Vol.2010-MUS-86 No.20
2010/7/29

多重奏音響信号中の演奏をユーザー指定の旋律に差し替えるフレーズ置換システム

安良岡 直希[†1] 糸山 克寿[†1] 吉岡 拓也[†2]
高橋 徹[†1] 駒谷 和範[†1,*1]
尾形 哲也[†1] 奥乃 博[†1]

フレーズ置換とは，多重奏音響信号から特定パート演奏をユーザー指定の別楽譜による演奏に差し替えるものである．これは，1) 元々のフレーズ演奏成分を除去する音源分離の課題と，2）元演奏の音色や演奏表情を新しい演奏上で再現する演奏合成の課題からなる．我々は調波非調波 Gaussian Mixture Model (GMM) による置換対象演奏モデルと Nonnegative Matrix Factorization による伴奏モデルを用いて音源分離を行い，同時に調波非調波 GMM から得た基本周波数，倍音強度などの音響特徴を新しい演奏楽譜の MIDI 音源音響信号に転写することで元演奏の音響特性を持つ新しい演奏を合成する．本フレーズ置換法に対し 1) 元の演奏が正しく除去されるか，2) 新しい演奏は元演奏の特徴を保持しているか，の 2 点を客観評価し，提案法の有効性を示す．

Phrase Replacing System for Polyphonic Music Waveforms

NAOKI YASURAOKA,[†1] KATSUTOSHI ITOYAMA,[†1]
TAKUYA YOSHIOKA,[†2] TORU TAKAHASHI,[†1]
KAZUNORI KOMATANI,[†1,*1] TETSUYA OGATA[†1]
and HIROSHI G. OKUNO [†1]

This paper presents a music manipulating system that enables a user to replace an instrument performance phrase in polyphonic audio mixture. Two technical problems must be solved to realize this system: 1)separating the melody part from accompaniment, and 2)synthesizing a new instrument performance that has timbre and expression of the original one. Our method first performs the separation using statistical model integrating harmonic and inharmonic Gaussian mixture and nonnegative-matrix-factorization. Then our method synthesizes a new instrument performance by adding the acoustic characteristics given by Gaussian mixture parameters to a MIDI synthesizer-generated sound. Two evaluations confirm the effectiveness of the proposed method.

1. はじめに

近年，専門知識や設備を持たない一般の人々が作曲・楽器演奏などを行い創作したコンテンツ (Consumer Generated Media: CGM などと呼ばれる) を web 等で公開する事例が急増している．これは，MIDI 音源など従来高価なハードウェアとして提供されていたものが比較的安価なソフトウェアとして利用可能になった他，歌唱音声合成ソフト[1)] などの新しいツールの登場により，音楽の制作・編曲の敷居が下がったためである．特に，既存楽曲のギター演奏を真似て自分の演奏音を重ねるなど二次創作，三次創作を楽しむユーザーが増えている．もし，市販 CD のような多旋律・多重奏のオーディオデータをユーザーが自由に改変できる技術が実現すれば，オリジナルのギター演奏を消去し自分の演奏に差し替えたり，既存曲からお気に入りのドラムフレーズだけを抽出し自作曲に混ぜたりと，単なるオーディオデータの切り貼りを超えた楽曲編集が可能になる．

我々は，フレーズ置換という全く新しい楽曲編集技術を報告する．フレーズ置換とは，市販 CD のような複数楽器パートが混在する音響信号に対し，i) ある楽器パートのフレーズ演奏成分を混合音中から除去し，ii) 代わりにユーザーが指定する別の楽譜に対応する演奏を合成する，ことで演奏フレーズを差し替えるものである．この技術は，例えばソロ演奏の別バージョンを作ることで，バンドのライブパフォーマンスにおいてギタリストが CD 収録のものとは全く異なるソロを披露する状況を個人的に再現することができる．すなわち，プロ，アマチュアを問わずポストプロダクションや二次創作の幅を飛躍的に広げる可能性を持つ．

フレーズ置換の主たる技術的課題は次の 2 点である．

(1) **音源分離：元々のフレーズ演奏を除去し，同時に演奏の音色，演奏表情を抽出**

(2) **演奏合成：元演奏の音色，演奏表情を新しい演奏上で再現**

ここで演奏表情とは，人間の実演奏が持つ音量や発音タイミング，音色等の揺らぎを指す．以下，これらの課題の詳細について，関連研究を引用して説明する．

(1) フレーズ置換のための音源分離法への要求として，単一パートを最小の事前情報で分

†1 京都大学 大学院情報学研究科
Kyoto University
†2 NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories
*1 現在，名古屋大学大学院工学研究科
Presently with Graduate School of Engineering, Nagoya University

離できることと，演奏合成時に用いる音色や演奏表情情報が同時に得られること，の二点がある．楽器単音のスペクトルモデルを立てて観測信号を分配する方法[2]は比較的正確に目的のフレーズ演奏を抽出できるものの，楽曲中の全てのパートの楽譜が事前情報として必要である．Independent Component Analysis[3]や Nonnegative Matrix Factorization (NMF)[4]に基づく音源分離法は，事前情報がほぼなくても良好な分離結果を得られるものの，分離結果が求めたい要素信号の単位と必ずしも一致しない．単一パート抽出をテーマとした研究[5],[6]も近年報告されているが，分離結果から音色や演奏表情を抽出するには後処理が必要である．

(2) 合成される演奏は「同一演奏者による演奏だ」と思えるよう元演奏の音色や演奏表情を保持している必要がある．楽譜からの演奏音響信号の合成自体は MIDI 音源を用いれば容易であるが，楽譜情報だけでは元演奏の情報を一切反映できない．そのため，楽譜構造と演奏表情のパターンが対応するという仮定の下，実演奏データからの統計学習や推論により未知楽譜に適切な演奏表情を付加する演奏表情付けと呼ばれる手法が研究されている[7]．しかし，一般の演奏表情付け法は MIDI 出力機能付き楽器の演奏から得る音量・発音タイミング情報の操作に基づき，音響的特性は一切扱われていない．一方，楽器音スペクトルモデルにより実演奏音響信号を直接分析し，モデルから新たな演奏音響信号を再合成する手法[8]も存在するが，伴奏や残響を含む演奏に対する分析結果は真の楽器音スペクトルに対して誤差があり，モデルからの音響信号再合成は音質に限界があった．

本稿では調波非調波 Gaussian Mixture Model と呼ぶ音モデルを用い，この 2 つの課題を同時解決するフレーズ置換の実現方法を報告する．調波非調波 GMM とは楽器音パワースペクトルの調波構造・非調波構造をそれぞれガウス関数の和で表現するモデルであり，基本周波数，倍音強度など重要な音響的特徴を直接表すパラメータによって楽器音を高精度に表現する．この楽器音モデルに基づいた音源分離により，高い精度で伴奏音からの分離ができると同時に元演奏の音高や倍音強度，非調波成分強度といった情報が得られるため，これを演奏表情と見なし合成する演奏に反映させるという処理でフレーズ置換を一つの大きな枠組みで定式化できる．より具体的には，以下のアイデアによって前述の 2 課題に対処する．

(1) **調波非調波 GMM と NMF モデルを組み合わせた音源分離**
置換対象演奏を表わす調波非調波 GMM と伴奏を表わす NMF に基づくスペクトルモデルの重ね合わせによって多重奏音響信号をモデル化し 1 パートの分離を実現する．調波非調波 GMM は楽譜情報を用いたパラメータの初期化を要するが，置換対象演奏のみをこれでモデル化し，一方伴奏音は NMF を用いて，伴奏が持つであろう少数の要素パターンの組み合わせという特徴を事前情報なしに表現する．この方法は事前情報が少ない点と分離結果から音響特徴を得やすい点とを兼ね備えており従来法[2),4)–6)]よりもフレーズ置換に適している．

(2) **MIDI 音源スペクトルへの調波非調波 GMM 特徴量の転写**
予め MIDI 音源で合成した演奏音響信号に対し，そのスペクトルを元演奏から推定された調波非調波 GMM パラメータが示す音色特徴に合わせて操作する．この方法は，元演奏が持つ音色・演奏表情を合成する演奏に反映されられるとともに，調波非調波 GMM が分離歪みやモデル化誤差を含む問題を緩和することができる．

2. 問題定義と手法の概要

本章では，フレーズ置換の具体的な問題定義と，これを解決するため提案するアルゴリズムの大まかな流れを示し，それが音源分離と演奏合成から構成されることを確認する．

2.1 問 題 定 義

フレーズ置換とは，市販 CD 等複数楽器パートが混在する音響信号から特定パートを除去し，代わりにユーザー指定の別楽譜による演奏を付加するものである．以後本論文では，もともと音響信号に含まれ，置換されることになるオリジナルの演奏を参照演奏，またユーザーが指定した代わりの楽譜に対応する演奏を置換演奏と呼ぶ．システムへの入力は，a) 操作元となるモノラル多重奏音響信号, b) 除去される参照演奏に対応する楽譜, c) 置換演奏に対応する楽譜 の三点である．楽譜は具体的には Standard MIDI File (SMF) の発音・消音タイミング及び音高情報とし，音響信号と時間の同期がとれているとする．

2.2 手法の概要

フレーズ置換手法は**図 1** のように，前章で述べた課題にそれぞれ対応する (1) 音源分離ステップと (2) 演奏合成ステップを直列に接続した形で構成される．

音源分離ステップでは参照演奏を多重奏音響信号から除去するとともに，参照演奏の音響特徴や演奏表情を抽出する．入力の多重奏音響信号の短時間フーリエ変換を $x_{n,f}$ とする．ここで，n は時間フレーム，f は周波数ビンである．参照演奏の除去とはすなわち $x_{n,f}$ から置換対象演奏のスペクトル成分を取り除くことである．伴奏音 $a_{n,f}$ と参照演奏音 $m_{n,f}$ のパワースペクトルを推定することができれば (それぞれの推定値を $\hat{A}_{n,f}$ 及び $\hat{M}_{n,f}$ とする), Wiener フィルターにより音響信号成分を分離し，伴奏成分 $\bar{a}_{n,f}$ を得られる．

$$\bar{a}_{n,f} = f^{(\mathcal{R})}\left(\frac{\hat{A}_{n,f}}{\hat{A}_{n,f}+\hat{M}_{n,f}} f^{-(\mathcal{R})}(x_{n,f})\right) \tag{1}$$

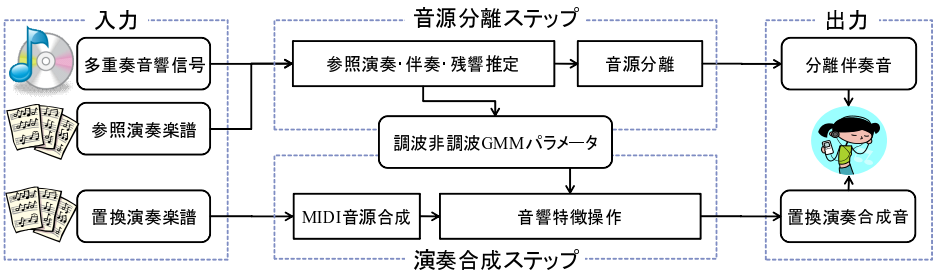

図 1 フレーズ置換の概要

ここで，$f^{(-\mathcal{R})}$ 及び $f^{(\mathcal{R})}$ はそれぞれ残響を除去する作用素と，逆に付加する作用素である．これらは，参照演奏の真のスペクトル成分 $m_{n,f}$ だけでなく，その残響成分も併せて除去する目的で導入され，既存の残響抑圧法[9] を統合することでこれを実現する．またこのとき，$\hat{M}_{n,f}$ が何らかの数理モデルで表されていれば，そのモデルパラメータ $\theta^{(\mathrm{m})}$ は参照演奏の音色・演奏表情を示す音響特徴と見なすことができる．

演奏合成ステップではユーザー指定の新たな演奏を合成する．前述のとおり，演奏合成の際に重要なことは元演奏の音色や演奏表情を新しい演奏上で再現することである．そこで，置換演奏の MIDI 音源合成音 $\dot{m}_{n,f}$ に $\theta^{(\mathrm{m})}$ が規定する音響特徴を反映させる処理を施し (その作用素を $f^{(\mathcal{S})}$ とする)，参照演奏の音色・演奏表情を持った合成演奏を得る．この置換演奏合成音に残響を足し戻し，最後に伴奏信号と加算すればフレーズ置換結果 $z_{n,f}$ が得られる．

$$z_{n,f} = f^{(\mathcal{R})}\left(f^{(\mathcal{S})}(\dot{m}_{n,f}, \theta^{(\mathrm{m})})\right) + \bar{a}_{n,f} \tag{2}$$

以上の処理に基づけば，フレーズ置換の実現のために必要なことは，伴奏音と参照演奏音のパワースペクトルのモデル，及び各作用素を具体的に定義し，それらを特徴付けるパラメータ θ を推定するアルゴリズムを設計することであると言える．

3. 音源分離ステップ

本章では，入力音響信号の複素スペクトログラム $x_{n,f}$ を生成する統計モデルを設計し，その最尤推定の結果を用いて音源分離を行う方法について述べる．

3.1 統計モデルの設計

伴奏音に相当する要素信号と参照演奏に相当する要素信号の重ね合わせ (音源信号と呼ぶ) が，自己回帰システムによって表現される残響の影響を受けて観測されるという仮定を置き，入力スペクトルの密度関数を導く．まず，以下の仮定を置く．

仮定 1. 音源信号を構成する参照演奏音 $m_{n,f}$ 及び伴奏音 $a_{n,f}$ は各時間フレーム，各周波数ビンにて独立に平均 0，分散 $M_{n,f}$，$A_{n,f}$ の複素正規分布 $\mathcal{N}_{\mathrm{C}}$ に従う．

$$m_{n,f} \sim \mathcal{N}_{\mathrm{C}}(0, M_{n,f}), \quad a_{n,f} \sim \mathcal{N}_{\mathrm{C}}(0, A_{n,f}) \tag{3}$$

なお $M_{n,f}$，$A_{n,f}$ はパワースペクトル密度 (PSD) と対応する．さらに，参照演奏 $m_{n,f}$ と伴奏信号 $a_{n,f}$ は互いに独立とする．このとき，正規分布に従う独立な確率変数の和正規分布に従い，その分散は各々の分布の分散の和となるから，

$$s_{n,f} = m_{n,f} + a_{n,f} \sim \mathcal{N}_{\mathrm{C}}(0, S_{n,f}), \qquad S_{n,f} = M_{n,f} + A_{n,f} \tag{4}$$

が成り立つ．$s_{n,f}$ は二つの要素信号の和であり，$S_{n,f}$ がその分散に対応する．

仮定 2. 観測信号 $x_{n,f}$ は音源信号 $s_{n,f}$ によって駆動される D 次の自己回帰システムによって生成される．これは[9] に基づく残響のモデル化である．

$$x_{n,f} = \sum_{d=1}^{D} g_{d,f} x_{n-d,f} + s_{n,f} \tag{5}$$

ここで，$g_{d,f}$ は第 f 周波数ビンの第 d 自己回帰係数であり，本稿では残響フィルタと呼ぶ．$s_{n,f}$ は線形予測の予測誤差の形で導かれる．

続いて，参照演奏のパワースペクトル密度 $M_{n,f}$，及び伴奏音のパワースペクトル密度 $A_{n,f}$ を，それぞれの PSD が持つであろう特性や，利用可能な事前知識，及び演奏合成処理への連携のしやすさを考慮して具体的に定義する．参照演奏のモデルには，楽器音 PSD を比較的正確に表現できる調波非調波 GMM を用いる．これは糸山らの音源分離法[2] を参考に設計し，**図 2** に示すように楽器音 PSD を，調波構造に対応する分散の小さい GMM (調波 GMM) と，非調波構造に対応する分散の大きい GMM (非調波 GMM) の線形混合で表す．

$$M_{n,f} = \sum_{j=1}^{J}\left(\sum_{k=1}^{K} H_{j,k,n,f} + \sum_{l=1}^{L} I_{j,l,n,f}\right) \tag{6}$$

$$H_{j,k,n,f} = \frac{u_{j,k,n}}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(\omega_f - k\mu_{j,n})^2}{2\sigma^2}\right], \quad I_{j,l,n,f} = \frac{v_{j,l,n}}{\sqrt{2\pi\gamma^2}} \exp\left[\frac{-(\omega_f - \nu_l)^2}{2\gamma^2}\right] \tag{7}$$

ここで，j は何番目の単音に対応する調波非調波 GMM であるかを示すインデックスであり，パラメータ $\theta^{(\mathrm{m})} = \{w_{j,n}^{(H)}, w_{j,n}^{(I)}, u_{j,k,n}, v_{j,l,n}, \sigma^2, \mu_{j,n}\}_{1\le j\le J, 1\le k\le K, 1\le l\le L, 0\le n\le N-1, 0\le f\le F-1}$ はそ

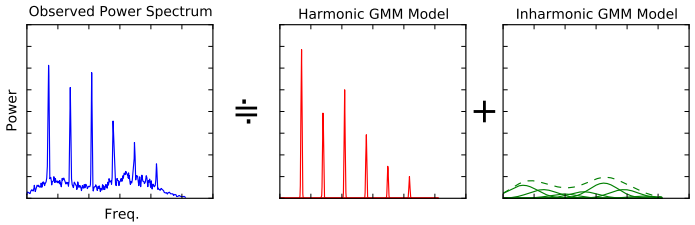

図 2 調波非調波 GMM: 楽器音パワースペクトルを調波構造用の GMM と非調波構造用の GMM の和で表現する

れぞれ第 j 音の調波成分パワー，非調波成分パワー，倍音相対強度，非調波サブバンド相対強度，調波構造ピーク分散，基本周波数に対応する．ω_f は周波数ビンから対応する Hz への写像であり，残りの ν_l，γ^2 は非調波サブバンドの形状を決める定数である．表現の任意性を避けるため，$\forall j,n:\sum_{k=1}^{K} u_{j,k,n}=1, \forall n:\sum_{l=1}^{L} v_{j,l,n}=1$ とする．このモデルは，基本周波数や倍音強度などの音響特徴をモデルパラメータとして持つため，推定結果を演奏合成ステップに簡単に利用できる．欠点として，尤度関数に多数の極値を持ち初期値依存性が高い点が挙げられるが，事前知識として楽譜情報から音高初期値を与えこの問題を軽減する．

伴奏音のモデル化には，少数の要素パターンの組み合わせと繰り返しからなるデータの表現に適した NMF モデルを用いる．なぜなら，伴奏音は一般的に和音の構成音を奏でることが多く，それは周波数軸上で少数の音高に集中するからであり，またリズムトラックは少数の打楽器音から成るからである．NMF は，少数の要素の組み合わせで構成される二次元データを圧縮する手法であり，n 行 f 列の値が $A_{n,f}$ である非負行列 $\mathrm{A}\in\mathbb{M}^{N\times F}$ をより小さい行列 $\mathrm{U}\in\mathbb{M}^{N\times C}$ と $\mathrm{V}\in\mathbb{M}^{C\times F}$, $(C\ll N,F)$ で表す．

$$\mathrm{A}=\mathrm{UV} \tag{8}$$

このとき，U は C 個の周波数基底を纏めた行列と見ることができ，V が各基底の各時間フレームにおけるアクティベーションに対応する．基底数 C によってモデルの自由度が決まり，適切な値を設定することで同じパターンの組み合わせという伴奏音の特徴を表現できる．$\theta^{(\mathrm{a})}=\{\mathrm{U},\mathrm{V}\}$ が推定すべきパラメータである．

以上より入力スペクトログラム $\mathbb{X}=\{x_{n,f}\}_{0\le n\le N-1,0\le f\le F-1}$ に対するモデルの尤度関数は

$$p(\mathbb{X};\theta)=\prod_{n=0}^{N-1}\prod_{f=0}^{F-1}\frac{1}{2\pi S_{n,f}}\exp\left(-\frac{|x_{n,f}-\sum_{d=1}^{D}g_{d,f}x_{n-d,f}|^2}{S_{n,f}}\right) \tag{9}$$

と具体的に決まる．残響に関するパラメータ $\{g_{d,f}\}_{1\le d\le D,0\le f\le F-1}$ を $\theta^{(\mathrm{g})}$ とおくと，この尤度関数は未知パラメータ $\theta=\{\theta^{(\mathrm{m})},\theta^{(\mathrm{a})},\theta^{(\mathrm{g})}\}$ からなり，最尤推定の枠組みで各パラメータを推定することで，二種類の要素信号の分離と残響の推定を一つのシンプルな枠組みで行う事ができる．また，この密度関数の負の対数をとると，以下に示す板倉-斎藤歪尺度と等価なコスト Q の最小化問題となっていることが分かる．

$$Q=\sum_{n=0}^{N-1}\sum_{f=0}^{F-1}\left(\log S_{n,f}+\frac{|x_{n,f}-\sum_{d=1}^{D}g_{d,f}x_{n-d,f}|^2}{S_{n,f}}\right) \tag{10}$$

3.2 EM アルゴリズムを用いたパラメータ推定

式 (10) の局所最小値を与える参照演奏・伴奏・残響モデルのパラメータ θ の最尤推定は，残響フィルタと音源 PSD の反復更新と，EM アルゴリズムに基づく音源 PSD 内の伴奏 PSD，参照演奏 PSD の反復更新という二つの反復アルゴリズムからなる．

参照演奏モデルと伴奏モデルのパラメータは各 PSD: $M_{n,f}$，$A_{n,f}$ を完全データと見なして EM アルゴリズム[10] により推定することができる．詳細な導出は[11] を参照されたいが，EM アルゴリズムを式 (10) に対して適用すると，以下の Q 関数の反復最小化問題に帰着する．

$$Q=-\sum_{n=0}^{N-1}\sum_{f=0}^{F-1}\left[\left(\log M_{n,f}+\frac{\Psi^{(\mathrm{m})}_{n,f}}{M_{n,f}}\right)+\left(\log A_{n,f}+\frac{\Psi^{(\mathrm{a})}_{n,f}}{A_{n,f}}\right)\right] \tag{11}$$

ただし，$\Psi^{(\mathrm{m})}_{n,f}$ 及び $\Psi^{(\mathrm{a})}_{n,f}$ はそれぞれ伴奏 PSD 及び参照演奏 PSD の推定値であり，

$$\Psi^{(\mathrm{m})}_{n,f}=M_{n,f}\left[1+\frac{M_{n,f}(|x_{n,f}-\sum_{d=1}^{D}g_{d,f}x_{n-d,f}|^2-S_{n,f})}{{S_{n,f}}^2}\right] \tag{12}$$

$$\Psi^{(\mathrm{a})}_{n,f}=A_{n,f}\left[1+\frac{A_{n,f}(|x_{n,f}-\sum_{d=1}^{D}g_{d,f}x_{n-d,f}|^2-S_{n,f})}{{S_{n,f}}^2}\right], \tag{13}$$

と導出される．式 (11) は各モデルと PSD 推定値との間の板倉-斎藤歪尺度の和となっている．E-ステップではこの $\Psi^{(\mathrm{m})}_{n,f}$ と $\Psi^{(\mathrm{a})}_{n,f}$ を算出し，M-ステップで $\theta^{(\mathrm{m})},\theta^{(\mathrm{a})}$ を更新する．

参照演奏 PSD 推定値 $\Psi^{(\mathrm{m})}_{n,f}$ は調波非調波 GMM パラメータ $\theta^{(\mathrm{m})}$ の更新に用いられる．ここで本来行うべきことは式 (11) に示す板倉-斎藤歪尺度を最小化するように $\theta^{(\mathrm{m})}$ を更新することであるが，本稿では以下の Kullback-Leibler (KL) 歪尺度に基づいて更新を行う．

$$Q'=\sum_{n=0}^{N-1}\sum_{f=0}^{F-1}\Psi^{(\mathrm{m})}_{n,f}\log\frac{\Psi^{(\mathrm{m})}_{n,f}}{M_{n,f}} \tag{14}$$

この方法を採る理由は，板倉-斎藤距離が調波非調波 GMM の推定に適しないからである．板倉-斎藤歪尺度は観測パワースペクトルを下回らないようモデルを適合する傾向があり，周波数方向に分散の広い非調波 GMM が観測スペクトルのトップエンベロープに適応し調波構造推定に失敗することがある．KL 歪尺度を用いる場合，Q' の最小化にさらに EM アルゴリズムを用いることができ，例えば基本周波数 $\mu_{j,n}$ は M ステップにて以下の式で更新される．

$$\mu_{j,n} = \frac{\sum_{k=1}^{K}\sum_{f=0}^{F-1} k\omega_f \Psi^{(\mathrm{H})}_{j,k,n,f}}{\sum_{k=1}^{K}\sum_{f=0}^{F-1} k^2 \Psi^{(\mathrm{H})}_{j,k,n,f}} \tag{15}$$

ここで, $\Psi^{(\mathrm{H})}_{j,k,n,f}$ は E ステップで求められる第 j 音の k 次高調波成分のパワー推定値である．

$$\Psi^{(\mathrm{H})}_{j,k,n,f} = \frac{H_{j,k,n,f}}{\sum_{j=1}^{J}\left(\dot{w}^{(H)}_{\lambda,n}\sum_{k=1}^{K} H_{j,k,n,f} + \dot{w}^{(I)}_{\lambda,n}\sum_{l=1}^{L} I_{j,l,n,f}\right)} \Psi^{(\mathrm{m})}_{n,f} \tag{16}$$

板倉斉藤歪尺度と KL 歪尺度は等価ではないが, 推定時にパラメータが大きく振動したりはせず，フレーズ置換全体してはより良い結果となることを実験的に確認した．

伴奏 PSD 推定値は NMF の各行列 U 及び V の更新に使われる．板倉-斎藤歪尺度に基づく NMF は，乗法更新則[4] と呼ばれる以下の更新式を反復適用することで現在の伴奏 PSD 推定値 $\Psi^{(\mathrm{a})}_{n,f}$ を要素に持つ行列 $\Psi^{(\mathrm{A})}$ に対する局所最適値を得ることができる．

$$\mathrm{U} = \mathrm{U}\cdot\frac{\mathrm{V}^{\mathrm{T}}\left((\mathrm{UV})^{\cdot -2}\cdot\Psi^{(\mathrm{A})}\right)}{\mathrm{V}^{\mathrm{T}}(\mathrm{UV})^{\cdot -1}}, \qquad \mathrm{V} = \mathrm{V}\cdot\frac{\left((\mathrm{UV})^{\cdot -2}\cdot\Psi^{(\mathrm{A})}\right)\mathrm{U}^{\mathrm{T}}}{(\mathrm{UV})^{\cdot -1}\mathrm{U}^{\mathrm{T}}} \tag{17}$$

ここで X^{T} は行列の転置を，$\cdot$ は各要素毎の演算を示す．すなわち $X\cdot X$ は要素積，$X^{\cdot z}$ は各要素の z 乗を表す．

音源パラメータの更新後，そのモデルが表す PSD を固定し今度は残響フィルタの更新を行う．これは従来法[9] と同様，以下の式によって行われる．

$$\begin{pmatrix} g_{1,f} \\ \vdots \\ g_{D,f} \end{pmatrix} = \begin{pmatrix} \sum_{n=0}^{N-1}\frac{x^*_{n+1,f}x_{n+1,f}}{S_{n,f}} & \cdots & \sum_{n=0}^{N-1}\frac{x^*_{n+1,f}x_{n+D,f}}{S_{n,f}} \\ \vdots & \ddots & \vdots \\ \sum_{n=0}^{N-1}\frac{x^*_{n+D,f}x_{n+1,f}}{S_{n,f}} & \cdots & \sum_{n=0}^{N-1}\frac{x^*_{n+D,f}x_{n+D,f}}{S_{n,f}} \end{pmatrix}^{-1} \begin{pmatrix} \sum_{n=0}^{N-1}\frac{x^*_{n+1,f}x_{n,f}}{S_{n,f}} \\ \vdots \\ \sum_{n=0}^{N-1}\frac{x^*_{n+D,f}x_{n,f}}{S_{n,f}} \end{pmatrix} \tag{18}$$

3.3 参照演奏除去の実行

以上のパラメータ推定の終了後，得られた各要素信号の PSD 推定結果 $\hat{M}_{n,f}$，$\hat{A}_{n,f}$ を用いて式 (1) を適用することで伴奏演奏の直接音に対応するスペクトログラムを $\bar{a}_{n,f}$ 得ることができる．ここで，残響を除去する関数 $f^{-(\mathcal{R})}$ は，残響フィルタの推定結果 $\hat{g}_{d,f}$ を用いた畳み込み $f^{-(\mathcal{R})}(x_{n,f}) \equiv x_{n,f} - \sum_{d=1}^{D} g_{d,f}x_{n-d,f}$ とし，残響を付加する関数 $f^{-(\mathcal{R})}$ は残響抑圧結果から観測信号 $x_{n,f}$ を得るような Finite Impulse Response フィルタを改めて推定し用いる．ただし，Wiener フィルターによって得た複素スペクトログラム $\bar{a}_{n,f}$ の位相は実在する時間領域信号に対応しないため，位相修正法[12] を残響付加前に適用することで音質劣化を防ぐ．

4. 演奏合成ステップ

本章では，フレーズ置換の後半に相当する，ユーザー指定の楽譜に対する演奏音響信号を合成し伴奏と足し合わせる処理について述べる．これは，類似する楽譜構造は類似する音色・演奏表情で演奏されるという仮定のもと，1) 合成する演奏楽譜の各単音にふさわしい調波非調波 GMM パラメータの算出，2)MIDI 演奏スペクトルへの調波非調波 GMM パラメータが示す音響特徴の転写，という手順で行われる．

4.1 連続 2 音の楽譜構造の類似性に基づくモデルパラメータ算出

本手法では，類似性を計る楽譜情報としてノートナンバーと音長を用いて，参照演奏のモデルパラメータ $\theta^{(\mathrm{m})}$ から置換演奏第 λ 音に対する調波非調波 GMM パラメータ ψ_λ を算出する．まず，合成する演奏の各単音 λ に対して，その前後関係がもっとも類似している二音を参照演奏中から以下の条件式により選出する．

$$q^-_\lambda = \operatorname*{argmin}_j \sum_{p=-1,0}\left(\left|\dot{\xi}_{\lambda+p} - \xi_{j+p}\right| + \alpha\left|\dot{\eta}_{\lambda+p} - \eta_{j+p}\right|\right) \tag{19}$$

$$q^+_\lambda = \operatorname*{argmin}_j \sum_{p=0,1}\left(\left|\dot{\xi}_{\lambda+p} - \xi_{j+p}\right| + \alpha\left|\dot{\eta}_{\lambda+p} - \eta_{j+p}\right|\right) \tag{20}$$

ここで，ξ_j，η_j は参照演奏の，$\dot{\xi}_\lambda$，$\dot{\eta}_\lambda$ は置換演奏のノートナンバーと音長であり，α はそれらの重みを操作する定数である．次に，得られた二つの単音のモデルパラメータを補間して，第 λ 音にふさわしい音モデルを算出する．置換演奏第 λ 音のモデルパラメータ中の時間フレーム n に対する部分を $\psi_{\lambda,n}$ と表すとすると，

$$\psi_{\lambda,n} = \begin{cases} \frac{\dot{\eta}^+_\lambda - n}{\dot{\eta}_\lambda}\theta^{(\mathrm{m})}_{q^-_\lambda,n}, + \frac{n-\dot{\eta}^-_\lambda}{\dot{\eta}_\lambda}\theta^{(\mathrm{m})}_{q^+_\lambda,n}, & \dot{\eta}^-_\lambda \le n \le \dot{\eta}^+_\lambda \\ 0, & \text{otherwise} \end{cases} \tag{21}$$

とする．ただし，$\theta^{(\mathrm{m})}_{q^-_\lambda,n}$，$\theta^{(\mathrm{m})}_{q^+_\lambda,n}$ をそれぞれ参照演奏の第 q^-_λ，q^+_λ 音のモデルパラメータを音

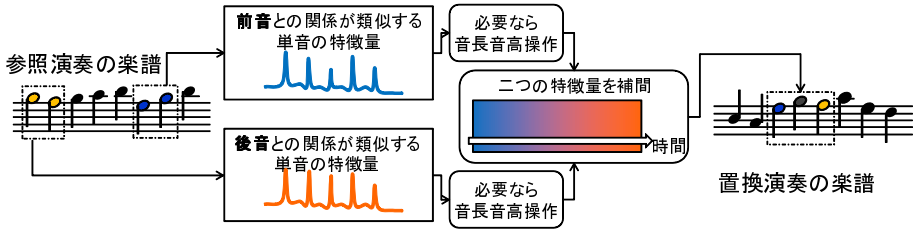

図 3 隣接二音の楽譜構造の類似性を用いて置換演奏の各単音に調波非調波 GMM パラメータを算出する

高が $\dot{\xi}_\lambda$，音長が $\dot{\eta}_\lambda$ となるように補間伸縮をしたものとし，その四則演算は各パラメータ同士のものと定義する．また $\dot{\eta}_\lambda^-$ 及び $\dot{\eta}_\lambda^+$ は第 λ 音の楽譜上の発音・消音時刻に対応する時間フレームである．この式は二つの音モデルパラメータの混合比を 1:0 から 0:1 へと時間変化させることを意味しており，$q_\lambda^+ = q_{\lambda+1}^-$ であることから，**図 3** 実演奏中で隣り合った音の組を合成する演奏の楽譜に合わせて次々と滑らかに連結させていく操作となる．

4.2 MIDI 演奏スペクトルの分離・重み操作

置換演奏各単音の調波非調波 GMM パラメータが得られたら，MIDI 音源を用いて合成した演奏音響信号のスペクトルをモデルパラメータに合わせて変形することにより演奏音響信号を合成する．この方法は，MIDI 演奏音響信号を直接用いる方法や，前節で求めた調波非調波 GMM が表すパワースペクトルをそのまま用いる方法に対して，元演奏が持つ音色・演奏表情を合成結果に反映されられると同時に，調波非調波 GMM の示すスペクトルが分離歪みやモデル化誤差を含む問題を緩和できる．

MIDI 音源から合成した演奏音響信号を $\dot{m}_{n,f}$ とすると，ここから楽譜情報を元に 3.2 節と全く同じ方法で各単音ごとの調波非調波 GMM モデル $\{\dot{w}_{\lambda,n}^{(H)}, \dot{w}_{\lambda,n}^{(I)}, \dot{H}_{\lambda,k,n,f}, \dot{I}_{\lambda,l,n,f}, \dot{\mu}_{\lambda,n}\}$ を得ることができる．さらに，その推定結果を用いて MIDI 演奏音響信号を第 λ 発音の第 k 調波成分 $\hat{H}_{\lambda,k,n,f}$，及び第 l 非調波サブバンド成分 $\hat{I}_{\lambda,l,n,f}$ に分離することができる．

$$\hat{H}_{\lambda,k,n,f} = \frac{\dot{H}_{\lambda,k,n,f}}{\dot{M}_{n,f}} \left|\dot{m}_{n,f}\right|^2, \quad \hat{I}_{\lambda,l,n,f} = \frac{\dot{I}_{\lambda,l,n,f}}{\dot{M}_{n,f}} \left|\dot{m}_{n,f}\right|^2 \tag{22}$$

$$\dot{M}_{n,f} = \sum_{\lambda=1}^{\Lambda} \left(\dot{w}_{\lambda,n}^{(H)} \sum_{k=1}^{K} \dot{H}_{\lambda,k,n,f} + \dot{w}_{\lambda,n}^{(I)} \sum_{l=1}^{L} \dot{I}_{\lambda,l,n,f} \right) \tag{23}$$

なお，これは EM アルゴリズム中の E ステップにおいて推定する各ガウス関数への分離 PSD そのものである．音色・演奏表情の反映は，これら分離スペクトルを，楽譜から推定されたモデルパラメータ ψ を元に重み付けることで実現される．具体的には，ψ 中の変数 $\tilde{w}_{\lambda,n}^{(H)}$, $\tilde{w}_{\lambda,n}^{(I)}$, $\tilde{u}_{\lambda,k,n}$, $\tilde{v}_{\lambda,l,n}$, $\tilde{\mu}_{\lambda,n}$ を用いて，次式より各分離スペクトルの強度と位置を変えたパワースペクトル $Y_{n,f}$ を得る．

$$Y_{n,f} = \sum_{\lambda=1}^{\Lambda} \left(\sum_{k=1}^{K} \frac{\tilde{w}_{\lambda,n}^{(H)} \tilde{u}_{\lambda,k,n}}{\dot{w}_{\lambda,n}^{(H)} \dot{u}_{\lambda,k,n}} \hat{H}_{\lambda,k,n,f}^{+} + \sum_{l=1}^{L} \frac{\tilde{w}_{\lambda,n}^{(I)} \tilde{v}_{\lambda,l,n}}{\dot{w}_{\lambda,n}^{(I)} \dot{v}_{\lambda,l,n}} \hat{I}_{\lambda,l,n,f} \right) \tag{24}$$

ただし，$\hat{H}_{\lambda,k,n,f}^{+}$ は $\hat{H}_{\lambda,k,n,f}$ を周波数方向に $(\tilde{\mu}_{\lambda,n}/\dot{\mu}_{\lambda,n})$ 倍に伸縮したスペクトルであり，これは音高を操作することに相当する．

4.3 音響信号の再構成

合成された置換演奏パワースペクトル $Y_{n,f}$ に MIDI 演奏音響信号 $\dot{m}_{n,f}$ の位相 $\angle\dot{m}_{n,f}$ を付加して，式 (2) 中の演奏合成結果の複素スペクトルが $f^{(S)}(\dot{m}_{n,f}, \theta^{(\mathrm{m})}) \equiv \sqrt{Y_{n,f}}\angle\dot{m}_{n,f}$ と具体的に求められる．ただし，このスペクトログラムについても位相修正を行う．

5. 評 価 実 験

フレーズ置換の能力は，1) 音源分離ステップにて参照演奏が正しく除去されるか，2) 演奏合成ステップにて置換演奏は実際の演奏者によるものと思えるように合成されているか，の二点によって決定付けられると考えられる．本章では，提案手法の有効性を検証するためにこの二点のそれぞれに着目した二つの評価実験について述べる．なお，二つの実験で共通する条件として，音響信号のサンプリング周波数は 44.1kHz，STFT 解析の窓関数は 1024 点ガウス関数，シフト幅は 256 点としている．

5.1 音源分離実験

5.1.1 実験目的と条件

第一の実験では，参照演奏が正しく除去されるかの評価として，音源分離ステップに対し，提案法「メロディパートに調波非調波 GMM，伴奏パートを NMF に基づくモデルを用いた音源分離」の評価を行った．提案する音源分離方法は，置換対象パートの楽譜のみ利用可能という条件下に適するように設計したため，以下の音源分離法との比較を行う．

(1) 参照演奏に対するモデルも NMF に基づくもの：楽器音の調波構造などを仮定しないため，楽譜情報が利用できるというメリットを十分に活かせない方法と見なせる．ただし問題の困難さを揃えるため，参照演奏楽譜の MIDI 音源音響信号を NMF によって分析した結果をこの NMF の初期値とする．

表 1 参照演奏除去実験にて用いた楽曲

ジャンル	曲番号と参照演奏とみなした楽器パート			
Jazz	#22	Trumpet	#33	Flute
	#24	Alto Sax	#34	Flute
	#32	Vibraphone	#41	Alto Sax
Classical	#12	Flute	#37	Violin
	#13	Cello	#39	Violin
	#16	Clarinet	#42	Harp

表 2 12 曲の音源分離結果の対数スペクトル距離平均 (小さいほど良い)

伴奏 \ 参照演奏	HIGMM	NMF
NMF10	**9.81**	13.68
NMF50	10.16	15.00
NP	10.34	11.14

(2) 伴奏演奏に対するモデルがノンパラメトリック (NP と呼ぶ) であるもの：伴奏のモデル化に NMF を用いた動機である「同じパターンの組み合わせ」といった制約を持たず，式 (12) で得られる PSD 推定値をそのままモデルの PSD と見なす．NMF の基底数を十分に大きくしたものとも言える．

参照演奏のモデルに対し提案法の調波非調波 GMM と比較法の NMF モデルの 2 種類，また伴奏演奏のモデルに対し基底の数 C を 10, 50 に設定した提案法の NMF モデルと比較法の NP モデルの 3 種類，それらの組み合わせで 6 種類の音源分離法を実践する．評価データは RWC Music Database: Jazz Music and Classic Music[13] の Jazz と Classic それぞれ 6 曲ずつ計 12 曲の SMF から，モデルの初期化に用いたものとは別の MIDI 音源を用いて合成した音響信号を用い，各曲の旋律演奏楽器パートを分離対象とした (**表 1** 参照)．なお，本実験では音源分離の能力のみに着目するため，MIDI 音源音響信号には残響を付加させず，残響フィルタパラメータ $\theta^{(\mathrm{g})}$ はすべて 0 で固定した．

分離の良し悪しは，処理対象の楽器パートを予め除いた SMF から合成された音響信号 $\mathbb{E} = \{e_{n,f}\}_{0\le n\le N-1, 0\le f\le F-1}$ を真値とし，伴奏音推定信号に相当する $\bar{\mathbb{A}} = \{\bar{a}_{n,f}\}_{0\le n\le N-1, 0\le f\le F-1}$ との間の対数スペクトル距離 $\mathrm{LSD}(\mathbb{E}, \bar{\mathbb{A}})$ によって評価する．

$$\mathrm{LSD}(\mathbb{E}, \bar{\mathbb{A}}) \equiv \sqrt{\sum_{n=0}^{N-1}\sum_{f=0}^{F-1}(20\log_{10}\left|\frac{e_{n,f}}{\bar{a}_{n,f}}\right|)^2 \Big/ NF} \tag{25}$$

値が小さい方が両信号が類似しており，0 のとき両信号は一致する．

5.1.2 実験結果と考察

図 2 に真値 (混合前の音響信号) と分離結果との間の対数スペクトル距離を全曲で平均したものを示す．各手法を比べると，調波非調波 GMM による参照演奏モデル・NMF による伴奏モデルの本手法が最も良い結果となっており，本手法が「分離消去したいパートの楽譜のみ利用可能」という条件下での音源分離に適していることが示されている．参照演奏モデルを NMF とした比較法では，曲目ごと・伴奏音 NMF 基底数ごとに対数スペクトル距離のばらつきが確認され，全体として提案法より劣る結果となった．ところで，双方に NMF を用いた比較法は Nonnegative Matrix Partial Co-Factorization (NMPCF)[5] と酷似している．これは，観測データの NMF と同時に reference データ (例えばメロディパート楽譜の MIDI 音響信号) の NMF を行い，周波数基底のうち一部を共有することで reference データに含まれる信号とよく似た信号成分を観測データから抽出するものである．NMPCF は周波数基底の共有が容易なドラム音の分離に有効という報告があるが，フレーズ置換で対象とする周波数依存性の強い旋律演奏に対しては調波非調波 GMM の方が効果的であることが示唆される．

本手法に残された課題としては，伴奏音 NMF の基底数を適切に設定する必要がある点が挙げられる．本実験では，MIDI 音源音響信号を用い，また Jazz と Classic という特定のジャンルのみを用いたため，結果では概ねどの曲でも基底数 10 のものが最良という結果となったが，伴奏音がより複雑に変化する楽曲ではより大きな基底数が適する可能性があり，任意の楽曲に対して最良の分離結果を得るには自動で最適な基底数を選択できるような枠組みが必要である．

5.2 演奏合成実験

5.2.1 実験目的と条件

第二の実験ではフレーズ置換の後半部分に相当する演奏合成ステップに対し，提案する演奏合成法が出力する演奏音響信号が「音響的に」どれほど実演奏に近いかを検証する．具体的には，単旋律の実演奏に対し，各曲の後ろ 4/5 の演奏音響信号と楽譜を用い実演奏の調波非調波 GMM パラメータ及び残響フィルタパラメータを取得し，これを用いて合成した前 1/5 の楽譜に対する演奏音響信号が実演奏とどれほど近いかを評価する実験を行う．提案法の新しさは，MIDI 演奏音響信号に参照演奏から取得した音色・演奏表情特徴を付加させるという方法により音質を損なわずに参照演奏の特徴反映を実現する点にある．そのため，以下に示す 4 つの演奏合成法を比較することによって提案法の有効性を検証する．

(1) Ours: 提案法
(2) Baseline1: 演奏合成ステップで算出する調波非調波 GMM が示すパワースペクトルから直接音響信号を合成する
(3) Baseline2: 式 (24) の代わりに音量のみを操作するスペクトル操作式

$$Y_{n,f} = \frac{\sum_{\lambda=1}^{\Lambda}(\tilde{w}^{(H)}_{\lambda,n} + \tilde{w}^{(I)}_{\lambda,n})}{\sum_{\lambda=1}^{\Lambda}(\hat{w}^{(H)}_{\lambda,n} + \hat{w}^{(I)}_{\lambda,n})}\left|\hat{m}_{n,f}\right|^2 \tag{26}$$

表 3 9 曲，4 種の MIDI 音源による演奏合成結果の対数スペクトル距離平均 (小さいほど良い)

MIDI 音源 \ 合成法	Ours	Baseline1	Baseline2	MIDI
A	**9.91**	10.74	12.24	12.14
B	**9.85**	10.71	12.60	13.14
C	**9.85**	10.70	12.48	13.81
D	**10.01**	10.67	13.97	16.60

を用いる．音量のみの演奏表情付けに相当する．

(4) MIDI: MIDI 音源による合成音響信号 $\grave{m}_{n,f}$ そのまま

本実験では演奏合成の能力にのみ着目するため，無伴奏演奏を用いまた伴奏モデルは導入しない．評価データは市販 CD 収録のプロによる演奏:Violin (VN)，Flute (FL)，Cello (VC) 各 3 曲の計 9 曲を用いた．これらはいずれもコンサートホールで収録したと思われる残響時間 1 秒前後の長い残響が含まれており，提案モデルの残響フィルタの長さ D は予備実験を元に 80 とした．合成の良し悪しは，置換演奏合成結果と元の実演奏との間の式 (25) による対数スペクトル距離の小ささをもって評価する．

5.2.2 実験結果

表 3 に，4 種類の MIDI 音源 A,B,C,D でパラメータ初期値設定と合成を行った際の合成演奏と実演奏の間の対数スペクトル距離を示す．いずれの音源の場合でも，提案法が最も良い結果となっている．これは，演奏合成において音色に関わる音響的特徴を再現することの有効性と，MIDI 音源からのスペクトル操作法が合成品質の点で優れていることを示している．

また 4 つの音源の比較として，表 3 中「MIDI」列の値のばらつきに対し，「Ours」列の値のばらつきが小さいことから，提案法による合成音は MIDI 音源の音の傾向に強くは依存しないことが分かる．従って，実演奏に近い合成演奏を得るために MIDI 音源を選択・調整する手間は小さいと言え，専門知識のない一般ユーザーが利用する状況に対して都合が良いと考えられる．

6. おわりに

本論文では，特定楽器パートのフレーズ演奏をユーザー指定の楽譜によるものに差し替えるフレーズ置換と呼ばれる全く新しい楽曲編集技術について報告した．提案法では調波非調波 GMM と呼ばれる楽器音スペクトルモデルを導入し，NMF に基づく伴奏モデルの併用により音響信号から元々の演奏成分を除去するとともに，調波非調波 GMM が示す音響特徴の MIDI 音源演奏への転写により元演奏の音色・演奏表情を合成演奏に転写した．

今後の課題には，5.1 節で述べたように NMF 最適基底数を自動決定することの他に，演奏表情付けの手法をより上級なものへと改善することが挙げられる．ユーザー指定の楽譜に対応する調波非調波 GMM パラメータを算出する方法について，実演奏の多様な変化を十分に学習・生成できるように，現在の隣接二音しか着目していないパラメータ算出法を改める必要がある．

謝辞: 本研究の一部は，グローバル COE プログラム，科研費 S，科学技術振興機構 CrestMuse プロジェクトによる支援を受けた．

参考文献

1) 剣持秀紀，大下隼人：歌声合成システム VOCALOID，情報処理学会研究報告 [音楽情報科学] 2007-MUS-72，pp.25–28 (2007).
2) Itoyama, K., Goto, M., Komatani, K., Ogata, T. and Okuno, H.G.: Parameter Estimation for Harmonic and Inharmonic Models by Using Timbre Feature Distributions, *IPSJ Journal*, Vol.50, No.7, pp.1757–1767 (2009).
3) Casey, M. and Westner, A.: Separation of Mixed Audio Sources by Independent Subspace Analysis, *Proc. ICMC*, pp.154–161 (2000).
4) Smaragdis, P. and Brown, J.: Non-negative Matrix Factorization for Polyphonic Music Transcription, *Proc. WASPAA*, pp.170–180 (2003).
5) Yoo, J., Kim, M., Kang, K. and Choi, S.: Nonnegative Matrix Partial Co-Factorization for Drum Source Separation, *Proc. ICASSP*, pp.1942–1945 (2010).
6) Smaragdis, P. and Mysore, G.: Separation By "Humming": User-Guided Sound Extraction From Monophonic Mixtures, *Proc. WASPAA* (2009).
7) 平賀瑠美，平田圭二，片寄晴弘：蓮根 : めざせ世界一のピアニスト，情報処理，Vol.43, No.2, pp.136–141 (2002).
8) Yasuraoka, N., Abe, T., Itoyama, K., Takahashi, T., Ogata, T. and Okuno, H.G.: Changing Timbre and Phrase in Existing Musical Performances as You Like, *Proc. ACM Multimedia*, pp.203–212 (2009).
9) Yoshioka, T., Nakatani, T. and Miyoshi, M.: An Integrated Method for Blind Separation and Dereverberation of Convolutive Audio Mixtures, *Proc. EUSIPCO* (2008).
10) Feder, M. and Weinstein, E.: Parameter Estimation of Superimposed Signals Using the EM Algorithm, *IEEE Trans. Acoust. Speech, Signal Process.*, Vol.36, No.4, pp.477–489 (1988).
11) Kameoka, H. and Kashino, K.: Composite Autoregressive System for Sparse Source-Filter Representation of Speech, *Proc. ISCAS*, pp.2477–2480 (2009).
12) Griffin, D.W. and Lim, J.S.: Signal Estimation from Modified Short-Time Fourier Transform, *IEEE Trans. Acoust. Speech, Signal Process.*, Vol.32, No.2, pp.236–243 (1984).
13) Goto, M., Hashiguchi, H., Nishimura, T. and Oka, R.: RWC Music Database: Popular, Classical, and Jazz Music Databases, *Proc. ISMIR*, pp.287–288 (2002).