ESTIMATION OF SOURCE-FILTER MODEL VIA ACOUSTICAL FEATURE EXTRACTION BY GA-LIKE ALGORITHM

Mizuki Ihara, Shin-ichi Maeda and Shin Ishii

Graduate School of Information Science,
Nara Institute of Science and Technology
8916-5 Takayama Ikoma, 630-0192 JAPAN

Abstract

This study presents an estimation method for a source-filter model, which takes a temporal continuity of pitch and amplitude into account and is useful, for example, for instrument identification.

We assume pitch and amplitude as hidden variables that tend to change continuously in time while the resonant property is fixed in order to reduce inherent indeterminacy in the source-filter model. In the observation process of this dynamical system, which models the generation of sound spectra from the hidden variables of the dynamics, pitch and amplitude are highly nonlinear and non-Gaussian, i.e., a nonlinear dynamical system. Therefore, it is intractable to analytically estimate the hidden variables as well as the model parameters which define the resonant property. For this parameter estimation, we employed a GA (Genetic Algorithm)-like algorithm. After the parameter of each instrument was estimated from isolated notes, we verified the possibility of this system identification method by reconstructing the spectrum and by whether synthesized log-spectrum are close to the original log-spectrum.

Index Terms – State space methods, Nonlinear acoustics, Nonlinear dynamics, Acoustic filters

1 Introduction

The estimation of elements in sound such as pitch, amplitude and timbre has many applications, including audio encoding with a small number of parameters, sound synthesis, extraction of instrument properties and music transcription. In this study, we in particular focus on one of those applications, instrument identification.

As an example of existing studies of instrument identification, Eronen used instrument features to create an instrument classifier in a hierarchical structure and evaluated it on solo tones from 30 instruments, achieving an identification rate for individual instruments of approximately 80 percent [1]. Brown presented a classifier with Gaussian mixture models, and obrained 75-85 percent accuracy in monophonic music instrument identification [2]. However, these studies could still not identify the instruments perfectly, especially for time-varying monophonic music.

For estimation of sound source or filter, Fant's source-filter model was sometimes used, which is originally used for modeling production processes of sound and speech [4]. In this model, it is assumed that the combination of a sound-source generation pattern and the filter that represents the resonant property of the target instrument produces observable power spectra. Simultaneous estimation of the time-varying sound-source pattern and the resonant property, however, suffers from the problem of indeterminacy; that is, observable spectra can be expressed in various ways. To determine the source-filter model, therefore, some kind of constraint is necessary.

Itakura and Saito attempted to solve this problem by identifying the filter part first. They modeled the short-term speech signal as a stationary Gaussian process and estimated the filter using maximum likelihood spectrum estimation [5]. The assumption of the stationary Gaussian process, however, ignores the effect of time-varying pitch and amplitude. Because of this assumption, the model does not include the continuity of pitch and amplitude, and then, it is not enough to express real sound-source characteristics or resonant properties.

In this study, we aim at simultaneous estimation of a sound-source and a filter with a less number of parameters. We propose an estimation method for the source-filter model that takes a temporal continuity into account by constructing a dynamical system model for the sound-source. In particular, dynamics of pitch and amplitude are considered. Additionally, we assume that the resonant property does not vary in time from the fact that the body of instrument itself

should be consistent in time.

2 The sound generative model

Using the source-filter model, unknown sound source generation G_t and resonant property (the filter part) H_t are both estimated from observable power spectra of sound s_t, whose generation process is described as

$$s_t = \phi(x_t) = G_t \odot H_t, \tag{1}$$

where s_t is a d-dimensional vector representing the spectrum amplitude of each digitized frequency, and the operation $\odot$ means the Hadamard product (element-by-element product). The representation of source-filter model has an inherent indeterminacy that we cannot identify G_t or H_t without additional constraints. To solve this problem, therefore, we introduce the continuity in sound source G_t as a nonlinear dynamical system and assume the resonant property H_t does not vary in time, written as H in the followings.

2.1 Nonlinear dynamical system

The nonlinear dynamical system with Markov properties consists of the observation process and the state transition process:

$$s_t = \phi(x_t) \odot n_1, \tag{2}$$

$$x_t = \psi(x_{t-1}) + n_2, \tag{3}$$

where the function $\psi(\cdot)$ describes acoustical dynamics. In these equations, $x_t = \{a_t, f_t\}$ is a two-dimensional hidden vector representing internal acoustical dynamics, where a_t and f_t denote log-amplitude and pitch, respectively. We express these equations as probabilistic models, $p(s_t|x_t, \theta)$ and $p(x_t|x_{t-1}, \theta)$.

2.2 Observation process

When the sound is assumed to be stationary Gaussian as in Linear Predictive Coding (LPC), Gaussian noise in the time domain is closely represented as multiplicative Chi-square distribution with $\gamma = 3$ in the frequency domain [5], so we employ that for the noise distribution. Since observable spectra can be written as the multiplication of noise and estimated with G_t and H: $s_t = \hat{s}_t \odot n$ where $\hat{s}_t = G_t \odot H$, we obtain

$$\begin{aligned}
&\log p(s_t|x_t, \theta) \\
&= \frac{1}{4\Gamma(1.5)\sigma_o \hat{s}_t(i)} \sum_{i=1}^{N} \left[\log\left(\frac{s_t(i)}{2\sigma_o \hat{s}_t(i)} \right) - \frac{s_t(i)}{\sigma_o \hat{s}_t(i)} \right] \\
&\simeq \frac{1}{4\Gamma(1.5)\sigma_o s_t(i)} \sum_{i=1}^{N} \left[\log\left(\frac{s_t(i)}{2\sigma_o \hat{s}_t(i)} \right) - \frac{s_t(i)}{\sigma_o \hat{s}_t(i)} \right].
\end{aligned} \tag{4}$$

Here, we define the time-fixed function of the resonant property as

$$\begin{aligned}
H(\tilde{\omega}) = 2^{1-p}\{&\sin^2 \frac{\tilde{\omega}}{2} \prod_{k=2,4,\cdots,p} (\cos\tilde{\omega} - \cos b_k)^2 \\
&+ \cos^2 \frac{\omega}{2} \prod_{k=1,3,\cdots,p-1} (\cos\tilde{\omega} - \cos b_k)^2\}^{-2},
\end{aligned} \tag{5}$$

which follows the one of Line Spectrum Pair (LSP) in LPC [6]. In equation (5), $\tilde{\omega}$ represents normalized frequencies $\tilde{\omega} = \frac{\omega \mathbf{Fs}}{2\pi}$, where **Fs** is the sampling frequency, and b_k $(k = 1, \cdots, p)$ is the parameter of H. The sound-source is time-dependent and represented as the sum of Gaussians whose peaks are located at harmonic frequencies:

$$\begin{aligned}
&G_t(\omega_i; a_t, f_t, K, \sigma_p, \tau) \\
&= \exp\left(a_t + A \exp\left(-\frac{\omega_i}{\tau} \right) \sum_k^K N(\omega_i | k f_t, {\sigma_p}^2) \right).
\end{aligned} \tag{6}$$

Here, $-\frac{\omega_i}{\tau}$ indicates exponential decay in frequency, A is an adjusting parameter for power, and $N(x|\mu, \sigma)$ denotes the Gaussian distribution of x with mean μ and variance σ. K and σ_p are the number of Gaussians in the resonant property and the variance of each Gaussian, respectively.

2.3 State transition

In addition to the constraint of time-invariance in the resonant property H, the hidden variables for sound source $x_t = \{a_t, f_t\}$ are assumed to change either continuously or discontinuously in time [7]. Whether continuous or discontinuous is modeled by means of a two-component Gaussian mixture:

$$\begin{aligned}
&p(x_t|x_{t-1}, \theta) = \bar{\eta} N(x_t; m_1, \sigma_1^2) + (1 - \bar{\eta}) N(x_t; m_2, \sigma_2^2) \\
&= \bar{\eta} \left(N(a_t; a_{t-1} + \log\rho, \sigma_{1_a}^2) N(f_t; f_{t-1}, \sigma_{1_f}^2) \right) \\
&+ (1 - \bar{\eta}) \left(N(a_t; m_{2_a}, \sigma_{2_a}^2) N(f_t; m_{2_f}, \sigma_{2_f}^2) \right),
\end{aligned} \tag{7}$$

where ρ is an attenuation constant ranging from 0 to 1, m_1 and m_2 are mean vectors of a_t and f_t. Covariance matrices for continuous transition are $\sigma_1^2 = \{\sigma_{1_a}^2, \sigma_{1_f}^2\}$, and those for discontinuous transitions are $\sigma_2^2 = \{\sigma_{2_a}^2, \sigma_{2_f}^2\}$. The first term corresponds to the continuous transition where amplitude decreases exponentially and pitch does not change much. The second transition corresponds to the discontinuous one, approximated as a Gaussian process with a large variance σ_2. Under these assumptions, the proportion of the state transition being either continuous or discontinuous is represented by $\bar{\eta}$ that takes a value from 0 to 1.

2.4 Joint distribution

For time-series of the observable variable s_t and the hidden variable x_t during $t = 1, \cdots, T$, which are wholely expressed as $S_{1:T} = \{s_1, s_2, \cdots, s_T\}$ and $X_{1:T} = \{x_1, x_2, \cdots, x_T\}$, respectively, the joint distribution for $X_{1:T}$ and $S_{1:T}$ is given by

$$p(X_{1:T}, S_{1:T}) = p(s_1|x_1, \theta)p(x_1|\theta)\prod_{t=2}^{T} p(s_t|x_t, \theta)p(x_t|x_{t-1}, \theta), \quad (8)$$

where θ is the parameter vector that defines the functions of G_t and H. Likelihood $p(S_{1:T})$ can be calculated by integrating this joint distribution with respect to $X_{1:T}$.

3 Parameter estimation

When the model has hidden variables, the EM algorithm has often been used for the parameter estimation. It requires the posterior probability of hidden variables to be calculated, but this calculation is often intractable in many nonlinear dynamical systems, like our case. Therefore, instead of the EM algorithm, we used an equivalent but practically different methodology, a coordinate descent of free energy.

3.1 The EM algorithm and free energy

Free energy is defined for any trial distribution of the hidden variable, $q(X_{1:T})$, as

$$F(q(X_{1:T}), \theta) = -\log p(S_{1:T}|\theta) + KL\left[q(X_{1:T})||p(X_{1:T}|S_{1:T}, \theta)\right]. \quad (9)$$

where $KL[q||p] = \int q(x) \log \frac{q(x)}{p(x)} dx$ is the Kullbuck-Leibler divergence. Apparently, minimizing the free energy with respect to the trial distribution $q(X_{1:T})$ yields the negative log-likelihood $-\log p(S_{1:T}|\theta)$, and in that case, $q(X_{1:T})$ is equal to $p(X_{1:T}|S_{1:T}, \theta)$ because of the positivity of the Kullbuck-Leibler divergence. Therefore, the maximum likelihood (ML) estimation is achieved by the simultaneous minimization with respect to $q(X_{1:T})$ and θ:

$$\hat{\theta} = \arg\max_\theta \log p(S_{1:T}|\theta) = \arg\min_\theta \left(\min_{q(X_{1:T})} \left(F(q(X_{1:T}), \theta)\right)\right). \quad (10)$$

In this optimization, we can employ alternate minimization of the free energy with respect to $q(X_{1:T})$ and θ, and it is known that the minimization of the free energy becomes identical to the ML estimation by the EM algorithm when we employ strict alternate minimization of the free energy. Instead of the intractable calculation of posterior distribution, however, we relax the strict alternate minimization as to restrict the trial distribution $q(X_{1:T})$ being a single Gaussian distribution, and then use a GA-like algorithm for the parameter estimation.

3.2 GA-like algorithm approximation

For the estimation of optimal parameters, we used GA-like algorithm known as a simplex method [8]. This method looks for an optimal point by moving to a new vertex whose function value is equal to or better than that of the previous vertex. When there is no such vertex, the current vertex is the locally optimal solution.

4 Experimental evaluation

4.1 Sound database

To verify the performance of the proposed method as an instrument identification application, we used isolated notes from five kinds of instruments, taken from the University of Iowa Electronic Music Studios samples [9]. The dataset consists of samples of flute, horn, trumpet, viola and cello. For each instrument, we prepared fifteen training data which consists of three data from five different tones, and six test data which contain the different tones with training data.

4.2 A sound-source and a filter estimation

The performance of the proposed system was first examined by seeing if a sound-source and a filter can be well estimated. Fig. 1 shows the original log-spectrum and the log-spectrum synthesized by the identified system for cello. In the log-spectrum after learning parameters, the characteristic of pitch expressed as peaks with constant intervals and the gradually decaying amplitude were well reconstructed.

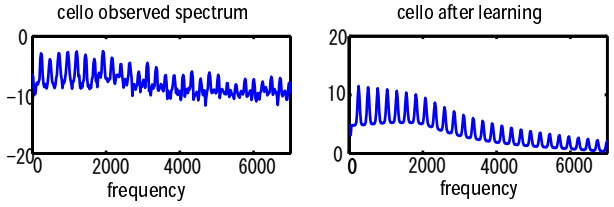

Figure 1: original log-spectrum and synthesized log-spectrum

4.3 Tracking of pitch and amplitude

We next evaluated whether the model can track the pitch and amplitude of the original cello sound. The result of pitch and amplitude tracking is shown in Fig. 2. In the left panel(frequency), since we used three

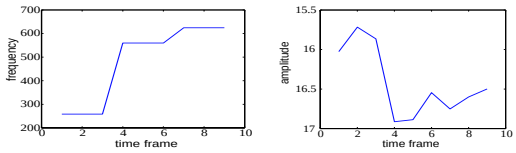

Figure 2: Pitch and amplitude tracking

data for each tone, frequency in each group of three are almost the same. That means the pitch tracking was successful.

4.4 Feature extraction

In addition, the classification performance was plotted with Local Fisher Discriminant Analysis(LFDA) [10]. From the results in Fig. 3 that each instrument tends to group, the system has the possibility to be used in instrument identification.

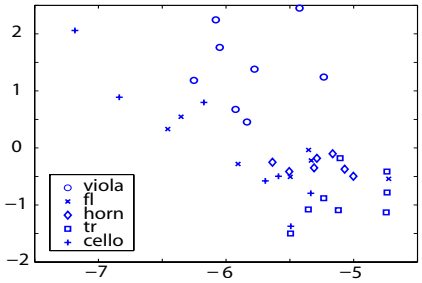

Figure 3: Instrument classification of isolated notes by learned characteristics of H

5 Conclusion and future work

We presented a system identification approach to the estimation of sound-source generation and resonant properties. To consider the time-varying phenomena, a nonlinear dynamical system was employed, while the filter representing the resonant property was fixed in time but estimated. A GA-like algorithm was used for identification of this complex model based on available data. This model well reconstructed the original sounds from the estimated sound source generation G_t and the resonant property filter H_t.

For the practical use of this model for instrument identification, we should evaluate this model with monophonic and polyphonic music to know the ability to identify instruments being played. Although the GA-like algorithm used for parameter estimation does not require explicit gradients of the objective function, it does not guarantee global minimization. In our future work, we also consider more strict free-energy minimization for the parameter estimation.

References

[1] A. Eronen and A. Klapuri, "Musical instrument recognition using cepstral coefficients and temporal features," *ICASSP*, 2000.

[2] J.C. Brown, O. Houix and S. McAdams, "Feature dependence in the automatic identification of musical woodwind instrument," *J. Acoust. Soc. Am.*, Vol.109(3), Mar. 2001.

[3] L.J. Lee, H. Attias, L. Deng, and P. Fieguth, "A multimodal variational approach to learning and inference in switching state space models," *ICASSP*, 2004.

[4] G. Fant, *Acoustical Theory of Speech Production*, The Hague, 1960.

[5] F. Itakura, and S. Saito, "A statistical method for estimation of speech spectral density and formant frequencies (in Japanese)," *Trans. IECE Jpn.*, Vol.53A, pp.36-43, 1967.

[6] N. Sugamura, and F. Itakura, "Speech Data Compression by LSP Speech Analysis-Synthesis Technique (in Japanese)," *Trans. IECE Jpn.*, Vol.64A(8), pp.599-606, 1981.

[7] R.A. Irizarry, "Local Harmonic Estimation in Musical Sound Signals," *Journal of American Statistical Association*, Vol.96(454), 2001.

[8] J.C. Lagarias, J.A. Reeds, M.H. Wright, and P.E. Wright, "Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions," *SIAM Journal of Optimization*, Vol.9(1), pp.112-147, 1998.

[9] The University of Iowa *Electronic Music Studios: MIS*, http://theremin.music.uiowa.edu/

[10] M. Sugiyama, *Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis*, Technical Report TR06-0008, Department of Computer Science, Tokyo Institute of Technology, Japan, 2006.