COMPLEX NMF: A NEW SPARSE REPRESENTATION FOR ACOUSTIC SIGNALS

Hirokazu Kameoka†*, Nobutaka Ono*‡*, Kunio Kashino*†*, Shigeki Sagayama*‡

† NTT Communication Science Laboratories, NTT Corporation,
3-1 Morinosato Wakamiya, Atsugi, Kanagawa 243-0198, Japan
‡ Graduate School of Information Science and Technology, The University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

ABSTRACT

This paper presents a new sparse representation for acoustic signals which is based on a mixing model defined in the complex-spectrum domain (where additivity holds), and allows us to extract recurrent patterns of magnitude spectra that underlie observed complex spectra and the phase estimates of constituent signals. An efficient iterative algorithm is derived, which reduces to the multiplicative update algorithm for non-negative matrix factorization developed by Lee under a particular condition.

Index Terms— Sparse signal representation, non-negative matrix factorization, sparse coding, data-driven approach

1. INTRODUCTION

The use of sparse representations including sparse coding (SC) [1] and non-negative matrix factorization (NMF) [2] in acoustic signal processing has gradually increased in recent years. Given a set of observed vectors, the goal is to find a set of basis vectors such that any observations can be succinctly represented as a linear combination of a small number of 'active' bases. While SC enforces the sparseness of basis activations by incorporating a sparsity cost in the objective function, NMF achieves sparseness as a result of a side effect caused by non-negativity constraints. Unlike certain other unsupervised basis learning techniques such as principal component analysis (PCA) and independent component analysis (ICA), the particularity of SC and NMF is that recurrent patterns underlying the observations are considered good bases owing to the sparse nature of the decompositions. [3]–[6] were among the first to apply this concept to an audio processing problem, and since then, NMF, in particular, has been applied extensively with considerable success to various problems including automatic music transcription, monaural source separation, speech denoising, bandwidth expansion, audio classification and speech dereverberation [7].

One reason for NMF being a powerful tool as regards extracting regularities or structural patterns from acoustic signals may be that the mixing model is defined in the magnitude spectrum domain. Because of the phase-invariant nature of magnitude spectra, NMF is able to project all signals that have the same spectral shape onto a single basis. This allows us to represent a variety of acoustic phenomena efficiently using a very compact set of spectrum bases. However, the mixing model implicitly assumes the additivity of magnitude spectra, which holds only approximately. Although attempts are made to mitigate the non-additivity problem with respect to NMF in certain recently published papers [8, 9], in these papers the additivity of power spectra is only verified in a statistical sense on the assumption that the constituent signals are Gaussian processes and perfectly independent. Another drawback of NMF is that it cannot estimate the phase spectra of underlying constituent signals, which certainly limits its range of applications.

Within the framework of SC or semi-NMF we can construct a mixing model defined in the time domain [10, 11] or in the complex-spectrum domain. Such a mixing model is valid in terms of defining an additive quantity as the basis, namely, a waveform signal. However, as pointed out in [6] because the phase coherence between frequency components can be easily destroyed as a result of many factors, it is difficult to capture high-level structural elements from observations through the use of waveform/complex-spectrum bases.

In this paper, we present a new framework "Complex NMF", which offers the advantages of the NMF and SC frameworks concurrently: Complex NMF, (1) is based on a mixing model defined in the complex-spectrum domain (where additivity holds), (2) can extract the recurrent patterns of magnitude spectra that underlie the observed complex spectra and the phase spectra of constituent signals, and (3) can be performed with an efficient iterative algorithm, which reduces to the multiplicative update algorithm developed by Lee [2] under a particular condition.

2. PRINCIPLE OF COMPLEX NMF

2.1. Sparse representation model

We start with the basic assumption that the short-term Fourier transform (STFT) of an arbitrary acoustic signal, $F_{x,t} \in \mathbb{C}$, consists of K complex-valued elements

$$F_{x,t} = \sum_{k=1}^{K} a_{k,x,t} = \sum_{k} |a_{k,x,t}| e^{j\phi_{k,x,t}}, \qquad (1)$$

where x and t are the frequency and frame indices, respectively. Here, we factorize the modulus, $|a_{k,x,t}|$, into the product of nonnegative parameters, $H_{k,x}$ and $U_{k,t}$,

$$|a_{k,x,t}| = H_{k,x} U_{k,t}, \;\; H_{k,x} \geq 0, \; U_{k,t} \geq 0, \qquad (2)$$

and assume

$$\sum_{x} H_{k,x} = 1 \;\; (k = 1, \cdots, K), \qquad (3)$$

in order to avoid an indeterminacy in the scaling. We therefore arrive at the following mixing model

$$F_{x,t} = \sum_k H_{k,x} U_{k,t} e^{j\phi_{k,x,t}}. \quad (4)$$

In this model, each element comprises a static magnitude spectral shape $H_{k,x}$, a time-varying activation coefficient $U_{k,t}$ and a time-varying phase spectrum $\phi_{k,x,t}$. Given an observed spectrum $Y_{x,t}$, the goal is to find a decomposition such that $Y_{x,t} \simeq F_{x,t}$ in which the basis activations are sparse. This basically means that any observed complex spectrum can be well represented using only a few active magnitude spectrum bases each of which is paired with an arbitrary phase spectrum. Combining the goal of a small reconstruction error with that of sparseness, we consider that the model will allow us to extract the recurrent patterns of magnitude spectra underlying observed spectra as with NMF and, simultaneously, the phase spectra of constituent signals. As seen from Eq. (4), the model cannot be expressed in matrix notation as with NMF and SC, it can clearly be considered a new class of sparse representation model.

2.2. Problem setting

Given an observed complex spectrum, $Y_{x,t} \in \mathbb{C}$, we would like to find the optimal estimates of $H_{k,x}$, $U_{k,t}$ and $\phi_{k,x,t}$. For simplicity of notation let $Y \equiv \{Y_{x,t}\}_{X\times T}$, $F \equiv \{F_{x,t}\}_{X\times T}$, $H \equiv \{H_{k,x}\}_{K\times X}$, $U \equiv \{U_{k,t}\}_{K\times T}$, $\phi \equiv \{\phi_{k,x,t}\}_{K\times X\times T}$. Now, we assume the following generative model

$$Y_{x,t} = F_{x,t} + \epsilon_{x,t}, \quad (5)$$

where the reconstruction error $\epsilon_{x,t}$ is assumed to be complex Gaussian white noise with mean 0 and variance σ^2. The likelihood of $\theta = \{H, U, \phi\}$ is thus given as follows:

$$P(Y|\theta) = \prod_{x,t} \frac{1}{\pi\sigma^2} \exp\left(-\frac{|Y_{x,t} - F_{x,t}|^2}{\sigma^2}\right). \quad (6)$$

We assume for convenience that the prior distributions for H, U and ϕ are independent, which yields $P(\theta) = P(H)P(U)P(\phi)$, and that $P(H)$ and $P(\phi)$ are uniform distributions. $P(U)$ corresponds to the sparsity cost, for which a natural choice is a generalized Gaussian prior

$$P(U) = \prod_{k,t} \frac{1}{2\Gamma\left(1+\frac{1}{p}\right)b} \exp\left(-\frac{|U_{k,t}|^p}{b^p}\right), \quad (7)$$

where p and b are the parameters that determine the shape of the distribution. When $0 < p < 2$, $P(U)$ becomes super-Gaussian and promotes sparsity if the norm of U is bounded. The likely values of H, U and ϕ can thus be inferred from the posterior density

$$P(\theta|Y) \propto P(Y|\theta)P(U). \quad (8)$$

From Eqs. (3)–(8), we are thus led to solve the following optimization problem:

$$\begin{aligned} \text{minimize} \quad & f(\theta) \equiv \sum_{x,t} \left|Y_{x,t} - F_{x,t}\right|^2 + 2\lambda \sum_{k,t} |U_{k,t}|^p \\ \text{subject to} \quad & \sum_x H_{k,x} = 1 \ \ (k = 1, \cdots, K) \end{aligned}$$

2.3. Iterative algorithm

We present an efficient algorithm that seeks to minimize $f(\theta)$. The auxiliary function concept is utilized for our derivation, similar to [2]. We use $G(\theta)$ to denote an objective function that we want to minimize w.r.t. θ, and define an auxiliary function of $G(\theta)$ as $G^+(\theta, \bar{\theta})$ if it satisfies

$$G(z) = \min_{\bar{z}} G^+(z, \bar{z}). \quad (9)$$

Theorem 1. *$G(\theta)$ is non-increasing under the updates, $\bar{\theta} \leftarrow \mathrm{argmin}_{\bar{\theta}} G^+(\theta, \bar{\theta})$ and $\theta \leftarrow \mathrm{argmin}_{\theta} G^+(\theta, \bar{\theta})$.*

Proof: Assume that we set θ at an arbitrary value θ_ℓ. Let $\bar{\theta}_{\ell+1} = \mathrm{argmin}_{\bar{\theta}} G^+(\theta_\ell, \bar{\theta})$ and $\theta_{\ell+1} = \mathrm{argmin}_{\theta} G^+(\theta, \bar{\theta}_{\ell+1})$. It is obvious that $G(\theta_\ell) = G^+(\theta_\ell, \bar{\theta}_{\ell+1})$. We deduce from $\theta_{\ell+1} = \mathrm{argmin}_{\theta} G^+(\theta, \bar{\theta}_{\ell+1})$ that $G^+(\theta_\ell, \bar{\theta}_{\ell+1}) \geq G^+(\theta_{\ell+1}, \bar{\theta}_{\ell+1})$. By definition, from $G^+(\theta_{\ell+1}, \bar{\theta}_{\ell+1}) \geq G(\theta_{\ell+1})$ we verify that $G(\theta_\ell) \geq G(\theta_{\ell+1})$. □

By iteratively updating θ and $\bar{\theta}$, $G(\theta)$ will converge to a stationary point. We can apply this concept to the minimization of $f(\theta)$ using the following theorem:

Theorem 2 (Auxiliary function)**.** *The function*

$$\begin{aligned} f^+(\theta, \bar{\theta}) \equiv & \sum_{k,x,t} \frac{\left|\bar{Y}_{k,x,t} - H_{k,x}U_{k,t}e^{j\phi_{k,x,t}}\right|^2}{\beta_{k,x,t}} \\ & + \lambda \sum_{k,t} \Big(p|\bar{U}_{k,t}|^{p-2} {U_{k,t}}^2 + 2|\bar{U}_{k,t}|^p - p|\bar{U}_{k,t}|^p \Big), \end{aligned} \quad (10)$$

with $\bar{\theta} = \{\bar{Y}, \bar{U}\}$, $\bar{Y} \equiv \{\bar{Y}_{k,x,t}\}_{K\times X\times T}$, $\bar{U} \equiv \{\bar{U}_{k,t}\}_{K\times T}$, is an auxiliary function for $f(\theta)$, if $\sum_k \bar{Y}_{k,x,t} = Y_{x,t}$ and $0 < p \leq 2$. $\beta_{k,x,t}$ can be any positive number satisfying $\sum_k \beta_{k,x,t} = 1$. $f^+(\theta, \bar{\theta})$ is minimized w.r.t. $\bar{\theta}$ when

$$\bar{Y}_{k,x,t} = H_{k,x}U_{k,t}e^{j\phi_{k,x,t}} + \beta_{k,x,t}\big(Y_{x,t} - F_{x,t}\big), \quad (11)$$

$$\bar{U}_{k,t} = U_{k,t}. \quad (12)$$

Proof: This follows from the following lemmas. □

Lemma 1. *With $\sum_k \bar{Y}_{k,x,t} = Y_{x,t}$ and for any $\beta_{k,x,t} > 0$ such that $\sum_k \beta_{k,x,t} = 1$,*

$$\left|Y_{x,t} - F_{x,t}\right|^2 \leq \sum_k \frac{\left|\bar{Y}_{k,x,t} - H_{k,x}U_{k,t}e^{j\phi_{k,x,t}}\right|^2}{\beta_{k,x,t}}. \quad (13)$$

Proof: By adding the Lagrange multiplier term to the right-hand side, and then taking the derivative w.r.t. $\bar{Y}^*_{k,x,t}$ and setting it at zero, we obtain Eq. (11). Substituting this to the right-hand side, we determine the minimum value of the right-hand side, which is equal to the left-hand side of the inequality. □

Lemma 2. *If $0 < p \leq 2$, for any $U_{k,t} \in \mathbb{R}$ and $\bar{U}_{k,t} \in \mathbb{R}$,*

$$|U_{k,t}|^p \leq \frac{p|\bar{U}_{k,t}|^{p-2}}{2} {U_{k,t}}^2 + |\bar{U}_{k,t}|^p - \frac{p|\bar{U}_{k,t}|^p}{2}. \quad (14)$$

Proof: Regarding $|U_{k,t}|^p$ as the function of $U_{k,t}$, the right-hand side of the inequality amounts to a convex quadratic function that is tangent to $|U_{k,t}|^p$ at argument $U_{k,t} = \pm\bar{U}_{k,t}$. □

As the update rule for $\bar{\theta}$ is shown in Eqs. (11), (12), we only need to derive the update rule for θ. Although it is necessary to take account of the normalization condition for $H_{k,x}$ and solve the Lagrange dual problem, here we simply describe a convenient approach, which consists of solving the unconstrained minimization of $f^+(\theta, \bar{\theta})$ w.r.t. $H_{k,x}$ and projecting its solution onto the constraint space. Differentiating $f^+(\theta, \bar{\theta})$ partially w.r.t. $H_{k,x}$ and $U_{k,t}$, and setting them at zero, we obtain update rules for $H_{k,x}$ and $U_{k,t}$:

$$H_{k,x} = \frac{\sum_t \frac{U_{k,t}}{\beta_{k,x,t}} \mathrm{Re}\left[\bar{Y}^*_{k,x,t} e^{j\phi_{k,x,t}}\right]}{\sum_t \frac{{U_{k,t}}^2}{\beta_{k,x,t}}}, \tag{15}$$

$$U_{k,t} = \frac{\sum_x \frac{H_{k,x}}{\beta_{k,x,t}} \mathrm{Re}\left[\bar{Y}^*_{k,x,t} e^{j\phi_{k,x,t}}\right]}{\sum_x \frac{{H_{k,x}}^2}{\beta_{k,x,t}} + \lambda p |\bar{U}_{k,t}|^{p-2}}. \tag{16}$$

We note that these updates are guaranteed to provide the minimum value for each coordinate because it is obvious from an inspection of their second derivatives. Next, we derive the update rule for $\phi_{k,x,t}$. By using c to denote the terms that do not depend on $\phi_{k,x,t}$, $f^+(\theta, \bar{\theta})$ can be simply written as follows

$$f^+(\theta, \bar{\theta}) = c - 2 \sum_{k,x,t} |A_{k,x,t}| \cos(\phi_{k,x,t} - C_{k,x,t}),$$

where $A_{k,x,t} = \bar{Y}_{k,x,t} H_{k,x} U_{k,t} / \beta_{k,x,t}$ and

$$\cos C_{k,x,t} = \frac{\mathrm{Re}\left[\bar{Y}_{k,x,t}\right]}{|\bar{Y}_{k,x,t}|}, \ \sin C_{k,x,t} = \frac{\mathrm{Im}\left[\bar{Y}_{k,x,t}\right]}{|\bar{Y}_{k,x,t}|}.$$

$f^+(\theta, \bar{\theta})$ is obviously minimized when $\cos(\phi_{k,x,t} - C_{k,x,t}) = \cos\phi_{k,x,t} \cos C_{k,x,t} + \sin\phi_{k,x,t} \sin C_{k,x,t} = 1$, that is, $\cos\phi_{k,x,t} = \cos C_{k,x,t}$ and $\sin\phi_{k,x,t} = \sin C_{k,x,t}$. This leads to the update formula for $e^{j\phi_{k,x,t}}$

$$e^{j\phi_{k,x,t}} = \frac{\bar{Y}_{k,x,t}}{|\bar{Y}_{k,x,t}|}. \tag{17}$$

Substituting Eq. (17) into Eqs. (15), (16), we notice that the non-negativity of $H_{k,x}$ and $U_{k,t}$ is preserved if we start with non-negative initial conditions for $H_{k,x}$ and $U_{k,t}$.

2.4. Condition for equivalence to NMF

The iterative algorithm presented in 2.3 reduces to the Lee's multiplicative update algorithm [2] under a particular condition. Let us consider a particular situation where we fix the value of ϕ, from the beginning of the iteration, at

$$e^{j\phi_{k,x,t}} = \frac{Y_{x,t}}{|Y_{x,t}|}. \tag{18}$$

Substituting this into Eq. (11),

$$\bar{Y}_{k,x,t} = Y_{x,t} \left[\frac{H_{k,x} U_{k,t}}{|Y_{x,t}|} + \beta_{k,x,t} \left(1 - \frac{\sum_n H_{n,x} U_{n,t}}{|Y_{x,t}|} \right) \right].$$

Substituting this result and Eq. (18) into Eq. (15),

$$H_{k,x} \leftarrow \frac{\sum_t \left[\frac{H_{k,x} U_{k,t}^2}{\beta_{k,x,t}} + U_{k,t} \left(|Y_{x,t}| - \sum_n H_{n,x} U_{n,t} \right) \right]}{\sum_t \frac{{U_{k,t}}^2}{\beta_{k,x,t}}}.$$

As $\beta_{k,x,t}$ is a parameter that can be chosen arbitrarily subject to $\beta_{k,x,t} > 0$ and $\sum_k \beta_{k,x,t} = 1$, the convergence of the iterative algorithm is still guaranteed even if we change its value in parallel with H and U in each iteration. If we decide to replace the value of $\beta_{k,x,t}$ at each iteration by

$$\beta_{k,x,t} = \frac{H_{k,x} U_{k,t}}{\sum_n H_{n,x} U_{n,t}}, \tag{19}$$

the update formula for $H_{k,x}$ leads eventually to

$$H_{k,x} \leftarrow H_{k,x} \frac{\sum_t U_{k,t} |Y_{x,t}|}{\sum_t U_{k,t} \sum_n H_{n,x} U_{n,t}}, \tag{20}$$

which amounts to Lee's multiplicative update rule for the Frobenius norm criterion. We obtain the same conclusion for $U_{k,t}$, if $\lambda = 0$. Based on this analysis, we can conjecture that

(1) the present algorithm works as efficiently and stably as the Lee's NMF algorithm,

(2) updating $\beta_{k,x,t}$ according to Eq. (19) is somewhat more effective than fixing it at a constant value, and

(3) we can stabilize the algorithm by running NMF at the beginning of the iteration, which can be performed simply by fixing the value of ϕ at Eq. (18).

The iterative algorithm is summarized as follows:

1. Initialize H, U and ϕ.
2. Update $\bar{\theta} = \{\bar{Y}, \bar{U}\}$ according to Eqs. (11), (12).
3. Update $\theta = \{H, U, \phi\}$ according to Eqs. (15), (16), (17) and $H_{k,x} \leftarrow H_{k,x} / \sum_n H_{n,x}$.
4. Update β according to Eq. (19) and return to 2.

3. EXPERIMENTS

In this section we report some results for speech data excerpted from the ATR B-set speech database. The aim of the experiments was to determine whether Complex NMF has an effect as with NMF to extract the underlying patterns of magnitude spectra from audio data. All speech data were monaural and sampled at 16kHz. STFT was computed using a Hanning window that was 32ms long with a 16ms overlap. p and λ were set at $p = 1.2$, $\lambda = \sum_{x,t} |Y_{x,t}|^2 / K^{1-p/2} \times 10^{-5}$. The

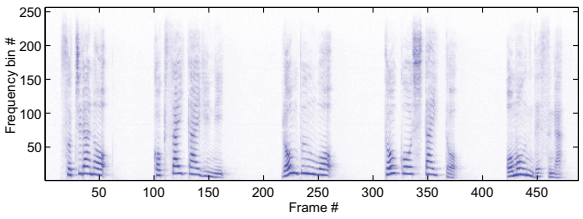

Fig. 1. Magnitude spectrogram of speech A

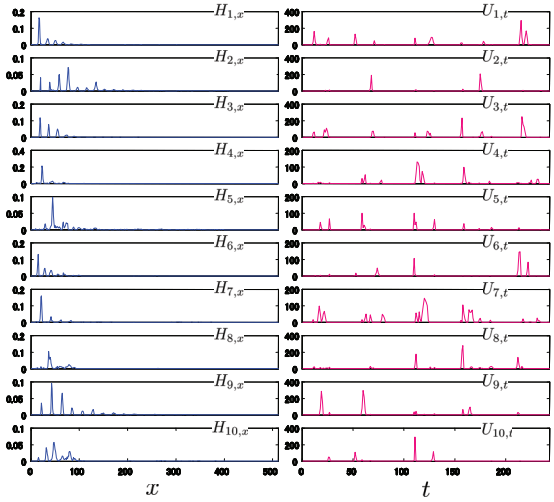

Fig. 2. 10 patterns of magnitude spectra and their activation patterns

algorithm was run for 30 iterations. H and U were initialized with random values and $\phi_{k,x,t}$ was initially set at $\arg(Y_{x,t})$.

For the first experiment we tested Complex NMF, with $K = 10$, on a single voice signal from a female speaker ("speech A"). Fig. 1 shows the spectrogram of speech A. The patterns of magnitude spectra discovered by using Complex NMF can be seen in Fig. 2 along with the corresponding activation patterns. As shown in Fig. 2, harmonic structures are vividly captured in H even without any prior knowledge.

For the second experiment we tested Complex NMF with $K = 30$ on a mixed voice signal (see Fig. 3). The mixed speech data was created by adding speech data from another female speaker ("speech B") to speech A. In this experiment, the aim was to ascertain whether each of the estimated patterns of magnitude spectra corresponds to a single-voice spectrum. We selected the pattern of magnitude spectrum that was closest to the true spectrum of each speaker per frame and then concatenated the framewise signals, each of which we constructed using the selected pattern and the corresponding activation coefficient and phase spectrum to synthesize the whole signal stream. If the synthesized signal contains two voices, it may suggest that the model is overfitting the observation. In contrast, if the synthesized signal does not sound like speech, it may suggest that the model is underfitting the observation. Seen in this light, we calculated SNR to see how well each speech signal could be restored based on the above procedure. As a result, we obtained 4.3dB and 4.0dB for the two speakers, which constituted an improvement from 0.87dB and -0.87dB, respectively. This result suggests that each estimated pattern of magnitude spectrum corresponds fairly appropriately to a single voice spectrum. The restored speech signal corresponding to speech A can be seen in Fig. 4.

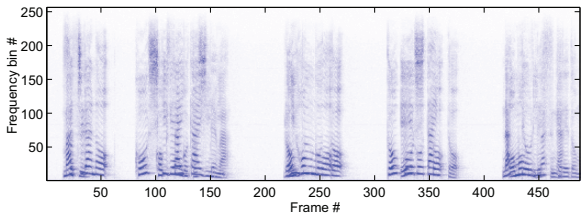

Fig. 3. Magnitude spectrogram of mixed speech

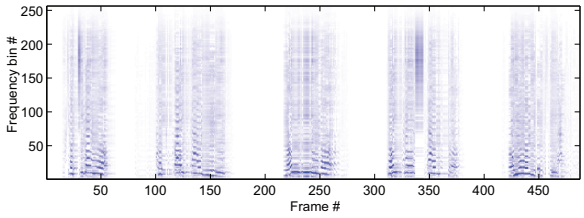

Fig. 4. Synthesized signal corresponding to speech A

4. CONCLUDING REMARKS

We developed a new framework for the sparse representation of acoustic signals. We named it "Complex NMF" (which may appear contradictory since the term "complex" conflicts with the term "non-negative" and the decomposition is no longer a matrix factorization), because, in common with NMF, it has the ability to generate non-negative matrices H and U, and is different in that the input matrix Y is assumed to be a complex matrix and it also generates a third-rank complex-valued tensor $e^{j\phi_{k,x,t}}$. As Complex NMF is capable of estimating the phase spectra of constituent signals, future work will include its extension to the formulation for microphone array processing. Motivated by [12], we also want to investigate a general class of the prior structure suited to the model introduced in this paper.

5. REFERENCES

[1] B.A. Olshausen and D.J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," Nature, **381**, 607–609, 1996.

[2] D.D. Lee and H.S. Seung, "Algorithms for nonnegative matrix factorization," Proc. NIPS'00, 556–562, 2000.

[3] M.D. Plumbley, S.A. Abdallah, J.P. Bello, M.E. Davies, G. Monti and M.B. Sandler, "Automatic music transcription and audio source separation," Cybernetics and Systems, **33**(6), 603–627, 2002.

[4] P. Smaragdis and J.C. Brown, "Non-negative matrix factorization for music transcription," Proc. WASPAA'03, 177–180, 2003.

[5] T. Virtanen, "Sound source separation using sparse coding with temporal continuity objective," Proc. ICMC'03, 231–234, 2003.

[6] S.A. Abdallah, M.D. Plumbley, "Unsupervised analysis of polyphonic music using sparse coding," IEEE Trans. NN, **17**(1), 179–196, 2006.

[7] H. Kameoka T. Nakatani and T. Yoshioka, "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms," Proc. ICASSP'09, to appear.

[8] R.M. Parry and I. Essa, "Phase-Aware Non-negative Spectrogram Factorization," Proc. ICA'07, 536–543, 2007.

[9] C. Févotte, N. Bertin and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence with application to music analysis," Technical Report TELECOM ParisTech 2008D006, 2008.

[10] T. Blumensath and M.E. Davies, "Sparse and shift-invariant representations of music," IEEE Trans. ASLP, **14**(1), 50–57, 2006.

[11] J. Le Roux, A. de Cheveigne and L.C. Parra, "Adaptive template matching with shift-invariant semi-NMF," Proc. NIPS'08, 2008.

[12] A.T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," Technical Report CUED/F-INFENG/TR.609, University of Cambridge, 2008.