調波非調波GMMに基づく MIDI演奏音響信号に対する音色・演奏表情操作

安良岡 直希　糸山 克寿　高橋 徹　駒谷 和範　尾形 哲也　奥乃 博

京都大学 大学院情報学研究科 知能情報学専攻

1. はじめに

本稿では，楽器演奏合成において，人間の実演奏に含まれる音響的な演奏表情を既存MIDI演奏に付加する新たな手法について報告する．楽譜に相当するデータから楽器演奏の音響信号を合成する技術において，人間の実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)を自動付加すること(演奏表情付け)は楽曲制作支援などへの応用が期待される大きな課題である．

音色情報に対する演奏表情付け(音響的演奏表情付けと呼ぶ)は困難な課題である．現在報告されている研究[1, 2, 3]のほとんどはMIDI出力機能付き楽器などから得た音量・発音タイミング情報の操作に基づいており，音色情報は対象としていない．一方，倍音相対強度などをとらえた音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を再合成する手法[4]は音色情報操作への拡張が可能である．しかし，従来法はモデルから直接音響信号を再合成するため，演奏信号に伴奏・残響などが含まれる場合はおろか，クリーンな信号を用いてもモデル化誤差や分析性能の影響により音質に限界がある．

我々は，予めMIDI音源で作成した合成演奏の「型」に対し，そのスペクトルを操作するという新たなアプローチを提案する．近年は実楽器音の高品質録音からなるMIDI音源が利用可能となり，これから合成した演奏スペクトルに対し，実演奏から得た音モデルに基づき倍音強度や非調波成分を操作することで，音響的演奏表情付けと音質の維持が同時に実現できる．このようなアプローチの課題は1) 音モデルの設計，2) 楽譜情報から音モデルパラメータの算出，3) MIDI演奏のスペクトル操作，である．本手法ではこれらをそれぞれ1) 調波非調波Gaussian Mixture Model (GMM)[5]の利用，2) 連続2音の楽譜構造類似性に基づくモデルパラメータ算出，3) 調波非調波GMMによる分離スペクトルの重み操作，により解決する．

2. 本手法における音響的演奏表情付け

2.1 問題設定

本稿では，学習用の実演奏音響信号とその楽譜が与えられる下で，未知の演奏の楽譜に対する演奏表情付きの演奏を合成するという問題を扱う．実演奏音響信号の短時間フーリエ変換を$x_{n,l}$とする．ここで，nは時間フレーム，lは周波数ビンである．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仮定し，音響信号と同期がとれているとする．

2.2 調波非調波GMMによる演奏分析

糸山らの音源分離法[5]をベースに調波非調波GMMと呼ぶ音モデルを設計する．これは楽器音のパワースペクトル$\lambda_{n,l}$を，調波構造に対応する分散の小さいGMMと，非調波構造に対応する分散の大きいGMMの線形混合で表す．

$$\lambda_{n,l}=\sum_{k=1}^{K}\left(w_{k,n}^{(H)}\sum_{m=1}^{M}H_{k,m,n,l}+w_{k,n}^{(I)}\sum_{i=1}^{I}I_{k,i,n,l}\right) \quad (1)$$

$$H_{k,m,n,l}=\frac{u_{k,m,n}}{\sqrt{2\pi\sigma^2}}\exp\left[\frac{-(\omega_l-m\mu_{k,n})^2}{2\sigma^2}\right] \quad (2)$$

$$I_{k,i,n,l}=\frac{v_{k,i,n}}{\sqrt{2\pi\gamma^2}}\exp\left[\frac{-(\omega_l-\nu_i)^2}{2\gamma^2}\right] \quad (3)$$

パラメータ$\theta=\{w_{k,n}^{(H)},w_{k,n}^{(I)},u_{k,m,n},v_{k,i,n},\sigma^2,\mu_{k,n}\}_{1\le k\le K,\ 1\le m\le M,\ 1\le i\le I}$はそれぞれ第$k$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_lは周波数ビンからHzへの写像であり，残りのν_i,γ^2は非調波サブバンドの形状を決める定数である．パラメータの推定は，観測パワースペクトル$|x_{n,l}|^2$を2次元確率空間上の頻度分布と見た場合の最尤推定によって実現される．これは通常のGMMの最尤推定と同様にEMアルゴリズムを用いて解くことができる．

2.3 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

新たに合成したい演奏の各単音のモデルパラメータは，分析した実演奏との楽譜構造の類似性に基づいて生成される．本手法では，合成演奏の第j音に対するモデルパラメータ$\tilde{\theta}$を，分析済実演奏中のノートナンバーNと音長Lの類似する単音のパラメータから算出する．

まず，合成する演奏の第j音に対して以下の条件を満たす分析済実演奏中の2音を選出する．

$$q_j^-=\underset{k}{\operatorname{argmin}}\sum_{p=-1,0}\left(|\bar{N}_{j+p}-N_{k+p}|+\alpha|\bar{L}_{j+p}-L_{k+p}|\right) \quad (4)$$

$$q_j^+=\underset{k}{\operatorname{argmin}}\sum_{p=0,1}\left(|\bar{N}_{j+p}-N_{k+p}|+\alpha|\bar{L}_{j+p}-L_{k+p}|\right) \quad (5)$$

ここで，N_k,L_kは実演奏の，$\bar{N}_j,\bar{L}_j$は合成する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混合して，第j音にふさわしい音モデルを算出する．

$$\tilde{\theta}_{j,n}=\begin{cases}\frac{\bar{L}_j^+-n}{\bar{L}_j}\theta_{q_j^-,n},+\frac{n-\bar{L}_j^-}{\bar{L}_j}\theta_{q_j^+,n}, & \bar{L}_j^-\le n\le\bar{L}_j^+\\ 0, & \text{otherwise}\end{cases} \quad (6)$$

ただし，$\tilde{\theta}_{j,n}$は第j音のモデルパラメータ中の時間フレームnに対するものであり，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，$\theta_{q_j^-,n}$, $\theta_{q_j^+,n}$はそれぞれ分析済実演奏の第q_j^-, q_j^+音のモデルパラメータを音高が$\bar{N}_j$，音長が$\bar{L}_j$となるように補間伸縮をしたものである．$\bar{L}_j^-$及び$\bar{L}_j^+$は第j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．

Performance and Timbre Rendering for MIDI-Synthesized Audio Signal by using Harmonic Inharmonic GMM: *Naoki Yasuraoka, Katsutoshi Itoyama, Toru Takahashi, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto Univ.)*

この式は二つの音モデルパラメータの混合比を 1 から 0 へと時間変化させることを意味しており，$q_j^+ = q_{j+1}^-$ であることから，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っている．

2.4 MIDI 演奏スペクトルの分離・重み操作

本手法における音モデル，特に非調波モデルは非常に「粗い」ものである．従ってこのモデルによる楽器音分析は楽器演奏音響信号の大局的傾向は推定するものの，スペクトルの微細構造は表現できないため，モデルが表すスペクトルをそのまま合成しても歪みは大きい．

このモデル化誤差の問題を軽減するため，本手法では，予め MIDI 音源を用いて合成演奏の型を作成し，そのスペクトルをモデルに合わせて変形する方法をとる．MIDI 音源から合成した演奏音響信号を $\bar{x}_{n,l}$ とすると，ここから楽譜情報を元に 2.2 節と全く同じ方法で各単音ごとの調波非調波 GMM モデル $\{\bar{w}_{j,n}^{(H)}, \bar{w}_{j,n}^{(I)}, \bar{H}_{j,m,n,l}, \bar{I}_{j,i,n,l}, \bar{\mu}_{j,n}\}$ 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 j 発音の第 m 調波成分 $\hat{H}_{j,m,n,l}$，及び第 i 非調波サブバンド成分 $\hat{I}_{j,i,n,l}$ に分離することができる．

$$\hat{H}_{j,m,n,l} = \bar{H}_{j,m,n,l} \left|\bar{x}_{n,l}\right|^2 / \bar{\lambda}_{n,l} \tag{7}$$

$$\hat{I}_{j,i,n,l} = \bar{I}_{j,i,n,l} \left|\bar{x}_{n,l}\right|^2 / \bar{\lambda}_{n,l} \tag{8}$$

$$\bar{\lambda}_{n,l} = \sum_{j=1}^{J} \left(\bar{w}_{j,n}^{(H)} \sum_{m=1}^{M} \bar{H}_{j,m,n,l} + \bar{w}_{j,n}^{(I)} \sum_{i=1}^{I} \bar{I}_{j,i,n,l} \right) \tag{9}$$

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信号そのものである．

音響的演奏表情付けは，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 $\tilde{\theta}$ を元に重み付けることで実現される．具体的には，$\tilde{\theta}$ 中の変数 $\tilde{w}_{j,n}^{(H)}$, $\tilde{w}_{j,n}^{(I)}$, $\tilde{u}_{j,m,n}$, $\tilde{v}_{j,i,n}$, $\tilde{\mu}_{j,n}$ 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 $\tilde{Y}_{n,l}$ を得る．

$$\tilde{Y}_{n,l} = \sum_{j=1}^{J} \left(\sum_{m=1}^{M} \frac{\tilde{w}_{j,n}^{(H)} \tilde{u}_{j,m,n}}{\bar{w}_{j,n}^{(H)} \bar{u}_{j,m,n}} \hat{H}_{j,m,n,l}^{+} + \sum_{i=1}^{I} \frac{\tilde{w}_{j,n}^{(I)} \tilde{v}_{j,i,n}}{\bar{w}_{j,n}^{(I)} \bar{v}_{j,i,n}} \hat{I}_{j,i,n,l} \right) \tag{10}$$

ただし，$\hat{H}_{j,m,n,l}^{+}$ は $\hat{H}_{j,m,n,l}$ を周波数方向に $(\tilde{\mu}_{j,n}/\bar{\mu}_{j,n})$ 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．再合成の際には MIDI 演奏音響信号 $\bar{x}_{n,l}$ の位相を付加して合成する．

3. 評価実験

本手法における MIDI 音響信号から音響的に演奏表情付けを行うというアプローチの有効性を示すために行った評価実験について述べる．市販 CD 収録のプロによる無伴奏単旋律演奏:Violin (VN), Flute (FL), Cello (VC) 各 3 曲の計 9 曲に対し，各曲の後ろ 4/5 の演奏音響信号と楽譜を用い演奏を分析し，その結果から前 1/5 の楽譜に対する演奏音響信号を合成する．合成演奏と元の実演奏との間の線形スペクトル距離の小ささを評価する．

本手法の有効性は以下に示す 4 つの手法による演奏合成結果を比較することによって検証する．

1. Ours: 本手法

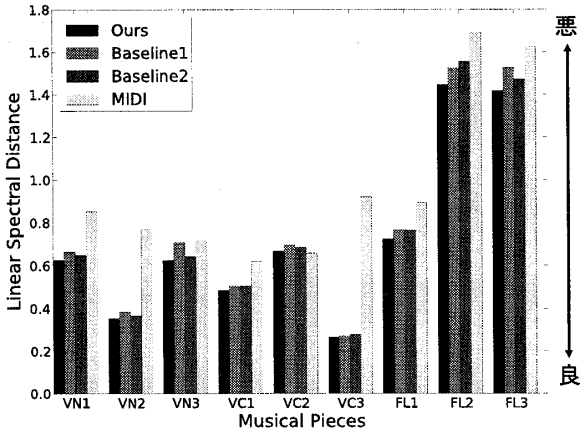

図 1: 曲目ごとの演奏合成実験結果

2. Baseline1: 式 (10) の代わりに音量のみを操作するスペクトル操作式

$$\tilde{Y}_{n,l} = \frac{\sum_{j=1}^{J} (\tilde{w}_{j,n}^{(H)} + \tilde{w}_{j,n}^{(I)})}{\sum_{j=1}^{J} (\bar{w}_{j,n}^{(H)} + \bar{w}_{j,n}^{(I)})} \left|\bar{x}_{n,l}\right|^2 \tag{11}$$

を用いる方法．音量のみの演奏表情付け法に相当*

3. Baseline2: MIDI 音源は用いず，楽譜情報から算出されるパラメータ $\tilde{\theta}$ から直接音響信号を合成する方法
4. MIDI: MIDI 音源による合成音響信号 $\bar{x}_{n,l}$ そのまま

図 1 に曲ごとの演奏合成実験結果を示す．VC2 を除く 8 曲にて本手法が最も実演奏に近いという結果となっている．これは，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によりモデル化誤差の影響を低減することの有効性を示している．

チェロ曲は比較的低音域での演奏からなり，短時間フーリエ変換に基づく分析がうまくいかないことから，MIDI 音源そのままの音の方が実演奏に近くなることがあると考えられる．

4. おわりに

本稿では，MIDI 音源で作成した合成演奏のスペクトルを操作するという新たなアプローチによる音響的演奏表情付け法を報告した．合成演奏の品質は，所望の音に MIDI 音源が近いほど良くなることはもちろんであるが，本手法が持つ倍音相対強度などの音色特徴の補正能力により，ある程度品質の劣る MIDI 音源を用いても十分な合成精度が得られることが予想される．今後は複数の MIDI 音源による合成実験を行いその能力を調査したい．なお，本研究は，科研費，GCOE, CREST-Muse の支援を受けた．

参考文献

[1] 平賀他. 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. Modeling the rational basis of musical expression. *Computer Music Journal*, Vol. 19, No. 2, pp. 76–96, 1995.
[3] 鈴木他. 事例に基づく演奏表情の生成. 情報学論, Vol. 41, No. 4, pp. 1134–1145, 2000.
[4] Yasuraoka et. al. Changing timbre and phrase in existing musical performances as you like: manipulations of single part using harmonic and inharmonic models. In *ACM Multimedia*, pp. 203–212, 2009.
[5] Itoyama et. al. Parameter estimation for harmonic and inharmonic models by using timbre feature distributions. *IPSJ Journal*, Vol. 50, No. 7, 2009.

*演奏表情のうちタイミング情報も従来法において盛んに研究されているが，ここでは考慮しない．