Vol.2009-MUS-81 No.10
2009/7/30

残差スペクトルモデルによる伴奏・残響成分抑制に基づいた楽器演奏分析合成の高精度化

安良岡直希[†1]　糸山克寿[†1]　高橋徹[†1]
尾形哲也[†1]　奥乃博[†1]

本報告書では，楽器演奏音響信号の分析合成における，入力中の伴奏音や残響成分を抑制した分析手法を報告する．対象演奏パートの楽譜情報に合致しないスペクトル成分を表現する残差スペクトルモデルを導入し，これを用いて伴奏や残響を含む音響信号から対象の演奏を効率よく分離する．調波非調波統合音モデルに用いた演奏分析をこの分離と同時に行い，分析された音モデルを用いて未知楽譜への演奏を合成する．評価実験では，伴奏付き演奏に対する分析精度が本手法によりスペクトル距離において平均 35.2%改善し，また残響を含む演奏に対する分析合成精度の低下を回避できる事が確認された．

Improvement of Performance Analysis-and-Symthesis Method by using Residual Spectrum Model for Reduction of Accompaniment or Sound Reverberation

Naoki Yasuraoka ,[†1] Katsutoshi Itoyama ,[†1]
Toru Takahashi ,[†1] Tetsuya Ogata [†1]
and Hiroshi G. Okuno [†1]

This paper presents a musical performance analysis-and-synthesis method using residual model for reduction of accompaniment or sound reverberation. The residual model is designed for representing spectrum that the score does not convey about the performance. This leads to an efficient extraction of a performed part from accompanied and/or reverberant audio source. The extraction is performed simultaneously with estimation of musical tone models that represent both harmonic and inharmonic sound of the performance. Using the estimated tone models, a new performance sound corresponding to a new given score is synthesized. An experiment showed that the spectral distance of one instrument part extracted from polyphonic audio source improved by 35.0 points by incorporating the residual model. Another result showed the effectiveness of our method under reverberant source.

1. はじめに

計算機による楽曲製作支援技術の一つに，楽器演奏合成がある．これは，楽譜に相当するデータを与えると，その演奏の音響信号を合成するものであり，実際の楽器演奏を録音することなく楽曲を製作することができる．特に，人間による楽器演奏に含まれる音量や発音タイミング，音色の揺らぎ (演奏表情と呼ぶ) を反映した演奏合成を行うことで，より自然で高品質な演奏が得られる．演奏表情の生成は，現在も熟練者による手作業のデータ入力で行われることが多く，演奏表情の自動生成が計算機によって可能になれば，楽曲製作現場での作業量の削減が期待される．実演奏からの演奏表情パターン獲得と，未知楽譜に対する演奏表情の生成についての研究はその需要に答えるものである．

演奏表情の中でも音量や発音タイミングは演奏知覚差を生む重要な要素であり，この二つの生成に関する研究は多い [1–3]．これらは，MIDI 出力機能付き楽器などを用いて実演奏の音量や発音タイミングを計測し，その揺らぎをモデル化することで，未知楽譜に対する音量・発音タイミングデータを生成する．しかし，この枠組みには実演奏の音響的な情報が欠落しており，音色に関わる演奏表情の再現は不可能であった．

音色に関わる演奏表情の再現を目指し，我々は演奏の音響信号そのものを用いた演奏分析合成の研究を進めている [4]．これは，単音のパワースペクトルを表現する数理モデル (以下，音モデルと呼ぶ) に基づき実演奏音響信号を分析し，各単音ごとの音モデルパラメータ値の差異を演奏表情と見なし楽譜構造と対応づけることで，未知楽譜に対する演奏音響信号を合成するものである．この方法により，演奏生成において音響的再現性の議論が可能になった．しかし，分析に用いる演奏音響信号は無伴奏でかつ残響が少ないものでなければならないという制約があった．

本報告書では，残差スペクトルモデルを導入した分析により伴奏・残響成分を抑制することで，それらを含む演奏音響信号を分析合成に用いた場合の音響的再現性の低下を軽減させる手法について述べる．対象演奏パートの楽譜情報に合致しないスペクトル成分を表現する残差スペクトルモデルを定義し，入力音響信号のパワースペクトルを分析対象演奏の音モデルとこの残差スペクトルモデルに適応的に分配する．音モデルは音源分離にも利用される調波非調波統合モデル [5] をもとに設計し，分配と音モデル推定を一つの枠組みで実現する．なお，必要な事前情報の増加を避けるため，伴奏についての楽譜は用いない．

†1 京都大学大学院情報学研究科知能情報学専攻
Dept. of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University

Vol.2009-MUS-81 No.10
2009/7/30

2. 音モデルの定義と残差スペクトルモデルを援用した実演奏分析

本章では, 実演奏音響信号の分析における, 残差スペクトルモデルと分析対象演奏の音モデル双方の同時推定方法について述べる.

2.1 音モデルの定義

単音のパワースペクトルを表現する数理モデルとして, 我々は音響心理学の知見 [6] に基づき楽器音の調波成分と非調波成分を別個に扱えるモデルを設計する. 単音のパワースペクトル $M(t,f)$ は, その調波成分に対するモデル (調波構造モデル) $M^{(H)}(t,f)$ と, 非調波成分に対するモデル (非調波構造モデル) $M^{(I)}(t,f)$ の和で表現される. 各重みを $w^{(H)}, w^{(I)}$ とすると,

$$M(t,f) = w^{(H)}M^{(H)}(t,f) + w^{(I)}M^{(I)}(t,f) \tag{1}$$

となる. ここで, t と f はそれぞれ時間と周波数を表す.

調波モデル $M^{(H)}(t,f)$ は, 各倍音ピークがそれぞれ一つのガウス関数に対応する混合ガウス分布でモデル化される.

$$M^{(H)}(t,f) = \sum_n F^{(H)}(n,t,f)E^{(H)}(n,t) \tag{2}$$

$$F^{(H)}(n,t,f) = v(n)\mathcal{N}\left(f-\mu(n,t),\sigma^2\right) \tag{3}$$

$$\mu(n,t) = n\mu(t)\sqrt{1+Bn^2} \tag{4}$$

ここで, n は倍音のインデックスを表す. $F^{(H)}(n,t,f)$ と $E^{(H)}(n,t)$ はそれぞれ調波成分の周波数方向及び時間方向のエンベロープに相当する (**図 1** 及び **図 2** 参照) [⋆1]. $\mu(n,t)$ は各時刻ごとに与えられるため, これは倍音ピークの時間軌跡を表す. B は非調和度を表す. 弦楽器音には, 各倍音の周波数が弦の剛性や長さに応じて基本周波数の厳密な整数倍よりも若干高くなる性質 (非調和性, インハーモニシティ) が存在する [7]. 式 (4) はインハーモニシティの理論式に基づき倍音ピークの周波数軸での配置間隔を拡げる.

非調波構造モデル $M^{(I)}(t,f)$ はノンパラメトリック表現の周波数方向相対強度 $F^{(I)}(t,f)$ 及び時間エンベロープ $E^{(I)}(t)$ の積で表現する.

$$M^{(I)}(t,f) = F^{(I)}(t,f)E^{(I)}(t), \tag{5}$$

非調波構造モデルは, 「調波構造モデルで表現されない成分」を表現する. これは非常に高い自由度を持つが, パラメータ推定時に $F^{(I)}(t,f)$ 及び $E^{(I)}(t)$ に制約条件を課すことで, 本来調波構造モデルが表現するべき調波成分をこちらで表現されることを避ける.

⋆1 本手法ではパワースペクトルを得るための Short-Time Fourier Transform (STFT) 解析時の窓掛けにガウス窓を用いるため, もともと周波数軸に対して鋭いピークであった調波成分はガウス関数の形状をとる.

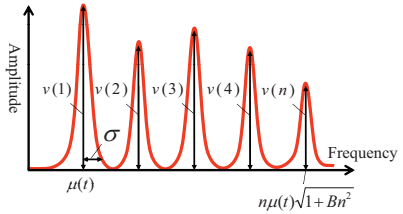

図 1 周波数方向エンベロープ

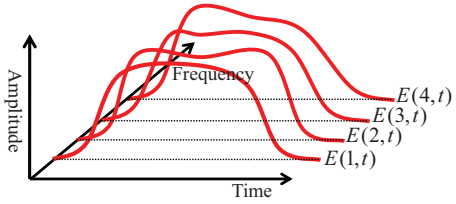

図 2 時間方向エンベロープ

各モデル内でパラメータ値に一意性を持たせるため, パラメータ間に以下の制約を課す.

$$\sum_n v(n) = 1, \qquad \forall n : \int E^{(H)}(n,t)\,dt = T \tag{6a}$$

$$\int E^{(I)}(t)\,dt = T, \qquad \forall t : \int F^{(I)}(t,f)\,df = 1 \tag{6b}$$

ここで T は時間長を表す. また, これまでの全てのパラメータは負にはならないとする.

2.2 残差スペクトルモデルを援用した実演奏分析

本節では, 伴奏や残響を含む混合音中から分析対象の演奏を分離すると同時に, 音モデルパラメータの推定を行う方法について述べる.

演奏分析の処理は以下の二行程を交互に繰り返すことで達成される.

(1) 分離：音モデル, 残差モデルの情報を用いて演奏音響信号のパワースペクトル $S(t,f)$ を, 対象パートの第 k 発音に対するスペクトル $S_k^{(M)}(t,f)$ と残差 $S^{(R)}(t,f)$ に分離する

(2) モデル推定：分離された各パワースペクトルの情報を用いて音モデル $M_k(t,f)$ 及び残差モデル $M^{(R)}(t,f)$ 内の各パラメータを更新する

以後, 音モデル, 残差モデルが表現するスペクトルをモデルスペクトル, 分配関数によって音響信号から分離されたスペクトルを分離スペクトルと呼ぶ.

$S(t,f)$ の分配は, これを各単音ごと分離する分配関数 $\Delta_k^{(M)}(t,f)$ と, 各単音をさらに調派成分と非調波成分とに分離する分配関数 $\Delta_k^{(H)}(n,t,f)$ 及び $\Delta_k^{(I)}(t,f)$ を用いて表現する, 即ち,

$$S_k^{(M)}(t,f) = \Delta_k^{(M)}(t,f)S(t,f) \tag{7a}$$

$$S_k^{(H)}(n,t,f) = \Delta_k^{(H)}(n,t,f)S_k^{(M)}(t,f) \tag{7b}$$

$$S_k^{(I)}(t,f) = \Delta_k^{(I)}(t,f)S_k^{(M)}(t,f) \tag{7c}$$

となる. ここで, $S_k^{(H)}(n,t,f)$ 及び $S_k^{(I)}(t,f)$ はそれぞれ一つの音モデル内での調波成分及び非調波成分に相当する分離スペクトルを表す. また, 分配関数は定義域で 0 以上 1 以下とする.

残差スペクトルモデル $M^{(R)}(t,f)$ とその分配関数 $\Delta^{(R)}(t,f)$ を導入する. $M^{(R)}(t,f)$ はノンパラメトリック関数として定義され, 分析対象演奏の音モデルが表現しきれないスペクトル成分, すなわち伴奏や残響を吸収するように, 分析の過程で更新される. $S(t,f)$ との関係は他

の分配関数と同様に以下の形をとる.

$$S^{(R)}(t,f) = \Delta^{(R)}(t,f)S(t,f) \tag{8}$$

分離スペクトルの間に以下の関係式を立て, パラメータ値の一意性を持たせる.

$$S(t,f) = \sum_k S_k^{(M)}(t,f) + S^{(R)}(t,f) = \sum_k \left(\sum_n S_k^{(H)}(n,t,f) + S_k^{(I)}(t,f) \right) + S^{(R)}(t,f) \tag{9}$$

このとき同時に, 分配関数は以下の条件を満たすことになる.

$$\forall t,f : \sum_k \Delta_k^{(M)}(t,f) + \Delta^{(R)}(t,f) = 1 \tag{10a}$$

$$\forall k,t,f : \sum_n \Delta_k^{(H)}(n,t,f) + \Delta_k^{(I)}(t,f) = 1 \tag{10b}$$

分離ステップでは, 音モデル及び残差モデルのパラメータを固定し, それらが表現するパワースペクトルに各々の分配スペクトルが近づくように分配関数 $\Delta_k^{(M)}(t,f), \Delta_k^{(H)}(n,t,f), \Delta_k^{(I)}(t,f)$ 及び $\Delta^{(R)}(t,f)$ を更新し, これを用いて入力スペクトルを分離する. モデル推定ステップでは逆に, 分離スペクトルを固定し, それらにモデルスペクトルが近づくように音モデル及び残差モデルのパラメータ $w_k^{(H)}, w_k^{(I)}, \sigma, \mu_k(t), B_k, w_k^{(I)}, E_k^{(I)}(t), F_k^{(I)}(t,f)$ 及び $M^{(R)}(t,f)$ を更新する. この二つの更新は, モデルスペクトルと分離スペクトルとの間の Kullback-Leibler 距離の和に基づく共通のコスト関数 Q の最小化問題として定義される.

$$\begin{aligned} Q = \sum_k \Bigg(& \sum_n \iint S_k^{(H)}(n,t,f) \times \log \frac{S_k^{(H)}(n,t,f)}{w_k^{(H)} E_k^{(H)}(n,t) F_k^{(H)}(n,t,f)} \, dt\, df \\ & + \iint S_k^{(I)}(t,f) \log \frac{S_k^{(I)}(t,f)}{w_k^{(I)} E_k^{(I)}(t) F_k^{(I)}(t,f)} \, dt\, df \\ & + \beta^{(HE)} \iint \tilde{E}_k^{(H)}(t) \log \frac{\tilde{E}_k^{(H)}(t)}{E_k^{(H)}(n,t)} \, dt\, df \\ & + \beta^{(IFS)} \iint \tilde{F}_k^{(I)}(t,f) \log \frac{\tilde{F}_k^{(I)}(t,f)}{F_k^{(I)}(t,f)} \, dt\, df \\ & + \beta^{(W)} \bar{w}_p^{(H)} / \bar{w}_p^{(I)} \log \frac{\bar{w}_p^{(H)} / \bar{w}_p^{(I)}}{w_k^{(H)} / w_k^{(I)}} + \beta^{(V)} \sum_n \bar{v}_p(n) \log \frac{\bar{v}_p(n)}{v_k(n)} \Bigg) \\ & + \iint S^{(R)}(t,f) \log \frac{S^{(R)}(t,f)}{\gamma M^{(R)}(t,f)} \, dt\, df \end{aligned} \tag{11}$$

これは, 制約条件 (6) 及び (10) の下での最適化問題であり, Lagrange の未定乗数法を用いて解くことができる. この分離とモデル推定の繰り返しは, Expectation-Maximization (EM) アルゴリズムと解釈でき, 局所最適状態へとモデルを収束させることができる. 収束時には, 分離スペクトル $S_k^{(M)}(t,f)$ とモデルスペクトル $M_k(t,f)$ がそれぞれ十分に等しくなる.

最終行の $\gamma\,(0 < \gamma \leq 1)$ は残差スペクトルへの分配重みである. 分離とモデル推定の反復過程において, γ を 0 に近い値から少し大きく設定していくことで, 音モデルが十分に推定されていない初期の段階で分析対象演奏のスペクトル成分を残差モデルが「奪う」のを避ける.

第 3 行から 5 行までは, 調波構造モデルと非調波構造モデルが, それぞれ目的のスペクトル成分に効率よく適応するように与えた制約条件であり, 定数として与える重み β を伴って定義している. 個々の制約式は, 式中に現れているモデルパラメータを対応する変数に近づける効果を持つ. $\tilde{E}_k^{(H)}(t)$ は調波成分の時間方向エンベロープ $E^{(H)}(n,t)$ を倍音方向に平均したものであり, 本来 $v(n)$ が表現すべき倍音ピーク間の相対強度が $E^{(H)}(n,t)$ の中に取り込まれてしまうのを抑制する. $\tilde{F}_k^{(I)}(t,f)$ は $F^{(I)}(t,f)$ を周波数方向に平滑化したものであり, 周波数方向に鋭い起伏を持つ調波成分が非調波構造モデルの中に取り込まれるのを抑制する. 残り 3 変数は, 各パラメータを同一演奏内の同一音高 p の単音全体で平均したものであり, 演奏中で大きく変化することはないと考えられるパラメータを補正する.

音モデルパラメータのうち, σ は主に STFT 解析時に用いる窓関数にのみ依存すると考えられ, すべての音モデルに対して共通の値を用いる. 同様に非調和度 B も, 同じ音高内では演奏表現には寄らず一定であるため, 同一音高内で共通の値 $\tilde{B}_p$ を用いる.

3. 演奏合成までの流れ

本章では, 分析した音モデルと楽譜構造との関係から, 与えられた未知楽譜の各単音に対する音モデルを算出し, 演奏を合成するまでの処理について述べる.

3.1 楽器音モデルと演奏表情モデル

実演奏から分析された音モデルをさらに以下の二要素に分解する.

(1) 楽器音モデル：実演奏から分析された複数の音モデルを音高ごとに平均したもの

(2) 演奏表情モデル：楽器音モデルと個々の音モデルとのパラメータ比を算出したもの

前者について, 楽曲中のあらゆる箇所で発せられた発音の音モデルを平均することで, 音モデル中の楽譜構造に依存しない成分を抽出する. これは主に楽器自体の音響特徴によって構成されると考えられるため, これを楽器音モデルと呼ぶ. 平均をとる範囲を音高[⋆1]毎としたのは, 楽器音の音色特徴に見られる音高依存性 [8]. を考慮するためである. 後者は各単音中の楽器音モデルで表現されなかった成分に相当する. こちらは楽譜構造に依存する演奏表情によって生じたと考えられるため, 演奏表情モデルと呼ぶ.

⋆1 実際には, 音名ごと, すなわち半音単位の集合を想定している

3.1.1 楽器音モデルの算出

楽器音モデル算出において, 異なる音長の音モデル平均化時の音色特徴保存のためにオフセット情報を用いる. 一般に楽器音は立ち上がり→定常 (または減衰) →立ち下がりという状態を持ち, 音長の違いは主に定常状態の長さの違いとなる. 立ち下がり時刻として, 一定以上の強さを持っていたエンベループが最も急降下した時刻をオフセット $t^{(\mathrm{off})}$ と定める.

$$t^{(\mathrm{off})} = \max\left(t|\mathrm{abs}[\frac{d\tilde{E}^{(H)}(t)}{dt}] \leq \varepsilon\tilde{E}^{(H)}(t) \geq \kappa\right) \tag{12}$$

ここで, κ は定常状態に相当するパワーを示す閾値である. なお, max を min で置き換えることで音の立ち上がり時刻に対応するオンセットとして定義でき, これは 3.2.1 節で用いる.

オフセット位置にて各単音の音モデルパラメータを分割したのち, 手前を単音の開始位置で, 後ろを単音の終了位置でアライメントして平均をとる. 音高が p である単音の集合を $\mathscr{P}(p)$ とし, この集合に対する平均モデルを $\bar{M}_p(t,f)$ とする. このモデル中の非時系列パラメータ $\bar{w}_p^{(H)}$, $\bar{w}_p^{(I)}$, $\bar{v}_p(n)$ については単純に加算平均をとる. 時系列パラメータ $\bar{\mu}_p(t)$, $\bar{E}_p^{(H)}(n,t)$, $\bar{E}_p^{(I)}(t)$, $\bar{F}_p^{(I)}(t,f)$ については, オフセットで分割し, 以降を集合中の最長の単音の音長 ($\bar{T}_p$ とする) でアライメントして平均化する. パラメータを便宜的に $\bar{\alpha}_p(t)$ と表すことにすると,

$$\bar{\alpha}_p(t) = \frac{\sum_{k\in\mathscr{P}(p)}\left(\alpha_k(t)W_k^{(1)}(t) + \alpha_k\big(t-(\bar{T}_p-T_k)\big)W_k^{(2)}(t)\right)}{\sum_{k\in\mathscr{P}(p)}\left(W_k^{(1)}(t)+W_k^{(2)}(t)\right)} \tag{13}$$

となる. ただし, T_k は 第 k 発音の時間長を表し,

$$W_k^{(1)}(t) = \begin{cases} 1 & (t < t_k^{(\mathrm{off})}) \\ 0 & (t_k^{(\mathrm{off})} \leq t) \end{cases}, \qquad W_k^{(2)}(t) = \begin{cases} 0 & (t < t_k^{(\mathrm{off})} + \bar{T}_p - T_k) \\ 1 & (t_k^{(\mathrm{off})} + \bar{T}_p - T_k \leq t) \end{cases} \tag{14}$$

である. 分母は各時刻の単音の加算数を表し, それで割ることで平均が計算される⋆1.

3.1.2 演奏表情モデルの算出

演奏表情モデル $\check{M}_p(t,f)$ は各単音のモデルパラメータの楽器音モデルパラメータ値からの比で定義する. 楽器音モデルの算出と同様に, 非時系列パラメータ $\check{w}_k^{(H)}$, $\check{w}_k^{(I)}$ については単純に比をとる. 時系列パラメータ $\check{\mu}_k(t)$, $\check{E}^{(H)}(n,t)$, $\check{E}^{(I)}(t)$, $\check{F}^{(I)}(t,f)$ については, 楽器音モデルのオフセット以降の系列を手前にシフトさせることで対象音と長さを揃えたモデルパラメータとの比をとる. 上記パラメータを便宜的に $\check{\alpha}_l(t)$ と表すことにすると,

$$\check{\alpha}_k(t) = \frac{\alpha_k(t)}{\bar{\alpha}_p(t)W_k^{(1)}(t) + \bar{\alpha}_p\big(t-(\bar{T}_p-T_k)\big)W_k^{(2)}(t)} \tag{15}$$

となる. 前節の方法で楽器音モデルを作成した場合, 楽器音モデルの音長は同一音高内の全単音の時間長に最大値に等しいため, 楽器音モデルの方が短いということは起こりえない.

⋆1 なお, これは特徴量系列を分断する処理であるため, 実際には分断した周辺で何かしらの平滑化処理を行うことを想定している.

3.2 未知楽譜が与えられたときの演奏合成

本節では, 任意の未知楽譜に対する音モデル算出方法, 及びその再合成方法について述べる.

3.2.1 音モデルパラメータの算出

未知楽譜中の各単音に対する音モデルは, 連続 2 音の音高遷移パターンの一致という指標に基づき演奏表情モデルを再構成し, それを楽器音モデルに掛け合わせることで得られる. 今, 未知楽譜中の第 l 発音:音高 p_l , 音長 T_l に対する音モデルの算出を考える. このとき, 直前の単音からの音高遷移パターン (p_{l-1}, p_l) と一致する分析済み演奏中の連続 2 音 (p_{k-1}, p_k) のうち, その 2 音の時間長差 $|T_{k-1}-T_{l-1}|+|T_k-T_l|$ が最小の単音 k を選出する. この単音に対する演奏表情モデルを $\check{M}_l(t,f)^{1}$ とする. 同様に未知楽譜中の第 l 発音とその直後 $l+1$ の組に対して同様に演奏表情モデル $\check{M}_l(t,f)^{(2)}$ が選出できる.

この二つの演奏表情モデルを用いて, 第 l 発音の演奏表情モデル $\check{M}_l(t,f)$ を次のように定義する. まず, 非時系列パラメータについては二つのモデルの平均値を用いる. 時系列パラメータについては, $\check{M}_l(t,f)^{(1)}$ の時間長を T_l にあわせたもの ($\check{\alpha}_l^{(1)}(t)$ とする) と $\check{M}_l(t,f)^{(2)}$ の時間長を T_l にあわせたもの ($\check{\alpha}_l^{(2)}(t)$ とする) へとその値が滑らかに変化するような新たな系列 $\check{\alpha}_l(t)$ を作成する.

$$\check{\alpha}_l(t) = \left\{\left(1-\frac{t}{T_l}\right)\check{\alpha}_l^{(1)}(t) + \frac{t}{T_l}\check{\alpha}_l^{(2)}(t)\right\} \tag{16}$$

楽器音モデルは当該単音の音高のもの $\bar{M}_p(t,f)$ を選出し, 時間長を T_l にあわせた後に $\check{M}_l(t,f)$ と掛け合わせる. この結果, 未知楽譜中の第 l 発音の音モデル $M_l(t,f)$ が得られる.

各モデルの時間長を伸縮させる際, 長さを伸ばす場合には, オンセットとオフセットの間のみを伸張させることで, 楽器音の立ち上がり及び立ち下がり部分の変化による音色特徴の崩れを抑える. また, 分析済み演奏に所望の音高遷移パターンが存在しない場合は, 最も近い遷移パターンを持つ部分を流用し, その基本周波数 $\mu(t)$ に適切に変えて用いる.

3.2.2 楽器演奏音響信号の合成

正弦波重畳モデルによって, 調波構造モデルから調波成分に相当する時間領域信号を合成する. なお, 調波構造モデルは STFT のシフト幅以上の分解能を持たないため, 時間領域での各サンプルごとの瞬時振幅はスプライン補間によって算出する. 非調波成分に相当する時間領域信号は, 非調波構造モデルの逆 STFT から合成する. この 2 つを足し合わせ, さらにそれを全単音に対して足し合わせることで, 未知楽譜に対する楽器演奏音響信号が合成される.

Vol.2009-MUS-81 No.10
2009/7/30

表 1 実験 1 で用いた楽曲と楽器パート

Musical pieces		Instrument part used as source		
Classical	No.12	FL, CB, VC, VL, VN1, VN2	No.37	PN, VN
	No.13	VC, VL, VN1, VN2	No.39	PN, VN
	No.16	CL, VC, VL, VN1, VN2	No.42	HP, VC
Jazz	No.22	TP, PN	No.33	FL, PN
	No.24	SA, PN	No.34	FL, PN
	No.32	VI, PN	No.41	SA, PN

FL:Flute, CB:Contrabass, VC:Cello, VL:Viola, VN:Violin,
CL:Clarinet, PN:Piano, TP:Trumpet, SA:Alto Sax, VI:Vibraphone

4. 評価実験

本章では, 本手法に対して行った 2 つの評価実験 (実験 1, 実験 2) について述べる. 実験 1 は本手法の 2 章で言及した分析手法を, 複数楽器混合音響信号を用いて評価する. 実験 2 では本手法全体を, 残響を持つ無伴奏実演奏音響信号を用いて評価する. どちらの実験も, 残差スペクトルモデルを用いない方法 (baseline) との比較を行っている. また, 手法の評価尺度として、参照すべき音響信号と合成信号と間のパワースペクトル距離を用いた. 実験条件は, サンプリング周波数 44100Hz, STFT フレーム長 1024, シフト幅 128 であり, また残差モデル初期重みは 0.1 とした.

4.1 実験 1:伴奏付き演奏音響信号の分離

4.1.1 実験条件

実験 1 では複数パート混合の楽器演奏を分析に用い, 残差スペクトルモデルの導入によって特定の単一パート分析合成精度がどのように変化するかを確認する. この実験における参照すべき音響信号とは混合前の単一パートの演奏音響信号を指す. 市販の実楽曲では単一パートの音響信号は得られないため, 本実験では standard MIDI file (SMF) を MIDI 音源で合成した音響信号を用いる. SMF は, RWC Music Database: Jazz Music and Classic Music [9] から Jazz と Classic でそれぞれ 6 曲ずつ選出し, その中のドラムを除く各楽器パート延べ 33 を実験単位とした. あるパートに対する実験をする際は, 同じ曲中の別のパートの情報は一切用いない. 例えば, ある曲 A のフルートパートの分析合成実験を行うときは, 曲 A 中のバイオリンパートは未知の伴奏として扱われ, バイオリンパート単独の音響信号やその楽譜は用いない. 具体的に使用した楽曲と楽器パートについて表 1 に纏めた.

4.1.2 実験結果と考察

図 3 は本手法, ベースライン手法の調波成分のみ (baseline-harmonic), ベースライン手法全体 (baseline-all), の 3 指標に対する線形スペクトル距離を, ジャンル毎・楽器毎 (2 回以上現れたもののみ) で平均した実験結果である[⋆1]. ベースライン手法では楽譜と合致しないスペクトル成分を無視できず, 入力混合音中の全成分は分析対象の音モデルで適応され, 特に比較的自由度の高い非調波成分へと吸収されるため, baseline-harmonic という指標を設けた.

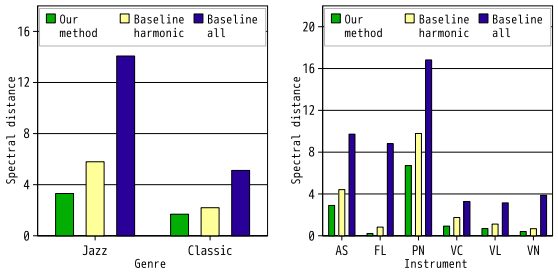

図 3 演奏分析実験結果 (スペクトル距離) ジャンル別 (左) 及び楽器別 (右)

本手法はベースライン手法と比較して Jazz 曲において 42.9%, Classic 曲において 23.1%, 全体で 35.0%の改善を示した. これは, 残差スペクトルモデルの導入によって, 分析対象の演奏をより正確に分離・分析できることを示している. ベースライン手法の調波成分のみと比較しても本手法の方が精度が高いことから, 残差スペクトルモデルの導入は音モデル全体の推定精度低下防止に貢献する. ジャンル別では, 3 つの評価結果すべてにおいて Jazz 曲の方が Classic 曲に比べてスペクトル距離は大きい. これは, 今回用いた Jazz 曲全てがドラムパートを含んでいるからであると考えている. ドラム音は一般に非調波成分を多く含み, 非調波成分を適切に分離するのは困難であるため, 全体の性能を下げていると予想される. また, 楽器別の評価結果では, ピアノ音の距離が全ての指標で大きい. ピアノは今回選定した曲の中では全て伴奏を担当しており, 両ジャンルで同じように距離が大きかったため, 伴奏演奏の分離・分析は困難であると考えられる.

4.2 実験 2:伴奏付き演奏音響信号の分析合成

4.2.1 実験条件

市販 CD 収録のプロによる無伴奏単旋律演奏から選出した Violin3 曲 (VN1〜3), Flute3 曲 (FL1〜3), Cello3 曲 (VC1〜3) の計 9 曲に対して演奏分析合成実験を行った. これらはいずれもコンサートホールを思わせる長い残響を含み, また同一楽器内ですべて同じ奏者になら

⋆1 表示されている距離は Jazz No.24 の AS における本手法の距離に基づいて正規化している.

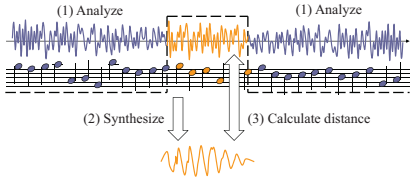

図 4 実験 2 の手順

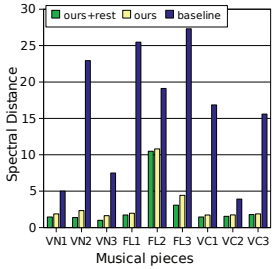

図 5 演奏合成実験結果

ないように選出したため, 奏者や楽器個体も異なる. 実験は, 各曲それぞれに対し, 曲の 4/5 の区間で分析を行い, その結果を用いて残り 1/5 の演奏を合成し, 合成演奏と元演奏との距離を評価する (**図 4**). 本手法 (ours) では合成演奏からは残響がある程度取り除かれているため, 合成演奏にその区間の残差[⋆1]を足し合わせた信号 (ours+rest) についての距離も評価した. 残差スペクトルモデルを用いない方法 (baseline) ではこの必要はない.

4.2.2 実験結果と考察

図 5 に, 実験 2 の結果を示す. どの楽曲に対しても, baseline 手法より本手法の方が飛躍的に距離が小さくなっている. baseline 手法の結果のうち, 距離の大きい楽曲はそもそも合成演奏として成立しえない程の音響的再現性であったが, それらの曲についても, FL2 を除いて小さい値に抑えることができた. FL2 を含むフルート曲の結果が全体的に良くないのは, フルート演奏に多く含まれる息づかいの音が影響していると考えられる. 息は主に非調波成分で構成され, 発音中ほぼ絶えず鳴っていると当時に, 楽器の音色特徴とは独立に変化する. 従って, 分析時の制約が返って悪影響を及ぼしたのではないかと考えている.

5. おわりに

本報告書では，楽器演奏音響信号の分析合成における，入力中の伴奏音や残響成分を抑制した分析手法を報告した．対象演奏パートの楽譜情報に合致しないスペクトル成分を表現する残差スペクトルモデルを定義し, 入力音響信号のパワースペクトルを分析対象演奏の音モデルとこの残差スペクトルモデルに適応的に分配するような定式化により, 伴奏や残響成分を事前情報なしである程度抑制できることを確認した.

本手法は伴奏音に対する事前情報は用いない問題設定であるが, 伴奏の正確な楽譜が場合はそれを援用することで分析精度の更なる向上が予想されるため, その場合との比較も今後行っていきたい. また, 残響については音声を対象とした統計モデルに基づく残響除去手法 [10] との統合によって更なる分析精度向上が期待できる.

演奏合成方法の品質向上のために, (1) 音モデルを HMM 等の確率過程でモデル化し, 音長の違いを吸収する, (2) 楽譜情報からの重回帰によって HMM の各パラメータ平均を操作する, という改良を考えている. これは音声合成の研究における発話スタイル制御 [11] に類似した発想である. これにより, 演奏表情を説明するために楽譜構造上の音長を含む複雑な特徴を用いる事ができると考えている.

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.

⋆1 演奏合成用の区間を別個に分析し, その結果得られた残差モデルを用いる.

参 考 文 献

1) 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).
2) Widmer, G.: Modeling the Rational Basis of Musical Expression, *Computer Music Journal*, Vol.19, No.2, pp.76–96 (1995).
3) 鈴木泰山，徳永健伸，田中穂積：事例に基づく演奏表情の生成，情処論， Vol.41, No.4, pp.1134–1145 (2000).
4) 安良岡直希，安部武宏，糸山克寿，高橋 徹，尾形哲也，奥乃 博：連続発音中の音色変化に着目した未学習譜面情への演奏信号生成，情処第 71 回全国大会, 4R-1 (2009).
5) 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，情処論， Vol.49, No.3, pp.1465–1479 (2008).
6) Grey, J.M.: Multidimensional Perceptual Scaling of Musical Timbres, *J. Acoust. Soc. Am.*, Vol.61, No.5, pp.1270–1277 (1977).
7) Fletcher, H., Blackham, E. and Stratton, R.: Quality of Piano. Tones, *J. Acoust. Soc. Am.*, Vol.34, No.6, pp.749–761 (1962).
8) 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).
9) Goto, M., Hashiguchi, H., Nishimura, T. and Oka, R.: RWC Music Database: Popular, Classical, and Jazz Music Databases, *Proc. ISMIR*, pp.287–288 (2002).
10) Yoshioka, T., Nakatani, T. and Miyoshi, M.: An integrated method for blind separation and dereverberation of convolutive audio mixtures, *Proc. EUSIPCO* (2008).
11) 宮永圭介，益子貴史，小林隆夫：HMM 音声合成における多様なスタイル実現のための制御法，信学技報. SP, 音声， Vol.104, No.30, pp.35–40 (2004).