GAUSSIAN MIXTURE MODELS FOR SCORE-INFORMED INSTRUMENT SEPARATION

Pablo Sprechmann,[1] *Pablo Cancela,*[2] *and Guillermo Sapiro*[1]

[1]University of Minnesota, USA; [2]Universidad de la República, Uruguay

ABSTRACT

A new framework for representing quasi-harmonic signals, and its application to score-informed single channel musical instruments separation, is introduced in this paper. In the proposed approach, the signal's pitch and spectral envelope are modeled separately. The model combines parametric filters enforcing an harmonic structure in the representation, with Gaussian modeling for representing the spectral envelope. The estimation of the signal's model is cast as an inverse problem efficiently solved via a maximum a posteriori expectation-maximization algorithm. The relation of the proposed framework with common non-negative factorization methods is also discussed. The algorithm is evaluated with both real and synthetic instruments mixtures, and comparisons with recently proposed techniques are presented.

Index Terms— Score-informed source separation, single channel source separation, audio modeling

1. INTRODUCTION

Single channel source separation (SCSS) is a classical problem in audio processing that arises naturally when dealing with musical signals. In this case, the main goal is to separate the different tracks corresponding to isolated musical instruments. Although important advances have been obtained throughout the years, this is still considered an open and difficult problem, in part due to its high degree of under-determination. It is then crucial to use all the available information to constraint instruments' separation in a meaningful way. Since musical scores are easily available and provide fundamental information about the musical piece, in this work we tackle the problem of *score-informed* SCSS in musical pieces [1, 2, 3]. The information extracted from the scores is used as prior information to initialize and guide our algorithm.

The decomposition of time-frequency representations, such as the power spectrogram, in terms of elementary atoms of a dictionary has become a popular tool in audio processing. In particular, non-negative matrix factorization (NMF), [4], leads to very good results in a variety of applications. SCSS via NMF is carried out by decomposing the magnitude spectrogram of the mixture signal and then performing reconstructions of groups corresponding to each single source. In the fully unsupervised setting, these methods brake down when the sources have a large time-overlap in the track. A considerable amount of work has been dedicated to add constraints to the factorization in order to include prior information guiding the challenging decomposition.

When NMF is applied to quasi-harmonic instrument sounds, the elementary components that are redundant throughout the piece will hopefully represent musical notes and thus have an harmonic structure. However, this cannot be guaranteed. Recent methods have proposed constraining the atoms to have particular designs following prior information about the signal in order to obtain physically meaningful atoms, and in particular, to mimic the harmonic structure present in the spectrograms of musical instruments [3, 5, 6, 7].

The variability in the spectrum of a musical instrument sound has two main components: the variation of the fundamental frequency and the changes in its spectral envelope. Standard NMF has shown good results when the characteristics of the sounds are stationary. In other words, NMF relies in the frame-to-frame redundancy. Slight changes in the fundamental frequency with constant spectral envelope produce severe changes in the spectrogram. The same happens when changes in the spectral envelope occur with a fixed pitch. Classical NMF will likely need several dictionary atoms to account for this variability, while the nature of these changes is rather simple.

In this paper we propose a simple model for representing the magnitude spectrogram of musical instruments that decouples between the information of pitch and spectral envelope. This allows to efficiently represent a great deal of variability using very simple models for each component. Specifically, let $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_N] \in \mathbb{R}^{F \times N}$ be the power spectrogram of a signal containing, for now, an isolated instrument. We can decompose this as

$$\mathbf{V} \approx \mathbf{H} \bullet \mathbf{E}, \tag{1}$$

where $\mathbf{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_N] \in \mathbb{R}^{F \times N}$ is a non-negative matrix modeling the spectral envelope and its evolution in time, and $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_N] \in \mathbb{R}^{F \times N}$ is a matrix with entries in the $[0, 1]$ interval enforcing the harmonic structure through an element-wise multiplication $\bullet$. With this representation, changes in pitch are captured in $\mathbf{H}$, while changes in the timbre appear in $\mathbf{E}$. The space of the spectral envelope is simpler and can be accurately represented via Gaussian modeling.

The proposed method is particularly well suited for score-informed SCSS, since the pitch of each source is (approximately) known beforehand, having a very good guess of the shape of $\mathbf{H}$ in (1). Given $\mathbf{H}$, finding $\mathbf{E}$ can be cast as an inverse problem. This formulation is inspired in part by the excellent results reported by [8] for a number of inverse problems in image processing.

Following [8], the Gaussian parameters and the signal representation are simultaneously estimated via an efficient MAP-EM (maximum a posteriori expectation-maximization) algorithm. Once the decomposition is obtained, we can construct a time-frequency mask source, recovering each source from the mixture by Wiener filtering.

In [3] the authors proposed a method based on NMF that also uses parametric templates to represent the harmonic structure of notes. Timbre is represented by the estimation of the amplitude of partials. The main difference with the proposed approach is that, for each instrument, all the notes are assumed to share the same fixed relative harmonics amplitudes. The Gaussian modeling allows a more flexible representation capturing the frequency-dependent resonance of the instruments and their variability. [1]

PS and PC contributed equally to this work.

[1]After this paper was submitted we became aware of a related work [7].

In Section 2 we present the proposed audio (instruments) signal modeling and describe the algorithm for performing score-informed SCSS. Section 3 discusses connections with PCA and factorization methods. In Section 4 we evaluate the method with synthetic and real data. In Section 5 we conclude the paper.

2. MODELING FRAMEWORK

We propose a model for mixtures of harmonic or quasi-harmonic instruments. The proposed framework is designed for solving the score-informed separation, in Section 5 we discuss possible extensions to the fully unsupervised case and speakers separation.

2.1. Single Signal Model

We assume that each source is stationary within a frame. This means that for the i-th frame, the quasi-harmonic part of the signal can be considered as harmonic with a fundamental frequency f_i. In the Fourier domain, most of the energy of $\mathbf{v_i}$ is concentrated in bins corresponding to frequencies of the form kf_i, with $k \in \mathbb{Z}$. Rewriting (1) in a frame basis we obtain

$$\mathbf{v}_i = \mathbf{h}_i \bullet \mathbf{e}_i + \mathbf{w}_i, \quad \text{for} \quad i = 1, \ldots, N,$$

where $\mathbf{h}_i$ is the power spectrum of a linear filter that enforces an harmonic constraint on the representation, $\mathbf{e}_i$ is the envelope of the spectral content, and $\mathbf{w}_i$ is a representation error. We consider $\mathbf{h}_i$ to be the spectral response of a comb filter with unit amplitude and parametrized by its fundamental frequency, $\mathbf{h}_i = \mathbf{h}(f_i)$. The point wise multiplication $\mathbf{h}_i \bullet \mathbf{e}_i$ corresponds to the filtering of $\mathbf{e}_i$. In every frame, $\mathbf{e}_i$ is assumed to be drawn from a Gaussian distribution with mean μ and covariance Σ, to be learned. The representation error is assumed to be also Gaussian with zero mean and known or estimated signal-independent isotropic covariance $\sigma^2 I_d$. See Figure 1 for an illustration of this model.

The available score provides a set of possible values for the fundamental frequency, $\mathbf{f}_0 = [f_{01}, \ldots, f_{0N}]$, which are a very good approximation of the true values of $\mathbf{f} = [f_1, \ldots, f_N]$. However, they cannot be assumed to be identical. A canonical example of such a situation is the vibrato, where the fundamental frequency slightly oscillates around a specific note. To deal with this variability we model the fundamental frequency as a time changing (real valued) Gaussian distribution centered at the fundamental frequency given by the score and with variance σ_0^2.[2]

2.2. Mixed Signal Model

Let's now assume that the signal is a mixture of c quasi-harmonic instruments. We want to decompose its power spectrum as

$$\mathbf{V} = \sum_{j=1}^{c} \mathbf{H}_j \bullet \mathbf{E}_j + \mathbf{W}. \tag{2}$$

The pitch and spectral envelope for the j-th instrument are modeled by the matrices $\mathbf{H}_j = [\mathbf{h}_{j1}, \ldots, \mathbf{h}_{jN}]$ and $\mathbf{E}_j = [\mathbf{e}_{j1}, \ldots, \mathbf{e}_{jN}]$ respectively, following the model presented in 2.1. As with NMF, the model estimation and the signal coding is done simultaneously:

- Estimating the Gaussian parameters $\mathcal{G} = \{(\mu_j, \Sigma_j)\}_{1 \le j \le c}$ for each instrument.
- Estimating the set of envelopes, $\{\mathbf{E}_j\}_{1 \le j \le c}$, and the real fundamental frequencies, $\{\mathbf{f}_j\}_{1 \le j \le c}$, from the spectrum $\mathbf{V}$ given, via the score, the set of corresponding fundamental frequencies,$\{\mathbf{f}_{0j}\}_{1 \le j \le c}$, and the Gaussian distributions for each instrument, $\mathcal{G}$.

To solve this non-convex problem we use an adaptation of the efficient MAP-EM algorithm presented for image processing in [8].

We could consider using a GMM to model the spectral content of the instruments instead of a single Gaussian distribution. This could help for example when modeling instruments with large timbre variability, i.e., plucked string sounds or singing voice. This is a natural and relatively simple extension of the framework here proposed and is part of our speaker modeling extension to be reported elsewhere.

2.3. Computational Algorithm

The MAP-EM algorithm is an iterative procedure that alternates between two steps. An *E-step* that estimates the spectral content of the sources assuming that the Gaussian parameters are known, and an *M-step* that reciprocally estimates the Gaussian parameters for each source while assuming that the spectral envelopes are known. To simplify the notation and without loss of generality, the Gaussians are assumed to have zero mean, since they can be centered with respect to the estimated mean. We use the available score to produce a good initial condition for the algorithm, see Section 2.3.3.

2.3.1. E-Step: Signal Estimation

The *E-step* can be performed independently for each frame. We obtain the estimates by solving the MAP,

$$\{\tilde{\mathbf{e}}_{ji}, \tilde{f}_{ji}\}_{1 \le j \le c} = \underset{\mathbf{e}_{ji}, f_{ji}}{\operatorname{argmax}} \log p(\mathbf{e}_{ji}, f_{ji} | \mathbf{v}_i, f_{0ji}, \mathcal{G}). \tag{3}$$

Maximizing the cost function in (3) is equivalent to maximizing

$$\log p(\mathbf{v}_i | \mathbf{e}_{ji}, f_{ji}) + \log p(\mathbf{e}_{ji} | \mathcal{G}) + \log p(f_{ji} | f_{0ji}). \tag{4}$$

The masking filter $\mathbf{h}(f_{ji})$ is a linear operator and can be written as $\mathbf{h}(f_{ji}) \bullet \mathbf{e}_{ji} = \mathbf{U}_{ji}\mathbf{e}_{ji}$, where $\mathbf{U}_{ji} = \operatorname{diag}(\mathbf{h}(f_{ji}))$. Using this notation and substituting with the corresponding Gaussian probability density functions in (4), we can rewrite the problem as,

$$\begin{aligned} \{\tilde{\mathbf{e}}_{ji}, \tilde{f}_{ji}\} \quad = \quad & \underset{\mathbf{e}_{ji}, f_{ji}}{\operatorname{argmin}} \|\mathbf{v}_i - \sum_{r=1}^{c} \mathbf{U}_{ri}\mathbf{e}_{ri}\|^2 + \\ & \sigma^2 \sum_{r=1}^{c} \mathbf{e}_{ri}^T \Sigma_r^{-1} \mathbf{e}_{ri} + \frac{\sigma^2}{\sigma_0^2} \sum_{r=1}^{c} |f_{ri} - f_{0ri}|^2. \end{aligned} \tag{5}$$

For ease of notation, we assume without loss of generality that the Gaussians have zero mean as the $\mathbf{v}_i$'s can be always be centered in zero. This sub-problem is non-convex when minimizing over both $\mathbf{e}_{ji}, f_{ji}$. We solve it by iteratively fixing one and optimizing over the other.

Fixing f_{ji} in (5), the problem is strictly convex and can be solved efficiently via Wiener filtering and in closed form,

$$\tilde{\mathbf{e}}_{ji} = \mathbf{U}_{ji}^T \Sigma_j \left(I_d \sigma^2 + \sum_{r=1}^{c} \mathbf{U}_{ri}^T \Sigma_r \mathbf{U}_{ri} \right)^{-1} \mathbf{v}_i, \tag{6}$$

where I_d is the identity matrix of the appropriate size. The natural initial condition is $f_{ji} = f_{0ji} \forall j$.

Since we are working with a finite resolution representation of the spectrogram, the set of distinguishable fundamental frequencies

[2] In all our experiments we considered σ_0 to be 1% of f_{0i}. However, σ_0^2 might be instrument dependent.

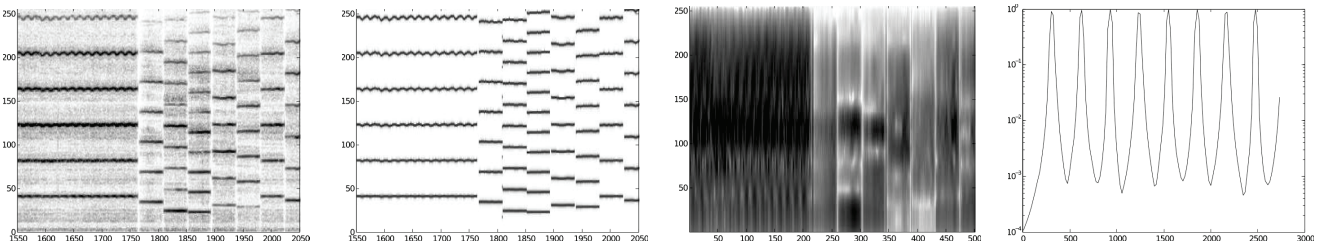

Fig. 1. In this figure we show, respectively, an example of a spectrogram $\mathbf{V}$, the corresponding matrices $\mathbf{H}$ and $\mathbf{E}$, and a comb filter $\mathbf{h}_i$.

is naturally discretized. Thus there is no need to consider the fundamental frequency as real valued for obtaining an accurate representation. Then, solving the problem in (5) with $\mathbf{e}_{ji}$ fixed, reduces to evaluating the cost function in a small number of candidates distributed around the ones provided by the score, $\{f_{0ji}\}_{1\leq j\leq c}$, choosing the one for which the minimum is obtained.

In oder to have a physically meaningful decomposition, the obtained $\{\tilde{\mathbf{e}}_{ji}\}_{1\leq j\leq c}$ in (6) need to be non-negative. When dealing with a large number of sources this might not happen in all cases due to the high degree of under-determination. We then add to (5) a non negativity constraint, $\mathbf{e}_{ji} \geq 0\ \forall j, i$. This constrained optimization is still convex and can be carried out using projected gradient methods [9]. In most of our experiments, and in most of the cases, the solution of the unconstrained problem is already non-negative. This has also been observed in various image processing inverse problems [8] (where the pixel intensity is also non-negative). Thus we use the solution given by (6) as a warm start to the constrained formulation, and in general, only a few (if any) iterations are needed to reach convergence.

2.3.2. M-Step: Model Estimation

After the estimation of the envelopes and the fundamental frequencies is performed, we recalculate the Gaussian parameters for all the instruments. This is done via the empirical mean and covariance,

$$\tilde{\mu}_j = \frac{1}{N}\sum_{i=1}^{N}\mathbf{e}_{ji}\,,\ \tilde{\Sigma}_j = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{e}_{ji}-\tilde{\mu}_j)(\mathbf{e}_{ji}-\tilde{\mu}_j)^T,\ 1\leq j\leq c.$$

Refer to [8] for a discussion on the optimality of this selection (see Section 5 for comments on how to remove this restriction).

2.3.3. Initialization

The initialization process for the MAP-EM goes as follows. First, the complete score gets synthesized producing an isolated track for each source. Then, each of these synthetic tracks is used to learn the Gaussian parameters for each instrument. In all the cases we assume that the signals are perfectly aligned to the available scores. The alignment problem can be done automatically using dynamic time warping, see [2] and references therein.

3. CONNECTIONS WITH PCA AND NMF

In this section we first present an interpretation of the proposed method that links it with structured principal component analysis (PCA). Then we discuss its relations with the NMF.

3.1. Structured Estimation in PCA Bases

Given a set of signals $\{\mathbf{e}_{ji}\}_{1\leq i\leq N}$, the PCA basis are defined as the matrix $\mathbf{B}_j = [\mathbf{b}_{j1},\ldots,\mathbf{b}_{jF}]$ that diagonalizes the corresponding covariance matrix, $\mathbf{\Sigma}_j = \mathbf{B}_j^T\mathbf{S}_j\mathbf{B}_j$, where $\mathbf{S}_j = \text{diag}(\lambda_1^j,\ldots,\lambda_F^j)$ is a diagonal matrix, whose diagonal elements are the sorted eigenvalues $\lambda_1^j \geq \lambda_2^j \geq \ldots \geq \lambda_F^j \geq 0$. The columns of $\mathbf{B}_j$ are orthonormal and represent the principal directions of variation of $\{\mathbf{e}_{ji}\}_{1\leq i\leq N}$. The magnitude of the eigenvalues measure the energy of the variation in the corresponding directions.

Working in the PCA basis rather than the canonical one, $\mathbf{a}_{ji} = \mathbf{B}_j^T\mathbf{e}_{ji}$, allows to significantly reduce the dimensionality of the data in a meaningful way. When representing the timbre of the instruments we can verify that they are highly compressible: the first few eigenvalues of the covariance matrices concentrate most of the total energy. We can then write our model stated in (2) as

$$\mathbf{V} = \sum_{j=1}^{c}\mathbf{H}_j \bullet \mathbf{B}_j\mathbf{A}_j + \mathbf{W} \approx \sum_{j=1}^{c}\mathbf{H}_j \bullet \hat{\mathbf{B}}_j\mathbf{A}_j + \mathbf{W}, \quad (7)$$

where the matrices $\hat{\mathbf{B}}_j = [\mathbf{b}_{j1},\ldots,\mathbf{b}_{jk}]$ conserve only the first $k \ll F$ principal directions. The MAP estimate (4) can be equivalently computed as (see also [8])

$$\{\tilde{\mathbf{a}}_{ji},\tilde{f}_{ji}\}_{1\leq j\leq c} = \underset{\mathbf{a}_{ji},f_{ji}}{\text{argmin}}\|\mathbf{v}_i - \sum_{r=1}^{c}\mathbf{U}_{ri}\hat{\mathbf{B}}_j\mathbf{a}_{ri}\|^2 + \sigma^2\sum_{r=1}^{c}\sum_{p=1}^{k}\frac{|a_{ri}[p]|^2}{\lambda_r^p} + \frac{\sigma^2}{\sigma_0^2}\sum_{r=1}^{c}|f_{ri}-f_{0ri}|^2. \quad (8)$$

Inside each PCA basis the atoms are pre-ordered by their corresponding eigenvalues. The weighting term in (8) privileges the coefficients corresponding to the principal directions with larger energy. This representation stabilizes the decomposition, which is crucial in these type of source separation ill-posed problems. We used $k = 25$ in all the experiments.

3.2. Relations with NMF

In standard NMF, the spectrogram of the mixture signal is decompose as the product of two non-negative matrices, $\mathbf{V} \approx \mathbf{WH}$, where $\mathbf{W} \in \mathbb{R}^{F\times Q}$, $\mathbf{H} \in \mathbb{R}^{Q\times N}$, and $Q \ll F, N$. The matrix $\mathbf{W}$ is the dictionary and each column represents an atom. The matrix $\mathbf{H}$ codes the activation of each atom in the dictionary throughout the frames. This representation is a low rank linear approximation of $\mathbf{V}$. Small variations in the pitch as the ones encountered in a vibrato for example, significantly increase the rank of $\mathbf{V}$, forcing to increase the number of atoms, Q. In SCSS this is highly undesirable, the models need to be as stable as possible in order to make sure that we can properly identify and reconstruct the multiple sources. Many solutions have been proposed to address this issue [7, 10].

In contrast to NMF, our proposed model is non-linear. We propose to use a linear model only to represent the timbre of each instrument, while using the set of parametric filters to model the harmonic

	PLCA train M1	PLCA train M2	PDA	sPCA train M1	sPCA train M2	Oracle
SIR	22.8	14.2	20.2	18.1	15.9	20.5
SAR	11.5	6.3	7.7	10.9	9.8	13.6
SDR	11.1	4.8	7.2	10.0	8.7	12.7

Table 1. Results with synthesis method M1 as testing signals.

	PLCA train M1	PLCA train M2	PDA	sPCA train M1	sPCA train M2	Oracle
SIR	12.4	20.1	12.6	14.9	15.8	19.6
SAR	4.5	10.6	3.3	6.8	7.7	12.3
SDR	3.1	10.1	2.1	5.9	6.8	11.5

Table 2. Results with synthesis method M2 as testing signals.

constraint, as shown in (7). The proposed model is clearly less general than NMF since it is restricted to quasi-harmonic signals.

4. NUMERICAL EXPERIMENTS

In this section we evaluate the performance of the proposed method using real and synthetic data. The performance of the instruments separation methods is evaluated in terms of the standard measures: Signal to Interference Ration (SIR), Signal to Artifact Ratio (SAR) and Signal to Distortion Ration [11].[3] All given ratios are averaged over all tested signals. Audio examples are available at `http://www.tc.umn.edu/~sprec009/icassp2012.html`

4.1. Synthetic Data

We used the publicly available database described in [3].[4] It contains 12 different string quartets (two violins, viola, and cello) by Bach, Beethoven and Boccherini. These pieces render important characteristics of real musical streams such as the overlap of the sources in the time-frequency domain. For each piece, the database provides a MIDI file containing the scores of the first 30 seconds of each piece, and two synthesized wave files for every individual instrument. The mixture signals are obtained by summing the individual tracks. The wave files are synthesized with one of two different method. We will refer to these methods as $M1$ and $M2$ (see [3] for a detailed description). The signals synthesized with the different methods present distinct characteristics. For instance, they have different vibratos and decay times. Also, method $M2$ includes a reverberation effect present in the sound-font, while $M1$ does not.

Tables 1 and 2 compare our (sPCA) against the ones produced by [3] and a method based on PLCA [2] (refered to as PDAand PLCA respectively).[5] We report the results obtained by the proposed algorithm and PLCA using both the testing and training signals to train the models. The proposed algorithm is less sensitive to the initialization than the PLCA-based, therefore has better generalization properties. We also report the results obtained using the true isolated tracks to generate the Wiener masks (refered as Oracle method).

4.2. Real Data

We used the *Development Set for MIREX 2007 MultiF0 Estimation Tracking Task*.[6] It contains a 52 second long musical piece played by different wind instruments. The separated tracks for each instrument are available and the mixture is done by summing them. Table 3 shows the performance ratio obtained for a mixture of 4 instruments (clarinet, flute, horn and oboe) with different training conditions. The obtained results show that each signal is well captured.

[3] We used the BSS_EVAL toolbox [11].

[4] Available at `http://perso.enst.fr/hennequi/database.zip`.

[5] The results for PDAand PLCA were copied from [3].

[6] `http://www.music-ir.org/mirex/wiki/`

	sPCA train SD	sPCA train RD	Oracle
SIR	17.4	17.5	18.2
SAR	11.1	11.2	13.2
SDR	10.1	10.6	11.7

Table 3. Results with MIREX 2007 MultiF0 database. Training using synthetic data (SD), real data from the database (RD), and (Oracle) as in 4.1.

5. CONCLUSIONS

In this paper, we introduced a new framework for representing quasi-harmonic audio signals that can be used to address audio source separation problems, and in particular score-informed source separation of musical mixtures. The method has the ability to model the pitch and envelope of a sound source independently. This permits the representation of signals with combinations of pitches and envelopes not previously observed. Moreover, it allows to easily incorporate meta-data such as the musical score indicating the signal to be represented. In this way, two or more musical instruments with the same characteristics can be separated. The characterized set of signals can be extended to incorporate non-harmonic sounds by defining filters $\mathbf{h_i}$ with the appropriate masking of spectral components. The method has been evaluated in the score-informed SCSS problem with synthetic and real data showing a performance comparable with those reported in the literature. Preliminary experiments, to be reported elsewhere, indicate that this basic model, with an GMM and MRF addition, is very efficient for the separation of speakers as well.

Acknowledgments. Work supported by CSIC, ONR, NGA, ARO, DARPA, and NSSEFF. We thank Dr. Guoshen Yu for very valuable discussions.

6. REFERENCES

[1] M.R. Every and J.E. Szymanski, "A spectral-filtering approach to music signal separation," in *DAFx*, 2004.

[2] J. Ganseman, P. Scheunders, G. J. Mysore, and J. S. Abel, "Evaluation of a score-informed source separation system," in *ISMIR*, 2010, pp. 219–224.

[3] R. Hennequin, B. David, and R. Badeau, "Score informed audio source separation using a parametric model of non-negative spectrogram," in *ICASSP*, May 2011.

[4] D.D. Lee and H.S. Seung, "Learning parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[5] S. K. Tjoa, M. C. Stamm, W. Sabrina Lin, and K. J. Ray Liu, "Harmonic variable-size dictionary learning for music source separation," in *ICASSP*, Dallas, TX, Mar. 2010, pp. 413–416.

[6] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription," *IEEE Trans. Audio, Speech & Lang. Process.*, vol. 18, pp. 538–549, 2010.

[7] J. Durrieu, B. David, and G. Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation," *IEEE Journal of Sel. Topics in Signal Process.*, vol. 5, no. 6, pp. 1180 –1191, oct. 2011.

[8] G. Yu, G. Sapiro, and S. Mallat, "Solving inverse problems with piecewise linear estimators: From gaussian mixture models to structured sparsity," *CoRR*, vol. abs/1006.3056, 2010, http://arxiv.org/abs/1006.3056.

[9] Y. Nesterov, "Gradient methods for minimizing composite objective function," CORE Discussion Papers 2007076, 2007.

[10] R. Jenatton, J.-Y. Audibert, and F. Bach, "Structured variable selection with sparsity-inducing norms," Tech. Rep., arXiv:0904.3523, 2009.

[11] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech & Lang. Process.*, vol. 14, pp. 1462–1469, 2006.