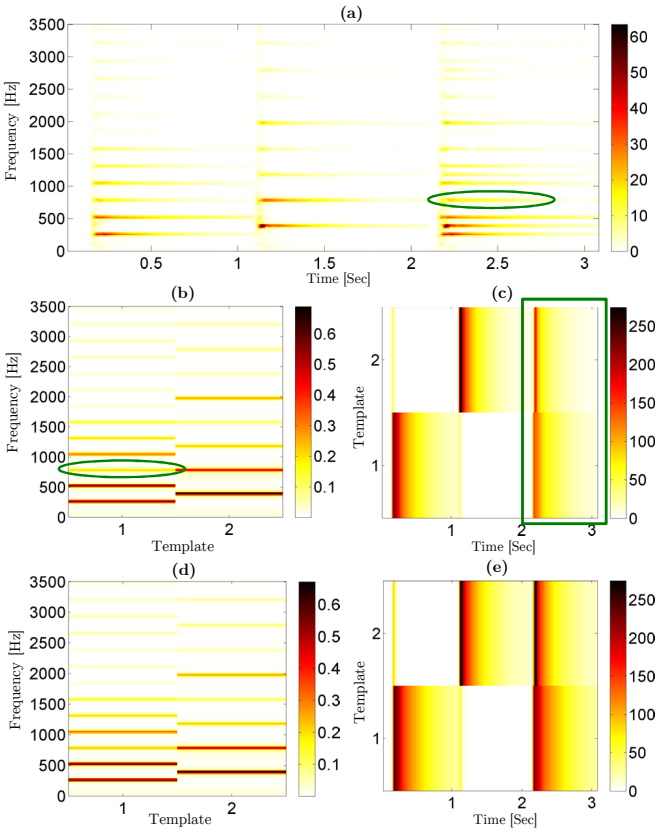

Fig. 2. Piano Example: **(a)** Spectrogram of the recording. **(b)/(c)** Dictionary and activation matrix learned via classic NMF. **(d)/(e)** Refined dictionary and activation matrix learned via proposed method.

In a second example, we used non-synthetic sounds taken from the RWC database [14]. We took isolated note recordings for a C4 and a G4 played on a piano and created a recording, where both notes are first played separately and then mixed together, see Fig. 2(a). The green circle highlights a region in the spectrogram where energy is lost due to a partial phase cancellation. The result of applying classic NMF on this recording is illustrated in Fig. 2(b) and Fig. 2(c), where the learned dictionary and activations are shown. We can observe, that the third partial in the first template has slightly less energy than it should have, see green circle in Fig. 2(b). More obvious though are the effects of the phase cancellation on the activations: While the isolated sounds were mixed together without changing their volume, the activations in the mix segment (green rectangle in Fig. 2(c)) are much lower compared to the activations for the isolated sounds and incorrectly indicate a change in loudness. However, applying our weighted-distance NMF, the isolated sound activations (between seconds 0 and 2 in Fig. 2(e)) are almost perfectly repeated in the segment containing the mixed sounds (between seconds 2 and 3 in Fig. 2(e)). This way, the refined activations reveal that the two sounds were indeed mixed without changing their volume. Also the third partial in the first template contains now slightly more energy. Overall, using our $D_{\mathcal{W}}$ enables NMF to focus on those parts of the spectrogram that are likely to be unaffected by phase cancellations and thus helps to refine learned dictionaries and activations.

With a final example, we illustrate a (positive) side-effect of our proposed method, which could be observed during our experiments.

In particular, our method relies on the heuristic that one sound is represented by a single template to detect locations of potential phase cancellations. While this is indeed enforced in some advanced NMF variants, there is no such constraint in general. In particular, in (classic) NMF one often finds the situation that a sounds that is not overlapped by other sounds (and thus phase cancellation does not occur) is represented by several templates. As an example, we consider in Fig. 3(a) a recording of a C5 and a G5 played on a flute. The C5 is repeated four times and after that the G5 is played once. Applying classic NMF with $K=2$ yields a factorization as shown in Fig. 3(b) and Fig. 3(c). Since most energy in the spectrogram is associated with the C5 note, the G5 note is not captured in the templates but instead both templates are used to (over-)represent the C5. Applying our proposed method, however, this behavior changes. As shown in Fig. 3(d) and Fig. 3(e), now one template is used for each note and the G5 is correctly modeled. In particular, since the C5 is represented by two templates our weighted distance measure attenuates the influence of some parts of the C5 notes on the distance measure such that the G5 energy becomes important enough to be represented by a template. However, this side-effect could not always be observed and behaved rather fragile. In particular, adding another repetition of the C5 to the recording shown in Fig. 3(a), and the dictionary learned using our proposed method is again almost identical to the one learned using classic NMF.

4. EXPERIMENTS

To evaluate the proposed method on more complex recordings, we conducted a simple source separation experiment. To this end, we used ten piano MIDI files obtained from the Mutopia website [15]. For our experiment, we derived two new MIDI files for each piece by partitioning the 88 keys on a piano into two subsets: one file contained only note events with a MIDI pitch in $P_1 := \{21, \ldots, 59\}$ (corresponding to A0 to B3), and the other one contained only note events with a MIDI pitch in $P_2 := \{60, \ldots, 108\}$ (corresponding to C4 to C8). We then synthesized the three MIDI files for each piece using the Timidity wavetable synthesizer. The task consists of splitting the audio recording synthesized using all MIDI note events into two recordings containing only the sounds corresponding to the sets P_1 and P_2, respectively. The two recordings that were synthesized using only the MIDI pitches in P_1 and P_2, respectively, were used as ground truth separation results.

For our experiment, we used the NMF variant described in [16, 17]. In this approach, the dictionary is not randomly initialized. Instead, each template is associated with a single MIDI pitch p. Then, a harmonic structure is enforced for each template by setting those entries to zero that are not in a neighborhood of an integer multiple of $f(p)$, where the function $f(p) = 12^{(p-69)/12} \cdot 440$ assigns a MIDI pitch to its corresponding frequency in Hertz. Using multiplicative update rules guarantees that these constraints remain valid during the subsequent learning process. In other words, zero-valued entries between the expected partials enforce the intended harmonic structure during the NMF learning process.

We used 88 templates, associated them with the MIDI pitches $\{21, \ldots, 108\}$ and initialized them according to [16]. After the NMF learning process, we derived two new activation matrices: for A_{P_1}, we kept only those activations from A that were associated with a MIDI pitch in P_1, and set the remaining rows to zero. The matrix A_{P_2} was derived accordingly. Using these two new activation matrices, we split the original spectrogram into two parts:

$$V_{P_1} := V \odot \frac{T \cdot A_{P_1}}{T \cdot A} \quad \text{and} \quad V_{P_2} := V \odot \frac{T \cdot A_{P_2}}{T \cdot A},$$

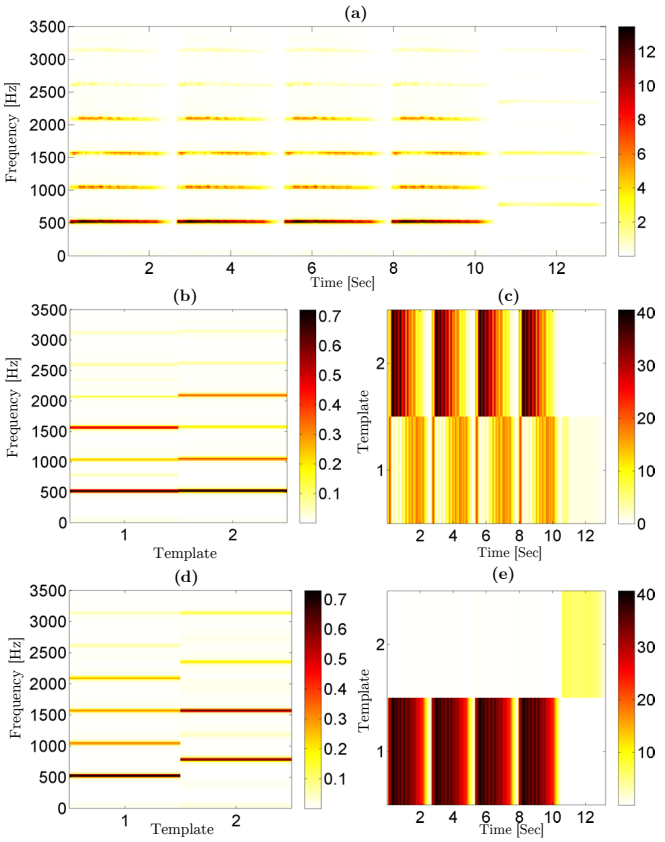

Fig. 3. Flute Example: **(a)** Spectrogram of the recording. **(b)/(c)** Dictionary and activation matrix learned via classic NMF. **(d)/(e)** Refined dictionary and activation matrix learned via proposed method.

where the division is understood point-wise. Finally, we used a spectrogram inversion method to create audio recordings corresponding to V_{P_1} and V_{P_2} [18].

Using the BBS-eval toolkit [19], we computed Signal-to-Distortion Ratio (SDR) values to quantify the quality of the separation results, comparing them to the recordings that we synthesized using only P_1 and P_2. On average, using classic NMF, this simple approach achieved an SDR of +2.8dB, while applying our proposed weighted NMF led to an SDR of +3.1dB. The slight increase in SDR demonstrates that our method is indeed useful also for complex polyphonic recordings. However, the overall separation performance remained rather small because the capabilities of our proposed method are limited by the robustness of the underlying NMF variant, which was not state-of-the-art in this experiment.

5. CONCLUSIONS

We presented a method for refining dictionaries learned via NMF-based methods. Our method exploits that (advanced) NMF methods typically yield robust signal models, which can be used to estimate where in a spectrogram phase cancellations could have occurred. By attenuating the influence of these positions in the spectrogram on the distance measures used in NMF, the method allows for refining dictionaries learned via NMF. In the future, we plan to further extend the idea of using weighted distances to guide the learning process in NMF-based methods.

6. REFERENCES

[1] Andrzej Cichocki, Rafal Zdunek, and Anh Huy Phan, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation*, John Wiley and Sons, 2009.

[2] Derry FitzGerald, Matt Cranitch, and Eugene Coyle, "Extended nonnegative tensor factorisation models for musical sound source separation (article id 872425)," *Computational Intelligence and Neuroscience*, vol. 2008, 2008.

[3] Tuomas Virtanen, Ali Taylan Cemgil, and Simon Godsill, "Bayesian extensions to non-negative matrix factorisation for audio signal modelling," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 1825–1828.

[4] Emmanuel Vincent, Nancy Bertin, and Roland Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 528–537, 2010.

[5] Cédric Févotte, Nancy Bertin, and Jean-Louis Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Computation*, vol. 21, pp. 793–830, 2009.

[6] Samer A. Abdallah and Mark D. Plumbley, "Unsupervised analysis of polyphonic music by sparse coding," *IEEE Transactions on Neural Networks*, vol. 17, no. 1, pp. 179–196, 2006.

[7] Tuomas Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[8] R Mitchell Parry and Irfan Essa, "Incorporating phase information for source separation via spectrogram factorization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2007, pp. 661–664.

[9] Paris Smaragdis, Bhiksha Raj, and Madhusudana Shashanka, "Missing data imputation for time-frequency representations of audio signals," *Journal of signal processing systems*, vol. 65, no. 3, pp. 361–370, 2011.

[10] Hirokazu Kameoka, Nobutaka Ono, Kunio Kashino, and Shigeki Sagayama, "Complex NMF: A new sparse representation for acoustic signals," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009, pp. 3437–3440.

[11] Roland Badeau, "Gaussian modeling of mixtures of non-stationary signals in the time-frequency domain (HR-NMF)," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, pp. 253–256.

[12] Daniel D. Lee and H. Sebastian Seung, "Algorithms for non-negative matrix factorization," in *Proceedings of the Neural Information Processing Systems (NIPS)*, Denver, CO, USA, 2000, pp. 556–562.

[13] Vincent D. Blondel, Ngoc-Diep Ho, and Paul van Dooren, "Weighted nonnegative matrix factorization and face feature extraction," *Image and Vision Computing (submitted)*, 2007.

[14] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka, "RWC music database: Popular, classical and jazz music databases," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Paris, France, 2002, pp. 287–288.

[15] Mutopia Project, "Music free to download, print out, perform and distribute," `http://www.mutopiaproject.org`, Retrieved 12.05.2009.

[16] Stanislaw Andrzej Raczynski, Nobutaka Ono, and Shigeki Sagayama, "Multipitch analysis with harmonic nonnegative matrix approximation," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Vienna, Austria, 2007, pp. 381–386.

[17] Sebastian Ewert and Meinard Müller, "Using score-informed constraints for NMF-based source separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, 2012, pp. 129–132.

[18] Daniel W. Griffin and Jae S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[19] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.