

フレーズ置換のための調波非調波 GMM・NMF に基づく音源分離・演奏合成

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

本論文では、多重奏音響信号中の特定の楽器パート演奏をユーザ指定の別楽譜による演奏に差し替える「フレーズ置換」システムのための音源分離・演奏合成法について報告する。本システムはまず上記特定の楽器パート演奏（置換元演奏と呼ぶ）を多重奏から分離除去し（音源分離）、次にユーザが指定した楽譜の演奏を合成し多重奏に挿入する（演奏合成）。自然なフレーズ置換のために、合成される演奏には置換元演奏の特徴を反映させる。本システムの技術的課題は、1) 置換元演奏楽譜のみを用いた音源分離、2) 置換元演奏特徴を持つ置換先演奏の高品質合成、の2点である。この課題に対処するため、次の2点に基づく音源分離・演奏合成法を設計した：1) 調波非調波 Gaussian Mixture Model (GMM) と Nonnegative Matrix Factorization (NMF) の統合モデルによる置換元演奏と伴奏の音源分離、2) MIDI 音源が合成した演奏への音色・演奏表情補正。本手法に対し i) 置換元演奏が正しく除去されるか、ii) 合成演奏は置換元演奏の特徴を保持しているか、の2点を客観評価した結果、それぞれ比較対象に対し 28.2%、11.5%対数スペクトル距離が改善された。

Musical Sound Separation and Synthesis Using Harmonic/Inharmonic GMM and NMF for Phrase Replacing System

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

This paper presents a sound separation and synthesis method for a new music manipulating system that facilitates a user to replace an instrument per-

formance phrase in polyphonic audio mixture. The system first separates the instrument part from polyphony using the original performance score, and then synthesizes a new instrument performance of the user-specified target score, keeping sound characteristics of the original one. Two technical problems must be solved to realize this system: 1) separating one instrument part without knowledge of the other parts, and 2) synthesizing a new performance from separated sound with high sound quality. We introduce a new sound separation and synthesis method for the phrase replacing system; 1) sound separation by harmonic/inharmonic Gaussian mixture and nonnegative-matrix-factorization, and 2) sound synthesis by modifying MIDI-synthesizer-generated sound to follow estimated timbre and expression of original performance. Two evaluations confirm the effectiveness of our method. The method separates the target part more accurately by 28.2% in log spectral distance, and synthesizes instrument performance more accurately by 11.5% in comparison with conventional methods.

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 などで公開する事例が急増している。この理由として、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ことに加え、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、楽曲の制作・編曲の敷居が下がった点があげられる。特に、既存楽曲のギター演奏を真似た自分の演奏音を元の楽曲に重ねるといった2次創作、3次創作を楽しむユーザが増えている²⁾。ただし、現状の普及技術では、既存楽曲の1区間を切り出したり、自分の演奏音を重ねたりするなどの編集にとどまる。もし、市販CD中の多旋律・多重奏の音響信号をユーザが自由に操作できる技術が実現できれば、楽曲からオリジナルのギター演奏だけを分離除去し自分の演奏に差し替える、既存曲からお気

^{†1} 京都大学大学院情報学研究所

Graduate School of Informatics, Kyoto University

^{†2} 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所

NTT Communication Science Laboratories, NTT Corporation

*1 現在、ヤマハ株式会社

Presently with Yamaha Corporation

*2 現在、名古屋大学大学院工学研究科

Presently with Graduate School of Engineering, Nagoya University

入りのドラムフレーズだけを抽出し自作曲に混ぜる、といったより自由な多重奏楽曲編集が可能になる。

我々の研究の目的は、多重奏音響信号中に含まれる特定パート演奏をユーザが指定した別楽譜による演奏に差し替える「フレーズ置換」システムの実現である。具体的には、フレーズ置換システムは、多重奏音響信号と置換したい楽器パートの元々の演奏『置換元演奏』の楽譜、および差し替える先の演奏『置換先演奏』の楽譜という3入力を受け取り、入力音響信号中の置換元演奏だけが置換先演奏に置き換えられた音響信号を出力する。フレーズ置換は音源分離と演奏合成の2行程から構成されると考えられる。まず置換元演奏の楽譜を用いて、置換元演奏を多重奏から分離除去し、次に置換先演奏の楽譜を用いて置換先演奏の音響信号を合成し、これを置換元演奏除去後の多重奏に挿入する。

自然なフレーズ置換のためには、合成される演奏は、置換元演奏の演奏特徴を反映することが重要である。ここで、演奏特徴とは、楽器固有の音色や残響などの音響的特徴、および演奏表情（楽譜に依存した音量などの変動）を指す。以下本論文では、新たに合成された演奏が置換元演奏特徴を保持していることを『演奏同一性を満たす』と表現する。

以上をまとめると、フレーズ置換を実現するには下記の2つの基本的な課題を解決する必要がある。

(課題1) 置換元演奏の楽譜のみを用いた音源分離と演奏特徴推定

(課題2) 演奏特徴推定の誤差に頑健な演奏合成

従来研究されてきた方法だけでは、これらの課題を十分に解決することができない。課題1に関連する方法として、Itoyamaらによって提案された音源分離法がある³⁾。この方法は多重奏音響信号に含まれる個々の楽器パートの信号を分離できる。しかし全楽器パートの楽譜が補助情報として必要でありまた計算時間が大きいという問題がある。フレーズ置換のような単一パート操作を目的とする場合、全楽器パートの楽譜がなくても効率的な操作が可能な手法が望まれる。課題2に関する方法には安部らによって提案された音合成法がある⁴⁾。この方法は、音色特徴を表現する楽器音モデルを用いて楽器音を分析・合成する。しかし、多重奏から推定された演奏特徴には推定誤りが不可避免的に含まれるため、この方法を直接用いることは有効でない。なおこれら2課題の一部ないしすべては、フレーズ置換システムに限らず多くの多重奏音響信号操作に共通するものであり、これを解決することは音楽情報処理分野において大きな意義がある。

本論文では上記の課題を解決する音源分離・演奏合成法を報告する。課題1については、調波非調波 Gaussian Mixture Model (GMM) と Nonnegative Matrix Factorization

(NMF)⁵⁾を統合した音源分離法を提案する。調波非調波 GMM は Itoyama らの音源分離法を参考に設計したモデルであり、これを置換元演奏のモデルとして用いる。一方、置換元演奏以外については楽譜情報が不要な NMF を用いたモデルを用いる。これによって、置換元演奏の楽譜だけを使って効率的に置換元演奏を分離抽出できるようにする。課題2については、MIDI音源で演奏音響信号を合成し、演奏同一性を満たすよう音色・演奏表情を補正するという演奏合成法を提案する。この方法は安部らの方法と異なり、置換元演奏推定結果からの直接再合成を避けることで分析歪みの影響を軽減できる。

以下、まず2章でフレーズ置換の問題定義を詳細に述べるとともに、解決すべき課題とその対処法の概要を説明する。3章では提案するフレーズ置換のための音源分離・演奏合成法の概要を述べる。4章では音源分離部、5章では演奏合成部についてそれぞれ詳細を述べる。6章で音源分離、演奏合成の各々について評価を行い本手法の有効性を示す。最後に7章で本論文のまとめを述べる。

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を述べ、その後前章で述べた課題を解決するための音源分離・演奏合成法のアイデアを示す。

2.1 問題定義

フレーズ置換とは、多重奏音響信号中の特定の楽器パート演奏をユーザ指定の別楽譜による演奏に差し替える処理である。具体的な手順は、i) 音源分離：多重奏からの置換元演奏の分離除去および置換元演奏特徴の推定、ii) 演奏合成：置換元演奏の特徴を反映した置換先演奏の合成、の2行程である。音源分離部と演奏合成部は図1のように組み合わせる。システムへの入力は、a) 操作元となる多重奏のモノラル音響信号、b) 除去される置換元演奏

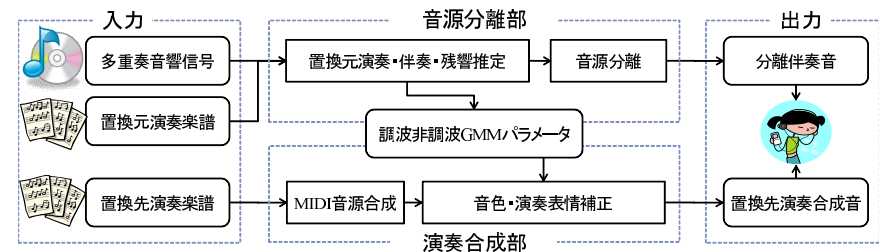


図1 フレーズ置換の概要

Fig. 1 Overview of the phrase replacing system.

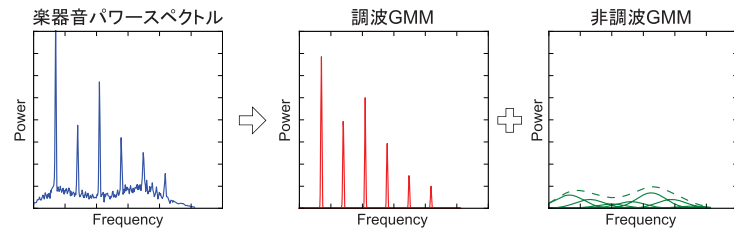


図 2 調波非調波 GMM: 楽器音パワースペクトルを調波構造用の GMM と非調波構造用の GMM の和で表現する
Fig.2 Harmonic/inharmonic GMM: the power spectrum of the instrument sound is represented as the sum of two GMMs for harmonic and inharmonic structures.

に対応する楽譜, c) 置換先演奏に対応する楽譜, の 3 点である。本論文の範囲では, 置換元演奏は調波楽器による主旋律かつ単旋律の演奏パートに限定し, 打楽器パートや歌唱パートは対象としない。また多重奏中の置換元演奏以外の楽器パートを本論文では伴奏と呼ぶ。楽譜は Standard MIDI File (SMF) の発音・消音時刻および音高情報とし, 音響信号と時間の同期がとれているとする。なお, SMF の時間情報を音響信号と同期させる課題は別途研究⁶⁾ されており, 本研究では扱わない。

2.2 調波非調波 GMM

本研究では音源分離・演奏合成の実装に Itoyama らの音源分離法³⁾ を参考に設計した調波非調波 GMM と呼ぶ楽器音モデルを用いる。調波非調波 GMM は図 2 のように, 楽器音の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る。このモデルは調波楽器音のパワースペクトルを少ないパラメータで精度良く近似できるため, 音源分離の際に置換元演奏のパワースペクトルを推定するのに利用される。また, モデルパラメータが基本周波数や倍音強度などの音響特徴と対応することから, 演奏合成の際には置換元演奏から推定した調波非調波 GMM が示す音響特徴を持つ置換先演奏音響信号を合成することで演奏同一性を保持できると期待される。倍音強度や非調波成分の各周波数帯域ごとの強度 (以後単に非調波成分強度と呼ぶ) が音色に対応し, 音量や基本周波数の変動情報が演奏表情に対応する。

2.3 解決すべき課題と解決法

ここではフレーズ置換システムの技術的課題とその解決法について述べる。以下に示す 2 点は, いずれも既存の音源分離・演奏合成法の組合せでは解決困難な課題と考えられる。

課題 1. 置換元演奏の楽譜のみを用いた音源分離と演奏特徴推定 置換元演奏 (すなわち一

部の楽器パート) の楽譜のみが利用可能という条件下では調波非調波 GMM のみに基づく音源分離は必ずしも効果的ではない。なぜなら, 調波非調波 GMM のパラメータをうまく推定するにはすべての楽器パートの楽譜などを用いておおよその音高情報を推定初期値として与える必要があるからである。実際, Itoyama らが報告している多重奏音響信号の各楽器パートごとの分離手法では, 伴奏を含む全楽器パートの楽譜を用いている。また本論文の問題では伴奏中の各単音それぞれの推定は不要であるのですべての楽器パートの演奏を分離する Itoyama らの方法は, 冗長である。一方楽譜情報がない条件下での音源分離法としては NMF⁷⁾ や Independent Component Analysis⁸⁾ に基づくものがあるが, これらは分離された各信号が各楽器パートと一般には対応しないことから, 置換元演奏の演奏特徴の推定は容易ではない。

解決策 1. 調波非調波 GMM と NMF の統合モデルによる音源分離 調波非調波 GMM によって置換元演奏をモデル化し, 一方 NMF によって伴奏をモデル化する。これにより置換元演奏だけの音源分離と演奏特徴推定を効率的に実現する。Itoyama らの手法と異なり, 置換元演奏に対してのみ調波非調波 GMM を用いることで, 置換元演奏の楽譜のみから音源分離を効果的に行える。一方 NMF のみを用いる音源分離法と比べると, 置換元演奏の楽譜情報を活用できるため, 置換元演奏を高精度に分離できる。さらに, 調波非調波 GMM のパラメータは, 演奏合成に直接利用できる。

課題 2. 演奏特徴推定の誤差に頑健な演奏合成 演奏同一性を満たす演奏合成法としては, 置換元演奏推定結果から何らかの方法を用いて置換先演奏に対する調波非調波 GMM パラメータを算出し, そのパワースペクトルに適当な位相を与え逆フーリエ変換する方法が考えられるが, 多くの場合, その品質はただちに合成音と判断できる程度にとどまる。なぜなら, モデルと実際のパワースペクトルの差異, すなわちモデル化誤差が存在するからである。モデル化誤差を回避するには, モデルの自由度を上げ楽器音のパワースペクトルをより精緻に表現する楽器音モデルの利用が考えられるが⁴⁾, その場合はモデルパラメータ推定がより困難になり, 伴奏や残響を含む多重奏からの推定に起因する誤差, すなわち推定誤差を生じやすくなる。

解決策 2. MIDI 音源合成演奏への音色・演奏表情補正に基づく演奏合成 MIDI 音源により合成された演奏音響信号を, 調波非調波 GMM が示す音色・演奏表情に基づき補正するという新しい演奏合成法を用いる。これにより調波非調波 GMM が示すパワースペクトルから音響信号を直接再合成することを避けられるので, モデル化誤差・推定誤差の影響を緩和できる。以下, MIDI 音源が合成した演奏音響信号のことを単に

『MIDI 音源合成演奏』と呼ぶ。演奏合成の具体的手順は次のとおりである。まず、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下、置換先演奏楽譜の各単音に期待される調波非調波 GMM パラメータを算出する。次に、置換先演奏楽譜に対する MIDI 音源合成演奏を用意し、そのパワースペクトルを上で計算された調波非調波 GMM パラメータに合わせて補正する*1。

2.4 残響への対応

残響は音色の知覚に深く関わるので、演奏同一性を保持するためには、置換元演奏と同様の残響を置換先演奏にも付加しなければならない。実際、残響を加えるエフェクターがエレキギタリストの間で一般的に用いられており、残響が演奏者がコントロールしたい情報の 1 つであると考えられる。

そこで提案手法に残響抑圧法⁹⁾を統合し、残響の推定を行う。具体的には提案音源分離法に残響抑圧法を統合することで、置換元演奏と伴奏を分離するだけでなく、残響成分も同時に推定する。また、残響推定結果から残響重畳フィルタを設計し、これを置換先演奏合成結果に適用することで、演奏同一性のうちの残響特性の再現を試みる。これは残響に冠する演奏同一性を保持するための効果的な実現方法と考えられるが、他の方法として、たとえばユーザが試行錯誤しながら付加する残響を調節することも考えられる。なお本論文で扱う残響は、室内の残響やこれを模擬した残響エフェクトを想定しており、楽器胴体の共鳴は含まない。

3. フレーズ置換のための音源分離・演奏合成法の概要

本章では、前章で述べたアプローチに基づく音源分離・演奏合成法について述べる。

音源分離部では入力された多重奏音響信号から置換元演奏を分離除去するとともに、置換元演奏特徴を推定する。入力音響信号を短時間フーリエ変換することで得られる複素スペクトル成分を $x_{n,f}$ とする。ここで、 n は時間フレーム、 f は周波数ビンである。本研究では、 $x_{n,f}$ は置換元演奏複素スペクトル成分 $m_{n,f}$ と伴奏複素スペクトル成分 $a_{n,f}$ 、およびそれらの残響成分から構成されると見なす。ただし、 $m_{n,f}$ 、 $a_{n,f}$ はすべて未知である。非残響音という意味で $m_{n,f}$ と $a_{n,f}$ を合わせて音源信号と呼び、これを $s_{n,f} = m_{n,f} + a_{n,f}$ と表記する。本論文で述べるシステムでは、まず置換元演奏と伴奏のパワースペクトル成

*1 MIDI 音源への制御情報を調波非調波 GMM パラメータに基づき生成するわけではない。一般の MIDI 音源には倍音強度などを柔軟に制御する機能はなく、音色に関する演奏同一性保持は MIDI 音源合成演奏の操作が必要である。

分を推定する(それぞれの推定値を $\hat{M}_{n,f}$ および $\hat{A}_{n,f}$ とする)。次に、残響成分を取り除き、Wiener フィルタにより置換元演奏を除去することで伴奏の複素スペクトル成分推定値 $\bar{a}_{n,f}$ を得る。

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

ここで、 $f^{(-\mathcal{R})}$ は残響を除去する作用素である。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 $\theta^{(m)}$ は置換元演奏の音色・演奏表情に対応する演奏特徴と見なすことができ、同様に $f^{(-\mathcal{R})}$ が数理モデルで表されていれば、そのモデルパラメータは残響特性に対応する演奏特徴となる。

本研究では、伴奏を含むすべての楽器音に対して同一の $f^{(-\mathcal{R})}$ で残響が除去できるとする。これはすなわち、すべての楽器が同じ位置で演奏され、同じ残響成分を持つと仮定していることに等しい。多くの場合、同様の仮定は厳密には成立しないものの、ステレオオーディオのサラウンド化では、この仮定の下でも一定の効果があることが報告されている¹⁰⁾。よって、簡単のため、本研究でもこの仮定を用いる。

演奏合成部ではユーザ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置換元演奏との演奏同一性を保持することである。そのために、置換先演奏の MIDI 音源合成演奏 $\hat{m}_{n,f}$ を $\theta^{(m)}$ が示す音響特徴に合わせて補正する処理を施し(その作用素を $f^{(s)}$ とする)、置換元演奏の音色・演奏表情を持った合成演奏を得る。この置換先演奏合成音を伴奏信号に足し戻し、最後に残響を重畳し戻せばフレーズ置換結果 $z_{n,f}$ が得られる。

$$z_{n,f} = f^{(\mathcal{R})} (f^{(s)}(\hat{m}_{n,f}, \theta^{(m)}) + \bar{a}_{n,f}) \quad (2)$$

ただし、 $f^{(\mathcal{R})}$ は残響を重畳する作用素である。

以上より、フレーズ置換のための音源分離・演奏合成の実現には、置換元演奏と伴奏のパワースペクトルのモデル ($M_{n,f}$, $A_{n,f}$)、および各作用素 ($f^{(-\mathcal{R})}$, $f^{(s)}$, $f^{(\mathcal{R})}$) の具体的な定義、およびそれらの特徴付けるパラメータを推定するアルゴリズムの設計が必要である。

4. 音源分離部

本章では音源分離部の実現方法について述べる。まず 4.1 節で、置換元演奏のパワースペクトル成分 $M_{n,f}$ のモデルとそのパラメータ $\theta^{(m)}$ 、伴奏のパワースペクトル成分 $A_{n,f}$ の

モデルとそのパラメータ $\theta^{(a)}$, 残響のモデルとそのパラメータ $\theta^{(g)}$ を定義する. 4.2 節で, 入力された多重重音響信号 $x_{n,f}$ および置換元演奏の楽譜を用いて上記パラメータを推定する方法を述べる. 最後に 4.3 節で推定結果から置換元演奏を分離除去する方法について述べる.

4.1 モデルの設計

置換元演奏のパワースペクトル $M_{n,f}$ は, Itoyama らの音源分離法³⁾ を参考に設計した調波非調波 GMM を用いる. このモデルは, 楽器音のパワースペクトルを精度良く表現できるという特徴がある. 調波非調波 GMM では図 2 で示したように楽器音のパワースペクトルを, 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.

$$M_{n,f} = \sum_{j=1}^J \left(\sum_{k=1}^K H_{j,k,n,f} + \sum_{l=1}^L I_{j,l,n,f} \right) \quad (3)$$

$$H_{j,k,n,f} = \frac{u_{j,k,n}}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{(\Omega_f - k\mu_{j,n})^2}{2\sigma^2} \right], \quad I_{j,l,n,f} = \frac{v_{j,l,n}}{\sqrt{2\pi\gamma^2}} \exp \left[-\frac{(\Omega_f - \nu_l)^2}{2\gamma^2} \right] \quad (4)$$

表 1 に各インデックスと各変数の意味を示す. また Ω_f は f 番目の周波数ピンの周波数 (Hz) を表す. 推定すべきパラメータは $\theta^{(m)} = \{u_{j,k,n}, v_{j,l,n}, \mu_{j,n}\}_{1 \leq j \leq J, 1 \leq k \leq K, 1 \leq l \leq L, 0 \leq n \leq N-1}$ である. 調波非調波 GMM 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

表 1 調波非調波 GMM のインデックス・パラメータ一覧

Table 1 Indices and parameters for the harmonic/inharmonic GMM.

記号	意味	範囲
j	置換元演奏何番目の単音かを示すインデックス	$1 \leq j \leq J$
k	調波構造を表現するガウス関数のインデックス	$1 \leq k \leq K$
l	非調波構造を表現するガウス関数のインデックス	$1 \leq l \leq L$
$u_{j,k,n}$	調波 GMM の第 k 倍音に対応するガウス関数の大きさ (倍音強度)	
$v_{j,l,n}$	非調波 GMM の第 l 番目のガウス関数の大きさ (非調波成分強度)	
$\mu_{j,n}$	基本周波数	
ν_l	非調波 GMM の第 l ガウス関数の中心周波数 (周波数軸上で均等に並ぶ値で固定)	
σ^2	調波 GMM の各ガウス関数の分散: 倍音ピークの周波数方向の広がりに対応 (STFT 解析に応じた適正值で固定)	
γ^2	非調波 GMM の各ガウス関数の分散 (非調波 GMM が周波数軸方向に滑らかになる値で固定)	

として持つので, 推定結果を演奏合成部に簡単に利用できる. 一方, パラメータ初期値をうまく設定しないと理想的な推定結果が得られないという欠点がある. フレーズ置換では置換元演奏の楽譜が入力として与えられるため, これを用いて初期値を定めることで, 局所解の問題を軽減することができる.

なお調波非調波 GMM の定義について, 非調波構造モデルに GMM を用いずノンパラメトリックに扱ったり^{4),11)}, 調波構造のインハーモニシティ¹²⁾ を吸収できるよう拡張すること¹¹⁾ が可能であるが, 本論文ではこれらのモデル定義を避けている. なぜならそのような拡張はモデルの自由度が上がり楽器音をより正確に表現できるようになるものの, 伴奏を含む混合音からのパラメータ推定がより困難となるからである.

伴奏のパワースペクトル $A_{n,f}$ は NMF に基づきモデル化する. すなわち, 伴奏のパワースペクトログラム全体 $\{a_{n,f}\}_{0 \leq n \leq N-1, 0 \leq f \leq F-1}$ に対し, 時刻によらずに決定される C 個の基底 (パワースペクトルパターン) $V_{f,c} > 0$ を考え (c は基底のインデックス), 各時刻のパワースペクトルが $V_{f,c}$ の時変重み付き和で決定されるとする.

$$A_{n,f} = \sum_{c=0}^{C-1} V_{f,c} U_{c,n} \quad (5)$$

ただし, $U_{c,n}$ は $U_{c,n} > 0$ を満たす時変重みである. 基底数 C がモデルの自由度を決定し, 適切な値を設定すれば, 同じスペクトルパターンが繰り返し出現するという伴奏の特徴を表現できる. $\theta^{(a)} = \{V_{f,c}, U_{c,n}\}_{0 \leq c \leq C-1, 0 \leq n \leq N-1, 0 \leq f \leq F-1}$ が推定すべきパラメータである.

残響は既存の残響抑圧法^{9),13)} と同じ方法でモデル化する. すなわち観測信号 $x_{n,f}$ は音源信号 $s_{n,f}$ によって駆動される $D-d+1$ 次の自己回帰システムによって生成されるとする.

$$x_{n,f} = \sum_{\tau=d}^D g_{\tau,f} x_{n-\tau,f} + s_{n,f} \quad (6)$$

ここで, $g_{\tau,f}$ は第 f 周波数ピンの第 τ 自己回帰係数であり, 本論文では残響フィルタと呼ぶ. d および D はそれぞれ初期反射の遅れと残響時間に対応する. 本研究で推定すべき残響, すなわち室内の残響や残響エフェクトの効果は, 通常, 初期反射と後部残響に分けることができる. 後部残響とは, 複数の反射が重なり合っ, 個々の反射が区別できない状態を指す. 後部残響の開始までには一定の時間を要する. 一方, 楽器胴体での共鳴は発音の直後に生じるので, 共鳴のエネルギーの大部分は室内の後部残響の開始よりも短時間のうちに発生しやすい. よって, d よりも長時間持続する楽器の共鳴のエネルギーは後部残響のエネル

ギーに比べて無視できるほど小さいと仮定すると、適切な d を選ぶことで室内の残響と楽器胴体の共鳴を区別しやすくなると考えられる。このことは、音声残響除去の文脈で、声道と室内残響の分離の観点から議論されている¹⁴⁾。残響に関する推定すべきパラメータは $\{g_{\tau,f}\}_{d \leq \tau \leq D, 0 \leq f \leq F-1}$ であり、これを $\theta^{(g)}$ とおく。

4.2 パラメータの推定方法

次に、多重奏音響信号の複素スペクトログラム $\{x_{n,f}\}_{0 \leq n \leq N-1, 0 \leq f \leq F-1}$ と置換元演奏の楽譜が与えられたときに、置換元演奏、伴奏、および残響の各パラメータ $\theta^{(m)}$ 、 $\theta^{(a)}$ 、 $\theta^{(g)}$ を推定する方法について述べる。

4.2.1 基本的な考え方

本論文では、パワースペクトルの加法性の仮定の下で、これらのパラメータが入力音響信号に適合する度合いに基づいてパラメータの推定値を求める。本手法のモデル化において、置換元演奏のパラメータ $\theta^{(m)}$ と伴奏のパラメータ $\theta^{(a)}$ を決めると、音源信号のパワースペクトルのモデル $S_{n,f}$ が、次式により得られる。

$$S_{n,f} = M_{n,f} + A_{n,f} \quad (7)$$

一方、残響のパラメータ $\theta^{(g)}$ を決めると、入力音響信号から残響を除去した信号、すなわち音源信号の推定値、のパワースペクトルを次式により得ることができる。

$$|r_{n,f}|^2 = \left| x_{n,f} - \sum_{\tau=d}^D g_{\tau,f} x_{n-\tau,f} \right|^2 \quad (8)$$

そこで、これら 2 種類のパワースペクトルの乖離度 Q を最小化することで、モデルパラメータ $\theta = \{\theta^{(m)}, \theta^{(a)}, \theta^{(g)}\}$ を求める。すなわち、以下の最適化問題を解くことになる。

$$\text{minimize} \sum_{n=0}^{N-1} \sum_{f=0}^{F-1} Q(S_{n,f}, |r_{n,f}|^2) \quad \text{w.r.t.} \quad \{\theta^{(m)}, \theta^{(a)}, \theta^{(g)}\} \quad (9)$$

乖離度 Q は、遂行するタスク（ここでは音源分離や残響推定）との相性や、パラメータ推定の容易さなどを考慮し具体的に設計することになる。音源分離の問題では以下で定義される I ダイバージェンス $Q^{(I)}$ がよく用いられる¹⁵⁾。

$$Q^{(I)} = \sum_{n=0}^{N-1} \sum_{f=0}^{F-1} \left(|r_{n,f}|^2 \log \frac{|r_{n,f}|^2}{S_{n,f}} - (|r_{n,f}|^2 - S_{n,f}) \right). \quad (10)$$

調波非調波 GMM や NMF によるモデルのパラメータは、I ダイバージェンスに関して効率的に最適化できることが知られている^{16),17)}。一方、残響除去の従来研究¹³⁾ では $S_{n,f}$ と $|r_{n,f}|^2$ の間の板倉斎藤ダイバージェンスと $|r_{n,f}|^2$ の対数平均の和 $Q^{(IS)}$ が使われている。

$$Q^{(IS)} = \sum_{n=0}^{N-1} \sum_{f=0}^{F-1} \left(\log \frac{S_{n,f}}{|r_{n,f}|^2} + \frac{|r_{n,f}|^2}{S_{n,f}} - 1 + \log |r_{n,f}|^2 \right), \quad (11)$$

本論文では、これを便宜上『IS ダイバージェンス』と呼ぶ。残響のパラメータは、IS ダイバージェンスに関して効率的に最適化できることが知られている。

本手法では上記 2 種類の乖離度の両方を用いて、適当な初期値からはじめ、置換元演奏、伴奏、残響の各パラメータ $\theta^{(m)}$ 、 $\theta^{(a)}$ 、 $\theta^{(g)}$ の更新を順番に繰り返し実施する。ここで、 $\theta^{(m)}$ と $\theta^{(a)}$ の更新では I ダイバージェンスを最小化し、 $\theta^{(g)}$ の更新では IS ダイバージェンスを最小化する。本方法は、実際的な計算コストでパラメータを推定可能である。2 種類の異なる乖離度を用いるので、パラメータ推定値の収束性は保証されないが、実験において良い収束性能を示すことを確認している。このパラメータ推定アルゴリズムの概要を表 2 に記す。

なお、パラメータ推定値の収束性を保証する方法として、すべてのパラメータを I ダイバージェンスのみ、あるいは IS ダイバージェンスのみに関して最適化する方法が考えられる。いずれの場合も、パラメータに関する偏微分が複雑な形式になるので、勾配法などの繰り返しアルゴリズムを援用する必要がある。I ダイバージェンスのみを用いる場合、残響のパラメータ $\theta^{(g)}$ を勾配法で最適化することになる。一般に音楽信号を対象とする場合、残響フィルタ長 D を大きくする必要があるので、勾配法による $\theta^{(g)}$ の最適化は著しく効率が悪くなる。一方、IS ダイバージェンスのみを用いる場合、調波非調波 GMM のパラメータ

表 2 パラメータ推定アルゴリズム
Table 2 Parameter estimation algorithm.

1. $\theta^{(m)}$ を楽譜情報に基づき初期化、 $\theta^{(g)}$ を 0 で初期化する。 $\theta^{(a)}$ についてはまず乱数で初期化し、その後入力音響信号のパワー $|x_{n,f}|^2$ から $\theta^{(m)}$ が示すパワースペクトル成分を減算した信号に対してパラメータ更新を行う。これにより伴奏用 NMF モデルが置換元演奏の成分を取り込むことを軽減する。
2. 音源のモデルパラメータを更新する。
 - 2.a 式 (15) に基づき置換元演奏分配パワースペクトル、伴奏分配パワースペクトルを算出する。
 - 2.b 置換元演奏分配パワースペクトルを用いて調波非調波 GMM パラメータを反復更新する。
 - 2.c 伴奏分配パワースペクトルを用いて NMF によるモデルのパラメータを反復更新する。
 - 2.d 適当な回数 2.a から反復する。
3. 残響フィルタを式 (20) に基づき更新する。
4. 各パラメータの更新時の変化が十分小さくなるまで 2 から反復する。

$\theta^{(m)}$ と NMF のパラメータ $\theta^{(a)}$ も IS ダイバージェンスに基づいて最適化しなければならない。しかし、IS ダイバージェンスは、調波非調波 GMM のパラメータ最適化に適さない。具体的には、IS ダイバージェンスは推定対象スペクトル（ここでは $|r_{n,f}|^2$ ）を下回らないようにパラメータを推定する傾向があり、周波数方向に広がり大きい非調波 GMM が推定対象スペクトルの包絡線に近い曲線をとるようになる。その結果、調波 GMM が調波構造に適應する余地がなくなり、最終的に音源分離と調波非調波 GMM パラメータ推定がうまくいかなくなる。

4.2.2 最適化アルゴリズム

置換元演奏および伴奏のモデルパラメータは、残響フィルタを固定し式 (10) に基づき更新する。式 (10) を最小化する $\theta^{(m)}$ と $\theta^{(a)}$ は解析的に得られないので、補助関数法¹⁸⁾ と呼ばれる反復アルゴリズムを用いる。補助関数法とは、最小化したい最適化規準 $Q(\theta)$ に対し次の条件：

$$Q(\theta) = \min_{\vartheta} Q^+(\theta, \vartheta) \quad (12)$$

を満たす補助関数 $Q^+(\theta, \vartheta)$ を設計し、 Q^+ に対し補助変数 ϑ に関する最小化と本来の変数 θ に関する最小化を反復することで、間接的に本来の最適化規準を単調減少させる手法である。 $Q^+(\theta, \vartheta)$ を最小化させる θ, ϑ がともに解析的に解けるように Q^+ を設計すればパラメータ推定は簡単化される。

今、式 (10) の第 1 項に対して、負の対数関数の凸性から Jensen の不等式

$$-|r_{n,f}|^2 \log(M_{n,f} + A_{n,f}) \leq -|r_{n,f}|^2 \left(\lambda_{n,f}^{(m)} \log \frac{M_{n,f}}{\lambda_{n,f}^{(m)}} + \lambda_{n,f}^{(a)} \log \frac{A_{n,f}}{\lambda_{n,f}^{(a)}} \right) \quad (13)$$

$$\text{ただし } \lambda_{n,f}^{(m)} + \lambda_{n,f}^{(a)} = 1, \quad \lambda_{n,f}^{(m)} > 0, \quad \lambda_{n,f}^{(a)} > 0 \quad (14)$$

が成り立つ。 $\lambda_{n,f}^{(m)}$ および $\lambda_{n,f}^{(a)}$ は補助関数法における補助変数である。このとき、右辺はモデルパラメータ $\theta^{(m)}$ 、 $\theta^{(a)}$ に関する偏微分を 0 とおいた式から各パラメータの解析的更新則が導けることに注意されたい。この不等式の等号成立は

$$\lambda_{n,f}^{(m)} = \frac{M_{n,f}}{M_{n,f} + A_{n,f}} \quad \text{かつ} \quad \lambda_{n,f}^{(a)} = \frac{A_{n,f}}{M_{n,f} + A_{n,f}} \quad (15)$$

のときである。補助関数法によると、i) 補助変数を式 (15) の値に設定し式 (13) の右辺に代入し、ii) 式 (13) 右辺を減少させるようモデルパラメータを更新する、という 2 行程を反復することでパラメータを局所最適解まで更新することができることが知られている。式 (10)

に対し式 (13) の左辺を右辺に置き換え、 $\theta^{(m)}$ 、 $\theta^{(a)}$ の更新に関与しない項を除くと、

$$\left(-\lambda_{n,f}^{(m)} |r_{n,f}|^2 \log M_{n,f} + M_{n,f} \right) + \left(-\lambda_{n,f}^{(a)} |r_{n,f}|^2 \log A_{n,f} + A_{n,f} \right) \quad (16)$$

となる。なお、 $\lambda_{n,f}^{(m)}$ および $\lambda_{n,f}^{(a)}$ を式 (15) のように設定したとき、 $\lambda_{n,f}^{(m)} |r_{n,f}|^2$ と $\lambda_{n,f}^{(a)} |r_{n,f}|^2$ はそれぞれ置換元演奏、伴奏のパワースペクトルの一時的な推定値と見なせる。このことから、この推定法はまず置換元演奏と伴奏のパワースペクトルを推定し、次に各パワースペクトル推定値に対する I ダイバージェンスを最小化する方法となっていることが分かる。

調波非調波 GMM パラメータ $\theta^{(m)}$ は現在の置換元演奏パワースペクトルの推定値 $\lambda_{n,f}^{(m)} |r_{n,f}|^2$ とモデルの I ダイバージェンスを減少させるよう更新される。この更新にも Jensen の不等式を利用した同様の反復推定法が使える。まず置換元演奏パワースペクトル推定値からさらに第 j 発音の第 k 調波成分および第 l 非調波成分パワースペクトル推定値を得る。

$$\hat{H}_{j,k,n,f} = \frac{H_{j,k,n,f}}{M_{n,f}} \lambda_{n,f}^{(m)} |r_{n,f}|^2, \quad \hat{I}_{j,l,n,f} = \frac{I_{j,l,n,f}}{M_{n,f}} \lambda_{n,f}^{(m)} |r_{n,f}|^2 \quad (17)$$

次に、この推定値を用いて $\{u_{j,k,n}, v_{j,l,n}, \mu_{j,n}\}$ に対する以下の更新式を得る。

$$u_{j,k,n} \leftarrow \frac{\sum_{f=0}^{F-1} \hat{H}_{j,k,n,f}}{\sum_{f=0}^{F-1} H_{j,k,n,f}}, \quad v_{j,l,n} \leftarrow \frac{\sum_{f=0}^{F-1} \hat{I}_{j,l,n,f}}{\sum_{f=0}^{F-1} I_{j,l,n,f}},$$

$$\mu_{j,n} \leftarrow \frac{\sum_{k=1}^K \sum_{f=0}^{F-1} k \Omega_f \hat{H}_{j,k,n,f}}{\sum_{k=1}^K \sum_{f=0}^{F-1} k^2 \hat{H}_{j,k,n,f}} \quad (18)$$

伴奏用 NMF モデルのパラメータ $V_{f,c}$ 、 $U_{c,n}$ は伴奏パワースペクトル推定値 $\lambda_{n,f}^{(a)} |r_{n,f}|^2$ とモデルが示すパワースペクトル $a_{n,f}$ との間の I ダイバージェンスを減少させるように更新する。I ダイバージェンス規準の NMF は、乗法更新則¹⁷⁾ と呼ばれる以下の更新式を反復適用することで、 $V_{f,c}$ 、 $U_{c,n}$ の局所最適値が得られるので、これを用いる。

$$V_{f,c} \leftarrow V_{f,c} \frac{\sum_{n=0}^{N-1} U_{c,n} \left(\lambda_{n,f}^{(a)} |r_{n,f}|^2 / \sum_{c=0}^{C-1} V_{f,c} U_{c,n} \right)}{\sum_{n=0}^{N-1} U_{c,n}} \quad (19a)$$

$$U_{c,n} \leftarrow U_{c,n} \frac{\sum_{f=0}^{F-1} V_{f,c} \left(\lambda_{n,f}^{(a)} |r_{n,f}|^2 / \sum_{c=0}^{C-1} V_{f,c} U_{c,n} \right)}{\sum_{f=0}^{F-1} V_{f,c}} \quad (19b)$$

音源パラメータの更新後、そのモデルが表すパワースペクトルを固定し今度は式 (11) に

基づき残響フィルタの更新を行う．これは従来法¹³⁾と同様，以下の式によって行われる．

$$[g_{d,f}, \dots, g_{D,f}]^T = R_f^{-1} r_f \quad (20)$$

ただし， R_f^{-1} と r_f はそれぞれ修正相関行列，修正相関ベクトルと呼ばれ，以下で定義される．

$$R_f = \begin{pmatrix} \sum_{n=0}^{N-1} \frac{x_{n-d,f}^* x_{n-d,f}}{S_{n,f}} & \dots & \sum_{n=0}^{N-1} \frac{x_{n-d,f}^* x_{n-D,f}}{S_{n,f}} \\ \vdots & \ddots & \vdots \\ \sum_{n=0}^{N-1} \frac{x_{n-D,f}^* x_{n-d,f}}{S_{n,f}} & \dots & \sum_{n=0}^{N-1} \frac{x_{n-D,f}^* x_{n-D,f}}{S_{n,f}} \end{pmatrix},$$

$$r_f = \begin{pmatrix} \sum_{n=0}^{N-1} \frac{x_{n-d,f}^* x_{n,f}}{S_{n,f}} \\ \vdots \\ \sum_{n=0}^{N-1} \frac{x_{n-D,f}^* x_{n,f}}{S_{n,f}} \end{pmatrix} \quad (21)$$

4.3 置換元演奏除去の実行

以上のパラメータ推定の終了後，得られた置換元演奏と伴奏のパワースペクトル推定結果 $\hat{M}_{n,f}$ ， $\hat{A}_{n,f}$ を用いて式 (1) を適用することで伴奏演奏に対応するパワースペクトル $\bar{a}_{n,f}$ を得る．ここで，残響を除去する関数 $f^{-(R)}$ は，残響フィルタの推定結果 $\hat{g}_{\tau,f}$ を用いた逆畳み込み $f^{-(R)}(x_{n,f}) \equiv x_{n,f} - \sum_{\tau=d}^D \hat{g}_{\tau,f} x_{n-\tau,f}$ とする．最後に，Griffin らによって提案された方法¹⁹⁾を用いて，複素スペクトログラム $\bar{a}_{n,f}$ から時間領域の伴奏信号を合成する．

5. 演奏合成部

本章では，フレーズ置換の後半に相当する，置換先演奏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．具体的には，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) 置換元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) 調波非調波 GMM パラメータに基づく MIDI 音源合成演奏の音色・演奏表情補正，という手順で行われる．

5.1 連続 2 音の楽譜構造の類似性に基づくパラメータ算出

置換元演奏の調波非調波 GMM パラメータ $\theta^{(m)}$ から，楽譜情報中の連続 2 音のノートナンバと音長の類似性に基づいて置換先演奏第 λ 音にふさわしい調波非調波 GMM パラメータ $\psi_\lambda = \{\bar{u}_{\lambda,k,n}, \bar{v}_{\lambda,l,n}, \bar{\mu}_{\lambda,n}\}$ を算出する．

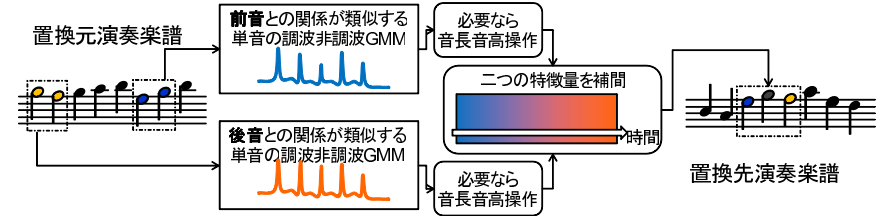


図 3 隣接 2 音の楽譜構造の類似性を用いて置換先演奏の各単音に調波非調波 GMM パラメータを算出する
Fig. 3 Calculate the harmonic/inharmonic GMM parameters of each note in the user specified score using estimated results.

$$q_\lambda^- = \operatorname{argmin}_j (|\hat{\xi}_{\lambda-1} - \xi_{j-1}| + |\hat{\eta}_{\lambda-1} - \eta_{j-1}| + |\hat{\xi}_\lambda - \xi_j| + |\hat{\eta}_\lambda - \eta_j|) \quad (22)$$

$$q_\lambda^+ = \operatorname{argmin}_j (|\hat{\xi}_\lambda - \xi_j| + |\hat{\eta}_\lambda - \eta_j| + |\hat{\xi}_{\lambda+1} - \xi_{j+1}| + |\hat{\eta}_{\lambda+1} - \eta_{j+1}|) \quad (23)$$

ここで， ξ, η は適当に規格化したノートナンバおよび音長を表し， ξ_j, η_j は置換元演奏の第 j 発音に対する値， $\hat{\xi}_\lambda, \hat{\eta}_\lambda$ は置換先演奏の第 λ 発音に対する値を示す．次に，得られた 2 つの単音に対するパラメータが滑らかに変化するように足し合わせる．置換先演奏第 λ 音の調波非調波 GMM 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，

$$\psi_{\lambda,n} = \begin{cases} \frac{\hat{\eta}_\lambda^+ - n}{\hat{\eta}_\lambda^+} \theta_{q_\lambda^-,n}^{(m)} + \frac{n - \hat{\eta}_\lambda^-}{\hat{\eta}_\lambda^-} \theta_{q_\lambda^+,n}^{(m)}, & \hat{\eta}_\lambda^- \leq n \leq \hat{\eta}_\lambda^+ \\ 0, & \text{otherwise} \end{cases} \quad (24)$$

とする．ただし， $\theta_{q_\lambda^-,n}^{(m)}, \theta_{q_\lambda^+,n}^{(m)}$ をそれぞれ置換元演奏の第 q_λ^-, q_λ^+ 音のパラメータを音高が $\hat{\xi}_\lambda$ ，音長が $\hat{\eta}_\lambda$ となるように伸縮したものとし，その和は各パラメータどうしに施すものと定義する．また $\hat{\eta}_\lambda^-$ および $\hat{\eta}_\lambda^+$ 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．この式は図 3 のように，2 つの調波非調波 GMM 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いる． $q_\lambda^+ + 1 = q_{\lambda+1}^-$ であることから，置換元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．

5.2 MIDI 音源合成演奏の音色・演奏表情補正

置換先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調波非調波 GMM 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、置換元演奏が持つ音色・演奏表情を合成結果に反映することができるだけでなく、調波非調波 GMM の示すスペクトルにモデル化誤差や推定誤差が入り込む余地を軽減することができる。

置換先演奏に対する MIDI 音源合成演奏を $\hat{m}_{n,f}$ とすると、ここから置換先楽譜を用いて 4.2 節とまったく同じ方法で各単音ごとの調波非調波 GMM パラメータ $\{\hat{u}_{\lambda,k,n}, \hat{v}_{\lambda,l,n}, \hat{\mu}_{\lambda,n}\}$ が得られ、同時に式 (17) による調波・非調波成分パワースペクトル推定値 $\{\hat{H}_{\lambda,k,n,f}, \hat{I}_{\lambda,l,n,f}\}$ が得られる。音色・演奏表情の反映は、これらパワースペクトル推定値を、前節で算出した調波非調波 GMM パラメータに基づき補正することで実現される。具体的には、 $\tilde{u}_{\lambda,k,n}, \tilde{v}_{\lambda,l,n}, \tilde{\mu}_{\lambda,n}$ を用いて、次式より各パワースペクトル推定値の強度と位置を変えたパワースペクトル $Y_{n,f}$ を得る。

$$Y_{n,f} = \sum_{\lambda=1}^{\Lambda} \left(\sum_{k=1}^K \frac{\tilde{u}_{\lambda,k,n}}{\hat{u}_{\lambda,k,n}} \hat{H}_{\lambda,k,n,f}^+ + \sum_{l=1}^L \frac{\tilde{v}_{\lambda,l,n}}{\hat{v}_{\lambda,l,n}} \hat{I}_{\lambda,l,n,f} \right) \quad (25)$$

ただし、 $\hat{H}_{\lambda,k,n,f}^+$ は $\hat{H}_{\lambda,k,n,f}$ を周波数方向に $(\tilde{\mu}_{\lambda,n}/\hat{\mu}_{\lambda,n})$ 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。この音高操作は簡易なものであり、フェーズボコーダ²⁰⁾ などの手法に比べ音高操作としての妥当性は低いが、ここでの音高操作量は同一音名内のピブラートなどによる微小変動分であるので小さく、問題はないと考えられる。

5.3 音響信号の再構成

合成された置換先演奏パワースペクトル $Y_{n,f}$ と置換先演奏に対する MIDI 音源合成演奏の位相 $\angle \hat{m}_{n,f}$ からなる複素スペクトログラムに対し、4.3 節と同様に Griffin らの方法で時間領域の置換先演奏を合成する。これに分離伴奏音を足し戻し、最後に残響を重畳する関数 $f^{(R)}$ を適用することでフレーズ置換結果を得る。 $f^{(R)}$ の実装には、残響除去 $f^{(-R)}$ と逆の効果を持つ FIR フィルタを推定し用いた。

6. 評価実験

フレーズ置換の性能は、1) 音源分離部にて置換元演奏と伴奏がよく分離されるか、2) 演奏合成部にて置換先演奏の合成結果は置換元演奏との演奏同一性を満たしているか、の 2 点によって決定付けられると考えられる。本章では、この 2 点のそれぞれに着目した 2 つの評価実験に基づき提案手法の有効性を検証し、またフレーズ置換処理全体に対する考察について述べる。各実験に共通する条件を表 3 に記す。

表 3 実験条件

Table 3 Experimental conditions.

STFT 分析条件	サンプリング周波数 STFT 分析窓関数 STFT 分析シフト幅	44.1 kHz 1,024 点ガウス関数 256 点
パラメータ	K : 調波構造を構成するガウス関数の数 L : 非調波ガウス関数の数 d : 残響フィルタ初期反射時刻 D : 残響フィルタ終了時刻	80 19 15 フレーム 80 フレーム

6.1 音源分離実験

6.1.1 実験目的と条件

第 1 実験では、音源分離部について、置換元演奏と伴奏が正しく分離されるかどうかを評価する。本論文で提案する音源分離法は、置換元演奏は調波非調波 GMM、伴奏には NMF を用いて推定を行い、さらに残響推定を統合したものである。そこで、提案法の有効性を検証するため、置換元演奏モデル・伴奏モデル・残響推定について以下に示す各条件を比較する。

- (1) 置換元演奏モデル：調波非調波 GMM を用いる（提案法）か NMF を用いるか
後者は前者に比べ、i) 少数の基底を用いる場合 F0 が頻繁に変化する楽器音の推定が困難、ii) 調波構造を直接モデル化しにくい、などの欠点があると考えられる。
- (2) 伴奏モデル：NMF を用いるか（提案法）調波非調波 GMM を用いるか
後者は前者に比べ、i) 性能が初期値に依存しやすい、ii) 計算時間が多い、などの欠点があると考えられる。
- (3) 残響推定：あり（提案法）かなしか
残響推定なしとは $\{g_{\tau,f}\} = 0$ と仮定することを意味する。

置換元演奏モデルに NMF を用いた手法では、置換元演奏の楽譜情報を有効利用できる Nonnegative Matrix Partial Co-Factorization (NMPCF)²¹⁾ の枠組みに沿って置換元演奏を推定する。NMPCF では元々の推定対象の多重奏音響信号に加え、あらかじめ MIDI 音源で合成した置換元演奏の音響信号を用い、これら両方の信号によく適合するような NMF のモデルのパラメータを推定する。また伴奏モデルに調波非調波 GMM を用いた手法では、問題設定上伴奏の楽譜をもとにモデルを初期化することはできないため、各時間フレームごとに、評価データの真の同時発音数を上回る個数の調波非調波 GMM を配置し、F0 はランダムに初期化する。本実験ではその個数を 8 とした。

表 4 音源分離実験で用いた楽曲

Table 4 List of musical pieces used in the separation experiment.

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

評価データは RWC Music Database: Jazz Music and Classic Music²²⁾ の Jazz と Classic それぞれ 6 曲ずつ計 12 曲を用いた。各曲の SMF から MIDI 音源を用いて音響信号を合成し、これに残響エフェクト用のインパルス応答（残響時間約 1 秒）を積み込んだものを入力音響信号とした。各曲について、旋律演奏楽器パートを分離対象、すなわち置換元演奏と見なした（表 4 参照）。

分離の良し悪しは、2 つの分離結果：i) 置換元演奏分離音 $\bar{m}_{n,f}$ ^{*1} へ残響を足し戻した $f^{(x)}(\bar{m}_{n,f})$ 、ii) 分離伴奏音 $\bar{a}_{n,f}$ へ残響を足し戻した $f^{(x)}(\bar{a}_{n,f})$ 、に対するそれぞれの真値からのずれを対数スペクトル距離により評価する。各分離結果の真値とは、置換元演奏のみの SMF・伴奏のみの SMF から入力音響信号作成時と同様の手順で合成した音響信号を意味する。分離結果 $\mathbb{B} = \{b_{n,f}\}_{0 \leq n \leq N-1, 0 \leq f \leq F-1}$ と真値 $\mathbb{E} = \{e_{n,f}\}_{0 \leq n \leq N-1, 0 \leq f \leq F-1}$ の間の対数スペクトル距離 $LSD(\mathbb{E}, \mathbb{B})$ は以下で定義する。

$$LSD(\mathbb{E}, \mathbb{B}) \equiv \sqrt{\sum_{n=0}^{N-1} \sum_{f=0}^{F-1} \left(20 \log_{10} \left| \frac{e_{n,f}}{b_{n,f}} \right| \right)^2} / NF \quad (26)$$

値が小さい方が両信号が類似し、0 のとき両信号は一致する。

6.1.2 実験結果と考察

図 4 および図 5 に真値と分離結果との間の対数スペクトル距離を全 12 曲における平均および標準偏差を示す。HIGMM は調波非調波 GMM を意味する。また NMF および NMPCF に付与された数字は各 NMF の基底数を意味する。

置換元演奏・伴奏のモデル化方法を比較すると、置換元演奏に調波非調波 GMM を用いた伴奏に NMF を用いた提案音源分離法が最良となっており、NMF のみを用いた手法に対し対数スペクトル距離は置換元演奏で 21.8%、伴奏で 11.93% 相対的に減少した。伴奏に調波

*1 分離伴奏音同様、 $\bar{m}_{n,f} = \frac{\bar{M}_{n,f}}{\bar{M}_{n,f} + \bar{A}_{n,f}} f^{(-x)}(x_{n,f})$ で算出する。

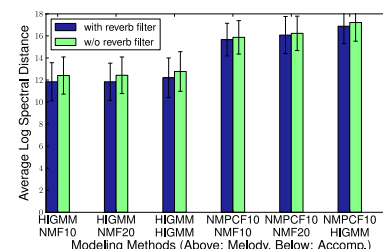


図 4 置換元演奏分離結果（小さいほど良い）
Fig. 4 Separation result of the target part.

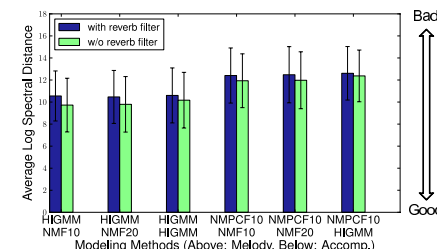


図 5 伴奏分離結果（小さいほど良い）
Fig. 5 Separation result of the accompaniment part.

非調波 GMM を用いた手法は NMF を用いた手法に近い分離性能を示したものの、我々の実装において 10 倍以上の計算時間を要した。この結果は「分離消去したいパートの楽譜のみ利用可能」という条件下の音源分離に対する提案法の有効性を示している。

残響推定の有無の比較では、残響成分を推定し分離後に再重畳することで置換元演奏の分離精度は向上したが、伴奏音の分離精度は悪化した。したがって、音源分離の観点からは、残響推定の必要性は明らかではない。このことは、残響除去そのものの精度とは直接関係ないことに注意されたい。実際、提案法のなかで用いた残響推定方法は、文献 9) で音楽音響信号に対しても一定の残響除去効果があることが報告されている。ただし、文献 9) では、本論文の方法と異なり、伴奏に対しても調波非調波 GMM を用いている。

また、評価尺度である対数スペクトル距離（の各曲の平均）について各手法間に有意差があるか検定を行った。これは平均の差の検定であり、対数スペクトル距離の等分散性の仮定を必要としないウェルチの t 検定を行った。結果は次のとおりで、上述の考察と合致する。

- (1) 置換元演奏・伴奏ともに調波非調波 GMM のものは提案法との間に有意差なし（ただし前述のとおり計算量の観点で提案法に優位性が認められる）
- (2) 残響推定の有無では有意差なし
- (3) NMPCF, NMF を用いた手法は提案法との間に有意差あり（置換元演奏分離結果では p 値 < 0.01, 伴奏分離結果では p 値 < 0.1）

6.2 演奏合成実験

6.2.1 実験目的と条件

第 2 実験ではフレーズ置換の後半部分に相当する演奏合成部に対し、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。ただし本

実験では演奏合成の能力にのみ着目するため、無伴奏演奏を用いた伴奏モデルは導入しない。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た演奏の調波非調波 GMM パラメータおよび残響フィルタパラメータを推定し、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。提案法の新しさは、MIDI 音源で合成した音響信号に置換元演奏から推定した音色・演奏表情特徴を付加させる方法により置換元演奏の特徴を反映した高品質な演奏を合成する点にある。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: 演奏合成部で算出する調波非調波 GMM が示すパワースペクトルから音響信号を合成する
- (3) Baseline2: 式 (25) の代わりに音量のみを操作するスペクトル操作式

$$Y_{n,f} = \frac{\sum_{\lambda=1}^{\Lambda} (\sum_{k=1}^K \tilde{u}_{\lambda,k,n} + \sum_{l=1}^L \tilde{v}_{\lambda,l,n})}{\sum_{\lambda=1}^{\Lambda} (\sum_{k=1}^K \hat{u}_{\lambda,k,n} + \sum_{l=1}^L \hat{v}_{\lambda,l,n})} |\hat{m}_{n,f}|^2 \quad (27)$$

を用いる。音量に関する演奏特徴のみを置換元演奏と同じにすることに相当する。

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

評価データは市販 CD 収録のプロによる演奏: Violin, Flute, Cello 各 3 曲の計 9 曲であり、4 つの MIDI 音源で上記の各演奏合成法を実行した。これらはいずれも残響時間 1 秒前後の長い残響が含まれている。合成の良し悪しは、置換先演奏合成結果と実演奏との間の式 (26) による対数スペクトル距離の小ささをもって評価する。

6.2.2 実験結果

図 6 に、9 曲各々に対し 4 つの MIDI 音源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との間の対数スペクトル距離の平均および標準偏差を示す。演奏合成法の比較では、MIDI 音源合成演奏の音色・演奏表情補正に基づく合成法「Ours」が最も対数スペクトル距離が小さいことが確認できる。また、残響推定の有無の比較では、残響フィルタを推定し合成時に同じ特性の残響を重畳することで対数スペクトル距離が小さくなることを確認できる。残響推定ありかつ「Ours」は最良の結果となり、残響推定なしの「Baseline1」からの改善量は 11.5%であった。これらの結果はいずれも、提案法が置換元演奏との演奏同一性を保持しつつ高品質に演奏を合成する能力に優れることを示している。各音響信号を聴取しても、提案法の音色は明らかに実演奏に近く、また「Baseline1」と比べ音質が良いと感じられた。なお、6.1.2 項と同様ウエルチの t 検定を行ったところ、「Ours」は他 3 手法

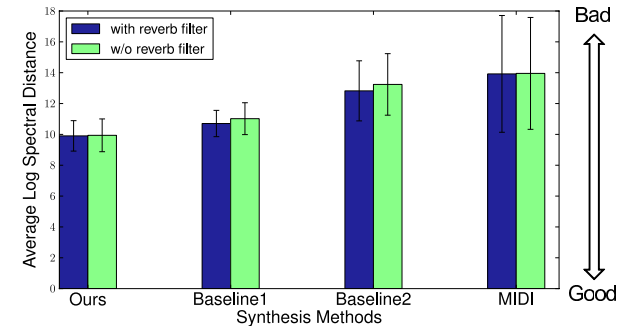


図 6 9 曲, 4 種の MIDI 音源による演奏合成結果の対数スペクトル距離平均 (小さいほど良い)
Fig. 6 LSD of the synthesized performances of nine pieces and four MIDI synthesizers.

と有意差が確認できた (p 値 < 0.01)。

また、図 6 中「MIDI」の標準偏差に対し、「Ours」の標準偏差が小さいことから、提案法による合成音は複数 MIDI 音源間の音の差異に強くは依存しないことが分かる。したがって、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いえ、専門知識に馴染みの薄い一般ユーザが利用する状況を想定すると都合が良いと考えられる。

残響推定の有無による対数スペクトル距離の違いは提案法において検定で有意差なしと判定されたが、聴感上は明確な違いが知覚された。実際、合成演奏を視聴してみると残響の有無という形で明確に違いを感じることができた。残響推定なしの合成結果からは残響音をほとんど感じられないのに対し、残響を推定し合成時に重畳する提案法による合成結果からは置換元演奏と同程度の長さの残響を聞き取ることができた。

6.3 フレーズ置換全体の考察

本節では本論文で示した手法を用いてフレーズ置換全体の処理を行った結果に対する著者自身の考察について述べる。本手法を用いて伴奏付きギター演奏を著者が作成した別フレーズの演奏に置換した例を文献 (23) に示す。主観的な印象としては、まず置換先演奏は置換元演奏の音色のまま合成されていると感じられる。一方置換元演奏と思われる成分が分離伴奏音にいくらか残っている。この残留成分は、特に分離伴奏音のみを試聴した場合に顕著に知覚できるが、置換先演奏を分離伴奏音に加えた後では、加える前と比べて残留成分はそれほど気にならない。とはいえ、まだ知覚できる程度の残留成分が残されているため、フレーズ置換のさらなる品質向上のためには、音源分離精度の改善が必要である。

7. おわりに

本論文では、特定パートのフレーズ演奏をユーザが指定した楽譜によるものに差し替えるフレーズ置換と呼ぶ新しい楽曲編集技術のための音源分離・演奏合成法について報告した。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と、NMF に基づく伴奏モデルを統合して用いることで音響信号から置換元演奏成分を除去するとともに、調波非調波 GMM により推定した置換元演奏の音色・演奏表情をもとに MIDI 音源音響信号を補正する演奏合成を行った。

今後の課題には、6.3 節で述べたように音源分離精度のさらなる向上があげられる。また、置換先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、実演奏の多様な演奏表情をより正確に学習・生成できるように、現在の隣接 2 音しか着目していないパラメータ算出法をより多様な情報を取り込むように改良していく必要がある。

謝辞 本研究の一部は、科研費基盤研究(S)、科学技術振興機構 CrestMuse プロジェクト、およびグローバル COE プログラム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) 濱野智史: インターネット関連産業, デジタルコンテンツ白書 2009, pp.118–124, 財団法人デジタルコンテンツ協会 (2009).
- 3) Itoyama, K., Goto, M., Komatani, K., Ogata, T. and Okuno, H.G.: Parameter Estimation for Harmonic and Inharmonic Models by Using Timbre Feature Distributions, *The Journal of Information Processing Society of Japan*, Vol.50, No.7, pp.1757–1767 (2009).
- 4) 安部武宏, 糸山克寿, 吉井和佳, 駒谷和範, 尾形哲也, 奥乃 博: 音色の音高依存性を考慮した楽器音の音高操作手法, 情報処理学会論文誌, Vol.50, No.3, pp.1054–1066 (2009).
- 5) Lee, D.D. and Seung, H.S.: Algorithms for Non-negative Matrix Factorization, *Proc. 2000 Neural Information Processing Systems Conference*, pp.556–562 (2001).
- 6) Dannenberg, R.B. and Hu, N.: Polyphonic Audio Matching for Score Following and Intelligent Audio Editors, *Proc. 2003 International Computer Music Conference*, pp.27–33 (2003).
- 7) Smaragdis, P. and Brown, J.: Non-negative Matrix Factorization for Polyphonic Music Transcription, *Proc. 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp.170–180 (2003).
- 8) Casey, M. and Westner, A.: Separation of Mixed Audio Sources by Independent Subspace Analysis, *Proc. 2000 International Computer Music Conference*, pp.154–161 (2000).
- 9) Yasuraoka, N., Yoshioka, T., Nakatani, T., Nakamura, A. and Okuno, H.G.: Music Dereverberation using Harmonic Structure Source Model and Wiener Filter, *Proc. 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp.53–56 (2010).
- 10) Kinoshita, K., Nakatani, T. and Miyoshi, M.: Blind Upmix of Stereo Music Signal Using Multi-Step Linear Prediction Based Reverberation Extraction, *Proc. 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp.49–52 (2010).
- 11) 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: Manipulations of Single Part Using Harmonic and Inharmonic Models, *Proc. 17th ACM International Conference on Multimedia*, pp.203–212 (2009).
- 12) Fletcher, H., Blackham, E. and Stratton, R.: Quality of Piano. Tones, *The Journal of the Acoustical Society of America*, Vol.34, No.6, pp.749–761 (1962).
- 13) Yoshioka, T., Nakatani, T. and Miyoshi, M.: An Integrated Method for Blind Separation and Dereverberation of Convolutional Audio Mixtures, *Proc. European Signal Processing Conference* (2008).
- 14) Kinoshita, K., Delcroix, M., Nakatani, T. and Miyoshi, M.: Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction, *IEEE Trans. Audio, Speech and Language Processing*, Vol.17, No.4, pp.534–545 (2009).
- 15) FitzGerald, D., Cranitch, M. and Coyle, E.: On the use of the beta divergence for musical source separation, *Proc. 20th Irish Signals and Systems Conference*, pp.1–6 (2009).
- 16) Kameoka, H., Ono, N. and Sagayama, S.: Speech Spectrum Modeling for Joint Estimation of Spectral Envelope and Fundamental Frequency, *IEEE Trans. Audio, Speech and Language Processing*, Vol.18, No.6, pp.1507–1516 (2010).
- 17) Nakano, M., Kameoka, H., Roux, J.L., Kitano, Y., Ono, N. and Sagayama, S.: Convergence-Guaranteed Multiplicative Algorithms for Nonnegative Matrix Factorization with Beta-Divergence, *Proc. 2010 IEEE International Workshop on Machine Learning for Signal Processing* (2010).
- 18) Kameoka, H., Ono, N. and Sagayama, S.: Auxiliary function approach to parameter estimation of constrained sinusoidal model for monaural speech separation, *Proc. 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing*,

pp.29–32 (2008).

- 19) Griffin, D.W. and Lim, J.S.: Signal Estimation from Modified Short-Time Fourier Transform, *IEEE Trans. Acoustics, Speech, & Signal Processing*, Vol.32, No.2, pp.236–243 (1984).
- 20) Dolson, M.: The Phase Vocoder: A Tutorial, *Computer Music Journal*, Vol.10, No.4, pp.14–27 (1986).
- 21) Yoo, J., Kim, M., Kang, K. and Choi, S.: Nonnegative Matrix Partial Co-Factorization for Drum Source Separation, *Proc. 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp.1942–1945 (2010).
- 22) Goto, M., Hashiguchi, H., Nishimura, T. and Oka, R.: RWC Music Database: Popular, Classical, and Jazz Music Databases, *Proc. 3rd International Conference on Music Information Retrieval*, pp.287–288 (2002).
- 23) 安良岡直希: フレーズ置換動作例, 入手先(<http://winnie.kuis.kyoto-u.ac.jp/members/yasuraok/ipsj2011.html>) (参照 2011-10).

(平成 22 年 12 月 28 日受付)

(平成 23 年 9 月 12 日採録)



安良岡直希 (正会員)

2009 年京都大学工学部情報学科卒業。2011 年同大学大学院情報学研究科知能情報学専攻修士課程修了。在学中は音楽音響信号の音源分離、残響推定、演奏分析合成の研究に従事。現在、ヤマハ株式会社勤務。日本音響学会会員。



吉岡 拓也

2004 年京都大学工学部情報学科卒業。2006 年同大学大学院情報学研究科知能情報学専攻修士課程修了。2010 年同大学院博士後期課程修了。京都大学博士 (情報学)。2006 年日本電信電話株式会社入社。NTT コミュニケーション科学基礎研究所にて残響除去、音声強調、耐雑音音声認識の研究に従事。2010 年日本音響学会第 28 回栗屋潔学術奨励賞、2011 年日本音響学会第 6 回独創研究奨励賞板倉記念、2010 年度音声研究会研究奨励賞各受賞。日本音響学会、IEEE 各会員。



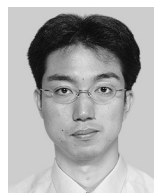
糸山 克寿 (正会員)

2006 年京都大学工学部情報学科卒業。2008 年同大学大学院情報学研究科知能情報学専攻修士課程修了。2011 年同大学院博士課程修了。博士 (情報学)。同年京都大学大学院情報学研究科助教。音源分離や音楽鑑賞インタフェース等の音楽情報処理の研究に従事。日本音響学会、IEEE 各会員。



高橋 徹 (正会員)

1996 年名古屋工業大学知能情報システム学科卒業。2004 年同大学大学院工学研究科電気情報工学専攻博士後期課程修了。博士 (工学)。和歌山大学システム工学部産学官連携研究員を経て、2008 年より京都大学大学院情報学研究科グローバル COE 助教。研究分野は、ロボット聴覚および音声コミュニケーション。音声による人間-ロボット間インタラクションのための音声認識・合成。ロボット聴覚ソフトウェア HARK、音声分析変換合成システム STRAIGHT の開発。RSJ、IEICE、ASJ 各会員。



駒谷 和範 (正会員)

1998 年京都大学工学部情報工学科卒業。2000 年同大学大学院情報学研究科知能情報学専攻修士課程修了。2002 年同大学院博士後期課程修了。京都大学博士 (情報学)。京都大学助手、助教を経て、2010 年より名古屋大学大学院工学研究科准教授。JST さきがけ「情報環境と人」領域研究員兼務。主に音声対話システムの研究に従事。2008 年から 2009 年までカーネギーメロン大学客員研究員。本会平成 16 年度山下記念研究賞、FIT2002 ヤングリサーチ賞等を受賞。電子情報通信学会、言語処理学会、人工知能学会、ACL、ISCA 各会員。



尾形 哲也 (正会員)

1993年早稲田大学理工学部機械工学科卒業。日本学術振興会特別研究員，早稲田大学理工学部助手，理化学研究所脳科学総合研究センター研究員，京都大学大学院情報学研究科講師を経て，2005年より同助教授（現准教授）。博士（工学）。JST さきがけ研究「情報環境と人」領域研究員（5年）。この間，早稲田大学ヒューマノイド研究所客員准教授，同大学理工学研究所客員准教授，理化学研究所脳科学総合研究センター客員研究員等を兼務。研究分野は人工神経回路モデルおよび人間とロボットのコミュニケーション発達を考えるインタラクション創発システム情報学。日本ロボット学会，日本機械学会，人工知能学会，計測自動制御学会，ヒューマンインタフェース学会，バイオメカニズム学会，IEEE 等各会員。



奥乃 博 (正会員)

1972年東京大学教養学部基礎科学科卒業。日本電信電話公社，NTT，JST，東京理科大学を経て，2001年より京都大学大学院情報学研究科知能情報学専攻教授。博士（工学）。この間，スタンフォード大学客員研究員，東京大学工学部客員助教授。人工知能，音環境理解，ロボット聴覚，音楽情報処理の研究に従事。1990年度人工知能学会論文賞，IEA/AIE-2001，2005，2010 最優秀論文賞，IEEE/RSJ IROS-2001，2006 Best Paper Nomination Finalist，IROS-2010 NTF Award for Entertainment Robots and Systems，第2回船井情報科学振興賞等受賞。本学会理事，人工知能学会，日本ロボット学会，日本ソフトウェア科学会，ACM，IEEE，AAAI，ASA 等各会員。