パワースペクトログラムの非線形伸縮に基づく音声・音楽信号の再生速度・音高変換*

◎水野優, 小野順貴, 嵯峨山茂樹 (東大院・情報理工)

1　はじめに

音声・音楽信号の再生速度・音高を変換する技術は, 個人の聴力に合わせた聴き取りやすい音声の生成や, 視覚障害者のための速聴システムなどに利用できるほか, 音楽を能動的に楽しむための技術の一つとしても有用である. 我々はパワースペクトログラムの伸縮と位相推定を用い, 多重音に対して高音質な加工が可能な手法を検討してきた [1, 2]. しかし, これまでは再生速度を変換する際に時間方向に一様に伸縮していたため, 人間が発話速度や演奏速度を変化させた際にはあまり長さが変化しない音声の子音部分や打楽器音までもが伸縮されてしまうことによる聴感上の違和感があった. 本稿では, パワースペクトログラムを非線形に伸縮させることで, より高音質な音声・音楽信号加工を可能とする手法を提案する.

2　パワースペクトログラムの伸縮に基づく信号加工

2.1　パワースペクトログラム伸縮に基づく信号加工

人間の聴覚系は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. よって聴覚に自然な音を生成するために, 音響信号のパワースペクトログラムを時間方向, もしくは周波数方向に伸縮させ, 変形させたパワースペクトログラムに対応する信号波形を合成することで再生速度・音高変換を行う, というのが我々のアプローチである [2]. まずこのアプローチで重要となる, 1) パワースペクトログラムの伸縮方法, 2) 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.

2.2　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 STFT(短時間フーリエ変換) 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.

同様に音高変換はパワースペクトログラムの周波数方向伸縮に相当する. これはスペクトルのリサンプリングと等価だが, STFT のフレーム長を変化させ, 高周波成分を間引きないし 0 詰めすることによって, より少ない計算量で実現できる. 更に, LPC(線形予測符号) 分析やラグ窓法と組み合わせ, スペクトルの包絡を保ったままピッチ成分のみを伸縮することで, 音色を保った音高変換を行うこともできる [2].

これらの手法を組み合わせることで再生速度と音高を独立かつ任意に変換し, またそれらを時々刻々と変化させることも可能である.

2.3　パワースペクトログラムからの波形合成

Griffin らの手法 [3] により適切な位相を反復的に推定し, ISTFT(逆短時間フーリエ変換) 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 (図 1). その手順を以下に示す.

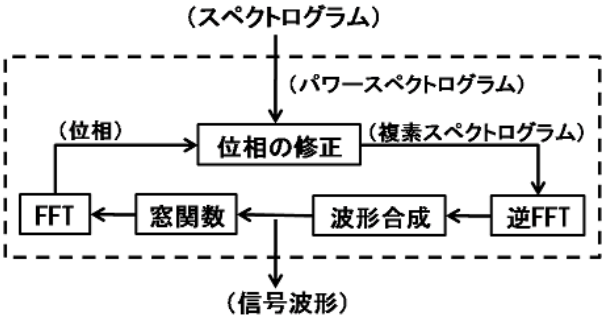

Fig. 1　位相推定アルゴリズム.

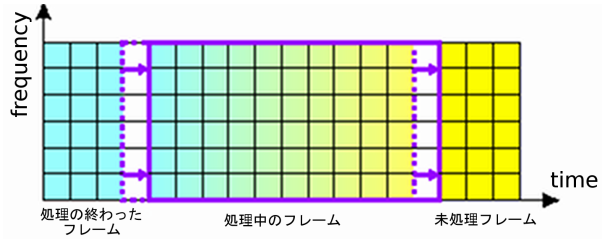

Fig. 2　スライディングブロック処理による効率化

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 (図 2) によって効率化・実時間化され [4], 音響信号加工の実用に耐えうるようになった.

3　スペクトログラムの非線形伸縮

3.1　音色の時間的変化を考慮した非線形伸縮

これまでに述べた我々の手法は多重音に対して実時間で高音質な変換を可能にする利点があるものの, 人間が速度を変えて行った発話や演奏と比べると聴感上の違和感があった. これは, 再生速度変換における時間方向の伸縮が一律であるため, 人間が発話速度・演奏速度を変えた場合にはあまり時間的に伸縮されない, 例えば子音や打楽器音のような音色の時間的変化が激しい部分も伸縮していることが原因と考えられる. そこで, そのような部分ではあまり再生速度を変化させず, 音色変化が小さい部分でより大きく再生速度を変化させ, 入力信号と出力信号の時間的対応関係が非線形になるような伸縮を行うことで聴感上の違和感を軽減し得ると考えられる.

---

* “Time-scale/Pitch Modification of a Speech/Music Signal by Non-linear Stretch of Power Spectrogram” by Mizuno Yuu, Ono Nobutaka, Sagayama Shigeki (The University of Tokyo).

Table 1　線形伸縮との比較. 線形伸縮より音が良いと判断された割合 (%).

速度変化の倍率	0.50	0.75	1.25	1.50	1.75	2.00
音声 (ΔCepstrum)	36.0	54.0	48.7	54.7	44.0	44.7
音声 (HPSS)	50.0	55.3	54.0	50.0	56.7	47.3
音楽 (ΔCepstrum)	52.0	53.3	44.0	48.0	42.7	57.3
音楽 (HPSS)	62.7	54.3	38.7	42.7	56.0	38.7

3.2　音色変化特徴量

音色変化の特徴量として, 本研究では以下の 2 種類を検討した.

$$D_1(m) = \sum_{t=0}^{T}\left(\sum_{i=-I}^{I} i\omega_i c_t(m+i)\right)^2 \quad (1)$$

$$D_2(m) = \sum_{n=0}^{N} P(m,n)^2 \Big/ \sum_{n=0}^{N} H(m,n)^2 \quad (2)$$

ここで, $c_t(m)$ は m 番目の分析フレームにおける t 次の LPC ケプストラム係数, ω_i は i 次の線形回帰係数であり, $D_1(m)$ は Δ ケプストラム [5] の 2 乗和を表す. また, $H(m,n)$, $P(m,n)$ はそれぞれ HPSS(調波打楽器音分離)[6] によって分離された m 番目の分析フレームにおける調波音・打楽器音のスペクトログラム (n は周波数座標) であり, $D_2(m)$ は調波音と打楽器音のパワーの比を表す.

Δ ケプストラムは LPC 法によって求められるケプストラム (対数スペクトルのフーリエ展開係数列) の時間微分ベクトルであり, 特に音声において声道の形状の変化とよく対応することが知られており, その 2 乗和 $D_1(m)$ は音色の変化が大きいほど大きくなる. 人間の発話時, 音色の変化に必要な時間は発話速度に関わらずほとんど変化しないため, $D_1(m)$ の大きいところは再生速度の変化を小さく (等倍速再生に近く) することで聴感上の違和感を軽減し得る.

一方, HPSS は, 音響信号を調波楽器音と打楽器音の混合信号と考え, 調波楽器音は比較的長時間一定の周波数の音が鳴り続けるためスペクトログラム上で時間方向に滑らか. 打楽器音は瞬間的に多くの帯域を占めるため周波数方向に滑らかであるという性質の違いから 2 つの信号を分離する手法である. 打楽器音の持続時間は基本的に演奏速度に依存しないため, 打楽器音が支配的な, $D_2(m)$ が大きい部分で再生速度変化を小さくすることで音質の改善が期待できる.

ケプストラムは打楽器がある部分では大きく変化し, また HPSS において子音は打楽器音成分に分離される傾向があるため, $D_1(m)$,$D_2(m)$ はどちらも音声・音楽の両方に対してある程度効果が期待できる.

3.3　非線形伸縮のアルゴリズム

2.2 節で述べたように, 再生速度は STFT の際のフレームシフトと対応するので前節の音色変化特徴量が大きいところではフレームシフトが等倍速の時のものに近くなり, また全体のフレーム数が目標再生速度に対応した数となるように各フレームシフトを決めればよい. 具体的にはシグモイド関数を用いて, あるフレームにおける音色変化特徴量を p としたとき, 次のフレームまでのフレームシフトは, 等倍速再生の時のフレームシフトを S として

$$aS + (1-a)S\left(\frac{2}{1+e^{-bp}} - 1\right) \quad (3)$$

で決定すればよい (a,b は定数). a は全体のフレーム数が目標値に合うよう決定する.

テンポを一定に保つことがより重要になる音楽信号の場合, 線形伸縮で一律な伸縮を行った場合 (このときテンポは完全に一定) でのフレーム位置から現在のフレーム位置を引いた値を D とし,

$$\begin{cases} S\left(\dfrac{2}{1+e^{-bp}} - 1\right) & (D < 0) \\ D + (S-D)\left(\dfrac{2}{1+e^{-bp}} - 1\right) & (D \geq 0) \end{cases} \quad (4)$$

をフレームシフトとすればよい.

4　評価実験

音声・音楽信号について, 1) 線形伸縮, 2)Δ ケプストラムを用いた非線形伸縮, 3)HPSS を用いた非線形伸縮, の 3 種類の手法で 6 種類の速度に変換したデータを生成し, 15 人によって評価を行った. 音声は ATR 音声データベース B セット [7] より男声・女声を各 5 種, 音楽は RWC ポピュラー音楽データベース [8] より 5 曲を用いた. 評価は, 上記 3 種類の方法で生成した音響信号をランダムに提示し, 音質の自然さの観点から被験者の主観で順位づけさせることにより行った. 線形伸縮よりも非線形伸縮の方が音が良いと判断された割合を表 1 に示す. これによると, 非線形伸縮による提案手法は必ずしも有効であるとは言えないが, 音声に対しては HPSS による手法が全体として線形伸縮に僅かながら勝っている. また, 音楽に対しては低速への変換に関しては非線形伸縮が優位であり, 特に, HPSS による手法は線形伸縮に対して有意水準 0.05 で有意に良い結果となっている一方, 高速への変換では HPSS による手法は線形伸縮に対して劣るという結果となった.

これは, 低速への変換時には打楽器音が持続しなくなったために非線形伸縮による変換がより自然に聞こえる一方, 高速への変換では打楽器音が相対的に強くなりすぎるために結果が悪くなったと考えられる.

5　結論

本稿では, スペクトログラムの非線形伸縮による音声・音楽信号の再生速度変換の高音質化について提案した. この手法は必ずしも全ての変換に対して有効なわけではないが, 音楽信号を低速に変換する場合については効果があることを実験的に確認した. 今後の発展として, 信号の内容に合わせた変換や, 音声のゆらぎ (ビブラート) を軽減することで音質を向上する手法について検討する.

参考文献

[1] L. Roux *et al.*, *Proc. SAPA*, Sep. 2008.
[2] 水野他, 音講論 (春), 843–844, 2009.
[3] Griffin, Lim, *Trans. ASSP*, 32(2), 236–243, 1984.
[4] Zhu *et al.*, *Trans. ASLP*, 15(5), 1645–1653, 2007.
[5] 嵯峨山, 板倉, 音講論 (春), 589–590, 1979.
[6] 宮本他, 音講論 (春), 903–904, 2008.
[7] Kurematsu *et al.*, *Speech Commun.*, 9(4), 357–363, 1990.
[8] 後藤他, 情処研報, 2001-MUS-42-6, 35–42, 2001.