GUIDING AUDIO SOURCE SEPARATION BY VIDEO OBJECT INFORMATION

Sanjeel Parekh[⋆†] *Slim Essid*[⋆] *Alexey Ozerov*[†] *Ngoc Q. K. Duong*[†] *Patrick Pérez*[†] *Gaël Richard*[⋆]

[⋆] LTCI, Télécom ParisTech, Université Paris–Saclay, 75013, Paris, France
[†] Technicolor, 975 avenue des Champs Blancs, CS 17616, 35576 Cesson Sévigné, France

ABSTRACT

In this work we propose novel joint and sequential multimodal approaches for the task of single channel audio source separation in videos. This is done within the popular non-negative matrix factorization framework using information about the sounding object's motion. Specifically, we present methods that utilize non-negative least squares formulation to couple motion and audio information. The proposed techniques generalize recent work carried out on NMF-based motion-informed source separation and easily extend to video data. Experiments with two distinct multimodal datasets of string instrument performance recordings illustrate their advantages over the existing methods.

Index Terms— Audio source separation, Nonnegative matrix factorization, Audio-visual objects, Motion, Multimodal analysis

1. INTRODUCTION

Several sounds in the real world are *visually indicated* through their relation to the sound-producing motion. This paper focuses on single channel audio source separation in audiovisual recordings of such sound mixtures by leveraging the accompanying motion information in the visual stream. Be it sound of people talking, playing an instrument or scratching a surface, both the audio and visual streams carry some common information about the physical interaction. In this study we attempt to identify, extract, and couple features from both modalities that represent this shared knowledge.

In the past, numerous frameworks have been developed for the task of single channel audio source separation, *e.g.* [1, 2]. The part-based decomposition of audio spectra into its spectral patterns and their activations obtained from nonnegative matrix factorization (NMF) makes it a particularly appealing and popular method for tackling this problem. In many cases, NMF-based frameworks have been used in a supervised manner, where spectral patterns are first learnt over clean examples. To perform source separation over a single mixture without any training step, several non-supervised (or blind) audio-only methods have been propounded. A generic Mel-spectra based spectral pattern clustering approach was proposed by Spiertz *et al.* [3]. Other methods involving shifted NMF or linear predictive coding were introduced subsequently [4, 5].

Use of auxiliary information for assisting source separation has also seen growing interest. Some examples include exploitation of score information for musical recordings [6, 7], text for speech [8], user-assistance [9], *etc.* For brevity, here we only elaborate on methods utilizing motion data.

In most cases, motion information is extracted from the video images and utilized within various settings including NMF. Early work by Fisher *et al.* [10] sought to learn a multimodal embedding through mutual information (MI) maximization. This is then used to tackle the task of user-assisted audio enhancement. However, Parzen window estimation for computing MI is complex and may suffer in quality when the data used to perform the estimation is limited. Some other works propose to do so in an unsupervised manner using sparse representations [11], audio-visual independent components [12], and onsets coincidence [13]. Some limitations of these methods include multiple parameter tuning and performance degradation in complex videos. Score information has also been used along with joint AV processing for source separation and player association in music videos [14]. Some recent studies [15], including ours [16], demonstrate the advantages of using motion within the NMF framework, however, their application to generic videos is not straightforward.

This work stems from the following intuition: motion features such as velocity, obtained from visual analysis, encode information about the physical excitation of a sounding object. On the other hand, for the audio modality, a representation of this excitation can be found in the spectral component activation matrix obtained after NMF decomposition. Thus, our hypothesis is that a set of audio activations would be "similar" to the velocity of *sound-producing* motion. We establish the idea's effectiveness for audio source separation through experiments on two very challenging multimodal string quartet performance datasets involving video and motion capture data. In particular, the proposed sequential and joint approaches extend and improve upon earlier work (i) by making it independent of specific inputs such as bow inclination in [16], or lip surface signals in [15] (as a result we eliminate the need to provide a preconstructed motion activation matrix), and (ii) by showing applicability of the proposed methods to complex videos.

The paper is organized as follows. We present an overview and technical details of our approach in Section 2 and 3 respectively. This is followed by experimental validation in Section 4 and concluding remarks in Section 5.

2. OVERVIEW

The problem of single channel audio source separation consists in obtaining an estimate for each of the J sources s_j forming the observed linear mixture $x(t)$:

$$x(t) = \sum_{j=1}^{J} s_j(t). \tag{1}$$

Using NMF we can decompose the mixture magnitude or power spectrogram $\mathbf{V} \in \mathbb{R}_+^{F\times N}$ consisting of F frequency bins and N short-time Fourier transform (STFT) frames, such that,

$$\mathbf{V} \approx \mathbf{WH}, \tag{2}$$

where $\mathbf{W} \in \mathbb{R}_+^{F\times K}$and $\mathbf{H} \in \mathbb{R}_+^{K\times N}$ are interpreted as the nonnegative audio spectral patterns and their activation matrices respectively. Here K is the total number of spectral patterns. Multiplicative update rules for estimating $\mathbf{W}$ and $\mathbf{H}$ can be obtained by minimizing a divergence cost function [17].

As already discussed, when dealing with only a single mixture, without a training step as in the supervised case, the source separation problem in the NMF framework reduces to assigning the appropriate spectral patterns, *i.e.*, each of the K components in the columns of $\mathbf{W}$, to the J sources. Here we propose methods to guide this assignment through the use of associated source-specific motion information. We discuss next each building block of our approach depicted in Fig. 1.

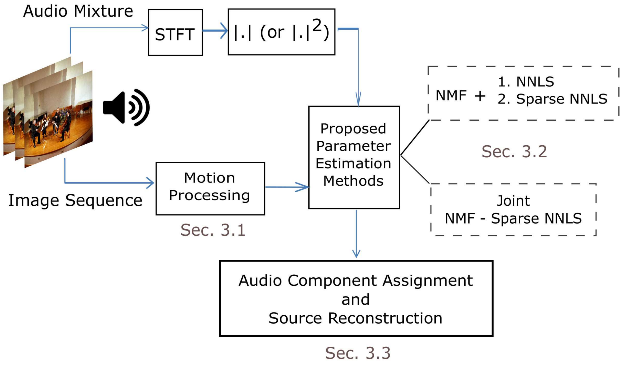
Audio Mixture
STFT
|.| (or |.|^2)
Image Sequence
Motion Processing
Sec. 3.1
Proposed Parameter Estimation Methods
NMF + 1. NNLS 2. Sparse NNLS
Sec. 3.2
Joint NMF - Sparse NNLS
Audio Component Assignment and Source Reconstruction
Sec. 3.3

Figure 1: Overview of our approach

3. TECHNICAL DETAILS

3.1. Motion Processing Unit

The motion data must be suitably processed to establish meaningful cross-modal relations. To begin with, for image sequences extracted from videos we assume that the spatial location of each moving AV object is known (in a user-assisted manner or otherwise), as shown through the bounding boxes in Fig. 2. Note that J audio sources correspond to the same number of visual objects in the images. For each such moving region we proceed as follows:

- First, motion trajectory segmentation is performed on the image sequence using a state-of-the-art multicuts-based formulation [18]. As shown in Fig. 2, here the idea is to cluster point trajectories with respect to their motion similarity.
- Next, for each trajectory, we compute the velocity by taking differences over consecutive frames in x-y directions.
- In the final step, average magnitude velocities over all trajectories in each cluster are computed frame-wise.

Thus, we get C_j motion clusters per audio source (each source being associated to a different performer's bounding box). Their velocity vectors are resampled to match the N STFT frames and arranged in the columns of a matrix $\mathbf{M} \in \mathbb{R}_+^{N \times C}$ where $C = \sum_{j=1}^{J} C_j$.

3.2. Proposed parameter estimation techniques using NMF and Nonnegative Least Squares

To illustrate the central idea of the methods given below assume that the magnitude trajectory of a violinist's bow velocity, given by $\mathbf{m}_{vbow} \in \mathbb{R}_+^N$ is known for a string quartet performance recording. We can then try to determine a linear transformation $\boldsymbol{\alpha}_{vbow} \in \mathbb{R}_+^K$ of the activation matrix $\mathbf{H}$ such that $\mathbf{H}^\top \boldsymbol{\alpha}_{vbow}$ is similar to $\mathbf{m}_{vbow}$ with respect to ℓ_2-norm based reconstruction error. This can be formulated as a nonnegative least squares (NNLS) cost function. Ideally, we expect $\boldsymbol{\alpha}_{vbow}$ to be sparse. In other words, to be concentrated on a few coefficients which indicate that few activations of spectral patterns are linked to bow velocity.

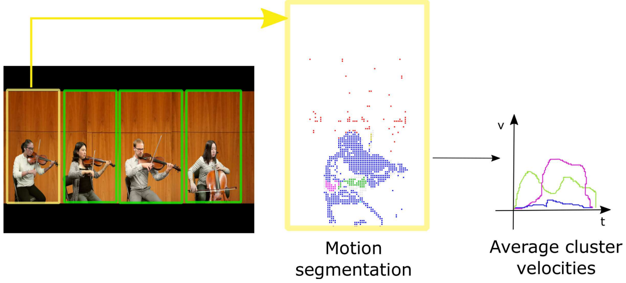

Figure 2: Motion Processing Unit: For each bounding box in the video (left), we compute motion segmentation using multicuts algorithm [18] (centre) and finally, average velocities over each cluster (right). Some clusters in pink, green, blue (foreground) and red (background) are visible. The graph on the right is only a sketch.

Thus, at this step of the algorithm, we determine this linear transformation, denoted by $\boldsymbol{\alpha}_c$, for each velocity vector $\mathbf{m}_c \in \mathbb{R}_+^N$ in $\mathbf{M}$, where $c = 1 \cdots C$, with the expectation that the ones corresponding to the sound-producing motion would be sparse as stated in the illustrative example above.

Defining the nonnegative linear combination coefficient matrix as $\mathbf{A} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_C]$, the following joint and sequential pathways could be taken for determining $\mathbf{A}$ while minimizing the Frobenius reconstruction error between $\mathbf{M}$ and $\mathbf{H}^\top \mathbf{A}$:

Sequential estimation

Two alternative schemes are considered here:

1. **NMF + NNLS**: After obtaining an NMF decomposition of the audio mixture, we perform NNLS where the objective is to determine $\mathbf{A}$ that best reconstructs $\mathbf{M}$ from the given audio activations $\mathbf{H}$. This can be written mathematically as:

$$\underset{\mathbf{A} \geqslant 0}{\text{minimize}} \quad \|\mathbf{M} - \mathbf{H}^\top \mathbf{A}\|_F^2. \tag{3}$$

The above formulation is equivalent to solving the NMF problem with $\mathbf{H}$ held constant.

2. **NMF + Sparse NNLS**: Within the previous formulation, concentration of $\boldsymbol{\alpha}_c$ on a few coefficients can be achieved by incorporating a sparsity constraint. This can be achieved through an ℓ_1-regularization term as follows:

$$\underset{\mathbf{A} \geqslant 0}{\text{minimize}} \quad \|\mathbf{M} - \mathbf{H}^\top \mathbf{A}\|_F^2 + \mu \|\mathbf{A}\|_1, \tag{4}$$

where μ is a positive constant. Equation (4) can be looked at as a sparse-NMF formulation where the basis vectors (here $\mathbf{H}^\top$) are held constant [19].

Joint NMF-Sparse NNLS

Here we propose a *novel* joint formulation where the cost functions for audio factorization and sparse-NNLS are simultaneously minimized:

$$C(\mathbf{W},\mathbf{H},\mathbf{A}) = D_{KL}(\mathbf{V}|\mathbf{WH}) + \lambda\|\mathbf{M} - \mathbf{H}^\top\mathbf{A}\|_F^2 + \mu\|\mathbf{A}\|_1, \quad (5)$$

where $D_{KL}(.|.)$ is the Kullback-Leibler divergence and λ is a regularization parameter. Note that it is trivial to minimize the cost function in absence of scaling constraint: $C(\gamma\mathbf{W},\mathbf{H}/\gamma,\mathbf{A}\gamma) < C(\mathbf{W},\mathbf{H},\mathbf{A})$ where γ is close to zero. Therefore, we constrain the columns of $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_K]$ to have unit norm:

$$\underset{\substack{(\mathbf{W},\mathbf{H},\mathbf{A})\geqslant 0 \\ \|\mathbf{w}_k\|=1,\ \forall k}}{\text{minimize}} \quad D_{KL}(\mathbf{V}|\mathbf{WH}) + \lambda\|\mathbf{M} - \mathbf{H}^\top\mathbf{A}\|_F^2 + \mu\|\mathbf{A}\|_1. \quad (6)$$

Details regarding the update rules for each variable and implementation are summarized in Algorithm 1. Specifically, we use multiplicative update heuristics to derive rules for $\mathbf{H}$, $\mathbf{W}$ and $\mathbf{A}$ given on line (8), (10) and (13) of Algorithm 1 respectively. Update rule for $\mathbf{W}$ is derived as in [19]. To avoid confusion and clutter we use $\mathbf{\Lambda} = \mathbf{WH}$. Product $\odot$ and exponents denote element-wise operations, $\mathbf{1}$ denotes a matrix with all entries equal to one and size given by context.

Algorithm 1 Joint NMF-Sparse NNLS

```
1: Input: $\mathbf{V}, \mathbf{M}, K, \lambda \geq 0, \mu \geq 0$
2: $\mathbf{W}$, $\mathbf{H}$, $\mathbf{A}$ initialized randomly
3: $\mathbf{H} \leftarrow \text{diag}(\|\mathbf{w}_1\|, \ldots, \|\mathbf{w}_K\|)\mathbf{H}$
4: $\mathbf{A} \leftarrow \text{diag}(\|\mathbf{w}_1\|^{-1}, \ldots, \|\mathbf{w}_K\|^{-1})\mathbf{A}$
5: $\mathbf{W} \leftarrow \mathbf{W}\text{diag}(\|\mathbf{w}_1\|^{-1}, \ldots, \|\mathbf{w}_K\|^{-1})$    ▷ Normalize
6: $\mathbf{\Lambda} = \mathbf{WH}$
7: repeat
8:     $\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^\top(\mathbf{V} \odot \mathbf{\Lambda}^{-1}) + \lambda\mathbf{A}\mathbf{M}^\top}{\mathbf{W}^\top\mathbf{1} + \lambda\mathbf{A}\mathbf{A}^\top\mathbf{H}}$
9:     $\mathbf{\Lambda} = \mathbf{WH}$
10:    $\mathbf{W} \leftarrow \mathbf{W} \odot \frac{(\mathbf{\Lambda}^{-1} \odot \mathbf{V})\mathbf{H}^\top + \mathbf{W} \odot (\mathbf{1}(\mathbf{W} \odot (\mathbf{1}\mathbf{H}^\top)))}{\mathbf{1}\mathbf{H}^\top + \mathbf{W} \odot (\mathbf{1}(\mathbf{W} \odot ((\mathbf{\Lambda}^{-1} \odot \mathbf{V})\mathbf{H}^\top)))}$
11:    $\mathbf{W} \leftarrow \mathbf{W}\text{diag}(\|\mathbf{w}_1\|^{-1}, \ldots, \|\mathbf{w}_K\|^{-1})$
12:    $\mathbf{\Lambda} = \mathbf{WH}$
13:    $\mathbf{A} \leftarrow \mathbf{A} \odot \frac{\lambda\mathbf{HM}}{\lambda\mathbf{H}\mathbf{H}^\top\mathbf{A} + \mu}$
14: until convergence
15: return $\mathbf{W}, \mathbf{H}, \mathbf{A}$
```

3.3. Audio spectral pattern assignment and reconstruction

Once we obtain $\mathbf{A}$, which contains $\boldsymbol{\alpha}_c$ for each of the C velocity clusters, the k-th basis vector is assigned to the j^{th} source if $\text{argmax}_c\, \alpha_{kc}$ belongs to the j^{th} source cluster. Once these assignments are made, each source is reconstructed by element-wise multiplication of the soft mask, given by $(\mathbf{W}_j\mathbf{H}_j)./(\mathbf{WH})$ where "./" stands for element-wise division, with the mixture spectrogram followed by an inverse STFT. Here $\mathbf{W}_j$ and $\mathbf{H}_j$ are the submatrices for spectral patterns and their activations assigned to the j^{th} source by the above-mentioned scheme.

4. RESULTS AND DISCUSSION

The performance of the proposed methods is evaluated through tests with two distinct multimodal datasets. General implementation details common to all experiments are detailed below. Some separation results and supplementary material is made available on our companion web page.[1] We evaluate with the following techniques (See Section 3.2):

- **LS**: NMF + NNLS;
- **spLS**: NMF + Sparse-NNLS with $\mu = 5$;
- **JLS Rand**: Joint NMF-Sparse NNLS with $\mathbf{W}$ and $\mathbf{H}$ initialized randomly, $\lambda = 0.01$ and $\mu = 0.1$;
- **JLS NMF**: Joint NMF-Sparse NNLS with $\mathbf{W}$ and $\mathbf{H}$ initialized using the output obtained after applying NMF to the mixture, $\lambda = 0.01$ and $\mu = 0.1$.

Hyperparameter values were decided after a crude grid-search using an example mixture.

General Implementation Details. For all the experiments audio spectrogram is computed with a Hamming window of size 4096 (92 ms) and 75% overlap. Thus, we have a $2049 \times N$ matrix where N is the number of STFT frames. Code provided by Févotte *et al.* [22] is used for standard NMF algorithms. *LS* and *spLS* formulations are implemented using publicly available sparse-NMF code [19], with sparsity set to zero for *LS*.

Evaluation metrics. We report Signal to Distortion Ratio (SDR) expressed in dB, computed using the BSS_EVAL Toolbox version 3.0 [23]. NMF for each of the methods is run for 200 iterations. For each mixture, all methods are run 5 times with different random initializations and the reconstruction is performed using a soft mask. SDR is averaged over all runs and mixtures of each set.

4.1. Experiments with Motion Capture Data

These experiments are performed with the publicly available Ensemble Expressive Performance (EEP) dataset[2] [20] which contains multimodal recordings of string quartet performances. We construct mixtures from four excerpts labeled from P1 to P4, exactly as in [16], using the available sources, namely: Violin, Viola and Cello. The acquired multimodal data consists of audio tracks and motion capture for each musician. Bow velocity descriptor for each source is used as a substitute for the velocities extracted from moving regions. This is meant to validate the proposed factorisation schemes with "ideal" motion features before considering the more challenging video scenario as described in Sec 3.1. Thus in this simple case, total number of clusters in $\mathbf{M}$ is equal to the number of sources in each mixture ($C = J$). For all the proposed methods the number of audio components is set to $(15 \times J)$, *e.g.* $K = 30$ for mixtures with 2 sources.

Setup: We evaluate over the following sets of mixtures:

1. **Set 1**: 4 trios of violin, viola and cello;
2. **Set 2**: 6 two-source combinations of the three instruments for pieces P1 - P2;
3. **Set 3**: 3 two-source combinations of the same instrument from different pieces, *e.g.*, a mix of 2 violins from P1 & P2.

We compare with the following earlier works:

[1] goo.gl/y7A5az

[2] http://mtg.upf.edu/download/datasets/eep-dataset

Mixtures	LS	spLS	JLS Rand	JLS NMF	sMcNMF	Mel NMF
Set 1 - 4 trios	1.23	1.27	0.82	**1.95**	1.35	1.22
Set 2 - 6 duos different instruments	4.54	4.52	3.89	**5.70**	3.88	5.08
Set 3 - 3 duos same instrument	0.86	0.80	0.55	0.93	**2.29**	-0.89

Table 1: **MoCap Dataset** [20]: SDR for different methods averaged over mixtures of each set. Best SDR is displayed in bold.

Mixtures	LS	spLS	JLS Rand	JLS NMF	Mel NMF
Duos	6.94	6.94	5.30	**7.14**	5.11
Trios	**3.30**	3.26	1.79	3.24	2.18
Quartet	0.97	1.00	-0.76	0.67	**1.01**

Table 2: **URMP Video** [21]: SDR for different methods averaged over mixtures of each set. Best SDR is displayed in bold.

1. **Mel NMF** [3]: This is a unimodal approach where basis vectors learned from the mixture are clustered based on the similarity of their mel-spectra. We take help of the example code provided online for implementation.[3] Like in all the proposed methods K is set to $(15 \times J)$.
2. **Soft Motion Coupled NMF (sMcNMF)** [16]: Audio and motion activations are coupled through a soft ℓ_1 constraint. The motion activations utilize quantized bow inclination. We retain the original parameter settings with number of basis vectors set to 4 for each instrument.

Since the MoCap data is sampled at 240 Hz, each of the selected descriptors is resampled to match the N STFT audio frames.

Discussion: From Table 1 we see that the joint approach with audio-NMF initialization outperforms all the other methods for the first two sets. Indeed, when compared with its random version, it seems to converge both faster and better. Moreover, it appears that sparsity for sequential NNLS does not provide significant improvements over LS. As expected, spLS attempts to concentrate weight on a few coefficients, but these are not very different from those yielded by LS. Also, as we are only interested in maximum values for component assignment, any existing differences are not visible in the reconstruction.

When confronted with sources having similar motion, the performance of the proposed methods is deemed to degrade. In this respect, EEP dataset is particularly challenging where we find multiple mixture segments with similar motion. In such cases information such as bow inclination (used by sMcNMF) proves to be quite useful. It is worth mentioning that for some mixtures, all the proposed methods outperform the baselines by a large margin.

4.2. Experiments with Videos

In this second series of experiments, we apply the proposed methods to videos. As no standard dataset exists for such a task, we consider the only publicly available example video from the URMP dataset [21].[4] We are provided with video recording of a string quartet performance and the separate audio tracks for each player. We consider a 5 sec. excerpt from 30-35s and compute the motion trajectory segmentation for each moving region bounding box (as depicted in Fig. 2) using publicly available binary from [18] with default parameter setting. The calculated velocity trajectories are resampled to match N STFT frames. We consider all two, three and four source combinations, denoted by **Duos** (6 mixtures), **Trios** (4 mixtures) and **Quartet** (1 mixture) in Table 2. We compare only with Mel-NMF as sMcNMF or the other NMF-based methods are not designed to deal with generic videos. As in the previous case, K is set to $(15 \times J)$ for all methods.

Discussion: Efficacy of the proposed methods is seen from the results in Table 2, with particular emphasis on good initialization for the joint approach. For the quartet with one mixture, though we see that Mel-NMF performs better, it is a particularly difficult case where none of the methods work consistently well over all five runs and the SDR variance is high.

Unlike the previous experiment, here we have multiple velocity clusters to choose from, which makes the problem considerably more difficult. We note that the methods deal reasonably well with low velocity unrelated/noisy clusters as they are not strongly related to any audio activation. Interestingly, it is possible to identify motion trajectory clusters responsible for the sound of each source using $\mathbf{A}$. As indicated by $\mathrm{argmax}_c\, \alpha_{kc}$, for each source we can determine the velocity cluster with maximum audio component assignments. These localization results are illustrated in Fig.3 for a mixture of violin and cello. This provides additional evidence for our hypothesis and the proposed estimation methods.

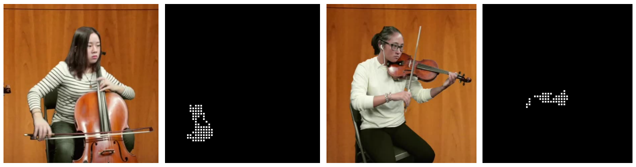

Figure 3: Localization results for the first frame of cello and violin. The clusters corresponding to the hand *i.e.* the bowing motion (in white) have been identified in both cases.

5. CONCLUSION

To summarize, we have proposed novel visually-assisted methods for source separation in audiovisual recordings. This was done by exploiting features encoding physical excitation information in both modalities. In addition to demonstrating the usefulness of this idea through sequential techniques, we present and derive algorithm for an original joint formulation. While the extension to audio denoising is straightforward, in their current form the methods cannot deal with high amplitude noisy visual motion. This can possibly be alleviated through group sparsity constraints over $\mathbf{A}$.

For the particular case of musical mixtures, score information would prove to be very beneficial in guiding source separation. It can certainly be incorporated within the present framework, which will be a topic for further study. Several other loss functions and non-linear methods for establishing similarity could be experimented with for better performance and wider applicability.

[3] http://www.ient.rwth-aachen.de/cms/dafx09/

[4] The full dataset is yet to be released. Sample video 32- The Art of the Fugue can be found at http://www.ece.rochester.edu/projects/air/projects/datasetproject.html

6. REFERENCES

[1] B. Wang and M. D. Plumbley, "Investigating single-channel audio source separation methods based on non-negative matrix factorization," in *Proc. ICA Research Network International Workshop*, 2006, pp. 17–20.

[2] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 1562–1566.

[3] M. Spiertz and V. Gnann, "Source-filter based clustering for monaural blind source separation," in *Proc. Int. Conf. on Digital Audio Effects DAFx09*, 2009.

[4] R. Jaiswal, D. FitzGerald, D. Barry, E. Coyle, and S. Rickard, "Clustering NMF basis functions using shifted NMF for monaural sound source separation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 245–248.

[5] X. Guo, S. Uhlich, and Y. Mitsufuji, "NMF-based blind source separation using a linear predictive coding error clustering criterion," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 261–265.

[6] S. Ewert, B. Pardo, M. Müller, and M. D. Plumbley, "Score-informed source separation for musical audio recordings: An overview," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 116–124, 2014.

[7] C. Dittmar and M. Müller, "Reverse engineering the amen break: score-informed separation and restoration applied to drum recordings," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 9, pp. 1531–1543, 2016.

[8] L. Le Magoarou, A. Ozerov, and N. Q. K. Duong, "Text-informed audio source separation. example-based approach using non-negative matrix partial co-factorization," *Journal of Signal Processing Systems*, vol. 79, no. 2, pp. 117–131, 2015.

[9] N. Q. K. Duong, A. Ozerov, L. Chevallier, and J. Sirot, "An interactive audio source separation framework based on non-negative matrix factorization," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 1567–1571.

[10] J. W. Fisher III, T. Darrell, W. T. Freeman, and P. Viola, "Learning Joint Statistical Models for Audio-Visual Fusion and Segregation," in *Advances in Neural Information Processing Systems*, no. Ml, 2001, pp. 772–778.

[11] A. L. Casanovas, G. Monaci, P. Vandergheynst, and R. Gribonval, "Blind audiovisual source separation based on sparse redundant representations," *Multimedia, IEEE Transactions on*, vol. 12, no. 5, pp. 358–371, Aug 2010.

[12] P. Smaragdis and M. Casey, "Audio/visual independent components," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, 2003, pp. 709–714.

[13] Z. Barzelay and Y. Y. Schechner, "Harmony in motion," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[14] B. Li, Z. Duan, and G. Sharma, "Associating players to sound sources in musical performance videos," *Late Breaking Demo, Intl. Soc. for Music Info. Retrieval (ISMIR)*, 2016.

[15] F. Sedighin, M. Babaie-Zadeh, B. Rivet, and C. Jutten, "Two multimodal approaches for single microphone source separation," in *EUSIPCO*, 2016.

[16] S. Parekh, S. Essid, A. Ozerov, N. Q. K. Duong, P. Perez, and G. Richard, "Motion informed audio source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, 2017.

[17] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[18] M. Keuper, B. Andres, and T. Brox, "Motion trajectory segmentation via minimum cost multicuts," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3271–3279.

[19] J. Le Roux, F. Weninger, and J. R. Hershey, "Sparse NMF–half-baked or well done?" *Mitsubishi Electric Research Labs (MERL), Cambridge, MA, USA, Tech. Rep., no. TR2015-023*, 2015.

[20] M. Marchini, R. Ramirez, P. Papiotis, and E. Maestre, "The sense of ensemble: a machine learning approach to expressive performance modelling in string quartets," *Journal of New Music Research*, vol. 43, no. 3, pp. 303–317, 2014.

[21] B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma, "Creating a musical performance dataset for multimodal music analysis: Challenges, insights, and applications," *arXiv preprint arXiv:1612.08727*, 2016.

[22] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[23] 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.