INFINITE SUPERIMPOSED DISCRETE ALL-POLE MODELING FOR MULTIPITCH ANALYSIS OF WAVELET SPECTROGRAMS

Kazuyoshi Yoshii[1] **Katsutoshi Itoyama**[1] **Masataka Goto**[2]
[1]Graduate School of Informatics, Kyoto University, Japan
[2]National Institute of Advanced Industrial Science and Technology (AIST), Japan
{yoshii,itoyama}@kuis.kyoto-u.ac.jp m.goto@aist.go.jp

ABSTRACT

This paper presents a statistical multipich analyzer based on a source-filter model that decomposes a target music audio signal in terms of three major kinds of sound quantities: pitch (fundamental frequency: F0), timbre (spectral envelope), and intensity (amplitude). If the spectral envelope of an isolated sound is represented by an all-pole filter, linear predictive coding (LPC) can be used for filter estimation in the linear-frequency domain. The main problem of LPC is that although only the amplitudes of harmonic partials are reliable samples drawn from the spectral envelope, the whole spectrum is used for filter estimation. To solve this problem, we propose an *infinite superimposed discrete all-pole* (iSDAP) model that, given a music signal, can estimate an appropriate number of superimposed harmonic structures whose harmonic partials are drawn from a limited number of spectral envelopes. Our nonparametric Bayesian source-filter model is formulated in the log-frequency domain that better suits the frequency characteristics of human audition. Experimental results showed that the proposed model outperformed the counterpart model formulated in the linear frequency domain.

1. INTRODUCTION

Statistical modeling of music audio signals based on machine learning techniques is a hot topic in the field of music signal analysis. In particular, nonnegative matrix factorization (NMF) has often been used for multiple fundamental frequency (F0) estimation (multipitch analysis) and source separation [1–4, 7, 14, 15, 17, 21–26]. The standard NMF approximates a nonnegative spectrogram (matrix) as the product of two nonnegative matrices: a set of basis spectra and a set of the corresponding activations. An efficient multiplicative-updating (MU) algorithm was proposed for minimizing a cost function that measures the approximation error [18]. This was later found to be maximum likelihood estimation of a particular probabilistic model [5].

Statistical source-filter models, which were inspired by the simplified model of the speech production mechanism,

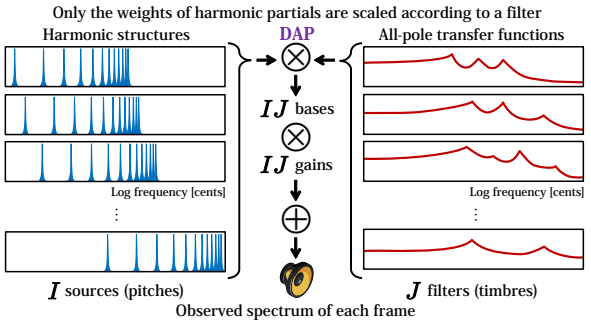

Figure 1. Overview of infinite superimposed discrete all-pole (iSDAP) modeling: We take the infinite limit as both the numbers of sources and filters, I and J, go to infinity.

have often been proposed for representing musical instrument sounds [7, 14, 25]. The pitches and timbres of musical instrument sounds are well characterized by fine structures (sources) and spectral envelopes (filters) in the frequency domain. Since the human auditory system is sensitive to spectral peaks and formants, the spectral envelope of each frame is usually modeled by an all-pole frequency transfer function (frequency response of an autoregressive (AR) filter) [14]. A classical method of all-pole spectral envelope estimation called linear predictive coding (LPC) [16] corresponds to maximum likelihood estimation of a particular probabilistic model under a strong assumption that source signals have the flat spectrum (white noise).

The composite autoregressive (CAR) modeling [17] is a promising statistical approach that overcomes the limitation of classical source-filter modeling in the framework of NMF. A given audio spectrogram is decomposed into specified numbers of fine structures (sources) and spectral envelopes (filters). A key feature of this approach is that source spectra themselves can be estimated (not limited to white noise) at the same time as all-pole spectral envelope estimation. The probabilistic interpretation of source-filter NMF makes it possible to formulate a nonparametric Bayesian extension called *infinite* CAR (iCAR) modeling that can automatically choose the appropriate numbers of sources and filters according to a given audio spectrogram [26]. Another useful extension is to restrict source spectra to harmonic structures by using parametric functions [26]. The F0s of source spectra can be estimated in a principled maximum-likelihood framework.

Conventional methods of source-filter NMF including CAR [7, 14, 17, 25, 26] have two major problems as follows:

1. All the frequency bins are taken into account for spectral envelope estimation although only the amplitudes of harmonic partials can be regarded as reliable samples from spectral envelopes.
2. Linear-frequency spectrograms given by short-time Fourier transform (STFT) are used for all-pole modeling although log-frequency spectrograms given by wavelet or constant-Q transform better suit the frequency characteristics of human audition.

To solve these problems, we propose a new statistical approach to source-filter NMF called *infinite superimposed discrete-all pole* (iSDAP) modeling. Our approach is based on a well-known technique called *discrete all-pole* (DAP) modeling [8] that takes into account only the peaks of harmonic partials for spectral envelope estimation. To deal with polyphonic music audio signals, however, we need to separate individual harmonic structures and estimate their F0s (positions of discrete harmonic partials). A major contribution of this study is to extend the DAP modeling for dealing with an arbitrary number of superimposed harmonic structures in a similar way to the iCAR modeling. This enables us to decompose a log-frequency spectrogram into appropriate numbers of pitches (F0s), timbres (spectral envelopes), and their volumes by leveraging the frequency-scale-free characteristics of the DAP modeling.

2. RELATED WORK

This section reviews probabilistic models of source-filter decomposition, NMF, and source-filter NMF as a basis of formulating the iSDAP model. Most conventional models are formulated in the linear frequency (STFT) domain.

2.1 Linear Predictive Coding (All-pole Modeling)

The linear predictive coding (LPC) [16] is a signal modeling method that can be used for estimating the spectral envelope of an observed spectrum. The underlying assumption is that the corresponding audio signal $\boldsymbol{x} = \{x_m\}_{m=-\infty}^{\infty}$ (a local signal $\{x_m\}_{m=1}^{M}$ is infinitely repeated) follows a P-order autoregressive (AR) process as follows:

$$x_m = -\sum_{p=1}^{P} a_p x_{m-p} + s_m \;\; \textit{i.e.}, \;\; \sum_{p=0}^{P} a_p x_{m-p} = s_m, \quad (1)$$

where $\boldsymbol{a} = [a_0, \cdots, a_P]^T$ is a vector of AR coefficients ($a_0 = 1$) and $\{s_m\}_{m=1}^{M}$ is a set of prediction errors. Eq. (1) can be interpreted in terms of source-filter modeling, *i.e.*, when $\boldsymbol{x}$ is a speech signal, $\boldsymbol{s}$ is an excitation signal generated by the vocal cords (source) and $\boldsymbol{a}$ represents the resonance characteristics of the vocal tract (filter).

Eq. (1) can be regarded as a linear system (governed by $\boldsymbol{a}$) that takes $\boldsymbol{s}$ as input and then gives $\boldsymbol{x}$ as output. Since Eq. (1) is a convolution of $\boldsymbol{a}$ with $\boldsymbol{x}$, we can say

$$A(z)X(z) = S(z) \;\; \textit{i.e.}, \;\; X(z) = S(z)F(z), \quad (2)$$

where $X(z)$ and $S(z)$ are the z-transforms of $\boldsymbol{x}$ and $\boldsymbol{s}$, respectively, which are given by

$$X(z) = \sum_{m=-\infty}^{\infty} x_m z^{-m} \text{ and } S(z) = \sum_{m=-\infty}^{\infty} s_m z^{-m}, \quad (3)$$

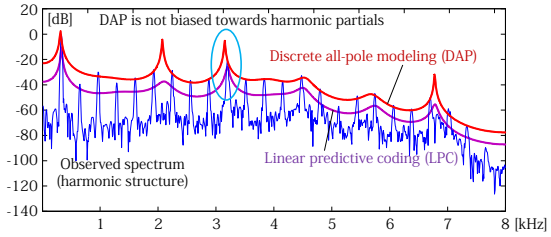

Figure 2. Spectral envelopes estimated by LPC and DAP.

and $F(z) \stackrel{\text{def}}{=} \frac{1}{A(Z)}$ is an all-pole transfer function given by

$$F(z) = \frac{1}{A(z)} = \frac{1}{\sum_{p=0}^{P} a_p z^{-p}}. \quad (4)$$

Letting $2\pi\frac{m}{M} = \omega_m$ and substituting $z = e^{i\omega_m}$ into Eq. (2), we get the Fourier-transform representation as follows:

$$X(e^{i\omega_m}) = S(e^{i\omega_m})F(e^{i\omega_m}), \quad (5)$$

where $\{X(e^{i\omega_m})\}_{m=1}^{M}$ is the complex spectrum of the observed signal $\boldsymbol{x}$, $\{S(e^{i\omega_m})\}_{m=1}^{M}$ is that of the source signal $\boldsymbol{s}$, and $\{F(e^{i\omega_m})\}_{m=1}^{M}$ is the frequency characteristics of the all-pole transfer function.

The goal of LPC is to estimate a set of AR coefficients $\boldsymbol{a}$ under a strong unrealistic assumption that the source signal $\boldsymbol{s}$ is Gaussian white noise. This means that $S(e^{i\omega_m})$ is complex Gaussian-distributed as follows:

$$S(e^{i\omega_m}) \sim \mathcal{N}_c(0, \sigma^2), \quad (6)$$

where σ^2 is the power of the white spectrum of the source signal $\boldsymbol{s}$. Using Eq. (5) and Eq. (6), we get

$$X(e^{i\omega_m}) \sim \mathcal{N}_c(0, \sigma^2|F(e^{i\omega_m})|^2). \quad (7)$$

Letting $X_m = |X(e^{i\omega_m})|^2$ and $F_m = |F(e^{i\omega_m})|^2$, we briefly rewrite Eq. (7) as follows:

$$X_m \sim \text{Exponential}(\sigma^2 F_m), \quad (8)$$

where $\{X_m\}_{m=1}^{M}$ is the *power* spectrum of the observed signal $\boldsymbol{x}$ and $\{F_m\}_{m=1}^{M}$ is the spectral envelope of $\{X_m\}_{m=1}^{M}$, as shown in Figure 2. Eq. (8) defines the probabilistic model of LPC. $\{F_m\}_{m=1}^{M}$ (*i.e.*, $\boldsymbol{a}$) and σ^2 can be estimated in a maximum-likelihood manner [16].

The main problem of LPC is that if we analyze a pitched sound derived from a periodic source signal (*e.g.*, vibration of strings), the estimated envelope $\{F_m\}_{m=1}^{M}$ loosely fits the observed spectrum $\{X_m\}_{m=1}^{M}$ and its peaks (formants) tend to be biased to the positions of harmonic partials. This is because all M frequency bins are used for all-pole modeling although in reality only the amplitudes of harmonic partials can be considered to be reliable samples from the spectral envelope.

2.2 Discrete All-pole Modeling

The discrete all-pole (DAP) modeling [8] is a well-known spectral envelope estimation method that was proposed for solving the problem of LPC. Since DAP is an extension of LPC, the probabilistic model of DAP has the same form as Eq. (8). A key feature of DAP is that Eq. (8) is defined over only a partial set of frequency bins, Ω, as follows:

$$X_m \sim \text{Exponential}(\sigma^2 F_m) \quad m \in \Omega, \quad (9)$$

where if $\Omega = \{1, \cdots, M\}$, DAP reduces to LPC. To estimate the spectral envelope of a harmonic spectrum, we can take into account only the discrete frequencies of harmonic partials. The estimated envelope passes close to the peaks of harmonic partials (Figure 2). To maximize the likelihood given by Eq. (9), an efficient algorithm was proposed for alternately optimizing $\boldsymbol{a}$ and σ^2 [8]. It was later found as a multiplicative updating algorithm [1, 14].

The main limitation of DAP is that the F0 and its overtones of an observed spectrum $\{X_m\}_{m=1}^M$ should be given in advance for defining a set of discrete frequencies to be considered, Ω. To analyze polyphonic music audio signals consisting of superimposed harmonic structures, we need to separate harmonic structures and estimate their F0s.

2.3 Composite Autoregressive Modeling

The composite autoregressive (CAR) modeling [17] is a variant of source-filter NMF that is used for decomposing a linear-frequency mixture spectrogram into I fine structures (sources) and J spectral envelopes (filters), as shown in Figure 3. Let $\boldsymbol{X}$ be an $M \times N$ power spectrogram, where M is the number of frequency bins and N is the number of frames. The nonnegative matrix $\boldsymbol{X}$ is factorized into three kinds of "factors" $\boldsymbol{S}$, $\boldsymbol{F}$, and $\boldsymbol{H}$ as follows:

$$X_{mn} \approx \sum_{i=1}^{I} \sum_{j=1}^{J} S_{im} F_{jm} H_{nij} \stackrel{\text{def}}{=} Y_{mn}, \tag{10}$$

where $\{S_{im}\}_{m=1}^M$ is the linear-frequency power spectrum of source i, $\{F_{jm}\}_{m=1}^M$ is that of filter j, and H_{nij} is the gain of a pair of source i and filter j at frame n. All these variables should be estimated from $\boldsymbol{X}$.

2.3.1 Original Formulation

The probabilistic model of CAR can be formulated by precisely modeling source signals in a statistical manner. To avoid the unrealistic assumption of LPC that each source signal is Gaussian white noise (Eq. (6)), we assume

$$S_i(e^{i\omega_m}) \sim \mathcal{N}_c\left(0, S_{im}\right), \tag{11}$$

where $\{S_i(e^{i\omega_m})\}_{m=1}^M$ is the complex spectrum of source i. Using Eq. (5) and Eq. (11), we get

$$X_{ijmn}(e^{i\omega_m}) \sim \mathcal{N}_c\left(0, S_{im} F_{jm} H_{nij}\right), \tag{12}$$

where $\{X_{ijmn}(e^{i\omega_m})\}_{m=1}^M$ is a *latent* complex spectrum generated from source i and filter j at frame n. Using the reproducing property of the Gaussian and Eq. (10), we get

$$X_{mn}(e^{i\omega_m}) \sim \mathcal{N}_c\left(0, Y_{mn}\right), \tag{13}$$

where $\{X_{mn}(e^{i\omega_m})\}_{m=1}^M$ is the *observed* complex spectrum at frame n. Eq. (13) is equivalent to

$$X_{mn} \sim \text{Exponential}\left(Y_{mn}\right), \tag{14}$$

where $\mathbb{E}[X_{mn}] = Y_{mn}$ is satisfied and $\{X_{mn}\}_{m=1}^M$ and $\{Y_{mn}\}_{m=1}^M$ are the *power* spectra of frame n.

This means that the Itakura-Saito (IS) divergence is theoretically justified as a cost function that evaluates the error between X_m and Y_m in Eq. (10) [17]. In general, however, optimization algorithms tend to get stuck in bad local minima because the IS divergence is not convex w.r.t. Y_{mn}.

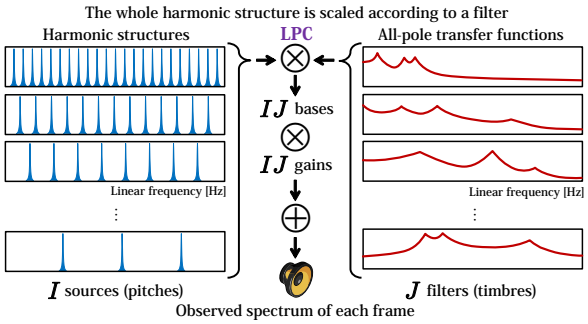

Figure 3. Overview of composite autoregressive (CAR) modeling defined in the linear frequency domain.

2.3.2 Several Extensions

Another probabilistic model of CAR was proposed by using the Kullback-Leibler (KL) divergence instead of the IS divergence as a cost function for a practical reason [26]. Instead of Eq. (14), we assume

$$X_{mn} \sim \text{Poisson}\left(Y_{mn}\right), \tag{15}$$

where $\mathbb{E}[X_{mn}] = Y_{mn}$ holds. $\{X_{mn}\}_{m=1}^M$ and $\{Y_{mn}\}_{m=1}^M$ are the *amplitude* spectra of frame n because KL-NMF models are usually formulated in the amplitude domain by assuming the amplitude additivity [10, 18].

A nonparametric Bayesian extension called *infinite* CAR enables us to automatically estimate appropriate numbers of sources and filters according to the observation $\boldsymbol{X}$ [26]. This technique is based on gamma process NMF [15].

Another extension of CAR is to force the amplitude spectrum of each source $\{S_{im}\}_{m=1}^M$ to have a harmonic structure [26]. If the source signal is a train of periodic impulses (an idealized model of the vocal chords), $\{S_{im}\}_{m=1}^M$ has a harmonic structure consisting of equally-spaced harmonic partials with the same weight. The optimal value of the F0 can be estimated such that the likelihood given by Eq. (15) is maximized. This technique of F0 estimation has a potential to solve the limitation of DAP.

3. PROPOSED MODEL

This section presents a nonparametric Bayesian approach called *infinite superimposed discrete all-pole* (iSDAP) modeling for source-filter decomposition of wavelet spectrograms. Our model can estimate multiple F0s at each frame and discover several kinds of instrument timbres (all-pole spectral envelopes) from polyphonic music audio signals. To achieve this, we integrate the technique of discrete all-pole (DAP) modeling [8] into the framework of composite autoregressive (CAR) modeling [17, 26] in a probabilistic manner. The iSDAP model can be regarded as a Bayesian extension of log-frequency source-filter NMF based on a single filter [19], and has all of the following features:

1. **Superimposed DAP modeling**: Our model can estimate the spectral envelope of each of harmonic structure contained in mixed sounds. The original DAP model can deal with only isolated sounds [8].
2. **Precise F0 modeling**: Each frame is allowed to contain a unique set of F0s (sources) for capturing fine

fluctuations of F0s (*e.g.*, vibrato). The original CAR models [17, 26] assume that a common set of source spectra (semitone-level F0s) is shared over all frames.

3. **Log-frequency modeling**: Our source-filter model can deal with wavelet spectrograms that suit the characteristics of human audition by leveraging an advantage of DAP modeling that only discrete frequencies are required for spectral envelope estimation.
4. **Bayesian nonparametrics**: Our model can estimate effective numbers of sources and filters according to a given spectrogram by allowing unbounded (infinite in theory) numbers of sources and filters to be used.

3.1 Model Formulation

We explain a probabilistic model of iSDAP. Let $\boldsymbol{X}$ be an $M \times N$ log-frequency amplitude spectrogram with M frequency bins and N frames. The nonnegative matrix $\boldsymbol{X}$ is factorized in a similar way to Eq. (10) as follows:

$$X_{mn} \sim \text{Poisson}\left(\sum_{i=1}^{I\to\infty}\sum_{j=1}^{J\to\infty}\theta_{ni}\phi_j W_{nijm}H_{nij}\right), \quad (16)$$

where θ_{ni} is the local weight of source i at frame n, ϕ_j is the global weight of filter j, and H_{nij} is the gain of a pair of source i and filter j at frame n. $\{W_{nijm}\}_{m=1}^{M}$ is the amplitude spectrum derived from the source-filter pair at frame n. Note that θ_{ni} and W_{nijm} are allowed to vary over time to represent the F0 fluctuation unlike Eq. (10). We aim to perform sparse learning of weight vectors $\boldsymbol{\theta}_n = [\theta_{n1}, \cdots, \theta_{nI}]^T$ and $\boldsymbol{\phi} = [\phi_1, \cdots, \phi_J]^T$ when the number of sources I and the number of filters J go to infinity.

3.1.1 Parametric Functions

As shown in Figure 4, we force the amplitude spectrum $\{W_{nijm}\}_{m=1}^{M}$ to have a harmonic structure as follows:

$$W_{nijm} = \sum_{r=1}^{R} S_{mnir}F_{nijr}, \quad (17)$$

where R is the number of harmonic partials and $\{S_{mnir}\}_{m=1}^{M}$ is the monomodal spectrum of the r-th harmonic partial of source i at frame n given by

$$S_{mnir} = \exp\left(-\frac{1}{2\sigma^2}\left(f_m - (\mu_{ni} + 1200\log_2 r)\right)^2\right), \quad (18)$$

where μ_{ni} is the F0 [cents] of source i at frame n, f_m is the log-frequency [cents] corresponding to the m-th bin, and σ^2 indicates energy diffusion around harmonic partials.

We then represent the weights of discrete harmonic partials, $\{F_{nijr}\}_{r=1}^{R}$, by using an all-pole transfer function in the log frequency domain as follows:

$$F_{nijr} = \frac{1}{\left|\sum_{p=0}^{P} a_{jp}e^{-\omega_{nir}pi}\right|} = \left(\boldsymbol{a}_j^T\boldsymbol{U}(\omega_{nir})\boldsymbol{a}_j\right)^{-\frac{1}{2}}, \quad (19)$$

where $\boldsymbol{a}_j \equiv [a_{j0}, \cdots, a_{jP}]^T$, ω_{nir} is a normalized frequency [rad] corresponding to the r-th harmonic partial of source i at frame n, and $\boldsymbol{U}(\omega)$ is a $(P+1)\times(P+1)$ matrix with $[\boldsymbol{U}(\omega)]_{pq} = \cos(\omega(p-q))$. Note that F_{nijr} indicates the value of amplitude (not power). The Poisson likelihood (KL-NMF) is considered to fit the amplitude domain rather than the power domain [19].

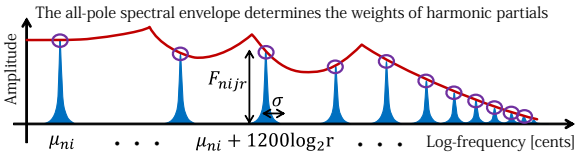

Figure 4. Composition of source i and filter j at frame n in the log-frequency domain.

3.1.2 Prior Distributions

We put gamma process (GaP) priors on infinite-dimensional vectors $\boldsymbol{\theta}_n$ and $\boldsymbol{\phi}$ as in [15, 26]. Specifically, we put independent gamma priors on elements of $\boldsymbol{\theta}_n$ and $\boldsymbol{\phi}$ as follows:

$$\theta_{ni} \sim \text{Gamma}\left(\frac{\alpha_\theta}{I}, \alpha_\theta\right), \quad \phi_j \sim \text{Gamma}\left(\frac{\alpha_\phi}{J}, \alpha_\phi\right), \quad (20)$$

where α_θ and α_ϕ are hyperparameters called concentration parameters. As J diverges to infinity, the vector $\boldsymbol{\phi}$ approximates an infinite vector drawn from a GaP with α_ϕ. It is proven that the effective number of filters, J^+, such that $\phi_j > \epsilon$ for some number $\epsilon > 0$ is almost surely finite [15]. If J is sufficiently larger than α_ϕ (J is often called a truncation level in weak-limit approximation to infinite modeling), the GaP can be well approximated. The same reasoning can be applied to the GaP on $\boldsymbol{\theta}_n$. On the other hand, we put a standard Gamma prior on H_{nij} as follows:

$$H_{nij} \sim \text{Gamma}(a_H, b_H), \quad (21)$$

where a_H and b_H are shape and rate hyperparameters.

3.2 Variational Inference

The posterior over random variables $p(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{H}|\boldsymbol{X};\boldsymbol{\mu},\boldsymbol{a})$ and parameters $\boldsymbol{\mu}$ and $\boldsymbol{a}$ are determined such that a *lower bound* $\mathcal{L}$ of the log-evidence $\log p(\boldsymbol{X};\boldsymbol{\mu},\boldsymbol{a})$ is maximized. Since this cannot be analytically computed, we use an approximate method called variational Bayes (VB), which restricts the posterior to a factorized form given by

$$q(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{H}) = \prod_{ni} q(\theta_{ni}) \prod_{j} q(\phi_j) \prod_{nij} q(H_{nij}). \quad (22)$$

Iteratively updating this posterior, we can monotonically increase a lower bound of the log-evidence given by

$$\begin{aligned}\log p(\boldsymbol{X};\boldsymbol{\mu},\boldsymbol{a}) &\geq \mathbb{E}[\log p(\boldsymbol{X}|\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{H};\boldsymbol{\mu},\boldsymbol{a})] \\ &+ \mathbb{E}[\log p(\boldsymbol{\theta})] + \mathbb{E}[\log p(\boldsymbol{\phi})] + \mathbb{E}[\log p(\boldsymbol{H})] \\ &- \mathbb{E}[\log q(\boldsymbol{\theta})] - \mathbb{E}[\log q(\boldsymbol{\phi})] - \mathbb{E}[\log q(\boldsymbol{H})] \equiv \mathcal{L}_0, \quad (23)\end{aligned}$$

where the first term can be further lower bounded by Jensen's inequality on the concave logarithmic function as follows:

$$\begin{aligned}&\mathbb{E}[\log p(\boldsymbol{X}|\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{H};\boldsymbol{\mu},\boldsymbol{a})] \\ &= \textstyle\sum_{mn} X_{mn}\mathbb{E}\left[\log\sum_{ijr}\theta_{ni}\phi_j S_{mnir}F_{nijr}H_{nij}\right] \\ &\quad - \textstyle\sum_{mnijr}\mathbb{E}\left[\theta_{ni}\phi_j S_{mnir}F_{nijr}H_{nij}\right] + \text{const.} \\ &\geq \textstyle\sum_{mnijr}\lambda_{mnijr}X_{mn}\mathbb{E}\left[\log\frac{\theta_{ni}\phi_j S_{mnir}F_{nijr}H_{nij}}{\lambda_{mnijr}}\right] \\ &\quad - \textstyle\sum_{mnijr}\mathbb{E}\left[\theta_{ni}\phi_j S_{mnir}F_{nijr}H_{nij}\right] + \text{const.} \quad (24)\end{aligned}$$

where λ_{mnijr} is a normalized auxiliary variable such that $\sum_{ijr} \lambda_{mnijr} = 1$. The equality holds (*i.e.*, the lower bound of $\mathcal{L}_0$ is maximized) if and only if

$$\lambda_{mnijr} \propto \exp(\mathbb{E}[\log(\theta_{ni}\phi_j S_{mnir} F_{nijr} H_{nij})]). \quad (25)$$

Using Eq. (24), the objective function of our model to be maximized, $\mathcal{L}$, is obtained as the lower bound of $\mathcal{L}_0$. For convenience, we define X_{mnijr} and Y_{mnijr} as

$$X_{mnijr} = \lambda_{mnijr} X_{mn}, \quad (26)$$
$$Y_{mnijr} = \mathbb{E}[\theta_{ni}\phi_j S_{mnir} F_{nijr} H_{nij}]. \quad (27)$$

3.3 Variational Bayesian Updating of $\boldsymbol{\theta}$, $\boldsymbol{\phi}$, and $\boldsymbol{H}$

The VB updating rules are given by

$$\begin{aligned} q(\boldsymbol{\theta}) &\propto \exp(\mathbb{E}_{q(\boldsymbol{\phi},\boldsymbol{H})}[\log p(\boldsymbol{X},\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{H};\boldsymbol{\mu},\boldsymbol{a})]), \\ q(\boldsymbol{\phi}) &\propto \exp(\mathbb{E}_{q(\boldsymbol{\theta},\boldsymbol{H})}[\log p(\boldsymbol{X},\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{H};\boldsymbol{\mu},\boldsymbol{a})]), \\ q(\boldsymbol{H}) &\propto \exp(\mathbb{E}_{q(\boldsymbol{\theta},\boldsymbol{\phi})}[\log p(\boldsymbol{X},\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{H};\boldsymbol{\mu},\boldsymbol{a})]). \end{aligned} \quad (28)$$

The variational posterior of each random variable is set to be the same family as its prior distribution as follows:

$$\begin{aligned} q(\theta_{ni}) &= \text{Gamma}(a^{\theta}_{ni}, b^{\theta}_{ni}), \quad q(\phi_j) = \text{Gamma}(a^{\phi}_{j}, b^{\phi}_{j}), \\ q(H_{nij}) &= \text{Gamma}(a^{H}_{nij}, b^{H}_{nij}). \end{aligned} \quad (29)$$

The variational parameters are given by

$$a^{\theta}_{ni} = \frac{\alpha_\theta}{I} + \sum_{mjr} X_{mnijr},\ b^{\theta}_{ni} = \alpha_\theta + \sum_{mjr} \mathbb{E}[\phi_j H_{nij}] W_{nijm},$$
$$a^{\phi}_{j} = \frac{\alpha_\phi}{J} + \sum_{mnir} X_{mnijr},\ b^{\phi}_{j} = \alpha_\phi + \sum_{mnir} \mathbb{E}[\theta_{ni} H_{nij}] W_{nijm},$$
$$a^{H}_{nij} = a_H + \sum_{mr} X_{mnijr},\ b^{H}_{nij} = b_H + \sum_{mr} \mathbb{E}[\theta_{ni}\phi_j] W_{nijm}.$$

To estimate the effective number of filters J^+, we perform sparse learning. If $\mathbb{E}[\phi_j]$ becomes sufficiently small for some filter j, we degenerate it and $J \leftarrow J-1$. A similar treatment is applied to $\mathbb{E}[\theta_{ni}]$. The proposed variational algorithm is gradually accelerated per iteration.

3.4 Multiplicative Updating of $\boldsymbol{\mu}$ and $\boldsymbol{a}$

To estimate parameters $\boldsymbol{\mu}$ and $\boldsymbol{a}$, we use the multiplicative update (MU) algorithm as in [1,14]. In general, to optimize x, we represent the partial derivative of a "cost" function $\mathcal{C}$ with respect to x as the difference of two positive terms, *i.e.*, $\frac{\partial \mathcal{C}}{\partial x} = R - R'$. An updating rule of x is then given by $x \leftarrow \frac{R'}{R} x$. Note that x becomes constant if the derivative is zero, and is updated in the opposite direction of the derivative. In this study the cost function is the negative lower bound of the log-evidence, $-\mathcal{L}$.

First, we represent the partial derivative of $-\mathcal{L}$ with respect to μ_{mi} as $\frac{-\partial \mathcal{L}}{\partial \mu_{ni}} = R_{ni} - R'_{ni}$, where R_{ni} and R'_{ni} are positive terms given by

$$R_{ni} = \sum_{mjr} (\mu_{ni} + 1200 \log_2 r) X_{mnijr} + f_m Y_{mnijr}, \quad (30)$$
$$R'_{ni} = \sum_{mjr} f_m X_{mnijr} + (\mu_{ni} + 1200 \log_2 r) Y_{mnijr}, \quad (31)$$

The updating rule of μ_{ni} is given by

$$\mu_{ni} \leftarrow R^{-1}_{ni} R'_{ni} \mu_{ni}. \quad (32)$$

As in [1, 14], we then represent the partial derivative of $-\mathcal{L}$ with respect to $\boldsymbol{a}_j$ as $\frac{-\partial \mathcal{L}}{\partial \boldsymbol{a}_j} = (\boldsymbol{R}_j - \boldsymbol{R}'_j)\boldsymbol{a}_j$, where $\boldsymbol{R}_j$ and $\boldsymbol{R}'_j$ are positive definite matrices given by

$$\boldsymbol{R}_j = \sum_{mnir} X_{mnijr} F^2_{nijr} \boldsymbol{U}(\omega_{nir}), \quad (33)$$
$$\boldsymbol{R}'_j = \sum_{mnir} Y_{mnijr} F^2_{nijr} \boldsymbol{U}(\omega_{nir}). \quad (34)$$

The updating rule of $\boldsymbol{a}_j$ is given by

$$\boldsymbol{a}_j \leftarrow \boldsymbol{R}^{-1}_j \boldsymbol{R}'_j \boldsymbol{a}_j. \quad (35)$$

Finally, we forcibly adjust the scale of the filter F_{nijr} such that $\alpha_{j0} = 1$ for normalizing the filter. Although this step violates the convergence of the optimization algorithm, it was empirically found to work well.

3.5 Binary Piano-roll Estimation

To perform multipitch analysis, *i.e.*, make a binary piano-roll representation, we need to judge the existence of each semitone-level pitch at each frame. Using a trained model, we calculate an activation matrix $\boldsymbol{V} = \{V_{kn}\}^{88,N}_{k=1,n=1}$ over pitch k and frame n (continuous-valued piano-roll representation *e.g.*, the middle figure of Figure 5) by accumulating the expected amplitude of the first partial of source i, $\sum_j \mathbb{E}[\theta_{ni}\phi_j F_{nij1} H_{nij}]$, into V_{kn} indicated by μ_{ni}. Finally, the activation matrix $\boldsymbol{V}$ is normalized such that all the elements sum to unity, *i.e.*, $\sum_{kn} V_{kn} = 1$.

There are several approaches to binary piano-roll estimation. The common approach is to make a binary decision based on a threshold η. Another approach is to define a hidden Markov model (HMM) and use the Viterbi-search algorithm for estimating a sequence of hidden binary states $\{Z_{kn}\}^N_{n=1}$ from a sequence of pitch-existence likelihoods $\{V^p_{kn}\}^N_{n=1}$ for each pitch k, where p controls the dynamic range. In our implementation, $p = 0.2$ and the transition matrix is $[0.8, 0.2; 0.01, 0.99]$ in the Matlab notation.

4. EVALUATION

We report comparative experiments that were conducted for evaluating the performance of the iSDAP model in multipitch analysis of piano music. Since the proposed model assumes that input mixture signals contain only harmonic sounds, we also tested the use of harmonic and percussive source separation (HPSS) [12] as a preprocessor.

4.1 Experimental Conditions

We used 30 pieces (labeled as "ENSTDkCl") selected from the MAPS database [9] that contain stereo signals sampled at 44.1 [kHz]. The audio signals were converted to monaural signals and truncated to 30 [s] from the beginning as in [2, 4, 21, 22, 24]. The amplitude spectrogram of each piece over the frequency bins ranging from 0 [cents] (16.325 [Hz]) to 12000 [cents] (16717 [Hz]) was obtained by performing the wavelet transform with a Gabor wavelet, a frequency interval of 10 [cents], and a shifting interval of 10 [ms], *i.e.*, $M = 1200$ and $N = 3000$. The other quantities were $I = 88$, $J = 3$, $R = 20$, $P = 13$, and $\sigma = 25$. The priors were set to be less informative, *i.e.*,

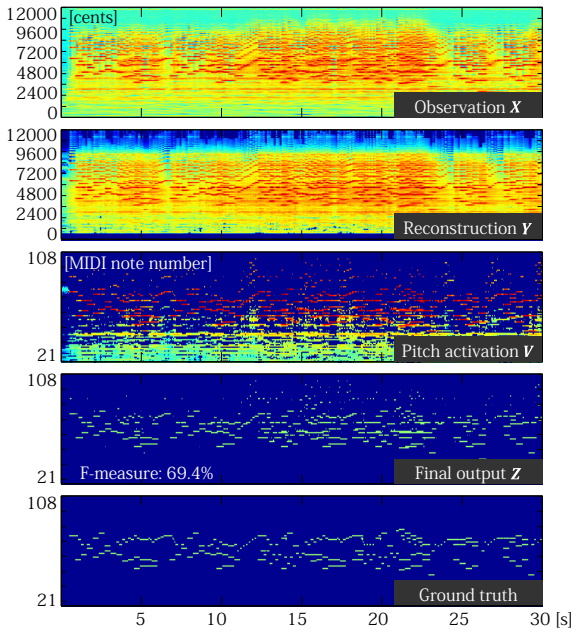

Figure 5. Analysis of MUS-mz_333_3_ENSTDkCl.

$\alpha_\theta = \alpha_\phi = a_H = 1$, and $b_H = \mathbb{E}_{\text{emp}}[X_{mn}]^{-1}$. Since $\boldsymbol{X}$ contained only piano sounds, the truncation level $J = 3$ worked well (two filters were degenerated in this experiment, *i.e.*, $J^+ = 1$). The values of $\{\mu_{ni}\}_{i=1}^I$ were initialized as the frequencies corresponding to the 88 keys ranging from 900 [cents] to 9600 [cents]. The value of each α_{jp} $(1 \le p \le P)$ was drawn from a Gaussian with a zero mean and a small variance of 0.01. The variational posteriors were initialized as the corresponding priors.

The proposed model was tested under possible combinations of preprocessing (with or without HPSS) and postprocessing (thresholding or Viterbi decoding). HPSS was performed in the log-frequency domain. The model with a single filter ($J = J^+ = 1$) was also tested in a supervised setting. A set of filter coefficients $\boldsymbol{a}_1$ was pretrained from 264 isolated sounds of the same or different piano (ENSTDkCl in a closed test or SptkBGCl in an open test) by using LPC, and kept constant during multipitch analysis.

The estimation results were evaluated in terms of the frame-level recall/precision rates and F-measure as in [24]:

$$\mathcal{R} = \frac{\sum_n c_n}{\sum_n r_n}, \quad \mathcal{P} = \frac{\sum_n c_n}{\sum_n e_n}, \quad \mathcal{F} = \frac{2\mathcal{R}\mathcal{P}}{\mathcal{R}+\mathcal{P}}, \qquad (36)$$

where r_n, e_n, and c_n are the numbers of ground truth, estimated and correct pitches on frame n, respectively. The threshold η was determined as $\eta = 10^{-1.3}$ without HPSS and $\eta = 10^{-1.5}$ with HPSS.

4.2 Experimental Results

The experimental results shown in Figure 5 and Table 1 indicate the great potential of the iSDAP model. The model supervised in the open condition (67.3%) significantly outperformed the iCAR model formulated in the linear frequency domain (48.4%) [26] and tied with the state-of-the-art methods, *e.g.*, harmonic NMF (67.7%) [24], NMF with group sparsity (71.3%) [21], and NMF with Hellinger

Filter learning	HPSS	HMM	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{F}$
Unsupervised			55.3	57.9	56.6
		✓	62.2	60.2	61.2
	✓		62.4	64.3	63.4
	✓	✓	**67.4**	**64.2**	**65.8**
Supervised	✓		62.4	67.0	64.4
(open test)	✓	✓	69.9	64.5	67.3
Supervised	✓		59.4	69.1	63.9
(close test)	✓	✓	67.4	67.8	67.6

Table 1. Experimental results of multipitch analysis for 30 piano pieces labeled as ENSTDkCl.

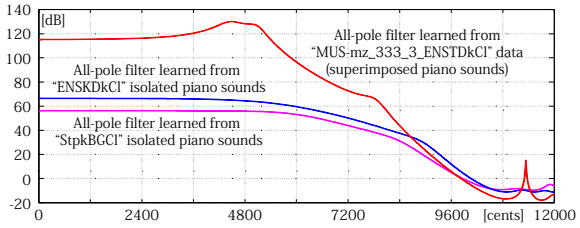

Figure 6. All-pole filters learned from isolated sounds or a piano piece (mixed sounds) in the log-frequency domain.

sparse coding (66.5%) [22]). While many recent methods need to pretrain a dictionary of basis spectra for reasonable decomposition [2,4,21,24], our model works well (65.8%) even in the completely unsupervised condition. As shown in Figure 6, a filter learned from a music signal dropped faster than the pretrained filters because the model failed to capture higher-order overtones even in the log-frequency domain due to the strong inharmonicity of piano sounds. Nonetheless, the learned filter acted as an effective constraint on the relative weights of harmonic partials.

There would be much room for improving the performance. KL-NMF [18] and IS-NMF [10] are special cases of β-divergence NMF [11, 20] with $\beta = 1, 0$, respectively. It was reported that the use of an intermediate divergence with $\beta = 0.5$ significantly improves the performance by about 5% [24]. Similar findings were reported in the context of source separation [13]. This calls for the use of the Tweedie likelihood instead of the Poisson likelihood [6].

5. CONCLUSION

We presented a new nonparametric Bayesian approach to source-filter NMF called infinite superimposed discrete all-pole (iSDAP) modeling that can decompose a *wavelet* spectrogram into three kinds of factors, *i.e.*, harmonic sources, all-pole filters, and time-varying gains of source-filter pairs. Our model clearly outperformed its counterpart called the iCAR model formulated in the linear frequency domain. One important research direction is to build a unified model of harmonic and percussive sounds. To bridge the gap between multipitch analysis and music transcription, we plan to incorporate a prior distribution on the time-frequency positions of musical notes into a Bayesian framework.

Acknowledgment: This study was partially supported by JST OngaCREST Project, JSPS KAKENHI 24220006, 26700020, and 26280089, and Kayamori Foundation.

6. REFERENCES

[1] R. Badeau and A. Ozerov. Multiplicative updates for modeling mixtures of non-stationary signals in the time-frequency domain. In *European Signal Processing Conference (EUSIPCO)*, 2013.

[2] E. Benetos, R. Badeau, T. Weyde, and G. Richard. Template adaptation for improving automatic music transcription. In *International Society for Music Information Retrieval Conf. (ISMIR)*, pages 175–180, 2014.

[3] N. J. Bryan, G. Mysore, and G. Wang. Source separation of polyphonic music with interactive user-feedback on a piano roll display. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 119–124, 2013.

[4] J. J. Carabias-Orti, T. Virtanen, P. Vera-Candeas, N. Ruiz-Reyes, and F. J. Cañadas-Quesada. Musical instrument sound multi-excitation model for non-negative spectrogram factorization. *IEEE Journal of Selected Topics in Signal Proc.*, 5(6):1144–1158, 2011.

[5] A. T. Cemgil. Bayesian inference for nonnegative matrix factorisation models. *Computational Intelligence and Neuroscience*, 2009:Article ID 785152, 2009.

[6] U. Şimşekli, A. Cemgil, and Y. K. Yılmaz. Learning the β-divergence in Tweedie compound Poisson matrix factorization models. In *International Conference on Machine Learning (ICML)*, pages 1409–1417, 2013.

[7] J.-L. Durrieu, G. Richard, B. David, and C. Févotte. Source/filter model for unsupervised main melody extraction from polyphonic audio signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):564–575, 2010.

[8] A. El-Jaroudi and J. Makhoul. Discrete all-pole modeling. *IEEE Transactions on Signal Processing*, 39(2):411–423, 1991.

[9] V. Emiya, R. Badeau, and B. David. Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *IEEE Transactions on Audio, Speech and Language Processing*, 18(6):1643–1654, 2010.

[10] C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis. *Neural Computation*, 21(3):793–830, 2009.

[11] C. Févotte and J. Idier. Algorithms for nonnegative matrix factorization with the beta-divergence. *Neural Computation*, 23(9):2421–2456, 2011.

[12] D. FitzGerald. Harmonic/percussive separation using median filtering. In *International Conference on Digital Audio Effects (DAFx)*, 2010.

[13] D. FitzGerald, M. Cranitch, and E. Coyle. On the use of the beta divergence for musical source separation. In *Irish Signals and Systems Conf.*, pages 1–6, 2008.

[14] R. Hennequin, R. Badeau, and B. David. NMF with time-frequency activations to model nonstationary audio events. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):744–753, 2011.

[15] M. Hoffman, D. Blei, and P. Cook. Bayesian nonparametric matrix factorization for recorded music. In *International Conference on Machine Learning (ICML)*, pages 439–446, 2010.

[16] F. Itakura and S. Saito. Analysis synthesis telephony based on the maximum likelihood method. In *International Congress on Acoustics (ICA)*, pages C17–C20, 1968.

[17] H. Kameoka and K. Kashino. Composite autoregressive system for sparse source-filter representation of speech. In *IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 2477–2480, 2009.

[18] D. Lee and H. Seung. Algorithms for non-negative matrix factorization. In *Neural Information Processing Systems (NIPS)*, pages 556–562, 2000.

[19] T. Nakamura, K. Shikata, N. Takamune, and H. Kameoka. Harmonic-temporal factor decomposition incorporating music prior information for informed monaural source separation. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 623–628, 2014.

[20] M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, and S. Sagayama. Convergence-guaranteed multiplicative algorithms for non-negative matrix factorization with beta divergence. In *International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 283–288, 2010.

[21] K. O'Hanlon and M. D. Plumbley. Polyphonic piano transcription using non-negative matrix factorisation with group sparsity. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2214–2218, 2014.

[22] K. O'Hanlon, M. Sandler, and M. D. Plumbley. Matrix factorisation incorporating greedy Hellinger sparse coding applied to polyphonic music transcription. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3112–3116, 2015.

[23] P. Smaragdis, C. Févotte, G. Mysore, N. Mohammadiha, and M. Hoffman. Dynamic source separation using nonnegative factorizations: A unified view. *IEEE Signal Processing Magazine*, 31(3):66–75, 2014.

[24] E. Vincent, N. Bertin, and R. Badeau. Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3):528–537, 2010.

[25] T. Virtanen and A. Klapuri. Analysis of polyphonic audio using source-filter model and non-negative matrix factorization. In *NIPS Workshop on Advances in Models for Acoustic Processing*, 2009.

[26] K. Yoshii and M. Goto. Infinite composite autoregressive models for music signal analysis. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 79–84, 2012.