INFINITE COMPOSITE AUTOREGRESSIVE MODELS FOR MUSIC SIGNAL ANALYSIS

Kazuyoshi Yoshii **Masataka Goto**
National Institute of Advanced Industrial Science and Technology (AIST), Japan
{k.yoshii, m.goto}@aist.go.jp

ABSTRACT

This paper presents novel probabilistic models that can be used to estimate multiple fundamental frequencies (F0s) from polyphonic audio signals. These models are nonparametric Bayesian extensions of nonnegative matrix factorization (NMF) based on the source-filter paradigm, and in them an amplitude or power spectrogram is decomposed as the product of two kinds of spectral atoms (sources and filters) and time-varying gains of source-filter pairs. In this study we model musical instruments as autoregressive systems that combine two types of sources—periodic signals (comb-shaped densities) and white noise (flat density)—with all-pole filters representing resonance characteristics. One of the main problems with such *composite autoregressive models* (CARMs) is that the numbers of sources and filters should be given in advance. To solve this problem, we propose nonparametric Bayesian models based on gamma processes and efficient variational and multiplicative learning algorithms. These *infinite* CARMs (iCARMs) can discover appropriate numbers of sources and filters in a data-driven manner. We report the experimental results of multipitch analysis on the MAPS piano database.

1. INTRODUCTION

Multiple fundamental frequency estimation (a.k.a. *multipitch analysis*) is the basis of various kinds of music content analysis. Recently, nonnegative matrix factorization (NMF) has gained a lot of popularity [1–13]. The standard NMF approximates an amplitude or power spectrogram (nonnegative matrix) as the product of two nonnegative matrices, one of which is a compact set of spectral bases and the other of which is a set of the corresponding time-varying gains [15, 16]. Such low-rank approximation is well justified by the fact that each musical piece consists of only limited kinds of sounds that repeatedly appear. In addition, a practical advantage of NMF is that the bases and gains can be alternately optimized by using efficient iterative algorithms called *multiplicative update* (MU) rules. The standard NMF, however, has three fundamental limitations:

1. The spectral bases are time-invariant, and only their gains vary over time. A large number of *independent* bases are needed to fully represent the timbral variations of instrument spectra (e.g., envelopes) even if these spectra share the same fundamental frequency (F0). Such an unconstrained increase of model complexity is likely to result in optimization algorithms easily getting stuck in bad local optima.
2. A post-processing step for estimating the F0s from individual bases is required because the F0s are not parameterized for representing the spectral bases. If the shapes of spectral bases are unconstrained, the resulting bases often deviate from natural harmonic spectra. This makes F0 estimation difficult and we need to judge the existence of an F0.
3. The number of bases (model complexity) should be carefully specified in advance because it has a strong impact on the decomposition results. Note that this limitation is closely related to the first. A naive solution is to exhaustively test all possible complexities and find an optimal value, but such *model selection* is often computationally impractical.

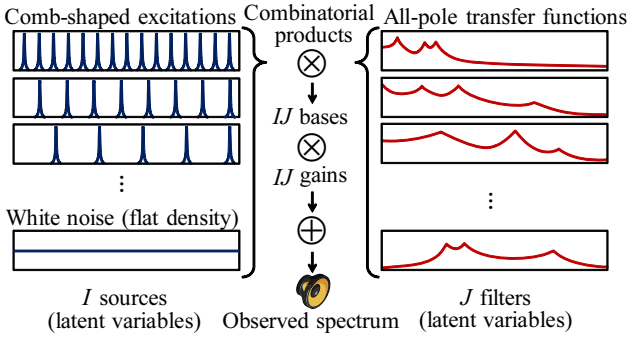

Figure 1. Overview of composite autoregressive models: The combinatorial products of I sources and J filters yield IJ spectral bases, which are activated according to the corresponding time-varying gains at each frame. We take the infinite limit as both I and J diverge to infinity.

As noted above, unconstrained NMF is too flexible for a set of musically-meaningful bases to be induced automatically. Although these limitations have partially been dealt with in previous studies [1–13], no study has overcome all of them simultaneously in a principled manner.

In this paper we propose *infinite composite autoregressive models* (iCARMs) (Fig. 1) developed for fusing the following techniques into a unified Bayesian framework:

1. **Source-filter factorization** (inspired by [1])
   We further factorize the spectral bases as the combinatorial products of sources and all-pole filters. This

idea originates in the autoregressive (AR) modeling of speech signals: various vowels can be generated by changing the shape of the vocal tract (filter) while keeping the same F0 (source). This factorization enables us to represent a wide variety of instrumental sounds in terms of two separate aspects (timbre and F0) with reasonable complexity.

2. **Harmonicity modeling** (inspired by [7] and [9])
   We represent each source as a comb-shaped function that uses an F0 parameter for representing equally-spanned harmonic partials of the same weight. Since such sources are multiplied by acoustically-inspired AR filters, the relative weights of partials of the bases are constrained to take realistic values as natural harmonic sounds. In addition, we can directly optimize the values of the F0s jointly with decomposition.
   As proposed in [7, 14], we additionally introduce a special source representing white noise (flat density). This enables us to deal with percussive and transient sounds having widely distributed spectra. Their timbres (envelopes) are characterized by AR filters.
3. **Bayesian nonparametrics** (inspired by [12])
   We build nonparametric Bayesian models that can automatically adjust the numbers of sources and filters needed to factorize a given spectrogram. Rather than these numbers being specified, the infinite limit of the conventional source-filter NMF [1] is taken as the numbers of sources and filters diverge to infinity. We perform sparse learning by introducing infinite-dimensional priors in such a way that only limited numbers of sources and filters are actually activated.

To optimize the iCARMs, we propose a new class of iterative algorithms that integrates a variational Bayesian (VB) technique with standard MU rules [8, 9].

The rest of this paper is organized as follows: Section 2 discusses the positioning of this study. Section 3 presents the iCARMs. Section 4 describes the evaluation. Section 5 concludes the paper with a mention of future work.

2. RELATED WORK

This section introduces two machine-learning (ML) techniques, i.e., NMF and Bayesian nonparametrics.

2.1 Nonnegative Matrix Factorization

NMF is a powerful tool for sparse decomposition of nonnegative matrix data [15]. It was first used for representing face images as linear combinations of a compact set of basis images corresponding to "local parts" such as eyes and noses. Such parts-based sparse representation is spontaneously induced by the nonnegativity constraint that allows only summation of basis images. Therefore, NMF fits naturally into audio spectrogram decomposition because the energy of harmonic sounds is concentrated at the discrete frequencies of harmonic partials.

2.1.1 Optimization Criteria

To perform NMF, we need some criterion for evaluating the "goodness-of-fit" of reconstructed data (linear combinations of spectral bases) to observed data (a spectrogram).

Method	Divergence	Sources	#	Filters	#
Kameoka [1]	IS	-	I	AR	J
Badeau [2]	IS	H	I	MA	J
Durrieu [3]	IS	(H)	I	-	J
Virtanen [4]	KL	-	I	-	J
Carabias-Orti [5]	KL	H	I	-	J
Heittola [6]	KL	-	I^*	-	J
Yasuraoka [7]	KL	H + N	I^*	AR	J
Hennequin [8]	Beta (0.5)	-	I	ARMA	J^*
Proposed	KL or IS	H + N	∞	AR	∞

(H: harmonic sources, N: noise source, -: others, $*$: varying over time)

Table 1. Several variants of source-filter NMF

As shown in Table 1, for example, Kullback-Leibler (KL) [15] and Itakura-Saito (IS) [16] divergences have been used intensively. Some studies used beta divergence [17], which includes KL and IS divergences as special cases.

In the context of audio modeling, although IS-NMF is justified in theory (see [16] and Section 3.2.1), KL-NMF often yields better results in a maximum likelihood estimation setting. One main reason is that the nonconvexity of IS divergence makes it difficult for gradient-descent-type optimization algorithms to find global optima. Note that no comparative tests have been conducted under a Bayesian estimation setting. In this paper we formulate two kinds of iCARMs, i.e., KL-iCARM and IS-iCARM.

2.1.2 Source-Filter Factorization

One extension is obtained with the source-filter paradigm, as listed in Table1. Kameoka and Kashino [1], for example, originally proposed the idea of the composite autoregressive model (CARM) using fixed numbers of unconstrained sources and autoregressive (AR) filters (all-pole transfer functions). Although similar models were devised by some researchers [3–5], filters were not acoustically constrained. Badeau *et al.* [2] used moving-average (MA) filters (all-zero transfer functions) with harmonic sources.

Some NMF variants allow sources or filters to vary over time to richly capture temporal variations of spectral bases at the cost of increasing complexity. Heittola *et al.* [6] and Yasuraoka and Okuno [7] used time-varying sources while a fixed number of filters was shared over time. Hennequin *et al.* [8], on the other hand, used time-varying ARMA filters that could be estimated by efficient MU rules.

2.1.3 Harmonicity Modeling

Another extension is based on harmonicity constraints on spectral bases or sources. For example, Vincent *et al.* [10] and Bertin *et al.* [11] assumed each basis as a weighted sum of narrowband template spectra consisting of a few adjacent harmonic partials. In the source-filter paradigm, Badeau *et al.* [2] represented each source as a binary vector whose elements are determined by independent Bernoulli trials, where particular elements corresponding to harmonic partials are more likely to take the value of 1. Yasuraoka and Okuno [7] and Hennequin *et al.* [9] represented each source as a parametric function based on a (weighted) sum of atomic functions (e.g., Gaussian functions) corresponding to harmonic partials. Carabias-Orti *et al.* [5] proposed to further factorize a set of partials' weights as a weighted sum of several patterns. Efficient MU rules for estimating the parameters of the function were proposed in [5, 9].

A key feature of [7] is to consider an additional source having a flat density. This idea was inspired by the speech production mechanism. Excitation signals produced by vocal cords are roughly categorized into periodic signals (harmonic comb-shaped spectra) and white noise (flat spectra). These signals are then articulated by the vocal tract whose resonance characteristics can be represented by AR filters. This assumption is widely accepted as reasonable to some extent for music signal modeling. In this study we model this generative process in a Bayesian framework.

2.2 Bayesian Nonparametrics

Another emerging ML technique is Bayesian nonparametrics [18], which is a generalization of the classical Bayesian technique. In the typical Bayesian framework, we put prior distributions on unknown random variables of interests and then, given observed data, estimate a posterior distribution over those variables. However, this framework cannot be used for determining model complexities (the numbers of sources and filters in this study) because these complexities are simply treated as hyperparameters. We thus have to use an expensive grid search for *combinatorial* model selection. Bayesian nonparametrics enables us to treat model complexities as random variables and estimate their optimal values jointly with posterior computation.

Bayesian modeling is being used in music signal analysis, and Bayesian extensions of NMF [19] have been used with great success for audio decomposition (source separation). An especially important breakthrough was recently made by Hoffman *et al.* [12]. They proposed a nonparametric Bayesian extension called the gamma-process NMF (GaP-NMF) that in theory allows an observed spectrogram to contain an infinite number of bases. A limited effective number of bases can be obtained by using an efficient variational inference algorithm. This extension is the basis of a more elaborate model that can consider infinite kinds of temporal variations of each basis [13].

3. PROPOSED MODELS

This section presents new nonparametric Bayesian models called infinite composite autoregressive models (iCARMs). The essential concept of these models is inspired by a composite autoregressive model (CARM) [1] that decomposes a *power* spectrogram into *fixed* numbers of sources and AR filters by using IS divergence as an optimization criterion. We formulate another CARM that decomposes an *amplitude* spectrogram by using KL divergence. To enforce harmonicity we explicitly represent each source—except for a single source that has a flat spectral density (white noise)—as a parametric comb-shaped function as proposed in [7]. Finally, both KL-CARM and IS-CARM are extended to in theory contain *infinite* numbers of sources and filters by using gamma processes as suggested in [12].

3.1 Overall Framework

We first define mathematical symbols as shown in Table 2. Let $\boldsymbol{X}$ be an $M \times N$ complex-valued spectrogram, where M is the number of frequency bins and N is the number of frames. Let I be the number of sources and J be the number of filters, which are assumed to go to infinity. Let the lower-case letters m, n, i, and j indicate the indices.

M	Number of frequency bins
N	Number of frames
I	Number of sources (diverges to infinity)
J	Number of filters (diverges to infinity)
X_{mn}	Amplitude (power) at m-th bin and n-th frame
Y_{mn}	Reconstructed value at m-th bin and n-th frame
θ_i	Global gain of i-th source
ϕ_j	Global gain of j-th filter
W_{im}	Amplitude (power) of i-th source at m-th bin
A_{jm}	Gain of j-th filter at m-th bin
H_{ijn}	Gain of i-th source & j-th filter pair at n-th frame

Table 2. Definition of mathematical symbols

In this paper we aim to factorize a nonnegative representation of $\boldsymbol{X}$ (amplitude or power spectrogram) into three kinds of "factors" $\boldsymbol{W}$, $\boldsymbol{A}$, and $\boldsymbol{H}$ as follows:

$$|X_{mn}| \text{ or } |X_{mn}|^2 \approx \sum_{i,j}^{I,J\to\infty} \theta_i \phi_j W_{im} A_{jm} H_{ijn} \qquad (1)$$

where W_{im}, A_{jm}, and H_{ijn} respectively indicate the amplitude (power) of the i-th source at the m-th bin, the gain of the j-th filter at the m-th bin, and the gain of the i-th source and j-th filter pair at the n-th frame. In addition, two kinds of variables, θ_i and ϕ_j, are introduced to respectively indicate the *global* gain of the i-th source and the *global* gain of j-th filter over all N frames. Even when I and J diverge to infinity, finite numbers of the elements of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are expected to be substantially greater than zero while all other elements are negligibly small. This makes it possible for the "effective" numbers of sources and filters, I^+ and J^+, to be estimated in a data-driven manner.

Our goal is, given the spectrogram $\boldsymbol{X}$, to compute a posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{H}|\boldsymbol{X}; \boldsymbol{W}, \boldsymbol{A})$ over random variables and estimate parameters that represent $\boldsymbol{W}$ and $\boldsymbol{A}$. We will discuss concrete forms of priors $p(\boldsymbol{\theta})$, $p(\boldsymbol{\phi})$, $p(\boldsymbol{H})$, likelihood $p(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{H}; \boldsymbol{W}, \boldsymbol{A})$, and parametric functions of $\boldsymbol{W}$ and $\boldsymbol{A}$ according to KL or IS divergence.

3.2 Mathematical Formulation

We explain the different formulations of iCARMs based on KL and IS divergences.

3.2.1 Observation Likelihoods for $\boldsymbol{X}$

We use KL or IS divergence as an optimization criterion. Let Y_{mn} be $\sum_{ij} Y_{mn}^{ij}$, where $Y_{mn}^{ij} = \theta_i \phi_j W_{im} A_{jm} H_{ijn}$. We aim to optimize Y_{mn} such that the KL or IS divergence between X_{mn} and Y_{mn} is minimized, as shown in Eq.(1). This is known to be equivalent to maximum likelihood estimation of a Poisson or exponential distribution having Y_{mn} as its parameter, given an observation X_{mn} [16]. We here introduce a complex-valued latent variable X_{mn}^{ij} that indicates the contribution of the i-th source and j-th filter pair in X_{mn} such that $X_{mn} = \sum_{ij} X_{mn}^{ij}$ is satisfied.

The KL-iCARM is based on an amplitude-additivity assumption; i.e., $|X_{mn}| = \sum_{ij} |X_{mn}^{ij}|$. This is obviously incorrect but is useful in practice. If $|X_{mn}^{ij}| \sim \text{Poisson}(Y_{mn}^{ij})$, the reproductive property of the Poisson distribution leads to $|X_{mn}| \sim \text{Poisson}(\sum_{ij} Y_{mn}^{ij})$, which means

$$|X_{mn}| \sim \text{Poisson}\,(Y_{mn}) \qquad (2)$$

The IS-iCARM is based on a complex-domain additivity assumption (see Section 3.2.5). If $X_{mn}^{ij} \sim \mathcal{N}_c(0, Y_{mn}^{ij})$, the reproductive property of the complex Gaussian leads to $X_{mn} \sim \mathcal{N}_c(0, \sum_{ij} Y_{mn}^{ij})$. This assumption, however, may be violated when the sources are *not* stationary Gaussian noise (see Section 3.2.4). We nonetheless assume

$$|X_{mn}|^2 \sim \text{Exponential}\,(Y_{mn}) \tag{3}$$

3.2.2 Gamma Process Priors on $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$

We put gamma process (GaP) priors on infinite-dimensional vectors $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. More specifically, we introduce independent gamma priors on elements of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ as follows:

$$\theta_i \sim \text{Gamma}\left(\frac{\alpha}{I}, \alpha\right), \quad \phi_j \sim \text{Gamma}\left(\frac{\gamma}{J}, \gamma\right) \tag{4}$$

As the truncation level I diverges to infinity, the vector $\boldsymbol{\theta}$ approximates an infinite sequence drawn from a GaP with shape parameter α. It is proven that the *effective* number of elements, I^+, such that $\theta_i > \epsilon$ for some number $\epsilon > 0$ is almost surely finite. If we set I to be sufficiently larger than α, we can expect that only a few of the I elements of $\boldsymbol{\theta}$ will be substantially greater than zero. This condensation property enables sparse learning in an infinite space. The same reasoning can be applied to the GaP on $\boldsymbol{\phi}$.

3.2.3 Gamma Chain Priors on $\boldsymbol{H}$

To impose smooth transitions on $\boldsymbol{H}$, we put a gamma chain prior [20] on a temporal sequence of gains of each source-filter pair. More specifically, $\boldsymbol{H}_{ij}$ is modeled as follows:

$$\begin{aligned} H_{ij1} &\sim \text{Gamma}\,(\beta, \beta/d) \\ G_{ijn} &\sim \text{Gamma}\,(\beta, \beta H_{ijn-1}) \\ H_{ijn} &\sim \text{Gamma}\,(\beta, \beta G_{ijn}) \end{aligned} \tag{5}$$

where β is a hyperparameter that controls the strength of the priors (degree of smoothness) and G_{ijn} is an auxiliary variable that imposes a positive correlation between temporally adjacent gains H_{ijn-1} and H_{ijn} ($\mathbb{E}_{\text{prior}}[G_{ijn}] = H_{ijn-1}^{-1}$ and $\mathbb{E}_{\text{prior}}[H_{ijn}] = G_{ijn}^{-1}$). Marginalizing G_{ijn} out, we obtain a positively correlated Markovian transition kernel as $p(H_{ijn}|H_{ijn-1}) = \frac{\Gamma(2\beta)}{2\Gamma(\beta)} \frac{(H_{ijn-1}H_{ijn})^\beta}{(H_{ijn-1}+H_{ijn})^{2\beta}} H_{ijn}^{-1}$.

3.2.4 Comb-shaped Functions for $\boldsymbol{W}$

We represent each harmonic source $\boldsymbol{W}_i$ as a comb-shaped function that is the sum of H Gaussian functions, where H is the number of harmonic partials. Specifically,

$$W_{im} = \sum_{h=1}^{H} \exp\left(-\frac{(m - h\mu_i)^2}{2\sigma^2}\right) \tag{6}$$

where μ_i indicates F0[1] and σ indicates an energy diffusion around the frequencies of partials. Note that only the last source is reserved as white noise, i.e., $W_{Im} = 1$.

3.2.5 All-pole Transfer Functions for $\boldsymbol{A}$

We assume each basis signal $\boldsymbol{x}^{ij} \equiv \{x_t^{ij}\}_{t=1}^{2M}$ in a frame to be represented as a P-order AR process as follows:

$$x_t^{ij} = -\sum_{p=1}^{P} a_p^j x_{t-p}^{ij} + s_t^i \tag{7}$$

where $\boldsymbol{s}^i \equiv \{s_t^i\}_{t=1}^{2M}$ is a signal of the i-th source and $\boldsymbol{a}_j \equiv \{a_0^j, \cdots, a_p^j\}^T$ is a coefficient vector of the j-th filter ($a_0^j = 1$). Let $\boldsymbol{w}^i \equiv \{w_t^i\}_{t=1}^{2M}$ be the autocorrelation of $\boldsymbol{s}^i$ and $\{W_{im}\}_{m=1}^{2M}$ be a complex (amplitude) spectrum density obtained by discrete Fourier transform (DFT) of $\boldsymbol{w}^i$. Let $\{X_m^{ij}\}_{m=1}^{2M}$ be a complex spectrum density obtained by DFT of $\boldsymbol{x}^{ij}$. *If the source signal $\boldsymbol{s}^i$ is a stationary Gaussian noise*, each X_m^{ij} is independently distributed as a complex Gaussian $\mathcal{N}_c(0, \Sigma_m^{ij})$, where $\Sigma_m^{ij} = W_{im}A_{jm}$ and

$$A_{jm} = \frac{1}{\left|\sum_{p=0}^{P} a_p^j e^{-2\pi \frac{m}{2M} pi}\right|^2} = \frac{1}{\boldsymbol{a}_j^T \boldsymbol{U}_m \boldsymbol{a}_j} \tag{8}$$

$\boldsymbol{U}_m$ is a $(P+1)\times(P+1)$ Toeplitz matrix with $[\boldsymbol{U}_m]_{pq} = \cos(2\pi \frac{m}{2M}(p-q))$. This means that $|X_m^{ij}|^2$ is distributed as an exponential distribution having $W_{im}A_{jm}$ as its scale parameter. In other words, maximum likelihood estimation of $\boldsymbol{a}_j$ for $\boldsymbol{x}^{ij}$ is equivalent to minimizing the IS divergence between $\{|X_m^{ij}|^2\}_{m=1}^M$ and $\{W_{im}A_{jm}\}_{m=1}^M$ [2].

In the iCARMs based on KL and IS divergences, the above discussion leads to the following formulations:

$$A_{jm}^{\text{KL}} = \sqrt{\frac{1}{\boldsymbol{a}_j^T \boldsymbol{U}_m \boldsymbol{a}_j}} \quad \text{or} \quad A_{jm}^{\text{IS}} = \frac{1}{\boldsymbol{a}_j^T \boldsymbol{U}_m \boldsymbol{a}_j} \tag{9}$$

A reason for taking the "root" in the KL-iCARM is that we assume an "amplitude" spectrogram as observed data.

3.3 Variational and Multiplicative Optimization

The posterior over random variables $p(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{H}|\boldsymbol{X}; \boldsymbol{W}, \boldsymbol{A})$ and $\boldsymbol{W}$ and $\boldsymbol{A}$ (parameters $\boldsymbol{\mu}$, σ, and $\boldsymbol{a}$) are determined such that the log-evidence $\log p(\boldsymbol{X}; \boldsymbol{W}, \boldsymbol{A})$ 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

$$q(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{H}) = \prod_i q(\theta_i) \prod_j q(\phi_j) \prod_{ijn} q(H_{ijn}) \tag{10}$$

and iteratively updates this form by monotonically increasing a *lower bound*[3] of the log-evidence, $\mathcal{L}$, given by

$$\begin{aligned} &\log p(\boldsymbol{X}; \boldsymbol{W}, \boldsymbol{A}) \geq \mathbb{E}[\log p(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{H}; \boldsymbol{W}, \boldsymbol{A})] \\ &+ \mathbb{E}[\log p(\boldsymbol{\theta})] + \mathbb{E}[\log p(\boldsymbol{\phi})] + \mathbb{E}[\log p(\boldsymbol{H})] \\ &- \mathbb{E}[\log q(\boldsymbol{\theta})] - \mathbb{E}[\log q(\boldsymbol{\phi})] - \mathbb{E}[\log q(\boldsymbol{H})] \equiv \mathcal{L} \end{aligned} \tag{11}$$

The iterative update rules are

$$\begin{aligned} q(\boldsymbol{\theta}) &\propto \exp(\mathbb{E}_{q(\boldsymbol{\phi},\boldsymbol{H})}[\log p(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{H}; \boldsymbol{W}, \boldsymbol{A})]) \\ q(\boldsymbol{\phi}) &\propto \exp(\mathbb{E}_{q(\boldsymbol{\theta},\boldsymbol{H})}[\log p(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{H}; \boldsymbol{W}, \boldsymbol{A})]) \\ q(\boldsymbol{H}) &\propto \exp(\mathbb{E}_{q(\boldsymbol{\theta},\boldsymbol{\phi})}[\log p(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{H}; \boldsymbol{W}, \boldsymbol{A})]) \end{aligned} \tag{12}$$

To optimize $\boldsymbol{W}$ and $\boldsymbol{A}$ ($\boldsymbol{\mu}$, σ, and $\boldsymbol{a}$), we use multiplicative update (MU) rules inspired by [8, 9]. A general rule is obtained from the partial derivative of a "cost" function, e.g., $-\mathcal{L}$. For example, if we can write the derivative as the difference of two positive terms, i.e., $\frac{-\partial \mathcal{L}}{\partial \mu_i} = G_{\mu_i} - F_{\mu_i}$, an update rule for μ_i is given by $\mu_i \leftarrow \mu_i \times \frac{F_{\mu_i}}{G_{\mu_i}}$. Note that μ_i becomes constant if the derivative is zero, and is updated in the opposite direction of the derivative. We omit detailed derivations and only describe update rules below.

[1] When the value of F0 is given by $\tilde{\mu}_i$ [Hz], $\mu_i = \tilde{\mu}_i/(r/2M)$ [bins], where r is a sampling rate and $2M$ is a window size of frequency analysis.

[2] In linear predictive coding (LPC), the source signal $\boldsymbol{s}^i$ is generally limited to white noise ($W_{im} = 1$). This is a conventional assumption.

[3] More specifically, a further lower bound of $\mathcal{L}$ should be computed.

3.3.1 Variational Updates for KL-iCARM

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

$$q(\theta_i) = \text{Gamma}(a_i^\theta, b_i^\theta), \quad q(\phi_j) = \text{Gamma}(a_j^\phi, b_j^\phi)$$
$$q(H_{ijn}) = \text{Gamma}(a_{ijn}^H, b_{ijn}^H) \tag{13}$$

The variational parameters are given by

$$\begin{aligned}
a_i^\theta &= \frac{\alpha}{I} + \textstyle\sum_{mnj} |X_{mn}| \lambda_{mnij} \\
b_i^\theta &= \alpha + \textstyle\sum_{mnj} \mathbb{E}[\phi_j W_{im} A_{jm} H_{ijn}] \\
a_j^\phi &= \frac{\gamma}{J} + \textstyle\sum_{mni} |X_{mn}| \lambda_{mnij} \\
b_j^\phi &= \gamma + \textstyle\sum_{mni} \mathbb{E}[\theta_i W_{im} A_{jm} H_{ijn}] \\
a_{ijn}^H &= 2\beta + \textstyle\sum_m |X_{mn}| \lambda_{mnij} \\
b_{ijn}^H &= \beta \mathbb{E}[G_{ijn} + G_{ijn+1}] + \textstyle\sum_m \mathbb{E}[\theta_i \phi_j W_{im} A_{jm}]
\end{aligned} \tag{14}$$

where λ_{mnij} is an auxiliary variable given by

$$\lambda_{mnij} \propto \exp(\mathbb{E}[\log(\theta_i \phi_j W_{im} A_{jm} H_{ijn})]) \tag{15}$$

3.3.2 Variational Updates for IS-iCARM

As proposed in [12], the variational posterior of each variable is given by a generalized inverse-Gaussian (GIG) distribution (see the Appendix) as follows:

$$q(\theta_i) = \text{GIG}(a_i^\theta, b_i^\theta, c_i^\theta), \quad q(\phi_j) = \text{GIG}(a_j^\phi, b_j^\phi, c_j^\phi)$$
$$q(H_{ijn}) = \text{GIG}(a_{ijn}^H, b_{ijn}^H, c_{ijn}^H) \tag{16}$$

The variational parameters are given by

$$\begin{aligned}
a_i^\theta &= \frac{\alpha}{I}, \quad b_i^\theta = \alpha + \textstyle\sum_{mnj} \frac{\mathbb{E}[\phi_j W_{im} A_{jm} H_{ijn}]}{\xi_{mn}} \\
c_i^\theta &= \textstyle\sum_{mnj} |X_{mn}|^2 \eta_{mnij}^2 \mathbb{E}\big[\frac{1}{\phi_j W_{im} A_{jm} H_{ijn}}\big] \\
a_j^\phi &= \frac{\gamma}{J}, \quad b_j^\phi = \gamma + \textstyle\sum_{mni} \frac{\mathbb{E}[\theta_i W_{im} A_{jm} H_{ijn}]}{\xi_{mn}} \\
c_j^\phi &= \textstyle\sum_{mni} |X_{mn}|^2 \eta_{mnij}^2 \mathbb{E}\big[\frac{1}{\theta_i W_{im} A_{jm} H_{ijn}}\big] \\
a_{ijn}^H &= 2\beta, \; c_{ijn}^H = \textstyle\sum_m |X_{mn}|^2 \eta_{mnij}^2 \mathbb{E}\big[\frac{1}{\theta_i \phi_j W_{im} A_{jm}}\big] \\
b_{ijn}^H &= \beta \mathbb{E}[G_{ijn} + G_{ijn+1}] + \textstyle\sum_m \frac{\mathbb{E}[\theta_i \phi_j W_{im} A_{jm}]}{\xi_{mn}}
\end{aligned} \tag{17}$$

where η_{mnij} and ξ_{mn} are auxiliary variables given by

$$\begin{aligned}
\eta_{mnij} &\propto \mathbb{E}[\tfrac{1}{\theta_i \phi_j W_{im} A_{jm} H_{ijn}}]^{-1} \text{ s.t.} \textstyle\sum_{ij} \eta_{mnij} = 1 \\
\xi_{mn} &= \textstyle\sum_{ij} \mathbb{E}[\theta_i \phi_j W_{im} A_{jm} H_{ijn}]
\end{aligned} \tag{18}$$

3.3.3 Multiplicative Updates for KL- and IS-iCARMs

The MU rules for $\boldsymbol{\mu}$, σ, and $\boldsymbol{a}$ are given by $\mu_i \leftarrow G_{\mu_i}^{-1} F_{\mu_i} \mu_i$, $\sigma^2 \leftarrow G_{\sigma^2}^{-1} F_{\sigma^2} \sigma^2$, and $\boldsymbol{a}_j \leftarrow \boldsymbol{G}_{\boldsymbol{a}_j}^{-1} \boldsymbol{F}_{\boldsymbol{a}_j} \boldsymbol{a}_j$, where

$$\begin{aligned}
F_{\mu_i} &= \textstyle\sum_{mnjh} h(m V_{mnij}^F + h \mu_i V_{mnij}^G) \exp\big(-\frac{(m - h\mu_i)^2}{2\sigma^2}\big) \\
G_{\mu_i} &= \textstyle\sum_{mnjh} h(m V_{mnij}^G + h \mu_i V_{mnij}^F) \exp\big(-\frac{(m - h\mu_i)^2}{2\sigma^2}\big) \\
F_{\sigma^2} &= \textstyle\sum_{mnijh} V_{mnij}^F (m - h\mu_i)^2 \exp\big(-\frac{(m - h\mu_i)^2}{2\sigma^2}\big) \\
G_{\sigma^2} &= \textstyle\sum_{mnijh} V_{mnij}^G (m - h\mu_i)^2 \exp\big(-\frac{(m - h\mu_i)^2}{2\sigma^2}\big) \\
\boldsymbol{F}_{\boldsymbol{a}_j}^{\text{KL}} &= \textstyle\sum_{mni} \theta_i \phi_j W_{im} H_{ijn} A_{jm}^3 \boldsymbol{U}_m \\
\boldsymbol{G}_{\boldsymbol{a}_j}^{\text{KL}} &= \textstyle\sum_{mni} |X_{mn}| \lambda_{mnij} A_{jm}^2 \boldsymbol{U}_m \\
\boldsymbol{F}_{\boldsymbol{a}_j}^{\text{IS}} &= \textstyle\sum_{mni} \frac{\mathbb{E}[\theta_i \phi_j W_{im} H_{ijn}]}{\xi_{mn}} A_{jm}^2 \boldsymbol{U}_m \\
\boldsymbol{G}_{\boldsymbol{a}_j}^{\text{IS}} &= \textstyle\sum_{mni} |X_{mn}|^2 \eta_{mnij}^2 \mathbb{E}\big[\frac{1}{\theta_i \phi_j W_{im} H_{ijn}}\big] \boldsymbol{U}_m
\end{aligned} \tag{19}$$

V_{mnij}^F and V_{mnij}^G are given by $V_{mnij}^F = \mathbb{E}[\theta_i \phi_j A_{jm} H_{ijn}]$ and $V_{mnij}^G = |X_{mn}| \lambda_{mnij} W_{im}^{-1}$ in the KL-iCARM. On the other hand, $V_{mnij}^F = \frac{\mathbb{E}[\theta_i \phi_j A_{jm} H_{ijn}]}{\xi_{mn}}$ and $V_{mnij}^G = |X_{mn}|^2 \eta_{mnij}^2 \mathbb{E}\big[\frac{1}{\theta_i \phi_j W_{im}^2 A_{jm} H_{ijn}}\big]$ in the IS-iCARM.

4. EVALUATION

We report comparative experiments that were conducted to evaluate the performance of the iCARMs based on KL and IS divergences as multipitch analyzers.

4.1 Experimental Conditions

We used thirty pieces of "ENSTDkCl" subset included in the MAPS piano database [21]. We truncated each piece to 30 s as in [5,11] and converted the original CD-quality signals into monaural signals sampled at 16 [kHz]. The spectrograms were obtained with short-time Fourier transform (STFT) with a window size of 2048 samples and a shifting interval of 10 [ms], i.e., $M = 1024$ and $N = 3000$. The amplitude or power spectrogram of each piece was scaled such that $\frac{1}{MN} \sum_{mn} |X_{mn}| = 1$ or $\max_{mn} |X_{mn}|^2 = 1$. The hyperparameters were specified as $I = 88{+}1$, $J = 10$, $\alpha{=}1$, $\beta = \gamma = 0.1$, $H = 20$, $P = 4$, and $d{=}\mathbb{E}_{\text{emp}}[|X_{mn}|]$ or $\mathbb{E}_{\text{emp}}[|X_{mn}|^2]$. Although $J = 10$ was too small to accurately approximate the GaP, it was sufficiently large in our experiments because the audio signals contain only piano sounds. We initialized $\{\mu_i\}_{i=1}^{88}$ as the frequencies corresponding to the 88 keys of the standard piano. The other parameters were initialized randomly.

Multiple F0s were detected at each frame in a thresholding process. If the gain of the i-th source, $\sum_j \theta_i \phi_j H_{ijn}$, was larger than the threshold, we judged that the n-th frame includes an F0 indicated by μ_i. The threshold was globally determined such that the frame-level precision and recall rates were balanced to yield the best average F-measure.

4.2 Experimental Results

We first tested our models on toy data obtained by extracting the first 4.9 s (490 frames) of the piece "alb_se2," which contains five different F0s and a polyphony level that increases one by one up to five (D4, +C#4, +C4, +A3, +F#3). As shown in Fig. 2, the KL-iCARM could successfully discover the correct number of sources (five harmonic sources + one white-noise source) in a data-driven manner. In addition, we could separate $\boldsymbol{X}$ into harmonic and noise components by computing $\mathbb{E}[Y_{mn}^i] = \sum_j \mathbb{E}[\theta_i \phi_j W_{im} A_{jm} H_{ijn}]$, which represents the component of the i-th source at the m-th bin and n-th frame.

As shown in Fig. 3, on the other hand, the IS-iCARM overestimated the numbers of sources and filters and made many octave errors (half-F0 errors). One reason is that IS divergence permits a reconstructed power to exceed an observed power with a smaller penalty. It is therefore difficult to reduce false alarms of harmonic partials.

We then used the 30 pieces for evaluation. The KL- and IS-iCARMs achieved the frame-level F-measures of 48.4% and 35.1% respectively. Although these preliminary results were not really impressive compared with the state-of-the-art results [5,11], we consider our framework to be promising because of its elegant nature of sparse learning over an infinite space. A main limitation of the KL-iCARM is that we still need to resort a thresholding process for temporal gains although limited numbers of sources and filters can be obtained by using GaPs. One solution would be to introduce *binary* latent variables that indicate note existences.

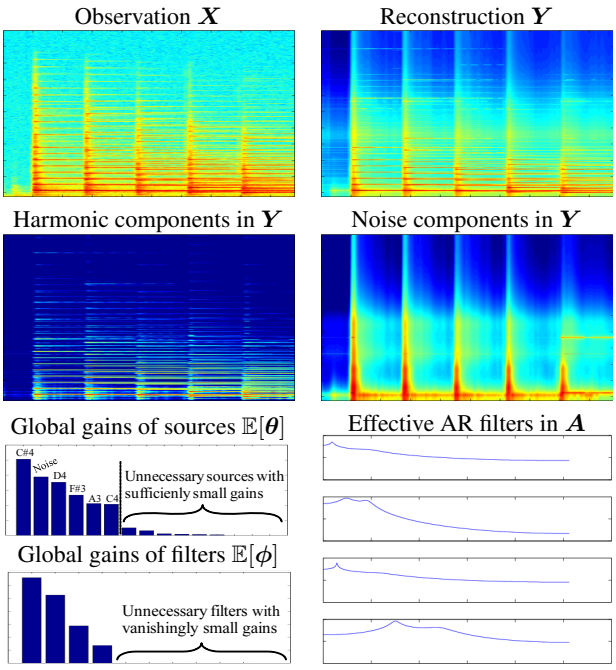

Figure 2. Decomposition results obtained by KL-iCARM

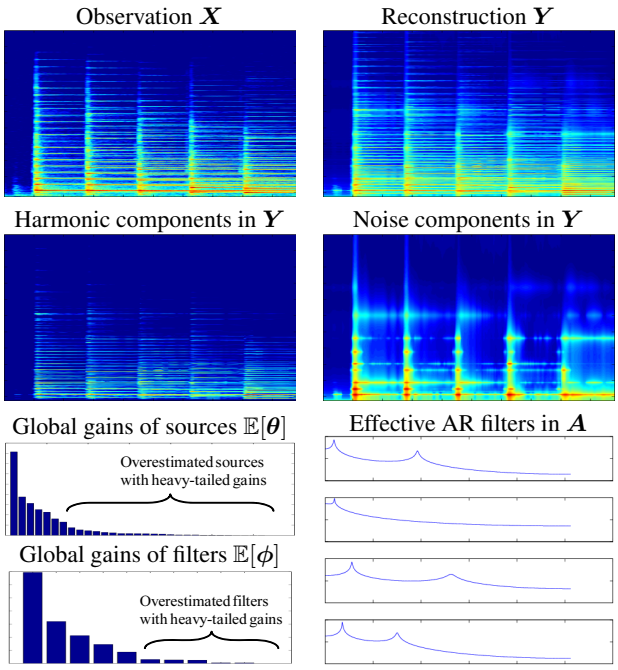

Figure 3. Decomposition results obtained by IS-iCARM

5. CONCLUSION

We presented nonparametric Bayesian models called infinite composite autoregressive models (iCARMs) that decompose an observed spectrogram into three kinds of factors, i.e., sources, filters, and time-varying gains of source-filter pairs. The experimental results showed that appropriate numbers of sources and filters can be discovered in a data-driven manner by using gamma processes for sparse learning. To improve the accuracy of multipitch analysis, we are considering the use of log-frequency spectrograms obtained by constant-Q or wavelet transform. We also plan to use these models for "timbre-based" source separation by distinguishing different resonance characteristics of instrument and vocal sounds by AR filters.

Acknowledgment: This study was supported by JSPS KAKENHI 23700184 and JST OngaCREST project.

6. REFERENCES

[1] H. Kameoka and K. Kashino. Composite autoregressive system for sparse source-filter representation of speech. *ICASSP*, pp. 2477–2480, 2009.

[2] R. Badeau, V. Emiya, and B. David. Expectation-maximization algorithm for multi-pitch estimation and separation of overlapping harmonic spectra. *ICASSP*, pp. 3073–3076, 2009.

[3] J.-L. Durrieu, G. Richard, B. David, and C. Févotte. Source/Filter model for unsupervised main melody extraction from polyphonic audio signals. *IEEE Trans. on ASLP*, 18(3):564–575, 2010.

[4] T. Virtanen and A. Klapuri. Analysis of polyphonic audio using source-filter model and non-negative matrix factorization. *NIPS Ws. on Adv. in Mod. for Acoust. Proc.*, 2009.

[5] 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 J. of Sel. Top. in Sig. Proc.*, 5(6):1144–1158, 2011.

[6] T. Heittola, A. Klapuri, and T. Virtanen. Musical instrument recognition in polyphonic audio using source-filter model for sound separation. *ISMIR*, pp. 327–332, 2009.

[7] N. Yasuraoka and H. G. Okuno. Musical audio signal modeling for joint estimation of harmonic, inharmonic, and timbral structure and its application to source sepatation. *SIG Technical Reports*, volume 2012-MUS-94, pp. 1–8, 2012.

[8] R. Hennequin, R. Badeau, and B. David. NMF with time-frequency activations to model nonstationary audio events. *IEEE Trans. on ASLP*, 19(4):744–753, 2011.

[9] R. Hennequin, R. Badeau, and B. David. Time-dependent parametric and harmonic templates in non-negative matrix factorization. *DAFx*, pp. 1–8, 2010.

[10] E. Vincent, N. Bertin, and R. Badeau. Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE Trans. on ASLP*, 18(3):528–537, 2010.

[11] N. Bertin, R. Badeau, and E. Vincent. Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription. *IEEE Trans. on ASLP*, 18(3):538–549, 2010.

[12] M. Hoffman, D. Blei, and P. Cook. Bayesian nonparametric matrix factorization for recorded music. *ICML*, 2010.

[13] M. Nakano *et al.* Bayesian nonparametric spectrogram modeling based on infinite factorial infinite hidden Markov model. *WASPAA*, pp. 325–328, 2011.

[14] B. Niedermayer. Improving accuracy of polyphonic music-to-score alignment. *ISMIR*, pp. 585–590, 2009.

[15] D. Lee and H. Seung. Algorithms for non-negative matrix factorization. *NIPS*, pp. 556–562, 2000.

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

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

[18] P. Orbanz and Y. W. Teh. Bayesian nonparametric models. *Encyclopedia of Machine Learning*. Springer, 2010.

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

[20] A. T. Cemgil and O. Dikmen. Conjugate gamma Markov random fields for modelling nonstationary sources. *ICA*, 2007.

[21] V. Emiya, R. Badeau, and B. David. Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *IEEE Trans. on ASLP*, 18(6):1643–1654, 2010.

A. PROBABILITY DISTRIBUTIONS

$$\text{Gamma}(x|a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} \quad \text{Exponential}(x|\lambda) = \frac{1}{\lambda} e^{-\frac{x}{\lambda}}$$

$$\text{Poisson}(x|\lambda) = \frac{\lambda^x}{x!} e^{-\lambda} \quad \text{GIG}(x|a,b,c) = \frac{(b/c)^{\frac{a}{2}} x^{a-1}}{2\mathcal{K}_a(2\sqrt{bc})} e^{-(bx+\frac{c}{x})}$$