SCALE-INVARIANT PROBABILISTIC LATENT COMPONENT ANALYSIS

Romain Hennequin, Roland Badeau and Bertrand David

Institut Telecom, Telecom ParisTech, CNRS LTCI
37-39 rue Dareau, 75014 Paris, France
<forename>.<surname>@telecom-paristech.fr

ABSTRACT

In this paper, we present a new method for decomposing musical spectrograms. This method is similar to shift-invariant Probabilistic Latent Component Analysis, but, when the latter works with constant Q spectrograms (i.e. with a logarithmic frequency resolution), our technique is designed to decompose standard short time Fourier transform spectrograms (i.e. with a linear frequency resolution). This makes it possible to easily reconstruct the latent signals (which can be useful for source separation).

Index Terms— Music signal processing, non-negative matrix factorization, probabilistic latent component analysis, shift-invariant decomposition.

1. INTRODUCTION

Non-negative decompositions are widely used for audio spectrograms processing: non-negative matrix factorization (NMF) [1] and PLCA [2] are both used to decompose spectrograms with applications such as source separation [3, 4] and automatic music transcription [5]. Shift-invariant models [6, 7] decompose constant-Q spectrograms [8] with a single frequency template for each harmonic instrument: with a log-frequency resolution, a frequency shift corresponds to a transposition. Then each note of a single instrument can be modeled as a base template shifted to the right pitch.

Unfortunately, constant-Q transforms (CQT) are difficult to use for sound source separation: even if a near-perfect inverse CQT transform was recently proposed [9], the variable resolution of the decomposition makes it difficult to apply time-frequency masking and source separation using CQT is still an open problem. Attempts were made to use shift-invariant decomposition of CQT for source separation [10] using a mapping between log-frequency and linear frequency resolution to avoid CQT inversion. This paper presents a new decomposition method inspired by shift-invariant decompositions, but which is designed to directly decompose STFT spectrograms. In a constant-Q spectrogram, a change of fundamental frequency approximately corresponds to a translation of the spectral template; in a standard STFT spectrogram, one can model such a change with an homothety of the spectral template. This approximation remains only valid for small modifications since:

- For real acoustic instruments, a note change usually leads to a different energy balance between the different harmonics. Even if it is widely used in Shift-Invariant decomposition, the assumption that this balance does not change is only exact for a few electronic instruments.
- The spectral width of harmonics (or partials) is affected by the homothety, in opposition with the real world case where it is mostly related to the analysis window type and length.

We call the new decomposition *scale-invariant PLCA* (SIPLCA). This scale-invariant model presents some new issues that were not encountered with shift-invariant models: an homothety on a set of integers does not yield integers. We propose a solution to this issue. In section 2, we remind the principle of standard PLCA and shift-invariant PLCA. We then present, in section 3, the new scale-invariant model and derive an algorithm to estimate the parameters. Some examples are presented in section 4 and we propose an application of single notes repitching in a polyphonic signal. Finally, we draw conclusions in section 5.

2. PROBABILISTIC LATENT COMPONENT ANALYSIS

The model that we used is inspired by shift-invariant PLCA [7] which is a probabilistic drawing model. In PLCA decompositions [2], a non-negative spectrogram $\mathbf{V}_{ft}$ is considered as an histogram obtained from a structured draw of a frequency random variable $f \in \{1, 2, \ldots, F\}$ and a time random variable $t \in \{1, 2, \ldots, T\}$ which follow the joint distribution $P(f,t)$. One can design $P(f,t)$ in different ways, which lead to different decompositions.

2.1. Standard PLCA

Standard PLCA (non shift-invariant) [2] leads to a decomposition similar to NMF. The draw is structured with a latent (hidden) random variable z which corresponds to a “component”, assuming that f and t are independent given z:

$$P(f,t) = \sum_{z=1}^{Z} P(z)P(f,t|z) = \sum_{z=1}^{Z} P(z)P(f|z)P(t|z).$$

The histogram $\mathbf{V}_{ft}$ is thus assumed to be obtained in the following way: first z is drawn following $P(z)$ and then f and t are drawn following respectively $P(f|z)$ and $P(t|z)$.

In an NMF framework, $P(f|z)$ corresponds to the spectral templates and $P(t|z)$ corresponds to the activations of each component. $P(z)$ is the relative weight of each component.

2.2. Shift-invariant PLCA

Shift-invariant PLCA introduces another latent random variable $f' \in \{1, 2, \ldots, F'\}$, the base frequency, in reference of which a

The research leading to this paper was supported by the CONTINT program of the French National Research Agency (ANR), as a part of the DReaM project (ANR-09-CORD-006-03), and by the Quaero Programme, funded by OSEO, French State agency for innovation.

transposition variable $\tau \in \mathbb{Z}$ is defined. f' and t are assumed independent given z, and τ and f' are also independent given z (but t and τ are not). f is obtained by a transposition of the base template: $f = f' + \tau$. $P(f,t)$ then takes the form:

$$P(f,t) = \sum_{z=1}^{Z} P(z) \sum_{f'=1}^{F'} P_K(f'|z) P_I(f - f', t|z).$$

P_K is called the kernel distribution: it corresponds to the base spectral templates which are shifted by the impulse distribution P_I.

3. SCALE-INVARIANT PLCA

3.1. Model

In standard short-time Fourier transform spectrograms (with a linear frequency resolution), transposition is no longer a shift but approximately corresponds to an homothety of the spectral template: we thus model it with a multiplication by a scalar $\lambda \in \mathbb{R}^+ \backslash \{0\}$. Let X be a discrete random variable taking its values in $\{0, 1, 2, ...K\}$ and λ a continuous positive random variable with a density function p. The density of $u = \lambda X$ is then:

$$p_{\lambda X}(u) = \sum_{k=1}^{K} \frac{P(X = k) p(\frac{u}{k})}{k} + \delta(u) P(X = 0) \qquad (1)$$

where $\delta(u)$ is a Dirac delta function.

In our model, one assumes that the continuous frequency random variable $f_c \in \mathbb{R}$ (the link between f_c and the observed variable f will be made clearer later) is obtained by multiplying the base frequency $f' \in \{0, 1, \ldots, F'\}$ (which is independent of t given z) with the transposition factor $\lambda \in \mathbb{R}^+ \backslash \{0\}$ (which depends on t but not on f' given z). Using (1) with $u = f_c$, $k = f'$ and $K = F'$, we get:

$$P(f_c, t|z) = \sum_{f'=1}^{F'} \frac{P_K(f'|z)}{f'} P_I\left(\frac{f_c}{f'}, t|z\right) + \delta(f_c) P_K(0|z).$$

We use the notation P_K for the kernel distribution and P_I for the impulse distribution, as for shift-invariant PLCA. However they do not represent the same object. P_K still corresponds to the base spectral templates but is now rescaled by the impulse distribution P_I.

In this paper, we consider that $P_K(0|z) = 0$ to avoid the singularity at the null frequency (there still can be energy in the frequency channel 0 introduced when scaling down the frequency template). We then get:

$$P(f_c, t) = \sum_{z=1}^{Z} P(z) \sum_{f'=1}^{F'} \frac{P_K(f'|z)}{f'} P_I\left(\frac{f_c}{f'}, t|z\right).$$

The random variable f_c is continuous, but the observed random variable $f \in \{0, 1, \ldots, F\}$ is discrete. Then we will suppose that $f = \text{round}(f_c)$ and consequently:

$$P(f,t) = \int_{f-\frac{1}{2}}^{f+\frac{1}{2}} P(f_c, t) df_c.$$

Consequently, $\forall f \in \{0, 1, \ldots, F\}, \forall t \in \{1, \ldots, T\}$:

$$P(f,t) = \sum_{z,f'} \frac{P(z) P_K(f'|z)}{f'} \int_{f-\frac{1}{2}}^{f+\frac{1}{2}} P_I\left(\frac{f_c}{f'}, t|z\right) df_c.$$

The parameters to be estimated are then: $\theta = \{P(z), P_K(f'|z), P_I(\lambda, t|z)\}$.

For practical purposes, one needs to discretize P_I (which is a continuous density function with respect to λ) in some way, in order to estimate θ. We propose to perform this discretization by parameterizing P_I assuming that $\lambda \mapsto P_I(\lambda, t|z)$ is piece-wise constant for all t and z. We select a family $\{\lambda_k\}_{k \in \{1,\ldots,K\}}$ (which does not depend on t and z). In this paper, we choose $\lambda_k = 2^{\frac{k-k_0}{12 n_{\text{st}}}}$: this exponential discretization is chosen to fit a transposition scale in subdivisions of the tone (n_{st} corresponds to the number of discretized values of λ for each semitone). We assume that P_I is given by:

$$\forall \lambda \in [\lambda_k 2^{-\frac{1}{24 n_{\text{st}}}}, \lambda_k 2^{\frac{1}{24 n_{\text{st}}}}], \quad P_I(\lambda, t|z) = P_I(\lambda_k, t|z).$$

Moreover, we assume that P_I is zero outside these intervals. The values $P_I(\lambda_k, t|z)$ (for all k, t and z) then completely describe P_I.

Then:

$$\int_{f-\frac{1}{2}}^{f+\frac{1}{2}} P_I(\frac{f_c}{f'}, t|z) df_c = f' \sum_{k=k_{\text{min}}^{f,f'}}^{k_{\text{max}}^{f,f'}} P_I(\lambda_k, t|z) \delta\lambda_k^{f,f'}$$

where $k_{\text{min}}^{f,f'}$ is chosen so that $\lambda_{k_{\text{min}}^{f,f'}} 2^{-\frac{1}{24 n_{\text{st}}}} < \frac{f-\frac{1}{2}}{f'} \leq \lambda_{k_{\text{min}}^{f,f'}} 2^{\frac{1}{24 n_{\text{st}}}}$ and $k_{\text{max}}^{f,f'}$ is chosen so that $\lambda_{k_{\text{max}}^{f,f'}} 2^{-\frac{1}{24 n_{\text{st}}}} \leq \frac{f+\frac{1}{2}}{f'} < \lambda_{k_{\text{max}}^{f,f'}} 2^{\frac{1}{24 n_{\text{st}}}}$ (with the following constraints: $1 \leq k_{\text{min}}^{f,f'} \leq K$ and $1 \leq k_{\text{max}}^{f,f'} \leq K$):

$$\begin{aligned} k_{\text{min}} &= \left\lceil k_0 - \frac{1}{2} + 12 n_{\text{st}} \log_2(\frac{f - \frac{1}{2}}{f'}) \right\rceil \\ k_{\text{max}} &= \left\lfloor k_0 + \frac{1}{2} + 12 n_{\text{st}} \log_2(\frac{f + \frac{1}{2}}{f'}) \right\rfloor \end{aligned}$$

where $\lceil . \rceil$ is the ceiling function and $\lfloor . \rfloor$ the floor function.

Thus $\delta\lambda_k^{f,f'}$ is given by $\delta\lambda_k^{f,f'} = \min(\lambda_k 2^{\frac{1}{24 n_{\text{st}}}}, \frac{f+\frac{1}{2}}{f'}) - \max(\lambda_k 2^{-\frac{1}{24 n_{\text{st}}}}, \frac{f-\frac{1}{2}}{f'})$. When $\delta\lambda_k^{f,f'}$ is not limited by constraints on f and f', we will denote $\delta\lambda_k = \lambda_k 2^{\frac{1}{24 n_{\text{st}}}} - \lambda_k 2^{-\frac{1}{24 n_{\text{st}}}}$.

The parameters to be estimated are then: $\theta = \{P(z), P_K(f'|z), P_I(\lambda_k, t|z) | z \in [\![1, Z]\!], f' \in [\![1, F']\!], k \in [\![1, K]\!], t \in [\![1, T]\!]\}$.

3.2. Expectation-Maximization algorithm

We intend to estimate the value of the parameter θ that maximizes the log-likelihood of observing $\mathbf{V}_{ft}$:

$$L((\bar{f}, \bar{t})|\theta) = \sum_{i \in I} \log P(f_i, t_i), \qquad (2)$$

where $\bar{f}$ and $\bar{t}$ correspond to the draws of f and t (draws are indexed by $i \in I = \{1, ..., N\}$ where N is the total number of draws). As the number of draws that leads to the value (f,t) is $\mathbf{V}_{ft}$, the log-likelihood can be rewritten:

$$L((\bar{f}, \bar{t})|\theta) = \sum_{f=1}^{F} \sum_{t=1}^{T} \mathbf{V}_{ft} \log P(f,t)$$

The estimation will be performed with the Expectation-Maximization (EM) algorithm with latent variables z and f' (it would be equivalent to consider z and λ as latent variables since $f = \lambda f'$). The completed log-likelihood is:

$$\begin{aligned} L((\bar{f},\bar{t},z,f')|\theta) &= \sum_{f=1}^{F}\sum_{t=1}^{T} \mathbf{V}_{ft} \log P(f,t,z,f') \\ &= \sum_{f,t} \mathbf{V}_{ft} \Bigg\{ \log P(z) + \log P_K(f'|z) \\ &+ \log \left(\int_{f-\frac{1}{2}}^{f+\frac{1}{2}} P_I(\frac{f_c}{f'},t|z) df_c \right) \Bigg\} + c \end{aligned}$$

where c is a constant that does not depend on θ.

The completed log-likelihood expectation is then:

$$\begin{aligned} Q(\theta|\theta^{(c)}) = \sum_{f',z,f,t} &\mathbf{V}_{ft} P(z,f'|f,t,\theta^{(c)}) \Bigg\{ \log P(z) \\ &+ \log P_K(f'|z) + \log \left(\int_{f-\frac{1}{2}}^{f+\frac{1}{2}} P_I(\frac{f_c}{f'},t|z) df_c \right) \Bigg\} + c \end{aligned} \tag{3}$$

where $\theta^{(c)}$ is the current value of the parameter.

We can get an expression for $P(z,f'|f,t,\theta^{(c)})$ with respect to $\theta^{(c)}$ using the Bayes theorem (E step):

$$P(z,f'|f,t,\theta^{(c)}) = \frac{P_K^{(c)}(f'|z)P^{(c)}(z) \int_{f-\frac{1}{2}}^{f+\frac{1}{2}} P_I^{(c)}(\frac{f_c}{f'},t|z)df_c}{f' P^{(c)}(f,t)}. \tag{4}$$

The notation $(.)^{(c)}$ refers to values computed from the current parameter: $\theta^{(c)} = \{P^{(c)}(z), P_K^{(c)}(f'|z), P_I^{(c)}(\lambda,t|z)\}$.

The completed expectation (3) will be maximized (M step) with respect to θ ($\theta^{(c)}$ being fixed). As θ is made up of probabilities that must sum to 1, the maximization is constrained. Thus we use the Lagrangian[1]:

$$\begin{aligned} H(\theta|\theta^{(c)}) &= Q(\theta|\theta^{(c)}) + \sum_z \rho_z \left(1 - \sum_{f'} P_K(f'|z) \right) \\ &+ \mu \left(1 - \sum_z P(z) \right) + \sum_z \tau_z \left(1 - \sum_t \int_{\lambda_{\text{min}}}^{\lambda_{\text{max}}} P_I(\lambda,t|z) \right) \end{aligned} \tag{5}$$

where ρ_z, μ, and τ_z are Lagrange multipliers.

3.2.1. Update of $P(z)$ and $P_K(f'|z)$

$\frac{\partial H(\theta|\theta^{(c)})}{\partial P(z)} = 0$ leads to the update rule of $P(z)$:

$$P(z) \leftarrow \frac{\sum_{f,f',t} \mathbf{V}_{ft} P(z,f'|f,t,\theta^{(c)})}{\sum_{z,f,f',t} \mathbf{V}_{ft} P(z,f'|f,t,\theta^{(c)})}. \tag{6}$$

In a similar way, we obtain the update rule of $P_K(f'|z)$:

$$P_K(f'|z) \leftarrow \frac{\sum_{f,t} \mathbf{V}_{ft} P(z,f'|f,t,\theta^{(c)})}{\sum_{f,f',t} \mathbf{V}_{ft} P(z,f'|f,t,\theta^{(c)})}. \tag{7}$$

3.2.2. Update of $P_I(\lambda_k,t|z)$

Because of the expression of the Lagrangian $H(\theta|\theta^{(c)})$ with respect to $P_I(\lambda_k,t|z)$, the update rule of $P_I(\lambda_k,t|z)$ is more complex to derive. We consider the following "fixed-point" update rule (iterated several times) which hopefully will converge to a zero of $\frac{\partial H}{\partial P_I(\lambda_k,t|z)}$:

$$P_I(\lambda_k,t|z) \leftarrow \sum_{f,f'} \frac{\mathbf{V}_{ft} P(z,f'|f,t,\theta^{(c)}) P_I(\lambda_k,t|z) \mathbb{1}_{[k_{\text{min}},k_{\text{max}}]}(k)}{\tau_z \delta\lambda_k \sum_{k'=k_{\text{min}}^{f,f'}}^{k_{\text{max}}^{f,f'}} P_I(\lambda_{k'},t|z)\delta\lambda_{k'}^{f,f'}} \delta\lambda_k^{f,f'}. \tag{8}$$

In equation (8), the division by τ_z is a normalization.

We have not managed to prove convergence of P_I under several iterations of the update rule (8) to a zero of $\frac{\partial H}{\partial P_I}$. However we observed it in practice.

As $Q(\theta|\theta^{(c)})$ (defined in equation (3)) and the constraints are $\mathcal{C}^1$, $Q(\theta|\theta^{(c)})$ is strictly concave and the constraints are affine, and the regularity conditions are satisfied, a stationary point of $H(\theta|\theta^{(c)})$ is necessarily the global maximum of $Q(\theta|\theta^{(c)})$ under the normalization constraints. Thus, $Q(\theta|\theta^{(c)})$ is effectively maximized at each iteration using the update rules (6), (7) and (8) and the EM algorithm makes the likelihood of the observed data increase at each iteration. More details about the calculation of the update rules can be found in [11].

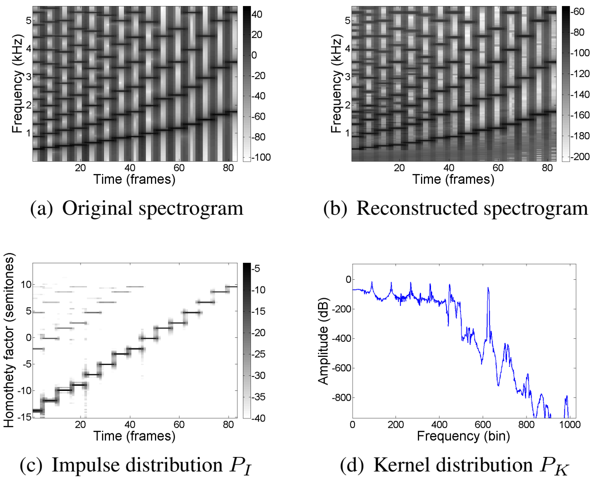
4. EXAMPLES AND APPLICATIONS

4.1. Toy example

In this section, we use our algorithm to decompose a simple spectrogram. This spectrogram is obtained by STFT (using a 1024 sample-long Hann window with 75% overlap and no zero-padding) of a short excerpt of synthesizer which plays the notes of a A major scale on two octaves (from A4 to A6) sampled at 11025Hz. The original spectrogram is pictured in figure 1(a). The decomposition provides the reconstructed spectrogram pictured in figure 1(b): the reconstructed spectrogram is very similar to the original one. The difference of maximum amplitudes between original and reconstructed spectrograms come from the normalization of $P(f,t)$ ($\mathbf{V}_{ft}$ is not normalized), but the dynamic remains the same in both spectrograms. High frequency harmonics of the reconstructed spectrogram are slightly larger than the original one.

Obtained kernel distribution P_K is represented in figure 1(d): we can see that the factorized template is actually harmonic. For high values of the frequency index, amplitudes of the templates tend to be very small. This comes from the fact that our model considers that values of the spectrogram outside the observed frequencies are zeros whereas the model spectrogram $P(f,t)$ can take positive values outside this range. This can be solved using the approach described in [12]. In practice, for signals of actual acoustic instruments, this is not a real issue, since the harmonic content of such

[1] We did not write non-negative constraints in the problem statement because our final solution guarantees that parameters remain non-negative, then these constraints are inactive.



(a) Original spectrogram (b) Reconstructed spectrogram

(c) Impulse distribution P_I (d) Kernel distribution P_K

Figure 1: Scale-invariant PLCA: decomposition of a diatonic scale.

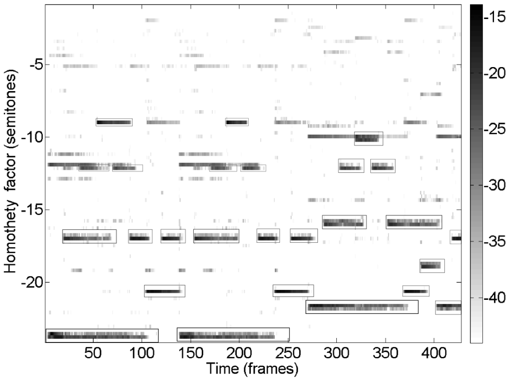


Figure 2: Impulse distribution of the introduction of *Because*.

signals is mostly smothered by noise from 5000Hz: thus a sampling rate of 22050Hz or more permits to reduce this issue.

Impulse distribution P_I is represented in figure 1(c): high probabilities clearly appear at actual notes relative positions. There are also some replicas of the notes at positions with high harmonic similitudes (octave, twelfth, double octave...). At onset times, P_I takes high values for many values of the homothety factor λ: this is caused by the flat shape of the spectrum of onsets which is matched here with several rescaled harmonic templates.

4.2. Real audio data

In this section, we present the SIPLCA decomposition of the spectrogram of the 10 first seconds of the song *Because* by the Beatles with a single template ($Z = 1$). The decomposed signal consists of a polyphonic harpsichord introduction recorded in real conditions. The original stereo signal was transformed into a mono signal (summing both channels) and downsampled to 22050Hz. The spectrogram was calculated with a STFT using a 2048 sample-long Hann window with 75% overlap and no zero-padding.

The obtained impulse distribution P_I is represented in figure 2: actually played notes are materialized by a rectangle in the figure. We can see that in all rectangles, P_I takes high values.

The impulse distribution P_I is thus very similar to the impulse distribution that can be obtained with shift-invariant decompositions. However, as our decomposition is computed on linear frequency resolution spectrograms, it has an important advantage: it is possible to generate time-frequency masks that can be directly used to separate different components with Wiener filtering. Thus, it is possible to isolate single notes in a polyphonic signal and to repitch them individually. Examples of processed sounds are available on the webpage [13].

5. CONCLUSION

In this paper, we proposed a new way of decomposing nonnegative music spectrograms: the decomposition is based on a few frequency templates that can be rescaled at each frame (which corresponds to a transposition). We presented examples of this decomposition on music spectrograms and showed how it can be used to modify individual notes in a polyphonic signal.

Future works will include a better representation of nonharmonic components of musical sounds, such as transients.

6. REFERENCES

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

[2] M. Shashanka, B. Raj, and P. Smaragdis, "Sparse overcomplete latent variable decomposition of counts data," in *NIPS*, Vancouver, BC, Canada, December 2007.

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

[4] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *ICA*, London, UK, September 2007.

[5] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *WASPAA*, New Paltz, NY, USA, October 2003, pp. 177 – 180.

[6] M. Schmidt and M. Mørup, "Nonnegative matrix factor 2-D deconvolution for blind single channel source separation," in *ICA*, vol. 3889, Paris, France, April 2006, pp. 700–707.

[7] P. Smaragdis, B. Raj, and M. Shashanka, "Sparse and shift-invariant feature extraction from non-negative data," in *ICASSP*, Las Vegas, Nevada, USA, March 2008, pp. 2069 – 2072.

[8] J. C. Brown, "Calculation of a constant Q spectral transform," *Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, January 1991.

[9] C. Schörkhuber and A. Klapuri, "Constant-Q transform toolbox for music processing," in *7th Sound and Music Computing Conference*, Barcelona, Spain, July 2010.

[10] D. Fitzgerald, M. Cranitch, and E. Coyle, "Shifted non-negative matrix factorisation for sound source separation," in *SSP*, Bordeaux, France, July 2005, pp. 1132 – 1137.

[11] R. Hennequin, B. David, and R. Badeau, "Scale-invariant probabilistic latent component analysis," Telecom ParisTech, Tech. Rep., 2011.

[12] P. Smaragdis, B. Raj, and M. Shashanka, "Missing data imputation for time-frequency representations of audio signals," *Journal of Signal Processing Systems*, August 2010.

[13] http://perso.telecom-paristech.fr/~hennequi/demoSIPLCA.html.