HARMONIC-TEMPORAL FACTOR DECOMPOSITION INCORPORATING MUSIC PRIOR INFORMATION FOR INFORMED MONAURAL SOURCE SEPARATION

Tomohiko Nakamura[†], **Kotaro Shikata**[†], **Norihiro Takamune**[†], **Hirokazu Kameoka**[†‡]
[†]Graduate School of Information Science and Technology, The University of Tokyo.
[‡]NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation.
{nakamura,k-shikata,takamune,kameoka}@hil.t.u-tokyo.ac.jp

ABSTRACT

For monaural source separation two main approaches have thus far been adopted. One approach involves applying non-negative matrix factorization (NMF) to an observed magnitude spectrogram, interpreted as a non-negative matrix. The other approach is based on the concept of computational auditory scene analysis (CASA). A CASA-based approach called the "harmonic-temporal clustering (HTC)" aims to cluster the time-frequency components of an observed signal based on a constraint designed according to the local time-frequency structure common in many sound sources (such as harmonicity and the continuity of frequency and amplitude modulations). This paper proposes a new approach for monaural source separation called the "Harmonic-Temporal Factor Decomposition (HTFD)" by introducing a spectrogram model that combines the features of the models employed in the NMF and HTC approaches. We further describe some ideas how to design the prior distributions for the present model to incorporate musically relevant information into the separation scheme.

1. INTRODUCTION

Monaural source separation is a process in which the signals of concurrent sources are estimated from a monaural polyphonic signal and is one of fundamental objectives offering a wide range of applications such as music information retrieval, music transcription and audio editing.

While we can use spatial cues for blind source separation with multichannel inputs, for monaural source separation we need other cues instead of the spatial cues. For monaural source separation two main approaches have thus far been adopted. One approach is based on the concept of computational auditory scene analysis (e.g., [7]). The auditory scene analysis process described by Bregman [1] involves grouping elements that are likely to have originated from the same source into a perceptual structure called an auditory stream. In [8, 10], an attempt has been made to imitate this process by clustering time-frequency components based on a constraint designed according to the auditory grouping cues (such as the harmonicity and the coherences and continuities of amplitude and frequency modulations). This method is called "harmonic-temporal clustering (HTC)."

The other approach involves applying non-negative matrix factorization (NMF) to an observed magnitude spectrogram (time-frequency representation) interpreted as a non-negative matrix [19]. The idea behind this approach is that the spectrum at each frame is assumed to be represented as a weighted sum of a limited number of common spectral templates. Since the spectral templates and the mixing weights should both be non-negative, this implies that an observed spectrogram is modeled as the product of two non-negative matrices. Thus, factorizing an observed spectrogram into the product of two non-negative matrices allows us to estimate the unknown spectral templates constituting the observed spectra and decompose the observed spectra into components associated with the estimated spectral templates.

The two approaches described above rely on different clues for making separation possible. Roughly speaking, the former approach focuses on the local time-frequency structure of each source, while the latter approach focuses on a relatively global structure of music spectrograms (such a property that a music signal typically consists of a limited number of recurring note events). Rather than discussing which clues are more useful, we believe that both of these clues can be useful for achieving a reliable monaural source separation algorithm. This belief has led us to develop a new model and method for monaural source separation that combine the features of both HTC and NMF. We call the present method "harmonic-temporal factor decomposition (HTFD)."

The present model is formulated as a probabilistic generative model in such a way that musically relevant information can be flexibly incorporated into the prior distributions of the model parameters. Given the recent progress of state-of-the-art methods for a variety of music information retrieval (MIR)-related tasks such as audio key detection, audio chord detection, and audio beat tracking, information such as key, chord and beat extracted from the given signal can potentially be utilized as reliable and useful prior information for source separation. The inclusion of auxiliary information in the separation scheme is referred to as informed source separation and is gaining increasing momentum in recent years (see e.g., among others, [5,15,18,20]). This paper further describes some ideas how to design the prior distributions for the present model to incorporate musically relevant information.

We henceforth denote the normal, Dirichlet and Poisson distributions by $\mathcal{N}$, Dir and Pois, respectively.

2. SPECTROGRAM MODEL OF MUSIC SIGNAL

2.1 Wavelet transform of source signal model

As in [8], this section derives the continuous wavelet transform of a source signal. Let us first consider as a signal model for the sound of the kth pitch the analytic signal representation of a pseudo-periodic signal given by

$$f_k(u) = \sum_{n=1}^{N} a_{k,n}(u) e^{\mathrm{j}(n\theta_k(u)+\varphi_{k,n})}, \tag{1}$$

where u denotes the time, $n\theta_k(u) + \varphi_{k,n}$ the instantaneous phase of the n-th harmonic and $a_{k,n}(u)$ the instantaneous amplitude. This signal model implicitly ensures not to violate the 'harmonicity' and 'coherent frequency modulation' constraints of the auditory grouping cues. Now, let the wavelet basis function be defined by

$$\psi_{\alpha,t}(u) = \frac{1}{\sqrt{2\pi}\alpha}\psi\left(\frac{u-t}{\alpha}\right), \tag{2}$$

where α is the scale parameter such that $\alpha > 0$, t the shift parameter and $\psi(u)$ the mother wavelet with the center frequency of 1 satisfying the admissibility condition. $\psi_{\alpha,t}(u)$ can thus be used to measure the component of period α at time t. The continuous wavelet transform of $f_k(u)$ is then defined by

$$W_k(\log\tfrac{1}{\alpha}, t) = \int_{-\infty}^{\infty} \sum_{n=1}^{N} a_{k,n}(u) e^{\mathrm{j}(n\theta_k(u)+\varphi_{k,n})} \psi^*_{\alpha,t}(u)\mathrm{d}u. \tag{3}$$

Since the dominant part of $\psi^*_{\alpha,t}(u)$ is typically localized around time t, the result of the integral in Eq. (3) shall depend only on the values of $\theta_k(u)$ and $a_{k,n}(u)$ near t. By taking this into account, we replace $\theta_k(t)$ and $a_{k,n}(t)$ with zero- and first-order approximations around time t:

$$a_{k,n}(u) \simeq a_{k,n}(t), \quad \theta_k(u) \simeq \theta_k(t) + \dot{\theta}_k(t)(u-t). \tag{4}$$

Note that the variable $\dot{\theta}_k(u)$ corresponds to the instantaneous fundamental frequency (F_0). By undertaking the above approximations, applying the Parseval's theorem, and putting $x = \log(1/\alpha)$ and $\Omega_k(t) = \log\dot{\theta}_k(t)$, we can further write Eq. (3) as

$$W_k(x,t) = \sum_{n=1}^{N} a_{k,n}(t)\Psi^*(ne^{-x+\Omega_k(t)}) e^{j(n\theta_k(t)+\varphi_{k,n})}, \tag{5}$$

where x denotes log-frequency and Ψ the Fourier transform of ψ. Since the function Ψ can be chosen arbitrarily, as with [8], we employ the following unimodal real function whose maximum is taken at $\omega = 1$:

$$\Psi(\omega) = \begin{cases} e^{-\frac{(\log\omega)^2}{4\sigma^2}} & (\omega > 0) \\ 0 & (\omega \leq 0) \end{cases}. \tag{6}$$

Eq. (5) can then be written as

$$W_k(x,t) = \sum_{n=1}^{N} a_{k,n}(t) e^{-\frac{(x-\Omega_k(t)-\log n)^2}{4\sigma^2}} e^{j(n\theta_k(t)+\varphi_{k,n})}. \tag{7}$$

If we now assume that the time-frequency components are sparsely distributed so that the partials rarely overlap each other, $|W_k(x,t)|^2$ is given approximately as

$$|W_k(x,t)|^2 \simeq \sum_{n=1}^{N} |a_{k,n}(t)|^2 e^{-\frac{(x-\Omega_k(t)-\log n)^2}{2\sigma^2}}. \tag{8}$$

This assumption means that the power spectra of the partials can approximately be considered additive. Note that a cutting plane of the spectrogram model given by Eq. (8) at time t is expressed as a harmonically-spaced Gaussian mixture function. If we assume the additivity of power spectra, the power spectrogram of a superposition of K pitched sounds is given by the sum of Eq. (8) over k. It should be noted that this model is identical to the one employed in the HTC approach [8].

Although we have defined the spectrogram model above in continuous time and continuous log-frequency, we actually obtain observed spectrograms as a discrete time-frequency representation through computer implementations. Thus, we henceforth use $Y_{l,m} := Y(x_l, t_m)$ to denote an observed spectrogram where x_l ($l = 1, \ldots, L$) and t_m ($m = 1, \ldots, M$) stand for the uniformly-quantized log-frequency points and time points, respectively. We will also use the notation $\Omega_{k,m}$ and $a_{k,n,m}$ to indicate $\Omega_k(t_m)$ and $a_{k,n}(t_m)$.

2.2 Incorporating source-filter model

The generating processes of many sound sources in real world can be explained fairly well by the source-filter theory. In this section, we follow the idea described in [12] to incorporate the source-filter model into the above model. Let us assume that each signal $f_k(u)$ within a short-time segment is an output of an all-pole system. That is, if we use $f_{k,m}[i]$ to denote the discrete-time representation of $f_k(u)$ within a short-time segment centered at time t_m, $f_{k,m}[i]$ can be described as

$$\beta_{k,m}[0]f_{k,m}[i] = \sum_{p=1}^{P} \beta_{k,m}[p] f_{k,m}[i-p] + \epsilon_{k,m}[i], \tag{9}$$

where i, $\epsilon_{k,m}[i]$, and $\beta_{k,m}[p]$ ($p = 0, \ldots, P$) denote the discrete-time index, an excitation signal, and the autoregressive (AR) coefficients, respectively. As we have already assumed in 2.1 that the F_0 of $f_{k,m}[i]$ is $e^{\Omega_{k,m}}$, to make the assumption consistent, the F_0 of the excitation signal $\epsilon_{k,m}[i]$ must also be $e^{\Omega_{k,m}}$. We thus define $\epsilon_{k,m}[i]$ as

$$\epsilon_{k,m}[i] = \sum_{n=1}^{N} v_{k,n,m} e^{jne^{\Omega_{k,m}} iu_0}, \tag{10}$$

where u_0 denotes the sampling period of the discrete-time representation and $v_{k,n,m}$ denotes the complex amplitude of the nth partial. By applying the discrete-time Fourier transform (DTFT) to Eq. (9) and putting $B_{k,m}(z) := \beta_{k,m}[0] - \beta_{k,m}[1]z^{-1} \cdots - \beta_{k,m}[P]z^{-P}$, we obtain

$$F_{k,m}(\omega) = \frac{\sqrt{2\pi}}{B_{k,m}(e^{j\omega})} \sum_{n=1}^{N} v_{k,n,m}\delta(\omega - ne^{\Omega_{k,m}}u_0), \tag{11}$$

where $F_{k,m}$ denotes the DTFT of $f_{k,m}$, ω the normalized angular frequency, and δ the Dirac delta function. The inverse DTFT of Eq. (11) gives us another expression of $f_{k,m}[i]$:

$$f_{k,m}[i] = \sum_{n=1}^{N} \frac{v_{k,n,m}}{B_{k,m}(e^{jne^{\Omega_{k,m}}u_0})} e^{jne^{\Omega_{k,m}} iu_0}. \tag{12}$$

By comparing Eq. (12) and the discrete-time representation of Eq. (1), we can associate the parameters of the source filter model defined above with the parameters introduced in 2.1 through the explicit relationship:

$$|a_{k,n,m}| = \left|\frac{v_{k,n,m}}{B_{k,m}(e^{jne^{\Omega_{k,m}}u_0})}\right|. \tag{13}$$

2.3 Constraining model parameters

The key assumption behind the NMF model is that the spectra of the sound of a particular pitch is expressed as a multiplication of time-independent and time-dependent factors. In order to extend the NMF model to a more reasonable one, we consider it important to clarify which factors involved in the spectra should be assumed to be time-dependent and which factors should not. For example, the F_0 must be assumed to vary in time during vibrato or portamento. Of course, the scale of the spectrum should also be assumed to be time-varying (as with the NMF model). On the other hand, the timbre of an instrument can be considered relatively static throughout an entire piece of music.

We can reflect these assumptions in the present model in the following way. For convenience of the following analysis, we factorize $|a_{k,n,m}|$ into the product of two variables, $w_{k,n,m}$ and $U_{k,m}$

$$|a_{k,n,m}| = w_{k,n,m}\sqrt{U_{k,m}}. \tag{14}$$

$w_{k,n,m}$ can be interpreted as the relative power of the nth harmonic and $U_{k,m}$ as the time-varying normalized amplitude of the sound of the kth pitch such that $\sum_{k,m} U_{k,m} = 1$. In the same way, let us put $v_{k,n,m}$ as

$$v_{k,n,m} = \tilde{w}_{k,n,m}\sqrt{U_{k,m}}. \tag{15}$$

Since the all-pole spectrum $1/|B_{k,m}(e^{j\omega})|^2$ is related to the timbre of the sound of the kth pitch, we want to constrain it to be time-invariant. This can be done simply by eliminating the subscript m. Eq. (13) can thus be rewritten as

$$w_{k,n,m} = \left|\frac{\tilde{w}_{k,n,m}}{B_k(e^{jne^{\Omega_{k,m}}u_0})}\right|. \tag{16}$$

We can use $\Omega_{k,m}$ as is, since it is already dependent on m.

To sum up, we obtain a spectrogram model $X_{l,m}$ as

$$X_{l,m} = \sum_{k=1}^{K} C_{k,l,m}, \quad C_{k,l,m} = \underbrace{\left(\sum_{n=1}^{N} w_{k,n,m}^2 e^{-\frac{(x_l-\Omega_{k,m}-\log n)^2}{2\sigma^2}}\right)}_{H_{k,l,m}} U_{k,m}, \tag{17}$$

where $C_{k,l,m}$ stands for the spectrogram of the kth pitch. If we denote the term inside the parenthesis by $H_{k,l,m}$, $X_{l,m}$ can be rewritten as $X_{l,m} = \sum_k H_{k,l,m}U_{k,m}$ and so the relation to the NMF model may become much clearer.

2.4 Formulating probabilistic model

Since the assumptions and approximations we made so far do not always hold exactly in reality, an observed spectrogram $Y_{l,m}$ may diverge from $X_{l,m}$ even though the parameters are optimally determined. One way to simplify the process by which this kind of deviation occurs would be to assume a probability distribution of $Y_{l,m}$ with the expected value of $X_{l,m}$. Here, we assume that $Y_{l,m}$ follows a Poisson distribution with mean $X_{l,m}$

$$Y_{l,m} \sim \mathrm{Pois}(Y_{l,m}; X_{l,m}), \tag{18}$$

where $\mathrm{Pois}(z;\xi) = \xi^z e^{-\xi}/\Gamma(z)$. This defines our likelihood function

$$p(\boldsymbol{Y}|\boldsymbol{\theta}) = \prod_{l,m} \mathrm{Pois}(Y_{l,m}; X_{l,m}), \tag{19}$$

where $\boldsymbol{Y}$ denotes the set consisting of $Y_{l,m}$ and $\boldsymbol{\Theta}$ the entire set consisting of the unknown model parameters. It should be noted that the maximization of the Poisson likelihood with respect to $X_{l,m}$ amounts to optimally fitting $X_{l,m}$ to $Y_{l,m}$ by using the I-divergence as the fitting criterion.

Eq. (16) implicitly defines the conditional distribution

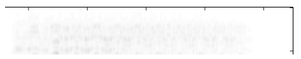


Figure 1. Power spectrogram of a violin vibrato sound.

$p(\boldsymbol{w}|\tilde{\boldsymbol{w}},\boldsymbol{\beta},\boldsymbol{\Omega})$ expressed by the Dirac delta function

$$p(\boldsymbol{w}|\tilde{\boldsymbol{w}},\boldsymbol{\beta},\boldsymbol{\Omega}) = \prod_{k,n,m} \delta\left(w_{k,n,m} - \left|\frac{\tilde{w}_{k,n,m}}{B_k(e^{jne^{\Omega_{k,m}u_0}})}\right|\right). \tag{20}$$

The conditional distribution $p(\boldsymbol{w}|\boldsymbol{\beta},\boldsymbol{\Omega})$ can thus be obtained by defining the distribution $p(\tilde{\boldsymbol{w}})$ and marginalizing over $\tilde{\boldsymbol{w}}$. If we now assume that the complex amplitude $\tilde{w}_{k,n,m}$ follows a circular complex normal distribution

$$\tilde{w}_{k,n,m} \sim \mathcal{N}_{\mathbb{C}}(\tilde{w}_{k,n,m}; 0, \nu^2), \; n = 1,\ldots,N, \tag{21}$$

where $\mathcal{N}_{\mathbb{C}}(z;0,\xi^2) = e^{-|z|^2/\xi^2}/(\pi\xi^2)$, we can show, as in [12], that $w_{k,n,m}$ follows a Rayleigh distribution:

$$w_{k,n,m} \sim \mathrm{Rayleigh}(w_{k,n,m}; \nu/|B_k(e^{jne^{\Omega_{k,m}}u_0})|), \tag{22}$$

where $\mathrm{Rayleigh}(z;\xi) = (z/\xi^2)e^{-z^2/(2\xi^2)}$. This defines the conditional distribution $p(\boldsymbol{w}|\boldsymbol{\beta},\boldsymbol{\Omega})$.

The F_0 of stringed and wind instruments often varies continuously over time with musical expressions such as vibrato. For example, the F_0 of a violin sound varies periodically around the note frequency during vibrato, as depicted in Fig. 1. Let us denote the standard log-F_0 corresponding to the kth note by μ_k. To appropriately describe the variability of an F_0 contour in both the global and local time scales, we design a prior distribution for $\boldsymbol{\Omega}_k := (\Omega_{k,1}, \Omega_{k,2}, \ldots, \Omega_{k,M})^\mathsf{T}$ by employing the product-of-experts (PoE) [6] concept using two probability distributions. First, we design a distribution $q_\mathrm{g}(\boldsymbol{\Omega}_k)$ describing how likely $\Omega_{k,1}, \ldots, \Omega_{k,L}$ stay near μ_k. Second, we design another distribution $q_\mathrm{l}(\boldsymbol{\Omega}_k)$ describing how likely $\Omega_{k,1}, \ldots, \Omega_{k,L}$ are locally continuous along time. Here we define $q_\mathrm{g}(\boldsymbol{\Omega}_k)$ and $q_\mathrm{l}(\boldsymbol{\Omega}_k)$ as

$$q_\mathrm{g}(\boldsymbol{\Omega}_k) = \mathcal{N}(\boldsymbol{\Omega}_k; \mu_k \boldsymbol{1}_M, \upsilon_k^2 \boldsymbol{I}_M), \tag{23}$$

$$q_\mathrm{l}(\boldsymbol{\Omega}_k) = \mathcal{N}(\boldsymbol{\Omega}_k; \boldsymbol{0}_M, \tau_k^2 D^{-1}), \tag{24}$$

$$D = \begin{bmatrix} 1 & -1 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -1 & 2 & -1 \\ 0 & \cdots & 0 & 0 & -1 & 1 \end{bmatrix}, \tag{25}$$

where $\boldsymbol{I}_M$ denotes an $M \times M$ identity matrix, D an $M \times M$ band matrix, $\boldsymbol{1}_M$ an M-dimensional all-one vector, and $\boldsymbol{0}_M$ an M-dimensional all-zero vector, respectively. υ_k denotes the standard deviation from mean μ_k, and τ_k the standard deviation of the F_0 jumps between adjacent frames. The prior distribution of $\boldsymbol{\Omega}_k$ is then derived as

$$p(\boldsymbol{\Omega}_k) \propto q_\mathrm{g}(\boldsymbol{\Omega}_k)^{\alpha_\mathrm{g}} q_\mathrm{l}(\boldsymbol{\Omega}_k)^{\alpha_\mathrm{l}} \tag{26}$$

where α_g and α_l are the hyperparameters that weigh the contributions of $q_g(\boldsymbol{\Omega}_k)$ and $q_1(\boldsymbol{\Omega}_k)$ to the prior distribution.

2.5 Relation to other models

It should be noted that the present model is related to other models proposed previously.

If we do not assume a parametric model for $H_{k,l,m}$ and treat each $H_{k,l,m}$ itself as the parameter, the spectrogram model $X_{l,m}$ can be seen as an NMF model with time-varying basis spectra, as in [14]. In addition to this assumption, if we assume that $H_{k,l,m}$ is time-invariant (i.e., $H_{k,l,m} = H_{k,l}$), $X_{l,m}$ reduces to the regular NMF model [19]. Furthermore, if we assume each basis spectrum to have a harmonic structure, $X_{l,m}$ becomes equivalent to the harmonic NMF model [16, 21].

If we assume that $\Omega_{k,m}$ is equal over time m, $X_{l,m}$ reduces to a model similar to the ones described in [17, 22]. Furthermore, if we describe $U_{k,m}$ using a parametric function of m, $X_{l,m}$ becomes equivalent to the HTC model [8, 10].

With a similar motivation, Hennequin *et al.* developed an extension to the NMF model defined in the short-time Fourier transform domain to allow the F_0 of each basis spectrum to be time-varying [4].

3. INCORPORATION OF AUXILIARY INFORMATION

3.1 Use of musically relevant information

We consider using side-information obtained with the state-of-the-art methods for MIR-related tasks including key detection, chord detection and beat tracking to assist source separation.

When multiple types of side-information are obtained for a specific parameter, we can combine the use of the mixture-of-experts and PoE [6] concepts according to the "AND" and "OR" conditions we design. For example, pitch occurrences typically depend on both the chord and key of a piece of music. Thus, when the chord and key information are obtained, we may use the product-of-experts concept to define a prior distribution for the parameters governing the likeliness of the occurrences of the pitches. In the next subsection, we describe specifically how to design the prior distributions.

3.2 Designing prior distributions

The likeliness of the pitch occurrences in popular and classical western music usually depend on the key or the chord used in that piece. The likeliness of the pitch occurrences can be described as a probability distribution over the relative energies of the sounds of the individual pitches.

Since the number of times each note is activated is usually limited, inducing sparsity to the temporal activation of each note event would facilitate the source separation. The likeliness of the number of times each note is activated can be described as well as a probability distribution over the temporal activations of the sound of each pitch.

To allow for designing such prior distributions, we decompose $U_{k,m}$ as the product of two variables: the pitch-wise relative energy $R_k = \sum_m U_{k,m}$ (i.e. $\sum_k R_k = 1$), and the pitch-wise normalized amplitude $A_{k,m} = U_{k,m}/R_k$ (i.e. $\sum_m A_{k,m} = 1$). Hence, we can write

$$U_{k,m} = R_k A_{k,m}. \tag{27}$$

This decomposition allows us to incorporate different kinds of prior information into our model by separately defining prior distributions over $\boldsymbol{R} = (R_1, \ldots, R_K)^\mathsf{T}$ and $\boldsymbol{A}_k = (A_{k,1}, \ldots, A_{k,M})^\mathsf{T}$. Here we introduce Dirichlet distributions:

$$\boldsymbol{A}_k \sim \mathrm{Dir}(\boldsymbol{A}_k; \boldsymbol{\gamma}_k^{(A)}), \quad \boldsymbol{R} \sim \mathrm{Dir}(\boldsymbol{R}; \boldsymbol{\gamma}^{(R)}), \tag{28}$$

where $\mathrm{Dir}(\boldsymbol{z}; \boldsymbol{\xi}) \propto \prod_i z_i^{\xi_i}$, $\boldsymbol{\gamma}_k^{(A)} := (\gamma_{k,1}^{(A)}, \ldots, \gamma_{k,M}^{(A)})^\mathsf{T}$, and $\boldsymbol{\gamma}^{(R)} := (\gamma_1^{(R)}, \ldots, \gamma_K^{(R)})^\mathsf{T}$. For $p(\boldsymbol{R})$, we set $\gamma_k^{(R)}$ at a reasonably high value if the kth pitch is contained in the scale and vice versa. For $p(\boldsymbol{A}_k)$, we set $\gamma_{k,m}^{(A)} < 1$ so that the Dirichlet distribution becomes a sparsity inducing distribution.

4. PARAMETER ESTIMATION ALGORITHM

Given an observed power spectrogram $\boldsymbol{Y} := \{Y_{l,m}\}_{l,m}$, we would like to find the estimates of $\boldsymbol{\Theta} := \{\boldsymbol{\Omega}, \boldsymbol{w}, \boldsymbol{\beta}, V, \boldsymbol{R}, \boldsymbol{A}\}$ that maximizes the posterior density $p(\boldsymbol{\Theta}|\boldsymbol{Y}) \propto p(\boldsymbol{Y}|\boldsymbol{\Theta})p(\boldsymbol{\Theta})$. We therefore consider the problem of maximizing

$$\mathcal{L}(\boldsymbol{\Theta}) := \ln p(\boldsymbol{Y}|\boldsymbol{\Theta}) + \ln p(\boldsymbol{\Theta}), \tag{29}$$

with respect to $\boldsymbol{\Theta}$ where

$$\ln p(\boldsymbol{Y}|\boldsymbol{\Theta}) =_c \sum_{l,m} (Y_{l,m} \ln X_{l,m} - X_{l,m}) \tag{30}$$

$$\begin{aligned} \ln p(\boldsymbol{\Theta}) &= \ln p(\boldsymbol{w}|\boldsymbol{\beta}, \boldsymbol{\Omega}) + \sum_k \ln p(\boldsymbol{\Omega}_k) \\ &\quad + \ln p(\boldsymbol{R}) + \sum_k \ln p(\boldsymbol{A}_k). \end{aligned} \tag{31}$$

$=_c$ denotes equality up to constant terms. Since the first term of Eq. (30) involves summation over k and n, analytically solving the current maximization problem is intractable. However, we can develop a computationally efficient algorithm for finding a locally optimal solution based on the auxiliary function concept, by using a similar idea described in [8, 12].

When applying an auxiliary function approach to a certain maximization problem, the first step is to define a lower bound function for the objective function. As mentioned earlier, the difficulty with the current maximization problem lies in the first term in Eq. (30) . By using the fact that the logarithm function is a concave function, we can invoke the Jensen's inequality

$$Y_{l,m} \ln X_{l,m} \geq Y_{l,m} \sum_{k,n} \lambda_{k,n,l,m} \ln \frac{w_{k,n,m}^2 e^{-\frac{(x_l - \Omega_{k,m} - \log n)^2}{2\sigma^2}} U_{k,m}}{\lambda_{k,n,l,m}}, \tag{32}$$

to obtain a lower bound function, where $\lambda_{k,n,l,m}$ is a positive variable that sums to unity: $\sum_{k,n} \lambda_{k,n,l,m} = 1$. Equality of (32) holds if and only if

$$\lambda_{k,n,l,m} = \frac{w_{k,n,m}^2 e^{-\frac{(x_l - \Omega_{k,m} - \log n)^2}{2\sigma^2}} U_{k,m}}{X_{l,m}}. \tag{33}$$

Although one may notice that the second term in Eq. (30) is nonlinear in $\Omega_{k,m}$, the summation of $X_{l,m}$ over l can be approximated fairly well using the integral $\int_{-\infty}^{\infty} X(x, t_m) \mathrm{d}x$, since $\sum_l X_{l,m}$ is the sum of the values at the sampled points $X(x_1, t_m), \ldots, X(x_L, t_m)$ with an equal interval, say Δ_x. Hence,

$$\begin{aligned} \sum_l X_{l,m} &\simeq \frac{1}{\Delta_x} \int_{-\infty}^{\infty} X(x, t_m) \mathrm{d}x \\ &= \frac{1}{\Delta_x} \sum_{k,n} w_{k,n,m}^2 U_{k,m} \int_{-\infty}^{\infty} e^{-\frac{(x - \Omega_{k,m} - \log n)^2}{2\sigma^2}} \mathrm{d}x \end{aligned}$$

$$= \frac{\sqrt{2\pi}\sigma}{\Delta_x}\sum_k U_{k,m}\sum_n w_{k,n,m}^2. \qquad (34)$$

This approximation implies that the second term in Eq. (30) depends little on $\Omega_{k,m}$.

An auxiliary function can thus be written as

$$\mathcal{L}^+(\Theta,\lambda) \underset{c}{=} \sum_{l,m} Y_{l,m}\sum_{k,n}\lambda_{k,n,l,m}\ln\frac{w^2_{k,n,m}e^{-\frac{(x_l-\Omega_{k,m}-\ln n)^2}{2\sigma^2}}U_{k,m}}{\lambda_{k,n,l,m}}$$

$$-\frac{\sqrt{2\pi}\sigma}{\Delta_x}\sum_m\sum_k U_{k,m}\sum_n w^2_{k,n,m} + \ln p(\Theta). \qquad (35)$$

We can derive update equations for the model parameters, using the above auxiliary function. By setting at zero the partial derivative of $\mathcal{L}^+(\boldsymbol{\Theta},\boldsymbol{\lambda})$ with respect to each of the model parameters, we obtain

$$w^2_{k,n,m} \leftarrow \frac{\sum_l Y_{l,m}\lambda_{k,n,l,m} + 1/2}{\sqrt{2\pi}R_kA_{k,m}\sigma/\Delta_x + \nu^2/(2|B_k(e^{jne^{\Omega_{k,m}}u_0})|^2)}, \qquad (36)$$

$$\boldsymbol{\Omega}_k \leftarrow \left(\frac{\alpha_1}{\tau^2}D + \frac{\alpha_{\mathrm{g}}}{\upsilon_k^2}\boldsymbol{I}_M + \sum_{n,l}\mathrm{diag}(\boldsymbol{p}_{k,n,l})\right)^{-1}$$

$$\times\left(\mu_k\frac{\alpha_{\mathrm{g}}}{\upsilon_k^2}\mathbf{1}_M + \sum_{n,l}(x_l-\ln n)\boldsymbol{p}_{k,n,l}\right), \qquad (37)$$

$$R_k \propto \frac{\sum_{l,m}Y_{l,m}\sum_n\lambda_{k,n,l,m} + \gamma_k^{(R)} - 1}{\sum_{m,n}A_{k,m}w^2_{k,m,n}}, \qquad (38)$$

$$A_{k,m} \propto \frac{\sum_l Y_{l,m}\sum_n\lambda_{k,n,l,m} + \gamma_{k,m}^{(A)} - 1}{R_k\sum_n w^2_{k,m,n}}, \qquad (39)$$

$$\boldsymbol{p}_{k,n,l} := \frac{1}{\sigma^2}\left[Y_{l,1}\lambda_{k,n,l,1}, Y_{l,2}\lambda_{k,n,l,2},\cdots,Y_{l,M}\lambda_{k,n,l,M}\right]^{\top}, \qquad (40)$$

where diag($\boldsymbol{p}$) converts a vector $\boldsymbol{p}$ into a diagonal matrix with the elements of $\boldsymbol{p}$ on the main diagonal.

As for the update equations for the AR coefficients $\boldsymbol{\beta}$, we can invoke the method described in [23] with a slight modification, since the terms in the auxiliary function that depend on $\boldsymbol{\beta}$ has the similar form as the objective function defined in [23]. It can be shown that $\mathcal{L}^+$ can be increased by the following updates (the details are omitted owing to space limitations):

$$\boldsymbol{h}_k \leftarrow \hat{C}_k(\boldsymbol{\beta}_k)\boldsymbol{\beta}_k, \quad \boldsymbol{\beta}_k \leftarrow C_k^{-1}\boldsymbol{h}_k, \qquad (41)$$

where C_k and $\hat{C}_k(\boldsymbol{\beta}_k)$ are $(P+1)\times(P+1)$ Toeplitz matrices, whose (p,q)-th elements are

$$C_{k,p,q} = \frac{1}{MN}\sum_{m,n}\frac{w^2_{k,m,n}}{2\nu}\cos[(p-q)ne^{\Omega_{k,m}}u_0],$$

$$\hat{C}_{k,p,q}(\boldsymbol{\beta}_k) = \frac{1}{MN}\sum_{m,n}\frac{1}{|B_k(e^{jne^{\Omega_{k,m}}u_0})|^2}\cos[(p-q)ne^{\Omega_{k,m}}u_0].$$

$$(42)$$

5. EXPERIMENTS

In the following preliminary experiments, we simplified HTFD by omitting the source filter model and assuming the time-invariance of $w_{k,m,n}$.

5.1 F_0 tracking of violin sound

To confirm whether HTFD can track the F_0 contour of a sound, we compared HTFD with NMF with the I-divergence, by using a 16 kHz-sampled audio signal which were artificially made by mixing D♭4, F4 and A♭4 violin vibrato sounds from the RWC instrument database [3]. In this paper, the F_0 of the pitch name A4 was set at 440 Hz. The power spectrogram of the mixed signal is shown in Fig. 2. To convert the signal into a spectrogram, we employed the fast approximate continuous wavelet transform [9] with a 16 ms time-shift interval. $\{x_l\}_l$ ranged 55 to 7040 Hz per 10 cent. The parameters of HTFD were set at $\boldsymbol{\gamma}_k^{(A)} = (1-3.96\times10^{-6})\mathbf{1}_I$, $(\tau_k,\nu_k) = (0.83, 1.25)$ for all k, $(N,K,\sigma,\alpha_g,\alpha_s) = (8, 73, 0.02, 1, 1)$, and $\boldsymbol{\gamma}^{(R)} = (1-2.4\times10^{-3})\mathbf{1}_K$. $\{\mu_k\}_k$ ranged A1 to A♯7 with a chromatic interval, i.e. $\mu_k = \ln(55) + \ln(2)\times(k-1)/12$. The number of NMF bases were set at three. The parameter updates of both HTFD and NMF were stopped at 100 iterations.

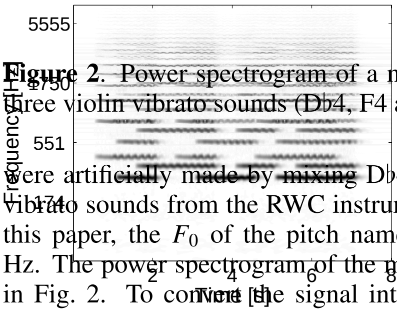


Figure 2. Power spectrogram of a mixed audio signal of three violin vibrato sounds (D♭4, F4 and A♭4).

While the estimates of spectrograms obtained with NMF were flat and the vibrato spectra seemed to be averaged (Fig. 3 (a)), those obtained with HTFD tracked the F_0 contours of the vibrato sounds appropriately (Fig. 3 (b)), and clear vibrato sounds were contained in the separated audio signals by HTFD.

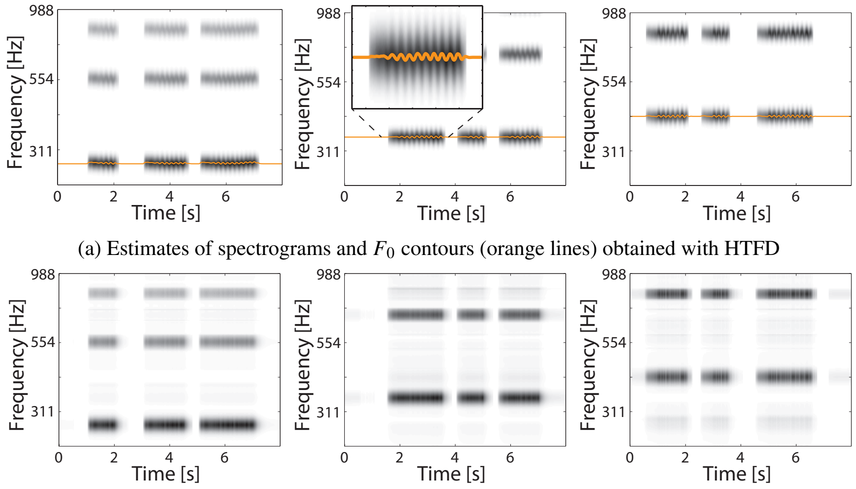
5.2 Separation using key information

We next examined whether the prior information of a sound improve source separation accuracy. The key of the sound used in 5.1, was assumed as D♭ major. The key information was incorporated in the estimation scheme by setting $\gamma_k^{(R)} = 1 - 2.4\times10^{-3}$ for the pitch indices that are not contained in the D♭ major scale and $\gamma_k^{(R)} = 1-3.0\times10^{-3}$ for the pitch indices contained in that scale. The other conditions were the same as 5.1.

With HTFD without using the key information, the estimated activations of the pitch indices that were not contained in the scale, in particular D4, were high as illustrated in Fig. 4 (a). In contrast, those estimated activations with HTFD using the key information were suppressed as shown in Fig. 4 (b). These results thus support strongly that incorporating prior information improve the source separation accuracy.

5.3 Transposing from one key to another

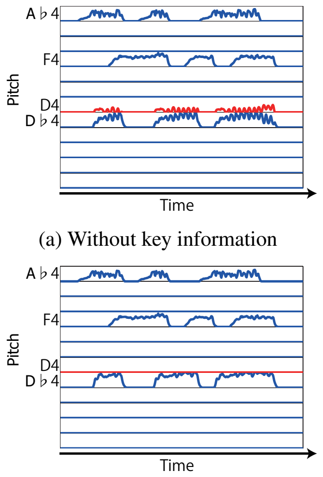
Here we show some results of an experiment on automatic key transposition [11] using HTFD. The aim of key transposition is to change the key of a musical piece to another key. We separated the spectrogram of a polyphonic sound into spectrograms of individual pitches using HFTD, transposed the pitches of the subset of the separated components, added all the spectrograms together to construct a pitch-modified polyphonic spectrogram, and constructed a



(a) Estimates of spectrograms and F_0 contours (orange lines) obtained with HTFD

(b) Estimates of spectrograms obtained with NMF

Figure 3. Estimated spectrogram models by harmonic-temporal factor decomposition (HTFD) and non-negative matrix factorization (NMF). In left-to-right fashion, the spectrogram models are for D♭4, F4 and A♭4.



(a) Without key information

(b) With key information

Figure 4. Temporal activations of A3–A♭4 estimated with HTFD using and without using prior information of the key. The red curves represent the temporal activations of D4.

time-domain signal from the modified spectrogram using the method described in [13]. For the key transposition, we adopted a simple way: To transpose, for example, from A *major* scale to A *natural minor* scale, we changed the pitches of the separated spectrograms corresponding to C♯, F♯ and G♯ to C, F and G, respectively.

Some results are demonstrated in `http://hil.t.u-tokyo.ac.jp/~nakamura/demo/HTFD.html`.

6. CONCLUSION

This paper proposed a new approach for monaural source separation called the "Harmonic-Temporal Factor Decomposition (HTFD)" by introducing a spectrogram model that combines the features of the models employed in the NMF and HTC approaches. We further described some ideas how to design the prior distributions for the present model to incorporate musically relevant information into the separation scheme.

7. ACKNOWLEDGEMENTS

This work was partly supported by JSPS Grant-in-Aid for Young Scientists B Grant Number 26730100 and the Tateishi Science and Technology Foundation.

8. REFERENCES

[1] A. S. Bregman: *Auditory Scene Analysis*, MIT Press, Cambridge, 1990.

[2] J. S. Downie, D. Byrd, and T. Crawford: "Ten years of ISMIR: Reflections on challenges and opportunities," *Proc. ISMIR*, pp. 13–18, 2009.

[3] M. Goto: "Development of the RWC Music Database," *Proc. ICA*, pp. l–553–556, 2004.

[4] R. Hennequin, R. Badeau, and B. David: "Time-dependent parametric and harmonic templates in non-negative matrix factorization," *Proc. DAFx*, pp. 246–253, 2010.

[5] R. Hennequin, B. David, and R. Badeau: "Score informed audio source separation using a parametric model of non-negative spectrogram," *Proc. ICASSP*, pp. 45–48, 2011.

[6] G. E. Hinton: "Training products of experts by minimizing contrastive divergence," *Neural Comput.*, vol. 14, no. 8, pp. 1771–1800, 2002.

[7] G. Hu, and D. L. Wang: "An auditory scene analysis approach to monaural speech segregation," *Topics in Acoust. Echo and Noise Contr.*, pp. 485–515, 2006.

[8] H. Kameoka: *Statistical Approach to Multipitch Analysis*, PhD thesis, The University of Tokyo, Mar. 2007.

[9] H. Kameoka, T. Tabaru, T. Nishimoto, and S. Sagayama: *(Patent) Signal processing method and unit*, in Japanese, Nov. 2008.

[10] H. Kameoka, T. Nishimoto, and S. Sagayama: "A multipitch analyzer based on harmonic temporal structured clustering," *IEEE Trans. ASLP*, vol. 15, no. 3, pp. 982–994, 2007.

[11] H. Kameoka, J. Le Roux, Y. Ohishi, and K. Kashino: "Music Factorizer: A note-by-note editing interface for music waveforms," *IPSJ SIG Tech. Rep.,* 2009-MUS-81-9, in Japanese, Jul. 2009.

[12] H. Kameoka: "Statistical speech spectrum model incorporating all-pole vocal tract model and F_0 contour generating process model," *IEICE Tech. Rep.*, vol. 110, no. 297, SP2010-74, pp. 29–34, in Japanese, Nov. 2010.

[13] T. Nakamura and H. Kameoka: "Fast signal reconstruction from magnitude spectrogram of continuous wavelet transform based on spectrogram consistency," *Proc. DAFx*, 40, to appear, 2014.

[14] M. Nakano, J. Le Roux, H. Kameoka, Y. Kitano, N. Ono, and S. Sagayama: "Nonnegative matrix factorization with Markov-chained bases for modeling time-varying patterns in music spectrograms," *Proc. LVA/ICA*, pp. 149–156, 2010.

[15] A. Ozerov, C. Févotte, R. Blouet, and J. L. Durrieu: "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," *Proc. ICASSP.*, pp. 257–260, 2011.

[16] S. A. Raczyński, N. Ono, and S. Sagayama: "Multipitch analysis with harmonic nonnegative matrix approximation," *Proc. ISMIR*, pp. 381–386, 2007.

[17] D. Sakaue, T. Otsuka, K. Itoyama, and H. G. Okuno: "Bayesian nonnegative harmonic-temporal factorization and its application to multipitch analysis," *Proc. ISMIR*, pp. 91–96, 2012.

[18] U. Simsekli and A. T. Cemgil: "Score guided musical source separation using generalized coupled tensor factorization," *Proc. EUSIPCO*, pp. 2639–2643, 2012.

[19] P. Smaragdis and J. C. Brown: "Non-negative matrix factorization for polyphonic music transcription," *Proc. WASPAA*, pp. 177–180, 2003.

[20] P. Smaragdis and G. J. Mysore: "Separation by "humming": User-guided sound extraction from monophonic mixtures," *Proc. WASPAA*, pp. 69–72, 2009.

[21] E. Vincent, N. Bertin, and R. Badeau: "Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription," *Proc. ICASSP*, pp. 109–112, 2008.

[22] K. Yoshii and M. Goto: "Infinite latent harmonic allocation: A nonparametric Bayesian approach to multipitch analysis," *Proc. ISMIR*, pp. 309–314, 2010.

[23] A. El-Jaroudi, J. Makhoul: "Discrete all-pole modeling," *IEEE Trans. SP*, vol. 39, no. 2, pp. 411–423, 1991.