Harmonic-Temporal Factor Decomposition for Unsupervised Monaural Separation of Harmonic Sounds

Tomohiko Nakamura, *Member, IEEE*, and Hirokazu Kameoka, *Senior Member, IEEE*

***Abstract*—We address the problem of separating a monaural mixture of harmonic sounds into the audio signals of individual semitones in an unsupervised manner. Unsupervised monaural audio source separation has thus far been mainly addressed by two approaches: one rooted in computational auditory scene analysis (CASA) and the other based on non-negative matrix factorization (NMF). These approaches focus on different clues for making source separation possible. A CASA-based method called harmonic-temporal clustering (HTC) focuses on a local time-frequency structure of individual sources, whereas NMF focuses on a global time-frequency structure of music spectrograms. These clues do not conflict with each other and can be used to achieve a more reliable audio source separation algorithm. Hence, we propose a monaural audio source separation framework, harmonic-temporal factor decomposition (HTFD), by developing a spectrogram model that encompasses the features of the models used in the NMF and HTC approaches. We further incorporate a source-filter model to build an extension of HTFD, source-filter HTFD (SF-HTFD). We derive efficient parameter estimation algorithms of HTFD and SF-HTFD based on the auxiliary function principle. We show, through music source separation experiments, the efficacy of HTFD and SF-HTFD compared with conventional methods. Furthermore, we demonstrate the effectiveness of HTFD and SF-HTFD for automatic musical key transposition.**

***Index Terms*—Computational auditory scene analysis, harmonic-temporal clustering, monaural audio source separation, non-negative matrix factorization.**

I. Introduction

AUDIO source separation, a technique of separating a mixture audio signal into individual source signals, has a wide variety of applications, including automatic music transcription and music editing/remixing. Audio source separation has thus far been tackled by many researchers but is still challenging since the problem is inherently ill-posed with infinitely many possible solutions if there is no prior information available. While spatial clues can be useful for multi-channel inputs, they are usually unavailable when it comes to monaural inputs.

In supervised settings, deep neural networks (DNNs) have shown promising performance when a sufficient number of training examples of individual source signals are available [1]–[5]. However, this is not always the case. For example, since the audio signal of each musical note in a musical piece is not always easily accessible, supervised approaches may perform unsatisfactorily in such tasks as notewise audio source separation. One solution would be to build databases large enough for individual tasks, but this can be a costly and painstaking process. Another possible way would be to adopt an unsupervised approach.

There are two major approaches to unsupervised monaural audio source separation. One is based on computational auditory scene analysis, e.g., [6]–[8]. The aim of this approach is to implement the process of grouping spectrogram-like elements that are likely to originate from the same auditory stream based on a set of auditory grouping cues, described for instance in [9]. One attempt that has been made to imitate this process is harmonic-temporal clustering (HTC) [7], [8]. HTC makes it possible to cluster time-frequency components originating from the same audio stream based on a constraint designed according to auditory grouping cues, such as harmonicity and the coherence and continuity of amplitude and frequency modulations.

The other approach is based on non-negative matrix factorization (NMF) [10]. The core idea is to interpret an observed spectrogram as a non-negative matrix and assume that each spectrum can be represented as a sum of a limited number of spectral templates scaled by time-varying amplitudes. This assumption amounts to approximating the observed spectrogram by a product of two non-negative matrices: one containing a different spectral template in each column and the other containing the associated time-varying amplitudes in each row. Thus, factorizing the observed spectrogram into the two non-negative matrices amounts to estimating unknown spectral templates and mixing weights that best explain the observed spectra. This approach has been used with notable success particularly for music source separation. One reason for this is that a musical piece typically consists of a limited number of recurring note events, so similar spectral patterns frequently appear in a music spectrogram.

Manuscript received April 2, 2020; revised August 28, 2020 and October 5, 2020; accepted October 17, 2020. Date of publication November 16, 2020; date of current version December 14, 2020. This work was supported in part by JST CREST under Grant JPMJCR19A3 and in part by JSPS KAKENHI under Grants JP26730100, JP15J09992, and JP20K19818. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Stefan Bilbao. *(Corresponding author: Tomohiko Nakamura.)*

Tomohiko Nakamura is with the Graduate School of Information Science, and Technology, University of Tokyo, Tokyo 113-8656, Japan (e-mail: tomohiko.nakamura.jp@ieee.org).

Hirokazu Kameoka is with the NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, Kanagawa 243-0198, Japan (e-mail: hirokazu.kameoka.uh@hco.ntt.co.jp).

Digital Object Identifier 10.1109/TASLP.2020.3037487

The above two approaches use different clues for making audio source separation possible. Roughly speaking, the former approach focuses on a local time-frequency structure of individual sources, whereas the latter approach focuses on a relatively global time-frequency structure of music spectrograms. Since these clues do not conflict with each other, we believe that they can be useful in achieving a more reliable audio source separation algorithm.

In this paper, by developing a spectrogram model that encompasses the features of the models used in the HTC and NMF approaches, we propose a monaural audio source separation framework, which we call harmonic-temporal factor decomposition (HTFD). As in the HTC approach, we model the spectrogram of each source in the continuous wavelet transform (CWT) domain and describe the spectral shape and the fundamental frequency (F_0) contour using an independent set of parameters. The spectral shape is characterized by a set of the relative magnitudes of harmonics scaled by a time-varying amplitude in a way similar to the NMF approach. Our model allows us to describe continuous spectral variations caused by F_0 changes reasonably well and handle musically expressive sounds, such as slur, vibrato and portamento sounds, many of which cannot be expressed very satisfactorily by the regular NMF model.

We further extend the HTFD spectrogram model by incorporating a source-filter model, which can describe the generating processes of many sound sources fairly well. We call the extension source-filter HTFD (SF-HTFD). The source-filter model describes an audio signal as the output of a linear filter excited by an excitation input. As for musical instrument sounds, the excitation signal expresses a vibrating object such as a violin string and the linear filter refers to the resonance structure of the instrument. The excitation signal and filter characterize the pitch and timbre of an instrument sound, respectively.

Several attempts have already been made to incorporate the source-filter model into NMF to enhance the performance of audio source separation and multipitch analysis [11]–[18]. In these studies, a spectrum of each source is simply represented as the product of excitation and filter spectra. We call this representation the excitation-filter product representation to distinguish it from the one we introduce in this paper. If there is no frame overlap and the filter length is significantly shorter than the frame length, this representation can be justified in the short-time Fourier transform (STFT) domain owing to the orthogonality of the Fourier transform. However, since the CWT is not an orthogonal transform, using the excitation-filter product representation in the CWT domain is not well justified.

To overcome this issue, we derive an explicit parameter relationship between the HTFD spectrogram model and a source-filter model defined in the discrete time domain, following the idea described in [19]. This relationship allows us not only to model spectral changes associated with pitch and timbre separately but also to reveal the underlying meanings of the excitation-filter product representation in the CWT domain.

We employ a generative modeling approach for HTFD and SF-HTFD and design reasonable prior distributions for F_0s, spectral envelopes, and temporal activations of sources. We then derive parameter estimation algorithms of HTFD and SF-HTFD based on the auxiliary function principle [8], [20], [21]. The contributions of this paper are summarized as follows:

- We propose a monaural audio source separation framework, HTFD, by developing a CWT spectrogram model that offers the features of the HTC and NMF spectrogram models concurrently.
- We further propose a source-filter extension of HTFD, SF-HTFD, by deriving an explicit parameter relationship between the HTFD spectrogram model and the source-filter model defined in the discrete time domain.
- We derive parameter estimation algorithms of HTFD and SF-HTFD based on the auxiliary function principle.
- We reveal the meaning of assuming the excitation-filter product representation in the CWT domain.
- We show the efficacy of HTFD and SF-HTFD through unsupervised music source separation experiments and demonstrate the usefulness of the proposed frameworks for automatic musical key transposition.

The rest of this paper is organized as follows. We derive the HTC spectrogram model and present a generative model of HTFD in Section II. We design prior distributions of the parameters of the HTFD generative model in Section III, and we derive a parameter estimation algorithm based on the auxiliary function principle in Section IV. We extend HTFD to SF-HTFD by deriving an explicit parameter relationship between the HTFD spectrogram model and the source-filter model in Section V, and we present a parameter estimation algorithm of SF-HTFD in Section VI. In Section VII, we review related studies in comparison with HTFD and SF-HTFD. Through experiments on unsupervised monaural separation of harmonic sounds, we show the efficacy of HTFD and SF-HTFD and that they can be used to devise an automatic musical key transposition system in Section VIII. We finally conclude this paper in Section IX. It should be noted that the present model is designed for harmonic sounds and cannot be applied to percussive sounds. Hence, in this paper, we focus on harmonic sounds only.

Note that this paper is partially based on an international conference paper written by the authors [22]. This paper has five additional contributions compared with the conference paper. (i) We reformulate a SF-HTFD model for enhancing the model expressivity and (ii) present a novel parameter optimization algorithm based on variational approximation and the auxiliary function principle. (iii) We have conducted systematic quantitative experiments and analyses of HTFD and SF-HTFD, while only a qualitative experiment was reported in the conference paper. (iv) We present HTFD and an efficient parameter estimation algorithm. In the work described in the conference paper, although we used HTFD in the experiment, we omitted the definition and derivation of the HTFD model and algorithm. (v) We define the proposed models in the magnitude spectrogram domain instead of the power spectrogram since we have found it to enhance the separation performance.

II. Spectrogram Model of HTFD

A. Continuous Wavelet Transform of Source Signal Model

In this section, we introduce the CWT of a source signal model as in [8]. Let K be the number of sources and $k = 1, \ldots, K$ be the source index. Suppose that a sound of source k can be represented as an analytic signal representation of a pseudo-periodic signal with N harmonic partials, which is defined in the continuous time domain, $u \in \mathbb{R}$:

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

where j is the imaginary unit and $n = 1, \ldots, N$ is the partial index. Here $d_{k,n}(u) \geq 0$ and $n\theta_k(u) + \varphi_{k,n} \in \mathbb{R}$ denote the instantaneous amplitude and phase of partial n. This signal model implicitly ensures that *harmonicity* and *coherent frequency modulation* constraints of the auditory grouping cues are not violated.

Let $x \in \mathbb{R}$ be the logarithm of the angular frequency and $t \in \mathbb{R}$ be continuous time. The CWT of a time-domain signal is given by the inner products of the signal and wavelet bases, which are determined by an analyzing wavelet function $\xi(u) \in \mathbb{C}$ that satisfies the admissible condition. As with [8], by using the analyzing wavelet function whose Fourier transform is defined by

$$\Xi(\omega) = \begin{cases} e^{-\frac{(\ln \omega)^2}{2\sigma^2}} & (\omega > 0) \\ 0 & (\omega \leq 0) \end{cases}, \tag{2}$$

where $\omega \in \mathbb{R}$ is the angular frequency, we can approximately describe the CWT of $s_k(u)$ as

$$W_k(x,t) = \sum_{n=1}^{N} d_{k,n}(t) e^{-\frac{(x-\Omega_k(t)-\ln n)^2}{2\sigma^2}} e^{j(n\theta_k(t)+\varphi_{k,n})}. \tag{3}$$

Here, $\Omega_k(t)$ denotes the logarithm of the time derivative of $\theta_k(t)$, i.e., the logarithmic F_0. The detailed derivation of Eq. (3) is written in Appendix A.

Assuming that the partials rarely overlap each other, the magnitude of $W_k(x,t)$ is approximately given as

$$|W_k(x,t)| \simeq \sum_{n=1}^{N} d_{k,n}(t) G_{k,n}(t), \tag{4}$$

where

$$G_{k,n}(t) := e^{-\frac{(x-\Omega_k(t)-\ln n)^2}{2\sigma^2}}. \tag{5}$$

This assumption implies that the magnitude spectra of partials can be approximately additive. A time slice of the spectrogram model given by Eq. (4) at time t is expressed as a harmonically spaced Gaussian mixture function as shown in Fig. 1. Note that the spectrogram model is essentially identical to the one used in the HTC approach [8].

Although we have thus far defined the spectrogram model in the continuous time and continuous log-frequency domain, observed spectrograms are actually given in the discrete time and discrete log-frequency domain through computer implementations. Let us define L uniformly quantized log-frequency points and M uniformly quantized time points as $\{x_l\}_{l=1}^{L}$ and $\{t_m\}_{m=1}^{M}$, respectively. Here, $l = 1, \ldots, L$ and $m = 1, \ldots, M$ are the indices of the quantized log-frequency and time. To simplify notation, we denote $G_{k,n}(t_m)$, $\Omega_k(t_m)$, and $d_{k,n}(t_m)$ by $G_{k,n,m}$, $\Omega_{k,m}$, and $d_{k,n,m}$, respectively.

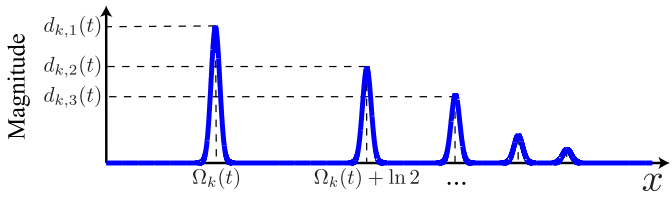


Fig. 1. Illustration of HTC model at time t.

B. Parameter Decomposition Into Time-Dependent and Time-Independent Factors

In this section, we incorporate the feature of the NMF approach into the spectrogram model defined in Eq. (4). The key assumption of NMF is that source spectra are decomposed into time-dependent and time-independent factors: the scales of the spectra and the spectral shapes, respectively. In order to extend the NMF model to develop 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_0s of string and wind instrument sounds, particularly with musical expressions such as vibrato and portamento, frequently varies continuously. To accurately capture these sounds, the parameters associated with F_0s should be time-dependent. As with NMF, the scale of the spectra should also be time-dependent, whereas the timbre of musical instruments can be considered relatively static.

To incorporate these assumptions, we factorize $d_{k,n,m}$ into a product of a time-dependent factor $U_{k,m} \geq 0$ and a time-independent factor $w_{k,n} \geq 0$:

$$d_{k,n,m} = w_{k,n} U_{k,m}. \tag{6}$$

The parameter $U_{k,m}$ represents the temporal activation of source k at frame m, and the parameter $w_{k,n}$ represents the normalized relative magnitude of partial n of source k such that $\sum_n w_{k,n} = 1$ for all k. Substituting Eq. (6) into Eq. (4) and assuming the additivity of magnitude spectra as with NMF and HTC, we can obtain an observed spectrogram model $X_{l,m}$ as the sum of K source spectrogram models, $S_{k,l,m}$:

$$X_{l,m} = \sum_k S_{k,l,m}, \quad S_{k,l,m} = \underbrace{\left(\sum_n w_{k,n} G_{k,n,m} \right)}_{H_{k,l,m}} U_{k,m}. \tag{7}$$

If we denote the term in the parenthesis of Eq. (8) by $H_{k,l,m}$, $X_{l,m}$ is rewritten as $X_{l,m} = \sum_k H_{k,l,m} U_{k,m}$, which makes the relation to NMF clear.

C. Probability Distribution of Observed Spectrogram

An observed spectrogram $Y_{l,m}$, in reality, may deviate from the assumptions and approximations we have made thus far. One possible way of handling this deviation would be to use the idea of probabilistic generative modeling. We assume $Y_{l,m}$ follows a Poisson distribution with mean $X_{l,m}$:

$$Y_{l,m} \sim \mathrm{Poisson}(Y_{l,m}; X_{l,m}). \tag{8}$$

Note that the maximum likelihood estimation of the Poisson distribution amounts to fitting $X_{l,m}$ to $Y_{l,m}$ with the generalized Kullback–Leibler divergence. The choice of the Poisson distribution allows us to derive a fast-converging parameter optimization algorithm as we will show in Section IV.

D. Relation to NMF and HTC Models

Our spectrogram model can be reduced to several variants of NMF and HTC models. If we treat each $H_{k,l,m}$ as a free parameter and assume it to vary in time according to a Markov chain, the spectrogram model $X_{l,m}$ can be seen as an NMF model with time-varying basis spectra as in [23]. By treating each $H_{k,l,m}$ as a free parameter and assuming it to be time-invariant, $X_{l,m}$ reduces to the regular NMF model [10]. If we further assume each basis spectrum to have a harmonic structure, $X_{l,m}$ becomes equivalent to the harmonic NMF (HNMF) model [24], [25].

If $\Omega_{k,m}$ is assumed to be time-invariant, $X_{l,m}$ reduces to a model similar to the ones described in [26], [27]. By further describing $U_{k,m}$ with a parametric function of m, $X_{l,m}$ becomes equivalent to the HTC model [7], [8].

III. Design of Prior Distributions of HTFD

A. Prior Distribution of $\Omega_{k,m}$

The F_0s of string and wind instrument sounds frequently varies in time continuously during slurs, vibrato, and portamento. For example, during vibrato, the F_0s of a violin sound vary periodically around the standard F_0 of the performed note. The F_0s of these sounds tend to be located around the standard F_0s of the corresponding notes globally, whereas the F_0 contours of the sounds smoothly vary in time locally. These global and local properties can be simultaneously incorporated by designing a prior distribution of $\boldsymbol{\Omega}_k = [\Omega_{k,1}, \ldots, \Omega_{k,M}]^\mathsf{T}$ based on the product-of-experts (PoE) concept [28].

Let μ_k denote the logarithm of the standard F_0 associated with source k. To describe how likely $\Omega_{k,m}$ is to be located near μ_k globally, we design a probability distribution $P_\mathrm{g}(\boldsymbol{\Omega}_k; \mu_k, \upsilon_k^2)$ as a multivariate normal distribution with mean $\mu_k \mathbf{1}_M$ and covariance $\upsilon_k^2 I_M$:

$$P_\mathrm{g}(\boldsymbol{\Omega}_k; \mu_k, \upsilon_k^2) = \mathrm{Normal}(\boldsymbol{\Omega}_k; \mu_k \mathbf{1}_M, \upsilon_k^2 I_M), \tag{9}$$

where $\mathbf{1}_M$ is the M-dimensional all-one vector and I_M is the $M \times M$ identity matrix. To describe how likely $\boldsymbol{\Omega}_k$ is to be locally smooth along time, we design a probability distribution $P_\mathrm{l}(\boldsymbol{\Omega}_k; \tau_k^2)$ as

$$P_\mathrm{l}(\boldsymbol{\Omega}_k; \tau_k^2) = \mathrm{Normal}(\boldsymbol{\Omega_k}; \mathbf{0}_M, \tau_k^2 (D^\mathsf{T} D)^{-1}), \tag{10}$$

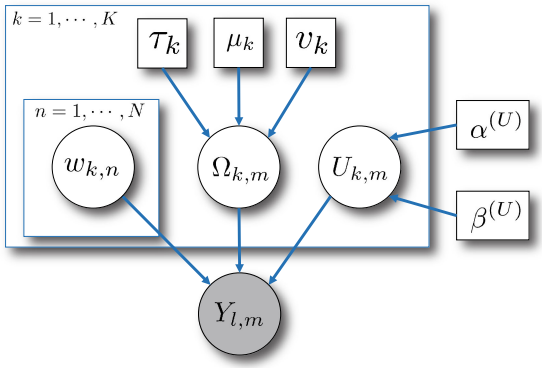


Fig. 2. Plate notation of HTFD generative model.

where τ_k is the standard deviation of time differences of F_0s and $\mathbf{0}_M$ is an M-dimensional all-zero vector. Here, D is an $(M-1) \times M$ band matrix given by

$$D = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & -1 \end{bmatrix}. \tag{11}$$

Following the PoE concept, the prior distribution of $\boldsymbol{\Omega}_k$ can then be defined by

$$P(\boldsymbol{\Omega}_k; \mu_k, \upsilon_k^2, \tau_k^2) \propto (P_\mathrm{g}(\boldsymbol{\Omega}_k; \mu_k, \upsilon_k^2))^{\alpha_\mathrm{g}} (P_\mathrm{l}(\boldsymbol{\Omega}_k; \tau_k^2))^{\alpha_\mathrm{l}}, \tag{12}$$

where α_g and α_l are the hyperparameters controlling the contributions of $P_\mathrm{g}(\boldsymbol{\Omega}_k; \mu_k, \upsilon_k^2)$ and $P_\mathrm{l}(\boldsymbol{\Omega}_k; \tau_k^2)$ to the prior distribution, respectively.

B. Prior Distribution of $U_{k,m}$

In popular and classical western music, the number of times each musical note is activated is usually limited. Thus, we can assume sparsity of the temporal activations. To promote sparseness, we employ a Gamma distribution as a prior distribution on U:

$$U_{k,m} \sim \mathrm{Gamma}(U_{k,m}; \alpha^{(U)}, \beta^{(U)}), \tag{13}$$

where $\alpha^{(U)} > 0$ and $\beta^{(U)} > 0$ are the shape and rate parameters. This prior distribution promotes sparsity when $\alpha^{(U)} < \beta^{(U)}$. In summary, the overall generative model is depicted in plate notation in Fig. 2.

IV. Parameter Estimation Algorithm of HTFD

A. Maximum a Posterior Estimation Problem

The parameters of interest in our model are

$w = \{w_{k,n}\}_{k,n}$: relative magnitude of partial n,

$U = \{U_{k,m}\}_{k,m}$: temporal activation,

$\Omega = \{\Omega_{k,m}\}_{k,m}$: F_0 of source k at frame m.

We denote the entire set of the above parameters as Θ. Given an observed magnitude spectrogram $Y = \{Y_{l,m}\}_{l,m}$, we wish to

find the estimate of Θ that maximizes the posterior $P(\Theta|Y) \propto P(Y|\Theta)P(\Theta)$, or equivalently,

$$\begin{aligned}\ln P(Y|\Theta) + \ln P(\Theta) &\underset{c}{=} \sum_{l,m} Y_{l,m} \ln X_{l,m} - \sum_{l,m} X_{l,m} \\ &+ \sum_{k} \ln P(\boldsymbol{\Omega}_k; \mu_k, \upsilon_k^2, \tau_k^2) \\ &+ \sum_{k,m} \ln P(U_{k,m}; \alpha^{(U)}, \beta^{(U)}),\end{aligned} \tag{14}$$

where $\underset{c}{=}$ represents equality up to a constant. We denote the right-hand side of Eq. (14) by $\mathcal{I}(\Theta)$.

$X_{l,m}$ includes the sums over k and n, and the first term of Eq. (14) involves these sums in the logarithm function. Directly maximizing $\mathcal{I}(\Theta)$ is thus intractable. However, local optima can be found by using the auxiliary function principle [8], [20], [21].

B. Auxiliary Function Principle

An auxiliary function approach consists of two steps. The first step is to introduce auxiliary variables and construct a lower bound of the objective function (the auxiliary function) that is tangent to the objective function at some point and can be maximized in closed form. The second step is to maximize the auxiliary function by alternately updating the parameters and the auxiliary variables. At each iteration the objective function is guaranteed to be nondecreasing.

Since the logarithm function is a concave function, by using the Jensen inequality, the first term of Eq. (14) can be lower-bounded as

$$Y_{l,m} \ln X_{l,m} = Y_{l,m} \ln \sum_{k,n} w_{k,n} G_{k,n,m} U_{k,m} \tag{15}$$

$$\geq Y_{l,m} \sum_{k,n} \lambda_{k,n,l,m} \left(\ln w_{k,n} + \ln U_{k,m} + \ln G_{k,n,m} - \ln \lambda_{k,n,l,m}\right), \tag{16}$$

where $\lambda_{k,n,l,m}$ is a non-negative auxiliary variable such that $\sum_{k,n} \lambda_{k,n,l,m} = 1$ for all l and m. The equality holds if and only if

$$\lambda_{k,n,l,m} = \frac{w_{k,n} G_{k,n,m} U_{k,m}}{X_{l,m}}. \tag{17}$$

Although one may notice that the second term of Eq. (14) is nonlinear in $\Omega_{k,m}$, this term can be well approximated by the integral $\int_{-\infty}^{\infty} X(x, t_m)$, where $X(x, t_m)$ is given by

$$X(x, t_m) = \sum_{k,n} w_{k,n} e^{-\frac{(x - \Omega_{k,m} - \ln n)^2}{2\sigma^2}} U_{k,m}, \tag{18}$$

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 a uniform interval, say Δ_x. Hence,

$$\sum_l X_{l,m} \simeq \frac{1}{\Delta_x} \int_{-\infty}^{\infty} X(x, t_m) \mathrm{d}x \tag{19}$$

$$= \frac{\sqrt{2\pi\sigma^2}}{\Delta_x} \sum_{k,n} w_{k,n} U_{k,m}. \tag{20}$$

This approximation implies that the second term in Eq. (14) depends little on $\Omega_{k,m}$. The choice of the Poisson distribution allows us to use the approximation and derive update rules in closed form. Denoting the set of the auxiliary variables $\{\lambda_{k,n,l,m}\}_{k,n,l,m}$ by λ, we derive the auxiliary function of $\mathcal{I}(\Theta)$ as

$$\begin{aligned}\mathcal{I}^+(\Theta, \lambda) = &\sum_{l,m} Y_{l,m} \sum_{k,n} \lambda_{k,n,l,m} \left(\ln w_{k,n} + \ln U_{k,m} \right. \\ &\left. + \ln G_{k,n,m} - \ln \lambda_{k,n,l,m}\right) \\ &- \frac{\sqrt{2\pi}\sigma}{\Delta_x} \sum_{k,n,m} w_{k,n} U_{k,m} \\ &+ \sum_{k} \ln P(\boldsymbol{\Omega}_k; \mu_k, \upsilon_k^2, \tau_k^2) \\ &+ \sum_{k,m} \ln P(U_{k,m}; \alpha^{(U)}, \beta^{(U)}).\end{aligned} \tag{21}$$

C. Update Rules

By taking account of the normalization constraint of $w_{k,n}$ and using the method of Lagrange multipliers, we can obtain update rules of the parameters. The update rule of $w_{k,n}$ is derived as

$$w_{k,n} \leftarrow \frac{\sum_{l,m} Y_{l,m} \lambda_{k,n,l,m}}{\sum_m U_{k,m}}, \tag{22}$$

followed by normalization:

$$w_{k,n} \leftarrow \frac{w_{k,n}}{\sum_{n'} w_{k,n'}}. \tag{23}$$

The update rule of $U_{k,m}$ is derived as

$$U_{k,m} \leftarrow \frac{\sum_l Y_{l,m} \sum_n \lambda_{k,n,l,m} + \alpha^{(U)} - 1}{\frac{\sqrt{2\pi\sigma^2}}{\Delta_x} + \beta^{(U)}}. \tag{24}$$

To ensure $U_{k,m}$ to be non-negative, we set $U_{k,m} = 0$ if the value of $U_{k,m}$ updated by Eq. (24) is negative. We experimentally found this step not to disturb the convergence of the algorithm.

The update rule of $\boldsymbol{\Omega}_k$ is derived as

$$\begin{aligned}\boldsymbol{\Omega}_k \leftarrow &\left(\frac{\alpha_{\mathrm{l}}}{\tau_k^2} D^{\mathsf{T}} D + \frac{\alpha_{\mathrm{g}}}{\upsilon_k^2} \boldsymbol{I}_M + \sum_{n,l} \mathsf{diag}(\boldsymbol{\eta}_{k,n,l})\right)^{-1} \\ &\times \left(\mu_k \frac{\alpha_{\mathrm{g}}}{\upsilon_k^2} \mathbf{1}_M + \sum_{n,l} (x_l - \ln n) \boldsymbol{\eta}_{k,n,l}\right),\end{aligned} \tag{25}$$

where diag is a function that converts a vector into a diagonal matrix with the entries of the vector on the main diagonal and $\boldsymbol{\eta}_{k,n,l}$ is given by

$$\boldsymbol{\eta}_{k,n,l} = \frac{1}{\sigma^2} \left[Y_{l,1} \lambda_{k,n,l,1}, Y_{l,2} \lambda_{k,n,l,2}, \ldots, Y_{l,M} \lambda_{k,n,l,M}\right]^{\mathsf{T}}. \tag{26}$$

Algorithm 1: Iterative Algorithm of HTFD.

Input: Observed spectrogram $\{Y_{l,m}\}_{l,m}$ and the number of iterations $N^{(\mathrm{iter})}$
Output: w, U and Ω
Initialize w, U and Ω
for $i = 1$ to $N^{(\mathrm{iter})}$
  Compute λ according to Eq. (17)
  Update w according to Eq. (22) followed by Eq. (23)
  Compute λ according to Eq. (17)
  Update U according to Eq. (24)
  forall k and m
  $U_{k,m} \leftarrow \max\{0, U_{k,m}\}$
  endfor
  Compute λ according to Eq. (17)
  Update Ω according to Eq. (25)
endfor

The algorithm of HTFD is summarized in Algorithm 1.

λ is a large array with $KNLM$ elements, so Eq. (17) can be computationally expensive. However, $\lambda_{k,n,l,m}$ can be approximated fairly well by computing only the entries around $\Omega_{k,m} + \ln n$ and dealing with the other entries as zeros since $\lambda_{k,n,l,m}$ is localized around $\Omega_{k,m} + \ln n$. Through preliminary experiments, we decided to compute only the entries within $[\Omega_{k,m} + \ln n - 3\sigma, \Omega_{k,m} + \ln n + 3\sigma]$.

V. Incorporation of Source-Filter Model Into HTFD

A. Parameter Relationship Between Source-Filter Model and HTFD Spectrogram Model

As described in Section I, we incorporate a source-filter model into HTFD in a different way from the excitation-filter product representation. To this end, we first derive a parameter relationship between the CWT spectrogram model defined in Eq. (4) and a source-filter model via the analytic signal model defined in Eq. (1), following the idea described in [19].

Let us consider $s_k(u)$ within a short frame centered at time t_m and denote its discrete-time representation by $s_{k,m}[i]$, where i is the discrete-time index. One well-founded way to describe a source-filter model would be to use an all-pole system. Assuming that $s_{k,m}[i]$ is the output of an all-pole system with $P+1$ filter coefficients, a.k.a. linear predictive coding (LPC) coefficients, $s_{k,m}[i]$ satisfies

$$a_{k,m,0} s_{k,m}[i] = -\sum_{p=1}^{P} a_{k,m,p} s_{k,m}[i-p] + \epsilon_{k,m}[i], \qquad (27)$$

where $\epsilon_{k,m}[i]$ is the excitation signal, $p = 0, \ldots, P$ is the filter coefficient index, and $a_{k,m,p}$ is the pth LPC coefficient such that $a_{k,m,0} \neq 0$. Since $s_{k,m}[i]$ is assumed to be a periodic signal with F_0 of $e^{\Omega_{k,m}}$ consisting of N partials, $\epsilon_{k,m}[i]$ must be a periodic signal with the same F_0:

$$\epsilon_{k,m}[i] = \sum_{n=1}^{N} c_{k,n,m} e^{\mathrm{j}ne^{\Omega_{k,m}} i\Delta_u}, \qquad (28)$$

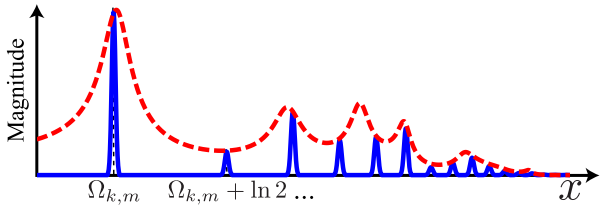


Fig. 3. Illustration of SF-HTFD spectrogram model at time t_m. The dashed line denotes the all-pole spectrum.

where $c_{k,n,m}$ is the complex amplitude of partial n and Δ_u is the sampling period of $s_{k,m}[i]$.

Putting $A_{k,m}(\omega) := \sum_p a_{k,m,p} e^{-\mathrm{j}p\omega}$ and applying the discrete-time Fourier transform (DTFT) to Eq. (27) yield

$$\hat{s}_{k,m}(\omega) = \frac{1}{A_{k,m}(\omega)} \sum_{n=1}^{N} c_{k,n,m} \delta(\omega - ne^{\Omega_{k,m}} \Delta_u), \qquad (29)$$

where $\hat{s}_{k,m}(\omega)$ is the DTFT of $s_{k,m}[i]$ and δ is the Dirac delta function. By applying the inverse DTFT to both sides of Eq. (29), we obtain another expression of $s_{k,m}[i]$ as

$$s_{k,m}[i] = \sum_{n=1}^{N} \frac{c_{k,n,m}}{A_{k,m}(ne^{\Omega_{k,m}} \Delta_u)} e^{\mathrm{j}ne^{\Omega_{k,m}} i\Delta_u}. \qquad (30)$$

Finally, comparing this expression with Eq. (1), we can obtain the following explicit parameter relationship between the CWT spectrogram model and the all-pole system:

$$d_{k,n,m} = \left| \frac{c_{k,n,m}}{A_{k,m}(ne^{\Omega_{k,m}} \Delta_u)} \right|. \qquad (31)$$

Eq. (31) means that the spectral envelope of each of the HTC source spectra is determined by the all-pole spectrum $1/A_{k,m}(ne^{\Omega_{k,m}} \Delta_u)$ as shown in Fig. 3.

B. Generative Model of SF-HTFD

In the model presented in Section II, we have assumed the relative magnitudes $w_{k,n}$ of partials to be time-invariant based on an assumption that the timbre of a musical instrument sound is almost static. However, since the timbre of a pitched sound may be better characterized by its spectral envelope than harmonic magnitudes, we consider it reasonable to assume the shape of the spectral envelope to be static. Here, we derive an extension of the generative model of HTFD in this section based on this idea.

Let us introduce a filter index $f = 1, \ldots, F$ and a pitch index $r = 1, \ldots, R$. In our conference paper, we assigned only one filter to each pitch r. However, under this constraint, it becomes difficult for the model to express the spectrum of concurrent sounds with the same pitch. In addition, we have experimentally found that this constraint tended to make the parameter estimation algorithm numerically unstable. To address these problems, we assign a pair (f, r) to each source k and hereafter replace the subscript k with a subscript pair (f, r).

In a way similar to Eq. (6), we factorize $c_{f,r,n,m}$ as

$$c_{f,r,n,m} = \tilde{c}_{f,r,n,m}\tilde{U}_{f,r,m}, \tag{32}$$

where $\tilde{c}_{f,r,n,m} \in \mathbb{C}$ is the scaled complex amplitude of partial n and $\tilde{U}_{f,r,m} \geq 0$ is the temporal activation. The all-pole spectrum $1/A_{f,r,m}(ne^{\Omega_{f,r,m}}\Delta_u)$ does not change the F_0s, and we can assume $a_{f,r,m,p}$ to be independent of time and pitch. To implement this, we hereafter remove the subscripts m and r in $A_{f,r,m}(\omega)$ and $a_{f,r,m,p}$, i.e., $A_f(\omega)$ and $a_{f,p}$. Substituting Eq. (32) into Eq. (31) yields

$$d_{f,r,n,m} = \tilde{w}_{f,r,n,m}\tilde{U}_{k,m}, \tag{33}$$

$$\tilde{w}_{f,r,n,m} := \left|\frac{\tilde{c}_{f,r,n,m}}{A_f(ne^{\Omega_{f,r,m}}\Delta_u)}\right|. \tag{34}$$

In this way, the relative magnitude of each partial, assumed to be time-invariant in the HTFD model, now becomes time-variant.

Eq. (34) implicitly defines a conditional probability distribution of $\tilde{w}_{f,r,n,m}$ as

$$P(\tilde{w}_{f,r,n,m}|\tilde{c}_{f,r,n,m};\boldsymbol{a}_f,\Omega_{f,r,m}) = \delta\left(\tilde{w}_{f,r,n,m} - \left|\frac{\tilde{c}_{f,r,n,m}}{A_f(ne^{\Omega_{f,r,m}}\Delta_u)}\right|\right), \tag{35}$$

where $\boldsymbol{a}_f = [a_{f,0},\ldots,a_{f,P}]^\mathsf{T}$. If we assume that $\tilde{c}_{f,r,n,m}$ follows an isotropic complex normal distribution with a standard deviation of 1 as in [19], $|\tilde{c}_{f,r,n,m}|$ follows a Rayleigh distribution with a scale parameter of 1. Hence $\tilde{w}_{f,r,n,m}$ also follows a Rayleigh distribution as

$$\tilde{w}_{f,r,n,m} \sim \text{Rayleigh}\left(\tilde{w}_{f,r,n,m}; \frac{1}{|A_f(ne^{\Omega_{f,r,m}}\Delta_u)|}\right), \tag{36}$$

where $\text{Rayleigh}(z;\gamma) = (z/\gamma^2)e^{-z^2/(2\gamma^2)}$. This prior distribution promotes the spectral envelopes of the source spectrogram model to resemble the all-pole spectrum. Note that we do not directly assume the prior distribution of $\tilde{w}_{f,r,n,m}$ but that of $\tilde{c}_{f,r,n,m}$. The maximum likelihood estimation of the prior distribution of $\tilde{w}_{f,r,n,m}$ given by Eq. (36) amounts to fitting the all-pole spectrum $|A_f(ne^{\Omega_{f,r,m}}\Delta_u)|^2$ to $\tilde{w}^2_{f,r,n,m}$ with the Itakura–Saito divergence. Introducing the prior distribution of $\tilde{c}_{f,r,n,m}$ allows us to derive an efficient update rule of $\boldsymbol{a}_f$ as was done in [29], which we will show in Section VI-B.

As with HTFD, the observed spectrogram model $\tilde{X}_{l,m}$ is given as the sum of FP source spectrogram models $\tilde{S}_{f,r,l,m}$:

$$\tilde{X}_{l,m} = \sum_{f,r}\tilde{S}_{f,r,l,m}, \tag{37}$$

$$\tilde{S}_{f,r,l,m} = \left(\sum_n \tilde{w}_{f,r,n,m}G_{f,r,n,m}\right)\tilde{U}_{f,r,m}. \tag{38}$$

By adopting the same probability distributions as Eqs. (8), (12), and (13) for $Y_{l,m}, \Omega_{f,r,m}$, and $\tilde{U}_{f,r,m}$, we can obtain the generative model of SF-HTFD as shown in Fig. 4.

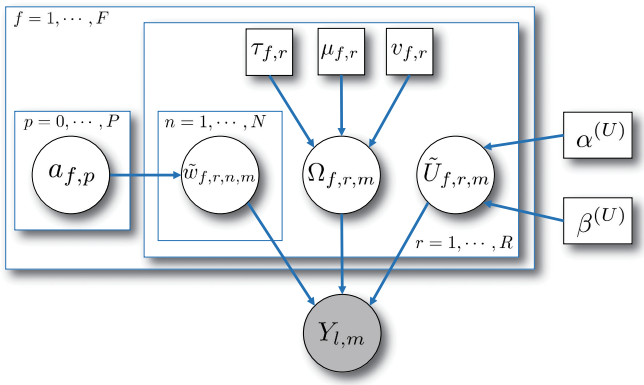


Fig. 4. Plate notation of SF-HTFD generative model.

VI. Parameter Optimization Algorithm of SF-HTFD

A. Derivation of Objective Function

1) Problem Formulation: The parameters of interest in the generative model of SF-HTFD are

$\tilde{w} = \{\tilde{w}_{f,r,n,m}\}_{f,r,n,m}$: relative magnitude of partial n,
$a = \{a_{f,p}\}_{f,p}$: LPC coefficient of filter f,
$\tilde{U} = \{\tilde{U}_{f,r,m}\}_{f,r,m}$: temporal activation,
$\Omega = \{\Omega_{f,r,m}\}_{f,r,m}$: F_0 of filter f and pitch r.

We use $\tilde{\Theta}$ to denote the set of $a, \tilde{U}$ and Ω. Although we have already derived an algorithm that searches for the parameters that maximize the posterior $P(\tilde{w},\tilde{\Theta}|Y)$ in [22], we have found that this algorithm is numerically unstable, particularly in terms of $\tilde{w}$ and a since $\tilde{w}$ heavily depends on the values of a. To alleviate this instability, here we develop an improved algorithm that approximately computes the posterior of $\tilde{w}$ based on the auxiliary function principle.

Let us consider the problem of finding the estimate of $\tilde{\Theta}$ that maximizes the posterior density $P(\tilde{\Theta}|Y)$, or equivalently

$$\mathcal{J}(\tilde{\Theta}) := \ln P(Y|\tilde{\Theta}) + \ln P(\tilde{\Theta}) \tag{39}$$

$$= \ln\int_{\mathbb{W}} P(Y,\tilde{w}|\tilde{\Theta})\mathrm{d}\tilde{w} + \ln P(\tilde{\Theta}), \tag{40}$$

where $\mathbb{W}$ denotes the domain of $\tilde{w}$. Since the first term of Eq. (40) involves the intractable integral in the logarithm function, the current maximization problem is difficult to solve analytically.

2) Derivation of Lower Bound of $\mathcal{J}(\tilde{\Theta})$: Now, in a similar way to Section IV, we derive a lower bound of $\mathcal{J}(\tilde{\Theta})$. The first term of Eq. (40) can be lower-bounded by the Jensen inequality on the concave logarithm function as

$$\ln\int_{\mathbb{W}} P(Y,\tilde{w}|\tilde{\Theta})\mathrm{d}\tilde{w} \geq \int_{\mathbb{W}} q(\tilde{w})\ln\frac{P(Y,\tilde{w}|\tilde{\Theta})}{q(\tilde{w})}\mathrm{d}\tilde{w}, \tag{41}$$

$$= \mathbb{E}_{\tilde{w}}[P(Y|\tilde{w},\tilde{\Theta})] + \mathbb{E}_{\tilde{w}}[P(\tilde{w}|\tilde{\Theta})] + \mathbb{E}_{\tilde{w}}[q(\tilde{w})], \tag{42}$$

where $q(\tilde{w})$ is a non-negative auxiliary distribution of $\tilde{w}$ such that $\int_{\mathbb{W}} q(\tilde{w})\mathrm{d}\tilde{w} = 1$. Here $\mathbb{E}_{\tilde{w}}[g(\tilde{w})]$ is given by

$$\mathbb{E}_{\tilde{w}}[g(\tilde{w})] = \int_{\mathbb{W}} q(\tilde{w})g(\tilde{w})\mathrm{d}\tilde{w}. \tag{43}$$

The equality holds if and only if $q(\tilde{w}) = P(\tilde{w}|Y)$. If we can iteratively update $q(\tilde{w})$ according to this equality condition and $\tilde{\Theta}$ so that it maximizes Eq. (42) plus the log-prior term $\ln P(\tilde{\Theta})$, the original objective function $\mathcal{J}(\tilde{\Theta})$ is guaranteed to be nondecreasing at each iteration according to the auxiliary function principle.

Although computing the exact posterior $P(\tilde{w}|Y)$ is intractable, we can use the idea of variational inference [30] to approximate the distribution using a so-called variational distribution $q(\tilde{w})$. Here, we restrict $q(\tilde{w})$ to a factorized form:

$$q(\tilde{w}) = \prod_{f,r,n,m} q(\tilde{w}_{f,r,n,m}), \tag{44}$$

where $q(\tilde{w}_{f,r,n,m})$ satisfies $\int_0^\infty q(\tilde{w}_{f,r,n,m})\mathrm{d}\tilde{w}_{f,r,n,m} = 1$ for all f, r, n and m.

The first term of Eq. (42) is written as

$$\mathbb{E}_{\tilde{w}}[P(Y|\tilde{w},\tilde{\Theta})] \underset{c}{=} \sum_{l,m} Y_{l,m}\mathbb{E}_{\tilde{w}}[\ln \tilde{X}_{l,m}] - \sum_{l,m} \mathbb{E}_{\tilde{w}}[\tilde{X}_{l,m}], \tag{45}$$

and the first term of Eq. (45) can be lower-bounded by the Jensen inequality as in Eq. (16):

$$Y_{l,m}\mathbb{E}_{\tilde{w}}[\ln \tilde{X}_{l,m}] \geq \sum_{f,r,n} Y_{l,m}\tilde{\lambda}_{f,r,n,l,m}(\mathbb{E}_{\tilde{w}}[\ln \tilde{w}_{f,r,n,m}] + \ln \tilde{U}_{f,r,l,m} + \ln G_{f,r,n,m} - \ln \tilde{\lambda}_{f,r,n,l,m}). \tag{46}$$

Here $\tilde{\lambda}_{f,r,n,l,m}$ is a non-negative variable such that $\sum_{f,r,n} \tilde{\lambda}_{f,r,n,l,m} = 1$ for all l and m. The right-hand side of Eq. (46) is maximized when

$$\tilde{\lambda}_{f,r,n,l,m} \propto e^{\mathbb{E}_{\tilde{w}}[\ln \tilde{w}_{f,r,n,m}]} G_{f,r,n,m}\tilde{U}_{f,r,m}. \tag{47}$$

Since $\tilde{w}_{f,r,n,m}$ can be seen as a tangent line of a quadratic function of $\tilde{w}_{f,r,n,m}$, the second term of Eq. (45) is bounded as

$$-\mathbb{E}_{\tilde{w}}[\tilde{X}_{l,m}] = -\mathbb{E}_{\tilde{w}}[\tilde{w}_{f,r,n,m}]G_{f,r,n,m}\tilde{U}_{f,r,m} \tag{48}$$

$$\geq -\frac{1}{2}\left(\frac{\mathbb{E}_{\tilde{w}}[\tilde{w}^2_{f,r,n,m}]}{\chi_{f,r,n,m}} + \chi_{f,r,n,m}\right) G_{f,r,n,m}\tilde{U}_{f,r,m}, \tag{49}$$

where $\chi_{f,r,n,m}$ is a non-negative variable. The right-hand side of Eq. (49) is maximized when

$$\chi_{f,r,n,m} = \sqrt{\mathbb{E}_{\tilde{w}}[\tilde{w}^2_{f,r,n,m}]}. \tag{50}$$

3) Approximation of Lower Bound of $\tilde{J}(\Theta)$: Furthermore, applying the approximation given by Eq. (19) to the right-hand side of Eq. (49) yields

$$-\sum_l \frac{1}{2}\left(\frac{\mathbb{E}_{\tilde{w}}[\tilde{w}^2_{f,r,n,m}]}{\chi_{f,r,n,m}} + \chi_{f,r,n,m}\right) G_{f,r,n,m}\tilde{U}_{f,r,m}$$

$$\simeq -\sqrt{\frac{\pi\sigma^2}{2\Delta_x^2}}\left(\frac{\mathbb{E}_{\tilde{w}}[\tilde{w}^2_{f,r,n,m}]}{\chi_{f,r,n,m}} + \chi_{f,r,n,m}\right)\tilde{U}_{f,r,m}. \tag{51}$$

This approximation allows us to simplify the variational posterior $q(\tilde{w})$ and derive an update rule of Ω in closed form.

To sum up, a lower bound of $\mathcal{J}(\tilde{\Theta})$ is approximately obtained as

$$\begin{aligned}
&\mathcal{J}^+(\tilde{\lambda}, \chi, q(\tilde{w}), \tilde{\Theta}) \\
&\underset{c}{=} \sum_{l,m,f,r,n} Y_{l,m}\tilde{\lambda}_{f,r,n,l,m}(\mathbb{E}_{\tilde{w}}[\ln \tilde{w}_{f,r,n,m}] + \ln \tilde{U}_{f,r,l,m} \\
&\quad + \ln G_{f,r,m} - \ln \tilde{\lambda}_{f,r,n,l,m}) \\
&\quad - \sqrt{\frac{\pi\sigma^2}{2\Delta_x^2}} \sum_{f,r,n,m} \mathbb{E}_{\tilde{w}}\left[\left(\frac{\tilde{w}^2_{f,r,n,m}}{\chi_{f,r,n,m}} + \chi_{f,r,n,m}\right)\tilde{U}_{f,r,m}\right] \\
&\quad + \sum_{f,r,n,m} \mathbb{E}_{\tilde{w}}\left[\ln P(\tilde{w}_{f,r,n,m}; \boldsymbol{a}_f, \Omega_{f,r,m})\right] + \mathbb{E}_{\tilde{w}}[q(\tilde{w})] \\
&\quad + \sum_{f,r} \ln P(\boldsymbol{\Omega}_{f,r}; \mu_{f,r}, \upsilon^2_{f,r}, \tau^2_{f,r}) \\
&\quad + \sum_{f,r,m} \ln P(\tilde{U}_{f,r,m}; \alpha^{(U)}, \beta^{(U)}),
\end{aligned} \tag{52}$$

where $\tilde{\lambda} := \{\tilde{\lambda}_{f,r,n,l,m}\}_{f,r,n,l,m}$ and $\chi := \{\chi_{f,r,n,m}\}_{f,r,n,m}$. Since the equalities of inequalities (46) and (49) do not always hold at Eqs. (47) and (50), respectively, $\mathcal{J}^+(\tilde{\lambda}, \chi, q(\tilde{w}), \tilde{\Theta})$ is no longer an auxiliary function of $\mathcal{J}(\tilde{\Theta})$. However, the inequalities (42), (46), and (49) hold for any $q(\tilde{w})$, $\tilde{\lambda}$, and χ, respectively, and the bound property is preserved up to the approximation (51). We maximize $\mathcal{J}^+(\tilde{\lambda}, \chi, q(\tilde{w}), \tilde{\Theta})$ instead of the original objective function.

B. Update Rules

1) Update Rule of $q(\tilde{w}_{f,r,n,m})$: By taking into account the normalization constraint of $q(\tilde{w}_{f,r,n,m})$ and using the method of Lagrange multipliers, we can obtain the following update rule of $q(\tilde{w}_{f,r,n,m})$.

$$q(\tilde{w}_{f,r,n,m}) \leftarrow \mathrm{Nakagami}(\tilde{w}_{f,r,n,m}; \rho_{f,r,n,m}, \zeta_{f,r,n,m}), \tag{53}$$

where $\mathrm{Nakagami}(z; \rho, \zeta) \propto z^{2\rho-1}e^{-\rho z/\zeta}$ and

$$\rho_{f,r,n,m} := \frac{\sum_{l,m} Y_{l,m}\tilde{\lambda}_{f,r,n,l,m}}{2} + 1, \tag{54}$$

$$\zeta_{f,r,n,m} := \frac{\sum_{l,m} Y_{l,m}\tilde{\lambda}_{f,r,n,l,m} + 2}{\frac{\sqrt{2\pi\sigma^2}\tilde{U}_{f,r,m}}{\chi_{f,r,n,m}\Delta_x} + |A_f(ne^{\Omega_{f,r,m}}\Delta_u)|^2}. \tag{55}$$

The expectations associated with $q(\tilde{w})$ are computed as

$$\mathbb{E}_{\tilde{w}}[\ln \tilde{w}_{f,r,n,m}] = \frac{1}{2}\left(\psi(\rho_{f,r,n,m}) - \ln \frac{\rho_{f,r,n,m}}{\zeta_{f,r,n,m}}\right), \tag{56}$$

$$\mathbb{E}_{\tilde{w}}[\tilde{w}^2_{f,r,n,m}] = \zeta_{f,r,n,m}, \tag{57}$$

$$\mathbb{E}_{\tilde{w}}[\tilde{w}_{f,r,n,m}] = \frac{\Gamma(\rho_{f,r,n,m}+\frac{1}{2})}{\Gamma(\rho_{f,r,n,m})}\sqrt{\frac{\zeta_{f,r,n,m}}{\rho_{f,r,n,m}}}, \tag{58}$$

where ψ denotes the digamma function.

2) Update Rule of $\boldsymbol{a}_f$: For $\boldsymbol{a}_f$, we can use the multiplicative update algorithm as in [29]. Since

$$|A_f(ne^{\Omega_{f,r,m}}\Delta_u)|^2 = \boldsymbol{a}_f^{\mathsf{T}} C(ne^{\Omega_{f,r,m}}\Delta_u)\boldsymbol{a}_f, \tag{59}$$

where $C(\omega)$ is a $(P+1)\times(P+1)$ Toeplitz matrix whose (p,q)th element is given by $\cos(\omega(p-q))$, the partial derivative of $-\mathcal{J}^+(\tilde{\lambda},\chi,q(\tilde{w}),\tilde{\Theta})$ with respect to $\boldsymbol{a}_f$ can be represented as

$$-\frac{\partial \mathcal{J}^+(\tilde{\lambda},\chi,q(\tilde{w}),\tilde{\Theta})}{\partial \boldsymbol{a}_f} = (V_f - V_f')\boldsymbol{a}_f, \tag{60}$$

where V_f and V_f' are $(P+1)\times(P+1)$ positive definite matrices defined by

$$V_f = \sum_{p,n,m}\frac{1}{\boldsymbol{a}_f^{\mathsf{T}} C(ne^{\Omega_{f,r,m}}\Delta_u)\boldsymbol{a}_f} C(ne^{\Omega_{f,r,m}}\Delta_u), \tag{61}$$

$$V_f' = \sum_{p,n,m}\frac{\mathbb{E}_{\tilde{w}}[\tilde{w}_{f,r,n,m}^2]}{2} C(ne^{\Omega_{f,r,m}}\Delta_u). \tag{62}$$

The update rule of $\boldsymbol{a}_f$ is given by

$$\boldsymbol{a}_f \leftarrow V_f^{-1}V_f'\boldsymbol{a}_f. \tag{63}$$

To adjust the scale of $1/|A_f(ne^{\Omega_{f,r,m}}\Delta_u)|$, we normalize $\boldsymbol{a}_f$ such that $\boldsymbol{a}_f \leftarrow \boldsymbol{a}_f/a_{f,0}$ for all f after Eq. (63) is applied.

3) Update Rules of $\tilde{U}_{f,r,m}$ and $\boldsymbol{\Omega}_{f,r}$: In a similar way to Section IV-C, the update rule of $\tilde{U}$ is derived as

$$\tilde{U}_{f,r,m} \leftarrow \frac{\sum_{l,n} Y_{l,m}\tilde{\lambda}_{f,r,n,l,m} + \alpha^{(U)} - 1}{\sqrt{\frac{\pi\sigma^2}{2\Delta_x^2}}\left(\sum_n \frac{\mathbb{E}_{\tilde{w}}[\tilde{w}_{f,r,n,m}^2]}{\chi_{f,r,n,m}}\right) + \beta^{(U)}}, \tag{64}$$

followed by $\tilde{U}_{f,r,m} \leftarrow \max\{0,\tilde{U}_{f,r,m}\}$, which ensures $\tilde{U}_{f,r,m}$ to be non-negative. If the term related to the all-pole spectrum in the objective function, or $\mathbb{E}_{\tilde{w}}[\ln P(\tilde{w}_{f,r,n,m};\boldsymbol{a}_f,\Omega_{f,r,m})]$, can be assumed to be negligible when updating $\boldsymbol{\Omega}_{f,r}$, we can use Eq. (25) in which $\boldsymbol{\eta}_{f,r,n,l}$ is replaced with $\tilde{\boldsymbol{\eta}}_{f,r,n,l}$ to update $\boldsymbol{\Omega}_{f,r}$:

$$\boldsymbol{\Omega}_{f,r} \leftarrow \left(\frac{\alpha_{\mathrm{l}}}{\tau_{f,r}^2}D^{\mathsf{T}}D + \frac{\alpha_{\mathrm{g}}}{\upsilon_{f,r}^2}\boldsymbol{I}_M + \sum_{n,l}\mathsf{diag}(\tilde{\boldsymbol{\eta}}_{f,r,n,l})\right)^{-1} \times \left(\mu_{f,r}\frac{\alpha_{\mathrm{g}}}{\upsilon_{f,r}^2}\boldsymbol{1}_M + \sum_{n,l}(x_l - \ln n)\tilde{\boldsymbol{\eta}}_{f,r,n,l}\right), \tag{65}$$

where $\tilde{\boldsymbol{\eta}}_{f,r,n,l}$ is a M-dimensional vector whose mth entry is given by $Y_{l,m}\lambda_{f,r,n,l,m}/\sigma^2$. Although the above assumption used when deriving Eq. (65) and the normalization steps of $\boldsymbol{a}_f$ and $\tilde{U}_{f,r,m}$ may affect the convergence of the algorithm, we have experimentally found that the algorithm works well.

The algorithm of SF-HTFD is summarized in Algorithm 2.

Algorithm 2: Iterative Algorithm of SF-HTFD.

Input: Observed spectrogram Y and the number of iterations $N^{(\mathrm{iter})}$
Output: $q(\tilde{w}), a, \tilde{U}$ and Ω
Initialize $q(\tilde{w}), a, \tilde{U}$ and Ω
for $i = 1$ to $N^{(\mathrm{iter})}$ **do**
  Compute λ and χ according to Eqs. (47) and (50)
  Update $q(\tilde{w})$ according to Eq. (53)
  Compute λ and χ according to Eqs. (47) and (50)
  Update $\tilde{U}$ according to Eq. (64)
  for all f, r and m **do**
    $U_{f,r,m} \leftarrow \max\{0, U_{f,r,m}\}$
  end for
  Compute λ and χ according to Eqs. (47) and (50)
  Update a according to Eq. (63)
  for all f **do**
    $\boldsymbol{a}_f \leftarrow \boldsymbol{a}_f/a_{f,0}$
  end for
  Compute λ and χ according to Eqs. (47) and (50)
  Update Ω according to Eq. (65)
end for

Using $q(\tilde{w}), a, \tilde{U}$ and Ω obtained with the algorithm, the estimated spectrogram of filter f and pitch r is calculated as $\mathbb{E}_{\tilde{w}}[\tilde{S}_{f,r,l,m}]$. To reduce the computational cost, as in HTFD, we computed the entries of $\check{\lambda}$ only within $[\Omega_{f,r,m} + \ln n - 3\sigma, \Omega_{f,r,m} + \ln n + 3\sigma]$.

VII. Related Studies

A. Separation of Harmonic Sounds With Time-Varying F_0s

In this section, we briefly review related studies and highlight our contributions. The proposed methods can be reduced to HNMF as described in Section II-D. HNMF forces the spectral templates to have the harmonic structures by initializing the entries of each template with zero except for those around the corresponding F_0 and its harmonics. The update rule of the spectral templates is multiplicative and the zero-initialized entries remain zeros after the update. Although HNMF works well for the separation of nearly stationary harmonic sounds, time-varying fluctuations such as vibrato decrease the separation performance of HNMF due to its assumption of the time-invariance of the spectral templates [31], [32].

To better capture sounds having time-varying fluctuations, a variant of NMF (Hennequin model) has been developed by allowing the F_0 of each basis spectrum to be time-varying [31]. The Hennequin model can be seen as an STFT version of the HTFD spectrogram model, with the exception of the additional penalty terms. Although it may appear that the Hennequin model can also be used to express CWT spectrograms by naively using it in the log-frequency domain, this may not work properly since the interpretation of the model parameters becomes unclear.

By contrast, since our HTFD model is derived based on the time-domain signal model, the interpretation of the model

parameters remains clear, regardless of the domain in which it is expressed. Namely, our spectrogram model, represented in either the CWT domain or STFT domain, is designed so that the model parameters are associated with those of the time-domain signal representation. In fact, an STFT version of the HTFD model can easily be derived. Specifically, the STFT version of the HTFD model can be written by replacing $G_{k,n,m}$ in Eq. (7) with $G^{(\mathrm{STFT})}_{k,n,m}$ given by

$$G^{(\mathrm{STFT})}_{k,n,m} = e^{-\frac{(e^x - n e^{\Omega_{k,m}})}{2\sigma^2}}. \tag{66}$$

B. Source-Filter Model Representation in CWT Domain

After our conference paper [22] had been published, a spectrogram model called an infinite superimposed discrete all-pole (iSDAP) model was proposed for multipitch analysis [17]. The iSDAP model adopts the the excitation-filter product representation in the CWT domain. As described in Section I, this way of modeling CWT spectrograms is not justified since CWT is not an orthogonal transform. In addition, the interpretation of the model parameters is unclear for the same reason as the Hennequin model when used as a CWT spectrogram model.

Indeed, the SF-HTFD model can be reduced to one using the excitation-filter product representation. If we assume that $c_{f,r,n,m}$ in Eq. (32) follows $\delta(c_{f,r,n,m} - 1)$ instead of the isotropic complex normal distribution when deriving the SF-HTFD model, $\tilde{w}_{f,r,n,m}$ equals $1/|A_f(ne^{\Omega_{f,r,m}}\Delta_u)|$. The spectrum model of each f and r pair is then represented by a product of the all-pole spectrum and the excitation spectrum whose partial amplitudes equal $\tilde{U}_{f,r,m}$. Thus, adopting the excitation-filter product representation in the CWT domain implies that the partials of the excitation spectrum are assumed to have the same magnitude, which can be too restrictive for audio source separation. We will show the effect of this assumption on the separation performance in Section VIII. To our knowledge, this is the first time in the literature to reveal the flaws and limitations of the excitation-filter product representation assumed in the CWT domain. Although the NMF-based model presented in [18] also uses the excitation-filter product representation in the CWT domain, we did not consider this model for comparison since it is for supervised settings.

C. Unsupervised Monaural Audio Source Separation

While the proposed methods perform the source separation in the CWT domain, a few recent methods of unsupervised monaural audio source separation have explored modulation-based representations of signals [32]–[34]. In [32], a variant of the modulation spectrogram (the common fate transform) has been presented, which has explicit dimensions corresponding to the spectral and temporal modulations. The source separation is achieved by factorizing the common fate transform of a mixture audio signal into that of the sources with nonnegative tensor factorization (NTF), which was originally introduced for multichannel audio source separation [35]. Since, for the sources with different modulation patterns, their common fate transforms are likely to be less heavily overlapped than the STFTs, the modulation-based representations have been reported to be advantageous for separating mixtures of vibrato sounds played with a same pitch into those of different musical instruments [32], [33]. Although the use of NTF assumes that each source has a unique modulation pattern, the modulation patterns are often changed, depending on musical expressions (e.g., legato and vibrato), players, and their performance skills. Additionally, since these methods are designed to separate the concurrent instrument sounds played with the same pitch, they are difficult to directly use for separation into the sounds of individual pitches, which we address. By contrast, although HTFD and SF-HTFD are not designed for instrument-wise separation, they do not impose such an assumption and can work well for real performances as shown in Section VIII.

In [16], a NTF-based framework was presented, which has the flexibility of encompassing prior distributions of its parameters and the source-filter model. However, this framework assumes the excitation-filter product representation. As shown in Section VII-B, this representation is reasonable in the STFT domain but is not well justified in the CWT domain. In addition, although NTF can be used for the monaural audio source separation of the modulation-based representations as described in the above, in such a domain, the excitation-filter product representation is apparently inappropriate for expressing the source-filter model. For example, the common fate transform is computed by dividing the complex STFT into overlapping rectangular patches and applying the two-dimensional discrete Fourier transform to these patches, and the common fate transform of the source is not represented by a product of the common fate transforms of the excitation and filter components. On the other hand, the proposed methods can appropriately encompass the source-filter model in the CWT domain without assuming the excitation-filter product representation.

VIII. Experimental Evaluations

A. Data Preparation and Separation Procedure

We assessed the separation performance of HTFD and SF-HTFD through experiments on separating music audio signals into the signals of individual semitones. We first describe common experimental settings in the following two sections.

For the test data, we used the first 30 seconds of seven musical pieces, Classic Music No. 1 to 7 from the RWC music database [36]. To determine the hyperparameters of the proposed and conventional methods, we used the first 30 seconds of Classic Music No. 13 to 15 from the RWC music database as the development data. We synthesized mixed and ground truth audio signals from MIDI files of the excerpts with a MIDI synthesizer called FluidSynth [37] and a high-quality GeneralUser GS 1.4 soundfont to make the synthesized signals as realistic as possible. The sampling frequency was set to 16 kHz, i.e., $\Delta_u = 1/16000$. All control messages contained in the MIDI files were preserved, but the drum tracks were muted. The number of semitones and input signal-to-distortion ratios (SDRs) of each excerpt are as shown in Table I.

We first transformed the mixed signals into spectrograms using the fast approximate CWT algorithm [38], [39] with a

TABLE I
STATISTICS OF EXCERPTS

Excerpt	#semitones	Input SDR [dB]				
		Ave.	Std.	Min.	Med.	Max.
No. 1	24	−16.9	6.9	−31.0	−15.8	−7.9
No. 2	41	−18.7	5.4	−29.6	−18.4	−10.4
No. 3	40	−20.3	7.2	−39.2	−19.7	−6.5
No. 4	24	−16.4	6.7	−32.9	−14.7	−8.1
No. 5	39	−21.3	9.0	−45.5	−19.4	−4.7
No. 6	18	−16.1	7.0	−27.6	−15.0	−3.1
No. 7	35	−18.4	6.3	−34.2	−16.2	−7.4

time resolution of around 2 ms and frequency bins from 27.5 to 7040 Hz per $100/3$ cents, i.e. $x_l = \ln(2\pi \times 27.5) + (l-1)\ln(2)/36$ for $l = 1, \ldots, L$. We used the analyzing wavelet defined by Eq. (2) and set $\sigma = \ln(2)/60$, which corresponds to one fifth of a semitone interval. To reduce the computation time, we decimated the magnitude spectrogram of each mixed signal so that $\Delta_t = t_m - t_{m-1}$ was around 10 ms. We scaled the decimated magnitude spectrogram such that $\sum_{l,m} X_{l,m}/(LM) = 1$, as in [15], and applied the separation methods to it. After the separation, we linearly interpolated the separated spectrograms up to the original time resolution and converted the interpolated ones into the time-domain signals using the inverse fast approximate CWT algorithm with the phase of each mixed signal.

B. HTFD Experiments

1) Comparison With HNMF and HTC: To evaluate the effect induced by the features of the HTC and NMF spectrogram models, we first compared HTFD with HNMF [24] and HTC [7]. For fair comparison, we used the Gamma distribution given by Eq. (13) as a prior distribution of temporal activations. To maximize the separation performance of HNMF and HTFD, we performed a grid search on $\beta^{(U)} = 1.0 \times 10^{\gamma}$ among $\gamma = -5, -4, -3, -2, -1$ and 0 while fixing $\alpha^{(U)} = 1$. For these methods, the number of iterations $N^{(\mathrm{iter})}$ was set to 100.

The other parameters of these methods were set as follows.

HTFD: The hyperparameters were set to $K = 88$, $N = 20$, $\alpha_{\mathrm{g}} = \alpha_{\mathrm{l}} = 100$, $\tau_k = \ln(2)\Delta_t/72$ (vibrato frequency of 6 Hz), $\upsilon_k = \ln(2)/36$ (one third of semitone interval), and $\mu_k = \ln(2\pi \times 27.5) + (k-1)\ln(2)/12$ for $k = 1, \ldots, 88$ (A0 to C8). $U_{k,m}$ were randomly initialized, and $\Omega_{k,m}$ and $w_{k.n}$ were initialized at $\Omega_{k,m} = \mu_k$ and $w_{k,n} \propto e^{-n}$.

HNMF: We used 88 spectral templates, which were associated with semitones from A0 to C8. The kth template was initialized such that its entries within a range of $[\mu_k + \ln n - 3\sigma, \mu_k + \ln n + 3\sigma]$ were proportional to $\sum_n e^{-n} e^{-(x_l - \ln n - \mu_k)^2/(2\sigma^2)}$ for all ns and the other entries were set to zero. The temporal activations of HNMF were initialized randomly.

HTC: The parameters were set and initialized as in [7] except for the number of initial models since we experimentally found that the number of initial models strongly affected the separation performance. Through a grid search, we determined the number of initial models as 600. Although HTC originally takes the CWT power spectrograms $|Y_{l,m}|^2$ as inputs, we instead used the CWT magnitude spectrograms $|Y_{l,m}|$, which significantly improved the separation performance of HTC.

TABLE II
AVERAGE AND MEDIAN SDR IMPROVEMENTS OF PROPOSED HTFD AND CONVENTIONAL METHODS OVER ALL SOURCES. TFR AND IRM DENOTE THE TIME FREQUENCY REPRESENTATION AND THE IDEAL RATIO MASK

Method	TFR	Window Length	γ	Average	Median
IRM	CWT	-	-	25.38	24.58
Hennequin [31]	STFT	64 ms	−2	9.15	10.42
		128 ms	−5	10.44	11.33
		256 ms	−1	8.57	10.32
HTC [7]	CWT	-	-	9.58	11.18
HNMF [24]		-	−2	16.27	16.80
HTFD		-	−5	**17.53**	**17.88**

Table II summarizes the average and median SDR improvements obtained with all methods under the best γ settings, where SDRs were computed using the BSS Eval toolbox [40]. As a reference objective, we also provided the results for the ideal ratio mask (IRM), which computes the best mask with the ground truth sources. In terms of average and median SDR improvements, HTFD outperformed HTC by more than 8 dB and HNMF by more than 1 dB. For all the excerpts, HTFD consistently achieved the highest average and median SDR improvements as shown in Table III. We also examined the statistical significance of the SDR differences between HTFD and HNMF, which gave the highest SDR improvements in the conventional methods, by performing a paired t-test on the sources of all the excerpts. Since the p-value was 2.00×10^{-5}, HTFD significantly outperformed HNMF in SDR improvements. These results clearly show the effect induced by the features of the spectrogram models used in the HTC and NMF approaches.

2) Comparison With STFT Domain Model: Next, we compared HTFD with the Hennequin model presented in [31]. As in HNMF, we also used the gamma distribution as a prior distribution of the temporal activations instead of the penalty terms used in the literature and conducted a grid search on γ to maximize the separation performance. We set the other parameters as in [31] except for the harmonic amplitudes of the spectral templates since we observed that initializing the partial amplitudes in the same way as HTFD performed better. For STFT, we used a Gaussian window with a hopsize of 10 ms length, where the standard deviation of the Gaussian window was set to one sixth of the window length.

As shown in Table III, HTFD achieved higher average and median SDR improvements than the Hennequin methods when the time windows were 64,128 and 256 ms lengths. This result supports that CWT is more suitable for music source separation than STFT, which is consistent with the literature [41].

C. SF-HTFD Experiments

1) Comparison With HTFD: To evaluate the effect of incorporating the source-filter model into HTFD, we compared SF-HTFD with HTFD. We further compared it with the iSDAP model [17] to evaluate the advantage of the proposed modeling frameworks over the excitation-filter product representation in the CWT domain. For SF-HTFD and the iSDAP model, the filter degree was set at $P = 16, 32, 48$ and 64, and the number of filters at $F = 1, 3, 5$, and 7. Since these models adopt gamma

TABLE III
SEPARATION PERFORMANCES OF PROPOSED HTFD AND CONVENTIONAL METHODS FOR EACH EXPERT. THE LABELS ARE THE SAME AS TABLE II

Method	Window Length	Excerpt No. 1	No. 2	No. 3	No. 4	No. 5	No. 6	No. 7
IRM	-	24.98/21.46	25.62/25.96	22.93/23.56	25.45/22.79	27.81/27.61	26.95/26.19	24.58/23.40
Hennequin [31]	64 ms	9.08/9.76	9.75/11.00	7.31/7.06	7.98/8.00	9.22/11.45	9.97/10.80	10.89/12.20
	128 ms	9.50/10.02	11.22/11.94	9.20/7.97	9.52/9.95	10.65/12.80	10.94/11.62	11.68/12.63
	256 ms	−6.59/ − 3.41	11.46/12.37	9.85/9.66	10.31/11.03	9.99/11.26	10.80/11.65	10.20/12.15
HTC [7]	-	13.38/11.36	10.44/10.95	10.68/10.98	8.82/10.80	9.07/12.21	11.70/13.30	4.68/8.75
HNMF [24]	-	18.47/15.76	15.10/15.60	13.26/14.77	17.10/16.95	16.08/17.17	20.91/20.96	16.82/17.14
HTFD	-	**19.21/16.28**	**16.99/18.35**	**13.75/17.19**	**18.23/17.78**	**18.01/17.92**	**22.27/22.24**	**17.89/17.62**

TABLE IV
γ VALUES OF SF-HTFD AND ISDAP MODEL FOR EACH PAIR OF P AND F

	iSDAP				SF-HTFD			
P \ F	1	3	5	7	1	3	5	7
16	−1	−1	−1	−1	0	−1	−2	−1
32	−1	−1	−1	−1	−5	−4	−1	−1
48	−3	−4	0	−3	−4	−2	−5	−1
64	−5	−5	0	−4	0	−3	−5	−2

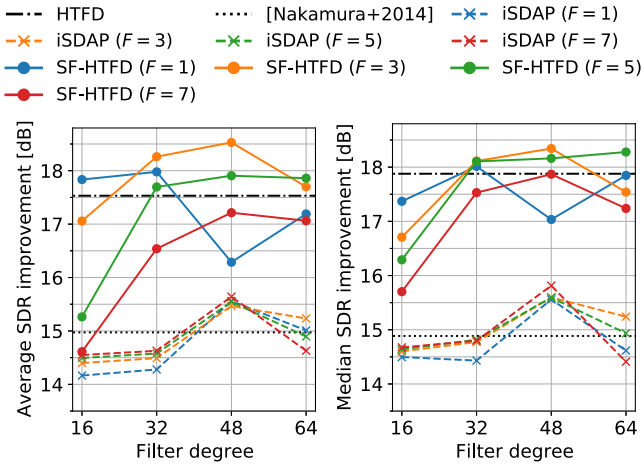
distributions as the prior distributions of temporal activations, we conducted a grid search on γ for each F and R pair to maximize the separation performance. The chosen values of γ are summarized in Table IV. The other parameters of SF-HTFD and the iSDAP model were set as follows.

SF-HTFD: We set $\mu_{f,r} = \ln(2\pi \times 27.5) + (r-1)\ln(2)/12$ for $r = 1, \ldots, 88$ and $f = 1, \ldots, F$, and $\alpha_{\mathrm{g}} = \alpha_{\mathrm{l}} = 1$. The LPC coefficients $\boldsymbol{a}_f$ were initialized by randomly sampling P values in the range of [0.09,0.11] and then using the sampled values as the poles of the all-pole system. This initialization ensures that the all-pole spectral functions monotonically decrease in the frequency direction. The variational posterior $q(\tilde{w}_{f,r,n,m})$ was initialized at the corresponding prior distribution. The other parameters were set and initialized as in HTFD.

iSDAP: We initialized the parameters and variational posteriors of the iSDAP model as in [17] except for the LPC coefficients, which were initialized as in SF-HTFD for fair comparison.

Fig. 5 shows the average and median SDR improvements of these models. SF-HTFD achieved the highest average and median SDR improvements when $(P, F) = (48, 3)$ and $(P, F) = (48, 5)$, respectively, and SF-HTFD with several P and F pairs outperformed HTFD, demonstrating the effect of incorporating the source-filter model.

The performance gain of SF-HTFD over HTFD was greater on average than median. We observed that SF-HTFD produced less standard deviation of the average SDR improvements over the excerpts than HTFD: The standard deviations were 6.36 dB for SF-HTFD and 8.26 dB for HTFD. These results may imply that the incorporation of the source-filter model had a larger effect on the sources that HTFD failed to separate than on those that it could separate successfully. To verify this, we compared the distributions of the SDR improvements of HTFD and SF-HTFD. Fig. 6 clarifies that SF-HTFD had fewer outliers of SDR improvements in all the excerpts and around 0.8 dB larger 25th percentiles than HTFD averagely. The 75th percentiles of SF-HTFD were slightly lower than those of HTFD in the excerpts



(a) Average SDR improvement (b) Median SDR improvement

Fig. 5. Separation performance of proposed and iSDAP models. The equations for F in the parentheses of the label names represent the number of filters. The chain, dotted, solid, and dashed lines correspond to HTFD, the model presented in our conference paper, SF-HTFD, and the iSDAP model, respectively. The blue, orange, red, and green colors correspond to $F = 1, 3, 5$, and 7, respectively.

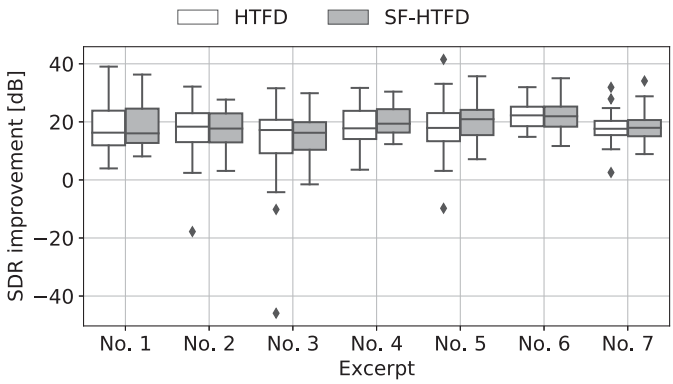


Fig. 6. Boxplots of SDR improvements for HTFD and SF-HTFD with $(P, F) = (48, 3)$. The parameters of HTFD are the same as in Table II.

No. 1 and 2, but in the other excerpts, were equal to or slightly greater than those of HTFD.

2) Comparison With iSDAP Model: SF-HTFD outperformed the iSDAP model for all the F and R pairs, demonstrating the advantage of our model over the excitation-filter product representation in the CWT domain. Due to this representation, the iSDAP model implicitly assumes the partials of the excitation spectra to have the same magnitude, representing the spectral

shape of each semitone by a superposition of only F all-pole spectra. This can be too restrictive for audio source separation, and this could be why the iSDAP model could not perform better than the HTFD model.

3) Comparison With Our Previous Model: In addition to the above models, we compared the present model with the model presented in our conference paper [22], which we call *[Nakamura+2014]*. The parameters of this model were initialized in a similar way to SF-HTFD. Through a grid search on γ and P, we decided to use $P = 2$ and $\gamma = 0$, which gave the highest average SDR. During the grid search, we occasionally encountered numerical errors when updating $\boldsymbol{a}_f$, particularly with $P \geq 16$. When such errors occurred, we chose to adopt the values obtained just before the last iteration as the current estimates of $\boldsymbol{a}_f$. As shown in Fig. 5, SF-HTFD notably outperformed [Nakamura+2014], clearly showing the advantages of the reformulation and the novel inference algorithm we have presented in this paper.

D. Demonstration on Automatic Musical Key Transposition

In this section, we demonstrate how well the proposed methods perform in an automatic musical key transposition task using real audio signals. The task is to change the key of an input music audio signal to another key. For this experiment, we developed an automatic musical key transposition system, which works by performing the following four steps. (i) Given a music audio signal and its key, we first separated the magnitude spectrogram of the input signal into that associated with each semitone by using HTFD, SF-HTFD, or HNMF. (ii) We selected the semitones to be transposed according to the source and target key and shifted only the separated magnitude spectrograms corresponding to the selected semitones in the log-frequency direction. In the equal-tempered scale, the F_0s of different semitones are equally spaced in log-frequency. Thus, shifting a spectrogram by a fixed offset corresponds to pitch transposition. For example, when the key of the input signal is A major, we can convert its scale into the A natural minor scale by shifting the separated magnitude spectrograms corresponding to C♯, F♯, and G♯ down by one semitone so that they correspond to C, F and G. (iii) We added all the separated spectrograms, including the shifted ones, together to construct the spectrogram of a pitch-transposed version and (iv) finally converted it into a time-domain signal using the phase reconstruction algorithm for fast approximate CWT presented in [39].

Several separated and key-transposed results obtained with the above system are available at http://tomohikonakamura.github.io/Tomohiko-Nakamura/demo/key_transposition/index.html , where we experimentally used the initialization that $w_{k,n} \propto e^{-0.3n}$ for HTFD and HNMF. We can confirm from these examples that the audio signals generated using HTFD and SF-HTFD had less dissonance than those obtained with HNMF, especially in vibrato sounds. When listening to the separated audio signals of the pitches that should not be transposed, we notice that the separated signals of SF-HTFD have less squeaking noise than those of the other methods and that SF-HTFD yielded perceptually better separation. These observations show that HTFD and SF-HTFD can not only perform well even on real audio signals but also track the F_0s of the performed notes more accurately than HNMF.

IX. Conclusion

In this paper, we presented a monaural audio source separation framework called HTFD and its extension incorporating the source-filter model, called SF-HTFD. The spectrogram model of HTFD concurrently offers the advantages of the HTC and NMF models, in which the regularities underlying both the local and global time-frequency structures of music spectrograms are exploited. The efficient parameter estimation algorithm of HTFD was derived based on the auxiliary function principle. To incorporate the source-filter model, we derived the explicit parameter relationship between the HTFD spectrogram model and the all-pole system and extended the generative model of HTFD to that of SF-HTFD. This extension allows for separately modeling fine structures and envelopes of source spectra and designing the reasonable prior distributions on F_0s and spectral envelopes. Furthermore, we revealed that assuming the excitation-filter product representation in the CWT domain as in the iSDAP model corresponding to assuming the partial amplitudes of the excitation spectrum to be the same. As with HTFD, we derived a parameter estimation algorithm consisting of closed-form update rules based on the auxiliary function principle. Through music audio separation experiments, we showed the efficacy of HTFD and SF-HTFD and demonstrated their effectiveness to automatic musical key transposition.

Some improvements remain as future work. Many musical pieces include percussive sounds, e.g., drums and attack parts of piano sounds, and dealing with these sounds would broaden the applications of the proposed methods. Since the regular NMF works well for the separation of percussive sounds, one possible way to deal with these sounds is to add the regular NMF model to the HTFD and SF-HTFD models as in [31]. Another way to deal with percussive sounds would be to preprocess an input signal with harmonic-percussive source separation [42].

Appendix A
Derivation of Continuous Wavelet Transform of Source Signal Model

In this section, we derive the CWT of the analytic signal model given by Eq. (1). The wavelet bases are computed by scaling and time-shifting the analyzing wavelet $\xi(u)$:

$$\xi_{\alpha,t}(u) = \frac{1}{\alpha}\xi\left(\frac{u-t}{\alpha}\right), \tag{A1}$$

where $\alpha > 0$ is the scale parameter. The CWT of $s_k(u)$ is written as

$$W_k\left(\ln\frac{1}{\alpha}, t\right) = \sum_{n=1}^{N}\int_{-\infty}^{\infty} d_{k,n}(u)e^{\mathrm{j}(n\theta_k(u)+\varphi_{k,n})}\xi^*_{\alpha,t}(u)\mathrm{d}u, \tag{A2}$$

where $\xi^*_{\alpha,t}(u)$ is the complex conjugate of $\xi_{\alpha,t}(u)$. The dominant part of $\xi^*_{\alpha,t}(u)$ is typically localized around time t, and the results of the integrals in Eq. (A2) shall depend only on the values of

$\theta_k(u)$ and $d_{k,n}(u)$ near t. For this reason, we introduce zeroth- and first-order approximations of $\theta_k(u)$ and $d_{k,n}(u)$ around time t given by

$$d_{k,n}(u) \simeq d_{k,n}(t), \quad \theta_k(u) \simeq \theta_k(t) + \dot{\theta}_k(t)(u-t), \quad \text{(A3)}$$

where $\dot{\theta}_k(u)$ is the time derivative of $\theta_k(u)$, a.k.a. the instantaneous fundamental frequency. By undertaking the above approximations, applying Parseval's theorem, and putting $\Omega_k(t) = \ln \dot{\theta}_k(t)$, Eq. (A2) is reduced to

$$W_k\left(\ln \frac{1}{\alpha}, t\right) = \sum_{n=1}^{N} d_{k,n}(t) \Xi^*(n e^{-x+\Omega_k(t)}) e^{j(n\theta_k(t)+\varphi_{k,n})}. \quad \text{(A4)}$$

Since the function $\Xi(\omega)$ can be chosen arbitrarily, we can use the wavelet given by Eq. (2). Since this wavelet has a center frequency of 1, α equals the reciprocal of the angular frequency, and we can put $x = \ln(1/\alpha)$. We can thus write Eq. (A4) as Eq. (3).

ACKNOWLEDGMENT

The authors would like to thank Kotaro Shikata and Norihiro Takamune for their fruitful discussions.

REFERENCES

[1] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE Int. Conf. Acoust, Speech Signal Process.*, Mar. 2016, pp. 31–35..

[2] N. Takahashi and Y. Mitsufuji, "Multi-scale multi-band densenets for audio source separation," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2017, pp. 21–25.

[3] N. Takahashi, N. Goswami, and Y. Mitsufuji, "Mmdenselstm: An efficient combination of convolutional and recurrent neural networks for audio source separation," in *Proc. Int. Workshop Acoust. Signal Enhancement*, Sep. 2018, pp. 106–110.

[4] F. R. Stöter, A. Liutkus, and N. Ito, "The 2018 signal separation evaluation campaign," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, Jul. 2018, pp. 293–305.

[5] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, May 2019.

[6] G. Hu and D. Wang, "An auditory scene analysis approach to monaural speech segregation," in *Topics Acoustic Echo and Noise Control*. Berlin, Germany: Springer, 2006, pp. 485–515.

[7] H. Kameoka, T. Nishimoto, and S. Sagayama, "A multipitch analyzer based on harmonic temporal structured clustering," *IEEE Trans. Acoust., Speech, Lang. Process.*, vol. 15, no. 3, pp. 982–994, Feb. 2007.

[8] H. Kameoka, "Statistical Approach to Multipitch Analysis," Ph.D. dissertation, The University of Tokyo, Mar. 2007.

[9] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA, USA: MIT press, 1994.

[10] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Apr. 2003, pp. 177–180.

[11] T. Virtanen and A. Klapuri, "Analysis of polyphonic audio using source-filter model and non-negative matrix factorization," in *Proc. Adv. Models Acoustic Process., Neural Inform. Process. Syst. Workshop*, Dec. 2006.

[12] H. Kameoka and K. Kashino, "Composite autoregressive system for sparse source-filter representation of speech," in *Proc. IEEE Int. Symp. Circuits and Syst.*, Jun. 2009, pp. 2477–2480.

[13] J.-L. Durrieu, G. Richard, and B. David, "An iterative approach to monaural musical mixture de-soloing," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2009, pp. 105–108.

[14] R. Hennequin, R. Badeau, and B. David, "NMF with time-frequency activations to model nonstationary audio events," *IEEE Trans. Acoust., Speech, Lang. Process.*, vol. 19, no. 4, pp. 744–753, May 2011.

[15] K. Yoshii and M. Goto, "Infinite composite autoregressive models for music signal analysis," in *Proc. Int. Soc. Music Inform. Retrieval Conf.*, Oct. 2012, pp. 79–84.

[16] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Trans. Acoust., Speech, Lang. Process.*, vol. 20, no. 4, pp. 1118–1133, May 2012.

[17] K. Yoshii, K. Itoyama, and M. Goto, "Infinite superimposed discrete all-pole modeling for multipitch analysis of wavelet spectrograms," in *Proc. Int. Soc. Music Inform. Retrieval Conf.*, 2015, pp. 86–92.

[18] T. Nakamura and H. Kameoka, "Shifted and convolutive source-filter non-negative matrix factorization for monaural audio source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2016, pp. 489–493.

[19] H. Kameoka, "Statistical speech spectrum model incorporating all-pole vocal tract model and f_0 contour generating process model," in *IEICE Tech. Rep.*, vol. 110, no. 297, Nov. 2010, pp. 29–34, in Japanese.

[20] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*. Society for Industorial and Applied Mathematics, 1970, no. 30.

[21] D. R. Hunter and K. Lange, "Quantile regression via an MM algorithm," *J. Comput. Graph. Statist.*, vol. 9, no. 1, pp. 60–77, Feb. 2000.

[22] T. Nakamura, K. Shikata, N. Takamune, and H. Kameoka, "Harmonic-temporal factor decomposition incorporating music prior information for informed monaural source separation," in *Proc. Int. Soc. Music Inform. Retrieval Conf.*, Oct. 2014, pp. 623–628.

[23] 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," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, Sep. 2010, pp. 149–156.

[24] S. A. Raczyński, N. Ono, and S. Sagayama, "Multipitch analysis with harmonic nonnegative matrix approximation," in *Proc. Int. Conf. Music Infom. Retrieval*, Sep. 2007, pp. 381–386.

[25] E. Vincent, N. Bertin, and R. Badeau, "Harmonic and inharmonic non-negative matrix factorization for polyphonic pitch transcription," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2008, pp. 109–112.

[26] K. Yoshii and M. Goto, "Infinite latent harmonic allocation: A nonparametric bayesian approach to multipitch analysis," in *Proc. Int. Soc. Music Inform. Retrieval Conf.*, Aug. 2010, pp. 309–314.

[27] D. Sakaue, T. Otsuka, K. Itoyama, and H. G. Okuno, "Bayesian non-negative harmonic-temporal factorization and its application to multipitch analysis," in *Proc. Int. Soc. Music Inform. Retrieval Conf.*, Oct. 2012, pp. 91–96.

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

[29] R. Badeau and A. Ozerov, "Multiplicative updates for modeling mixtures of non-stationary signals in the time-frequency domain," in *Proc. Eur. Signal Process. Conf.*, Sep. 2013, pp. 1–5.

[30] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *J. Amer. Statist. Assoc.*, vol. 112, no. 518, pp. 859–877, Jul. 2017.

[31] R. Hennequin, R. Badeau, and B. David, "Time-dependent parametric and harmonic templates in non-negative matrix factorization," in *Proc. Int. Conf. Digital Audio Effects*, 2010, pp. 246–253.

[32] F. Stöter, A. Liutkus, R. Badeau, B. Edler, and P. Magron, "Common fate model for unison source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 126–130.

[33] E. Creager, N. Stein, R. Badeau, and P. Depalle, "Nonnegative tensor factorization with frequency modulation cues for blind audio source separation," in *Proc. Int. Soc. Music Inform. Retrieval Conf.*, 2016, pp. 211–217.

[34] F. Pishdadian and B. Pardo, "Multi-resolution common fate transform," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 2, pp. 342–354, Feb. 2019.

[35] D. FitzGerald, M. Cranitch, and E. Coyle, "Non-negative tensor factorisation for sound source separation," in *Proc. Irish Signals Syst. Conf.*, 2005, pp. 8–12.

[36] M. Goto, "Development of the RWC Music Database," in *Proc. Int. Congr. Acoust.*, vol. 1, 2004, pp. 553–556.

[37] "FluidSynth," Accessed: Mar. 5, 2020. [Online]. Available: http://www.fluidsynth.org/

[38] H. Kameoka, T. Harada, T. Nishimoto, and S. Sagayama, "Signal processing method and unit," (in Japanese), JP Patent JP2008-281898, Nov. 20, 2008.

[39] T. Nakamura and H. Kameoka, "Fast signal reconstruction from magnitude spectroggram of continuous wavelet transform based on spectrogram consistency," in *Proc. Int. Conf. Digit. Audio Effects*, Sep. 2014, pp. 129–135.

[40] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Acoust., Speech, Langu. Process.*, vol. 14, no. 4, pp. 1462–1469, Jun. 2006.
[41] J. J. Burred and T. Sikora, "Comparison of frequency-warped representations for source separation of stereo mixtures," in *Proc. Audio Eng. Soc. Conv.*, Oct. 2006. [Online]. Available: https://www.aes.org/e-lib/online/browse.cfm?elib=13758
[42] H. Tachibana, N. Ono, and S. Sagayama, "Singing voice enhancement in monaural music signals based on two-stage harmonic/percussive sound separation on multiple resolution spectrograms," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 1, pp. 228–237, Jan. 2014.

Tomohiko Nakamura (Member, IEEE) received the B.E., M.S., and Ph.D. degrees from the University of Tokyo, Tokyo, Japan, in 2011, 2013, and 2016, respectively. He was with SECOM Intelligent Systems Laboratory, Tokyo, Japan, in 2016, and has been a Project Research Associate with the University of Tokyo, Tokyo, Japan, since 2019. He is currently a Project Research Associate with the University of Tokyo. His research interests include audio signal processing, music signal and information processing, and machine learning.

Hirokazu Kameoka (Senior Member, IEEE) received B.E., M.S., and Ph.D. degrees, all from the University of Tokyo, Japan, in 2002, 2004, and 2007, respectively. He is currently a Senior Distinguished Researcher with NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation and an Adjunct Associate Professor with the National Institute of Informatics. From 2011 to 2016, he was an Adjunct Associate Professor with the University of Tokyo. He is the author or co-author of about 150 articles in journals and peer-reviewed conference proceedings. His research interests include audio, speech, and music signal processing and machine learning. He has been an Associate Editor for the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING since 2015, a Member of IEEE Audio and Acoustic Signal Processing Technical Committee since 2017, and a Member of IEEE Machine Learning for Signal Processing Technical Committee since 2019. He was the recipient of 17 awards, including the IEEE Signal Processing Society 2008 SPS Young Author Best Paper Award.