2008－MUS－78 (15)
2008／12／20

ダイナミクスを考慮したソースフィルタモデルの推定

井原 瑞希† 前田 新一†† 石井 信††

† 奈良先端科学技術大学院大学，情報科学研究科
†† 京都大学大学院，情報学研究科

あらまし 調波楽器音の生成過程は，しばしば音源信号の生成と共鳴による調音フィルタを分離したソースフィルタモデルで表現される．音源特性を同定することによりピッチの推定ができ，調音フィルタを同定することで楽器の種別を推定するのに役立てられる．しかしこのモデルでは，用いた観測可能な音波形から2つの未知量である音源特性と調音フィルタの推定を行うことを必要とし，このモデルの仮定だけでは音源特性と調音フィルタは特定できないという不定性を有する．そこで本研究では，音源特性と調音フィルタをパラメトリックな関数として表現し，さらに音源特性の滑らかな時間変化，調音フィルタの時間不変性を仮定した確率モデルを構築することで不定性の除去を試みた．この確率モデルに変分EMアルゴリズムによるパラメータ学習を行うことで，調音フィルタのパラメータを利用した楽器推定の精度が高まることが示された．

Estimation of the Source-Filter Model Considering Temporal Dynamics

Mizuki Ihara†, Shin-ichi Maeda††, and Shin Ishii††

† Graduate School of Information Science, Nara Institute of Science and Technology
†† Graduate School of Informatics, Kyoto University

Abstract The harmonic instrument sound production is often represented by a source-filter model, which separately represents the source signal generation and the synthesis filter. The source signal estimation makes pitch identification possible as well as the synthesis filter estimation helps instrument identification. However, this source-filter model has inherent ill-posedness since the estimation with this model requires the estimation of two unknowns: the unknown source signal and the unknown synthesis filter from the observed sound signal. In this study, we assume that the source signal corresponds to the time-variant pitch and amplitude, in addition to that the time-invariant synthesis filter contains the information of the instruments. With those assumptions, we constructed the probabilistic sound production model. After the model parameters are learned by the variational EM algorithm that minimizes the free energy, those are utilized to reconstruct the spectrum envelope and to identify the instruments. The experimental results suggested that the learning scheme could achieve simultaneous estimation of both source signal and synthesis filter.

1. はじめに

音の生成過程を音源特性と調音フィルタの線形接続で表現できると仮定したソースフィルタモデル[1]は，音声合成や音声特徴抽出などに使用されている．このモデルは，音の発生源パターン G と，共鳴管の共振，反共振特性を表す調音フィルタ H の畳み込みによって，音信号のスペクトル s が生成されると仮定している．例えば，バイオリンの音のスペクトル s においては，弦振動が音源信号であり，楽器の筐体がフィルタとしてはたらく．一方，音声の場合，音源信号は声帯振動によって生成され，声道特性に依存して異なる音声が発声される．この音源特性と調音フィルタを推定することができれば，楽音合成のみならず，楽器音の符合化や採譜に利用できると考えられる．本報告では，音のダイナミクスを考慮した確率モデルに基づき，音源特性と調音フィルタを同時に推定する手法を提案し，推定されたソースフィルタモデルを利用して単旋律楽曲における楽器同定を行った結果について報告する．

ソースフィルタモデルは，調波楽器音の生成メカニズムを簡単かつ，比較的正確に表現することのできるモデルであるが，複数の異なる G と H の組み合わせが同一の観測スペクトル s を

表現できるという不定性をもつ．そのため，ソースフィルタモデルの同定には，この不定性を解消するための適切な制約の付加が必要である．

従来研究では，何らかの方法でソースフィルタモデルの不定性を解消し，音源特性と調音フィルタの推定を行っている．板倉らは，ソースフィルタモデルの不定性を解消するための制約として，短時間音声信号が定常ガウス過程にしたがって生成され，スペクトル密度は零点をもたないという仮定をおいた．その上で，自己回帰モデルのパラメータを最尤スペクトル推定法によって求め，自己回帰モデルによって得られる包絡スペクトルを調音フィルタとし同定した後，スペクトルの予測残差から音源信号の情報を抽出した[2]．しかし，この定常ガウス過程の仮定は，音の音高や音量の非定常な時間変化を無視する問題をもつ．

Klapuri は，ソースフィルタモデルを用いて，音信号のスペクトルを音源特性と調音フィルタの基底関数の線形結合で表現することを提案した[3]．この研究では，音信号のスペクトルを 4 つの部分，定常ハーモニクス，定常調音フィルタ，時変損失フィルタ，時変モデル誤差に分ける．ここで時変損失フィルタはスペクトル領域での周波数依存の減衰を表現している．音源特性（ハーモニクス）の基底関数は，あらかじめ主成分分析（PCA）によって求めておき，調音フィルタと損失フィルタは，それぞれ同じ三角の形状のバンドパスフィルタを一様に，隣のフィルタと 50%ずつ重なるようにしておいたものを用いた．これが適切に楽器の時変スペクトル密度を表現できているかを 33 種類の楽器において多旋律楽曲からメロディラインを抜き出すことで評価した．このモデルは計算量の扱いやすさから線形モデルを仮定しているが，実際の音生成には非線形な要素があると考えられる．

本研究では，音の生成過程がソースフィルタモデルで近似できるとし，音源信号が音高と音量の情報を，調音フィルタが楽器の特徴を含むと仮定したパラメトリックな確率モデルを構築し、学習データから確率モデルのパラメータを学習させることでソースフィルタモデルを同定する．ソースフィルタモデルに含まれる不定性を緩和するために，ダイナミクスをもつ確率モデルとし，音の時間的連続性を考慮した音源信号と調音フィルタの推定を行う．具体的には単旋律楽曲で時間的に変化する音高と音量の時間変化情報を音源特性のモデルに取り入れる一方，楽器の形状に依存すると考えられる調音フィルタは，演奏中に楽器の形状がほとんど変化しないことから定常とする．調音フィルタは，上記の板倉らのモデルと同様のモデルを用いてパラメトライズした．モデルパラメータの学習には，EM アルゴリズムに類似したアルゴリズムである変分 EM アルゴリズムを用いた．データから学習した調音フィルタのパラメータが仮定した通り、楽器固有のパラメータとして得られているかどうかを確認するために、調音フィルタのパラメータを用いて単旋律楽曲の楽器を推定し，提案モデルの妥当性を評価した．

2. 音響生成モデル

2.1 モデルの定式化

提案モデルでは，ソースフィルタモデルの不定性をやわらげるための制約として，音高，音量の連続的な時間変化と調音フィルタの定常性を仮定する．

2.1.1 非線形ダイナミカルシステム

時刻 t における d 次元短時間振幅スペクトルベクトルを s_t で表し，時刻 t における対数音圧と基本周波数をそれぞれ a_t, f_t で表す．a_t と f_t の組みは，2 次元ベクトル $x_t = [a_t, f_t]^{\mathrm{T}}$ (ここで T は転置を表す) で表す．時系列 $\{s_1, \cdots, s_T\}$ と $\{x_1, \cdots, x_T\}$ をそれぞれ，$S_{1:T}, X_{1:T}$ で表す．

このとき，$S_{1:T}, X_{1:T}$ の同時分布を以下のように表現する．

$$
\begin{aligned}
&p(X_{1:T}, S_{1:T}|\theta) \\
&= p(x_1, s_1|\theta) \prod_{t=2}^{T} p(x_t, s_t|X_{1:t-1}, S_{1:t-1}, \theta) \\
&= p(s_1|x_1, \theta) p(x_1|\theta) \prod_{t=2}^{T} p(s_t|x_t, \theta) p(x_t|x_{t-1}, \theta). \qquad (1)
\end{aligned}
$$

以降，この同時分布を構成する各分布の説明を行う．

2.2 観測過程の確率モデル

対数音圧 a_t と基本周波数 f_t が与えられたもとで，観測スペクトル s_t が得られる確率分布 $p(s_t|x_t, \theta)$ は，以下の自由度 3 のカイ二乗分布とした[4]．

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

s_t と $\hat{s}_t$ が十分近い場合，分布は次のように近似できる．

$$
\approx \prod_{i=1}^{d} \frac{1}{2\Gamma(1.5) \hat{s}_t(i) \sigma_o} \left(\frac{1}{2\sigma_o} \frac{s_t(i)}{\hat{s}_t(i)}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2\sigma_o} \frac{s_t(i)}{\hat{s}_t(i)}\right), \qquad (3)
$$

ここで $\hat{s}_t$ は，ソースフィルタモデルに基づく推定スペクトルであり，音源特性を $G_t(x_t; \theta)$, 調音フィルタを $H_t(\theta)$ として $\hat{s}_t = G_t(x_t; \theta) \odot H_t(\theta)$ で表現される．$\odot$ は行列のアダマール積を表す．また，θ は分布や関数の形状を決めるパラメータである．この観測分布に対数をとった関数は，$\sigma_o = 1$ の場合，真のスペクトル s_t と推定スペクトル $\hat{s}_t$ との距離を測る Itakura-Saito 距離[5] と等しくなる．

2.2.1 LSF パラメータによるフィルタ H の表現

調音フィルタ $H_t(\theta)$ は，スペクトル包絡を表すものとして，しばしば線スペクトル対（line spectrum frequencies; LSF）によって推定される[6]．楽器推定においては，LSF は線形予測分析（linear predictive coding; LPC）係数やメル周波数ケプストラム係数（mel-frequency cepstrum coefficients; MFCCs）などの他の楽器特徴抽出方法と比較して，より高精度な楽器特徴抽出が可能

であることが楽器判別によって実証されている[7]．LSFによるスペクトル包絡の表現は次式によって与えられる．

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

ここで $\tilde{\omega}$ は，Fs をサンプリング周波数，ω を周波数として $\tilde{\omega}=\frac{\omega \mathrm{Fs}}{2\pi}$ で定義される正規化角周波数である．本報告では，周波数 ω は離散周波数を考えるため，$\tilde{\omega}$ は離散値をとる．LSFの次数 p が小さい場合，LSFによって推定されたスペクトルはなめらかになるため，スペクトル包絡を表現することとなる．ここでは，従来研究[8]を参考にLSF係数を用いた楽器分類時の計算量と楽器推定の精度のトレードオフを考えて $p=12$ と設定した．この式において b_n（$n=1,\cdots,p$）が調音フィルタを表現するパラメータである．

2.2.2 時間変化情報を含んだ音源信号 G の生成

音源特性 $G_t(x_t;\theta)$ は，基本周波数 f_t の整数倍の周波数 kf_t（$k=1,\cdots,K$）にピークをもち，高周波数になるにつれてスペクトルの値が減衰するような関数として以下のように表現する．

$$G_t(\omega_i;a_t,f_t,K,\sigma_p,\tau,A) = \exp(a_t + A\exp(-\frac{\omega_i}{\tau})\sum_{k=1}^{K}\mathrm{Gauss}(\omega_i;kf_t,{\sigma_p}^2)). \quad (5)$$

ここで ω_i はインデックスが i の離散周波数，A は振幅調整のパラメータ，τ は周波数減衰調整のパラメータ，K は倍音周波数成分の数である．また，$\mathrm{Gauss}(x;\mu,\sigma^2)$ は平均 μ，分散 σ^2 の x のガウス分布関数を示す．

2.3 状態遷移の確率モデル

対数音圧 a_t と基本周波数 f_t は，それぞれ確率的に時間変化するものとして定式化する．実際の楽器音では，ある一音が演奏されたとき，音の高さはしばらく時間ほぼ変化せず持続し，音の大きさは時間とともに小さくなる．また，違う音が演奏されたときには，音の高さも大きさも，前の時刻の音にそれほど強く依存せず大きく変化しうる．この事実から，$x_t=[a_t,f_t]^T$ の時間変化を，連続的に変化する場合（$\eta=1$）と非連続的に変化する場合（$\eta=0$）に分けて，それぞれ確率 $\bar{\eta}$ と $1-\bar{\eta}$ で生じると仮定する．

$$p(x_t|x_{t-1},\theta) = \bar{\eta}p(x_t|x_{t-1},\eta=1,\theta)+(1-\bar{\eta})p(x_t|x_{t-1},\eta=0,\theta) \quad (6)$$

ここでは，$\bar{\eta}$ は0.8とした．

2.3.1 連続的な遷移

対数音圧と基本周波数の連続的な遷移はそれぞれ独立に起こると仮定し，以下のような確率モデルで記述できるとする．

$$p(x_t|x_{t-1},\eta=1,\theta) = \mathrm{Gauss}(a_t;\bar{a}_{t-1}+\log\rho,\Sigma_a)\,\mathrm{Gauss}(f_t;f_{t-1},\Sigma_f), \quad (7)$$

一つ目のガウス分布が対数音圧，二つ目が基本周波数の状態遷移分布を表現し，Σ_a と Σ_f はそれらの分散を示す．

2.3.2 非連続的な遷移

連続的な変化の場合と同様に，非連続的な変化の場合も基本周波数と対数振幅の遷移が独立に起こるとして以下のような確率モデルで表現する．

$$p(x_t|x_{t-1},\eta=0,\theta) = \mathrm{Gauss}(a_t;m_a,\sigma_a^2)\,\mathrm{Gauss}(f_t;m_f,\sigma_f^2) \quad (8)$$

ここで，m_a と m_f，σ_a^2 と σ_f^2 はそれぞれ a_t と f_t の平均と分散である．

2.4 初期分布の確率モデル

隠れ状態の初期分布は，以下のように基本周波数と対数振幅とで独立なガウス分布で与える．

$$p(x_1|\theta)=\mathrm{Gauss}(a_1;m_a^1,(\sigma_a^1)^2)\,\mathrm{Gauss}(f_1;m_f^1,(\sigma_f^1)^2) \quad (9)$$

3. 変分EMアルゴリズムによるパラメータ推定

式1で与えられる同時分布を隠れ変数 $X_{1:T}$ に関して積分することにより，観測されるスペクトル時系列の尤度 $p(S_{1:T}|\theta)=\int p(X_{1:T},S_{1:T}|\theta)dX_{1:T}$ が求まる．ここで θ は確率分布を特徴付けるパラメータをまとめたものであり，

$$\theta=\{b_0,b_1,\cdots,b_p,K,A,\tau,\sigma_p,\sigma_{\{a,f\}},m_{\{a,f\}},\Sigma_{\{a,f\}},m^1_{\{a,f\}},\Sigma^1_{\{a,f\}}\}. \quad (10)$$

で与えられる．これらのうち $\sigma_{\{a,f\}}$, $m_{\{a,f\}}$, $\Sigma_{\{a,f\}}$, $m^1_{\{a,f\}}$, $\sigma^1_{\{a,f\}}$ はあらかじめ適当な値を手で与えた．残りの係数 $b_1,\cdots,b_p$ と，K,σ_p,A,τ に関して学習による推定を行った．なお，学習の際，係数 $b_1,\cdots,b_p$ の初期値にはLSFで求めた係数を用いた．パラメータの学習は，尤度最大化による最尤推定によって行うことができるが，モデルが隠れ変数を含んでおり尤度の直接評価が不可能な場合，そのような推定を行うことができない．その場合，しばしばEMアルゴリズムが用いられる．このEMアルゴリズムは隠れ変数の事後確率の計算を必要とするが，それができればEステップ，Mステップを繰り返すことで尤度を単調に増加させることが出来る．しかしながら，提案モデルでは事後確率の計算も困難である．その際に利用されるのが，自由エネルギーとよばれる関数である．この自由エネルギーを最小化することはEMアルゴリズムと等価であることが示されている[9]．

3.1 自由エネルギー

自由エネルギーは，

$$F(q(X_{1:T}),\theta)$$
$$= -\log p(S_{1:T}|\theta) + \mathrm{KL}\left[q(X_{1:T})||p(X_{1:T}|S_{1:T},\theta)\right], \quad (11)$$

と定義される．ここで $q(X_{1:T})$ は隠れ変数 $X_{1:T}$ の試験分布であり，$\mathrm{KL}[q(\cdot)||p(\cdot)]$ は二つの確率分布 $p(\cdot)$，$q(\cdot)$ 間のカルバックライブラー（KL）擬距離

$$\mathrm{KL}[q||p] = \int q(x)\log\frac{q(x)}{p(x)}dx. \quad (12)$$

である．KL 擬距離の非負性から，自由エネルギーを試験分布 $q(x_{1:T})$ について最小化した場合，自由エネルギーは負の対数尤度と一致し，θ に関する自由エネルギーの最小化は最尤推定と一致する．つまり，

$$\hat{\theta} = \arg\max_{\theta}\log p(S_{1:T}|\theta)$$
$$= \arg\min_{\theta}(\min_{q(X_{1:T})}(F(q(X_{1:T}),\theta))) \text{ である} \quad (13)$$

この自由エネルギーの試験分布 $q(x_{1:T})$ に関する最小化が EM アルゴリズムの E ステップに対応し、パラメータ θ に関する最小化が M ステップに対応する。しかし実際は，試験分布に関する自由エネルギー最小化は困難であるため，試験分布をある分布族に制約した上で自由エネルギーの最小化を行う．ここでは試験分布を平均 $\mu_{1:T}$，分散 $\Sigma_{1:T}$ のガウス分布で表現した．ここで，共分散行列 $\Sigma_{1:T}$ は各時刻の相関 $\mathrm{E}[a_t{}^2]$，$\mathrm{E}[f_t{}^2]$，$\mathrm{E}[a_t f_t]$ に隣接する時刻間の対数音圧と基本周波数の相関 $\mathrm{E}[a_{t-1}a_t]$，$\mathrm{E}[f_{t-1}f_t]$ のみを考え，他の相関についてはゼロとした．$\mathrm{E}[\cdot]$ は試験分布 $q(X_{1:T})$ に関する期待値である．

3.2 滑降シンプレックス法によるパラメータ推定

自由エネルギーを最小化するために，Nelder-Mead の滑降シンプレックス法 [10] を用いた．この方法は，制約なしの非線形最適化アルゴリズムであり，目的関数の微分値の計算を必要としないという利点がある．

シンプレックスとは，N 次元パラメータ空間上で $N+1$ 個の頂点をもつ単体のことを指す．例えば 3 次元空間で，シンプレックスは四面体となる．滑降シンプレックス法は，この四面体の 4 つの頂点の目的関数の値を初めに評価する．次に，目的関数の値が現在の値より小さくなるような点を探す方法として，4 種類の操作，reflection，expansion，contraction，shrinkage を行う．この 4 つの操作を繰り返すことで，最適点がシンプレックス内に入るようにする方法がシンプレックス法である．

4. 楽器判別方法：SVM

学習によって得られたパラメータのうち，調音フィルタのパラメータである係数 $b_1,\cdots,b_p$ を用いて楽器判別を行った．判別手法にはサポートベクタマシン（Support vector machines; SVM）を用いた．SVM は幅広い分野で使用されているが，音分野に限っても，単旋律楽器判別のみならず，多旋律や音声の推定にも用いられており，その有効性が示されている [11].

SVM はいわゆる，マージン最大化を行うことでクラス分類を可能にしている．ここでいうマージンとは，分類境界と分類境界に近いサンプル点との距離のことである．サンプルが線形分離可能である場合，学習データのサンプルを正しく分離する超平面が複数存在してしまうため，境界を一意に定めることができない．そこで，SVM はテストデータのサンプルに対する汎化能力の向上のためにマージン最大化をおこなう．また，サンプルが線形分離不可能な場合でも，サンプル空間を高次元の特徴空間に射影することで線形分離可能となる場合がある．この高次元空間への射影を陽に扱わず，カーネル関数を利用していることも SVM の特徴である [12]．本研究における実験では，LIBSVM の C 言語のインターフェースを用いた [13].

5. 実　　験

我々は，同一楽器の調音フィルタ H は，音高によらず近い関数となることを想定した．実際に，そのような調音フィルタを学習しているかどうかを確認するため，提案手法による学習後のモデルに対して以下の 3 つの実験をおこなった．

（1） 楽器は同じだが音高（音源信号）が異なっている場合に，同じような調音フィルタのパラメータが抽出できているか，つまり，調音フィルタ H によるスペクトル包絡形状が近いかどうかを確かめる．

（2） 複数の楽器の様々な楽音データから推定した調音フィルタのパラメータを特徴空間に写像したとき，これらのパラメータがどのように分布するかを可視化する．

（3） 調音フィルタのパラメータを用いて楽器同定が可能かどうかを SVM を用いて調べる．

5.1 実験データ

単旋律の楽器推定に関する共通データベースがないために複数の商用 CD や単音データベースから学習サンプルとテストサンプルを用意した．楽器は 5 種類の楽器（ビオラ，フルート，ホルン，トランペット，チェロ）を考え，これら 5 種類の楽器の演奏を含む商用 CD を各楽器 4 枚づつとアイオワ大学の単音データベース [14] を音源とした．これらのデータのサンプリング周波数はすべて 44.1kHz である．

5.2 スペクトル前処理

それぞれの楽器の音源からランダムに約 1 秒（サンプリング点：40960 点）のサンプルを 60 個抜き出す．無音区間や音量の小さいものはあらかじめ除いたため，そのようなサンプルは含まれていない．5 種類の楽器を考えているので，合計 300 サンプルあることになる．それらを半分に分けて学習データとテストデータとする．各サンプルを 20 フレームに分割し，各フレームごとに離散フーリエ変換することで振幅スペクトルを算出する．1 フレームは，約 50 ミリ秒である．通常は，高周波よりも低周波の情報が楽器判別に重要であると言われているため，今回の実験では 1-11020Hz の周波数のみのスペクトル情報を使用した．振幅スペクトルを 10Hz ごとに平滑化を行い，1102 次元

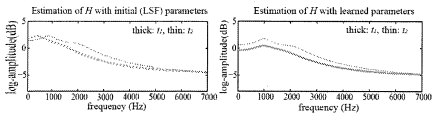

図 1　時刻 t_1（太線）と t_2（細線）の学習前（LSF 係数，左図）と学習後（右図）のパラメータから再構成したスペクトル包絡の一例

のスペクトルベクトルを取得とし，対数振幅を考えるために対数をとり，観測されるスペクトルベクトル s_t とした．

5.3　実験 1：ソースとフィルタの推定

調音フィルタが楽器音の時間に依存しない情報を保持していると仮定しているため，調音フィルタによるスペクトル包絡は楽器ごとに等しくなるべきである．各サンプルの 20 フレーム内では，調音フィルタのパラメータは同一としているが，サンプルが異なるときにこの調音フィルタが一致する保障はない．音高に依存せず，同一の調音フィルタが存在すると仮定した提案モデルによる学習によって，異なるサンプル間でも類似した調音フィルタが得られるかどうかを調べた．学習は 3 節で示した自由エネルギーを滑降シンプレックス法を用いて最小化することで行う．初期パラメータについては，b_n は LSF 係数とし，A と τ は各楽器個別に手で与えた．図 1 はトランペットの異なるサンプル点から，ある時刻 t_1（太線）と t_2（細線）におけるスペクトル包絡を示したものであり，左図が LSF 係数から再構成したスペクトル包絡，右図が学習後のパラメータから再構成したスペクトル包絡である．

時刻 t_1 と t_2 の二つのスペクトル包絡を比較すると，提案モデルによって生成されたスペクトル包絡どうしの方が，初期パラメータの包絡よりもお互いに近い形状をとっている．t_1 と t_2 のスペクトル包絡上のピークのうち最も低い周波数上にあるピークはそれぞれ学習前 500Hz，900Hz から学習後は共に 1000Hz 付近に移動している．図 2 は，学習前（上図）と学習後（下図）の音源信号と調音フィルタによるスペクトル包絡を示したものである．左図が時刻 t_1，右図が時刻 t_2 に対応する．

時刻 t_1，t_2 のそれぞれの学習前のスペクトル包絡は，音源信

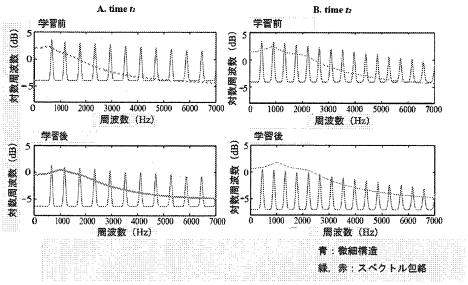

図 2　学習前と学習後の t_1（A）と t_2（B）の音源信号と調音フィルタ

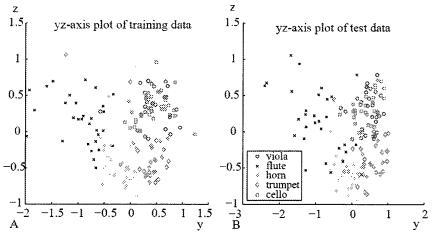

図 3　学習データ（A）とテストデータ（B）の LFDA による 12 次元 LSF 学習後パラメータから 3 次元空間への射影した yz 軸．

号の倍音成分に影響され，倍音周波数上にピークがある．しかし倍音周波数は，音高で変わるために時間変化する．このような場合でも，ダイナミカルシステムに基づくソースフィルタモデルによって学習したパラメータを用いれば，同楽器ではほぼ一致する調音フィルタを得ることがわかる．

5.4　実験 2：LFDA によるパラメータの次元削減と可視化

調音フィルタを特徴付けるパラメータが楽器の特徴表現能力を保ったまま，パラメータを次元圧縮できかどうかを考える．また，次元圧縮をおこなうことで可視化を可能にする．代表的な次元圧縮の方法として線形判別分析（linear discriminant analysis; LDA）と主成分分析（principal component analysis; PCA）がある [15]．これらは線形変換をおこなう手法であるが，前者は識別がしやすい特徴空間へ変換するために教師ラベルを用いた教師あり学習，後者は教師ラベルを用いない教師なし学習である．楽器判別は分類問題であるために LDA の次元削減が適切にみえるが，LDA は特徴次元が識別ラベル数に比べて少ないときや，ある一つの識別ラベルに対する入力データ分布のモードが複数ある際には次元圧縮がうまく働かない [16]．

このことから，本実験では局所フィッシャー判別分析（linear discriminant analysis; LFDA）を採用する．LFDA は LDA や PCA と同様に線形射影をおこなう教師あり次元圧縮法であるが，局所的なクラスタを考えることでマルチモーダルな分布に対応している点で LDA と異なる．図 3 は，提案モデルにより学習した 12 次元の学習後の調音フィルタパラメータを，LFDA によって 3 次元に圧縮し，クラスごとの分布が一番明確であった yz 軸を示したものである．学習データとテストデータの y 軸のスケールは見やすさのために変えてある．この図から，わずか 2 次元の表示であっても各楽器がクラスタを形成しており，また，弦楽器，管楽器どうしが近い場所にクラスタを形成することがわかった．

5.5　実験 3：楽器推定

楽器推定には 12 次元の学習前と学習後の調音フィルタパラメータ，LFDA によって 6，3 次元に次元圧縮を行ったパラメータを用いた．提案手法と従来手法を SVM によって判別した結果を表 1 に示す．同じデータセットを用いた比較ではないために

必ずしも的確な比較とは言えないが，参考のために従来手法の結果も掲示する．この表から，学習前の LSF パラメータが従来

	楽器数	特徴数	判別精度 (%)
初期パラメータ	5	12	84.67
(6 次元)	5	6	83.33
(3 次元)	5	3	71.33
学習後のパラメータ	5	12	87.33
(6 次元)	5	6	82.67
(3 次元)	5	3	74
Marques, 1999 [12]	8	16	70
Eggink, 2003 [17]	5	120	66
Livshin, 2004 [18]	7	62	88
Jinachitra, 2004 [19]	5	28	66
Essid, 2006 [20]	10	70	87

表 1　楽器推定結果の比較

手法と同程度，もしくはより優れていることだけでなく，提案手法で用いた特徴パラメータの数は従来手法よりもはるかに少ないにも関わらず，提案手法によるモデルパラメータの学習によってさらに良い楽器識別が可能なパラメータが得られることがわかった．また，LFDA によって 6 次元に圧縮してもほとんど楽器判別精度は変わらず，3 次元に次元圧縮しても 70%以上の判別率を保っていることがわかった．

6. ま と め

本報告では，音高と音量の時間変化情報を考慮した音源特性と調音フィルタの同時推定をするシステム同定方法を提案した．連続的なダイナミクスを取り入れた確率モデルを構築し，パラメータは自由エネルギーの最小化によって求めた．調音フィルタの初期パラメータは LSF で求め，自由エネルギーを小さくするようなパラメータ値を推定した．また，学習したモデルパラメータを用いて，楽器を推定することで，パラメータの評価をおこなった．

初期モデルパラメータである LSF でも少ない次元で十分に楽器推定が出来ることがわかったが，提案手法で学習したパラメータを用いた場合の楽器推定精度の方が高くなることがわかった．それに加えて，LFDA によってパラメータ次元を落としても，精度が劣化しないことも示された．

このモデルを幅広い楽器音判別に適用するためには少なくとも二つの課題がある．一つ目は，このモデルは多旋律や音声信号には直接適用できないため，観測過程を修正する必要があることである．二つ目の課題は，計算量の問題の解決である．提案手法のパラメータ推定にはシンプレックス法が用いられているが，収束までに時間がかかる．今後，多旋律へ適用できるようにするためのモデルの拡張と，勾配に基づく非線形最適化手法の採用による収束の改善を行いたいと考えている．

文　献

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

[2] 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).

[3] 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.

[4] K. Ito 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).

[5] 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).

[6] Z. Yuan, "The weighted sum of the line spectrum pair for noisy speech," Master's thesis, Helsinki University of Technology, 2003.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

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

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

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

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

[17] 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.

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

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

[20] 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.