NAIST-IS-DD0761004

Doctoral Dissertation

Statistical Approach to the Single-Channel Sound Source Extraction

Mizuki Ihara

February 4, 2010

Department of Bioinformatics and Genomics
Graduate School of Information Science
Nara Institute of Science and Technology

A Doctoral Dissertation
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of ENGINEERING

Mizuki Ihara

Thesis Committee:

Professor Kazushi Ikeda	(Supervisor)
Professor Kiyohiro Shikano	(Co-supervisor)
Associate Professor Tomohiro Shibata	(Co-supervisor)
Assistant Professor Shin-ichi Maeda	(Co-supervisor)

Statistical Approach to the Single-Channel Sound Source Extraction*

Mizuki Ihara

Abstract

Extracting only the essential sound attributes from sounds is one of the fundamental issues of computational auditory scene analysis. Especially, the estimation of sound source characteristics is difficult since the corresponding physical quantity is not defined. This dissertation proposes two music information retrieval methods: the instrument feature extraction assuming the timbre space and the probabilistic model of sounds considering the source-filter model and dynamics.

In the former part of this dissertation, an instrument feature extraction method with a combination of linear projection methods is developed. For monophonic music instrument identification, various feature extraction and selection methods have been proposed. Although raw power spectra have enough information for accurate instrument identification, their dimensionality is too high and redundant. It is important to find non-redundant instrument specific characteristics that maintain information essential for high-quality instrument identification to apply them to various instrumental music analyses. As such a dimensionality reduction method, a combination of linear projection methods is introduced: principal component analysis (PCA) and local Fisher discriminant analysis (LFDA). Additionally, the reason why linear projection algorithms are suitable for instrument identification is explained by the geometrical analysis of algorithms. After experimentally clarifying that raw power spectra are actually good for instrument classification, the feature dimensionality is reduced by PCA followed by LFDA. The reduced features achieved reasonably high identification performance that is comparable or higher than those by the power spectra and those

*Doctoral Dissertation, Department of Bioinformatics and Genomics, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD0761004, February 4, 2010.

obtained by existing studies. These results suggest that the proposed PCA-LFDA can successfully extract low-dimensional instrument features that preserve the characteristic information of instruments.

In the latter part, a probabilistic model that represents our assumption with an extension of the source-filter model is introduced to estimate three elements of sounds: pitch, loudness and instrument-specific characteristics. The source-filter model, originally devised to represent a sound production process, has been widely used to estimate both of the source signal and the synthesis filter. This model suffers from an indeterminacy problem. To resolve it, three constraints are included in the model: harmonics, smoothness and sparseness. In detail, the source signal and synthesis filter contain the time-varying fundamental frequency and amplitude information and time-invariant instrument-specific information, respectively in the source-filter model. A probabilistic model that represents those assumptions with an extension of the source-filter model is constructed. For learning of model parameters, an EM-like minimization algorithm of a cost function called the free energy is introduced. Reconstruction of the spectrum with the estimated source signal and synthesis filter and instrument identification by using the model parameters of the estimated synthesis filter are performed to evaluate the proposed approach, showing that this learning scheme could achieve simultaneous estimation for the source signal and the synthesis filter.

Keywords:

Sound source identification, Acoustical feature extraction, Time series analysis, System identification, Pattern recognition, Variational EM algorithm

Table of Contents

List of Figures

List of Tables

CHAPTER I

Introduction

1 Motivation

People can guess who is speaking or singing from sounds in a daily life. Here, they easily solve the task of sound source estimation. However, to simulate estimation of sound sources on machines is a complex task because the sound signals vary depending on the syllables, pitch or tones. Moreover, in the real environment, there are noises and some other present sound sources such as background music and other people's conversation that interfere the identification process. The solution of the human auditory system to this problem suggests that there are time-constant source-specific characteristics that are independent of from other acoustical features, which is known as timbre constancy [1]. We utilize these characteristics to identify a speaker or an instrument. This ability of audition helps to understand the process that we usually sense the source-specific time-constant features from the time-varying sound signals.

It is also known that humans perceive sounds with decomposing into three elements: pitch, loudness and timbre [2]. However, corresponding physical attributes are not discovered yet while it is known that pitch and loudness correspond to the fundamental frequency and the sound intensity, respectively [3]. In this research, the ultimate goal is to extract sound-source characteristics (in this study, the sound source is limited to a musical instrument) by constructing a mathematical model that can extract

time-invariant instrument-specific features like “timbre” and to identify the instrument with the extracted sound-source features.

The current common approach to the problem of sound-source identification is to select some useful features from pre-assumed features of instruments; however, the number of features is too large and too complex to determine which of the feature sets are related to the sound-source. In order to represent sound-source information efficiently, the source-filter model is employed since it consists of small number of parameters.

2 Goals and approach

The final goal of this research is to extract only the certain information from sounds. There are many kinds of studies about it, for example, the blind source separation. The difference between the blind source separation and the study on this paper, the single-channel sound source decomposition is supervised or semi-supervised, so available information about the source is incorporated in the model to decompose sounds.

The aim of this dissertation is not to separate the sounds but to estimate the three elements of sounds, pitch, intensity and timbre. To contribute it, two studies are introduced in this dissertation. The first one focuses on the instrument feature extraction. Existing feature extraction algorithms depend on the global low-level features, which model all instruments together. The method proposed here first projects the instrument features to the instrument feature space to scale them with ignoring other information such as pitch and amplitude. On that instrument feature space, features of each instrument are described with the parameters. The efficiency of the extracted features is evaluated by instrument identification. The second one models the sound to extract three elements of the sound. Sounds have characteristics of harmonicity, smoothness and sparsity. The knowledge about sounds is considered to solve the indeterminacy of the model.

3 Contributions

Although the ultimate focus of this research is to investigate sound-source identification system, it also contributes to other research fields.

- Sound-source identification:
  As mentioned before, the source-filter model consists of a source generation component and a filter component. The model is considered to represent the actual sound generation mechanism well. For example, in the case of the violin, the oscillation of a string produces a source signal generation and the instrument itself works as a filter, while in the case of human speech, generation of the source signal corresponds to the oscillation of vocal cords and the filter corresponds to the vocal tract resonance. Thus, the filter part of the source-filter model is considered to have important information about the sound-source. In this thesis, whether an instrument can be identified by the parameters of the filter component is verified.

  Though the experiment in this dissertation focuses only on instrument identification, the estimation of the source-filter model is also used for identification of a speaker [4]. Speaker verification technology is useful for voiceprint verification known as one division of biometrics.

- Data compression and encoding:
  Since the source-filter model is a highly structured model, we need only a few parameters to represent the sound signal. It serves as a basis for an excellent data compression system, especially for low bit-rate speech compression. Actually, various vocoders that rely on the source-filter model have been proposed [5]. These methods are often combined with other compression techniques.

- Music information retrieval and transcription:
  To estimate pitch, loudness and timbre from a sound, it helps to annotate music. Music annotation gives information about music similarities, so this could be applied into music retrieval like Goto did in his system, Musicream [6]. For music transcription, it is apparent that the tracking of time-varying pitch is important. Since the source generation part can be characterized by fundamental frequency if it exists, the estimation of the source-filter model will serve as useful information for music transcription.

4 Dissertation overview

In this dissertation, two statistical approaches to the sound source extraction from the single-channel monophonic music are proposed. In Chapter II, Instrument identification on monophonic music with low-dimensional instrument features, the combination of machine learning based dimensionality reduction methods is introduced as an instrument feature set followed by that feature set evaluation with instrument identification. Chapter III, Probabilistic harmonic model for single-channel sound decomposition, introduces the probabilistic model for sound decomposition, which considers the sound dynamics of pitch, amplitude and timbre. The dissertation is closed with the concluding chapter, Chapter IV.

CHAPTER II

Instrument identification on monophonic music with low-dimensional instrument features

In this chapter, the method to extract instrument features is proposed, and the features' representation ability is evaluated with instrument identification. This study proposes a method that extracts compact and temporally consistent instrument features to classify instruments at high accuracy rates even from short excerpts. The compact instrument features are extracted by applying a set of linear projection techniques to the log-power spectra. Instrument feature extraction is important for constructing an instrument sound production model, which is useful for music information retrieval, musical transcriptions and sound synthesis. One such application, instrument identification, has three categories of input sources: isolated notes, monophonic music and polyphonic music. In this study, we concentrate on instrument identification tasks for monophonic music. Tasks for polyphonic music are regarded as more important as practical application; however, this problem is more complex than those for monophonic music. To simplify discussion, the topic is focused on the single instrument identification problem to explore a feature extraction technique for identifying single instruments. Such a feature extraction technique will also be useful for polyphonic music source identification because the compact but high quality of instrument characteristic representation is beneficial to reduce the uncertainty in the identification of sources in polyphonic music.

Concerning spectrum-based features such as linear predictive coding coefficients (LPCs), LSF coefficients and MFCCs have been effective in instrument identification in many existing studies. In this study, firstly how much the (short-term) log-power spectrum could be used as instrument features in the context of instrument identification is confirmed, even though such non-stationary temporal information as sound attack or decay is discarded. Also, instrument features are extracted from the log-power spectra, which are redundant but sufficient features for distinguishing each instrument. Note here that most existing spectrum-based features are also reduced features from the power spectra. Since these features are heuristically obtained, the dimensionality reduction criteria are not clear; however, there is almost no quantitative way to evaluate whether they are sufficient or redundant. Although such sufficiency and redundancy should be dependent on the amount of data used to extract the features, heuristically-determined features cannot incorporate the effect of the amount of such available information. In this study, the feature extraction criterion is clarified by quantitatively defining the model of feature extraction and incorporated the effect of the amount of available information into the feature extraction method; features are extracted in an *adaptive* manner based on available training data instead of obtained by a prescribed *analytic* procedure.

Here, two kinds of dimensionality reduction techniques are employed, principal component analysis (PCA) and local Fisher discriminant analysis (LFDA), both of which work in an adaptive manner with the available data. We call the combination of PCA and LFDA as *PCA-LFDA* and also call the extracted (low-dimensional) features PCA-LFDA features. Based on the mathematical properties of the LFDA algorithm, it is assumed that the PCA-LFDA features well represent the instrument characteristics. The performance of the SVM classifiers employing those features was compared with those employing the existing spectrum-based features such as MFCC. Note that many of the existing spectrum-based features are obtained by applying a certain nonlinear transformation to the log-power spectra; however, PCA-LFDA features are restricted in the space of the linear transformation of the log-power spectra but are adjustable because the transformation is learned based on the available training data.

1 Introduction of instrument feature extraction and instrument identification

To represent a sound-source such as a speaker or a musical instrument with a small number of features is important not only for sound or speech compression but also for music information retrieval and music transcription. A sound-source estimation problem, i.e., how to find the parameters that represent the sound-source well, is, however, considered to be difficult because sound signals are usually very high-dimensional and are supposed to change nonlinearly. For instance, when the sampling frequency is 44.1kHz as in commercial compact disks, the length of an input vector becomes as long as 44,100 per second. Also, pitch and musical performance techniques greatly affect the waveform of instrument sounds, and variations on syllables, pitch intonation and other acoustical features of a given speaker affect the speech signals.

The advantage of experiments on single notes is that there are many databases for isolated notes such as [7, 8, 9, 10]. Additionally, instrument features are obtained easily with less computation than those of music or speech. On the other hand, there are drawbacks: it is impossible to obtain the features on note transitions with isolated notes; it is not practical because what humans hear in a daily life is not single notes but mostly continuous music, sound or speech; and even still humans' ability to identify isolated notes' is far higher than those results [11]. For example, when the system tracks fundamental frequency information simultaneously along with instrument identification as in my experiment, it is very easy to determine the fundamental frequency on isolated notes but not at all on polyphonic music. Since general sound-source estimation problems would suffer from these difficulties, this study focuses on instrument identification problems in monophonic music.

For monophonic instrument identification, some existing studies used only one feature extracted from sound signals, while others used a combination of temporal, cepstral, spectral and other acoustic features. Livshin achieved high identification accuracy by using a combination of a total of 62 temporal, energy, spectral, harmonic and perceptual features. Even though the number of features was reduced to 20, the system kept nearly the same recognition rate [12]. In the same way, Essid obtained 70 useful features that were reduced from the originally prepared 160 features of various acoustic kinds, achieving 87% correctness on average in classifying ten instruments

[13]. In these studies, recognition systems employed feature parameters chosen on the basis of a heuristically prepared physical model or acoustical features selected by feature extraction methods.

Instrument identification consists of two phases: an instrument feature extraction phase and a class estimation phase based on the extracted features. For the latter phase, instrument class estimation, Gaussian mixture models (GMM) and support vector machines (SVM) have been widely used. Marques compared the classification accuracies of GMM and SVM and concluded that SVM are superior when identifying isolated notes [14]. Likewise, Agostini compared several classification methods and concluded that SVM is more accurate than other classification methods [15]. Considering these experimental results, SVM is employed as a classifier in all the classification experiments in this dissertation.

There have been several studies of instrument feature extraction followed by instrument identification on monophonic music. Many used certain spectrum-based features as instrument features [14, 16, 17]. When extracting spectrum-based features, not only direct approximation methods of the power spectrum but also those for spectrum envelope estimation have been employed. Marques used 16 mel-frequency cepstral coefficients (MFCCs) for the classification of eight instruments [14], and Essid compared classifiers to classify five woodwind instruments based on ten MFCCs [18]. Line spectral frequencies (LSFs) were also proposed as instrument features that actually showed the identification accuracy of 86% for six instruments which is superior to MFCCs [17].

In contrast to the above approaches, which are based mainly on cepstral and spectral features, in this study we utilize the raw log-amplitude spectrum of given sound signals, which discards time-varying information such as attack or decay, is useful for instrument identification. The result of high accuracy of instrument identification with the raw log-power spectrum reminds us that most of the features used in existing methods such as LSFs or MFCCs are calculated from spectral information. We then reduce the dimensionality of the log-amplitude spectrum by applying a combination of linear transformations based on classical machine learning techniques: principle component analysis (PCA) and local Fisher discriminant analysis (LFDA). Then, we show the relation of algorithms based on the linear discriminant analysis and instrument feature extraction with the geometrical views of those algorithms. Finally, we

compare the identification done by our method with identification carried out by the physical model-based feature extraction methods (LPCs and LSFs), which can be regarded as nonlinear feature extraction methods of the spectrum, in order to see how the model-free and linear-feature-based methods (LFDA and PCA) work.

2 Existing spectrum-based global features

In existing studies, most features are extracted in a cepstral or spectral domain. Here, we call them spectrum-based features. In this section, those feature extraction methods are reviewed.

First, three feature extraction methods are reviewed. They have been shown effective in previous instrument identification experiments: LPC, LSF and mel-frequency cepstrum encoding. These three methods are commonly used to estimate the rough form of spectra called the spectrum envelope. Since the spectrum envelope highly affect human perception of sound sources [19, 20, 21], methods for spectrum envelope estimation are included in instrument feature extraction.

2.1 Linear predictive coding (LPC)

LPC is a popular model-based method that can estimate the spectrum envelope. Assuming that the current sample s_t ($t = 1, \cdots, T$) is represented by a linear combination of p previous samples:

$$s_t = -\sum_{i=1}^{p} \alpha_i s_{t-i} + \epsilon_t, \tag{2.2.1}$$

where α_i ($i = 1, \cdots, p$) is a set of LPC coefficients, and ϵ_t is time-independent (white) noise. The higher the order of LPC coefficients p is, the closer the estimated spectrum comes to the original spectra [22]. On the other hand, the estimated spectrum envelope works for instrument identification when order p is suitably small [14, 16, 23].

LPC coefficients are estimated to suitably represent the original signal by minimizing the total squared error of the signal from T_0 to T_1,

$$\mathcal{E} = \sum_{t=T_0}^{T_1} {\epsilon_t}^2 = \sum_{t=T_0}^{T_1} \sum_{i=0}^{p} \sum_{j=0}^{p} \alpha_i \alpha_j x_{t-i} x_{t-j}. \tag{2.2.2}$$

In practice, to solve the partial differentiation with respect to α_j, there are two major methods: the covariance method and the auto-correlation method. The covariance method limits the range between T_0 and T_1 as p to $N-1$, while the correlation method sets the range as $-\infty$ to $+\infty$ with the regulation $x_t = 0$ for $t < 0$ and $N \leq t$.

By the maximum likelihood spectrum estimation on the frequency domain, which is equivalent to LPC with auto-correlation method on the time domain, the LPC power spectrum is reconstructed with

$$H(\tilde{\omega}) = \frac{\sigma^2}{2\pi} \frac{1}{A_0 + 2\sum_{i=1}^{p} A_i \cos(i\tilde{\omega})}, \tag{2.2.3}$$

where a set of LPC coefficients is computed with $A_i = \sum_{j=0}^{p-|i|} \alpha_j \alpha_{j+|i|}$, $\alpha_0 = 1$ ($i = 0, \pm 1, \cdots, \pm p$). Here, σ^2, p and $\tilde{\omega}$ denote the scale factor, the number of LPC coefficients and a normalized angular frequency $\tilde{\omega} = \frac{\omega \mathrm{Fs}}{2\pi}$ ($-\pi \leq \tilde{\omega} \leq \pi$) where ω and Fs denote the frequency and sampling frequency, respectively [24, 25].

LPC compresses the spectra to adequately maintain the smooth shapes of the spectrum envelopes with a small number of parameters; since there are round-off errors in the floating-point values, LPC suffers from a risk of instability [26]. In addition, the way to choose the order (number) of LPC coefficients is not clear since the optimal order may vary depending on the instrument, the fundamental frequency, the playing style and so on.

2.2 Line spectral frequencies (LSF)

LSF coefficients are based on another representation of LPC coefficients that show better compression rates than LPC because of the improved robustness to the round-off errors. The spectrum envelope is estimated by the following function [27, 28]:

$$H(\tilde{\omega}) = 2^{1-p} \left\{ \sin^2 \frac{\tilde{\omega}}{2} \prod_{n=2,4,\cdots,p} (\cos\tilde{\omega} - \cos b_n)^2 + \cos^2 \frac{\tilde{\omega}}{2} \prod_{n=1,3,\cdots,p-1} (\cos\tilde{\omega} - \cos b_n)^2 \right\}^{-2}, \tag{2.2.4}$$

where $\{b_n\}$ is a set of LSF coefficients. Employing LSF parameters as instrument features on monophonic music, Chétry obtained an average identification accuracy of

86.3% on average [17]. Krishna showed that LSF parameters were superior to other features (linear prediction cepstral coefficients and MFCCs) in isolated notes' spectra compression [29]. These results suggest that LSF can capture instrument's characteristics as a compact representation. Note that LSF also has a problem in the determination of the optimal order (number) of coefficients like LPC.

2.3 Mel-frequency cepstrum encoding

Mel-frequency cepstral coefficients (MFCCs) are calculated in the cepstral domain by applying a few transformations to the spectrum in the frequency domain. Similar to the above frequency domain analyses, power spectra are computed by applying a Fourier transformation to the sound waveform in each time frame. In the frequency domain, linear frequency $f(\mathrm{Hz})$ is converted to mel frequency $\mathcal{M}$ with

$$\mathcal{M} = 1127 \log_e \left(1 + \frac{f}{700}\right). \tag{2.2.5}$$

After that, a logarithm is taken to each mel-spectrum followed by cosine transformation. The lower MFCCs can often be regarded as instrument characteristics because they show nice correspondence to the spectrum envelope. While some studies reported that LSFs outperformed MFCCs as instrument features [29], other studies reported that MFCCs worked relatively well as instrument features compared with LPCs and cepstral coefficients when applied to isolated notes, monophonic music and polyphonic music [14, 16].

3 Features based on linear-projection to the local timbre space

The extraction methods of instrument features are reviewed in the previous section. Although these features help to classify instruments correctly, such a manually determined features might have missed the other useful features. Alternatively, the linear projection methods are utilized in this study to extract useful and compact features from a large number of labeled spectrum data concerning a timbre space similar to [30]. In this section, three machine learning-based linear projection methods are reviewed followed by the proposed instrument feature extraction framework.

3.1 Principal component analysis (PCA)

PCA orthogonally projects the given input data into a low-dimensional linear space. Let $\mathbf{x}$ denote an N-dimensional input vector and $\mathbf{y}$ be a D-dimensional projected output vector, where $N \geq D$. The PCA projection is linear:

$$\mathbf{y} = \mathbf{W}^{\mathrm{T}}\mathbf{x}, \tag{2.3.1}$$

where $\mathbf{W}$ is an $N \times D$ linear transformation matrix and T denotes the transpose. Transformation matrix $\mathbf{W}_{\mathrm{pca}}$ in PCA is obtained by maximization of the total variance of $\mathbf{\Sigma}$ [31]:

$$\mathbf{W}_{\mathrm{pca}} = \underset{\mathbf{W}}{\mathrm{argmax}} \left| \mathbf{W}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{W} \right|, \tag{2.3.2}$$

under the orthogonal constraint $\mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I}$. PCA gives an optimal solution that minimizes the mean squared reconstruction error among all possible linear transformations [32]. Using the mean of all sound samples $\boldsymbol{\mu}$, a covariance matrix of all samples $\mathbf{\Sigma}$ is given by

$$\mathbf{\Sigma} = \sum_{k=1}^{K} (\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu})^{\mathrm{T}}. \tag{2.3.3}$$

The factor loading vector (a row vector of $\mathbf{W}_{\mathrm{pca}}$) that maximizes the total variance of the input data is assigned as the first projection axis; the following axes project the input data to maximize the variance of the data projected onto the subspace, which is orthogonal to all the previous axes. The number of principal components (feature dimensionality) D is usually determined based on the total sum of the contribution for explaining data variance with the first K principal components, called the cumulative contribution ratio.

3.2 Linear Discriminant Analysis (LDA)

While PCA extracts features from the mere input data (unsupervised learning), linear discriminant analysis (LDA) also considers class labels (supervised learning). Similar to PCA, LDA is a linear transformation of input space $\mathbf{x}$ into output space $\mathbf{y}$. In LDA, the linear transformation matrix is determined to maximize the between-class variance

and to maximize the within-class variance. When N-dimensional input space is reduced to D-dimensional output space with $D \times N$ linear transformation matrix $\mathbf{W}$, $\mathbf{W}$ is given by

$$\mathbf{W}_{\mathrm{lda}} = \underset{\mathbf{W}}{\operatorname{argmax}} \operatorname{tr}\left((\mathbf{W}^{\mathrm{T}}\boldsymbol{\Sigma}_W\mathbf{W})^{-1}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\Sigma}_B\mathbf{W})\right), \tag{2.3.4}$$

where $\boldsymbol{\Sigma}_B$ and $\boldsymbol{\Sigma}_W$ denote between-class and within-class covariance matrices, respectively. Let C_i, $\boldsymbol{\mu}_i$ and n_i be the set, the mean and the number of samples of class i, respectively, C the total number of classes and $\boldsymbol{\mu}$ the mean vector of all input samples. Given N-dimensional input vector $\mathbf{x}$, the between-class covariance matrix is the sum of the covariance matrices of each class, and the between-class and the within-covariance matrices are given by

$$\boldsymbol{\Sigma}_B = \sum_{i=1}^{|C|} n_i(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^{\mathrm{T}}, \tag{2.3.5}$$

$$\boldsymbol{\Sigma}_W = \sum_{i=1}^{|C|} \sum_{x \in C_i} (\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_i)^{\mathrm{T}}. \tag{2.3.6}$$

$\mathbf{W}_{\mathrm{lda}}$, defined in (6) is obtained analytically [32]. Unlike PCA, LDA considers the label information in a supervised manner. Since this is more suitable for extracting the features important for label classification, it is actually used in speech recognition [33], e.g. However, LDA may give undesirable results especially when training data size is small or the feature dimensionality is higher than the number of classes, due to its supervised nature. In addition, it does not work well when the input distribution of a certain class has multiple modes [34, 32].

3.3 Local Fisher Discriminant Analysis (LFDA)

Considering the weaknesses of LDA, local Fisher discriminant analysis (LFDA) combines LDA with locality preserving projection (LPP), a linear unsupervised dimensionality reduction method [35]. The projection matrix of LPP, W_{lpp}, is given by minimization of weighted squared error,

$$\begin{aligned}\mathbf{W}_{\mathrm{lpp}} &= \underset{\mathbf{W}}{\operatorname{argmin}} \left(\frac{1}{2}\sum_{jk}^{N} ||\mathbf{W}^{\mathrm{T}}\mathbf{x}_j - \mathbf{W}^{\mathrm{T}}\mathbf{x}_k||^2 A_{jk}\right) \\ &= \underset{\mathbf{W}}{\operatorname{argmin}} \left(\mathbf{A}^{\mathrm{T}}\mathbf{X}\mathbf{L}\mathbf{X}^{\mathrm{T}}\mathbf{A}\right)\end{aligned} \tag{2.3.7}$$

under the constraint, $\mathbf{A}^{\mathrm{T}}\mathbf{X}\mathbf{D}\mathbf{X}^{\mathrm{T}}\mathbf{A} = 1$, where $\mathbf{X}$ is a matrix of all N samples, and $\mathbf{x}_j$ and $\mathbf{x}_k$ are the j^{th} and k^{th} column vectors of $\mathbf{X}$, respectively. $\mathbf{L}$ is the Laplacian matrix, $\mathbf{L} = \mathbf{D} - \mathbf{A}$, with a N-dimensional diagonal matrix $D_{jj} = \sum_{k=1}^{N} A_{jk}$ ($A_{jk} \in [0, 1]$). There are several ways to choose values of the affinity matrix, A_{jk}. Here, A_{jk} is defined by the local scaling of data, which is

$$A_{jk} = \exp\left(-\frac{||\mathbf{x}_j - \mathbf{x}_k||^2}{\sigma_j \sigma_k}\right), \tag{2.3.8}$$

where $\sigma_{\{j,k\}} = ||\mathbf{x}_{\{j,k\}} - \mathbf{x}_{\{j,k\}}^{(m)}||$ with the m^{th} nearest neighbor of $\mathbf{x}_{\{j,k\}}$. A_{jk} takes a large value when the originally neighboring points $\mathbf{x}_j$ and $\mathbf{x}_k$ are projected closely. This method projects samples closely located in the original space to the close output position; in other words, it maintains the locality.

The LFDA transformation matrix is based on that of LDA, Equation (2.3.4). The only difference is that the locality Q_{jk}, Equation 2.3.12, is added to between- and within-class matrices, which are

$$\tilde{\mathbf{\Sigma}}_B = \frac{1}{2}\sum_{j,k=1}^{N} \tilde{Q}_{jk}^{(B)}(\mathbf{x}_j - \mathbf{x}_k)(\mathbf{x}_j - \mathbf{x}_k)^{\mathrm{T}}, \tag{2.3.9}$$

$$\tilde{\mathbf{\Sigma}}_W = \frac{1}{2}\sum_{j,k=1}^{N} \tilde{Q}_{jk}^{(W)}(\mathbf{x}_j - \mathbf{x}_k)(\mathbf{x}_j - \mathbf{x}_k)^{\mathrm{T}}, \tag{2.3.10}$$

where

$$\tilde{Q}_{jk}^{(B)} = \begin{cases} A_{jk}\left(\frac{1}{N} - \frac{1}{n_c}\right) & (x_j = x_k = c) \\ \frac{1}{N} & (x_j \neq x_k) \end{cases} \tag{2.3.11}$$

$$\tilde{Q}_{jk}^{(W)} = \begin{cases} \frac{A_{jk}}{n_c} & (x_j = x_k = c) \\ 0 & (x_j \neq x_k) \end{cases} \tag{2.3.12}$$

LFDA linearly transforms input samples to output samples so that the transformed output samples are separated well when they belong to different classes and they make a local cluster when they belong to the same class [36]. The adoption of LPP allows input distribution of each class to have multiple modes as well as the output dimensionality to be higher than the number of classes [37]. On the other hand, it may over-fit the training data, especially when the data size is small, which is a similar drawback to LDA.

3.4 Proposed Feature Extraction Structure (PCA-LFDA)

The PCA-LDA combination has often been used in practical applications, e.g., face recognition [38, 39], image retrieval [40]. In this study, however, to manage data multimodality, which both LDA and PCA-LDA cannot extract, LFDA is used instead of LDA. Moreover, to eliminate the possible noise that incurs in the input space and to avoid over-fitting due to the possible lack of training data, we conducted PCA-LFDA, which applies LFDA to the dimensionality-reduced space obtained by PCA.

4 Geometrical interpretation of linear-projection algorithms

A sound production of an instrument is well approximated by the source-filter model [1, 41]. In a frequency domain, the source-filter model assumes that the sound spectrum at frequency ω is represented by a product of the source $\boldsymbol{\alpha}(\omega)$ and the filter $\mathbf{R}(\omega)$ as

$$\mathbf{S}(\omega) = \boldsymbol{\alpha}(\omega)\mathbf{R}(\omega), \tag{2.4.1}$$

where the source oscillation is determined according to the fundamental frequency of the sound, and the filter property is determined by the instrument-specific resonant property.

Since every source is assumed to have harmonics on multiples of the fundamental frequency, any source $\boldsymbol{\alpha}'(\omega)$ is well represented by a certain fixed source $\boldsymbol{\alpha}(\omega)$ as $\boldsymbol{\alpha}'(\omega) = \boldsymbol{\alpha}(\kappa\omega)$, where κ denotes a scalar coefficient; e.g., $\kappa = 2$ when the values of fundamental frequency of $\boldsymbol{\alpha}'(\omega)$ is twice of that of $\boldsymbol{\alpha}(\omega)$ [42].

Let $\boldsymbol{\alpha}$ be the vector of the discritized spectrum along the log-frequency axis, and $\mathbf{R}$ be the diagonal matrix. When we consider a discrete spectrum along with the log-scale frequency, the pitch modulated source vector $\boldsymbol{\alpha}'$ is represented by a certain element-wise shift of $\boldsymbol{\alpha}$. The example of the shift is shown in Figure 2.4.1.A to Figure 2.4.1.B. The elements of the diagonal matrix $\mathbf{R}$ hold a filter property, corresponding to the shift invariant on the log-frequency axis (Figure 2.4.1.C). In the following, we write $\mathbf{R}$ of instrument i as $\mathbf{R}_i$ and the basis spectrum of instrument i as $\mathbf{R}_i\boldsymbol{\alpha}$ (Figure 2.4.1.D).

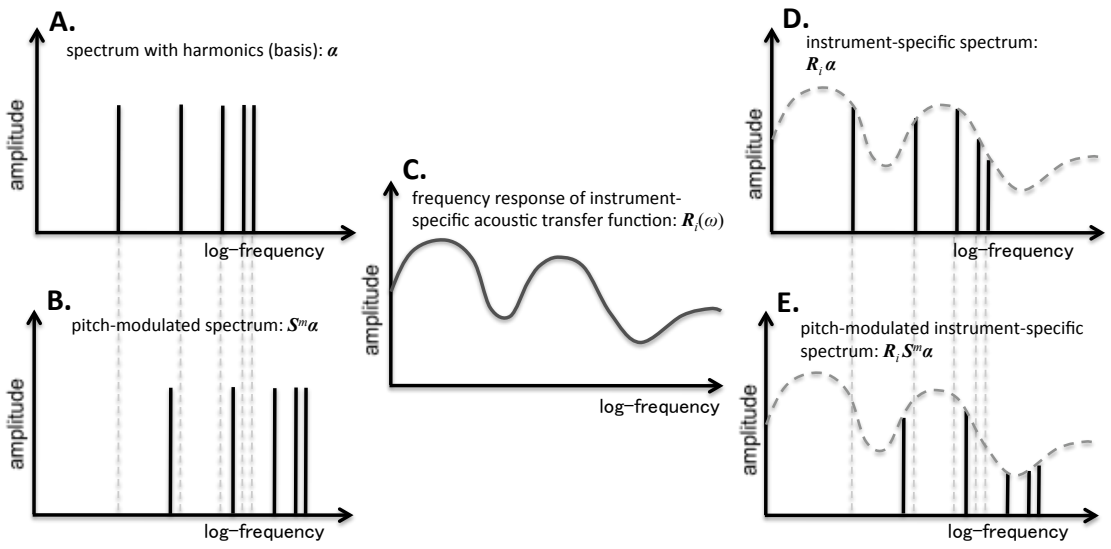

Figure 2.4.1. Basis spectrum (A) and its pitch-modulated spectrum (B) of one instrument through the instrument-specific transfer function (C) resulted in the instrument-specific spectrum (D) and the pitch-modulated instrument-specific spectrum (E) in the log-frequency scale

Given the $N \times N$ cyclic shift matrix

$$\mathbf{S} = \begin{pmatrix} 0 & 1 & 0 & \dots & \dots \\ 0 & 0 & 1 & \ddots & \dots \\ \vdots & \vdots & \ddots & \ddots & \ddots \\ 0 & 0 & \dots & \ddots & 1 \\ 1 & 0 & \dots & \dots & 0 \end{pmatrix}, \tag{2.4.2}$$

the shift is represented by the power of the shift matrix $\mathbf{S}^m$, i.e., $\boldsymbol{\alpha}' = \mathbf{S}^m\boldsymbol{\alpha}$ where m is a certain integer assuming the vector on the log-frequency is discretized at equal intervals. Then, the spectrum is represented with $\mathbf{R}_i\mathbf{S}^m\boldsymbol{\alpha}$ (Figure 2.4.1.E), where $\mathbf{R}_i$ is a diagonal matrix whose diagonal elements denote the filter property of the corresponding log-frequency scale shift invariantness of instrument i.

By the left multiplication of the vector $\mathbf{R}_i\mathbf{S}^m\boldsymbol{\alpha}$ with $\mathbf{R}_i^{-1}$, an inverse matrix of $\mathbf{R}_i$, we can obtain only the source information, $\mathbf{R}_i^{-1}\mathbf{R}_i\mathbf{S}^m\boldsymbol{\alpha} = \mathbf{S}^m\boldsymbol{\alpha}$, which does not depend on the pitch modulation m, as geometrically shown in Figure 2.4.2.A. In other words, the source information is that the spectra of a certain instrument are represented as a basis spectrum $\boldsymbol{\alpha}$ with the shift $\mathbf{S}^m$ for any pitch. Therefore, the projections of all instrument samples of all pitches to the unit vector $\mathbf{1}$ agree, i.e. $\mathbf{1}^{\mathrm{T}}\mathbf{R}_i\boldsymbol{\alpha} = \mathbf{1}^{\mathrm{T}}\mathbf{R}_i\mathbf{S}^m\boldsymbol{\alpha}$. The above discussion suggests that in-class variance of sam-

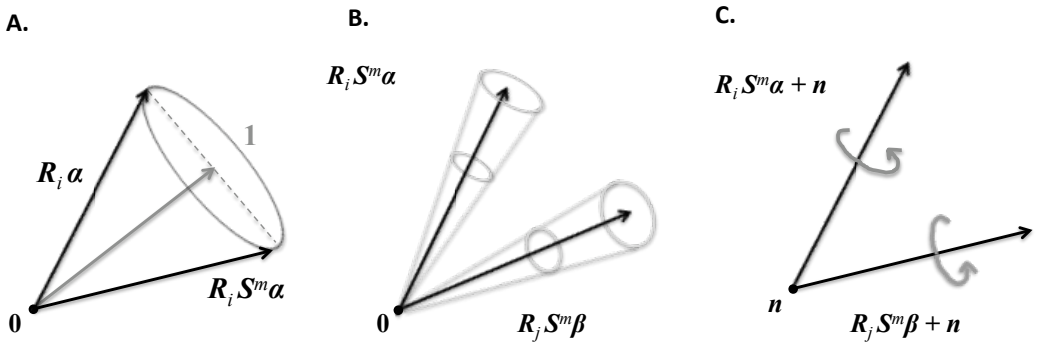

Figure 2.4.2. Geometrical view of the projections of the basis spectrum $\boldsymbol{\alpha}$ and the pitch modulated spectrum $\mathbf{S}^m\boldsymbol{\alpha}$ along the unit vector $\mathbf{1}$ (A), the projections of two instruments with spectra $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ (B) and those with additive noise (C)

ples transformed with $\mathbf{1}^{\mathrm{T}}\mathbf{R}_i^{-1}$ is ideally zero when the label of classes is the kind of instruments.

In the same way, we introduce the instrument j, whose basis spectrum is $\mathbf{R}_j\boldsymbol{\beta}$. Here, $\boldsymbol{\beta}$ is the discritized source spectrum in the log-frequency scale. The projection of vector $\mathbf{R}_i\mathbf{S}^m\boldsymbol{\alpha}$ along the instrument-specific space $\mathbf{R}_i^{-1}\mathbf{1}$ is given by $\mathbf{1}^{\mathrm{T}}\mathbf{R}_i^{-1}\mathbf{R}_i\mathbf{S}^m\boldsymbol{\alpha}$, $\mathbf{1}^{\mathrm{T}}\mathbf{S}^m\boldsymbol{\alpha}$, which is shift-variant to $\boldsymbol{\alpha}$. When samples from the instrument with the spectrum vector $\boldsymbol{\alpha}$ are projected on $\mathbf{R}_i^{-1}\mathbf{1}$ as well as the instrument with that of $\boldsymbol{\beta}$ on $\mathbf{R}_j^{-1}\mathbf{1}$ ideally, the variance of projected samples become small, which allows us determine the boundaries of different instruments.

The projection of LDA is based on the criterion which maximizes the between-class variance and minimizes the within-class variance. This is equivalent to estimate $\mathbf{1}^{\mathrm{T}}\mathbf{R}_i^{-1}$ in $\mathbf{1}^{\mathrm{T}}\mathbf{R}_i^{-1}\mathbf{S}^m\boldsymbol{\alpha}$ from given samples. The geometrical view of LDA projections of two instruments are in the Figure 2.4.2.B.

Assuming that the additive observation noise, $\mathbf{n}$ are added to the sets of instrument samples $\mathbf{R}_i\mathbf{S}^m\boldsymbol{\alpha}$ and $\mathbf{R}_j\mathbf{S}^m\boldsymbol{\beta}$, respectively, it only causes the parallel translation. Considering this additive noise spectrum as a bias in the geometrical interpretation, the vectors shift from the origin of the coordinate axes as from Figure 2.4.2.B to Figure 2.4.2.C. The algorithm of LFDA is explained as consideration of the locality to that of LDA. LFDA keeps the shift-invariance since the within-class variance (Equation 2.3.10) does not change even if the bias is added to the samples from the same class $\mathbf{x}_j$ and $\mathbf{x}_k$. Since LFDA also holds the translation-invariance, it achieves the instrument feature projection with the minimal within-class variance.

It suggests that we can find the transformation that minimizes the between-class variance. Since the projections of LDA and LFDA are based on the criteria which maximizes the between-class variance and minimizes the within-class variance, LDA and LFDA can find a similar transformation matrix to the matrix whose column vector is $\mathbf{1}^{\mathrm{T}}\mathbf{R}^{-1}$, where $\mathbf{R}$ varies according to the instrument.

5 Classification method: SVM

Support Vector Machine (SVM), which is one of the most popular pattern classification methods in the field of machine learning. This method is extensively applied to various tasks including not only instrument identification of isolated notes, monophonic music

[14, 17, 13] and polyphonic music but also speaker identification.

SVM are associated with two important concepts: margin maximization and kernel tricks. When samples are linearly separable between different classes, more than one boundary hyper-plane exists, which accurately separate those samples. As a criterion to determine the best boundary among them, Vapnik's idea maximizes the distance between the boundary and the samples closest to the boundary, which are called support vectors. On the other hand, a linearly separable space could be found by projecting the input samples onto a possibly high-dimensional space even if the input samples are not linearly separable in the original input space. In SVM, a kernel function

$$k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \tag{2.5.1}$$

is assumed, and the left-hand side is directly defined instead of defining mapping function $\phi(\cdot)$ [43]. Here, $\mathbf{x}_i$ and $\mathbf{x}_j$ are i^{th} and j^{th} input samples, respectively. $k(\cdot)$ is a kernel function, and $\phi(\cdot)$ is a (no more important) mapping function. Learning and classification by SVM can be employed by kernel function $k(\cdot)$ without caring about mapping $\phi(\cdot)$. This is called the kernel trick.

For the implementation of SVM, matlab interface of libsvm package available at National Taiwan University [44] was used in the following experiments. Though SVM is originally a binary classifier, the combination of it allows us to solve the multi-class problem. There are several strategies for multi-class SVM. The popular ones are "one-against-one" and "one-against-all." Implementation in libsvm adopted one-against-one because it takes less time than that of one-against-all for training. The SVM type used in the experiments is C-SVC with the linear kernel.

6 Experiments and results

Applying to several monophonic instrument identification tasks, we examined how well the proposed method, PCA-LFDA, works to extract compact instrument features. Due to the lack of standard databases of instrument identification tasks for monophonic music, however, directly comparing the performance of the proposed method with existing methods is difficult, unless we implement those methods. If the experimental conditions, such as the number and the types of instruments to be classified, the time length of the input sound and the content of the sound data itself, differ between existing studies, direct comparison of their classification performance can be meaningless.

In this study, therefore, we implemented some of the existing methods and collected data for evaluation (and also for training) from various sources so that the comparison is meaningful.

6.1 Monophonic music sound database

Sound samples were collected from monophonic commercial CDs of various genres as well as various recording environments. Eight instruments were chosen to include at least two from three instrument categories: violin (vn.), cello (vc.), guitar (gt.) and piano (pf.) from strings; flute (fl.) and oboe (ob.) from woodwinds; and horn (hr.) and trumpet (tp.) from brass instruments. Monophonic songs (sampling frequency: 44.1 kHz) were divided into 0.046 sec. without overlapping. Silent samples were removed in advance. The number of total samples, which were taken from 30 different CDs with 47 different sources (vn: 6, vc: 6, gt: 5, pf: 7, fl: 7, ob: 6, hr: 5, tp: 5; some CDs include more than one instruments), is 38507 (vn: 6612, vc: 5005, gt: 6524, pf: 5366, fl: 5783, ob: 3008, hr: 2498, tp: 3711). CDs and instruments played in CDs are listed in appendix C.

In Experiment 1 and Experiment 2, half of available data for each sample are chosen randomly as training data and the rest as test data. In Experiment 3, we evaluate the test instrument identification rate averaged over the sources when none is included in the training dataset; the test data samples are taken from one source, and the training data samples are taken from the remaining 46 sources. Then, the averaged performance is evaluated by changing the source for the test dataset. We call the first sample set as *mixed-CD* and the second one as *leave-1CD-out*.

The reason two sample sets are prepared is to verify effects of sound recording environment to the identification results. As mentioned in [**?**], difference in recording environment might affects values of parameters. One solution is to substract the long-term average from the parameters in each time-frame [**?**]. In this study, *leave-1CD-out* takes test samples and training samples from different sources. Thus, this would be free from source effects more than the that of *mixed-CD* results.

6.2 Signal processing

The sound data spectra were obtained as follows. First, the log-power spectra are calculated. We used the Fourier transformation after applying a Hamming window to obtain a 1,024-dimensional spectrum. The dimensionality of the features extracted by PCA-LFDA was temporarily set as ten (dimensionality effect is evaluated later), which was small enough compared to the original spectrum dimension, 1,024, but included enough information for instrument identification tasks.

6.3 Instrument Identification Experiments

To evaluate the effectiveness of PCA-LFDA, our feature extractions method, classification performance based on the extracted features was compared with those by the existing methods. We performed three experiments: the first two were preparatory, followed by a main experiment. We used several sources in the experiments, but the test dataset was always independent from the training dataset to avoid the "information leak."

Experiment 1 - Log-power spectrum classification

In the first experiment, a basic thing is examined: whether the 1,024-dimensional raw log-power spectrum itself has enough information for instrument classification. With 10 test trials of instrument identification, we obtained 96.11% correctness with 0.300 standard deviation whose training and test datasets were chosen randomly and independently from all sound samples in each trial. This result suggests that each instrument can be well characterized solely by its short-term spectra, without other information, like temporal change in the spectra.

Experiment 2 - Feature Dimensionality

Since the previous experiment suggested that the raw log-power spectrum contains enough information for instrument identification, it is investigated further whether we can reduce the data dimensionality by PCA-LDA and the proposed feature extraction method, PCA-LFDA. First, the dimensionality of features is fixed LFDA to ten and then investigated how much the first-step PCA reduced the dimensionality before the

second-step dimensionality reduction by either LDA or LFDA. Figure 2.6.1 shows the average identification rate by 10-fold cross-validation against the cumulative contribution ratio of the principal components obtained by the first-step PCA. Since the dimensionality extracted by the second-step LDA or LFDA was constant (ten), this experiment examined how the first-step PCA worked for the whole feature extraction process. The best identification rates for LDA and LFDA were obtained using 310 or

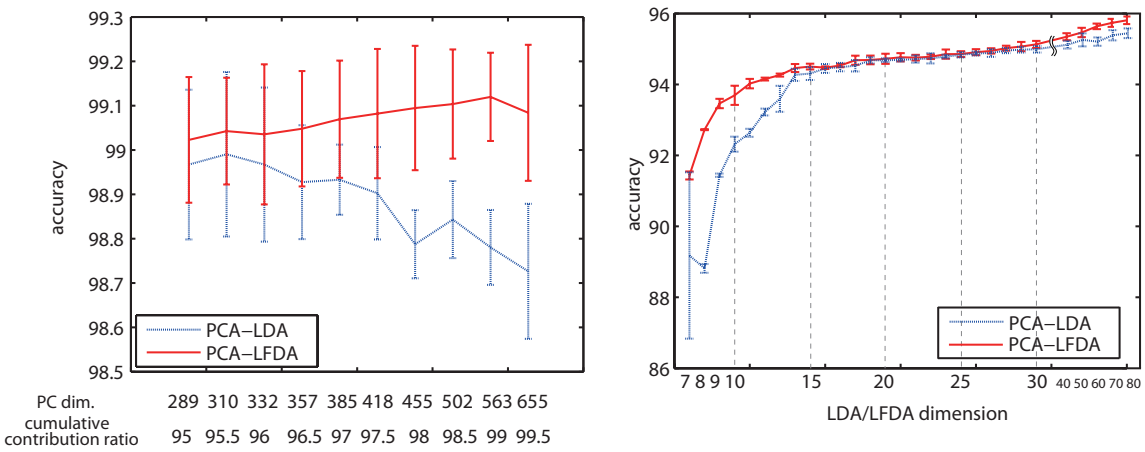

Figure 2.6.1. Instrument identification rates of PCA-LDA (red, solid) and PCA-LFDA (blue, dotted) with various principal components (A; number of LDA/LFDA features was fixed at 10) and with various LDA or LFDA dimension (B; PCA dimension is fixed at 310 for PCA-LDA and 563 for PCA-LFDA)

563 principal components with 95.5% and 99% as the cumulative contribution ratio, respectively. When the second-step LFDA extracted ten features from these 310 or 563 principal components, the identification rate was higher than that of the 1,024-dimensional raw log-power spectrum, suggesting that we can deftly represent the instrument features by PCA-LDA or PCA-LFDA without degradation of classification accuracy. Then we investigated the influence of the dimensionality of the feature extracted by LDA or LFDA on instrument identification.

In the same way, the feature dimensionality of LDA and LFDA is evaluated with the fixed number of principal components obtained by the first-step PCA, 310 or 563. The dimensionality over 10 gives the accuracies higher than 98% for LDA and 99% for

LFDA. By considering the trade-off between complexity and accuracy, fixing the feature dimensionality extracted by PCA-LDA or PCA-LFDA around ten seems reasonable. In the following experiments, therefore, we evaluated several feature extraction methods by fixing the feature dimensionality at ten.

Experiment 3 - Monophonic Music Instrument Identification

As explained at the beginning of this section, we prepared total 47 sound sources. Note that our experimental condition is not exactly consistent with the existing studies. For example, some studies only used samples from normal playing styles [45] and samples from limited pitch ranges [46, 47], whereas we collected samples from all instrument playing styles to extract the instrument's characteristic features, which remain constant despite playing-style variations.

From the result of Experiment 2, the PCA-LFDA instrument identification accuracy does not change much for all LFDA dimensionality over ten; in addition, the feature dimensionalities of many of existing studies are between 10 and 20. Therefore, we extract 10-dimensional features for each feature extraction methods. In other words, the coefficients of methods to compare, LPC [14], LSF [17] and MFCC [18] are ten, and dimensionality is reduced to ten with PCA, LDA, or LFDA. In the cases of PCA-LDA and PCA-LFDA, from Experiment 2, we first reduce the spectrum dimensionality with PCA to 385 (for PCA-LDA) or 563 (for PCA-LFDA) followed by dimensionality reduction into 10 with LDA or LFDA. The leave-1CD-out SVM classification results are shown in the left panel of Figure 2.6.2 and Table 2.6.1.

MFCC, *LPC* and *LSF* denote 10-dimensional mel-frequency cepstral coefficients, linear predictive coefficients and line spectral frequencies, respectively, and *LP* denotes the original raw log-power spectra. In Figure 2.6.2, blue bars are results with training data and red bars are results with test data. Identification rates shown in Table 2.6.1 are average test accuracies over 10 different trials. Note that the test dataset was taken independently of the training dataset.

To verify whether the source-dependent factor affected the instrument identification results, we compared the *leave-1CD-out* results with the results by the following *mixed-CD* validation method. Different from leave-1CD-out training and test datasets, the samples of *mixed-CD* were taken from all 40 mixed sources, but avoiding sample overlap between the training and test datasets. The right panel of Figure 2.6.2 and

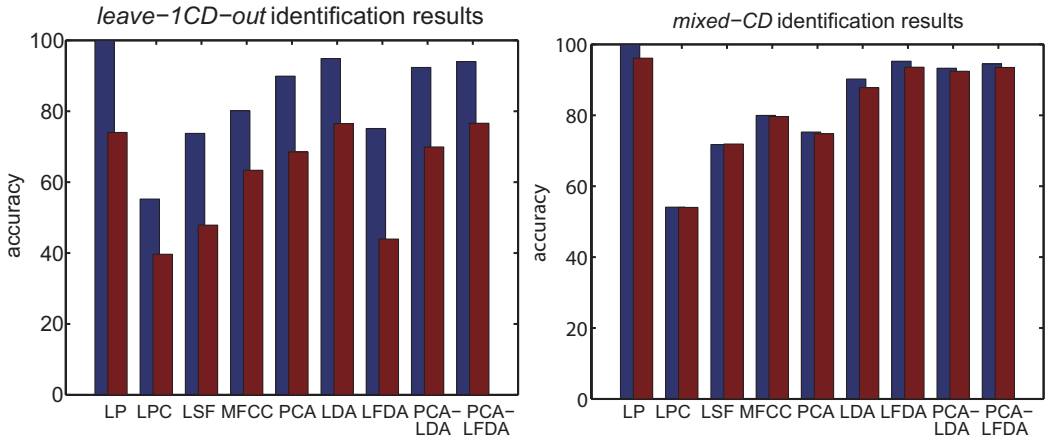

Figure 2.6.2. Leave-1CD-out (left) and mixed-CD (right) average classification results of 10 trials with our 8-instrument-class datasets for training (blue) and test (red) samples

Table 2.6.2 show the training and test accuracies evaluated by the *mixed-CD* method.

Comparing the results of *leave-1CD-out* and *mixed-CD*, the results of *mixed-CD* are relatively higher than those of *leave-1CD-out*. This suggests that in *mixed-CD*, perhaps not only instrument-specific characteristics but also source-dependent features have been extracted, which was also discussed in previous studies [14, 12]. For the leave-1CD-out datasets, the proposed method (PCA-LFDA) achieved the highest identification rate among all ten-dimensional features we compared. Moreover, the relatively small difference between the training and test accuracies suggests that PCA-LFDA does not suffer much from the over-fitting that actually occurred in LFDA. Accordingly, these encouraging results lead to the conclusion that our proposed feature extraction method, PCA-LFDA, is useful to extract essential instrument features from monophonic music excerpts.

Some existing studies failed to clearly state the preparation method of the training and test datasets [48, 16, 46, 47], and others suggested the importance of using training and test datasets taken from different sound sources in instrument identification [14, 12, 17]. We experimented following the presumption suggested by the latter class of studies. As a reference, we summarize the classification results of the existing studies in Table 2.6.3. In the table, the "y" in the "source mixed" column denotes that the

instrument features	training data accuracy (%)	standard deviation	test data accuracy (%)	standard deviation
raw spectrum	100	0	74.01	4.167
LPC (used in [14])	55.23	0.291	39.66	8.056
LSF (used in [17])	73.75	0.301	47.87	9.365
MFCC (used in [18])	80.10	0.413	63.34	5.690
PCA	89.90	0.109	68.52	4.305
LDA	94.85	0.077	76.51	3.970
LFDA	75.13	0.466	43.96	12.633
(proposed methods)				
PCA-LDA	92.34	0.117	69.92	7.074
PCA-LFDA	94.00	0.080	76.61	3.575

Table 2.6.1. Comparison of the instrument identification rates of leave-1CD-out training and test data using proposed features and their standard deviation

training and the test samples were taken from the same source (including when both were taken from the RWC database [49, 50]), and "n" denotes that they were taken from different sources, which is [46]. When both letters are found, the experiments were done for both of the above two settings, but the results were shown only for the source-mixed setting. The identification accuracies of [17, 12] are higher than our *leave-1CD-out*, although they selected training and test samples from different sources. Because their experimental conditions are different from ours, direct comparison is not very meaningful. Livshin proposed a sample outlier omission process for simplifying dataset treatment. Chétry showed high instrument accuracy with LSF features, whereas based on our experiment the accuracy with the LSF features was around 50% because the time length of each sample in Chétry's experiment was much longer (300 sec.) than the samples used in our experiment (0.046 sec.). If we utilized longer excerpts, our classification performance would improve.

instrument features	training data accuracy (%)	standard deviation	test data accuracy (%)	standard deviation
raw spectrum	100	0	96.10	0.024
LPC (used in [14])	54.10	0.374	54.03	0.114
LSF (used in [17])	71.72	0.173	71.89	0.239
MFCC (used in [18])	79.97	0.149	79.66	0.013
PCA	75.29	0.414	74.79	0.038
LDA	90.22	0.024	87.78	0.134
LFDA	95.26	0.044	93.57	0.033
(proposed methods)				
PCA-LDA	93.28	0.114	92.42	0.046
PCA-LFDA	94.47	0.022	93.50	0.035

Table 2.6.2. Comparison of the instrument identification rates of mixed-CD training and test data using proposed features and their standard deviation

6.4 Analysis of PCA-LFDA results

Leave-1CD-out confusion matrix

The confusion matrix and its percentage in Table 2.6.4 and Table 2.6.5, respectively, shows in all 47 leave-1CD-out test trials (eight instruments times five sources), when the test samples were classified based on the PCA-LFDA features. [1] "T" and "E" in the upper left box denote the true and estimated instruments, respectively. This table shows that instrument identification error occurs mainly between instruments in the same instrument categories, string, woodwind and brass, because these instruments have overlapping ranges of fundamental frequencies or their sound production mechanisms are similar.

[1] In a confusion matrix of size $l \times l$, where l is the number of classes, the column and row represent the instances of a predicted and an actual class respectively [51]. It represents how accurately the test datum is classified.

Authors	number of instruments	number of features	accuracy (%)	confidence intervals	source mixed	sample length (sec.)
leave-1CD-out	8	10	76.61	73.77-83.39	n	0.046
mixed-CD	8	10	93.50	89.92-94.41	y	0.046
Marques [14]	8	16	70	n/a	n	0.2
Livshin [12]	7	62	88	81-94	n	1.0
Chétry [17]	6	16	86	72-98	n	300
Eggink [46]	6	120	66	56-85	n,y	2-10
Essid [13]	10	19	87	66-100	y	0.5
Jinachitra [47]	6	28	66	n/a	y	0.5
Ventura [16]	5	12	99	97-100	y	10
Brown [48]	4	10	n/a	79-84	y	2.0-7.8

Table 2.6.3. Summary of classification results of existing studies (*y*: sources mixed, *n*: sources not mixed)

Mixed-CD visualization

Samples of each instrument were plotted in a two-dimensional space reduced by PCA-LFDA. PCA-LFDA dimension reduction was done in the same manner as in the previous experiments except that the dimensionality reduced by LFDA (LFDA dimension: two). For intelligible visualization, we randomly chose 50 samples from each instrument sample set (Figure 2.6.3). Even though only two features were used for this visualization, each instrument's sample was gathered to form a cluster, and the clusters of instruments belonging to the same instrument category are also closely clustered: string instruments (violin, cello, guitar and piano: triangles), woodwind instruments (flute and oboe: circles) and brass instruments (trumpet and horn: squares). Moreover, three instruments (flute, oboe and trumpet), which are frequently confused in human listening tests [2], are all located at the right-bottom corner. Such human misclassification is consistent with our experiment shown in Table III.

[2]This human listening test was informal; we simply presented music samples to subjects and asked them to identify the instrument.

T \ E	vn.	vc.	gt.	pf	fl.	ob.	hr.	tp.
vn.	4985	38	36	6	82	69	349	1047
vc.	111	3597	1033	90	4	1	4	165
gt.	18	1999	3121	1175	0	0	1	210
pf.	2	140	1276	3557	10	0	2	379
fl.	89	0	0	2	2571	184	684	84
ob.	8	0	0	0	1747	1011	595	254
hr.	55	0	0	0	398	1414	822	55
tp.	561	81	162	221	535	122	645	2700

Table 2.6.4. Confusion matrix for leave-1CD-out test samples with linear-kernel SVM (T: true instrument label, E: estimated instrument label)

7 Summary and discussion

In this study, we presented a machine learning-based method (PCA-LFDA) to extract low-dimensional features to characterize each instrument and its effectiveness in geometrical interpretation and instrument classification tasks. We first evaluated how accurately the instruments were classified with only the log-power spectra information and suggested that spectra contain enough information of instrument characteristics. Next, we reduced the dimensionality of the log-power spectra by combining two machine learning-based linear dimensionality reduction methods: PCA and LFDA. Even when the feature dimensionality was reduced to ten, instrument classification by SVM based on the reduced features was as accurate as that based on the raw log-power spectra.

Most existing studies used heuristically pre-determined features as instrument features or features not directly selected from the log-power spectrum. In contrast, our features were extracted by linear feature projection techniques in an adaptive manner to the given dataset and still achieved a rather high instrument identification performance compared with the existing methods. Moreover, because our feature extraction method is based on linear transformation of the spectra, the extracted instrument features will be useful for instrument identification in polyphonic music in which the

T \ E	vn.	vc.	gt.	pf	fl.	ob.	hr.	tp.
vn.	75.4	0.6	0.6	0.1	1.2	1.0	5.3	15.8
vc.	2.2	71.9	20.0	1.8	0.1	0.0	0.1	3.3
gt.	0.3	30.6	47.8	18.0	0	0	0.0	3.2
pf.	0.0	2.6	23.7	66.0	0.2	0	0.0	7.0
fl.	2.5	0	0	0.1	72.8	5.2	19.4	2.4
ob.	0.2	0	0	0	48.3	28.0	16.5	7.0
hr.	2.0	0	0	0	14.5	51.5	30.0	2.0
tp.	11.1	1.6	3.2	4.4	10.6	2.4	12.8	53.7

Table 2.6.5. Confusion matrix percentage for leave-1CD-out test samples with linear-kernel SVM (T: true instrument label, E: estimated instrument label)

spectra are well approximated by a linear summation of the spectra of the constituent instruments.

Another important observation of this study is that the classification results are substantially affected by the sound source; namely, the test instrument identification rate was poor when the training datasets were taken from isolated notes or different sound sources from the test datasets, which was also discussed in previous studies [14, 12]. One plausible explanation is that the feature extraction methods extracted such source dependent features as recording circumstances that do not correspond to the characteristic features of instruments. To avoid such unwanted over-fitting to the source dependent features, the feature extraction methods should be trained and evaluated on large datasets taken from a large variety of independent sources.

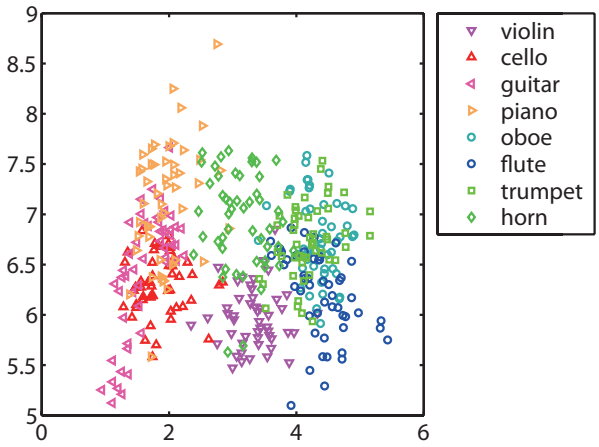

Figure 2.6.3. Distribution of leave-1CD-out randomly chosen 30 test samples of eight instruments in a two-dimensional PCA-LFDA feature space

CHAPTER III

Probabilistic harmonic model for single-channel sound decomposition

For the sound-source identification problem, some traditional approaches have been based on Fant's source-filter model that was originally proposed for modeling production processes of sound and speech [41]. This model assumes that the combination of a sound-source generation pattern G and the synthesis filter H, which represents the resonant property of the target instrument, produce sound signals, whose power spectrum is represented as s. In the case of a violin, for example, the oscillation of a string generates source signals, and the body of the instrument works as a filter. In the case of human speech, generation of source signals is due to the oscillation of vocal cords, and the filter corresponds to vocal tract resonance.

Non-negative matrix factorization (NMF) is widely used for the single-channel sound source decomposition with various constraints. In NMF, the observed signals are represented with factorization of two matrices, for instance, $\mathbf{V}$ and $\mathbf{W}$, the source basis matrix and gain matrix. Some studies simply use the original NMF with additional penalty terms, which would represent sound characteristics such as sparseness [52, 53, 54], continuity [55, 56, 54, 57], fixed source [58, 59] and harmonicity [60].

Both of the source-filter based models and the matrix factorization based models suffer from the problem of indeterminacy; that is, observed spectra can be expressed by various combinations of G and H for the source-filter model and $\mathbf{V}$ and $\mathbf{W}$ for

the matrix factorization based models. In order to estimate those combinations, some constraints are necessary to relax the indeterminacy.

In this chapter, the model assuming that the sound generated by an instrument is well approximated by the source-filter model is introduced. On these assumptions, instrument-specific features and pitch information are extracted from monophonic music excerpts. The model requires to estimate two elements, the sound source and the synthesis filter. To reduce the indeterminacy in the source-filter model, assumptions of harmonics, temporal continuity and sparseness are taken into account. In particular, dynamics of pitch and loudness are considered. For learning of model parameters, the variational EM algorithm is employed. The numerical experiments for instrument identification show that the proposed algorithms work well.

1 Source-filter model and single-channel sound decomposition state-of-art

Several previous studies have assumed models similar to that in the source-filter model. Itakura attempted to solve the source-filter problem, in particular for speech signals, by identifying the synthesis filter first. They modeled the short-term speech signal as a stationary Gaussian process and estimated the filter by using maximum likelihood spectrum estimation [61]. Many of speech separation or recognition systems are based on this source-stationary assumption [62]. For example, in Weiss's work, as a prior knowledge, phonetical transitions are assumed to be based on Markov process. Then, the source-dependent characteristics on the eigenvoice speaker model is estimated via variational EM algorithm to separate speech mixtures [63]. The assumption of the stationary Gaussian process, however, ignored the time-varying characteristics of pitch and loudness. Because of this assumption, it was not sufficient to reproduce real sound-source or resonant properties.

Klapuri modeled both filters and sources by a linear combination of basis functions. A spectrum is divided into four parts in that model: time-invariant harmonics, time-invariant body response filter, time-variant loss filter representing the frequency-dependent decay and time-variant modeling error. The basis functions for the harmonics (source) were obtained by means of principal component analysis (PCA) in advance, while those for the body response filter and the loss filter were given as a set

of triangular band-pass filters distributed uniformly on the critical band scale with 50% overlapping. This model was successful in extracting the time-constant characteristics (assumed to be instrument characteristics) from both source signal and the synthesis filter [64]. Klapuri then used his source-filter model for instrument feature extraction by assuming that each instrument has its own source and filter characteristics. Although this model was simplified as a linear model for computational tractability, the real sound generation process should include high non-linearity.

Additionally, Kitahara solved the illposed problem by extracting harmonic structure with the model of feature weighting given note information [65]. Vincent modeled polyphonic sound signals as a summation of the power spectra of each note to track two melodies at once [66] and proposed a basis-gain model as a sound decomposition model [60]. This model assumes that the time-frequency spectrogram, a matrix $\mathbf{S}$, is generated with the factorization of a basis matrix $\mathbf{V}$ and a gain matrix $\mathbf{W}$, which is similar to the non-negative matrix factorization (NMF). Many of recent studies about estimation of the source and/or the filter are based on NMF, which assumes that the original time-series signals can be factorized into two matrices. Similar to the source-filter model, this factorization has the inherit illposedness since observation is only the original signals. Recent studies introduced the sparseness constraint [52], harmonics [57, 60] as well as the pre-learned and fixed dictionary, which holds the harmonic structure, for the model simplicity [58, 53].

We talk a little about the related topic, speaker identification. A speaker here corresponds to the instrument in the meaning of "source." Because the resonances vary a lot in speaker identification compared with the changes in instruments, this would be the hardest problem in four categories. As a speech model, a dynamical system is often employed. As an example of existing studies, Lee attempted to express the speech dynamics constructing a switching state-space (SSS) model. For the approximation of posterior distribution, a variational Expectation-Maximization (EM) algorithm [67] was used [68, 69]. In Deng's studies such as [70], the filter model is determined by a certain number of resonant frequencies and their bandwidths and is based on LPC. It allows the model to learn parameters of both a filter part (vocal tract resonances, VTR) and a source signal part (residual). However, it has still not been explored enough to apply to real speech because it requires many parameters.

2 Sound generative model and illposedness

2.1 Source-filter model

Many traditional approaches to the sound-source identification problem are based on the Fant's source-filter model [41, 1]. This model was originally proposed to represent the sound- and speech-production process. In this model, the power spectrum s is modeled as the combination of source signal generation G and a synthesis filter H, which modulates the generated source signal, shown in Figure 3.2.1. The examples of

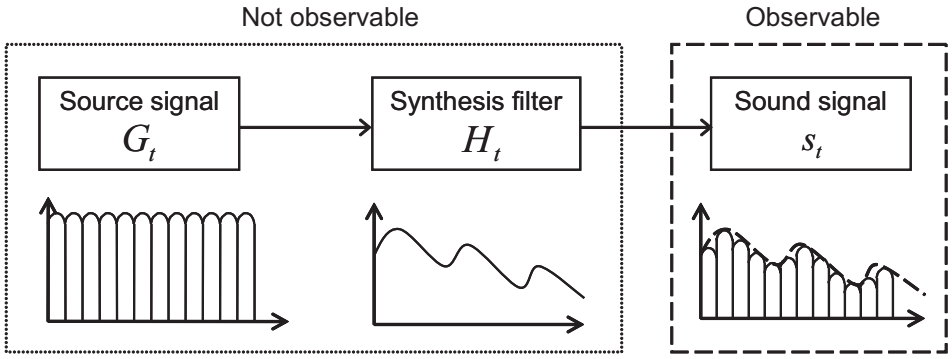

Figure 3.2.1. The source-filter model of sound production

a source signal and a synthesis filter are the vocal cords and the vocal tract for speech as well as pitch and instrument specific characteristics for instrument. The filter part in the source-filter model mainly contributes to the spectrum envelope, which forms a rough curve over the spectrum shown in the middle part of Figure 3.2.1. Since the power of the spectrum typically decreases from low frequencies to high frequencies, the spectrum envelope tends to have a gradual decrease in a frequency domain. It is often assumed that the synthesis filter contains information to specify the instrument [20, 71] or information on both voice characteristics and pronounced syllables in speech [72, 73]. On the other hand, the source signal mainly contributes to the fine structure of the spectrum. It forms a small jagged line in a frequency domain as in the left part of Figure 3.2.1. Those peaks are apt to be on the multiples of the fundamental frequency known as harmonics or overtones [74]. Since this model assumes the source and the filter are independent, it allows us to examine sound-source information

without the influence of other acoustical features such as pitch or loudness.

2.2 Illposedness

Although this model is widely used as a generative model to estimate sound characteristics such as instrument-specific features or pitch because of its simplicity, it has indeterminacy. In other words, there is more than one way to express the observed signal. [1] For instance, suppose the source-filter model is expressed in one-dimension, it is equivalent to a multiplication of two subjects, G and H. When we want to estimate the values of a source and a filter from the observed signal equals to six, there is an infinite number of combinations that can express that data, for example, one by six or two by three.

$$
\begin{aligned}
s &= G \times H \\
6 &= 1 \times 6 \\
6 &= 2 \times 3
\end{aligned}
\tag{3.2.1}
$$

The representation of the source-filter model has an inherent indeterminacy since it is impossible to estimate the source signal generation G_t and the synthesis filter H_t at each time t ($t \in \{1, \cdots, T\}$) without additional constraints.

2.3 Constraints

To relax this uncertainty, this study considers three kinds of constraints from the sound properties as well as the source-filter model: harmonicity, dynamics and sparsity.

- Harmonicity
  Several recent studies introduce the harmonicity assumption into the model [57, 76]. Most of instruments have harmonic overtones in their power spectrum, one example of a power spectrum is shown in Figure 3.2.2. From this sound property, in this study, harmonicity is assumed in the source-filter model in the

[1] Problems are called ill-posed when they do not satisfy the following three properties of a well-posed problem: 1. A solution exists. 2. The solution is unique. 3. The solution depends continuously on the data, in some reasonable topology.
This is introduced by J. Hadamard [75].

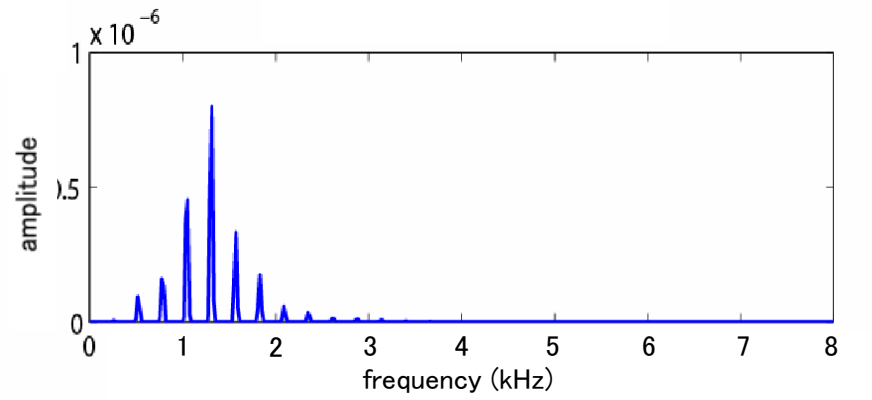

Figure 3.2.2. The example of harmonicity in the power spectrum

observation process of the probabilistic model which is explained Section IV.3.1. In this model, the phase information is not modeled from the fact that the human auditory system is insensible to the phase information [19].

- Continuity or discontinuity transitions of pitch and amplitude
  In the proposed model, the time-varying source and the time-invariant filter are assumed in the source-filter model. Smoothness of fundamental frequency is also considered in other existing studies such as [57, 76]; however, in this study, the case of discontinuous is also included in the state transition of the stochastic dynamical system model (Section IV.3.2).

- Sparsity
  Models based on the nonnegative matrix factorization (NMF) are often together with the sparsity constraint [53, 55, 77, 78, 79]. Since the proposed model is probabilistic, the sparse distribution is employed in the state transition part (Section IV.3.2).

3 Dynamical system model formulation

The joint distribution of the set of observed spectra $S_{1:T}$ and the set of hidden variables $X_{1:T}$ is

$$\begin{aligned} p\left(S_{1:T}, X_{1:T}\right) &= p(s_T, x_T|S_{1:T-1}, X_{1:T-1}, \theta)p(S_{1:T-1}, X_{1:T-1}, \theta) \\ &= p(s_T, x_T|S_{1:T-1}, X_{1:T-1}, \theta)p(s_{T-1}, x_{T-1}|S_{1:T-2}, \theta) \\ &= p(s_1, x_1|\theta)\prod_{t=2}^{T} p(s_t, x_t|S_{1:t-1}, X_{1:t-1}, \theta) \\ &= p(s_1|x_1, \theta)p(x_1|\theta)\prod_{t=2}^{T} p(s_t|x_t, \theta)p(x_t|x_{t-1}, \theta). \end{aligned} \tag{3.3.1}$$

Its graphical representation is shown in Figure 3.3.1. The joint distribution can be

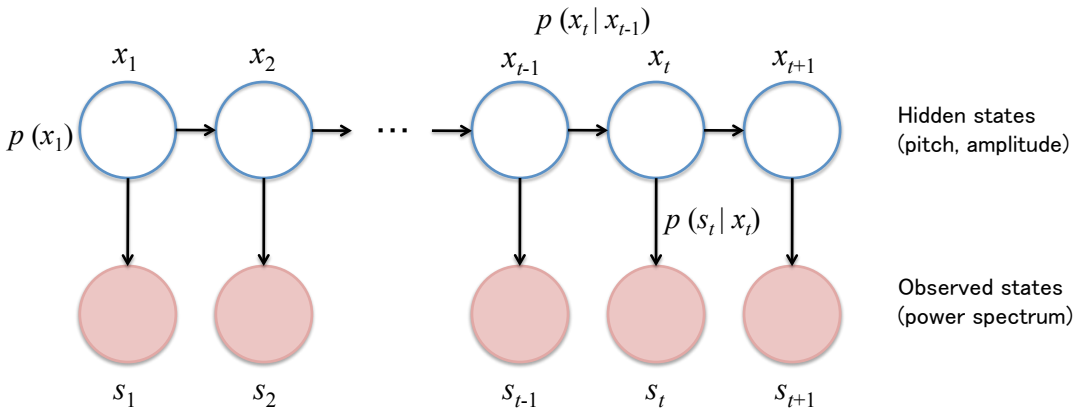

Figure 3.3.1. Graphical representation of dynamical system

computed given the observation process $p(s_t|x_t)$, the state transition $p(x_t|x_{t-1})$ and the initial state $p(s_1, x_1|\theta)$. The likelihood $p(S_{1:T})$ can be calculated by integrating this joint distribution with respect to $X_{1:T}$; however, in general, it cannot be performed analytically. The solution is provided in the Section IV.4. Followings are the explanation of distributions for the observation process, the state transition and the initial state.

3.1 Observation process

Acoustic distortion measures

There are several distortion measures to evaluate how much the spectrum generated from the model is close to the observed spectrum; the Euclidean distance and the Kullback-Leibler divergence are widely used. Recently, NMF with Itakura-Saito divergence is introduced [80], which successfully separate the three different instruments in Jazz music [81]. This measure is known to be close to perceptual measure of sounds since it emphasizes the distortion of the peaks of spectra more than that of the valleys [82]. In this study, the Itakura-Saito distortion measure is employed for the formalization of the observation process because the observation process describes how likely the observed spectrum is produced from the model spectrum by a random observation noise. Roughly speaking, it measures the divergence between observed spectrum and model spectrum. As a probabilistic model, Itakura-Saito divergence is considered to impose a Chi-square distribution for the observation noise.

Itakura-Saito distortion measures to Chi-square distribution

We consider the noise distribution in the frequency domain rather than in the time domain. Now, we show the equivalence between the Itakura-Saito distortion measure and the Chi-square distribution. When the observation spectrum is represented along the continuous frequency axis as in the Itakura-Saito distortion, the summation is replaced by the integral, i.e.,

$$d_{IS} = \mathbb{E}\left[\frac{\hat{s}_t(\tilde{\omega})}{s_t(\tilde{\omega})}\right] = 2\int_{-\pi}^{\pi}\left\{\log\frac{\hat{s}_t(\tilde{\omega})}{s_t(\tilde{\omega})} + \frac{s_t(\tilde{\omega})}{\hat{s}_t(\tilde{\omega})} - 1\right\}d\tilde{\omega} \tag{3.3.2}$$

where $\hat{s}(\tilde{\omega})$ is the estimated power spectrum, $s(\tilde{\omega})$ is the true short-term power spectrum, $\tilde{\omega} = \frac{\omega \mathrm{Fs}}{2\pi}$ is the normalized angular frequency ($-\pi \leq \tilde{\omega} \leq \pi$), and ω and Fs are the digitized frequency and the sampling frequency, respectively.

A Chi-square distribution (degree of freedom: 3) is given by

$$f(n) = \frac{1}{2\Gamma(1.5)}\left(\frac{n}{2}\right)^{\frac{1}{2}}\exp\left(-\frac{n}{2}\right). \tag{3.3.3}$$

By taking the logarithm of this equation, we have

$$\begin{aligned}\log f(n) &= -\log\Gamma(1.5) - \frac{1}{2}\log\frac{1}{n} - \frac{1}{2}\log 2 - \frac{n}{2} \\ &\equiv -\frac{1}{2}\left(\text{const.} + \log\frac{1}{n} + n\right).\end{aligned} \tag{3.3.4}$$

Substituting

$$\begin{aligned}s_t &= \hat{s}_t \odot n_o \\ n &= \frac{s(\tilde{\omega})}{\hat{s}(\tilde{\omega})}\end{aligned} \tag{3.3.5}$$

into 3.3.4, we obtain

$$\log\ f\left(\frac{s(\tilde{\omega})}{\hat{s}(\tilde{\omega})}\right) \equiv \log\frac{\hat{s}(\tilde{\omega})}{s(\tilde{\omega})} + \frac{s(\tilde{\omega})}{\hat{s}(\tilde{\omega})} + \text{const.} \tag{3.3.6}$$

In Equation 3.3.5, $\odot$ is the Hadamard product, the element-by-element product. Since we have assumed that the noise is generated independently from a Chi-square distribution for each frequency, the joint log-probability of the observation noise becomes a summation of 3.3.6 over frequencies, which is equivalent to the Itakura-Saito distortion 3.3.2.

The Chi-square distribution as a noise distribution

The probabilistic distribution for the observed spectrum s_t, given the amplitude a_t and the fundamental frequency f_t, $p(s_t|x_t, \theta)$, is defined with the Chi-square distribution with three degree of freedom as:

$$p(s_t|x_t,\theta) = \prod_{i=1}^{d} \frac{1}{2\Gamma(1.5)s_t(i)\sigma_o}\left(\frac{1}{2\sigma_o}\frac{s_t(i)}{\hat{s}_t(i)}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2\sigma_o}\frac{s_t(i)}{\hat{s}_t(i)}\right). \tag{3.3.7}$$

When s_t and $\hat{s}_t$ are close enough, this observation process can be approximated as

$$\approx \prod_{i=1}^{d} \frac{1}{2\Gamma(1.5)\hat{s}_t(i)\sigma_o}\left(\frac{1}{2\sigma_o}\frac{s_t(i)}{\hat{s}_t(i)}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2\sigma_o}\frac{s_t(i)}{\hat{s}_t(i)}\right). \tag{3.3.8}$$

Here, $\hat{s}_t$ is the estimated spectrum based on the above-mentioned source-filter model, represented by $\hat{s}_t = G_t H_t$.

Details of the source-filter model in an observation process

Suppose we have the observed d-dimensional short-term log-power spectrum s_t and the hidden variable x_t at time t. The set of time series spectra and the set of time series hidden variables are $S_{1:T} = \{s_1, \cdots, s_T\}$ and $X_{1:T} = \{x_1, \cdots, x_T\}$, respectively.

The observed log-power spectrum s_t is represented by two independent function as $\hat{s}_t = G_t(x_t; \theta) \odot H_t(\theta)$ with the source signal $G_t(x_t; \theta)$ and the synthesis filter $H_t(\theta)$. The hidden variable x_t is a vector whose elements are the hidden log-amplitude a_t and the frequency f_t written as $x_t = [a_t,\ f_t]^{\mathrm{T}}$, as well as θ here is a set of parameters for the model which determines the forms of the distributions or functions.

In this model, the time-invariant function for the synthesis filter is defined as

$$H(\tilde{\omega}) = 2^{1-p} \left\{ \sin^2 \frac{\tilde{\omega}}{2} \prod_{n=2,4,\cdots,p} (\cos\tilde{\omega} - \cos b_n)^2 \right\} + \left\{ \cos^2 \frac{\tilde{\omega}}{2} \prod_{n=1,3,\cdots,p-1} (\cos\tilde{\omega} - \cos b_n)^2 \right\}^{-2}, \tag{3.3.9}$$

which is exactly the same as the parameterization with LSF [83]. In this equation, b_n $(n = 1, \cdots, p)$ is a set of parameters, which determines the synthesis filter.

On the other hand, the source signal generation has time dependence, which is represented as the sum of Gaussians whose peaks are located at harmonic frequencies kf_t $(k = 1, \cdots, K)$,

$$G_t(\omega_i; a_t, f_t, K, \sigma_p, \tau, A) = \exp\left(a_t + A \exp\left(-\frac{\omega_i}{\tau}\right) \sum_{k=1}^{K} \mathcal{N}(\omega_i; kf_t, {\sigma_p}^2) \right). \tag{3.3.10}$$

Here, $-\left(\frac{\omega_i}{\tau}\right)$ represents the exponential decay along the frequency axis where ω_i is a digitized frequency, and A is an amplitude parameter. $\mathcal{N}(x; \mu, \sigma)$ denotes the Gaussian distribution of x with mean μ and variance σ. K and σ_p represent the number of Gaussians and the variance of each Gaussian in the synthesis filter, respectively. Details of each parameter are explained in Figure 3.3.2.

3.2 State transition of fundamental frequency and amplitude

The state transition includes the Markov assumption on dynamics. From the fact that most of the played notes are present for certain time as well as the current note does

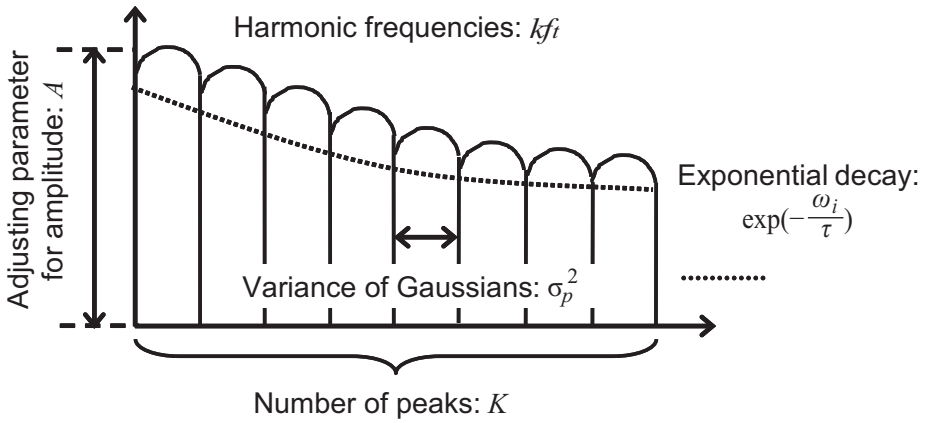

Figure 3.3.2. Correspondence between parameters of G_t and the form of the source spectrum

not depend on the previous note when the note shifts to the another note $\mathbf{x}_t$ is assumed to shift either continuously or discontinuously. With this assumption, we separately model two cases as the state transition with a scale mixture [84] of two distributions; one for a continuous term ($\eta = 1$), and the other is for a discontinuous term ($\eta = 0$):

$$p(x_t|x_{t-1}) = \bar{\eta}p(\mathbf{x}_t|\mathbf{x}_{t-1}, \eta = 1) + (1 - \bar{\eta})p(\mathbf{x}_t|\mathbf{x}_{t-1}, \eta = 0), \tag{3.3.11}$$

where $\bar{\eta}$ is the mixing rate, which is the probability to be continuous. In the experiment, η is set as 0.8 in advance. Details of the continuous and discontinuous terms are as follows.

- Continuous transition term, $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \eta = 1)$:
  This term represents the case that one note is played and kept for a while. The model assumes that the transition of the fundamental frequency and the log-amplitude occur independently. It is known that when one note is played, the value of the fundamental frequency does not change much while the amplitude decays exponentially as the sound progresses [85, 74], so the log-amplitude decays linearly. Considering such sound properties, the transitions of the fundamental frequency f_t and the log-amplitude a_t are defined as

$$f_t = f_{t-1} + n_f, \tag{3.3.12}$$

and

$$\frac{\exp(a_t)}{\exp(a_{t-1})} = \rho \exp(n_a)$$
$$a_t = a_{t-1} + \log \rho + n_a, \tag{3.3.13}$$

where ρ is an attenuation constant ranging from 0 to 1. In this equation, n_a and n_f are Gaussian noise with small variances, Σ_a for the log-amplitude and Σ_f for the fundamental frequency. Under these assumptions, the continuous transition can be written as

$$p(x_t|x_{t-1}, \eta = 1, \theta) = \bar{\eta} \left[\mathcal{N}\left(a_t; a_{t-1} + \log \rho, \Sigma_a\right) \mathcal{N}\left(f_t; f_{t-1}, \Sigma_f\right) \right], \tag{3.3.14}$$

where $\mathcal{N}(x; \mu, \sigma)$ is a Gaussian distribution whose mean and variance are μ and σ, respectively. The first Gaussian represents an exponential decay of the log-amplitude, and the second Gaussian represents the constancy of the fundamental frequency.

- Discontinuous transition term, $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \eta = 0)$:
  This term represents the change of the notes in music. Same as the model of the continuous transition, fundamental frequency and amplitude transitions are assumed to occur independently. Their transitions are approximated by two independent Gaussian distributions with large variances $\sigma^2 = [\sigma_a^2,\ \sigma_f^2]$, i.e.

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}, \eta = 0, \theta) = (1 - \bar{\eta}) \left[\mathcal{N}(a_t; m_a, \sigma_a^2) \mathcal{N}(f_t; m_f, \sigma_f^2) \right], \tag{3.3.15}$$

  where m_a and m_f are mean vectors of a_t and f_t, respectively.

  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. As a whole, the state transition is defined as

$$\begin{aligned} p(x_t|x_{t-1}, \theta) &= \bar{\eta}(\text{continuous}) + (1 - \bar{\eta})(\text{discontinuous}) \\ &= p(\mathbf{x}_t|\mathbf{x}_{t-1}, \theta, \eta = 1) + p(\mathbf{x}_t|\mathbf{x}_{t-1}, \theta, \eta = 0) \\ &= \bar{\eta} \left[\mathcal{N}(a_t; a_{t-1} + \log \rho, \Sigma_a) \mathcal{N}(f_t; f_{t-1}, \Sigma_f) \right] \\ &\quad + (1 - \bar{\eta}) \left[\mathcal{N}(a_t; m_a, \sigma_a^2) \mathcal{N}(f_t; m_f, \sigma_f^2) \right]. \end{aligned} \tag{3.3.16}$$

3.3 Initial distribution

The initial distribution of the hidden state is given by two independent Gaussian distributions as

$$p(x_1|\theta) = \mathcal{N}(a_1; m_a^1, (\sigma_a^1)^2)\mathcal{N}(f_1; m_f^1, (\sigma_f^1)^2), \tag{3.3.17}$$

4 Parameter estimation with free energy minimization

4.1 Maximum likelihood and free energy minimization

The maximum likelihood estimation is one of the methods to estimate parameters of probabilistic models. However, the likelihood cannot be computed when the model contains hidden variables. In those cases, the EM algorithm is often employed. The EM algorithm can increase the likelihood only if the posterior of hidden variables can be computed. It is not also possible to compute the posterior of hidden variables in the proposed method; therefore, instead, so-called free energy function is introduced. It is proved that maximization of likelihood can be reformulated as minimization of the free energy [86].

Given the trial distribution of hidden variables $X_{1:T}$ as $q(X_{1:T})$, the free energy is defined as

$$\mathcal{F}(q(X_{1:T}), \theta) = -\log p(S_{1:T}|\theta) + \mathrm{KL}\left[q(X_{1:T})||p(X_{1:T}|S_{1:T}, \theta)\right]. \tag{3.4.1}$$

Here, $\mathrm{KL}[q(\cdot)||p(\cdot)]$ is the KL divergence of two probability distributions: $p(\cdot)$ and $q(\cdot)$,

$$\mathrm{KL}[q||p] = \int q(x) \log \frac{q(x)}{p(x)} dx. \tag{3.4.2}$$

The KL divergence takes the value of zero only when the two probability distributions are equal and otherwise takes the positive values. From the nonnegativity property of the KL divergence, the free energy takes the minimum value when $p(\cdot) = q(\cdot)$. Thus, minimization of the free energy with respect to the trial distribution $q(X_{1:T})$ is equivalent to the maximization of the log-likelihood with respect to the set of parameters θ.

4.2 Free energy minimization and variational EM algorithm

Minimization of the free energy with respect to the trial distribution $q(X_{1:T})$ corresponds to the E-step of the EM algorithm while that with respect to the model parameters θ corresponds to the M-step. In practice, since it is difficult to minimize the free energy with respect to the trial distribution, we minimize the free energy with restricting of the trial distribution to the certain distribution family.

Alternate minimization of the free energy with respect to the trial distribution and the model parameters is equivalent to the variational EM algorithm with the following update rules:

$$\text{E-step: } q^k = \underset{q}{\operatorname{argmin}}\, \mathcal{F}(q, \theta^{k-1}) \tag{3.4.3}$$

$$\text{M-step: } \theta^k = \underset{\theta}{\operatorname{argmin}}\, \mathcal{F}(q^k, \theta) \tag{3.4.4}$$

Here, k is the index for the update. The variational EM algorithm minimizes the free energy based on the coordinate-descent algorithm, which updates the set of parameters of either q or θ with fixing the other one. The problem in this algorithm is that the speed of learning parameters becomes slow when two sets of parameters to be learned are highly correlated. Thus, we parameterize the trial distribution with κ to allow the simultaneous update of parameters of the model and the trial distribution. That is $\min_{\theta,\kappa} \mathcal{F}\left(q(X_{1:T}|\kappa), \theta\right)$. Total parameters to be learned and the initial values (set manually in advance) are as follows. The trial distribution q is modeled with a single Gaussian with mean $\mu_{1:T}$ and variance $S_{1:T}$, which is

$$q(X_{1:T}) = \mathcal{N}\left(X_{1:T}; \mu, S\right), \tag{3.4.5}$$

where $\kappa = \{\mu, S\}$. The covariance matrix $S_{1:T}$ has the values only on $\mathbb{E}[a_{t-1}, a_t]$ and $\mathbb{E}[f_{t-1}, f_t]$, which are adjacent to the correlation of the fundamental frequency and the amplitude at each time, $\mathbb{E}[{a_t}^2]$, $\mathbb{E}[{f_t}^2]$. Other values are set to be zero. $\mathbb{E}[\cdot]$ is the expectation of the trial distribution, $q(X_{1:T})$.

All the parameter set θ of this sound generative model is

$$\theta = \left\{\bar{\eta}, \sigma_{\{a,f\}}, m_{\{a,f\}}, \Sigma_{\{a,f\}}, m^1_{\{a,f\}}, \Sigma^1_{\{a,f\}}, A, \tau, \sigma_p, K, b_1, \cdots, b_p\right\}. \tag{3.4.6}$$

In the experiment, parameters, $\sigma_{\{a,f\}}$, $m_{\{a,f\}}$, $\Sigma_{\{a,f\}}$, $m^1_{\{a,f\}}$, $\Sigma^1_{\{a,f\}}$ are set manually. In addition, the probability to take continuous transition η is set as 0.8. In summary,

sets of parameters to be estimated are

$$\theta_{\text{est}} = \{b_1, \cdots, b_p, K, \sigma_p, A, \tau\} \tag{3.4.7}$$

$$\kappa = \{\mu, S\}. \tag{3.4.8}$$

The initial values of coefficients for spectrum envelope, $b_1, \cdots, b_p$, were determined by LSF, which we had assumed as the synthesis filter by Equation 2.2.4 which the number of LSF coefficients being $p = 12$. Other parameters, K, σ_p, A and τ were chosen to reproduce well the log-power sectra of instruments we examined.

4.3 Free energy revisited

Calculation of the free energy is as follows.

$$\begin{aligned}
&\mathcal{F}(q(X_{1:T}|\kappa), \theta) \\
&= -\int\cdots\int q(X_{1:T}|\kappa) \log p(S_{1:T}, X_{1:T}|\theta) dX_{1:T} \\
&\qquad + \int\cdots\int q(X_{1:T}|\kappa) \log q(X_{1:T}|\kappa) dX_{1:T}
\end{aligned} \tag{3.4.9}$$

We substitute the entropy

$$\mathcal{H}(p(x)) = -\int p(x) \log p(x) dx, \tag{3.4.10}$$

into the second term of the free energy 3.4.9; then, the second term can be written as

$$\int\cdots\int q(X_{1:T}|\kappa) \log q(X_{1:T}|\kappa) dX_{1:T} = -\mathcal{H}(q(X_{1:T}|\kappa)), \tag{3.4.11}$$

and replacing the second term of the free energy resulted in

$$\mathcal{F}(q(X_{1:T}|\kappa), \theta) = -\int\cdots\int q(X_{1:T}|\kappa) \log p(S_{1:T}, X_{1:T}|\theta) dX_{1:T} - \mathcal{H}\left(q(X_{1:T}|\kappa)\right). \tag{3.4.12}$$

Substitute the joint distribution $p(S_{1:T}, X_{1:T}|\theta)$, Equation 3.3.1, into this free energy, we obtain

$$
\begin{aligned}
&\mathcal{F}(q(X_{1:T}|\kappa), \theta) \\
&\quad = -\int\cdots\int q(X_{1:T}|\kappa) \log \left(p(s_1|x_1,\theta)p(x_1|\theta) \prod_{t=2}^{T} p(s_t|x_t,\theta)p(x_t|x_{t-1},\theta) \right) dX_{1:T} \\
&\qquad - \mathcal{H}\left(q\left(X_{1:T}|\kappa\right)\right) \\
&\quad = -\int q(x_1|\kappa) \log p(s_1|x_1,\theta) dx_1 - \int q(x_1|\kappa) \log p(x_1|\theta) dx_1 \\
&\qquad - \int q(x_t|\kappa) \log \sum_{t=2}^{T} p(s_t|x_t,\theta) dx_t \\
&\qquad - \iint q(x_t, x_{t-1}|\kappa) \log \sum_{t=2}^{T} p(x_t|x_{t-1},\theta) dx_t \, dx_{t-1} - \mathcal{H}\left(q\left(X_{1:T}|\kappa\right)\right) \\
&\quad = -\int q(x_1|\kappa) \log p(x_1|\theta) dx_1 - \sum_{t=1}^{T} \int q(x_t|\kappa) \log p(s_t|x_t,\theta) dx_t \\
&\qquad - \sum_{t=2}^{T} \iint q(x_t, x_{t-1}|\kappa) \log p(x_t|x_{t-1},\theta) dx_t \, dx_{t-1} - \mathcal{H}\left(q\left(X_{1:T}|\kappa\right)\right).
\end{aligned}
\tag{3.4.13}
$$

5 Downhill simplex method approximation

The Nelder-Mead's downhill simplex method or sometimes known as an amoeba method [87, 88] is employed for the minimization of the approximated free energy. [2] This is an unconstraint non-linear optimization algorithm that minimizes an objective function. Compared with other non-linear optimization methods that use gradients such as the conjugate gradient method, it does not require any derivatives of the objective function but only the objective function evaluations. This algorithm is adopted to eliminate the difficulty in the differentiation of this proposed particular free energy function.

A simplex in this approximation is a polytope of $N + 1$ affine independent ver-

[2] This method is conceptually similar to GA. Both of them utilize only a value of the cost function, keep several candidates for the variables to be optimized and select the new candidate based on the present candidates [89].

tices in the N-dimensional parameter space. In two and three dimensions, the figures of a simplex are a triangle and a tetrahedron, respectively. For example, in a three-dimensional space, the downhill simplex method first evaluates the objective function on four vertices. To search and to converge to the point that gives a smaller value of the objective function than a current value, this method takes four operations: reflection, expansion, contraction and shrinkage [90]. Figure 3.5.1 shows the operations of the Nelder-Mead's downhill simplex method in a two-dimensional parameter space. We

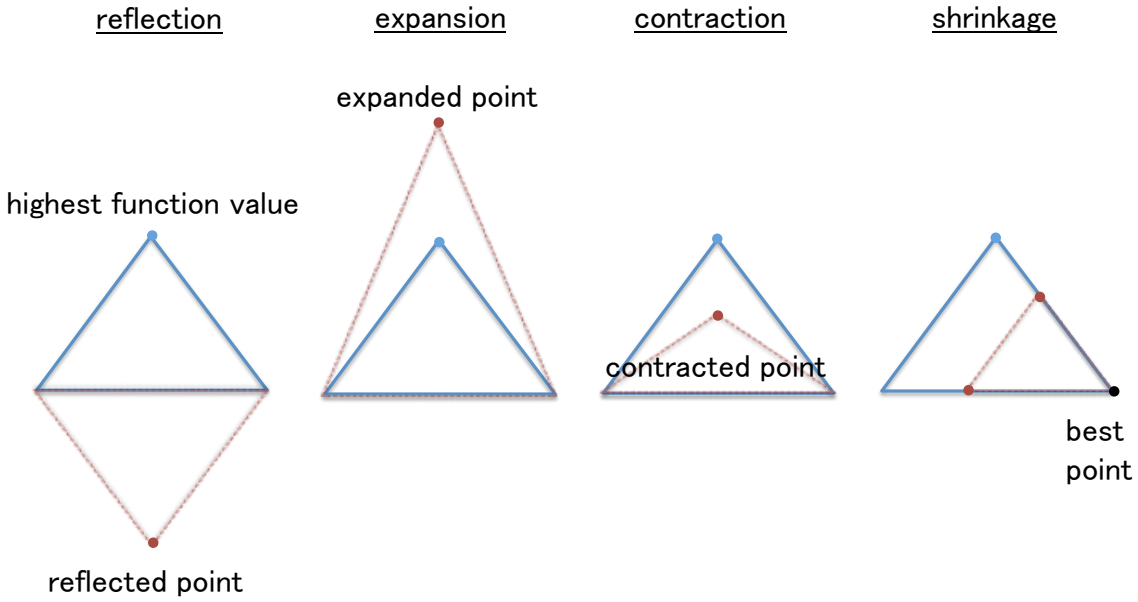

Figure 3.5.1. The four operations of the Nelder-Mead's downhill simplex method

take three vertices in two-dimensional space. The function values are evaluated in each vertex. The worst point, which has the highest function value among three, is replaced by a point with a lower value of the cost function with the Nelder-Mead's downhill simplex method.

If the first operation, reflection, finds the lowest value among three vertices, we go on to the second operation, expansion; otherwise, we skip that. We keep the new value searched by an expansion operation when the expansion finds the best; we keep the one searched by a reflection operation when the value of the expansion is still worst. If the reflection operation fails to find the best value, the algorithm tries a contraction

operation. If all operations fail, the simplex shrinks around the best point among three vertices. With iterations of those four operations, the Nelder-Mead's downhill simplex method searches the optimal point.

6 Experimental evaluation

To verify whether the proposed method can achieve simultaneous estimation of source-signal generation G_t and synthesis filter H, as well as find sound source-specific filter parameters, i.e., LSFs, we examined whether the spectrum envelopes represented by the synthesis filter H show similar patterns (that is, consistency) when the sound source (instrument) is the same. Moreover, we utilized the LSFs, which are estimated by our method to identify the instrument.

6.1 Sound data

In monophonic music instrument identification, it is difficult to make an objective evaluation since there is no public database. As is done in the most of the existing studies of music instrument identification, we then chose training and test samples from several commercial CDs whose sampling frequency was $44.1kHz$. We prepared samples of five instruments: viola, flute, horn, trumpet and cello. Samples that were silent or too low in amplitude were removed, but no restriction was applied to the fundamental frequency.

6.2 Spectrum analysis

Each of the training and test data sets was constructed by extracting 30 excerpts of one second per sample from each recording, totaling 150 data samples. Each sample for one second was divided into twenty frames, and in each frame the power spectrum was computed by discrete Fourier transform with a sampling of 2048 points per frame. In general, the low frequency range, more than the high range, is considered to include the timbre information, so the frequency range was set as 1-11020Hz. By smoothing the spectrum over every 10Hz, the dimensionality of spectrum information can be reduced from 11,020 to 1,102. To use a log-amplitude as spectrum information, logarithmic transformation was applied to the 1,102-dimensional vector. According to the LSF

method, the spectrum vector was compressed into 12 dimensions, which we assumed as instrument features. Determination of the dimensionality of instrument features has trade-offs of complexity and accuracy of instrument identification. An existing study of speaker identification argued that ten-dimensional LPC coefficients were able to represent well the standard speech waves whose sampling frequency was $8kHz$ [91]. According to that study, we set the number of instrument features at ten.

6.3 Feature extraction

As mentioned in Chapter II Section 2, there are many kinds of feature extraction methods. Typical instrument features are either temporal or spectral features or both, which especially consider information on pitch variation [46, 92, 45]. Especially, the set of MFCCs is one of the simple and efficient feature extraction methods in existing studies [14, 18, 47, 93]. Here, we employ the set of LSFs from its high ability of instrument identification in [17].

6.4 Experiment 1: The source signal and the synthesis filter estimation

Since we have assumed that the filter constitutes a temporally invariant part of the instrument characteristics, it is expected that the spectrum envelope is consistent for each instrument. Typical forms of spectrum envelope of each instrument are as follows:

viola	one gradual peak around 1,000 Hz
flute	a sharp peak around 500-1,000 Hz
horn	two sharp peaks around 500 and 1000-1,500 Hz
trumpet	one or two gradual peaks around 1,500-2,000 Hz
cello	a very smooth peak around 300-800 Hz

which are shown in Figure 3.6.1. The downhill simplex method was used to find the optimal parameters by evaluating the free energy 1,000 times for the parameters A and τ, followed by evaluating it 1,000 times for a set of parameters in the synthesis filter, b_n ($n = 1, \cdots, K$). The initial values of b_n were set to the LSF coefficients, which were obtained by the existing Levinson-Durbin algorithm [24, 25]. The initial parameters for A and τ were set manually for each instrument.

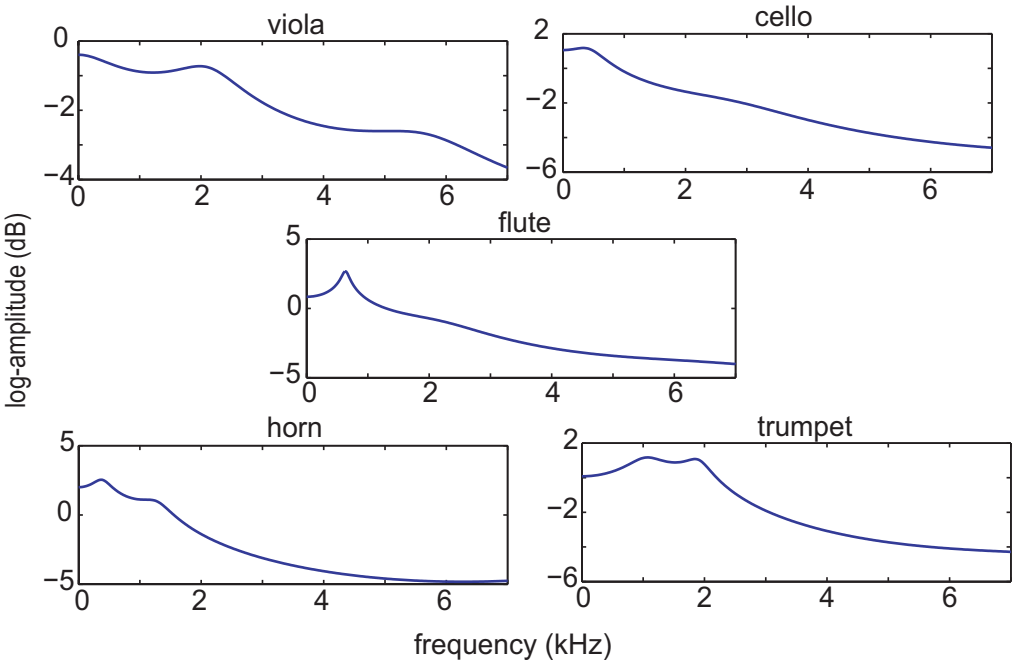

Figure 3.6.1. Forms of spectrum envelope of five instruments (x-axis: frequency, y-axis: log-amplitude)

Figure 3.6.2.A shows the original spectrum (red, lower) and the reconstructed spectrum from the estimated source and filter with initial parameters (blue, upper). After the parameter learning, the model spectrum is closer to the original spectrum than that with the initial parameters shown in Figure 3.6.2.B.

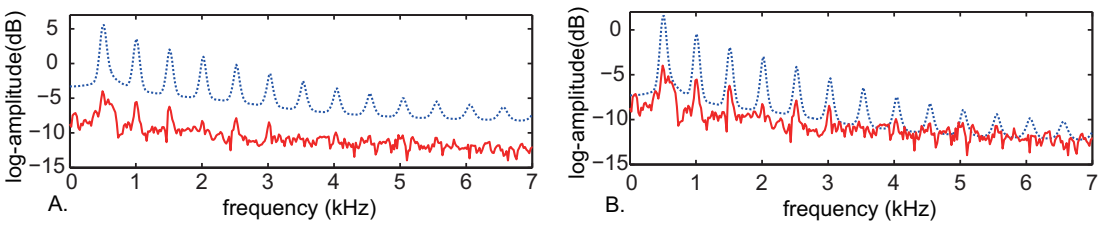

Figure 3.6.2. Original log-amplitude spectrum and estimated model spectrum with initial parameters (A) and with learned parameters (B)

Figure 3.6.3 shows spectrum envelopes of the trumpet with initial (LSF) parameters (A) and with learned parameters (B) at randomly-chosen times t_1 (thick) and t_2 (thin). Comparing the spectrum envelopes at t_1 and t_2, the spectrum envelopes reproduced by

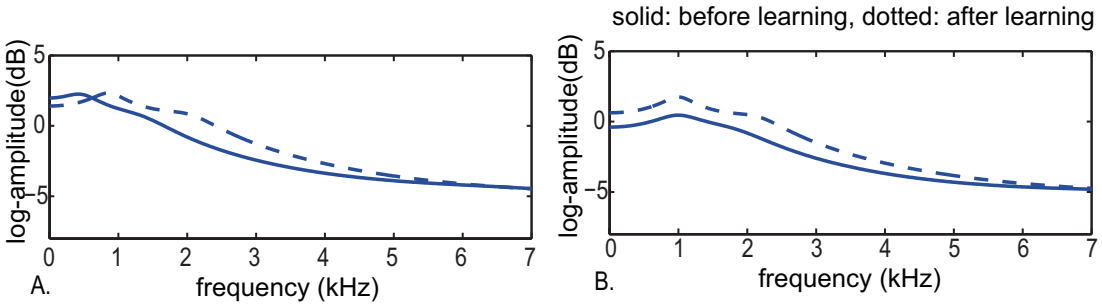

Figure 3.6.3. Example spectrum envelopes of the trumpet, reproduced by the model with the parameter before learning (LSF coefficients; A) and that after learning (B) at randomly-chosen times t_1 (thick) and t_2 (thin).

the model with the parameters after learning (Figure 3.6.3.B) are closer to each other than those with the initial parameters.

The first (lowest-frequency) peaks of spectrum envelopes at t_1 and t_2 shifted from 500 and 900 respectively to around 1,000Hz for both after parameter learning. Figure

3.6.4 illustrates the source signal and the spectrum envelope estimated by the model at t_1 and t_2 with the parameters before and after learning. Both initial and learned

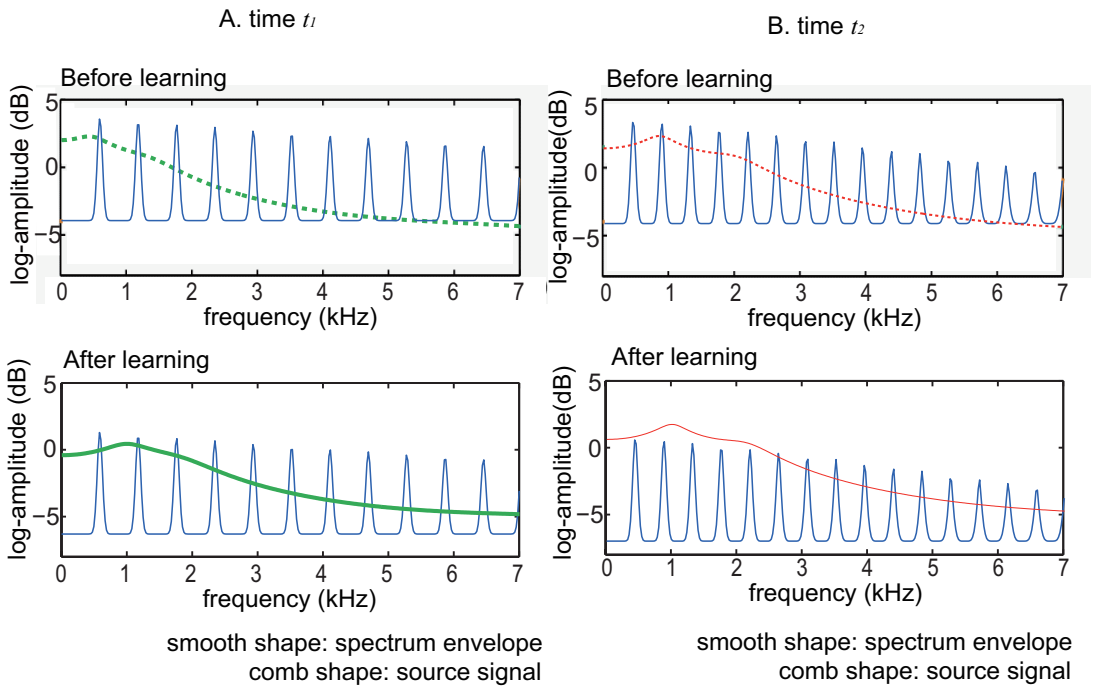

Figure 3.6.4. Model source signal (thin) and spectrum envelope (thick) before learning at time t_1 (A) and before learning at time t_2 (B)

spectrum envelopes at t_1 and t_2 in Figure 3.6.3 are displayed together in Figure 3.6.5, enabling clear comparison of envelopes.

The top two panels of Figure 3.6.4 show the source signal and the spectrum envelope estimation of at times t_1 and t_2 before learning, whose spectrum envelops are the same as in Figure 3.6.3.A. Similarly, the spectrum envelopes in the bottom panels of Figure 3.6.4 correspond to those in Figure 3.6.3.B. Before learning, the first peaks of the spectrum envelopes match one of the harmonic frequencies. Because the harmonic frequencies vary with time, the frequencies at which the spectrum envelopes have their first peaks differ between times t_1 and t_2, and consequently, we cannot obtain similar forms of spectrum envelope from the same instrument (spectra at times t_1 and t_2 are from the same instrument) with LSF. Even in such a situation, our dynamical system-based source-filter model has been successful in making the estimation consistent over

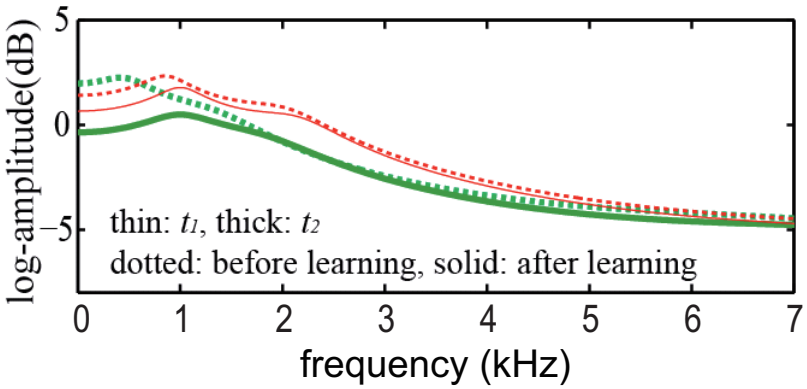

Figure 3.6.5. Spectrum envelopes before (dotted) and after (solid) learning at t_1 (thin) and t_2 (thick)

time, because the parameter learning has fully utilized the consistency constraint of the source filter of a single instrument.

6.5 Experiment 2: Parameter reduction and visualization with LFDA

Considering the trade-offs between the number of parameters that represent the instrument characters and accuracy in instrument identification based on the parameters, we examined if it is possible to reduce the number of feature parameters under maintaining the accuracy. LDA and PCA are classical parameter reduction methods as used in [12, 13, 65]. They are supervised and unsupervised dimension reduction methods based on linear transformation, respectively. Since instrument identification is basically a classification task, LDA may be a more appropriate dimension reduction method than PCA. However, as mentioned in Section 2, LDA sometimes underperforms PCA when the number of features is smaller than the number of classes or when data distribution has multiple modes [34, 37].

Figure 3.6.6 shows the LSFs obtained by our method, reduced from 12 to three dimensions. Note that the scale of the y-axis is different for the training and test data, for visibility. We can see that distributions of individual instruments are consistent when comparing the training and test data sets. Especially, flute (x) and horn (+) are clearly separated from the other three instruments even in this reduced three-dimensional space. Although this is a simple visualization task, the fact that the distri-

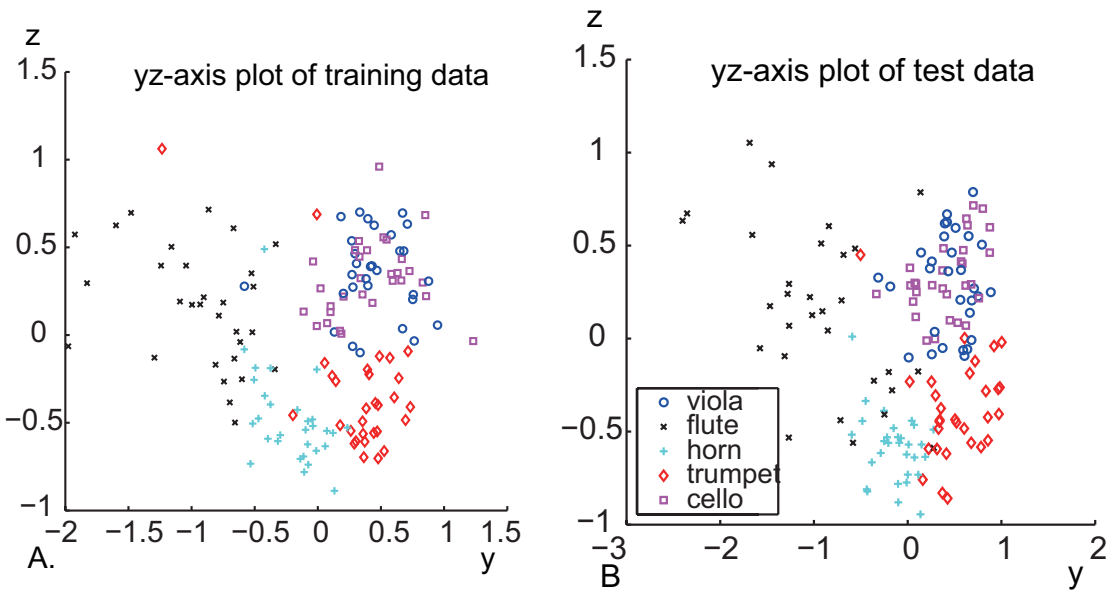

Figure 3.6.6. Three-dimensional projection by LFDA from the 12-dimensional LSFs obtained by our method. For visibility, the projection onto the yz-axis is shown for the training data (A) and the test data (B)

butions of different instrument are characteristic even in the reduced three-dimensional space is encouraging regarding the applicability of the returned features to instrument identification task.

6.6 Experiment 3: Verification of Instrument Features with Instrument Identification

Twelve LSFs were used as the features for instrument identification. The number was determined so that the computational cost would not be expensive in the real sound analysis. Table 3.6.1 summarizes the results of the SVM classification by using LSFs obtained by our method and those of the former instrument identification experiments. Note that these experiments do not share the same data sets, which limits the significance of this comparison. This table suggests that the employment of the LSFs achieved similar or higher identification performance although the dimensionality of

	Number of instruments	Number of features	Accuracy (%)
Initial parameters	5	12	84.67
(6 dim.)	5	6	83.33
(3 dim.)	5	3	71.33
(2 dim.)	5	2	58.67
Learned parameters	5	12	87.33
(6 dim.)	5	6	82.67
(3 dim.)	5	3	74
(2 dim.)	5	2	52
Marques, 1999 [14]	8	16	70
Eggink, 2003 [46]	5	120	66
Livshin, 2004 [12]	7	62	88
Jinachitra, 2004 [47]	5	28	66
Essid, 2004 [18]	5	10	67
Essid, 2006 [13]	5	70	87

Table 3.6.1. Classification results of original and reduced feature space with initial and learning parameters, comparing with the other instrument identification experiments

the LSFs used in our study was much smaller than those in the other experiments. The training of the parameters is additionally effective for identification. Especially when the number of features is reduced to as few as three, our model parameter estimation is effective in preventing the performance from degrading.

Table 3.6.2 and Table 3.6.3 show the confusion matrices for the initial and learned LSF parameters. The numbers shows the classification results when the number of

	Viola	Flute	Horn	Trumpet	Cello
Viola	24 (0)	0 (1)	2 (3)	3 (14)	1 (12)
Flute	0 (0)	19 (22)	7 (6)	0 (1)	1 (1)
Horn	0 (0)	0 (0)	30 (25)	0 (5)	0 (0)
Trumpet	0 (0)	0 (1)	2 (1)	28 (28)	0 (0)
Cello	3 (0)	0 (0)	1 (6)	0 (11)	26 (13)

Table 3.6.2. Monophonic music confusion matrix for five instruments with initial parameters

	Viola	Flute	Horn	Trumpet	Cello
Viola	25 (26)	0 (1)	1 (2)	2 (1)	2 (0)
Flute	2 (1)	22 (23)	5 (5)	0 (1)	1 (0)
Horn	0 (0)	0 (0)	30 (27)	0 (3)	0 (0)
Trumpet	0 (24)	0 (1)	1 (1)	29 (2)	0 (2)
Cello	5 (20)	0 (0)	0 (5)	0 (5)	25 (0)

Table 3.6.3. Monophonic music confusion matrix for five instruments with learned parameters

LSF coefficients is 12, and numbers inside parentheses are those when the number of LSF coefficients is reduced to 2. These confusion matrices show that the error occurs mostly between string instruments, viola and cello, which have similar resonance structures.

7 Discussion

7.1 Summary

In this study, we proposed a system identification approach to the simultaneous estimation of the source signal generation and the synthesis filter, based on an additional assumption of temporal continuity of pitch and loudness. The probabilistic model was constructed so as to represent the continuity dynamics, and the parameters were estimated by the minimization of the free energy. The synthesis filter was initially parameterized by LSFs, and was further modified so as to minimize the free energy. After the learning of the model parameters, instrument identification was carried out by using the optimized model parameters.

Although we also found that the initial model parameters could provide enough information for instrument identification with a small number of parameters, in contrast to the existing methods, the optimized parameters showed further effectiveness for instrument identification. In addition, the accuracy was not degraded so much even when parameter dimensionality was significantly reduced by applying LFDA to the LSF parameters.

For practical use of this model for general sound identification, there are at least two problems. First, generalization to polyphonic music or speech signals is not straightforward; it is necessary to modify the observation process to deal with polyphonic music. The second problem is computational complexity. Although the simplex method currently used for parameter estimation does not require explicit gradient of the objective function, it takes much computation time to converge. To introduce some non-linear optimization methods based on gradient information is important for further applicability of our method.

7.2 Issues

The log-amplitude spectrum and model spectrum of one of the incorrectly classified samples in both the initial condition and after learning are shown in Figure 3.7.1. The Figure 3.7.1.A shows the model synthesis filter and the Figure 3.7.1.B represents the estimated source signal from the original sound. In Figure 3.7.1.C, the model spectrum (blue, dotted), which combined the first two figures and the original spectrum (red,

solid), is drawn. Compared with correctly classified samples (See Figure 3.6.2 for

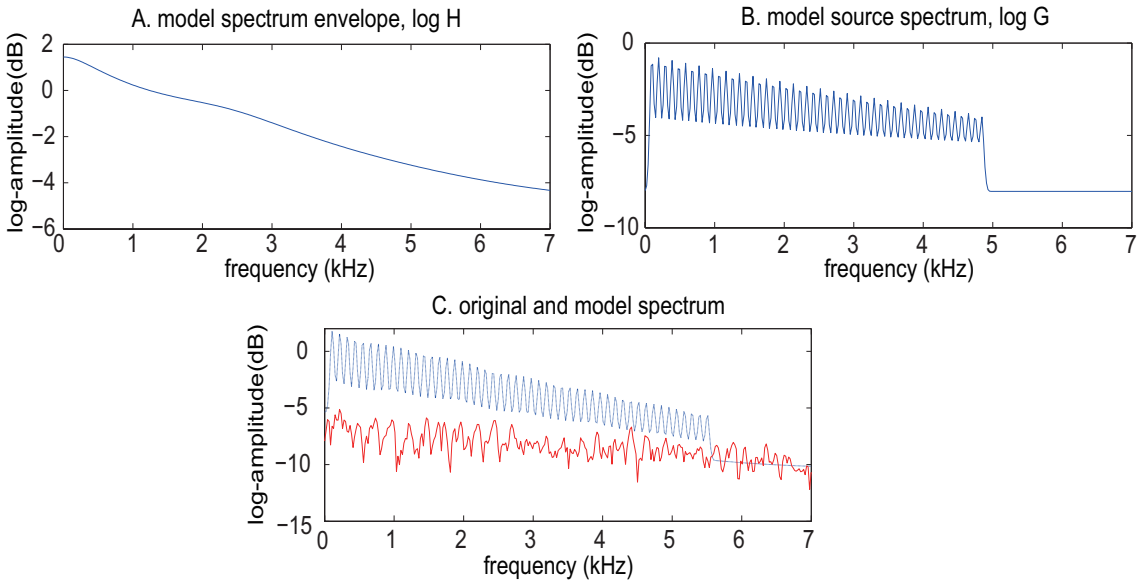

Figure 3.7.1. One example of source signal (A), spectrum envelope (B) and estimated spectrum (C) of wrong-classified sound (cello)

the typical original and estimated spectrum of correctly classified samples), most of misclassification tends to occur at the time that the whole spectrum power is low, and it is difficult to estimate the fundamental frequency from this power spectrum since the harmonics tend to collapse.

CHAPTER IV

Conclusion

1 Summary and contributions

In this dissertation, two studies which contribute to the music information retrieval and the sound source decomposition are presented.

The former study proposes the instrument feature representation with the linear projection parameters. A combination of two linear projection methods, PCA and LFDA, is employed to extract the sets of instrument features. The efficiency is evaluated with identification of instruments, which results in the high accuracy rates than those of existing studies. Additionally, from the visualization of the parameters of proposed method, PCA-LFDA, it is found that the same structured instruments or ones that are in the same instrument categories are projected closely in the PCA-LFDA space.

The latter study introduces a system identification approach to the estimation of source signal generation and the synthesis filter for the resonant property. The model takes into account the temporal continuity of pitch and intensity. The probabilistic model is constructed for the dynamical model, and the parameters are estimated by minimization of the free energy. The synthesis filter is parameterized by LSF. They are further modified so that the free energy is minimized. After the learning of model parameters, samples are classified using the trained model parameters. It is found that the initial model parameters could provide enough information for instrument iden-

tification with a smaller number of parameters compared with existing methods, and trained parameters show a little improvement for instrument identification. In addition, further reduction of the parameters could be achieved using LFDA without degrading the accuracy of the identification much.

2 Issues and future development

One possible extension of both instrument identification and the source-filter model estimation is applying it to instrument identification on polyphonic music. For ensemble music, Vincent's research on non-linear Independent Subspace Analysis (ISA), in which short-term-time log-power spectra of more than one instrument can be expressed as a sum of each instruments' spectrum [66], can be applied to the current model, and this would result in higher estimation accuracy of musical sound-sources.

Additionally, this research can be extended to the identification of speaker's characteristics. As with [68, 94], many researchers are currently attempting to apply the dynamical system to the speech, and research on speaker identification would be more important than that of instruments since they have many applications. The reason why speech is gaining attention as a target is that it has many applications. In order to identify a speaker's individuality with the proposed model, other kinds of constraints should be assumed since the important elements of instrument sound and speech are different from each other.

With the source-filter model, time-variation of the filter H should be considered. Since the transitions of H can be expressed by the smooth articulation of vowels, it can be modeled using the proposed dynamical system. In detail, in the case of the Japanese language, there are five vowels. Among five filters, we find the closest vowel structure to the original speech using structural matching introduced in [95]. Similarly, we assume some consonant structures to the source signal generation component. With these assumptions, we will try to estimate source and filter simultaneously.

References

[1] A.S. Bregman, Ed., *Auditory Scene Analysis: the perceptual organization of sound*, The MIT Press, Hoboken, New Jersey, Sept. 1994.

[2] R. Plomp, *Timbre as A Multidimensional Attribute of Complex Tones, Frequency Analysis and Periodicity Detection in Hearing*, Sijthoff, Leiden, 1970.

[3] S. Namba, "Definition of timbre," *Journal of Acoustical Society of Japan*, vol. 49(11), pp. 823–831, 1993, (in Japanese).

[4] O. Turk, "New methods for voice conversion," Master's thesis, Bogazici University, 2003.

[5] L.B. Almeida and J. M. Tribolet, "Nonstationary spectral modeling of voiced speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. ASSP-31, no. 3, pp. 664–678, June 1983.

[6] M. Goto and T. Goto, "Musicream: new music playback interface for streaming sticking, sorting, and recalling musical pieces," in *Proc. of European Conference on Signal Processing (EUSIPCO)*, 2005, pp. 404–411.

[7] M. Goto, "Development of the RWC music database," in *Proc. of International Congress on Acoustics (ICA)*, Apr. 2004, pp. I–553–556.

[8] The University of Iowa, "Electronic music studios: Musical instrument samples," http://theremin.music.uiowa.edu/MIS.html.

[9] McGill University, "Master samples," http://www.music.mcgill.ca/resources/mums/html/mums.html.

[10] IRCAM, "Studio On Line," http://www.ircam.fr/.

[11] K.D. Martin, *Sound-source recognition: A theory and computational model*, Ph.D. thesis, Massachusetts Institute of Technology, June 1999.

[12] A. Livshin and X. Rodet, "Musical instrument identification in continuous recordings," in *Proc. of International Conference on Digital Audio Effects (DAFx)*, Oct. 2004.

[13] S. Essid, G. Richard, and B. David, "Musical instrument recognition by pairwise classification strategies," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1401–1412, July 2006.

[14] J. Marques and P.J. Moreno, "A study of musical instrument classification using Gaussian mixture models and support vector machines," Tech. Rep., Compaq Computer Corporation, June 1999.

[15] G. Agostini, M. Longari, and E. Pollastri, "Musical instrument timbres classification with spectral features," in *Proc. of European Conference on Signal Processing (EUSIPCO)*, 2003, vol. 1, pp. 5–14.

[16] R. Ventura-Miravet, F. Murtagh, and J. Ming, "Pattern recognition of musical instruments using hidden markov models," in *Stockholm Music Acoustics Conference (SMAC)*, Aug. 2003, pp. 667–670.

[17] N. Chétry and M. Sandler, "Linear predictive models for musical instrument identification," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2006, vol. 5, pp. 5083–5086.

[18] S. Essid, G. Richard, and B. David, "Musical instrument recogntion on solo performance," in *Proc. of European Conference on Signal Processing (EUSIPCO)*, Vienna, Austria, Sept. 2004.

[19] H.K.F von Helmholtz, *On the Sensation of Tone - As a physiological basis for the theory of music*, Dover, New York, 2nd English edition, 1954, originally written in 1877. Translated by A.J. Ellis from 4th German edition.

[20] J.R. Miller and E.C. Carterette, "Perceptual evaluation of synthesized musical instrument tones," *Journal of Acoustical Society of America*, vol. 58, no. 3, pp. 711–720, Sept. 1975.

[21] K. Itoh and S. Saito, "Effects of acoustical feature parameters of speech on perceptual identification of speaker," *Transaction of IEICE of Japan*, vol. J65-A, no. 1, pp. 101–108, Jan. 1982, (in Japanese).

[22] L. Deng, *Speech Processing: A Dynamics and Optimization-Oriented Approach*, Marcel Dekker, Inc., New York, 2003.

[23] J.J. Burred, A. Robel, and X. Rodet, "An accurate timbre model for musical instruments and its application to classification," in *Proc. of Workshop on Learning the Semantics of Audio Signals (LSAS)*, 2006.

[24] F. Itakura and S. Saito, "A statistical method for estimation of speech spectral density and formant frequencies," *Transaction of IEICE of Japan*, vol. 53-A, pp. 36–43, 1970, (in Japanese).

[25] B.S. Atal and S.L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *Journal of Acoustical Society of America*, vol. 50, no. 2, pp. 637–655, Apr. 1971.

[26] F.K. Soong and B. Juang, "Line spectrum pair (LSP) and speech data compression," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Mar. 1984, vol. 9, pp. 37–40.

[27] F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signals," *Journal of Acoustical Society of Japan*, vol. 57, pp. S35, Apr. 1975, (in Japanese).

[28] N. Sugamura and F. Itakura, "Line spectrum representation of linear predictor coefficients of speech signal and its statistical properties," *Transaction of IEICE of Japan*, vol. J64-A, no. 4, pp. 323–330, Apr. 1981, (in Japanese).

[29] A.G. Krishna and T.V. Sreenivas, "Music instrument recognition: From isolated notes to solo phrases," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2004, vol. 4, pp. 265–268.

[30] G. De Poli and P. Prandoni, "Sonological models for timbre characterization," *Journal of New Music Research*, vol. 26, no. 2, pp. 170–197, 1997.

[31] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, John Wiley and Sons, Inc., Canada, 2nd edition, Oct. 2000.

[32] C.M. Bishop, *Pattern Recognition and Machine learning*, Springer Science+Business Media, LLC, New York, NY, Feb. 2006.

[33] Q. Jin and A. Waibel, "Application of LDA to speaker recognition," in *Proc. of International Conference on Spoken Language Processing (ICSLP)*, Oct. 2000, vol. 5, pp. 250–253.

[34] K. Fukunaga, Ed., *Introduction to Statistical Pattern Recognition*, Academic Press. Inc., Boston, 2nd edition, 1990.

[35] X. He and P. Niyogi, "Locality preserving projections," *Advances in Neural Information Processing Systems*, vol. 16, 2004.

[36] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis," Tech. Rep., Department of Computer Science, Tokyo Institute of Technology, Japan, 2006.

[37] A.M. Martínez and A.C. Kak, "PCA versus LDA," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, Feb. 2001.

[38] P.N. Belhumeur, J.P. Hepanha, and D.J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transcations on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, July 1997.

[39] Q. Li, J. Ye, and C. Kambhamettu, "Linear projection methods in face recognition under unconstrained illuminations: a comparative study," in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004, vol. 2, pp. 474–481.

[40] W. Hwang, T. Kim, and S. Kee, "LDA with subgroup PCA method for facial image retrieval," in *International workshop on image analysis for multimedia interactive services (WAMIS)*, Apr. 2004.

[41] G. Fant, *Acoustical Theory of Speech Production: With Calculations based on X-Ray Studies of Russian Articulations*, The Hague, Mouton, 1970.

[42] S. Sagyama, K. Takahashi, H. Kameoka, and T. Nishimoto, "Specmurt anasylis: A piano-roll-visualization of polyphonic music signal by deconvolution of log-frequency spectrum," in *Workshop on Statistical and Perceptual Audio Processing*, Oct. 2004, p. 128.

[43] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, 1995.

[44] C.C. Chang and C.J. Lin, *LIBSVM: a library for support vector machines*, 2001, http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[45] T. Kitahara, M. Goto, and H.G. Okuno, "Pitch-dependent identification of musical instrument sounds," *Applied Intelligence*, vol. 23, no. 3, pp. 267–275, Dec. 2005.

[46] J. Eggink and G.J. Brown, "Application of missing feature theory to the recognition of musical instruments in polyphonic audio," in *Proc. of International Conference on Music Information Retrieval (ISMIR)*, Oct. 2003, pp. V–553–556.

[47] P. Jinachitra, "Polyphonic instrument identification using independent subspace analysis," in *Proc. of International Conference on Multimedia and Expo (ICME)*. June 2004, IEEE Computer Society.

[48] J.C. Brown, O. Houix, and S. McAdams, "Feature dependence in the automatic identification of musical woodwind instrument," *Journal of Acoustical Society of America*, vol. 109, no. 3, pp. 1064–1072, Mar. 2001.

[49] "RWC music database," http://staff.aist.go.jp/m.goto/RWC-MDB.

[50] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Music genre database and musical instrument sound database," in *Proc. of International Conference on Music Information Retrieval (ISMIR)*, 2003, pp. 229–230.

[51] R. Kohavi and F. Provost, "Glossary of terms," 1998.

[52] P.O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *The Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, Dec. 2004.

[53] A. Cont and S. Dubnov, "Realtime multiple-pitch and multiple-instrument recognition for music signals using sparse non-negative constraints," in *Proc. of International Conference on Digital Audio Effects (DAFx)*, Sept. 2007, pp. 85–92.

[54] P.D. O'Grady and S.T. Rickard, "Compressive sampling of non-negative signals," in *Proc. of IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Oct. 2008.

[55] ," .

[56] P. Smaragdis, "Non-negative matrix factor deconvolutionl extraction of multiple sound sources from monophonic inputs," in *Proc. of International Conference on Independent Component Analysis and Blind Signal Separation (ICA)*, Sept. 2004.

[57] D. FitzGerald, M. Cranitch, and E. Coyle, "Extended nonnegative tensor factorisation models for musical sound source separation," *Computational Intelligence and Neuroscience*, p. 15, 2008, Article ID 872425.

[58] S.A. Abdallah and M.D. Plumbley, "Polyphonic transcription by non-negative sparse coding of power spectra," in *Proc. of International Conference on Music Information Retrieval (ISMIR)*, 2004, pp. 318–325.

[59] A. Cont, "Realtime audio to score alignment for polyphonic music instruments using sparse non-negative constraints and hierarchical hmms," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2006.

[60] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in bayesian non-negative matrix factorization applied to polyphonic music transcription," Tech. Rep., Jan. 2009.

[61] F. Itakura and S. Saito, "Speech information compression based on the maximum likelihood spectral estimation," *Journal of Acoustical Society of Japan*, vol. 27, no. 9, pp. 463–472, 1971, (in Japanese).

[62] Seiichi Nakagawa, "A survey on automatic speech recognition," *IEICE Transactions on Information and Systems*, vol. J83-D-2, no. 2, pp. 433–457, May 2000, (in Japanese).

[63] R.J. Weiss and D.P.W. Ellis, "A variational EM algorithm for learning eigenvoice parameters in mixed signals," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Apr. 2009, pp. 113–116.

[64] A. Klapuri, "Analysis of musical instrument sounds by source-filter model," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Apr. 2007, vol. 1, pp. I53–I56.

[65] T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H.G. Okuno, "Instrument identification in polyphonic music: Feature weighting to minimize influence of sound overlaps," *EURASIP Journal on Applied Signal Processing*, , no. 51979, pp. 1–15, 2007.

[66] E. Vincent and R. Xavier, "Instrument identification in solo and ensemble music using independent subspace analysis," in *Proc. of International Conference on Music Information Retrieval (ISMIR)*, 2004, pp. 576–581.

[67] Z. Ghahramani and G.E. Hinton, "Variational learning for switching state-space models," *Neural Computation*, vol. 12:44, pp. 831–864, 2000.

[68] L.J. Lee, H. Attias, and L. Deng, "Variational inference and learning for segmental switching state space models of hidden speech dynamics," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2003.

[69] L.J. Lee, H. Attias, L. Deng, and P. Fieguth, "A multimodal variational approach to learning and inference in switching state space models," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2004, pp. 920–923.

[70] L. Deng, A. Acero, and I. Bazzi, "Tracking vocal tract resonances using a quantizes nonlinear function embedded in a temporal constraint," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 425–434, Mar. 2006.

[71] J.M. Grey, "Multidimensional perceptual scaling of musical timbres," *Journal of Acoustical Society of America*, vol. 61, no. 5, pp. 1270–1277, May 1977.

[72] R. Plomp and H.J.M. Steeneken, "Place dependence of timbre in reverberant sound fields," *Acoustica*, vol. 28, pp. 50–59, 1973.

[73] S. Furui, "Key issues in voice individuality," *Journal of Acoustical Society of Japan*, vol. 51, no. 11, pp. 876–881, 1995, (in Japanese).

[74] C. Dodge and T. A. Jerse, Eds., *Computer Music: Synthesis, composition, and performance*, Schirmer, 2nd edition, July 1997.

[75] M.M. Lavrentév and L.Y. Savelév, "Linear operators and ill-posed problems," *American Mathematical Society*, vol. 34, no. 2, pp. 193–196, Apr. 1997, translated from Russian by Nanka Publishers, Moscow, 1995, xii+382.

[76] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in bayesian non-negative matrix factorization applied to polyphonic music transcription," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 99, 2009.

[77] E. Benetos, M. Kotti, and C. Kotropoulos, "Musical instrument classification using non-negative matrix factorization algorithms and subset feature selection," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2006.

[78] P. Leveau, D. Sodoyer, and L. Daudet, "Automatic instrument recognition in a polyphonic mixture using sparse representations," in *Proc. of International Conference on Music Information Retrieval (ISMIR)*, 2007.

[79] T. Heittola, A. Klapuri, and T. Virtanen, "Musical instrument recognition in polyphonic audio using source-filter model for sound separation," in *Proc. of International Conference on Music Information Retrieval (ISMIR)*, 2009.

[80] A. Cichocki and A. Phan, "Fast local algorithms for large scale nonnegative matrix and tensor factorizations," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E92.A, no. 3, pp. 708–721, 2009.

[81] C. Févotte, N. Bertin, and J.L. Durreiu, "Nonnegative matrix factorization with the itakura-saito divergence: with application to music analysis," Tech. Rep., Mar. 2009.

[82] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," in *international congress on acoustics*, Aug. 1968.

[83] N. Sugamura and F. Itakura, "Speech data compression by LSP speech analysis-synthesis technique," *Transaction of IEICE of Japan*, vol. 64A, no. 8, pp. 599–606, Aug. 1981, (in Japanese).

[84] D.R. Andrews and C.L. Mallows, "Scale mixtures of normal distributions," *Journal of the Royal Statistical Society*, vol. B, no. 36, pp. 99–102, 1974.

[85] N.H. Fletcher and T.D. Rossing, *The physics of instruments*, Springer-Verlag, New York, 2nd edition, 1998.

[86] R.M. Neal and G.E. Hinton, "A view of the EM algorithm that justifies incremental, sparse and other variants," *Learning in Graphical Models*, pp. 355–368, 1998.

[87] J.A. Nelder and R. Mead, "A simplex method for function minimization," *Computer Journal*, , no. 7, pp. 308–313, 1965.

[88] J.A. Reeds, M.H. Wright, J.C. Lagarias, and P.E. Wright, "Convergence properties of the Nelder-Mead simplex method in low dimensions," *SIAM Journal of Optimization*, vol. 9, no. 1, pp. 112–147, 1998.

[89] G.S. Young and S.E. Haupt, "Going nonlinear: towards automated puff intercept," Jan. 2007.

[90] W. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, *Numerical Recipes in C*, Cambridge University Press, 1994.

[91] P.J. Campbell and T.E. Tremain, "Voiced/unvoiced classification of speech with applications to the U.S. government LPC-10E algorithm," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Apr. 1986, pp. 473–476.

[92] J. Marozeau, A. de Cheveigné, S. McAdams, and S. Winsberg, "The dependency of timbre on fundamental frequency," *Journal of Acoustical Society of America*, vol. 114, pp. 2946–2957, 2003.

[93] E. Benetos, M. Kotti, C. Kotropoulos, J.J. Burred, G. Eisenberg, M. Haller, and T. Sikora, "Comparison of subspace analysis-based and statistical model-based algorithms for musical instrument classification," in *Workshop on Immersive Communication and Broadcast Systems (ICOB)*, Berlin, Germany, Oct. 2005.

[94] C. Li, *Non-Gaussian, non-stationary, and nonlinear signal processing methods - with applications to speech processing and channel estimation*, Ph.D. thesis, Aalborg University, Feb. 2006.

[95] N. Minematsu, T. Nishimura, K. Nishinari, and K. Sakuraba, "Theorem of the invariant structure and its derivation of speech gestalt," in *International Workshop on Speech Recognition and Intrinsic Variations*, May 2006, pp. 47–52.

Acknowlegement

本研究をすすめるにあたり，多くの方々に御指導，ならびに御鞭撻を賜った．

主指導教員である，論理生命学講座 池田和司 教授には研究に集中できる研究環境を与えていただいた．私はなかなか研究結果を出すことができなかったが，情報幾何や機械学習の理論研究の立場から多くのことを教えていただいた．謹んで感謝の意を表する．

音情報処理学講座 鹿野清宏 教授には，博士前期課程の頃から，音情報処理の立場から貴重なご意見をいただき，普段私自身が気づかない点を指摘してくださったことに対して謝意を表する．

論理生命学講座 柴田智広 准教授には，認知科学やシステム制御の観点からのご意見をいただいた．研究もスポーツもそれ以外も自分が満足いくまでされており，いい意味で研究者の壁を壊してくれた．充実した研究生活を送ることができたのも柴田先生のおかげである．五年間かげながら気づかっていただいたことに感謝する．

京都大学 情報学研究科 論理生命学分野 前田新一 助教は，確率モデルや情報理論に全く無知であった私に対して，私のレベルに合わせて基礎から一つずつ丁寧に説明して下さった．京都大学に転任されてからも論文締め切り直前に夜遅くまで遠隔で添削に付き合っていただき，否定的な私をいつも励ましてくださった．前田先生のような研究熱心な方の元で研究できたことを嬉しく思う．ここに深くお礼を申し上げる．

博士前期課程と博士後期課程の初めの 1 年間指導してくださった 京都大学 情報学研究科 論理生命学分野 石井信 教授には，私が研究室を移ってからも論文の添削をしていただいた．難しい問題をわかりやすい問題に置き換えて誰にでもわかるように説明する石井先生の話には何度感動させられたかわからない．表立って意見を言われることはほとんどなかったが，研究をしたいと思っている学生に

対してこっそりと他人には気づかれぬようサポートしておられたように思う．ここに深謝の意を表する．

本学 論理生命学講座より，京都大学 情報学研究科 論理生命学分野へと移られた方々からは，研究の基礎的知識を多く教わった．特に，統計輪講において，質問の言い方を変えたりヒントを出して私が自分自身で答えにたどり着けるまで付き合ってくださった 大羽成征 講師，計算機に関する基礎的知識だけでなく日本語や英語の表現方法まで教えてくださった 兼村厚範 博士，困っているときに適切な情報を提供してくださり基礎の基礎から説明してくださった 植野剛 氏，唯一音楽情報処理に関する研究に関するざっくばらんな話を共有することのできた 寺村佳子 氏に深謝する．

本学 論理生命学講座の皆様には，全く研究内容が違うにも関わらず研究に関して多くのアドバイスをいただき，日々の研究生活を助けていただいた．特に，博士前期課程から統計輪講での発表や私の研究に対して鋭い指摘をしてくださった 竹之内高志 助教，研究に関してこっそりゼミ後にフランクな意見を言ってくれただけでなくスポーツを通して私の集中力や忍耐力を上げさせてくれたであろう 為井智也 博士，研究室の廊下にあるホワイトボードの前で研究についてのアイデアを夜中まで一緒に語った 喜多いずみ 氏にも深くお礼を申し上げる．

同じ研究科の先輩方からの助言にもしばしば助けられた．特に，研究室修士時代より研究や学生生活について多くの知恵を教えていただき，博士論文を執筆するにあたっては経験からのアドバイスをいただいた，東京農工大学 工学部 電気電子工学科 中村幸紀 助教，ならびに，立命館大学 総合理工学院 情報理工学部 奥健太 助教には深く感謝の意を表したい．

困ったときにはいつも同級生が助けてくれたように思う．わずか一年しか同じ場所で研究できなかったが，面倒だと言いながらも確率統計について教えてくれ，研究と私生活両方に関して感じたことを率直に言ってくれた 神経計算学講座 中野高志 氏，音全般の質問や疑問に嫌な顔ひとつせずに彼なりの解答をしてくれ研究生活に関する悩みごとに対する少々厳しいが的を射た意見をくれた 音情報処理学講座 高橋祐 氏，機械学習の話だけでなく感情的な私に対して常に冷静かつ客観的な意見をくれ，いつも長話につきあってくれた 自然言語処理学講座 渡邉陽太郎 氏，いろいろな研究所の話を聞かせてくれたり悩み相談にのってくれた 自然言語処理学講座 小町守 氏にも感謝を申し上げたい．

また，秘書の仕事だけではなく英文校正までしてくださった 論理生命学講座 元秘書 藤澤祐子さん，書類で分からないところがあったら丁寧に教えてくださっ

ただけでなくいつも私の雑談につきあってくださった 論理生命学講座 秘書 谷本史さんにも謝意を表したい.

最後に，小さな頃から常に人と違う経験をさせてくれた母 千秋と，学びたいなら思う存分しなさいと言ってくれた父 弘実には，長年の私のわがままを許してくれたことに感謝している.

Appendix A: Theorem and probabilistic distributions

- 1-dimensional Gaussian distribution

$$\mathcal{N}(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{A-1}$$

- D-dimensional Gaussian distribution

$$\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \tag{A-2}$$

- Laplace distribution

$$\mathcal{LA}(x|\mu,b) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right) \tag{A-3}$$

- Multivariate (D-dimensional) Laplace distribution

$$\mathcal{LA}(\mathbf{x}|\boldsymbol{\mu},\mathbf{B}) = \prod_{d=1}^{D}\frac{1}{2B_d}\exp\left(-\frac{|x_d-\mu_d|}{B_d}\right), \tag{A-4}$$

$$\text{where}\quad B_d = \frac{1}{N}\sum_{n=1}^{N}|x_{nd}-\mu_d|. \tag{A-5}$$

  N, $\boldsymbol{\mu}$ and $\mathbf{B}$ denote the number of samples, a median vector and a scale parameter vector, respectively.

- Jensen's inequality

$$\mathbb{E}[f(X)] \leq f(E[X]) \rightarrow \mathbb{E}[\log X] \leq \log \mathbb{E}[X] \tag{A-6}$$

Appendix B: Free energy calculation in variational EM

In the calculation of the free energy, we assume the trial distribution $q(X_{1:T}|\kappa)$ is a single Gaussian distribution,

$$q(X_{1:T}|\kappa) = \mathcal{N}(X_{1:T}; \mu, S), \tag{B-1}$$

where $\kappa = \{\mu, S\}$. The free energy then becomes

$$\begin{aligned}
\mathcal{F} &= (q(X_{1:T}|\kappa)) && \text{(B-2)}\\
&= -\int \cdots \int q(X_{1:T}|\kappa) \log p(X_{1:T}, S_{1:T}|\theta) dX_{1:T} \\
&\quad + \int \cdots \int q(X_{1:T}|\kappa) \log q(X_{1:T}|\kappa) dX_{1:T} && \text{(B-3)}\\
&= -\int \cdots \int q(X_{1:T}|\kappa) \left(\log \left(p(s_1|x_1,\theta) p(x_1|\theta) \prod_{t=2}^{T} p(s_t|x_t,\theta) p(x_t|x_{t-1},\theta) \right) \right) dx_{1:T} \\
&\quad + \int \cdots \int q(X_{1:T}|\kappa) \log q(X_{1:T}|\kappa) dX_{1:T} && \text{(B-4)}\\
&= -\int q(x_1|\kappa) \log p(x_1|\theta) dx_1 - \sum_{t=1}^{T} \int q(x_t|\kappa) \log p(s_t|x_t,\theta) dx_t \\
&\quad - \sum_{t=2}^{T} \int q(x_t, x_{t-1}|\kappa) \log(x_t|x_{t-1},\theta) dx_t dx_{t-1} - \mathrm{H}\left(q(X_{1:T}|\kappa)\right) && \text{(B-5)}\\
&= \mathcal{F}_1 + \mathcal{F}_2 + \mathcal{F}_3 + \mathcal{F}_4. && \text{(B-6)}
\end{aligned}$$

B.1 The term for initial state

The first term is corresponds to the initial state, which is

$$-\int q(x_1|\kappa) \log p(x_1|\theta) dx_1. \tag{B-7}$$

From the likelihood of

$$\begin{aligned}
p(x_1|\theta) &= \mathcal{N}(a_1; m_a^1, (\sigma_a^1)^2)\, \mathcal{N}(f_1; m_f^1, (\sigma_f^1)^2) \\
&= \mathcal{N}(x_1; m_1, (\sigma_1)^2), && \text{(B-8)}
\end{aligned}$$

the first term results in

$$\begin{aligned}\mathcal{F}_1 &= -\int q(x_1|\kappa)\log p(x_1|\theta)dx_1 \\ &= -\int \mathcal{N}(x_1;\mu_1,S_1)\log \mathcal{N}(x_1;m_1,(\sigma_1)^2)dx_1 \\ &= \frac{1}{2}\left(\mathrm{Tr}\left((\sigma_1)^{-2}S_1\right)+(\mu_1-m_1)^{\mathrm{T}}(\sigma_1)^{-2}(\mu_1-m_1)\right)+\frac{1}{2}\log|(\sigma_1)^2|+\log(2\pi).\end{aligned} \tag{B-9}$$

B.2 The term for observation process

The second term is

$$-\int q(x_t|\kappa)\log p(y_t|x_t,\theta)dx_t, \tag{B-10}$$

where the likelihood is

$$\begin{aligned}p(y_t|x_t) &= \prod_{i=1}^{N} p(y_t(i)) \\ &= \prod_{i=1}^{N}\frac{1}{\sqrt{2\pi}\sigma_o y_t(i)}\exp\left(-\frac{1}{2\sigma_o^2}(\log y_t(i)-\log\hat{y}_t(i))^2\right).\end{aligned} \tag{B-11}$$

As a result of substitution, it becomes

$$\begin{aligned}\mathcal{F}_2 &= -\sum_{t=1}^{T}\int q(x_t|\kappa)\log p(y_t|x_t,\theta)dx_t \\ &= -\sum_{t=1}^{T}\int \mathcal{N}(x_1;\mu_1,S_1)\sum_{i=1}^{N}\log\frac{1}{\sqrt{2\pi}\sigma_o y_t(i)}\exp\Big(-\frac{1}{2\sigma_o^2}(\log y_t(i)-\log\hat{y}_t(i))^2\Big)dx_t.\end{aligned} \tag{B-12}$$

Substitute $\hat{y}_t(i) = H(i)G_t(i)$ and $q = \mathcal{N}(x_t; \mu_t, S_t)$ into this equation, then we obtain

$$
\begin{aligned}
&= -\sum_{t=1}^{T} \int \mathcal{N}(x_t; \mu_t, S_t) \sum_{i=1}^{N} \left(-\frac{1}{2}\log(2\pi) - \log(\sigma_o) - \log(y_t(i)) \right. \\
&\quad \left. -\frac{1}{2\sigma_o^2} \left(\log y_t(i) - \log H(i) \log G_t(i)\right)^2 \right) dx_t \\
&= \sum_{t=1}^{T} \left[\frac{N}{2}\log(2\pi) + \sum_{i=1}^{N} \log(s_t(i)) + N\log(\sigma_o) + \frac{1}{2\sigma_o^2} \sum_{i=1}^{N} (\log y_t(i) - \log H(i) \log G_t(i))^2 \right. \\
&\quad \left. + \frac{1}{2\sigma_o^2} \sum_{i=1}^{N} \left(S_{at} + \mu_{at}^2 + \sqrt{\frac{2}{\pi}} A(\omega(i)) \mu_{at} K_{\exp}(i) + \frac{A(\omega(i))^2}{2\pi\sigma_p} \mathrm{KL}_{\exp}(i) \right) \right] \\
&\quad - \frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(\log \frac{y_t(i)}{H(i)} \left(\mu_{at} + \frac{A(\omega(i))}{\sqrt{2\pi}} K_{\exp}(i) \right) \right), \qquad \text{(B-13)}
\end{aligned}
$$

where

$$
K_{\exp}(i) = \sum_{k}^{K} \frac{1}{\sqrt{k^2 S_{ft}} + \sigma_p^2} \exp\left(-\frac{1}{2} \left(S_{ft} + \frac{\sigma_p^2}{k^2} \right)^{-1} \left(\mu_{ft} - \frac{\omega(i)}{k} \right)^2 \right), \qquad \text{(B-14)}
$$

$$
\begin{aligned}
\mathrm{KL}_{\exp}(i) = \sum_{k,l} & \frac{1}{\sqrt{\left((k^2+l^2)S_{ft} + \sigma_p^2\right)}} \\
& \exp\left(-\frac{1}{2} \left(\left(S_{ft} + \frac{\sigma_p^2}{k^2+l^2} \right)^{-1} \left(\mu_{ft} - \frac{k+l}{k^2+l^2}\omega(i) \right)^2 + \frac{(k-l)^2}{k^2+l^2} \frac{(\omega(i))^2}{\sigma_p^2} \right) \right), \qquad \text{(B-15)}
\end{aligned}
$$

$$
H(\tilde{\omega}, b_1, \cdots, b_p) \qquad \text{(B-16)}
$$

$$
= 2^{1-p} \left(\sin^2 \frac{\tilde{\omega}}{2} \prod_{n=2,4,\cdots,p} (\cos\tilde{\omega} - \cos b_n)^2 + \cos^2 \frac{\tilde{\omega}}{2} \prod_{n=1,3,\cdots,p-1} (\cos\tilde{\omega} - \cos b_n)^2 \right)^{-2}, \qquad \text{(B-17)}
$$

$$
\tilde{\omega} = \frac{\mathrm{Fs}\omega}{2\pi}. \qquad \text{(B-18)}
$$

B.3 The term for state transition

The third term is

$$-\int q(x_t, x_{t-1}|\kappa) \log p(x_t|x_{t-1}, \theta) dx_t dx_{t-1}. \tag{B-19}$$

Using the likelihood

$$p(x_t|x_{t-1}, \theta) = p(a_t|a_{t-1}, \theta) p(f_t|f_{t-1}, \theta), \tag{B-20}$$

for the calculation of the state equation results in

$$\begin{aligned}
\mathcal{F}_3 &= -\sum_{t=2}^{T} \int\int q(x_t, x_{t-1}|\kappa) \log p(x_t|x_{t-1}, \theta) dx_t dx_{t-1} \\
&\approx \sum_{t=2}^{T} \left[\frac{1}{2} \log \upsilon_a + \log \Gamma\left(\frac{\upsilon_a}{2}\right) - \log \Gamma\left(\frac{\upsilon_a + 1}{2}\right) + \frac{1}{2} \log \upsilon_f + \log \Gamma\left(\frac{\upsilon_f}{2}\right) \right. \\
&\quad - \log \Gamma\left(\frac{\upsilon_f + 1}{2}\right) + \frac{\upsilon_a + 1}{2} \left(\frac{\beta_a^4 U_a}{\upsilon_a^2} + \frac{\beta_a^2 V_a}{\upsilon_a} + abc_{a,t} \right) \\
&\quad \left. + \frac{\upsilon_f + 1}{2} \left(\frac{\beta_f^4 U_f}{\upsilon_f^2} + \frac{\beta_f^2 V_f}{\upsilon_f} + abc_{f,t} \right) + \log \beta_a + \log -\beta_f + \log \pi \right],
\end{aligned} \tag{B-21}$$

where

$$\begin{aligned}
U_a &= a_{a,t} \left(3\Sigma_{u,a}^2 + e_{u,a}^4 + 6\Sigma_{u,a} e_{u,a}^2 \right), \\
U_f &= a_{f,t} \left(3\Sigma_{u,f}^2 + \mu_{u,f}^4 + 6\Sigma_{u,f} \mu_{u,f}^2 \right), \\
V_a &= (2a_{a,t} + b_{a,t}) \left(\Sigma_{u,a} + e_{u,a}^2 \right), \\
V_f &= (2a_{f,t} + b_{f,t}) \left(\Sigma_{u,f} + \mu_{u,f}^2 \right), \\
e_{u,a} &= \mu_{u,a} - \log \rho, \\
e_{u,f} &= \mu_{u,f}, \\
abc_{*,t} &= a_{*,t} + b_{*,t} + c_{*,t}, \\
a_{*,t} &= \frac{w_{a1}}{x_{\max,*,t} - 1} + w_{a2}, \\
b_{*,t} &= -\frac{w_{b1}}{\sqrt{x_{\max,*,t}} - 1 + w_{b3}} + w_{b2}, \\
c_{*,t} &= w_{c1} \log(x_{\max,*,t} - 1) + w_{c2},
\end{aligned}$$

$$
\begin{aligned}
x_{\max,*,t} &= 1 + \frac{(\mu_{u,a} + k_a\sqrt{S_{u,*}})^2 - \log\rho}{\upsilon_a}, \\
x_{\min} &= 1, \\
\mu_{u,*} &= \frac{1}{\sqrt{2}}(\mu_{*,t} - \mu_{*,t-1}), \\
S_{u,*} &= \frac{1}{2}(S_{*,t} + S_{*,t-1} - 2S_{*,t,t-1}), \\
e_{u,a} &= \mu_{u,a} - \log\rho, \\
e_{u,f} &= \mu_{u,f}.
\end{aligned}
$$

Note that $*$ represents the set of a and f.

B.4 The term for entropy

The last term expresses the entropy which is

$$
\mathcal{F}_4 = -\mathrm{H}\big(q(X_{1:T}|\kappa)\big). \tag{B-22}
$$

The calculation results in

$$
\begin{aligned}
\mathcal{F}_4 &= -\mathrm{H}(q(X_{1:T}|\kappa)) \\
&\quad - \mathrm{H}(\mathcal{N}(X_{1:T}|\mu, S)) \\
&= -\frac{1}{2}(n + n\log(2\pi) + \log|S|).
\end{aligned} \tag{B-23}
$$

Appendix C: A list of 30 CDs used in experiments

Table 4.2.1-4.2.2 show the list of CDs used in the experiment. All RWC CDs are regarded as one source. There are 30 CDs including 47 different musical performances (violin: 6, cello: 6, guitar: 5, piano: 7, flute: 7, oboe: 6, horn: 5, trumpet: 5).

Title	player/composer	instruments
The 18 monologues for instruments	Erland von Koch (composer)	all
Sequenzas I-XIV for solo instruments	Luciano Berio (composer)	all except hr
J.S. Bach/ B.A. Zimmermann	Thomas Demanga	vn, vc
RWC-MDB-C-M06 (classic)		vn, vc
Solo violin sonatas	Eugéne Ysaÿe	vn
Sonata for solo violin SZ 117	Béla Bartók (composer)	vc
Virtuoso music for violoncello	Janos Starker	vc
Bach: the unaccompanied cello suits	Yo-Yo-MA	vc
RWC-MDB-J-M01 (jazz)		gt, pf
RWC-MDB-C-M05 (classic)		gt, pf
RWC-MDB-C-M04 (classic)		pf
RWC-MDB-G-M06 (music genre)		pf
Jazz piano ever	V.A.	pf
B.C. A.D.	T-square	pf
RWC-MDB-G-M08 (music genre)		gt
Six suites for violoncello solo	Wangenheim	gt
SOLOS: solo workds of Daniel Asia	Daniel Asia (composer)	ob
Oboe music	Helen Jahren (composer)	ob
Oboe solo	Yeon-Hee Kwak	ob
Noctune	Hansjörg Schellenberger	ob
Teleman: tweleve fantasies for oboe solo	Heinz Holliger	ob

Table 4.2.1. list of CDs used for the experiment in Chapter II

Music for flute	Brian Ferneyhough	fl
Bach music for solo flute	Wilbert Hazelzet	fl
Manuera plays French solo flute music	Mauela Wiesler	fl
In lines of dazziling light	John Buckley (composer)	fl, hr
First chairs: cantos for solo instruments	Samuel Aduler	tp, hr
Bach: works for trumpet	Alison Balsom	tp
Bach cello suites on trumpet	David Coopen	tp
Southwest chamber music	Richard Derby	hr
J.S. Bach: 3 suits	Radek Baborák	hr

Table 4.2.2. list of CDs used for the experiment in Chapter II (continued)

List of publications

Journal papers

- Mizuki Ihara, Shin-ichi Maeda and Shin Ishii, ”Solo Instrumental Music Analysis using the Source-Filter Model as a Sound Production Model Considering Temporal Dynamics,” Neural Computing and Applications, Vol.18, Issue 1, pp.3-14, 2009.

International conferences (reviewed)

- Mizuki Ihara, Shin-ichi Maeda and Shin Ishii, ”Instrument Identification in Monophonic Music Using Spectral Information,” IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), pp.607-611, Dec. 2007.
- Mizuki Ihara, Shin-ichi Maeda and Shin Ishii, ”Estimation of the Source-Filter Model Using Temporal Dynamics,” International Joint Conference on Neural Networks (IJCNN), pp.3098-3103, Aug. 2007.
- Mizuki Ihara, Shin-ichi Maeda and Shin Ishii, ”Estimation of the Source-Filter Model via Acoustical Feature Extraction by GA-like Algorithm,” International Symposium on Artificial Life and Robotics (AROB), GS12-4, Jan. 2007.

Others

- 井原瑞希，池田和司，前田新一．判別分析の幾何的解釈と楽器特徴抽出法の考察．電子情報通信学会技術報告，SIP2010.

- 井原瑞希，池田和司，前田新一．判別分析の幾何的解釈と楽器特徴抽出への適用．情報処理学会研究報告，2010-MUS-84.
- 井原瑞希，前田新一，石井信．ダイナミクスを考慮したソースフィルタモデルの推定．情報処理学会研究報告，2008-MUS-78.
- 井原瑞希，前田新一，石井信．単旋律楽曲における機械学習を用いた楽器音の音源同定．第 26 回 AI チャレンジ研究会，SIG-Challenge-A702, 2007.
- 井原瑞希，前田新一，石井信．線スペクトル対を用いた楽器分類．電子情報通信学会技術報告，NC2006-207，pp.115-118.