情報処理学会論文誌 Vol. 50 No. 3 1054–1066 (Mar. 2009)

音色の音高依存性を考慮した楽器音の音高操作手法

安部武宏[†1] 糸山克寿[†1] 吉井和佳[†2]
駒谷和範[†1] 尾形哲也[†1] 奥乃博[†1]

本稿では，ある音高を持つ楽器音をもとにして，音色の歪みを抑えながら任意の音高を持つ楽器音を合成する手法について述べる．我々は音色の聴感上の差に関する音響心理学的知見に基づき，楽器音のスペクトログラム上で観察される音色特徴量として，(i) 倍音ピーク間の相対強度，(ii) 非調波成分の分布，(iii) 時間方向の振幅エンベロープの 3 つを定義する．まず，もとになる楽器音の音色特徴量を分析するため，糸山らの調波・非調波統合モデルを用いて楽器音を調波構造と非調波構造に分離する．音高操作時には，特徴量 (i)，(ii) の音高依存性を考慮しなければならない．そのため，音高に対する特徴量を 3 次関数で近似し，所望の音高における特徴量の値を予測する．32 種類の楽器に対して音高操作を試みたところ，音高依存性を考慮しない場合と比べて合成音と実際の楽器音との距離が，スペクトル距離尺度では 64.70%，MFCC 距離尺度では 32.31%減少し，手法の有効性が確かめられた．

An Analysis-and-synthesis Approach for Manipulating Pitch of a Musical Instrument Sound Considering Pitch-dependency of Timbral Characteristics

Takehiro Abe,[†1] Katsutoshi Itoyama,[†1]
Kazuyoshi Yoshii,[†2] Kazunori Komatani,[†1]
Tetsuya Ogata[†1] and Hiroshi G. Okuno[†1]

This paper presents a synthesis method that can generate musical instrument sounds with arbitrary pitches from a given musical instrument sound while constraining distorting timbral characteristics. Based on the psychoacoustical knowledge on auditory effects of timbre, we define timbral features on the spectrogram of a musical instrument sound as (i) relative amplitudes of harmonic components, (ii) distribution of inharmonic components, and (iii) temporal envelopes of harmonic components. First, to analyze timbral features of a seed, it is separated into harmonic and inharmonic components by using Itoyama's integrated model. In pitch manipulation, it is necessary to take into account the relation of pitch and features (i) and (ii). Therefore, we predict the values of each feature by using a cubic polynomial that approximates the feature distribution over pitches. Experimental results showed the effectiveness of our method; the spectral and MFCC distances between synthesized sounds and real sounds of 32 instruments were reduced by 64.70% and 32.31%, respectively.

1. はじめに

従来のイコライザとは音響信号全体の周波数特性を変化させるものであったが，近年，音楽音響信号に特化し，楽器単位での音量の操作や音色の置き換えが可能な楽器音イコライザと呼ばれる新技術が開発されてきている[1)–3)]．多くのオーディオプレーヤに実装されているイコライザは周波数帯域の操作によって楽曲の音響を変化させるが，楽器音イコライザが提供する楽器単位の操作によって音楽鑑賞の幅はさらに広がると期待される．Yoshii らの Drumix[2)] では，スネアドラムやバスドラムといった打楽器単位での音量の操作と音色の置き換えを実現している．一方，糸山らの楽器音イコライザ[3)] では打楽器だけではなく，すべての楽器の音量を操作させることができるが，Drumix で実現されていた音色の置き換えは扱われていない．

我々の最終目標は，任意の楽器パートをユーザの好みの楽器音に置き換えるイコライザの開発である．これが実現できれば，たとえば，ロック風の楽曲を構成するギター，ベース，キーボードなどの楽器音を，ヴァイオリン，ウッドベース，ピアノなどの楽器音で置き換えることで，ユーザはその楽曲をクラシック風にアレンジして楽しむことができる．また，好きなギタリストが演奏した楽曲からギター音を抽出し，別の楽曲のギターパートをそのギター音で置き換えることで，ユーザはそのギタリストに様々なフレーズを演奏させることもできる．

上記イコライザを実現するための技術的課題として以下の 2 つがあげられる．

(1) 混合音中からユーザが置き換えに用いたい楽器音を抽出するため，**混合音から任意の楽器音を分離**する．

(2) 任意のフレーズを演奏するため，分離された楽器音をもとにして**任意の音高・音長を**

†1 京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University
†2 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)

持つ楽器音を合成する．

前者は，糸山らを含め多くの研究者によって継続的に取り組まれており，その成果が報告されている[4)–6)]．一方，分離された音の応用についてはほとんど議論されてこなかった．そこで，我々は後者，特に音高操作に着目し，分離された複数の単音を入力とした楽器音の音高操作に取り組む．

楽器音の合成に関する従来研究は，以下の 3 つに大別できる．

(1)　**ボコーダ**

音声合成における最も古い研究としてボコーダがあげられ，近年でも高品質な音声合成に向けてさかんに研究がされている[7),8)]．主に音声を入力の対象として扱われる手法であるが，楽器音への適用も可能である．音高の操作は，スペクトルをソースとフィルタ構造に分離し，所望の基本周波数を持つソースにフィルタ構造を畳み込むことによって実現される．ボコーダは入力された楽器音の調波・非調波成分を区別せずにフィルタ構造を推定して，操作を加えるため，音色の特徴に関係がある非調波成分が歪められる．また，楽器音の音色特徴を明確にパラメータ表現していないため，その特徴を解析することは困難である．

(2)　**フェーズボコーダ**

フェーズボコーダは歴史の長い楽器音合成方式であり，多くの派生的な手法が報告されている[9)–11)]．音高を操作するには，まず最初に入力となる分析対象音のスペクトログラムを時間方向に伸縮し，隣接するフレーム間の整合性がとれる位相を人工的に与えてフーリエ逆変換を行うことで，音長操作を行う．次に音長操作された音源を伸縮率の逆数を乗算したサンプリングレートで標本化しなおせばよい．また，音長操作と同様に，スペクトログラムを周波数方向に伸縮する手法もある[12)]．ボコーダと同様に，フェーズボコーダもまた入力された楽器音の調波・非調波成分を区別せずに操作を加えるので，非調波成分が歪められるという問題がある．

(3)　**正弦波重畳モデル**

正弦波重畳モデルは，音声や楽器音をよく表現するモデルとして有名である[13),14)]．本方式では，まず，分析対象音のスペクトログラムに現れるローカルピークをトラッキングし，各ピーク成分の瞬時周波数，瞬時振幅，位相を抽出する．次にフレーム単位で分析された瞬時振幅を補間することによって得られるサンプル単位の瞬時振幅と，分析されたピークの周波数を持つ正弦波の積を足し合わせることで音の合成を行う．音高の操作はトラッキングしたピークの周波数間隔を伸縮することによって行われる．フェーズボコーダとは異なり，楽器音のピークのみに注目すればよく，異なる楽器音のピークの対応付けが容易なため，楽器音のモーフィングにも利用されている[15)]．また，音源分離にも応用されており，様々なモデルパラメータ推定手法が報告されている[16)–18)]．正弦波重畳モデルでは，分析対象音からトラッキングしたピーク成分を減算した残差信号によって非調波成分を扱えるが，明示的に音色の特徴がパラメータとして定義はされていない．複数の楽器音の分析は，中間音を合成するモーフィングといった手段でしか扱われておらず，音色の音高依存性などの音色の性質を分析するまでには至っていない．

以上の従来手法における課題をふまえて，本稿では，楽器音の音色から音色特徴量を分析し，複数の楽器音の単音から音高による音色の変化を学習することで，音色の音高依存性を考慮した音高操作手法を報告する．

本稿の構成は以下のとおりである．2 章で音色の特徴量を定義し，音高操作における問題と解決法について議論する．3 章で実装方法について述べる．4 章で本手法の評価実験について報告する．5 章で従来の音高操作手法として知られているボコーダの 1 種である STRAIGHT[8)] と比較評価した結果を報告する．最後に 6 章でまとめとする．

2. 音色の特徴を考慮した音高操作

本研究の目的は，ある楽器個体の実際の音（seed と呼ぶ）がいくつか得られているとき，それらをもとにして同個体の任意の音高を持つ音を合成することである．このとき重要な点は，音色の音響的特徴が歪まないようにすることである[⋆1]．たとえば，ある音高を持つ楽器音から他の音高を持つ音を合成したとき，これら 2 つの音は同一個体から発せられる音であると感じることができなければならない[⋆2]．

音色の音響的特徴の歪みを抑えて楽器音を合成するには，音色の特徴量を数学的に定義し，これを分析する必要がある．音響心理学の分野では，音色の聴感上の知覚の差はおもに，(i) 高周波数領域での倍音ピークの有無，(ii) 発音時に発生する非調波成分，(iii) 各ピークの時間方向における振幅の変動，の 3 つに起因する傾向があるとの報告がある[19)]．我々はこれらの要因を以下の 3 つの特徴量にそれぞれ対応付ける．

(i)　**倍音ピーク間の相対強度**

⋆1 本稿では音色の歪みを，音量と音高の要因を除外したスペクトル上での実楽器から発せられる楽器音との差異と定義する．

⋆2 ただし，同一楽器から発せられる音と感じられるかどうかを工学的に評価することは困難であるので，本稿の実験では，実楽器から発せられた楽器音の特定の音高の音と，評価すべき同じ音高の合成音に関して，スペクトル距離に基づき定められる尺度で音色の歪みを評価することにする．

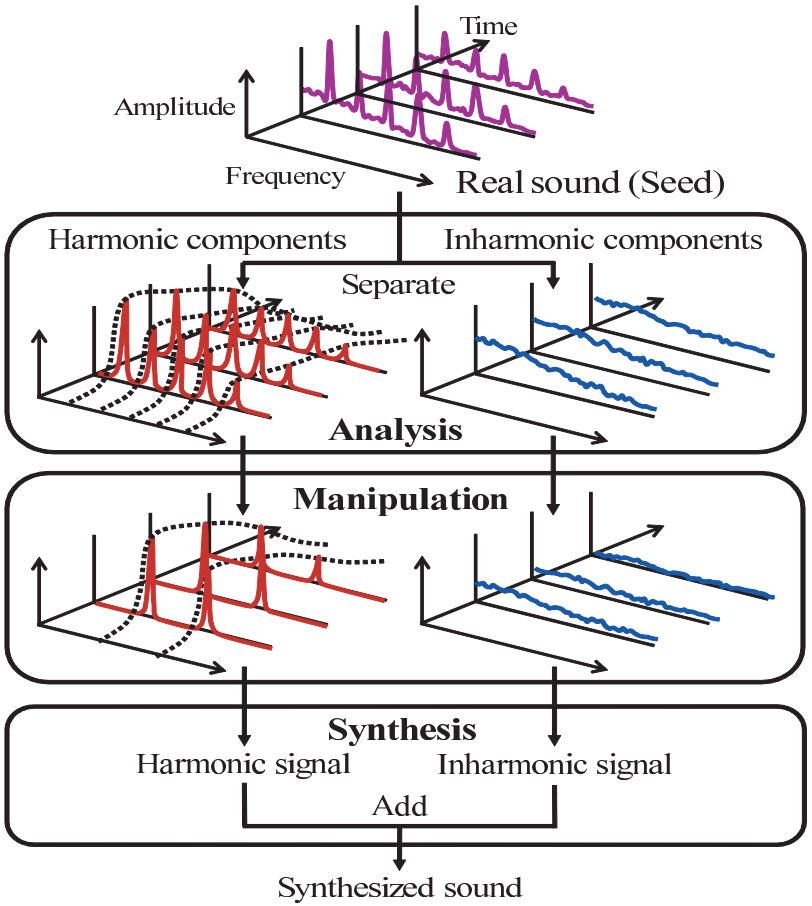


図 1　本手法の概要
Fig. 1　Overview of our method.

(ii)　**非調波成分の分布**

(iii)　**時間方向の振幅エンベロープ**

図 1 に本分析・合成手法の概要を示す．特徴量 (i) および (iii) は調波成分に関するもの，特徴量 (ii) は非調波成分に関するものであるので，まず，seed の調波成分と非調波成分を分離し，各特徴量を分析する．

次に，音色の歪みを抑えるように音高操作を行う．JIS では，音色は「聴感上の音の性質の一つで，2 音の大きさおよび高さがともに等しくてもその 2 音が異なった感じを与えるとき，その相違に対応する性質」と定義されている[20]．この定義では，音色は音高と音量とは独立の音の性質として扱われており，これに従えば，音色を歪みを抑えつつ音高操作するには定義した音色特徴量を保持したまま行うべきである．しかし，音色には音高への依存性があることが知られており[21]，音高によって変化するべき特徴量を保持したまま音高操作を行うと操作された楽器音に歪みが生じる．また，音色に関係する物理量としてスペクトル包絡が知られているが，1 つのスペクトル包絡だけで異なる音高の倍音ピーク間の相対強度を正確に表現することはできない．これら音色特徴量のみで音色の特徴をとらえられるとはいいがたい．そこで我々は音色特徴量とそれらの依存関係を分析しなければ，音色の特徴をとらえることができないという立場で，音色特徴量に加え，複数の楽器音から音色特徴量の音高依存性を分析することで，楽器個体の音色を扱うことを試みている．すなわち，操作は音色特徴量の音高依存性を考慮して行われる．最後に，調波成分・非調波成分を別々に再合成し，足し合わせる．

2.1　楽器音の分析

音色特徴量を分析するためには，調波成分と非調波成分とを明示的に分けて取り扱い，それぞれにおける特徴量を定義する必要がある．この問題を解決するため，我々は糸山らが提案した調波・非調波統合モデル[3] を利用して楽器音を表現することを試みる．すなわち，seed のスペクトログラム $M(f,r)$ に対し，調波成分に対応するパラメトリックモデル $M_H(f,r)$ と非調波成分に対応するノンパラメトリックモデル $M_I(f,r)$ を w_H および w_I で重み付けした混合モデルをフィッティングさせる．

$$M(f,r) = w_H M_H(f,r) + w_I M_I(f,r) \tag{1}$$

ここで，f と r はそれぞれ周波数とスペクトログラムのフレーム番地を表す．また，$\sum_{f,r} M_I(f,r) = 1$ という制約が与えられているので，重み w_I は非調波成分のエネルギーと考えることができ，$w_I M_I(f,r)$ は非調波成分のスペクトログラムそのものを表す．一方，$M_H(f,r)$ は，各倍音 n に対するパラメトリックモデルの重み付き混合モデルとして表現される．

$$M_H(f,r) = \sum_n F_n(f,r) E_n(r) \tag{2}$$

ここで，$F_n(f,r)$ および $E_n(r)$ は，**図 2** と**図 3** に示すような周波数エンベロープおよび時

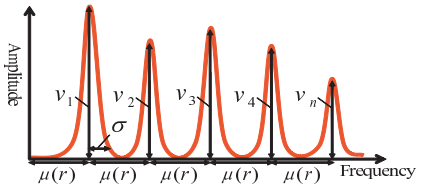


図 2　周波数エンベロープ
Fig. 2　Spectral envelope.

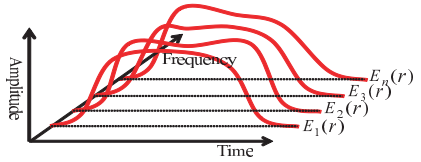


図 3　時間エンベロープ
Fig. 3　Temporal envelopes.

間エンベロープのモデルとなっている．

$F_n(f,r)$ は混合正規分布を構成する 1 つの要素の正規分布に混合比を乗じたものとして表現される．

$$F_n(f,r) = \frac{v_n}{\sqrt{2\pi\sigma^2}} \exp -\frac{(f - n\mu(r))^2}{2\sigma^2} \tag{3}$$

ここで，σ は倍音ピークの周波数方向への分散を表す．$\mu(r)$ は seed の音高軌跡である．また，v_n は $\sum_n v_n = 1$ を満たす重みである．

一方，$E_n(r)$ は $\sum_r E_n(r) = 1$ を満たすノンパラメトリックな関数である．糸山らは $E_n(r)$ に対しても $F_n(f,r)$ と同様のパラメトリックモデルを構成していたが，より詳細な分析を可能とするため本稿ではこのような方法をとった．

この統合モデルにおいて，音色の特徴量 (i)，(ii) および (iii) は，それぞれ v_n，$w_I M_I(f,r)$ および $E_n(r)$ に対応する．これらの求め方は 3.1 節において述べる．

2.2　音 高 操 作

音高を操作するには，音高軌跡 $\mu(r)$ に所望の倍率を乗算すればよいが，このとき音色特徴量の値を変化させずにそのまま利用することはできない．なぜなら，音色は音高依存性を持つことが知られており[21]，音高の操作が大きくなるにつれて音色の歪みは増加するからである．**図 4** に示すように，音高を $\mu(r)$ から $\mu'(r)$ に変化させる場合には，相対強度を v_n から v'_n へと適切に変化させる必要がある．

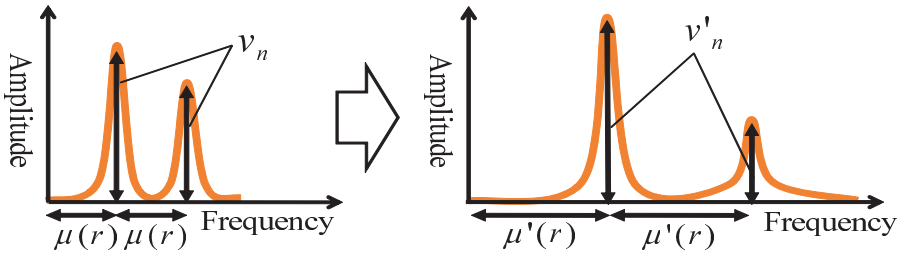


図 4　周波数エンベロープの操作
Fig. 4　Manipulation of spectral envelope.

この問題を解決するため，我々は北原らの提案した音高依存性を考慮した楽器音識別手法[22] に着目する．彼らは音高に対する音響的特徴量を 3 次関数を用いて近似し，音高依存性を除去したあとの特徴量分布を学習することで，楽器音識別率が向上したと報告している．

音色が音高に依存する理由として以下が知られている[23]．

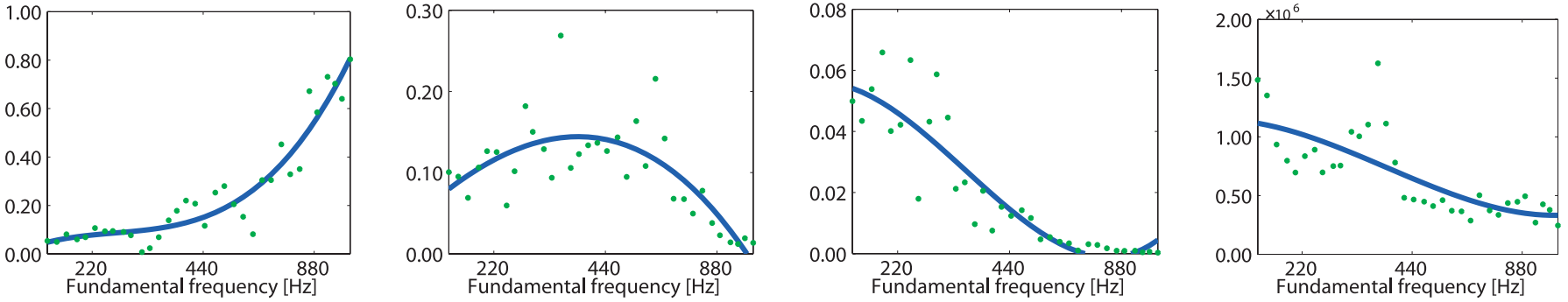
(1)　音高が低くなれば，発音体は大きくなる．発音体の質量が大きくなると慣性も大きくなり，発音の立ち上がりや減衰により多くの時間を要する．
(2)　音高が高くなると振動損失が大きくなるために，高次の高調波は発生されにくくなる．
(3)　一部の楽器では音高により発音体が異なり，各発音体は異なる材質からできている．

これらの知見から，楽器の音色は低域から高域にいくに従って連続的に変わるといえる．よって，本研究では音高よりも奏法に依存すると考えられる特徴量 (iii) 時間エンベロープを除き，音高に対する特徴量 (i) 倍音ピーク間の相対強度，(ii) 非調波成分の分布を n 次関数（**音高依存特徴関数**と呼ぶ）で近似する．本稿では音高依存特徴関数の次数に 3 次を用いた．この次数は，限られた学習データから音色の音高依存性を学習でき，音色特徴量の音高による変化を十分に扱えるという基準を設け，予備実験より決定した．

具体的には，以下の 2 つのパラメータに着目する．

(1)　各倍音の倍音ピーク間の相対強度 v_n
(2)　調波成分のエネルギーに対する非調波成分のエネルギーの比 w_H/w_I

(1) の v_n に関しては，n ごとに独立に音高依存特徴関数を作成する．これによって，必ずしも v_n に関する制約 $\sum_n v_n = 1$ は満たされなくなるが，この場合でも $\sum_n v_n$ の値はほぼすべての音高に対して 0.9 – 1.1 程度に収まっており，生成される楽器音の音色がこれに



a) トランペットの第 1 次倍音の相対強度　b) トランペットの第 4 次倍音の相対強度　c) トランペットの第 10 次倍音の相対強度　d) トランペットの調波と非調波のエネルギー比

図 5　トランペットの音高依存特徴関数．点と実線はそれぞれ，音高ごとに分析された音色の特徴量と，導出された音高依存特徴関数である
Fig. 5　Pitch-dependent feature functions for trumpet. The dots and the lines are the analyzed timbral features and the approximated pitch-dependent feature functions.

よって大きく変化することはないと考える．異なった音高を持つ複数の seed が与えられれば，それらの音色特徴量を分析し，最小二乗法によって音高依存特徴関数を求めることができる．得られた音高依存特徴関数を用いれば，所望の音高における音色特徴量を予測することができる．例として，**図 5** にトランペットの第 1 次倍音，第 4 次倍音，第 10 次倍音の相対強度，および調波成分と非調波成分のエネルギー比の音高特徴依存関数を示す．

2.3　楽器音の合成

調波成分に対応する音響信号 $s_H(t)$ を合成するには，特徴量 (i) および (iii) をパラメータとする正弦波重畳モデルを用いる．非調波成分に対応する音響信号 $s_I(t)$ を合成するには，特徴量 (ii) を入力とするオーバラップ加算法を用いる．

3.　処理系の実装

本章では，2 章で述べた手法の具体的な実装について説明する．

3.1　楽器音の分析

ここで問題となるのは，2.1 節で示した統合モデルにおける未知パラメータ w_H，w_I，$F_n(f,r)$，$E_n(r)$，v_n，$\mu(r)$，σ，$M_I(f,r)$ を推定することである．そのため，糸山らは統合モデルと seed のスペクトログラムとの Kullback-Leibler Divergence（KLD）を減少させるようにパラメータを反復更新する手法を提案している．この反復過程は Expectation-Maximization（EM）アルゴリズムと解釈でき，効率的にパラメータを推定することができる．

3.2　音 高 操 作

音高操作を行うには，周波数エンベロープを構成する音高軌跡 $\mu(r)$ に対して，実数 α（音高を低くする場合：$0 \leq \alpha < 1$，音高を高くする場合：$1 < \alpha$）を乗算する．ここで，$\mu'(r)$ は所望する操作後の音高とすると以下が成り立つ．

$$\mu'(r) = \alpha\mu(r) \tag{4}$$

たとえば，α を 2 とすれば，seed の 1 オクターブ上の音高の楽器音が合成できる．操作後の楽器音の倍音ピーク間の相対強度 v'_n は，音高依存特徴関数から予測される各倍音ごとの倍音ピーク間の相対強度を制約条件 $\sum_n v_n = 1$ より正規化することで得られる．また，操作後の楽器音の非調波成分のエネルギー w'_I は，調波成分のエネルギー w_H を音高特徴依存関数から予測される調波成分に対する非調波成分のエネルギーの比 w_H/w_I で割ることで得られる．

3.3　楽器音の合成

調波モデルから調波信号 $s_H(t)$ を，非調波モデル $s_I(t)$ から非調波信号を合成し，以下のように重ね合わせることで最終的な楽器音 $s(t)$ を合成する．

$$s(t) = s_H(t) + s_I(t) \tag{5}$$

ここで，t はサンプリングされた信号のサンプル番地を表す．

3.3.1 調波信号の合成

調波信号 $s_H(t)$ を合成するには，次式によって表現される正弦波重畳モデルを用いる．

$$s_H(t) = \sum_n A_n(t) \exp[j\phi_n(t)] \tag{6}$$

ここで，$A_n(t)$，$\phi_n(t)$ とはそれぞれ n 番目の正弦波の瞬時振幅と瞬時位相である．このモデルでは，各正弦波の振幅と周波数が定常性を持っていることが仮定されている．瞬時位相は，フレーム単位で分析されている音高軌跡をスプライン補間によってサンプル単位を補間したもの $\hat{\mu}(\tau)$ を積分することによって得られる．

$$\phi_n(t) = \phi_n(0) + n \int_0^t \hat{\mu}(\tau) d\tau \tag{7}$$

ここで，$\phi_n(0)$ は任意の初期位相である．正弦波重畳モデルではトラッキングしたピークを瞬時振幅として用いる．調波構造の概形をモデル化した調波モデルにおいては，周波数エンベロープを構成する各ガウス関数の平均に時間エンベロープと調波エネルギーを積算したものをトラッキングしたピークと見なすことができる．特徴量抽出のモデルと楽器音合成のモデルが異なるために合成音が持つ倍音の相対強度は分析対象の楽器音のものとは必ずしも一致しないが，実験的にはこの操作を経ても特徴量が大きく変化することはなかったため，モデルの違いの音色への影響は小さいと考える．よって，瞬時振幅は次式から求めることができる．

$$A_n(t) = \frac{w_H \hat{E}_n(t) v'_n}{\sqrt{2\pi}\sigma} \tag{8}$$

ここで，$\hat{E}_n(t)$ はフレーム単位で求められた $E_n(r)$ をスプライン補間を用いてサンプル単位にしたものである．

3.3.2 非調波信号の合成

非調波信号 $s_I(t)$ を合成するには，オーバラップ加算法を用いる．このとき，非調波成分のエネルギーを乗算した非調波モデルをスペクトログラムと見なして信号に変換する．位相は seed のものをそのまま利用する．

4. 評価実験

本章では，音高依存特徴関数を用いない本手法の枠組みを使った方法をベースライン手法として，本手法と比較評価する．

4.1 評価条件

合成した楽器音の品質を評価するため，合成音と実楽器音との距離を 2 種類の評価尺度を用いて算出した．

(1)　スペクトル距離尺度

$$D_S = \sum_{f,r} (S_{syn}(f,r) - S_{real}(f,r))^2 / R \tag{9}$$

(2)　メル周波数ケプストラム係数（MFCC）距離尺度

$$D_M = \sum_{d,r} (\mathrm{MFCC}_{syn}(d,r) - \mathrm{MFCC}_{real}(d,r))^2 / R \tag{10}$$

ここで，$S_i(f,r)$ と $\mathrm{MFCC}_i(f,d)$ はそれぞれスペクトログラムと MFCC を表し，添え字 $_{syn}$ と $_{real}$ は合成音と実楽器音に対応する．d は MFCC の次元のインデックス，R は楽器音の時間長である．これらの距離が小さいほど，合成音が実楽器音に近いことを示す．スペクトル距離尺度では時間軸・周波数軸ともに線形スケールで表されるので，調波成分に含まれる各倍音ピークの差を主に評価できる．MFCC 距離尺度は聴感上の尺度としてしばしば用いられる．周波数軸は対数スケールで表されるため，調波成分に含まれる各倍音ピークの差だけでなく，それらよりずっと小さなエネルギーしかない非調波成分の差も含めて評価できる．MFCC の次元は 12 次元とした．

評価実験に用いた実楽器音には，RWC 研究用音楽データベースの楽器音データベース RWC-MDB-I-2001 に登録されている楽器音を利用した[24]．このデータベースに含まれる楽器音は，単独発音を半音ごとに収録（サンプリング周波数：44.1 kHz，16 ビット量子化，モノラル）されている．このデータベースから 32 種類の楽器ごとに 3 個体を選択し，フォルテで通常の奏法で演奏されたものを実験に用いた[⋆1]．実験に用いた楽器音の内訳を**表 1** に示す．

各楽器の個体ごとに無作為に 10 等分してのクロスバリデーションによって合成音と実楽器音との距離を求めた．なお，10 等分されたデータのうち，9 つのグループに含まれる楽器音のうちの 1 つを $\mathrm{seed}(S_{syn}(f,r), \mathrm{MFCC}_{syn}(d,r))$ として用い，残りの 1 つのグループに含まれる楽器音のうちの 1 つを距離を測る対象となる実楽器音 $(S_{real}(f,r), \mathrm{MFCC}_{real}(d,r))$ とした．本手法にて楽器音を合成するとき，音高依存特徴関数を近似するための学習データ

⋆1 ビブラート奏法やスタッカート奏法などを除く一般的な奏法を指す．ただし，RWC 研究用音楽データベースのバイオリン音においては，ビブラート奏法が通常奏法とタグ付けされているため，「ビブラート無」とタグ付けされているバイオリン音を用いる．

表 1　本稿の実験で用いた楽器音の内訳
Table 1　Musical instrument sounds used in this paper.

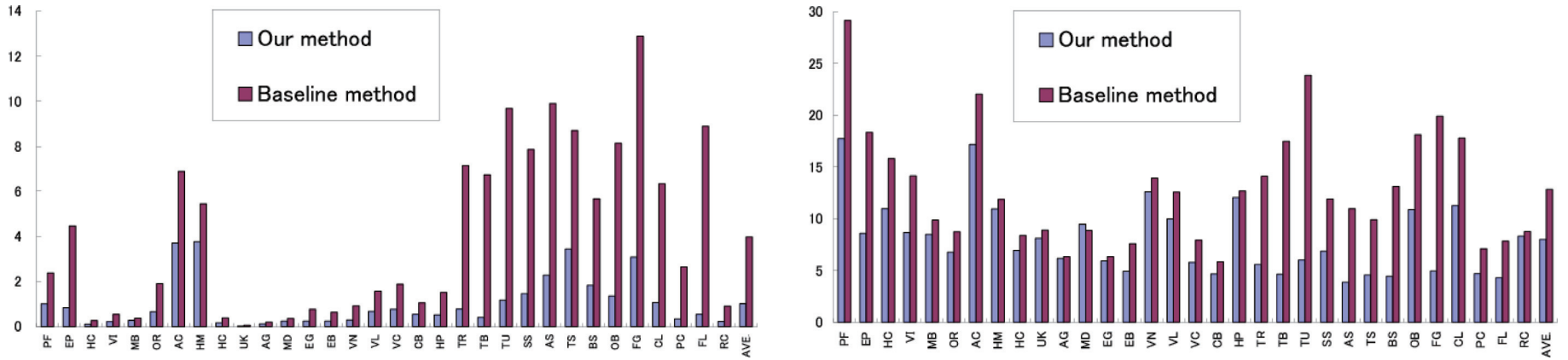
楽器名 （楽器記号）	ピアノ（PF），エレクトリックピアノ（EP），ハープシコード（HC），ビブラフォン（VI）， マリンバ（MB），オルガン（OR），アコーディオン（AC），ハーモニカ（HM）， クラシックギター（HC），ウクレレ（UK），アコースティックギター（AG）， マンドリン（MD），エレキギター（EG），エレキベース（EB）， バイオリン（VN），ビオラ（VL），チェロ（VC），コントラバス（CB）， ハープ（HP），トランペット（TR），トロンボーン（TB），チューバ（TU）， ソプラノサックス（SS），アルトサックス（AS），テナーサックス（TS）， バリトンサックス（BS），オーボエ（OB），ファゴット（FG）， クラリネット（CL），ピッコロ（PC），フルート（FL），リコーダー（RC）
楽器個体数	3 個体
強さ	フォルテのみ．
奏法	通常の奏法のみ
データ数	PF: 264，EP: 206，HC: 178，VI: 104，MB: 145，OR: 178，AC: 141，HM: 101， HC: 111，UK: 71，AG: 111，MD: 123，EG: 111，EB: 88，VN: 138，VL: 126， VC: 134，CB: 111，HP: 241，TR: 103，TB: 96，TU: 90，SS: 99，AS: 99， TS: 98，BS: 98，OB: 96，FG: 120，CL: 120，PC: 98，FL: 111，RC: 75

には，9 つのグループに含まれるすべての楽器音を用いた．たとえば，個体ごとに 88 種類の音高別の楽器音が用意されているピアノの場合，seed および学習データは 79 個，あるいは 80 個となり，距離を測る対象となる実楽器音は 8 個，あるいは 9 個となり，6,968 回が 1 個体内での試行回数となる．上記の評価方法を 32 種類，3 個体の楽器のすべてに適用し，合計 447,772 回の試行を行った．

合成音と実楽器音の間にある，奏者のそのときどきによる発音方法の違いと音高の違いを解消するために，距離を測る対象となる実楽器音の時間エンベロープ $E_n(r)$ と音高軌跡 $\mu(r)$ を事前に分析し，これら 2 つを，距離の比較対象となる合成音のパラメータ値として用いた．さらに，合成音と実楽器音との音量の違いもなくすために，これら楽器音のスペクトログラムの各周波数ビン，各フレームを合計した値が同じになるよう，合成音の音量を調節した．

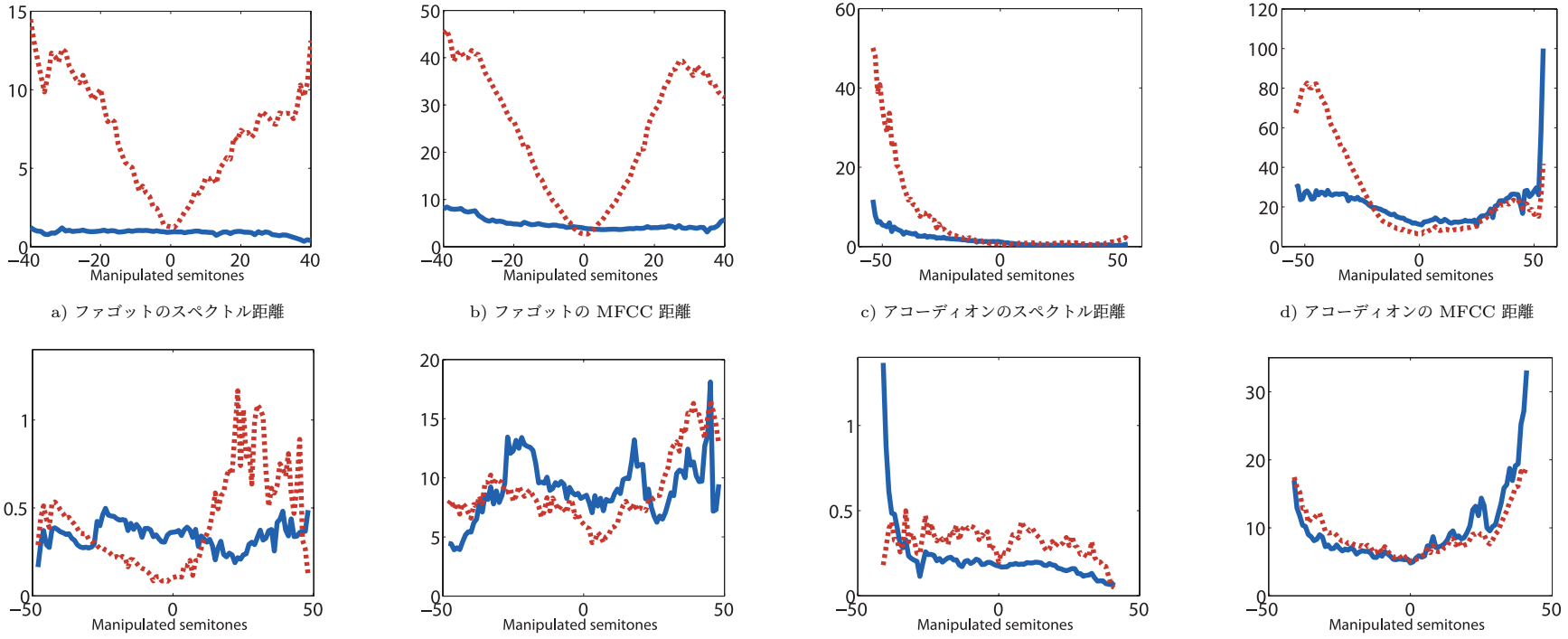
4.2　実験結果・考察

図 6 に本手法とベースライン手法によって合成した楽器音の実楽器音に対する距離を示す．これらの値は楽器個体ごとに得られた距離を平均して示してある．32 種類の楽器のうち，スペクトル距離においてはすべての楽器で減少している．また，MFCC 距離もマンドリンを除いてすべての楽器において減少している．全楽器で平均すると，ベースラインと



a) スペクトル距離　　b) MFCC 距離

図 6　本手法とベースライン手法におけるスペクトル距離と MFCC 距離．スペクトル距離は本手法におけるピアノの距離で正規化している
Fig. 6　Spectrum distance and MFCC distance with our method and the baseline method. Spectrum distance was normalized by distance of piano in our method.



a) ファゴットのスペクトル距離　b) ファゴットの MFCC 距離　c) アコーディオンのスペクトル距離　d) アコーディオンの MFCC 距離

e) マリンバのスペクトル距離　f) マリンバの MFCC 距離　g) マンドリンのスペクトル距離　h) マンドリンの MFCC 距離

図 7　本手法とベースライン手法における半音操作ごとの距離．実線と破線はそれぞれ本手法での距離とベースライン手法での距離を示す
Fig. 7　Distances of semitone in our method and the baseline method. The solid line and the dashed line indicate the distances in our method and the baseline method, respectively.

比較して本手法を用いたことによるスペクトル距離と MFCC 距離の相対的な減少率（改善率）はそれぞれ 64.70%，32.31%となり，音色の音高依存性を扱った本手法の有効性が示されている．

距離の改善率が高かった例として，ファゴット（スペクトル距離：76.02%，MFCC 距離：75.17%）における半音操作ごとの距離を**図 7** a)，b) に示す．ベースライン手法では，音高操作の変化量が増えるにつれて，スペクトル距離と MFCC 距離の両方が増加するのを確認できる．一方，本手法では音高操作の変化量に対しての距離の増加がみられない．音高操作の半音数が小さいときは距離の大きさが手法で逆転しているが，この原因は，音色の特徴量を関数近似することによって発生する誤差である．以下に述べる距離の改善の大きかった楽器においても，音高操作の変化量にかかわらず，距離の増加が抑えられることを確認した．

楽器によって距離の改善が大きくなったことについて，考えられる理由を以下で述べる．

(1)　**楽器の材質の特性**

トランペットやトロンボーン，チューバといったブラス材質の楽器の大きな改善は，材質の特性が強い音高依存性を引き出しているためと考えられる．

(2)　**楽器の発音機構が複雑さ**

ピアノの大きな改善は，88 種類の音高すべてが独立である複雑な発音機構が強い音色の音

高依存性を引き出しているためと考えられる．

一方，アコーディオン（スペクトル距離：46.19%，MFCC 距離：22.08%），マリンバ（スペクトル距離：24.49%，MFCC 距離：13.99%），マンドリン（スペクトル距離：31.21%，MFCC 距離：−6.64%）を例にあげるように，いくつかの楽器では距離の改善がさほどされなかった．上記の楽器の音高操作の変化量に対応する距離を図 7 c)〜h) に示す．

(1)　音色の音高依存性が小さい場合の改善

アコーディオン（スペクトル距離：46.19%，MFCC 距離：22.08%）における音高操作の変化量に対応する距離を図 7 c)，d) に示す．音高を下げる操作では，本手法により両方の距離が改善されている．一方，音高を上げる操作では，両方の距離において本手法による改善がみられない．ベースライン手法においても音高をあげる操作では距離の増加がないところから，アコーディオンのような楽器では高い音高では，音色の音高依存性をさほど持たないためと考えられる．

(2)　音色の音高依存性の高次元への対応

マリンバ（スペクトル距離：24.49%，MFCC 距離：13.99%）における音高操作の変化量に対応する距離を図 7 e)，f) に示す．両手法において音高操作の変化量に対しての距離が複雑に変化しており，マリンバの音色の音高依存性を学習できていないことが分かる．マリンバの音は打楽器音の要素を含み，さらに発音機構がピアノのように各音高で独立である．本稿では，音色の特徴量を 3 次関数で近似することで，音高依存特徴関数を導出したが，マリンバのように音高ごとに複雑に音色が変わる楽器では，3 次関数で近似するのでは不十分であると考えられる．これの解決方法として近似関数の次数を上げると，学習が正確に行われない可能性もある．本稿では，予備実験より全楽器に対して適切な音高依存性を表現できる次数を 3 次としたが，各楽器に対しての次数は今後検討する．

(3)　非調波成分における強い音高依存性への対応

マンドリン（スペクトル距離：31.21%，MFCC 距離：−6.64%）における音高操作の変化量に対応する距離を図 7 g)，h) に示す．スペクトル距離の改善は確認できるが，MFCC 距離においては改善がみられなかった．スペクトル距離の改善は，本手法によって音高に対する倍音ピーク間の相対強度の変化を学習に起因すると考えられる．一方，MFCC 距離の改悪は，合成音の非調波成分の分布の実楽器音のものとの異なりが，原因と考えられる．本手法では，調波成分と非調波成分のエネルギーの比の音高特徴量関数によって，これの音高依存性を扱ってはいるが，分布自体の違いまでは扱っていない．他の撥弦楽器においても，MFCC 距離の改善はさほどみられなかった．撥弦楽器は発音時に高周波数領域に多くの非調波成分を含むことが知られている[19]．この知見から，撥弦楽器の非調波成分の音高依存性が強いことが分かる．いくつかの撥弦楽器の合成音を試聴実験により，調波成分は実楽器音の特徴をとらえたものになっていたが，非調波成分は不自然な音に合成されていることを確認した．

本手法によって合成された楽器音は下記の URL で参照することができる．
http://winnie.kuis.kyoto-u.ac.jp/~abe/IPSJ-SP-09/

5. STRAIGHT との比較

本章では，本手法が従来手法と比べての有効性を評価するために，ボコーダの 1 種である STRAIGHT[8] の最新版と比較評価する．

実験データには打弦・撥弦楽器としてピアノとアコースティックギター，擦弦楽器としてバイオリンとコントラバス，吹奏楽器としてトランペットとアルトサックスの計 6 種類を選択した．これら実験データに対して 4 章と同じくクロスバリデーションを行う．評価尺度には 4 章の評価実験で用いたスペクトル距離と MFCC 距離のほかに，対数スペクトル距離尺度を加える．

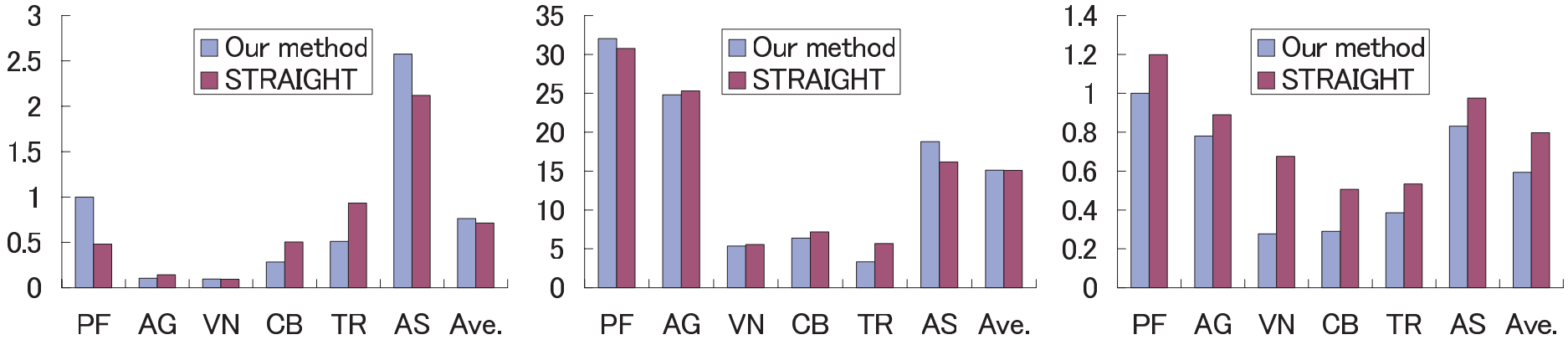
$$D_L = \sum_{f,r} (20 \log_{10} S_{syn} - 20 \log_{10} S_{real})^2 / R \tag{11}$$

パワーが対数スケールでとられている対数スペクトル距離尺度は，調波成分に比べて微小な非調波成分の分布の違いを評価することが可能である．MFCC 距離尺度でも非調波成分の差を含めて評価できるが，メルフィルタバンク内のパワーの合計値をとるため，非調波成分の分布の違いを対数スペクトル距離尺度ほど反映しない．

なお，本実験では 4 章のように奏者のそのときどきによる発音方法の違いを解消することはできないため，実楽器音の音高軌跡のみを合成音のパラメータ値として用いた．

5.1　実験結果・考察

実験結果を **図 8** に示す．これらの値は楽器個体ごとに得られた距離を平均して示してある．スペクトル距離，MFCC 距離においては，楽器によって STRAIGHT に対しての本手法での距離が改善されているものや劣化しているものがあり，全楽器の平均で差は出なかった．全楽器で平均すると，本手法を用いたことによるスペクトル距離と MFCC 距離の改善率はそれぞれ −6.95%，−0.03%となった．一方，対数スペクトル距離尺度はすべての楽器において改善されている．全楽器で平均すると，本手法を用いたことによる対数スペクトル距離の改善率は 25.45%となった．対数スペクトル距離における改善は，本手法によって合



a) スペクトル距離尺度　b) MFCC 距離尺度　c) 対数スペクトル距離尺度

図 8　本手法と STRAIGHT におけるスペクトル距離と MFCC 距離と対数スペクトル距離．スペクトル距離と対数スペクトル距離は本手法におけるピアノの距離で正規化している

Fig. 8　Spectrum distance, MFCC distance and log-spectrum distance with our method and STRAIGHT. Spectrum distance and log-spectrum distance were normalized by distance of piano in our method.

成された音の非調波成分の分布が実楽器音のものと近いことを示している．

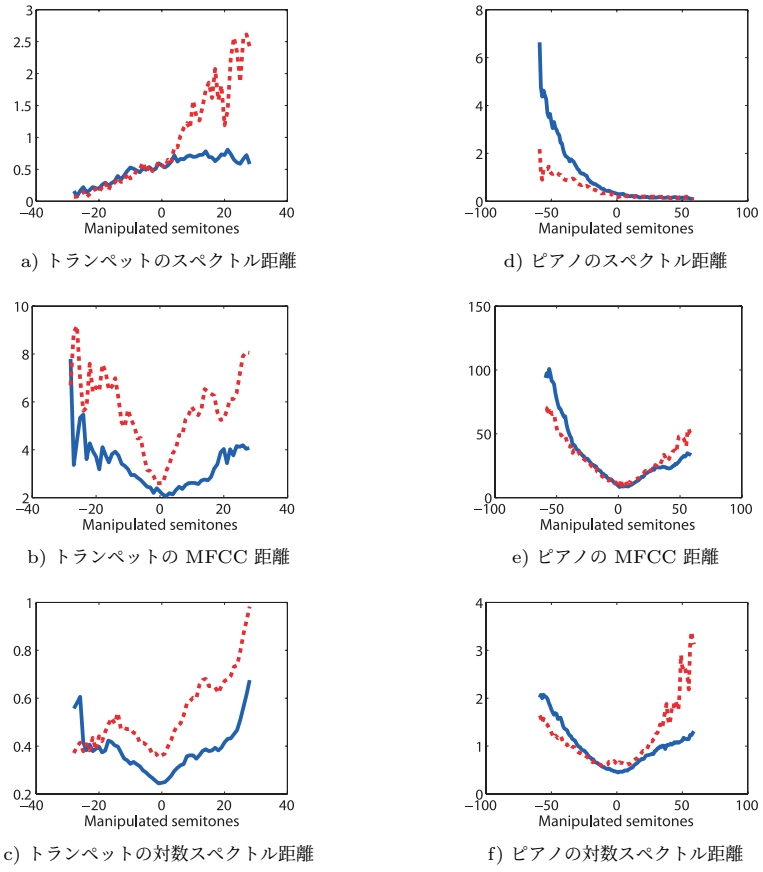
距離の改善率が高かった例として，トランペット（スペクトル距離：45.31%，MFCC 距離：41.15%，対数スペクトル距離距離：0.27%）における半音操作ごとの距離を**図 9** a)，c)，e) に示す．音高を下げる操作では両手法においてスペクトル距離が抑えられているが，音高を上げる操作では本手法に比べて STRAIGHT におけるスペクトル距離が著しく増加している．この原因は，高い音高ではトランペットのスペクトル包絡が大きく変化するためであると考えられる．STRAIGHT ではスペクトル包絡が音高によらず一定と仮定しているので，音高によってスペクトル包絡が変化する楽器音の音色の特徴をとらえることは困難である．一方，本手法ではスペクトル包絡の変化を学習できているために距離が低く抑えられていると考えられる．MFCC 距離はスペクトル距離と対数スペクトル距離の結果を包括したものとなっている．

距離の改善率が低かった例として，ピアノ（スペクトル距離：45.31%，MFCC 距離：41.15%，対数スペクトル距離距離：0.27%）における半音操作ごとの距離を図 9 b)，d)，f) に示す．音高を上げる操作では STRAIGHT に比べて本手法におけるスペクトル距離，MFCC 距離，対数スペクトル距離が抑えられているが，音高を下げる操作では距離が増加している．この原因は，本手法が低音の分析が十分にできていないのが原因である．低音ほど各ピークが周波数軸上で近接し，本手法のモデルでの分析が困難になる．また，ピアノ音および弦楽器のピークは厳密な倍音構造をとらないことが知られている[25]．この現象はインハーモニシティと呼ばれ，弦のスティフネスが大きくなるほど，倍音ピークが厳密な調波構造から離れていく．すなわち，低音にスティフネスの高い弦が使われているピアノの場合，厳密な調波構造を仮定している本手法では低音の分析がより困難となる．分析に用いた本手法のモデルを拡張し，さらに詳細な分析が可能になればこれらの距離の改善を期待できる．

6. おわりに

本稿では，音色の特徴量を数学的なパラメータとして定義し，これを考慮した楽器音の音高操作手法について報告した．我々は Grey が報告した音色の聴感上の知覚の差に対応するスペクトルの要因を参考に，音色特徴量として，(i) 倍音ピーク間の相対強度，(iii) 時間方向の振幅エンベロープ，(ii) 非調波成分の 3 つを定義した．音高操作時には，特徴量 (i)，(iii) の音高依存性を考慮しなければならない．そのため，音高に対する特徴量を 3 次関数で近似し，所望の音高における特徴量の値を予測した．

今後は，音長操作手法についても検討していく．そして，実際に楽曲から分離された楽器音に対して，本手法を適用することを検討していく．このとき，分離音には様々なノイズが



a) トランペットのスペクトル距離　d) ピアノのスペクトル距離

b) トランペットの MFCC 距離　e) ピアノの MFCC 距離

c) トランペットの対数スペクトル距離　f) ピアノの対数スペクトル距離

図 9　本手法と STRAIGHT における半音操作ごとの距離．実線と破線はそれぞれ本手法での距離と STRAIGHT での距離を示す

Fig. 9　Distances of semitone in our method and STRAIGHT. The solid line and the dashed line indicate the distances in our method and STRAIGHT, respectively.

含まれているので，できる限りクリーンな seed を選択することが重要になる．さらに，楽器ごとの適切な音高依存特徴関数の次数の検討やモデルの拡張にも取り組んでいきたい．

謝辞　本研究の一部は，科学研究費補助金（基盤研究（S）），グローバル COE プログラム「知識社会基盤構築のための情報学拠点形成」，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) Gillet, O. and Richard, G.: Extraction and Remixing of Drum Tracks from Polyphonic Music Signals, *Proc. WASPAA*, pp.315–318 (2005).
2) Yoshii, K., Goto, M. and Okuno, H.G.: Drumix: An Audio Player with Real-time Drum-part Rearrangement Functions for Active Music Listening, *IPSJ Journal*, Vol.48, No.3, pp.1229–1239 (2007).
3) 糸山克寿，後藤真孝，駒谷和範，尾形哲也，奥乃　博：楽譜情報を援用した多重奏音楽音響信号の音源分離と調波・非調波統合モデルの制約付パラメータ推定の同時実現，情報処理学会論文誌，Vol.49, No.3, pp.1465–1479 (2008).
4) Uhle, C., Dittmar, C. and Sporer, T.: Extraction of Drum Tracks from Polyphonic Music Using Independent Subspace Analysis, *Proc. ICA*, pp.834–848 (2003).
5) Helen, M. and Virtanen, T.: Separation of Drums from Polyphonic Music Using Non-negative Matrix Factorization and Support Vector Machine, *Proc. EUSIPCO* (2005).
6) Fitzgerald, D., Cranitch, M. and Coyle, E.: Sound Source Separation using Shifted Non-negative Tensor Factorization, *Proc. ICASSP*, Vol.V, pp.653–656 (2006).
7) Dudley, H.: Fundamentals of Speech Synthesis, *J. Audio Eng. Soc.*, Vol.3, pp.170–185 (1955).
8) 河原英紀：Vocoder のもう一つの可能性を探る—音声分析変換合成システム STRAIGHT の背景と展開，日本音響学会誌，Vol.63, No.8, pp.442–449 (2007).
9) Dolson, M.: The Phase Vocoder: A Tutorial, *Computer Music J.*, Vol.10, pp.14–27 (1986).
10) Laroche, J. and Dolson, M.: Improved phase vocoder timescale modification of audio, *IEEE Trans. Speech and Audio Processing*, Vol.7, No.3, pp.323–332 (1999).
11) Robel, A.: A New Approach to Transient Processing in the Phase Vocoder, *Proc. DAFx*, pp.344–349 (2003).
12) Laroche, J. and Dolson, M.: New Phase-Vocoder Techniques for Real-Time Pitch-Shifting, Chorusing, Harmonizing and Other Exotic Audio Modifications, *J. Audio Eng. Soc.*, Vol.47, No.11 (1999).
13) McAulay, R. and Quatieri, T.: Speech Analysis/Synthesis based on a Sinusoidal Representation, *IEEE Trans. Acoust., Speech, and Signal Processing*, pp.744–754 (1986).
14) Serra, X. and Smith, J.: Spectral Modeling Synthesis:A Sound Analysis/Synthesis Based on a Deterministic plus Stochastic Decomposition, *Computer Music J.*, Vol.14, pp.12–24 (1990).
15) Osaka, N.: Concatenation and stretch/squeeze of musical instrumental sound using

morphing, *Proc. ICMC* (2005).
16) Godsill, S. and Davy, M.: Bayesian Harmonic Models for Musical Pitch Estimation and Analysis, *Proc. ICASSP*, Vol.II, pp.1769–1772 (2002).
17) Jinachitra, P.: Constrained EM Estimates for Harmonic Source Separation, *Proc. ICASSP*, Vol.VI, pp.609–612 (2003).
18) 亀岡弘和，小野順貴，嵯峨山茂樹：正弦波重畳モデルのパラメータ最適化アルゴリズムの導出，電子情報通信学会技術研究報告，Vol.106, No.432, pp.49–54 (2006).
19) Grey, J.M.: Multidimensional perceptual scaling of musical timbres, *J. Acoust Soc. Am.*, Vol.61, No.5, pp.1270–1277 (1977).
20) 日本工業規格，JIS-Z8109 音響用語（音声聴覚・音楽）(1961).
21) Marozeau, J., Cheveigne, A., McAdams, S. and Winsberg, S.: The dependency of timbre on fundamental frequency, *J. Acoust Soc. Am.*, Vol.114, No.5, pp.2946–2957 (2003).
22) 北原鉄朗，後藤真孝，奥乃　博：音高による音色変化に着目した楽器音の音源同定：F0依存多次元正規分布に基づく識別手法，情報処理学会論文誌，Vol.44, No.10, pp.2448–2458 (2003).
23) 早坂寿雄：楽器の科学，電子情報通信学会 (1992).
24) Goto, M., Hashiguchi, H., Nishimura, T. and Oka, R.: RWC Music Database: Music Genre Database and Musical Instrument Sound Database, *Proc. ISMIR*, pp.229–230 (2003).
25) Fletcher, H., Blackham, E. and Stratton, R.: Quality of Piano. Tones, *J. Acoust. Soc. Am.*, Vol.34, No.6, pp.749–761 (1962).

(平成 20 年 5 月 14 日受付)
(平成 20 年 11 月 5 日採録)

安部　武宏（学生会員）
2007 年同志社大学工学部知識工学科卒業，現在，京都大学大学院情報学研究科知能情報学専攻修士課程に在籍中．音楽情報処理，音楽音響等の研究に従事．

糸山　克寿（学生会員）
2006 年京都大学工学部情報学科卒業，2008 年同大学情報学研究科知能情報学専攻修士課程修了．現在，同専攻博士後期課程に在籍中．2008 年より日本学術振興会特別研究員（DC1）．音楽情報処理，音楽鑑賞インタフェース等の研究に従事．

吉井　和佳（正会員）
2006 年京都大学大学院情報学研究科知能情報学専攻修士課程修了．2008 年同大学院情報学研究科博士後期課程修了．博士（情報学）．2008 年 4 月より産業技術総合研究所情報技術研究部門研究員．FIT2004 論文賞，インタラクション 2006 ベストインタラクティブ発表賞，情報処理学会山下記念研究賞等 9 件受賞．音楽情景分析，音楽推薦，音楽ロボット等の研究に従事．電子情報通信学会，IEEE 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．同年京都大学情報学研究科助手．2007 年より助教，現在に至る．京都大学博士（情報学）．音声対話システムの研究に従事．2008〜2009 年米国カーネギーメロン大学客員研究員．情報処理学会平成 16 年度山下記念研究賞，FIT2002 ヤングリサーチャー賞等受賞．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）

1969 年生．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）．この間，2005 年より早稲田大学ヒューマノイド研究所客員准教授，2006 年より理化学研究所脳科学総合研究センター客員研究員を兼務．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．2001 年日本機械学会論文賞，IEA/AIE-2005 最優秀論文賞等を受賞．RSJ，JSME，JSAI，SICE，IEEE 等会員．

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，科学技術振興事業団，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 最優秀論文賞，IEEE/RSJ IROS-2001，2006 Best Paper Nomination Finalist，IROS-2008 Award for Entertainment Robots and Systems Nomination Finalist 2 件，第 2 回船井情報科学振興賞等受賞．人工知能学会，日本ロボット学会，日本ソフトウエア科学会，ACM，IEEE，AAAI，ASA 等各会員．