Motion informed audio source separation

Sanjeel Parekh, Slim Essid, Alexey Ozerov, Ngoc Duong, Patrick Perez, Gaël Richard

MOTION INFORMED AUDIO SOURCE SEPARATION

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 paper we tackle the problem of single channel audio source separation driven by descriptors of the sounding object's motion. As opposed to previous approaches, motion is included as a soft-coupling constraint within the nonnegative matrix factorization framework. The proposed method is applied to a multimodal dataset of instruments in string quartet performance recordings where bow motion information is used for separation of string instruments. We show that the approach offers better source separation result than an audio-based baseline and the state-of-the-art multimodal-based approaches on these very challenging music mixtures.

Index Terms— audio source separation, nonnegative matrix factorization, motion, multimodal analysis

1. INTRODUCTION

Different aspects of an event occuring in the physical world can be captured using different sensors. The information obtained from one sensor, referred to as a modality, can then be used to disambiguate noisy information in the other, based on the correlations that exist between the two. In this context, consider the scene of a busy street or a music concert: what we hear in these scenarios is a mix of sounds coming from multiple sources. However, information received from the visual system in terms of movement of these sources over time is very useful for decomposing and associating them with their respective audio streams [1]. Indeed, often, there exists a corrrelation between sounds and the motion responsible for the production of those sounds. Thus, machines too could use joint analysis of audio and motion to perform computational tasks in either of the modalities which are otherwise difficult. In this paper we are interested in audio and motion modalities. Specifically, we demonstrate how information from sound-producing motion can be used to perform the challenging task of single channel audio source separation.

Several approaches have been proposed for monaural source separation in the unimodal case, *i.e.*, methods using only audio [2–5], in which nonnegative matrix factorization (NMF) has been the most popular one. Typically, source separation in the NMF framework is performed in a supervised manner [2], where magnitude or power spectrogram of an audio mixture is factorized into nonegative spectral patterns and their activations. In the training phase, spectral patterns are learnt over clean source examples and then factorization is performed over test examples while keeping the learnt spectral patterns fixed. In the last few years, several methods have been proposed to group together appropriate spectral patterns for source estimation without the need for a dictionary learning step. Spiertz *et al.* [6] proposed a promising and generic basis vector clustering approach using Mel-spectra. Subsequently methods based on shifted-NMF, inspired by western music theory and linear predictive coding were proposed [7, 8]. While the latter has been shown to work well with harmonic sounds, its applicability to percussive sounds will be limited.

In the single channel case it is possible to improve system performance and avoid the spectral pattern learning phase by incorporating auxiliary information about the sources. The inclusion of side information to guide source separation has been explored within task-specific scenarios such as text informed separation for speech [9] or score-informed separation for classical music [10]. Recently, there has also been much interest in user-assisted source separation where the side information is obtained by asking the user to hum, speak or provide time-frequency annotations [11–13].

Another trend is to guide audio source separation using video. In such cases, information about motion is extracted from the video images. One of the first works was that of Fisher *et al.* [14] who utilize mutual information (MI) to learn a joint audio-visual subspace. The Parzen window estimation for MI computation is complex and requires determining many parameters. Another technique which aims to extract audio-visual (AV) independent components [15] does not work well with dynamic scenes. Later, work by Barzeley *et al.* [16] considered onset coincidence to identify AV objects and subsequently perform source separation. They dileanate several limitations of their work, including: setting multiple parameters for optimal performance on each example and possible performance degradation in dense audio environments. Application of AV source separation work using sparse representations [17] is limited due to their method's dependence on active-alone regions to learn source characteristics. Also, they assume that all the audio sources are seen on-screen which is not always realistic. A recent work proposes to perform AV source separation and association for music videos using score information [18]. Some prior work on AV speech separation has also been carried out [19, 20], primary drawbacks being the large number of parameters and hardware requirements.

Thus, in this work we improve upon several limitations of the earlier methods. With the exception of a recently published study [21], to the best of our knowledge no previous work has incorporated motion into the NMF-based source separation systems. Moreover, as we demonstrate in Section 3, the applicability of methods proposed in [21] is limited. Our approach utilizes motion information within the NMF parameter estimation procedure through soft coupling rather than a separate step after factorization. This not only preserves flexibility and efficiency of the NMF system, but unlike previous motion-based approaches, significantly reduces the number of parameters to tune for optimal performance (to effectively just one). Particularly, we show that in highly non-stationary scenarios, information from motion related to the causes of sound vibration from each source can be very useful for source separation. This is demonstrated through the application of the proposed method to musical instrument source separation in string trios using bow motion information. To the best of our knowledge this paper describes the first study to use motion capture data for audio source separation.

The rest of the paper is organized as follows: In Section 2 we discuss our approach followed by the experimental validation in Section 3. Finally we conclude with a mention of ongoing and future work in Section 4.

2. PROPOSED APPROACH

Given a linear instantaneous mixture of J sources

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

the goal of source separation is to obtain an estimate for each of the J sources, s_j.

Within the NMF framework this is done by obtaining a low-rank factorization for the mixture magnitude or power spectrogram $\mathbf{V}_a \in \mathbb{R}_+^{F\times N}$ consisting of F frequency bins and N short-time Fourier transform (STFT) frames, such that,

$$\mathbf{V}_a \approx \hat{\mathbf{V}} = \mathbf{W}_a\mathbf{H}_a, \tag{2}$$

where $\mathbf{W}_a = (w_{a,fk})_{f,k} \in \mathbb{R}_+^{F\times K}$ and $\mathbf{H}_a = (h_{a,kn})_{k,n} \in \mathbb{R}_+^{K\times N}$ are interpreted as the nonnegative audio spectral patterns and their activation matrices respectively. Here K is the total number of spectral patterns. Matrices $\mathbf{W}_a$ and $\mathbf{H}_a$ can be estimated sequentially with multiplicative updates obtained by minimizing a divergence cost function [22].

2.1. Motion Informed Source Separation

We assume that we now have information about the causes of sound vibration of each source in the form of motion activation matrices $\mathbf{H}_{m_j} \in \mathbb{R}_+^{K_{m_j}\times N}$, vertically stacked into a matrix $\mathbf{H}_m \in \mathbb{R}_+^{K_m\times N}$:

$$\mathbf{H}_m = \begin{bmatrix} \mathbf{H}_{m_1} \\ \vdots \\ \mathbf{H}_{m_J} \end{bmatrix}, \text{ where } K_m = \sum_{j=1}^{J} K_{m_j}. \tag{3}$$

Following Seichepine *et al.*'s work [23], our central idea is to couple $\mathbf{H}_m$ with the audio activations, *i.e.*, to factorize $\mathbf{V}_a$ such that $\mathbf{H}_a$ is "similar" to $\mathbf{H}_m$. With such a constraint, the audio activations for each source $\mathbf{H}_{a_j}$ would automatically be coupled with their counterparts in the motion modality $\mathbf{H}_{m_j}$ and we would obtain basis vectors clustered into audio sources. For this purpose, we propose to solve the following optimization problem with respect to $\mathbf{W}_a, \mathbf{H}_a$ and $\mathbf{S}$:

$$\underset{\mathbf{W}_a,\mathbf{H}_a,\mathbf{S}}{\text{minimize}} \Big[D_{KL}(\mathbf{V}_a|\mathbf{W}_a\mathbf{H}_a) + \alpha\|\Lambda_a\mathbf{H}_a - \mathbf{S}\mathbf{H}_m\|_1 + \beta\sum_{k=1}^{K}\sum_{n=2}^{N}(h_{a,kn} - h_{a,k(n-1)})^2 \Big] \tag{4}$$

subject to $\mathbf{W}_a \geq 0, \mathbf{H}_a \geq 0$.

In equation (4), the first term is the standard generalized Kullback-Leibler (KL) divergence cost function such that $D_{KL}(x|y) = x\log(x/y) - x + y$. The second term enforces "similarity" between audio and motion activations, up to a scaling diagonal matrix $\mathbf{S}$, by penalizing their difference with the ℓ_1 norm. The last term is introduced to ensure ℓ_2 temporal smoothness of the audio activations. The influence of each of the last two terms on the overall cost function is controlled by the hyperparameters α and β, repectively. Λ_a is a diagonal matrix with k^{th} diagonal coefficient $\lambda_{a,k} = \sum_f w_{a,fk}$.

The cost function is minimized using a block coordinate majorization-minimization (MM) algorithm [23] where $\mathbf{W}_a$ and $\mathbf{H}_a$ are updated sequentially. Our formulation is a simplified variant of the previously proposed soft non-negative matrix co-factorization (sNMcF) algorithm [23], wherein two modalities are factorized jointly with a penalty term soft-coupling their activations. However, here we do not factorize the second modality (*i.e.*, the motion modality) and its activations are held constant in the update procedure. Note that, from the model's perspective, $\mathbf{H}_a$ and $\mathbf{H}_m$ need not contain the same number of components. So if $K \neq K_m$, then we can readily ignore some components when coupling. However, for this work we maintain $K = K_m$. The reader is referred to [23] for details about the algorithm. Reconstruction is done by performing pointwise multiplication between soft mask, $\mathbf{F}_j = (\mathbf{W}_{a_j}\mathbf{H}_{a_j})./(\mathbf{W}_a\mathbf{H}_a)$ and mixture STFT and finally taking its inverse. Here $\mathbf{W}_{a_j}$ and $\mathbf{H}_{a_j}$ represent the estimated spectral patterns and activations corresponding to the j^{th} source, respectively.

In the following section, we will discuss the procedure for obtaining motion activation matrices $\mathbf{H}_{m_j}$ for each source.

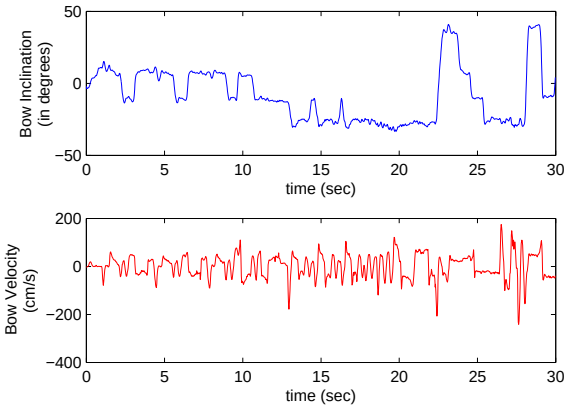


Fig. 1: An example of bow inclination and velocity data for violin.

2.2. Motion Modality Representation

While for audio, the classic magnitude spectrogram representation is used, motion information must be processed to obtain a representation that can be coupled with audio activations. The question now being: What motion features will be useful?

We work with a multimodal dataset of instruments in string quartet performance recordings. Thus, the motion information exists in the form of tracking data (motion capture or MoCap data) acquired by sensors placed on each instrument and the bow [24]. Now we immediately recognize that information about "where" and "how" strongly the sound-producing object is excited will be readily conveyed by bowing motion velocity and orientation in time. In this light, we choose to use bow inclination (in degrees) and bow velocity (cm/s) as features (as shown in Fig. 1), which can be easily computed from the raw motion capture data described in [24, 25]. These descriptors have been pre-computed and provided with the dataset. The bow inclination is defined as the angle between the instrument plane and the bow. The bow velocity is the time derivative of the bow transversal position. The motion activation matrix, $\mathbf{H}_{m_j}$ for $j \in (1, J)$ can then be built using the following simple strategy:

1. In the first step, we quantize the bow inclination for each instrument into 4 bins based on the maximum and minimum inclination values. A binary encoded matrix of size $4 \times N$ is then created where the row corresponding to the active bin is set to 1 and the rest to 0 for each frame.
2. With such a simple descriptor we already have information about the active string within each time window. We then do a pointwise multiplication of each component with the absolute value of the bow velocity. Intuitively, this gives us information about string excitation. Fig. 2 visualizes the effectiveness of this step, where Fig. 2a depicts the quantized bow inclination vector components, overlapped for two sources. Notice, especially in the third subplot, that there are several places where the components overlap and the contrast between the motion of these sources is difficult to see. However, once it is multiplied with the bow velocity (in Fig. 2b) the differences are much more visible.

3. EXPERIMENTAL VALIDATION

We conduct several tests over a set of challenging mixtures to judge the performance of the proposed approach.

3.1. Dataset

We use the publicly available Ensemble Expressive Performance (EEP) dataset[1] [26]. This dataset contains 23 multimodal recordings of string quartet performances (including both ensemble and solo). These recordings are divided into 5 excerpts from Beethoven's Concerto N.4, Op. 18. Four of these, labeled from P1 to P4 contain solo performances, where each instrument plays its own part in the piece. We use these solo recordings to create mixtures for source separation. Note that due to unavailability of microphone recording for the solo performance of the second violin in the quartet we consider mixtures of three sources, namely: Violin (vln), Viola (vla) and Cello (cel). The acquired multimodal data consists of audio tracks and motion capture for each musician's instrument performance.

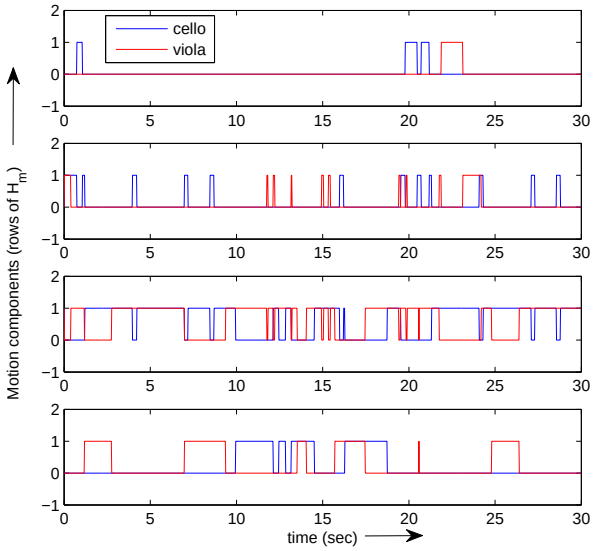
3.2. Experimental Setup

For evaluating the performance of the proposed methods in different scenarios we consider the following three different mixture sets:

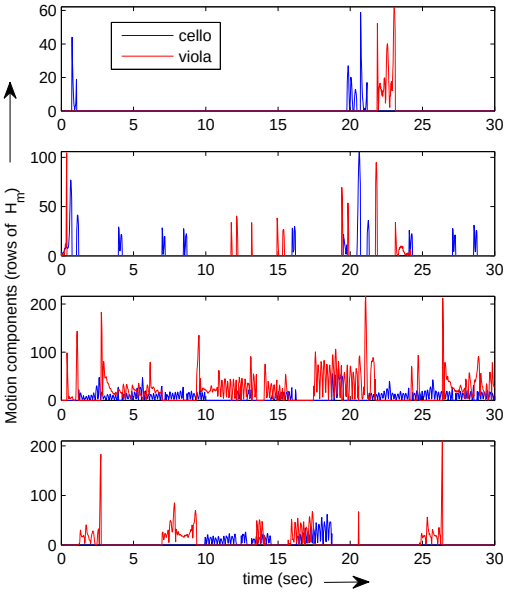
1. **Set 1** - 4 trios of violin, viola and cello, one for each piece denoted by P1, P2, P3, P4 in Table 1.
2. **Set 2** - 6 two-source combinations of the three instruments for pieces P1 - P2.
3. **Set 3** - 3 two-source combinations of the same instrument from different pieces, *e.g.*, a mix of 2 violins from P1 and P2.

Our approach is compared with the following baseline and state-of-the-art methods:

1. **Mel NMF** [6] – 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 of this baseline method.[2]



(a) Quantized bow inclination.



(b) Quantized components multiplied with bow velocity.

Fig. 2: Motion representation.

2. **MM Initialization** [21] – This is a multimodal method where the audio activation matrix is initialized with the motion activation matrix during the NMF parameter estimation.
3. **MM Clustering** [21] – Here, after performing NMF on audio, basis vectors are clustered based on the similarity between motion and audio activations. For details the reader is referred to [21].

Note that, for the latter two methods, as done by the authors, we utilize the Itakura-Saito (IS) divergence cost function. Code pro-

[1] http://mtg.upf.edu/download/datasets/eep-dataset
[2] http://www.ient.rwth-aachen.de/cms/dafx09/

	Mixtures	Proposed Method			MM Initialization			MM Clustering			Mel NMF		
		SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
Set 1	P1	**2.78**	6.06	6.60	-2.00	1.06	3.75	-7.25	-0.77	8.89	-1.15	1.48	5.45
	P2	-0.37	1.81	6.17	-1.79	1.87	3.25	-7.37	-1.30	9.31	**0.56**	3.55	6.56
	P3	**0.97**	3.85	5.81	-0.36	3.86	3.35	-6.45	-0.24	8.67	-2.64	0.32	4.80
	P4	**2.01**	4.79	6.52	-0.37	4.33	3.05	-6.86	-1.03	11.51	0.59	3.94	5.67
Set 2	P1 - vln + vla	**4.25**	6.90	8.48	0.55	3.25	5.89	0.22	4.40	8.82	0.44	2.67	7.74
	P1 - vln + cel	**7.22**	10.19	11.16	3.25	6.62	6.80	-3.99	1.77	18.77	3.56	7.30	8.40
	P1 - vla + cel	2.56	5.81	7.27	-1.17	0.97	5.53	-3.15	2.91	17.06	**2.87**	10.00	7.55
	P2 - vln + vla	0.12	1.75	7.40	-2.32	0.55	3.85	-1.01	4.03	12.59	**3.11**	6.32	8.39
	P2 - vln + cel	**5.97**	9.10	10.20	4.98	9.95	7.16	-3.67	3.64	24.26	4.55	9.79	9.79
	P2 - vla + cel	3.12	5.87	8.15	4.74	8.50	8.07	-3.52	3.08	17.49	**4.94**	9.23	8.49
Set 3	vln(P1) + vln(P2)	**3.57**	5.85	8.54	0.47	3.61	5.11	1.30	3.44	9.09	0.84	1.96	9.76
	vla(P1) + vla(P2)	**-0.35**	1.16	7.44	-1.37	0.43	6.13	-4.45	0.61	16.67	-1.71	0.82	4.73
	cel(P1) + cel(P2)	**3.66**	5.94	8.62	2.07	5.79	5.86	-4.60	2.32	24.10	-0.42	2.22	6.08

Table 1: SDR, SIR and SAR (measured in dB) for different methods on each mixture. Best SDR is displayed in bold.

vided by Févotte *et al.* [27] is used for standard NMF algorithms.

The audio is sampled at 44.1 kHz. We compute the spectrogram with a Hamming window of size 4096 (92 ms) and 75% overlap for each 30 sec excerpt. Thus, we have a $2049 \times N$ matrix. Here N is the number of STFT frames. Since the MoCap data is sampled at 240 Hz, each of the selected descriptors is resampled to match the N STFT audio frames. For all the runs the proposed method hyperparameters were set at $\alpha = 10$ and $\beta = 0.3$ after preliminary testing. As discussed in section 2.2, the number of components for each instrument is set to 4. NMF for each of the methods is run for 100 iterations. For each mixture, all the methods are run 5 times and the reconstruction is performed using a soft mask. The average of each evaluation metric over these runs is displayed in Table 1.

Evaluation metrics: the Signal to Distortion Ratio (SDR), the Signal to Interference Ratio (SIR) and the Signal to Artifacts Ratio (SAR) are computed using the BSS_EVAL Toolbox version 3.0 [28]. All the metrics are expressed in dB.

3.3. Results and Discussion

The results are as presented in Table 1, where the best SDR for each mixture is displayed in bold. Our method clearly outperforms the baselines and the state-of-the-art methods for highly challenging cases of trios (Set 1) and duos involving the same instrument (Set 3). For the third set of mixtures, audio only methods would not be able to cluster the spectral patterns well. Motion information clearly plays a crucial role for disambiguation and indeed the proposed method outperforms all the others by a large margin.

Particularly, notice that the multimodal baselines do not perform well. The MM initialization relies on setting to zero the coefficients where there is no motion. This might not prove to be the best strategy with such a dataset because even during the inactive period of the audio there is some motion of the hand. On the other hand, multimodal clustering depends on the similarity between source motion activation centroids and audio activations. As we observe during the experiments, such a similarity is not very obvious for the data we use and the method ends up assigning most vectors to a particular cluster.

Despite its overall good performance it is worth noting that for trio mixtures the proposed method performs poorly with P2. In fact, all the mixtures involving the viola from the second piece seem to have worse performance than others. We note that the separation for the viola suffers. One possible reason for this could be that, for P2, the motion descriptors of the viola with respect to the violin and the cello overlap in parts. As a consequence, the estimation of $\mathbf{W}_a$ for such cases is poor.

We must emphasize that the optimal value for α, which is held constant here, would differ for each recording. Thus, it should be possible to tune that parameter to gain the best performance, as could be achieved by an audio engineer through a knob controlling α, in a real world audio production setting. As an illustration, consider the mixture of viola and cello from P2: if we search for the best α in the mean SDR sense within the range $(1, 15)$, we find that mean SDR value of up to 5.97 dB can be reached at $\alpha = 1.5$. Also, note that we work with a limited number of components which is probably not well suited for some of these cases.

4. CONCLUSION

We have demonstrated the usefulness of exploiting sound-producing motion for guiding audio source separation. Formulating it as a soft constraint within the NMF source separation framework makes our approach very flexible and simple to use. We alleviate the shortcomings of previous works, such as multiple parameter tuning while making no unrealistic assumptions about the audio environment. The results obtained on the multimodal string instrument dataset are very encouraging and serve as a proof-of-concept for applying the method to separate any audio object accompanied with its sound-producing motion. The use of motion capture data is new and the proposed technique would apply to video data in a similar manner.

As part of ongoing work, we are investigating automatic extraction of motion activation matrix and ways to accommodate different number of basis components in both the modalities.

5. REFERENCES

[1] Jinji Chen, Toshiharu Mukai, Yoshinori Takeuchi, Tetsuya Matsumoto, Hiroaki Kudo, Tsuyoshi Yamamura, and Noboru Ohnishi, "Relating audio-visual events caused by multiple movements: in the case of entire object movement," in *Proc. fifth IEEE Int. Conf. on Information Fusion*, 2002, vol. 1, pp. 213–219.

[2] Beiming Wang and Mark 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.

[3] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis, "Deep learning for monaural speech separation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 1562–1566.

[4] Jean-Louis Durrieu, Bertrand David, and Gaël Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1180–1191, 2011.

[5] Olivier Gillet and Gaël Richard, "Transcription and separation of drum signals from polyphonic music," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 3, pp. 529–540, 2008.

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

[7] Rajesh Jaiswal, Derry FitzGerald, Dan Barry, Eugene Coyle, and Scott 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.

[8] Xin Guo, Stefan Uhlich, and Yuki 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.

[9] Luc Le Magoarou, Alexey Ozerov, and Ngoc 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.

[10] Joachim Fritsch and Mark D. Plumbley, "Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2013, pp. 888–891.

[11] Paris Smaragdis and Gautham J. Mysore, "Separation by humming: user-guided sound extraction from monophonic mixtures," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2009, pp. 69–72.

[12] Ngoc Q. K. Duong, Alexey Ozerov, Louis Chevallier, and Joël 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.

[13] Antoine Liutkus, Jean-Louis Durrieu, Laurent Daudet, and Gaël Richard, "An overview of informed audio source separation," in *14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, July 2013, pp. 1–4.

[14] John W. Fisher III, Trevor Darrell, William T. Freeman, and Paul Viola, "Learning Joint Statistical Models for Audio-Visual Fusion and Segregation," in *Advances in Neural Information Processing Systems*, 2001, number Ml, pp. 772–778.

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

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

[17] Anna L. Casanovas, Gianluca Monaci, Pierre Vandergheynst, and Rémi Gribonval, "Blind audiovisual source separation based on sparse redundant representations," *Multimedia, IEEE Transactions on*, vol. 12, no. 5, pp. 358–371, Aug 2010.

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

[19] Kazuhiro Nakadai, Ken-ichi Hidai, Hiroshi G Okuno, and Hiroaki Kitano, "Real-time speaker localization and speech separation by audio-visual integration," in *Proc. IEEE Int. Conf. on Robotics and Automation*, 2002, vol. 1, pp. 1043–1049.

[20] Bertrand Rivet, Laurent Girin, and Christian Jutten, "Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 96–108, 2007.

[21] Farnaz Sedighin, Massoud Babaie-Zadeh, Bertrand Rivet, and Christian Jutten, "Two multimodal approaches for single microphone source separation," in *EUSIPCO*, 2016.

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

[23] Nicolas Seichepine, Slim Essid, Cédric Févotte, and Olivier Cappé, "Soft nonnegative matrix co-factorization," *IEEE Transactions on Signal Processing*, vol. 62, no. 22, pp. 5940–5949, 2014.

[24] Marco Marchini, *Analysis of Ensemble Expressive Performance in String Quartets: a Statistical and Machine Learning Approach*, Phd thesis, Univesitat Pompeu Fabra, 2014.

[25] Esteban Maestre, *Modeling instrumental gestures: an analysis/synthesis framework for violin bowing*, Phd thesis, Universitat Pompeu Fabra, 2009.

[26] Marco Marchini, Rafael Ramirez, Panos Papiotis, and Esteban 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.

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

[28] 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.