USING SCORE-INFORMED CONSTRAINTS FOR NMF-BASED SOURCE SEPARATION

Sebastian Ewert

University of Bonn
Bonn, Germany

Meinard Müller

Saarland University and MPI Informatik
Saarbrücken, Germany

ABSTRACT

Techniques based on non-negative matrix factorization (NMF) can be used to efficiently decompose a magnitude spectrogram into a set of template (column) vectors and activation (row) vectors. To better control this decomposition, NMF has been extended using prior knowledge and parametric models. In this paper, we present such an extended approach that uses additional score information to guide the decomposition process. Here, opposed to previous methods, our main idea is to impose constraints on both the template as well as the activation side. We show that using such double constraints results in musically meaningful decompositions similar to parametric approaches, while being computationally less demanding and easier to implement. Furthermore, additional onset constraints can be incorporated in a straightforward manner without sacrificing robustness. We evaluate our approach in the context of separating note groups (e. g. the left or right hand) from monaural piano recordings.

Index Terms— Score-informed processing, non-negative matrix factorization, music synchronization, alignment.

1. INTRODUCTION

In recent years, methods for separating musically meaningful sound sources from monaural music recordings have been applied to many music processing tasks. For example, techniques to extract individual instrument tracks have been incorporated into approaches for instrument recognition [1] or instrument-wise equalization [2,3]. Most of these techniques rely on some variant of non-negative matrix factorization (NMF) [4], or an equivalent formulation such as probabilistic latent component analysis (PLCA) [5]. Here, the idea is to decompose the magnitude spectrogram of a given recording into a set of template (column) vectors and activation (row) vectors.

However, as discussed in more detail below, template vectors learnt by NMF-based approaches are often hard to interpret and lack explicit semantics. To obtain musically meaningful vectors, the original NMF can be modified such that each template vector is associated with a single musical pitch. To this end, many approaches specify the template vectors using a parametric model. For example, the template vectors in [1] are described by a source/filter model, in [6] by harmonic combs and in [2, 7, 8] by spectrally and/or temporally localized Gaussians. On the one hand, a parametric model allows for a straightforward integration of musical knowledge. For example, in [8] the authors extend their harmony-based approach with a percussive component exploiting the typical spectral shape around onsets. Furthermore, allowing only solutions that are valid within the model, the parametric approaches offer a high degree of robustness. On the other hand, the parameter estimation and the resulting spectrogram approximation can be inaccurate in the case that some model assumptions are violated. Additionally, the parameter estimation process is often computationally expensive.

In this paper, we present a method that combines the efficiency and flexibility of classic NMF with advantages of parametric approaches. Our method is based on a strategy originally presented in [9]. Here, the underlying idea is to enforce a harmonic structure for the template vectors by setting those entries to zero that are not in a neighborhood of an expected partial. Then, using multiplicative update rules guarantees that these constraints remain valid during the subsequent learning process. We extend this idea in several ways. First, opposed to previous methods, we simultaneously constrain the template vectors as well as the activations. This way, instead of just specifying *what* is expected we additionally specify *when* something is expected. The use of these double constraints becomes possible by exploiting available score information in the form of a MIDI file. Here, we use high-resolution synchronization techniques to align the MIDI file with a given recording. As a second extension, we exploit the robustness gained by the double constraints to integrate template vectors that represent percussive elements such as onsets. As our experiments show, these double constraints in combination with the onset template vectors stabilize the resulting separation results and lead to an increased overall separation quality. Altogether, the proposed method combines the expressive power of parametric approaches with the efficiency of classic NMF, while still being easy to implement.

The remainder of this paper is organized as follows. In Section 2 we introduce our NMF-based method using double constraints. Then, in Section 3, we report on systematic experiments, where we employ the proposed method to separate note groups (e. g. the left or right hand) from monaural piano recordings. Conclusions and prospects on future work are given in Section 4. Further related work is discussed in the respective sections.

2. SCORE-INFORMED CONSTRAINTS FOR NMF

Given the magnitude spectrogram of a music recording $V \in \mathbb{R}_{\geq 0}^{m \times n}$ and $k \in \mathbb{N}$, classical NMF derives two non-negative matrices $W \in \mathbb{R}_{\geq 0}^{m \times k}$ and $H \in \mathbb{R}_{\geq 0}^{k \times n}$ such that a distance $D(V, WH)$, typically a modified Kullback-Leibler divergence, is minimized. In this context, the columns of W are often referred to as *template vectors* and the rows of H as the corresponding *activities*. To compute a factorization one typically initializes W and H with random values and updates them iteratively using multiplicative rules [4]. Such rules offer several advantages over alternative approaches. Firstly, they are easy to implement. Secondly, each entry of W and H set to zero remains zero throughout the update process, which easily allows for imposing hard constraints on the factorization. Thirdly, consisting of only highly-parallelizable matrix operations, multiplicative update rules are computationally very efficient. In contrast, most parametric

This work has been supported by the German Research Foundation (DFG CL 64/6-1) and the Cluster of Excellence on Multimodal Computing and Interaction at Saarland University.

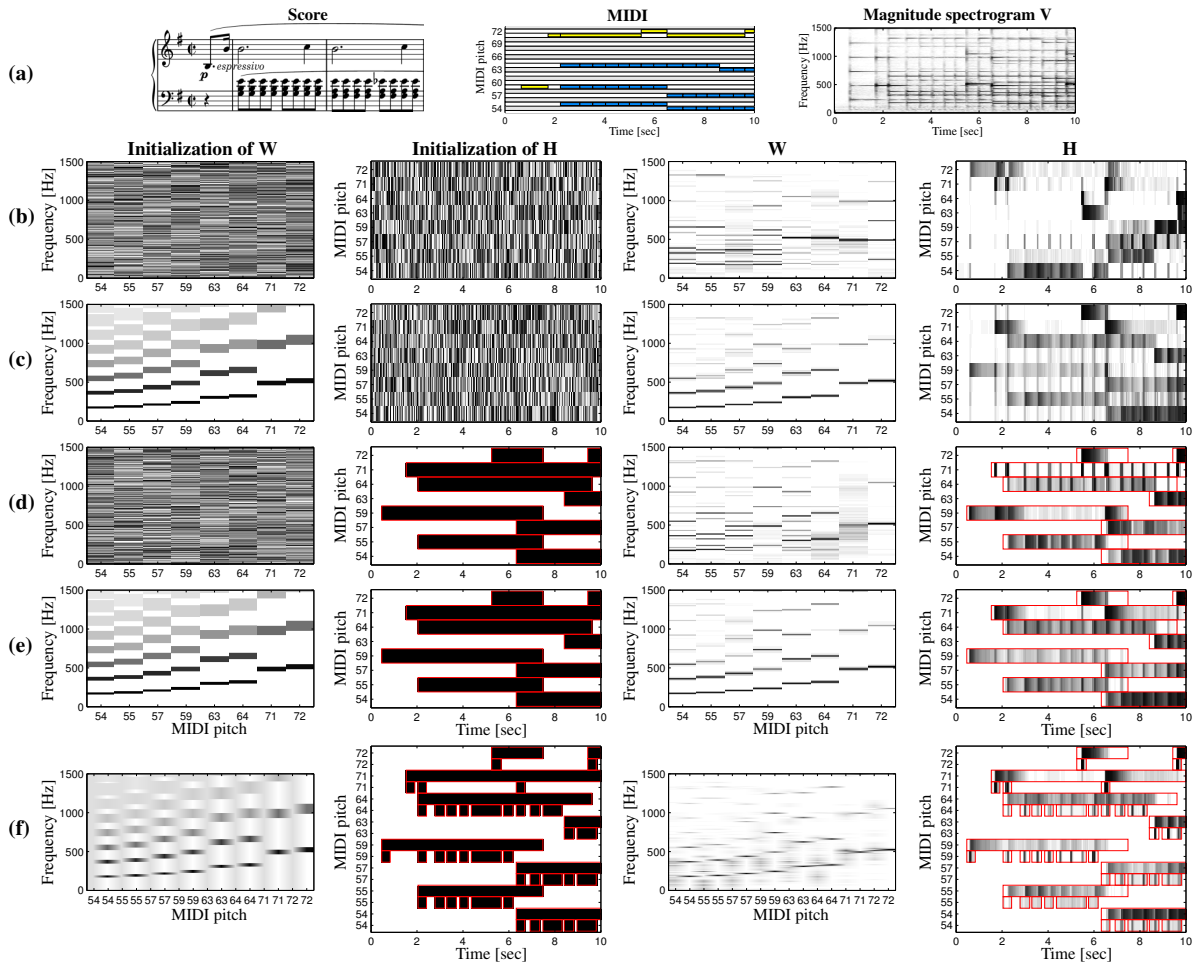

Fig. 1. NMF factorizations resulting from several initialization strategies for a recording of the first measures of Chopin's Op. 28 No. 4. **(a)** Score and MIDI representation of the piece as well as a magnitude spectrogram of the recording. **(b)** Standard random initialization. **(c)** I_{W}**:** Initialization of the template vectors. **(d)** I_{H}**:** Initialization of the activations. **(e)** I_{WH}**:** Combination of I_{W} and I_{H}. **(f)** $I_{\mathrm{WH}}^{\mathrm{O}}$**:** Extended variant of I_{WH} involving onset template vectors.

approaches usually require a computationally much more expensive learning procedure involving a large number of function evaluations. However, a major drawback of NMF is that the factorization is hard to control. As an example, Fig. 1(b) shows the initialization and resulting NMF factorization for a recording of the first measures of Chopin's Op. 28 No. 4. Here, k is set to the number of active pitches in this section of the piece. The example illustrates that the NMF template vectors and activations often have little to no musical meaning even though their product closely approximates V.

2.1. Constraints in NMF

To obtain a semantically more meaningful factorization most approaches enforce a certain structure for the template vectors. Here, one possibility is to describe the vectors via parametric models. While this approach allows for a direct integration of musical knowledge, it is computationally more expensive and susceptible to inappropriate model assumptions. In [9], an alternative approach is presented. Here, each template vector is initialized by a rough overtone model specifying the partials' energy distribution, see Fig. 1(c). Zero-valued entries between the expected partials enforce the intended structure during the refinement process. As a result, the learnt template vectors have an explicit harmonic structure, see Fig. 1(c). This is a significant gain in structure compared to the chaotic template vectors computed via standard NMF. In the following, we refer to this initialization strategy as I_{W}. Alternatively, another possibility is to constrain the activations instead of the template vectors. To this end, one can mark suitable regions in H where a given pitch is expected while setting the remaining entries to zero, see Fig. 1(d). This results in a similar factorization as the one using I_{W}. However, the results depend strongly on the input data. In our example, several pitches appear only in groups of two, such that the corresponding template vectors tend to be mixtures of those pitches. However, such conditions usually do not occur when using more extensive audio material instead of just short snippets. We refer to this initialization strategy in the following as I_{H}.

Opposed to previous methods, our main idea is to constrain both the template vectors and the activities, see Fig. 1(e). As to be expected, such double constraints lead to an increased stability and robustness of the factorization. While this will be experimentally shown in Section 3, it can also be observed in our example, see Fig. 1(e). Here, almost all template vectors have a well-defined harmonic structure. We refer to this combined strategy as I_{WH}. Furthermore, the robustness of I_{WH} even allows for introducing additional template vectors dedicated to describe onsets. This further stabilizes the factorization and leads to even more meaningful template vectors, see Fig. 1(f). In the next subsection, we describe this strategy, referred to as $I_{\mathrm{WH}}^{\mathrm{O}}$, in more detail.

2.2. Proposed Method

Overall, to use strategy $I^{\mathrm{O}}_{\mathrm{WH}}$, one needs to suitably initialize onset and harmony template vectors as well as their activations. After that, only the standard NMF updates rules have to be applied. For the harmony template vectors, our procedure essentially follows [9]. To this end, each vector is assigned to a pitch and then initialized such that only areas around the partials are non-zero. We choose the size of these areas relatively generous in order to be flexible in dealing with potential inharmonicities of the recorded instrument or non-standard tunings. More exactly, the area for the n-th partial of pitch p corresponds to the frequency range $(n \cdot f(p-\phi), n \cdot f(p+\phi))$, where ϕ is a parameter in semitones to control the size of these areas (we use $\phi = 1$ in our experiments). Here, $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$ defined by $f(p) := 2^{(p-69)/12} \cdot 440$ maps the pitch to the frequency scale. Furthermore, since the lower partials usually carry most of the energy, we set all entries in the n-th area to $1/n^2$, see Fig. 1(f). In a next step, we initialize the onset template vectors. Here, opposed to many other approaches, we take into account that the spectral shape for onsets is for many instruments (including the piano) not the same as for white noise but depends on the respective pitch with the energy being concentrated around the partials. Therefore, we use one onset vector for each pitch. Contrary to the harmonic templates, we do not enforce here any spectral constraints but initialize the onset templates uniformly and let the learning process derive their shape. To compensate for this lack of constraints, we apply more rigid restrictions on the activation side.

Next, to meaningfully initialize the activations, our method exploits available score information given in the form of a MIDI file. Here, instead of unrealistically assuming that a perfectly aligned MIDI file is available (as it is done in many of the previously described methods), we employ a high-resolution music synchronization approach to determine for each MIDI note event its corresponding position in the audio recording [10]. To impose the activation constraints, we essentially initialize H to look like a piano roll representation of the synchronized MIDI file. Starting with the activations for the harmony template vectors, we extract a pitch, as well as an onset and offset position from each MIDI event. Then, we set the corresponding entries in H to 1, while all remaining entries are set to zero. To account for possible alignment inaccuracies that occur using automatic synchronization procedures, we relax these constraints to some degree. To this end, we additionally set all entries in H in a $\mathrm{tol}_{\mathrm{on}}$-neighborhood around onsets and in a $\mathrm{tol}_{\mathrm{off}}$-neighborhood around offsets to 1 (in our experiments we use $\mathrm{tol}_{\mathrm{on}} = 0.2$ seconds and $\mathrm{tol}_{\mathrm{off}} = 1$ second). Then, in a final step, we initialize the activations for the onset template vectors. Here, we place more strict constraints by only setting entries in a $\mathrm{tol}_{\mathrm{on}}$-neighborhood around the MIDI onset positions to 1, see Fig. 1(f).

Comparing the $I^{\mathrm{O}}_{\mathrm{WH}}$ factorization to the others in Fig. 1, we see that the harmonic vectors of $I^{\mathrm{O}}_{\mathrm{WH}}$ have the clearest harmonic structure with most partials being very sharp in frequency direction. Here, a reason is that the percussive broadband energy is now well-described by the onset vectors, such that onsets have significantly less disturbing influence on the harmonic vectors. Furthermore, most onset vectors are activated in an impulse-like manner at the start of note events, which indeed indicates their use for representing onsets. Overall, making use of double constraints, the initialization strategy $I^{\mathrm{O}}_{\mathrm{WH}}$ allows for computing musically meaningful factorizations including a dedicated representation of onsets. It combines the expressive power of parametric approaches with the efficiency of classic NMF, while still being easy to implement. Furthermore, as shown in the next section, it is robust regarding smaller alignment errors as well as regarding potential inharmonicities of an instrument or non-standard tunings.

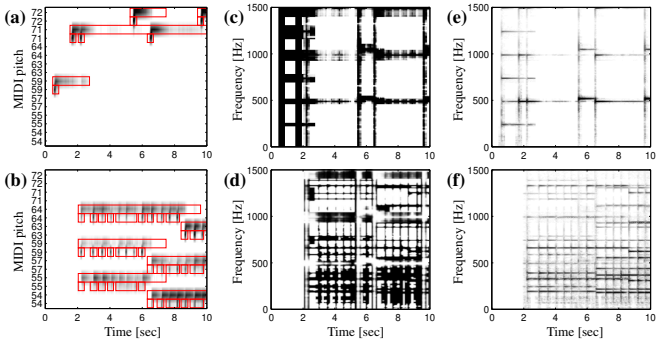

Fig. 2. Illustration of the separation process for the left and the right hand. (a)/(b): Partition of the activations matrix H (Fig. 1(e)) into H_L and H_R. (c)/(d): Masking matrices M_L and M_R. (e)/(f): Separated spectrograms.

3. EXPERIMENTS

In this section, we report on systematically conducted experiments to illustrate the potential of our method. To this end, we created a database consisting of ten pieces from the Western classical music repertoire. The database consists of four Bach pieces (mainly inventions) and six Chopin pieces (mainly preludes and mazurkas). Here, we used uninterpreted score-like MIDI files from the Mutopia Project[1], high-quality audio recordings from the Saarland Music Database (SMD)[2] as well as digitized versions of historical recordings from the Piano Society project[3]. In total, the database contains 24 minutes of music with each recording having a length between 30 seconds and 5 minutes.

In a first step, we indicate the quality of our approach quantitatively using synthetic audio data. To this end, we used the Mutopia MIDI files to create two additional MIDI files for each piece using only the notes of the left and the right hand, respectively. Using a wave table synthesizer, we then generated audio recordings from these MIDI files which are used as ground truth separation results in the following. A linear mix of these two recordings serves as input for all evaluated separation approaches. For the experiment, we compute a magnitude spectrogram of the mix and derive a factorization with one of the methods discussed in Section 2. To employ the factorization for the separation of the left and the right hand, we again make use of the available score information. While we could separate any user-defined group of notes, we exploit here that the used MIDI files specify which note event belongs to which hand. This way, we can partition the computed H into matrices H_L and H_R, see Fig. 2(a)/(b). From these matrices, we then derive masking matrices $M_L := (WH_L)/(WH+\epsilon)$ and $M_R := (WH_R)/(WH+\epsilon)$, where the division is understood pointwise and ϵ is a small positive constant to avoid a potential division by zero, see Fig. 2(c)/(d). Applying the masking matrices to the original mixture spectrogram V via pointwise multiplication yields a separate spectrogram for the left and the right hand, see Fig. 2(e)/(f). Finally, to yield the separated audio signals an inverse FFT in combination with an overlap-add technique is applied to the separated spectrograms using the phase of the original spectrogram.

To assess the quality of a separation result, we employ version 3.0 of the BSSEVAL toolkit [11] to compute *signal-to-distortion (SDR)* values. Fig. 3 shows SDR values for the initialization strategies I_{H}, I_{W}, I_{WH} and $I^{\mathrm{O}}_{\mathrm{WH}}$ separately for the left and the right hand as well as an average for both hands. All values are averaged over

[1] http://www.mutopiaproject.org
[2] http://www.mpi-inf.mpg.de/resources/SMD/
[3] http://pianosociety.com

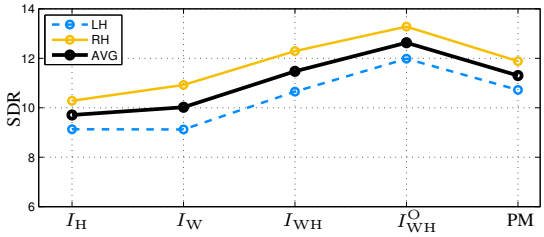

Fig. 3. Evaluation results given in SDR values for the left (LH) and the right hand (RH) as well as an average of both hands (AVG).

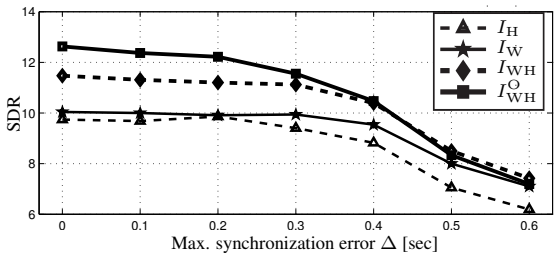

Fig. 4. Evaluation results given in SDR values for several separation approaches varying the synchronization error.

the ten pieces in our database. Note that the SDR values for the right hand are consistently higher than those for the left hand. Here, the main reason is that the right hand often corresponds to the main melody and is therefore played louder (level difference is 1.64 dB on average). As a consequence, there is more energy related to this hand in the mixture making the separation easier. Furthermore, we see in Fig. 3 that the strategies I_{H} and I_{W}, which initialize only one of the matrices, yield the lowest SDR values. Combining the two strategies (I_{WH}) we see a significant SDR-gain of almost 1.5 dB. Finally, integrating onset information leads to another substantial gain of 1.2 dB for the strategy $I^{\mathrm{O}}_{\mathrm{WH}}$. Here, the dedicated representation of the percussive sounds leads to a more coherent representation of the harmonic parts and consequently to a better separation quality.

To additionally indicate how a typical parametric model (PM) behaves in our scenario, we also include SDR values for a state-of-the-art approach based on spectrally localized Gaussians [7]. Similar to I_{WH}, this approach only models the harmonic part of a recording, i. e. no onset model is included, and, indeed, the average SDR values for both approaches are almost identical (11.3 dB and 11.47 dB SDR, respectively). However, the NMF factorization, using only simple matrix operations, can be computed more efficiently than the parameter estimation required for PM and additionally is easier to implement. For example, to process the whole database consisting of 24 minutes of music, our Matlab implementation takes about 6 minutes on an Intel W3530 for the synchronization and another 6 minutes for the factorization using 100 NMF iterations. Using more optimized implementations, both values could be reduced even further. This is in contrast to many parametric approaches, which often require several hours to process this amount of data. Furthermore, the straightforward integration of an onset model allows for a significant SDR-gain for the $I^{\mathrm{O}}_{\mathrm{WH}}$ strategy over PM.

Since the MIDI files were perfectly aligned to their sonifications in the first experiment, we also investigated how the synchronization accuracy affects the separation performance. To this end, we simulate a low accuracy alignment by shifting each note event randomly by $\pm\Delta$ seconds. Fig. 4 gives the SDR values for the four initialization strategies and varying values for Δ. Here, we see that all approaches are relatively stable as long as the synchronization error is not larger than $\mathrm{tol}_{\mathrm{on}}$ (200 ms in our experiments). Beyond $\mathrm{tol}_{\mathrm{on}}$, all SDR values drop significantly. However, it should be noted that even with very inaccurate onset information the strategy $I^{\mathrm{O}}_{\mathrm{WH}}$ stays on a similar level as I_{WH} demonstrating its overall robustness.

Since signal-to-distortion ratios and similar evaluation measures often do not capture the perceptual separation quality, we additionally provide a website with audible separation results[4]. Here, real, non-synthetic audio recordings from the SMD and Piano Society databases are used to give a realistic and perceptually meaningful impression of the quality of our approach in real world scenarios.

4. CONCLUSIONS

In this paper, we have presented an extended NMF variant that exploits available score information to guide the factorization. Based on the idea of simultaneously constraining both the template vectors as well as the activations, the method yields similar results as a state-of-the-art parametric approach. These results are further improved by integrating template vectors dedicated to representing onsets. In the future, we plan to further extend the idea of double constraining to integrate further model assumptions into the NMF framework and to apply our framework to other types of music.

5. REFERENCES

[1] T. Heittola, A. Klapuri, and T. Virtanen, "Musical instrument recognition in polyphonic audio using source-filter model for sound separation," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Kobe, Japan, 2009, pp. 327–332.

[2] K. Itoyama, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Instrument equalizer for query-by-example retrieval: Improving sound source separation based on integrated harmonic and inharmonic models," in *Proceedings of the International Conference for Music Information Retrieval (ISMIR)*, Philadelphia, USA, 2008, pp. 133–138.

[3] J. Ganseman, P. Scheunders, G. J. Mysore, and J. S. Abel, "Source separation by score synthesis," in *Proceedings of the International Computer Music Conference (ICMC)*, New York, USA, 2010, pp. 462–465.

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

[5] M. Shashanka, B. Raj, and P. Smaragdis, "Probabilistic latent variable models as nonnegative factorizations (article id 947438)," *Computational Intelligence and Neuroscience*, vol. 2008.

[6] R. Hennequin, R. Badeau, and B. David, "Time-dependent parametric and harmonic templates in non-negative matrix factorization," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, Graz, Austria, 2010, pp. 246–253.

[7] S. Ewert and M. Müller, "Score-informed voice separation for piano recordings," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Miami, USA, 2011.

[8] J. Wu, E. Vincent, S. A. Raczynski, T. Nishimoto, N. Ono, and S. Sagayama, "Multipitch estimation by joint modeling of harmonic and transient sounds," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, pp. 25–28.

[9] S. A. Raczynski, N. Ono, and S. Sagayama, "Multipitch analysis with harmonic nonnegative matrix approximation," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 2007, pp. 381–386.

[10] S. Ewert, M. Müller, and P. Grosche, "High resolution audio synchronization using chroma onset features," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009, pp. 1869–1872.

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

[4] http://www.mpi-inf.mpg.de/resources/MIR/ICASSP2012-ScoreInformedNMF