INFINITE KERNEL LINEAR PREDICTION FOR JOINT ESTIMATION OF SPECTRAL ENVELOPE AND FUNDAMENTAL FREQUENCY

Kazuyoshi Yoshii *Masataka Goto*

National Institute of Advanced Industrial Science and Technology (AIST)
{k.yoshii, m.goto}@aist.go.jp

ABSTRACT

This paper presents a new probabilistic formulation of linear prediction (LP) for jointly estimating the spectral envelope and fundamental frequency (F0) of a speech signal. A main problem of classical LP is that the peaks of the estimated envelope are highly biased toward the harmonic partials of a speech spectrum. To solve this problem, we propose a nonparametric Bayesian model called infinite kernel linear prediction (IKLP) based on a Gaussian process with multiple kernel learning. Our model can represent the periodicity of a speech signal by using a weighted sum of infinitely many periodic kernels that correspond to different F0s. We put a gamma process prior on the positive weights of those kernels and perform sparse learning to determine a predominant kernel indicating the F0 at the same time of spectral envelope estimation. The experimental results showed that our model can estimate spectral envelopes and F0s of speech and singing signals while identifying pitched segments.

Index Terms— Linear prediction, source-filter model, Bayesian nonparametrics, kernel methods, Gaussian and gamma processes.

1. INTRODUCTION

Spectral envelope estimation forms the basis for speech and singing analysis. The speech production mechanism is considered to be well explained by the source-filter theory, which assumes that an excitation signal generated by the vocal cords is modified acoustically by the vocal tract, *i.e.*, the source signal is convoluted by an impulse response of the filter. In the frequency domain, the fine structure and spectral envelope of a speech spectrum reflect the frequency characteristics of the source and filter, respectively (Fig. 1).

Linear prediction (LP) is a popular parametric approach to spectral envelope estimation [1]. In general, we assume speech signals to be autoregressive (AR), *i.e.*, the filter response can be represented by an all-pole transfer function. This assumption is widely accepted as reasonable because most phonemes have no anti-resonance and the human auditory system is sensitive to spectral peaks (formants) corresponding to poles. If the source signal is a white Gaussian noise, we can correctly estimate the filter coefficients by maximum likelihood estimation in a probabilistic framework [1]. When we analyze a pitched signal, however (*i.e.*, when the source signal is *periodic*), the estimated filter (spectral envelope) unnecessarily has very sharp peaks at the harmonic partials of the spectrum.

A lot of effort has been devoted to solving this problem. For example, El-Jaroudi and Makhoul [2] proposed a basic approach that fits the frequency response of an all-pole filter to a discrete set of harmonic partials. This discrete all-pole (DAP) modeling was extended by Badeau and David [3] for AR and moving-average (ARMA) modeling. Alternatively, Oudot *et al.* [4] took a regularization approach

This study was supported in part by the JSPS KAKENHI Grant Number 23700184 and the JST OngaCREST project.

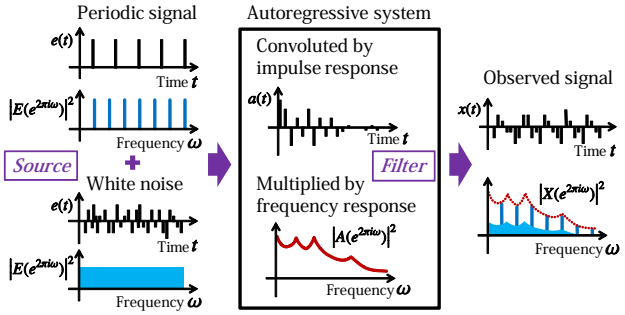

Fig. 1. A source-filter model of the speech production mechanism.

that imposes smoothness constraints on all-pole filters. Villavicencio *et al.* [5] applied iterative cepstral smoothing to all-pole filters. In a wider context of spectral envelope estimation not limited to LP, Kawahara *et al.* [6] proposed a speech analysis-and-synthesis system called STRAIGHT that can decompose a speech spectrum into the harmonic and non-harmonic spectra and the spectral envelope. Nakano and Goto [7] aimed to avoid the bias of fundamental frequencies (F0s) by averaging the spectral envelopes of neighboring frames. Although these methods are considered to yield good estimates, they all require that the F0s be known in advance.

A probabilistic approach provides a solid mathematical foundation for modeling F0s (sources) and envelopes (filters). For example, Sasou and Tanaka [8] proposed the periodicity of the source signal could be represented by using an AR hidden Markov model (HMM) with circularly-connected states. Toda and Tokuda [9] used a trajectory HMM for capturing the temporal dynamics of harmonic components derived from the source signal, although F0 information is required. Kameoka *et al.* [10, 11] pioneered probabilistic models for joint F0 and spectral-envelope estimation. In [10] they formulated multiple kernel LP (MKLP) based on a Gaussian process (GP) that is specified by a fixed number of periodic kernels corresponding to different F0s. To determine a predominant kernel, the maximum-a-posteriori (MAP) estimate of kernel weights is computed (F0 estimation) at the same time the all-pole filter is estimated. Our study is situated as the state of the art in this thread of research.

In this paper we propose a more sophisticated model called infinite kernel LP (IKLP) that combines the strengths of Bayesian nonparametrics and kernel methods in a principled manner. We take the limit of the MKLP model [10] as the number of kernels diverges to infinity, and put a gamma process (GaP) prior on the weights of those kernels. As a result of variational Bayesian (VB) inference, we can obtain a sparse estimate of infinitely many weights. We believe that our study could contribute not only to the field of signal processing by revealing the underlying probabilistic assumptions of LP, but also to the field of machine learning by providing a new efficient and convergence-guaranteed algorithm as a general solution to the problem of multiple kernel learning (MKL) [12].

2. LINEAR PREDICTION

This section introduces relevant probabilistic models for estimating spectral envelopes (coefficients of AR filters) of speech and singing signals. First we revisit a classical AR model based on a strong assumption that excitation signals are white Gaussian noise. We then explain kernelized models that can perform robust estimation even when excitation signals have a periodic nature.

2.1. Probabilistic Formulation

We review a standard formulation that assumes a target audio signal to locally follow an AR process. Let $\boldsymbol{x} = (x_1, x_2, \cdots, x_M)^T$ be M consecutive samples contained in a short segment (shifting window). If the local signal $\boldsymbol{x}$ follows a P-order AR process, we can write

$$x_m = \sum_{p=1}^{P} a_p x_{m-p} + \epsilon_m, \tag{1}$$

where $\boldsymbol{a} = (a_1, \cdots, a_P)^T$ is a set of P coefficients of the AR filter called *predictor coefficients*, and $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_M)^T$ is a noise term. In the source-filter modeling of speech signals, $\boldsymbol{\epsilon}$ represents an excitation signal generated by the vocal cords and $\boldsymbol{a}$ represents the resonance characteristics of the vocal tract. This AR model can be regarded as a linear system that takes ϵ_m as input and then outputs x_m according to an all-pole transfer function given by $A(z) = 1/(1 - a_1 z^{-1} \cdots - a_P z^{-P})$, *i.e.*, $X(z) = E(z)A(z)$, where $X(z)$ and $E(z)$ are z transforms of $\boldsymbol{x}$ and $\boldsymbol{\epsilon}$. The frequency response of the all-pole filter (spectral envelope) is therefore given by $|A(e^{2\pi i m/M})|^2$, where m here is the index of a frequency bin (Fig. 1).

Given the observed audio signal $\boldsymbol{x}$, our objective is to estimate the coefficients $\boldsymbol{a}$ in a probabilistic framework. To do this, we need to specify the characteristics of the excitation signal $\boldsymbol{\epsilon}$. A standard assumption is that $\boldsymbol{\epsilon}$ is a white Gaussian noise given by

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \nu \boldsymbol{I}), \tag{2}$$

where ν is a noise variance and $\boldsymbol{I}$ an identity matrix. This means that M elements of $\boldsymbol{\epsilon}$ are *independent and identically distributed* according to a Gaussian $\mathcal{N}(0, \nu)$. We let $\boldsymbol{\Psi}$ be an M-by-M approximate circulant matrix and $\boldsymbol{X}$ be an M-by-P matrix as follows:

$$\boldsymbol{\Psi} = \begin{bmatrix} 1 & & & & \\ -a_1 & \ddots & & \boldsymbol{0} & \\ \vdots & \ddots & \ddots & & \\ -a_P & & & & \\ & \ddots & & \ddots & \ddots \\ \boldsymbol{0} & & -a_P \cdots -a_1 & & 1 \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 0 & \cdots & 0 \\ & \ddots & \vdots \\ x_1 & & \\ \vdots & \ddots & 0 \\ & & \\ \vdots & & x_1 \\ & & \vdots \\ x_{M-1} & \cdots & x_{M-P} \end{bmatrix}. \tag{3}$$

We can compactly rewrite Eq. (1) as $\boldsymbol{\epsilon} = \boldsymbol{\Psi}\boldsymbol{x}$, which means

$$\boldsymbol{x} = \boldsymbol{\Psi}^{-1}\boldsymbol{\epsilon}. \tag{4}$$

Using Eqs. (2) and (4) gives a likelihood function of $\boldsymbol{x}$ as follows:

$$\boldsymbol{x} \sim \mathcal{N}(0, \nu \boldsymbol{\Psi}^{-1}\boldsymbol{\Psi}^{-T}). \tag{5}$$

This is a standard probabilistic model of LP. In maximum-likelihood (ML) estimation, the optimal values of $\boldsymbol{a}$ are obtained by solving a normal equation given by $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{a} = \boldsymbol{X}^T\boldsymbol{x}$.

This model works well when the excitation signal $\boldsymbol{\epsilon}$ is well approximated by an isotropic Gaussian defined as Eq. (2). Speech and singing signals, however, exhibit a clear periodicity that is derived from the periodic vibration of the vocal cords (source).

2.2. Kernelization based on a Gaussian Process

Kameoka *et al.* [10] generalized the probabilistic model specified as Eq. (5) in terms of Gaussian process (GP) regression. In Section 2.1, we assumed that $\boldsymbol{\epsilon}$ is a white Gaussian noise given by Eq. (2). Here we instead discuss a linear regression problem that aims to model an excitation signal (continuous function) $\epsilon(t)$ over time t. Let $\boldsymbol{t} = \{t_m\}_{m=1}^M$ be a set of times at which $\{\epsilon_m\}_{m=1}^M$ are sampled from $\epsilon(t)$. Our goal is to approximate $\epsilon(t)$ by the weighted sum of J basis functions $\{\phi_j(t)\}_{j=1}^J$ as follows:

$$\epsilon(t) = \sum_{j=1}^{J} w_j \phi_j(t) + \eta(t) = \boldsymbol{\phi}(t)^T \boldsymbol{w} + \eta(t), \tag{6}$$

where $\eta(t)$ is an error function and $\boldsymbol{\phi}(t) = (\phi_1(t), \cdots, \phi_J(t))^T$. Let $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_M)^T$ be a set of values sampled from $\eta(t)$ at $\boldsymbol{t}$ and let $\boldsymbol{\Phi} = (\boldsymbol{\phi}(t_1), \cdots, \boldsymbol{\phi}(t_M))^T$ be an M-by-J design matrix. Then, we can write the "marginal" of Eq. (6) on $\boldsymbol{t}$ as follows:

$$\boldsymbol{\epsilon} = \boldsymbol{\Phi}\boldsymbol{w} + \boldsymbol{\eta}. \tag{7}$$

We assume that both $\boldsymbol{\omega}$ and $\boldsymbol{\eta}$ are Gaussian distributed as follows:

$$\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \nu_w \boldsymbol{I}), \quad \boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}, \nu_e \boldsymbol{I}), \tag{8}$$

where ν_w and ν_e are scales of Gaussian variance. Then, we get

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \nu_w \boldsymbol{\Phi}\boldsymbol{\Phi}^T + \nu_e \boldsymbol{I}). \tag{9}$$

This is indeed a GP [13] because any marginal $\boldsymbol{\epsilon}$ sampled from $\epsilon(t)$ is Gaussian distributed with a positive semidefinite covariance matrix specified by a term $\boldsymbol{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}^T$ called a kernel matrix. Each element of $\boldsymbol{K}$ is given by the inner product of basis functions as follows:

$$K_{m,m'} = \boldsymbol{\phi}(t_m)^T \boldsymbol{\phi}(t_{m'}). \tag{10}$$

Instead, *any* positive semidefinite matrix can be used as a kernel matrix. This enables the *kernel trick*; we can directly calculate $\boldsymbol{K}$ without explicitly specifying basis functions as $K_{m,m'} = k(t_m, t_{m'})$, where $k(t_m, t_{m'})$ is a kernel function. From this kernelization and Eq. (4), we derive a likelihood function of $\boldsymbol{x}$ as follows:

$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}^{-1}(\nu_w \boldsymbol{K} + \nu_e \boldsymbol{I})\boldsymbol{\Psi}^{-T}). \tag{11}$$

This is a GP regression model [10], and it includes Eq. (5) as a special case. In fact, when $\boldsymbol{\Phi}\boldsymbol{\Phi}^T = \boldsymbol{I}$, *i.e.*, when J basis functions are independent of each other like a series of Dirac delta functions, we can recover Eq. (5) by assuming $\nu = \nu_w + \nu_e$.

2.3. Multiple Kernel Learning

We discuss how to design a kernel matrix $\boldsymbol{K}$ that reflects the characteristics of the excitation signal $\boldsymbol{\epsilon}$. If the observation $\boldsymbol{x}$ is a pitched signal having the F0, a natural choice is using a periodic kernel. For example, $k(t, t') = \exp(-2\sin^2(\pi \frac{t-t'}{T})/l^2)$ is a well-known kernel having the period T. This means that all basis functions are implicitly assumed to have the period T (the F0 is given by $1/T$). Alternatively, Kameoka *et al.* [10] designed a basis function having H harmonic components with equal power as follows:

$$\phi_j(t) = \sum_{h=1}^{H} \sin\left(2\pi h \frac{t - c_j}{T}\right) \quad (c_j \text{ is a phase}) \tag{12}$$

and used a kernel matrix $\boldsymbol{K}$ calculated using Eq. (10). A problem is that the true period T of the excitation signal $\boldsymbol{\epsilon}$ is unknown. Given the observation $\boldsymbol{x}$, one must therefore estimate $\boldsymbol{K}$ itself.

Multiple kernel learning (MKL) [12] is a powerful solution to this problem. More specifically, the kernel matrix $\boldsymbol{K}$ is defined as the weighted sum of I kernel matrices as follows:

$$\boldsymbol{K} = \sum_{i=1}^{I} \theta_i \boldsymbol{K}_i, \tag{13}$$

where $\boldsymbol{\theta} = \{\theta_1, \cdots, \theta_I\}$ is a set of kernel weights to be estimated from the observed signal $\boldsymbol{x}$ and $\boldsymbol{K}_i$ is a periodic kernel having the period T_i. The value of θ_i indicates a degree of the predominance of period T_i in $\boldsymbol{x}$ (the F0 is given by $1/T_i$). Plugging Eq. (13) into Eq. (11) gives a likelihood function of $\boldsymbol{x}$ as follows:

$$\boldsymbol{x} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Psi}^{-1}\left(\nu_w \textstyle\sum_{i=1}^{I} \theta_i \boldsymbol{K}_i + \nu_e \boldsymbol{I}\right)\boldsymbol{\Psi}^{-T}\right). \tag{14}$$

The parameters $\boldsymbol{a}$ and $\boldsymbol{\theta}$ can be estimated by using the expectation-maximization (EM) algorithm. In [10], several hundreds of kernels having different periods $\boldsymbol{T} = \{T_1, \cdots, T_I\}$ are prepared, and MAP estimation is then performed by putting a generalized Gaussian distribution on each θ_i as a prior distribution. However, the values of $\boldsymbol{\theta}$ do not become truly sparse in MAP estimation.

3. INFINITE KERNEL LINEAR PREDICTION

This section proposes a nonparametric Bayesian model, called infinite kernel linear prediction (IKLP), for jointly estimating the spectral envelope and F0 of a speech signal in a principled manner. First we take the limit of Eq. (14) as the number of kernels goes to infinity, *i.e.*, $I \to \infty$, in the framework of Bayesian nonparametrics. Then we put appropriate priors on unknown variables for a full Bayesian treatment. To approximate a posterior distribution of those variables, we derive an efficient and convergence-guaranteed algorithm.

3.1. Nonparametric Bayesian Formulation

A likelihood function of the observed signal $\boldsymbol{x}$ is given by

$$\boldsymbol{x} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Psi}^{-1}\left(\nu_w \textstyle\sum_{i=1}^{I\to\infty} \theta_i \boldsymbol{K}_i + \nu_e \boldsymbol{I}\right)\boldsymbol{\Psi}^{-T}\right). \quad (15)$$

We here put a gamma process (GaP) prior on the infinite-dimensional vector $\boldsymbol{\theta}$. More specifically, we introduce independent gamma priors on elements of $\boldsymbol{\theta}$ as follows:

$$\theta_i \sim \text{Gamma}(\alpha/I, \alpha). \quad (16)$$

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 I is 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 property gives a theoretical basis to sparse learning in an infinite space.

To complete the Bayesian formulation, we put gamma priors on the positive weights ν_w and ν_e as follows:

$$\nu_w \sim \text{Gamma}(a_w, b_w), \quad \nu_e \sim \text{Gamma}(a_e, b_e), \quad (17)$$

where a_* and b_* are hyperparameters indicating the shape and rate parameters of the gamma distribution. We also put a Gaussian prior on $\boldsymbol{a}$ thusly: $\boldsymbol{a} \sim \mathcal{N}(\boldsymbol{0}, \lambda \boldsymbol{I})$, where λ is a hyperparameter.

The F0 is given by identifying a predominant kernel $\boldsymbol{K}_i$ with the highest value of $\mathbb{E}[\theta_i]$ (the F0 is $1/T_i$). The spectral envelope is specified by the filter coefficients $\mathbb{E}[\boldsymbol{a}]$. In addition, our model can distinguish whether the target signal $\boldsymbol{x}$ is pitched or unpitched according to a degree of periodicity $\mathbb{E}[\nu_w]/\mathbb{E}[\nu_w + \nu_e]$ as in MKLP [10].

3.2. Variational Bayesian Inference

Given an observed signal $\boldsymbol{x}$, our goal is to compute a posterior distribution over random variables $p(\boldsymbol{\theta}, \boldsymbol{a}, \nu_w, \nu_e|\boldsymbol{x})$ by using the Bayes rule $p(\boldsymbol{\theta}, \boldsymbol{a}, \nu_w, \nu_e|\boldsymbol{x}) = p(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{a}, \nu_w, \nu_e)/p(\boldsymbol{x})$. Since analytical calculation of $p(\boldsymbol{x})$ is infeasible, we use a variational Bayesian (VB) method for approximating $p(\boldsymbol{\theta}, \boldsymbol{a}, \nu_w, \nu_e|\boldsymbol{x})$ by a variational distribution that can be factorized into four posteriors as follows:

$$q(\boldsymbol{\theta}, \boldsymbol{a}, \nu_w, \nu_e) = q(\boldsymbol{a})q(\nu_w)q(\nu_e)\textstyle\prod_i q(\theta_i). \quad (18)$$

Each posterior is then iteratively updated such that an *evidence lower bound (ELBO)* $\mathcal{L}$ is monotonically increased, where $\mathcal{L}$ is given by

$$\begin{aligned}\log p(\boldsymbol{x}) &\geq \mathbb{E}[\log p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{a}, \nu_w, \nu_e)] \quad (19)\\ &+ \mathbb{E}[\log p(\boldsymbol{\theta})] + \mathbb{E}[\log p(\boldsymbol{a})] + \mathbb{E}[\log p(\nu_w)] + \mathbb{E}[\log p(\nu_e)]\\ &- \mathbb{E}[\log q(\boldsymbol{\theta})] - \mathbb{E}[\log q(\boldsymbol{a})] - \mathbb{E}[\log q(\nu_w)] - \mathbb{E}[\log q(\nu_e)] \equiv \mathcal{L}.\end{aligned}$$

However, the first term is is still intractable. We therefore take a further lower bound $\mathcal{L}'$ such that $\mathcal{L} \geq \mathcal{L}'$. Note that $\mathcal{L}$ can be indirectly maximized by maximizing $\mathcal{L}'$. The updating formulas are

$$\begin{aligned}q(\boldsymbol{\theta}) &\propto p(\boldsymbol{\theta}) \exp(\mathbb{E}_{q(\boldsymbol{a}, \nu_w, \nu_e)}[\log q(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{a}, \nu_w, \nu_e)]),\\ q(\nu_w) &\propto p(\nu_w) \exp(\mathbb{E}_{q(\boldsymbol{\theta}, \boldsymbol{a}, \nu_e)}[\log q(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{a}, \nu_w, \nu_e)]),\\ q(\nu_e) &\propto p(\nu_e) \exp(\mathbb{E}_{q(\boldsymbol{\theta}, \boldsymbol{a}, \nu_w)}[\log q(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{a}, \nu_w, \nu_e)]), \quad (20)\end{aligned}$$

where $q(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{a}, \nu_w, \nu_e)$ is a lower bound of $p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{a}, \nu_w, \nu_e)$ (see Eq. (23)). For tractability we assume $q(\boldsymbol{a}) = \delta_{\boldsymbol{a}^*}(\boldsymbol{a})$, where $\delta_{\boldsymbol{a}^*}$ is a Dirac delta function taking infinity at a MAP point estimate $\boldsymbol{a}^*$.

3.2.1. Deriving Matrix-variate Inequalities

To derive the tractable lower bound $\mathcal{L}'$, we need to use some inequalities. We start with the following definitions:

Definition 1 (Positive semidefinite matrix)**.** *We say a symmetric matrix $\boldsymbol{A}$ is positive semidefinite iff $\boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z} \geq 0$ for any vector $\boldsymbol{z}$ or iff $\boldsymbol{A} = \boldsymbol{Z}^T\boldsymbol{Z}$ for some real matrix $\boldsymbol{Z}$.*

Definition 2 (Convex/concave matrix functions)**.** *We say a matrix-variate function $f(\cdot)$ is convex iff $\lambda f(\boldsymbol{A}) + (1-\lambda)f(\boldsymbol{B}) \geq f(\lambda\boldsymbol{A} + (1-\lambda)\boldsymbol{B})$ for any number $0 \leq \lambda \leq 1$. Conversely, we say $f(\cdot)$ is concave iff $\lambda f(\boldsymbol{A}) + (1-\lambda)f(\boldsymbol{B}) \leq f(\lambda\boldsymbol{A} + (1-\lambda)\boldsymbol{B})$.*

Let $\boldsymbol{V}$ be a positive semidefinite (PSD) matrix and $\boldsymbol{z}$ be any vector. Using these two definitions, we can derive the following lemmas (we omit the proofs because of space limitations):

Lemma 1. *A matrix-variate function $f(\boldsymbol{V}) = \log|\boldsymbol{V}|$ is concave.*

Lemma 2. *A matrix-variate function $g(\boldsymbol{V}) = \boldsymbol{z}^T\boldsymbol{V}^{-1}\boldsymbol{z}$ is convex.*

These lemmas lead to useful inequalities of matrix-variate functions. First, we can apply first-order Taylor expansion to $f(\boldsymbol{V})$ around an arbitrary PSD matrix $\boldsymbol{\Omega}$ as follows:

$$\log|\boldsymbol{V}| \leq \log|\boldsymbol{\Omega}| + \text{tr}(\boldsymbol{\Omega}^{-1}\boldsymbol{V}) - M, \quad (21)$$

where M is the size of $\boldsymbol{V}$. Second, we can apply a matrix inequality proposed by Sawada *et al.* [14] to $g(\boldsymbol{V})$ as follows:

$$\boldsymbol{z}^T\left(\textstyle\sum_{i=1}^I \boldsymbol{V}_i\right)^{-1}\boldsymbol{z} \leq \textstyle\sum_{i=1}^I \boldsymbol{z}^T\boldsymbol{\Upsilon}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{\Upsilon}_i\boldsymbol{z}, \quad (22)$$

where $\{\boldsymbol{V}_i\}_{i=1}^I$ is a set of PSD matrices and $\{\boldsymbol{\Upsilon}_i\}_{i=1}^I$ is a set of auxiliary matrices that sum to the identity matrix.

3.2.2. Deriving Evidence Lower Bound and Updating Formulas

We derive updating formulas by computing the tractable ELBO $\mathcal{L}'$. Let $\boldsymbol{K}$ be $\boldsymbol{K} = \nu_w \sum_i \theta_i\boldsymbol{K}_i + \nu_e\boldsymbol{I}$. Using Eqs. (21) and (22) enables us to compute the lower bound of $\mathbb{E}[\log p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{a}, \nu_w, \nu_e)]$ (the first term of $\mathcal{L}$ given by Eq. (19)) as follows:

$$\begin{aligned}&\mathbb{E}[\log p(\boldsymbol{x}|\cdot)] = -\frac{M}{2}\log(2\pi) - \frac{1}{2}\mathbb{E}[\log|\boldsymbol{K}|] - \frac{1}{2}\mathbb{E}[\boldsymbol{x}^T\boldsymbol{\Psi}^T\boldsymbol{K}^{-1}\boldsymbol{\Psi}\boldsymbol{x}]\\ &\geq -\frac{1}{2}\log|\boldsymbol{\Omega}| - \frac{1}{2}\textstyle\sum_i \mathbb{E}[\nu_w\theta_i]\,\text{tr}(\boldsymbol{\Omega}^{-1}\boldsymbol{K}_i) - \frac{1}{2}\mathbb{E}[\nu_e]\,\text{tr}(\boldsymbol{\Omega}^{-1}) + \text{const.}\\ &- \frac{1}{2}\textstyle\sum_i \mathbb{E}\left[\frac{1}{\nu_w\theta_i}\right]\boldsymbol{x}^T\boldsymbol{\Psi}^T\boldsymbol{\Upsilon}_i^T\boldsymbol{K}_i^{-1}\boldsymbol{\Upsilon}_i\boldsymbol{\Psi}\boldsymbol{x} - \frac{1}{2}\mathbb{E}\left[\frac{1}{\nu_e}\right]\boldsymbol{x}^T\boldsymbol{\Psi}^T\boldsymbol{\Upsilon}_0^T\boldsymbol{\Upsilon}_0\boldsymbol{\Psi}\boldsymbol{x}, \quad (23)\end{aligned}$$

where $\boldsymbol{\Omega}$ is a PSD matrix and $\boldsymbol{\Upsilon} = \{\boldsymbol{\Upsilon}_i\}_{i=0}^{I\to\infty}$ is a set of auxiliary matrices that sum to unity. By setting the partial derivatives equal to zero, we obtain the following optimal values of $\boldsymbol{\Omega}$ and $\boldsymbol{\Upsilon}$:

$$\boldsymbol{\Omega} = \mathbb{E}[\nu_w]\textstyle\sum_i \mathbb{E}[\theta_i]\boldsymbol{K}_i + \mathbb{E}[\nu_e]\boldsymbol{I}, \quad (24)$$

$$\boldsymbol{\Upsilon}_i = \mathbb{E}\left[\frac{1}{\nu_w\theta_i}\right]^{-1}\boldsymbol{K}_i\boldsymbol{S}^{-1}, \quad \boldsymbol{\Upsilon}_0 = \mathbb{E}\left[\frac{1}{\nu_e}\right]^{-1}\boldsymbol{S}^{-1}, \quad (25)$$

where $\boldsymbol{S} = \sum_i \mathbb{E}\left[\frac{1}{\nu_w\theta_i}\right]^{-1}\boldsymbol{K}_i + \mathbb{E}[\frac{1}{\nu_e}]^{-1}\boldsymbol{I}$. Note that Eq. (23) involves the expectations both of the parameters and of their reciprocals, *i.e.*, the sufficient statistics are x and $1/x$. Since the sufficient

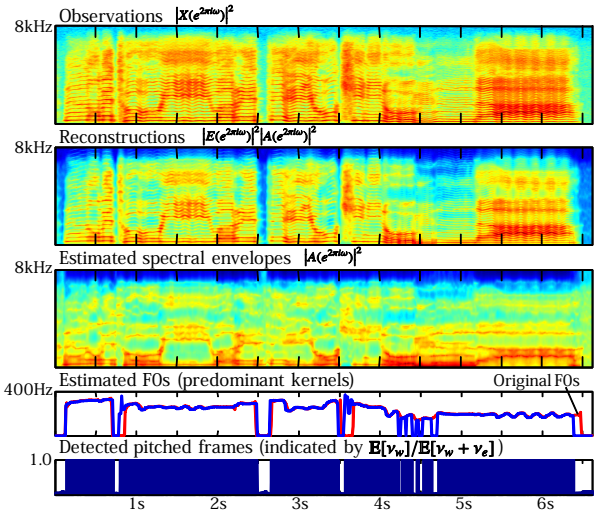

Fig. 2. Estimation results for a male-singing signal.

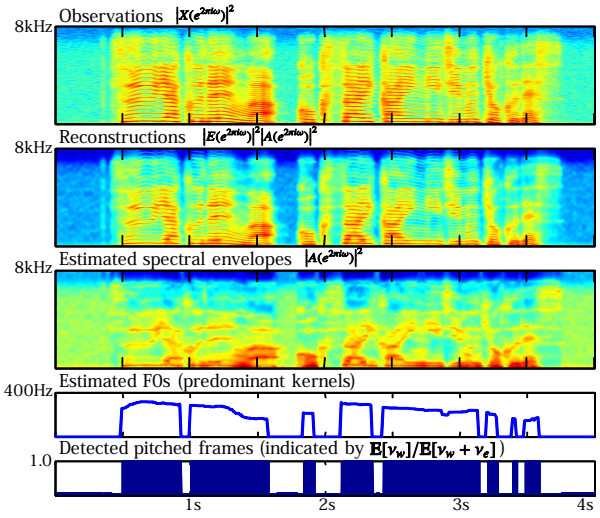

Fig. 3. Estimation results for a female-speech signal.

statistics of the gamma priors are $\log(x)$ and x, the generalized inverse Gaussian (GIG) distribution is a convenient choice as the functional form of posteriors [15]. We thus assume

$$q(\theta_i) = \mathrm{GIG}(\theta_i|\gamma_i,\rho_i,\tau_i),$$
$$q(\nu_w) = \mathrm{GIG}(\nu_w|\gamma_w,\rho_w,\tau_w),\ \ q(\nu_e) = \mathrm{GIG}(\nu_e|\gamma_e,\rho_e,\tau_e), \quad (26)$$

where $\mathrm{GIG}(x|\gamma,\rho,\tau) = \frac{(\rho/\tau)^{\gamma/2}}{2K_\gamma(\sqrt{\rho\tau})}x^{\gamma-1}e^{-(\rho x+\tau/x)/2}$. The updating formulas are given by

$$\gamma_i = \alpha/I,\ \ \rho_i = 2\alpha + \mathbb{E}[\nu_w]\,\mathrm{tr}(\boldsymbol{\Omega}^{-1}\boldsymbol{K}_i),$$
$$\tau_i = \mathbb{E}\Big[\tfrac{1}{\nu_w}\Big]\boldsymbol{x}^T\boldsymbol{\Psi}^T\boldsymbol{\Upsilon}_i^T\boldsymbol{K}_i^{-1}\boldsymbol{\Upsilon}_i\boldsymbol{\Psi}\boldsymbol{x}$$
$$\gamma_w = a_w,\ \ \rho_w = 2b_w + \textstyle\sum_i \mathbb{E}[\theta_i]\,\mathrm{tr}(\boldsymbol{\Omega}^{-1}\boldsymbol{K}_i),$$
$$\tau_w = \textstyle\sum_i \mathbb{E}\Big[\tfrac{1}{\theta_i}\Big]\boldsymbol{x}^T\boldsymbol{\Psi}^T\boldsymbol{\Upsilon}_i^T\boldsymbol{K}_i^{-1}\boldsymbol{\Upsilon}_i\boldsymbol{\Psi}\boldsymbol{x}$$
$$\gamma_e = a_e,\ \ \rho_e = 2b_e + \mathrm{tr}(\boldsymbol{\Omega}^{-1}),\ \ \tau_e = \boldsymbol{x}^T\boldsymbol{\Psi}^T\boldsymbol{\Upsilon}_0^T\boldsymbol{\Upsilon}_0\boldsymbol{\Psi}\boldsymbol{x}. \quad (27)$$

The MAP estimate of $\boldsymbol{a}$ is determined such that the partial derivative of $\mathcal{L}'$ is equal to zero. More specifically, $\boldsymbol{a}^*$ is a solution of the regularized normal equation $(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X} + \lambda\boldsymbol{I})\boldsymbol{a} = \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}$, where $\boldsymbol{\Sigma}^{-1} = \sum_i \mathbb{E}\big[\frac{1}{\nu_w\theta_i}\big]\boldsymbol{\Upsilon}_i^T\boldsymbol{K}_i^{-1}\boldsymbol{\Upsilon}_i + \mathbb{E}\big[\frac{1}{\nu_e}\big]\boldsymbol{\Upsilon}_0^T\boldsymbol{\Upsilon}_0$.

4. EXPERIMENTS

This section reports experiments that were conducted to evaluate the basic performance of the proposed model called IKLP.

4.1. Experimental Conditions

We used two audio signals sampled at 16 [kHz]; One is an unaccompanied solo of a male singer (RWC-MDB-G-2001 No.91) excerpted from the RWC Music Database [16]. We analyzed and resynthesized the signal with STRAIGHT [6] by using ground-truth F0s. The other is a female speech (FSUSA101) excerpted from the ATR Japanese speech database [17]. Since these signals have high F0s over 300Hz, higher-order LP is more difficult. We applied IKLP on a frame-by-frame basis with a window size of 2048 samples ($M = 2048$) and a shifting interval of 160 samples. The hyperparameters were set as $\alpha = 1.0$, $a_w = b_w = a_e = b_e = 1.0$, $P = 30$, and $\lambda = 0.1$. We prepared sufficiently many periodic kernels $\{\boldsymbol{K}_i\}_{i=1}^I$ that represent different F0s ranging from 100 [Hz] to 400 [Hz] at 6-cent intervals, *i.e.*, the truncation level was set as $I = 400 \gg \alpha$.

4.2. Experimental Results

The experimental results (Fig. 2 and Fig. 3) showed the potential of IKLP not only for joint spectral-envelope and F0 estimation but also for detection of pitched segments. We conjecture that speech signals can be analyzed more correctly than singing signals. In Fig. 2, IKLP failed to detect pitched segments around 4.2s–4.6s because of relatively weak harmonic partials. The F0s were estimated accurately in the pitched segments of both signals, even in vibrato around 6.0s in Fig. 2. The estimated filter spectra (envelopes) were little skewed by harmonic partials because the source spectra were explicitly modeled, as shown in the second and third rows of Fig. 2 and Fig. 3.

Although the IKLP model itself is theoretically sound, we found that half pitch errors tend to occur because the optimization method easily gets stuck in bad local optima when using periodic kernels. The likelihood function given by Eq. (15) imposes a larger penalty when the model variance $\boldsymbol{\Psi}^{-1}\boldsymbol{K}\boldsymbol{\Psi}^{-T}$ underestimates the observed variance $\boldsymbol{x}\boldsymbol{x}^T$, resulting in false (overestimated) harmonic partials with a smaller penalty. This implies a deep connection of Eq. (15) to the Itakura-Saito divergence (ISD), which in fact acts as a cost function when we restrict $\boldsymbol{K}$ to an identity matrix (IKLP reduces to ISD-based AR modeling [1]). It is known that the ISD is harder to optimize because of its nonconvexity than the Kullback-Leibler divergence (KLD) [18]. This problem could be solved by representing the temporal dynamics of F0s and/or learning multiple models with different initializations based on rough F0 estimates.

5. CONCLUSION

We presented a nonparametric Bayesian model for joint spectral-envelope and F0 estimation in the solid framework of multiple kernel learning. The experimental results showed that our model is robust to high-pitched signals and can detect unpitched segments. We plan to conduct more comprehensive experiments for comparison. There are several interesting directions of this research. One possibility would be to use more precise models of the source signal $\boldsymbol{\epsilon}$ [19–21] when designing basis functions $\boldsymbol{\Phi}$. The kernel parameters $\boldsymbol{T}$ could be optimized with type-II ML estimation (empirical Bayes). In addition, the deep connection of IKLP to classical ISD-based LP [1] opens up a door to the fundamental generalization of ISD-based nonnegative matrix factorization [15, 22] (known as useful for music signal separation) in that the covariance structure *within the elements of each basis vector* can be considered by using various kernels as in [23].

6. REFERENCES

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

[2] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *IEEE Transactions on Signal Processing*, vol. 39, no. 2, pp. 411–423, 1991.

[3] R. Badeau and B. David, "Weighted maximum likelihood autoregressive and moving average spectrum modeling," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008, pp. 3761–3764.

[4] M. Oudot, O. Cappé, and E. Moulines, "Estimation of the spectral envelope of voiced sounds using a penalized likelihood approach," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 469–481, 2001.

[5] F. Villavicencio, A. Röbel, and X. Rodet, "Improving LPC spectral envelope extraction of voiced speech by true-envelope estimation," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2006, pp. 869–872.

[6] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3–4, pp. 187–207, 1999.

[7] T. Nakano and M. Goto, "A spectral envelope estimation method based on F0," in *Statistical and Perceptual Audition (SAPA) - Speech Communication with Adaptive Learning (SCALE) Conference*, 2012, pp. 11–16.

[8] A. Sasou and K. Tanaka, "Robust LP analysis using glottal source HMM with application to high-pitched and noise corrupted speech," in *European Conference on Speech Communication and Technology (Eurospeech)*, 2001, vol. 4, pp. 2443–2446.

[9] T. Toda and K. Tokuda, "Statistical approach to vocal tract transfer function estimation based on factor analyzed trajectory HMM," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008, pp. 3925–3928.

[10] H. Kameoka, Y. Ohishi, D. Mochihashi, and J. Le Roux, "Speech analysis with multi-kernel linear prediction," in *Acoustical Society of Japan (ASJ) Spring Meeting, 2-Q-24 (in Japanese)*, 2010.

[11] H. Kameoka, N. Ono, and S. Sagayama, "Speech spectrum modeling for joint estimation of spectral envelope and fundamental frequency," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1507–1516, 2010.

[12] G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semidefinite programming," *Journal of Machine Learning Research*, vol. 5, pp. 27–72, 2004.

[13] C. E. Rasmussen and C. K. I. Williams, Eds., *Gaussian Processes for Machine Learning*, MIT Press, 2006.

[14] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Efficient algorithms for multichannel extensions of Itakura-Saito nonnegative matrix factorization," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012, pp. 261–264.

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

[16] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music database," in *International Conference on Music Information Retrieval (ISMIR)*, 2002, pp. 287–288.

[17] A. Kuramatsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kawabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, no. 4, pp. 357–363, 1990.

[18] N. Bertin, C. Févotte, and R. Badeau, "A tempering approach for Itakura-Saito non-negative matrix factorization. with application to music transcription," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009, pp. 1545–1548.

[19] M. Rothenberg, "A new inverse-filtering technique for deriving the glottal air flow waveform during voicing," *Journal of the Acoustical Society of America*, vol. 53, no. 6, pp. 1632–1645, 1973.

[20] D. H. Klatt and L. C. Klatt, "Analysis, synthesis and perception of voice quality variations among female and male talkers," *Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820–857, 1990.

[21] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 26, no. 4, pp. 1–13, 1985.

[22] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[23] A. Liutkus, R. Badeau, and G. Richard, "Gaussian processes for underdetermined source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 59, no. 7, pp. 3155–3167, 2011.