指揮／演奏追従再生システムのための 音楽音響信号の位相再構成法に基づく実時間テンポ変換*

☆水野優, 橘秀幸, 嵯峨山茂樹 (東大院・情報理工)

1　はじめに

近年, 音楽をより手軽に, 自由に楽しむための技術が盛んに研究されている. なかでも, 自動伴奏システム [1] や指揮システム [2, 3] は, これまで個人で行うには難しかった能動的な音楽体験を可能にする技術であり, 今後の発展が期待されている.

これらのシステムでは入力に同期した音楽の再生を行うために音楽の再生速度や演奏位置を実時間で自由に変化させることが必要となる. 臨場感のある音楽体験を提供するためには音質が重要な要素となるが, 多重音に対する変換を高音質で行うことは必ずしも容易ではない. これらのシステムではこれまで, 加工のしやすさから主に MIDI(Musical Instrument Digital Interface) が用いられきたが, 演奏表情に乏しく音質も十分でないという問題点があった. 一部には CUT & SPLITE 法や Phase Vocoder[4] を用いて実際の演奏を加工している研究もある [5] が, これらの手法も音質が十分とは必ずしも言えない.

本稿ではまず, 指揮／演奏追従再生システムに共通のフレームワークを提示し, そこで必要となる再生速度変換について述べる. 次に, パワースペクトログラムに対する位相再構成に基づく音楽音響信号加工の枠組みを示し, 多重音に対しての再生速度変換手法の音質について実験によって確認する. さらに, 自動伴奏システムや指揮追従演奏再生システムへの応用のため, 処理の実時間性についての検討を行う.

2　指揮／演奏追従再生システムの構成

自動伴奏システムや指揮追従演奏再生システムのような, ユーザーの入力に追従して音楽を再生するシステムは, 以下のような共通の枠組みで構成されていると捉える事が出来る (Fig. 1).

1. ユーザーの入力を解析し, 特徴量を抽出する.
2. 解析結果を用いて, 演奏位置と速度を推定する.
3. 再生する音楽を時間伸縮し, 音響信号を出力する.

Step. 1 は指揮システムで指揮動作から拍点を認識したり, 自動伴奏システムで何の音をいつ演奏しようとしたか推定することに相当する. Step. 2 は事前に用意された正解データを用い, 追従再生する音楽と入力の時間的対応関係を推定する部分で, 例えば [1] の自動伴奏システムや [2] の指揮システムでは隠れマルコフモデル (HMM) を用いている.

指揮／演奏追従再生システムにおいて, Step. 1 と Step. 2 についてはこれまで多くの研究がなされてきたが, その結果に基づいて再生速度変換を施した信号を出力する Step. 3 はこれまであまり注目されてこなかった. 入力に追従して音楽を再生するためには, Step. 3 で扱う再生速度変換はアルゴリズムの観点からも計算時間の観点からも実時間処理が可能であることが必要であり, かつ音質が良いことが望ましい. 我々は以下でこれらの条件を満たす再生速度変換について提案する.

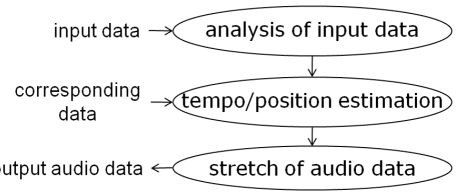

Fig. 1　指揮／演奏追従再生システムの構成.

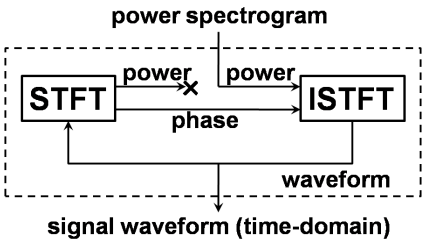

Fig. 2　位相推定アルゴリズム.

3　位相再構成法に基づく再生速度変換

3.1　スペクトログラムからの信号波形生成

我々は音響信号加工において, パワースペクトログラムからの信号生成に着目している. 人間の聴覚系は時間周波数領域におけるエネルギー分布, すなわちパワースペクトログラムに相当する特徴を抽出し, 音響信号を知覚していると考えられるため, パワースペクトログラム領域で音声を設計し, そのスペクトログラムを持つ波形を生成するというアプローチによって, 柔軟な加工が可能であると考えられる.

この考え方に基づき, 音響信号のパワースペクトログラムを時間方向に伸縮させ, それに対応する信号波形を合成することで聴感上自然な再生速度変換を実現する, というのが我々のアプローチである. まずこのアプローチで重要となる, パワースペクトログラムからの波形合成法について簡単にまとめる.

3.2　位相再構成法に基づく波形生成

Griffin らの手法 [6] により適切な位相を反復的に推定することで, 与えられた任意のパワースペクトログラムに最も近いスペクトログラムを持つ信号波形を合成できる (Fig. 2). その手順を以下に示す.

---

*“Real-time Time-scale Modification of a Music Signal Using Phase Reconstruction for Synchronous Playback in Conducting/Accompaniment System” by Mizuno Yuu, Hideyuki Tachibana, Sagayama Shigeki (The University of Tokyo).

Table 1　位相再構成法による提案手法と Phase Vocoder の比較 (再生速度 1.87 倍への変換の場合).

	iteration number						phase vocoder
	2	4	8	16	32	64	
RTF	0.052	0.074	0.118	0.206	0.383	0.735	0.039
SER (dB)	10.88	12.62	14.94	17.06	18.21	18.86	6.90

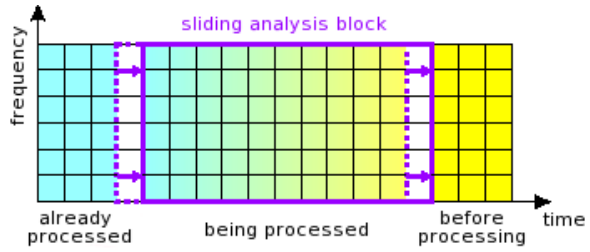

Fig. 3　スライディングブロック処理による効率化.

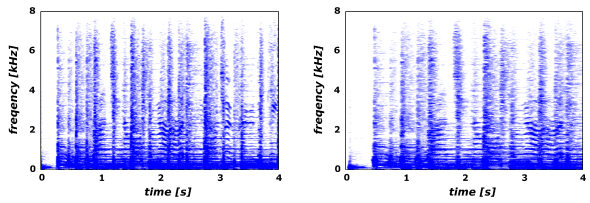

Fig. 4　スペクトログラム伸縮の例. 左: 原信号, 右: 2/3 倍テンポ変換信号

1. 与えられたパワースペクトログラムに対し適当な初期位相を与える.
2. 各フレームを逆フーリエ変換し, 合成窓関数をかけ, 波形信号を合成する (ISTFT).
3. 波形信号から分析窓関数を用いて信号を切り出し, 各フレームをフーリエ変換する (STFT).
4. 与えられたパワースペクトログラムに Step. 3 で得られたスペクトログラムの位相を与える.
5. Step. 2〜Step. 4 を繰り返す.

これは従来収束までに多くの計算量を必要としたが, 近年スライディングブロック処理 (Fig. 3) によって効率化・実時間化され [7], 音響信号加工の実用に耐えうるようになった.

3.3　パワースペクトログラムの伸縮方法

再生速度変換はパワースペクトログラムの時間方向伸縮に相当し, これは短時間フーリエ変換 (STFT) のフレームシフトを変化させることにより実現できる. 例えばフレームシフトを a 倍にすると全体のフレーム数は $1/a$ 倍になり, これは再生速度を a 倍にしたことに相当する (Fig. 4).

フレームシフトを時々刻々と変化させることで再生速度を自由に制御することが可能であり, またこの手法は音楽などの多重音にも適用可能である.

4　再生速度変換の評価実験

再生速度変換の音質と処理時間について, 従来手法である Phase Vocoder との比較評価を行った. 音源分離などと違い, これらの変換には正解となる信号が存在しないので, 音質については生成した波形信号を分析した時に得られるスペクトログラム $X'[mS,k]$ (mS は時間, k は周波数のインデックス) が目的とするスペクトログラム $X[mS,k]$ をどの程度再現出来ているかという観点で, SER(Signal-to-Error Ratio):

$$10\log\frac{\sum_{m=-\infty}^{\infty}\sum_{k=0}^{N-1}\left|X[mS,k]\right|^2}{\sum_{m=-\infty}^{\infty}\sum_{k=0}^{N-1}\left(\left|X[mS,k]\right|-\left|X'[mS,k]\right|\right)^2}$$

を用いて評価し, また処理時間については Real time factor(RTF) を用いて生成信号長との比較を行った. 原信号は RWC 音楽データベース [8] のジャンルの異なる楽曲 10 曲から 10 秒ずつ, 16kHz, モノラルにしたものを用い, STFT のフレーム長は 512, フレームシフトは 128 とした.

3.6GHz の Pentium4 プロセッサを搭載した PC での結果を示した Table. 1 によると, 位相再構成法に基づく我々の手法では, 少ない反復回数においても Phase Vocoder より SER が高く, 反復回数を増やすごとに SER は増加している. たとえば 16 回の反復の場合, 生成信号長の 1/5 程度の時間で 17dB の SER を実現している. この結果から, 計算機の処理能力に応じて反復回数を選択することにより, 我々の手法は指揮／演奏追従再生システムにおいて実時間で従来法より良い音質での再生速度変換を実現することが可能であるといえる.

5　まとめと今後の展望

本稿では, 自動伴奏システムや指揮システムなど, ユーザーの入力に追従して音楽を再生するシステムについて, 共通する枠組みを示し, そこで必要となる再生速度変換についてパワースペクトログラムの伸縮と位相再構成に基づく手法を用いることを提案した. また, その処理時間や音質について検討し, その結果, 実時間で従来手法よりも良い音質の変換が可能であることを確認した.

今後の展望としては, 実際にこの手法を用いた自動伴奏システムや指揮システムの構築が挙げられる.

参考文献

[1] H.Takeda *et al.*,*IPSJ Tech.Report*,66,109–116,2006.
[2] S. Usa *et al.*, *J. ASJ* 19(4), 275–287, 1998.
[3] 馬場他, 情処研報, 2010-MUS-86-26, 1–8, 2010.
[4] J. Laroche *et al.*, *IEEE SAP.*,7(3),323–332,1999.
[5] T. M. Nakra *et al.*, *NIME09*, 250–255, 2009.
[6] D.W.Griffin *et al.*,*IEEE ASSP*,32(2),236–243,1984.
[7] X. Zhu *et al.*, *IEEE ASLP*, 15(5), 1645–1653, 2007.
[8] 後藤他, 情処研報, 2001-MUS-42-6, 35–42, 2001.