連続発音中の音色変化に着目した未学習譜面上への演奏信号生成

安良岡 直希 † 安部 武宏 ‡ 糸山 克寿 ‡ 高橋 徹 ‡ 尾形 哲也 ‡ 奥乃 博 ‡
† 京都大学 工学部情報学科 ‡ 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, 与えられた楽譜に対する楽器演奏音響信号の自動生成について述べる. 本技術は, 事前に与えられた実演奏音響信号と楽譜のセットを元に, 任意の楽譜に対して実際に演奏されたような音響信号を生成することを目標とし, 音楽製作支援技術としての応用が見込まれる. 従来の楽器演奏生成 (演奏表情付け) [1, 2] では, 演奏情報として入出力されるデータは単音ごとの音量や発音時刻等の数値列であった. これらは音響信号への再合成が不可能な特徴量であり, 元演奏が持っていた音色のような音響的構造は扱えない. 一方, 従来の楽器音分析, 操作及び合成 [3] の研究では, 楽器音の音高・音長操作などを可能にしたが, これらは単発音を対象とし, 表情付けのための操作は考えられてこなかった.

我々はこの 2 分野を統合し, 音響的特徴量の上での演奏表情付けによって音響信号の演奏を出力するシステムを実現する. これを達成するためには, 1. 特徴量表現上での表情の分析, 2. 所与楽譜上での特徴量の再構成, という表情付けの課題を, 音響信号へ再合成可能な特徴量の上で新たに定義・実装する必要がある. 本稿では, 音色知覚に基づき設計した音響特徴量に対して, 1. この特徴量の平均からの変化率に基づく表情分析, 2. 音高遷移パターンと特徴量変化の対応付けに基づく表情構成, というアプローチを取ることでシステムを構築した.

2. 音響特徴量の定義と分析方法

演奏表情付けとしてのパラメータの操作と音響信号再合成の同時実現のために, 演奏知覚差の一要因である音色に着目した特徴量を用いる.

2.1 音色知覚に基づく音響特徴量

音響心理学の知見によると, 音色とは (i) 倍音ピークの有無, (ii) 非調波成分, (iii) 各ピークの時間振幅変動, によって特徴付けられる [4]. これに基づく音響特徴量で単音の分析・操作を行った安部らの研究 [3] を参考に, 本稿では以下の特徴量を設計した.

1) F0 軌跡 $\mu(t)$
2) 各倍音のパワーエンベロープ $M(t,n)$
3) 非調波成分のスペクトログラム $I(t,f)$

ここで, t は時間, n は倍音の次数, f は周波数を表す.

さらに, M を以下のように基準値, 時間方向の相対値, 周波数方向の相対値に展開する.

$$M(t,n) = \hat{m} \times \bar{m}(t) \times m(t,n) \quad (1)$$

$$m(t,n) = M(t,n) / \textstyle\sum_n M(t,n) \quad (2)$$

$$\bar{m}(t) = \textstyle\sum_n M(t,n) / \sum_r (\sum_n M(t,n)) \quad (3)$$

$\hat{m}$ が平均パワー, $\bar{m}$ がパワー振幅, m が倍音間相対パワーに対応し, 演奏表情をそれぞれの値で特徴づける事ができる. 非調波成分 I に対しても同様の展開を行い, $\hat{i}, \bar{i}, i$ とする. 特徴量のイメージを図 1 に示す.

Performance Rendering and Sound Synthesis considering the Timbral Deviation within Note Sequence: *Naoki Yasuraoka, Takehiro Abe, Katsutoshi Itoyama, Toru Takahashi, Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto Univ.)*

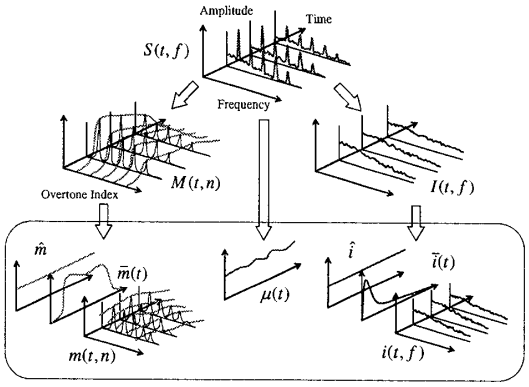

図 1: 音色知覚に基づく音響特徴量

2.2 演奏事例からの特徴量分析

実演奏では発音が連続しているので, 予め同期を取った楽譜を用いて演奏音響信号を単音毎に時間で区切る. 得られた単音のスペクトログラムを $S(t,f)$ とすると, 上記パラメータは調波・非調波統合モデル [5] に基づき以下の式から推定される.

$$S(t,f) = H(t,f) + I(t,f) \quad (4)$$

$$H(t,f) = \sum_n \frac{M(t,n)}{\sqrt{2\pi}\sigma} \exp\left[- \frac{(f - n\mu(t)\sqrt{1+Bn^2})^2}{2\sigma^2} \right] \quad (5)$$

ここで, σ は倍音ピークの周波数方向の分散に対応し, B は調波成分の非調和性を再現する. 以下, 第 k 発音の特徴量を $X_k \equiv \{\mu, m, \bar{m}, \hat{m}, i, \bar{i}, \hat{i}\}_k^{(X)}(t,n,f)$ で表す. また, 第 k 発音のノートナンバーを $p_k^{(X)}$ で, 音長を $T_k^{(X)}$ で表す.

2.3 特徴量からの音響信号再合成

これらの特徴量は正弦波重畳モデルによって音響信号へと再合成される [3]. 残差に対応する非調波成分 I はオーバーラップ加算により時間領域表現に戻し加算する.

3. 表情の分析

演奏の中で生じる特徴量の平均的状態からの揺らぎを本研究における「表情」と定義し, 本章でその分析方法を述べる. これはまず特徴量平均を算出し, 次に各単音に対しそこからの変化率を算出することで実現される.

3.1 平均特徴量の算出

学習データ全体から同じノートナンバーの単音を集め, 特徴量の平均値を算出する. ただし, 特徴量のほとんどは時系列情報であるため, 音長が異なる場合の特徴量平均化を定義する必要がある. 楽器音は一般に音の立ち上がりと立ち下がりの間の定常 (または減衰) 部分の長さのみが異なることが知られており, 従って図 2 のように, 特徴量系列を立ち下がりで分割し, 始点と終点を合わせ共通部分のみを加算平均する. この操作により, ノートナンバー p 毎の平均系列 U_p が得られる.

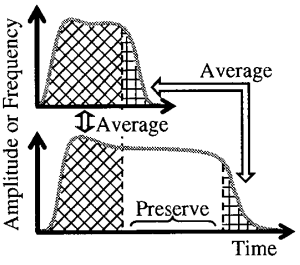

図 2: 特徴量系列の平均化

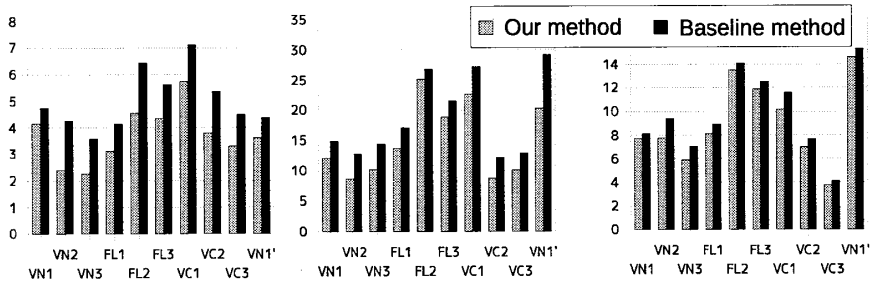

図 3: 評価結果 (左) 調波距離 (中) 非調波距離 (右) スペクトル距離

3.2 変化率の算出

X_k を平均特徴量 $U_{p_k^{(X)}}$ で割ることにより, 第 k 発音に含まれる特徴量変化率すなわち表情 Y_k が得られる.

非調波成分に対応する i については, メル周波数軸のフィルタバンクを用いて周波数方向に次元圧縮した量の間での変化率を用いる. これは, 実演奏から分析された i に含まれる推定誤差が時刻・周波数ごとの変化率計算により強調されることを回避するためである.

4. 所与の楽譜上への演奏の再構築

本手法では, 楽譜上の音高遷移パターンと特徴量とが対応付くとして, 学習した表情を所与楽譜上に再構成する. その理由は, 音高遷移が似ている音符列は一般的に似た方法 (運指, 息使い, アーティキュレーション等) で演奏されやすく, また本手法の時系列特徴量の連続性が単音を跨いでも保存されるからである. この場合, 所与楽譜上の第 j 発音に対する特徴量 $\tilde{X}_j$ は, そのノートナンバーを $p_j^{(\tilde{X})}$, 音長を $T_j^{(\tilde{X})}$ とすると以下の式より算出される.

$$\tilde{X}_j(t) = \left\{\left(1-\frac{t}{T_j^{(\tilde{X})}}\right)\tilde{Y}_{j,1}(t) + \frac{t}{T_j^{(\tilde{X})}}\tilde{Y}_{j,2}(t)\right\} U_{p_j^{(\tilde{X})}}(t) \quad (6)$$

ここで, $\tilde{Y}_{j,1}$ は音高遷移 $(p_{k-1}^{(X)}, p_k^{(X)}) = (p_{j-1}^{(\tilde{X})}, p_j^{(\tilde{X})})$ を満たす Y_k の中で, 音長差 $|T_{k-1}^{(X)} - T_{j-1}^{(\tilde{X})}| + |T_k^{(X)} - T_j^{(\tilde{X})}|$ が最も小さいものであり, $\tilde{Y}_{j,2}$ も $j, j+1$ に関して同様である.

特徴量系列の時間長が足りない場合は延長させる必要がある. μ, m 及び $\bar{m}$ はスプライン補間を用いて伸長させる. i 及び $\bar{i}$ は楽器音のアタック成分などに歪みを生じる伸縮操作は避け, 時間方向の中間で二つの領域に分けて開始位置と終了位置を合わせる. こうして得られた特徴量系列が 2.3 節の方法で演奏音響信号に再合成される.

5. 評価実験

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, 及び熟練者による VN1 の室内録音 1 曲 (VN1') に対し, それぞれ演奏を 5 区間に分け, 4 区間を学習データとして用い残り 1 区間の楽譜に対する音響信号を再現する実験を行い, 生成信号と実演奏の比較により本手法を評価した.

5.1 距離尺度について

本研究の目的に合わせた距離尺度として新たに「調波距離」と「非調波距離」を用いる. これは本手法で分析された特徴量上の距離で定義され, F0 や音量の微小差で値が大きく変動ぜず直感的な演奏の類似性を考察できると考える. 調波距離は以下の D_H を, 非調波距離は以下の D_I の値をすべての単音で平均化したものである.

$$D_H = \frac{1}{T}\sum_t \sqrt{(\bar{m}_r(t)-\bar{m}(t))^2 \sum_n (m_r(t,n)-m(t,n))^2} \quad (7)$$

$$D_I = \frac{1}{T}\sum_t \sqrt{(\bar{i}_r(t)-\bar{i}(t))^2 \sum_f (i_r(t,f)-i(t,f))^2} \quad (8)$$

ここで, 添字 r のある値は実演奏の分析で得た特徴量であり, T は単音の時間長を表す.

また, 合成された音響信号の時間周波数領域での誤差であるスペクトル距離でも評価した. こちらは分析精度の問題等を回避した純粋な信号再現性を考察するのが目的である. どちらも値が小さい方が良い.

5.2 結果と考察

評価実験の結果を図 3 に示す. 比較対象としたベースラインは「表情の分析再構成をせず学習データの最も音高音長の近い単音をそのまま配置する」という方法である§. 本手法の方が, 平均して調波距離が 26.3%, 非調波距離が 21.5%改善された. これは, 音響信号へ再合成可能な特徴量の上での演奏表情付けにより, 楽器演奏の再現能力が向上することを示す.

対してスペクトル距離は 9.2%の改善となり, 変化としては小さい. これは特徴量分析と再合成の精度が完全でないためと考えられる. 原因の一つに, 演奏を時刻で分割して分析している点が挙げられる. たとえ単旋律演奏でも実際には残響などにより, ある瞬間に複数の単音が混在し得るため, これを正しく分離できるよう改良することで分析合成の精度向上が期待される.

6. おわりに

本稿では, 音色知覚に基づく音響特徴量によって実演奏を分析し, その平均からの変化率操作による表情付けで音響信号を再合成する手法を報告した. 今後は, 実演奏の条件により適した楽器音分析や, より高度な表情再構成を検討し, より高品質な演奏生成の実現を目指す.

本研究は, 科研費, GCOE, CREST-Muse の支援を受けた.

参考文献

[1] 鈴木他. 事例に基づく演奏表情の生成. 情処論, Vol. 41, No. 4, pp. 1134–1145, 2000.
[2] 野口他. 演奏情報と楽譜情報の対からの演奏表情規則の獲得とその応用. 情処研報 98-MUS-26, pp. 109–114, 1998.
[3] 安部他. 音高による音色変化を考慮した楽器音の音高・音長操作手法. 情処研報 2008-MUS-76, pp. 155–160, 2008.
[4] J. M. Grey. Multidimensional perceptual scaling of musical timbres. *J. Acoust Soc. Am.*, Vol. 61, No. 5, pp. 1270–1277, 1977.
[5] 糸山他. 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.

§音響信号の出力を条件とする演奏表情付けが報告されてないため, 音合成の知見のみによる演奏生成に相当する処理を設定した.