Vol.2012-MUS-96 No.21
2012/8/11

フェーズボコーダと PSOLA による時間伸縮音の比較評価

田坂直季† 小坂直敏†

音響信号に対する時間長の伸縮やピッチ変換は，通信や音楽応用などにおいて重要である．また，これらの処理に優れた方式には，時間領域での TD-PSOLA 方式と，周波数領域でのフェーズボコーダ方式がある．本稿では，われわれが新たに提案したフェーズボコーダ方式と，TD-PSOLA 方式について，時間伸縮，ピッチ変換を適用し，品質を比較して評価する実験を行い，その性能を評価した．その結果，いずれの方式も変換の度合いが大きくなると品質が劣化すること，時間伸縮ではフェーズボコーダの方が音質がいいことが確認できた．

Evaluation of Time-Scale Modified Sound for a Phase-Vocoder and PSOLA

NAOKI TASAKA† NAOTOSHI OSAKA†

For time-scaling and pitch conversion of acoustic signal, TD-PSOLA in the time-domain and Phase-Vocoder in the frequency-domain are well known framework. We apply time-scaling and pitch conversion to newly proposed phase vocoder and TD-PSOLA, and run an evaluation test of sound quality for these two synthesis methods.

1. はじめに

音声や楽音に対する時間長の伸縮，およびピッチ変換は，通信や音楽応用などにおいて重要な音響合成技術である．これらを実現するために，従来よりフェーズボコーダや，PSOLA といった時間伸縮やピッチ変換に優れた分析合成方式が提案されてきた．しかし，このような分析合成方式を用いて合成した時間伸縮音，ピッチ変換音について，それぞれの品質の比較評価はあまりされてきていない．そこで，本研究では，三田らによって提案された分析時に狭帯域 FIR フィルタバンクを，合成時に AM-FM 信号を用いる新しいフェーズボコーダ[1][2]と，PSOLA の一つである TD-PSOLA[3]の二つの分析合成方式を用いて，それぞれの方式で音に時間伸縮とピッチ変換を適用し，品質の比較評価実験を行った結果について報告する．

本稿では，2 節で三田らによって提案された新しいフェーズボコーダの概要，分析合成法，時間伸縮とピッチ制御方法について述べる．3 節では，TD-PSOLA の概要，分析合成法，時間伸縮とピッチ制御法を述べる．4 節では，これらの二つの方式でそれぞれ音に時間伸縮とピッチ変換を適用し，音の品質を比較した実験についての詳細を述べる．

2. フェーズボコーダ

フェーズボコーダは，Flanagan らによってデジタル表現が提案された方式[4]である．さらに，Portnoff が短時間フーリエ変換を用いて定式化を行い[5]，現在では，さまざまな方式の研究が進められている．この方式は，フィルタバンクを用いて入力信号を狭帯域信号に分割した後に，それぞれの帯域内信号を瞬時周波数により表現し，その足し合わせとして出力信号を得る分析合成方式である．再合成を行った場合には入力信号と同等の出力信号が得られるが，時間長/ピッチ制御を行った場合では，“phasiness”という特有の残響感のある異音の混入により，品質劣化が起きるなどの問題点が存在する．この品質劣化を解決するために，三田らによって分析時に狭帯域 FIR フィルタバンクを，合成時に AM-FM 信号を用いる新しいフェーズボコーダ方式[1]が提案された．この項では，その分析合成法の概要，時間伸縮，およびピッチ変換の方式について詳説する．

2.1 狭帯域 AM/FM 信号による分析合成法

2.1.1 分析

FIR フィルタバンクを用いて入力信号を狭帯域信号に分割する．分割帯域幅は 100～400Hz 程度とする．分割された狭帯域信号を AM-FM 信号とみなし，AM 成分である瞬時振幅と FM 成分である瞬時周波数をヒルベルト変換によって求める．ヒルベルト変換は，時点 n の入力信号 $x(n)$ から(1)式に示す解析信号を求める．

$$A(n)e^{j\theta(t)} = x_r(n) + jx_i(n) \tag{1}$$

このとき，振幅エンベロープは

$$A(n) = \sqrt{x_r^2(n) + jx_i^2(n)} \tag{2}$$

であり，位相は

$$\theta(n) = \tan^{-1}\{x_i(n)/x_n(n)\} \tag{3}$$

である．瞬時周波数は $\theta(n)$ を時間微分することにより求

† 東京電機大学
Tokyo Denki University.

められる. この手順を, すべての狭帯域信号について行う.

2.1.2 合成

(4)式を用いて狭帯域信号を合成する.

$$s_k(n) = A_k(n)\cos\{\theta_k(n)\} \tag{4}$$

ここで, $A_k(n)$ は前項の分析によって得たチャネル k の瞬時振幅であり, $\theta_k(n)$ はチャネル k の瞬時周波数を積分して求めた位相である. この手順を全ての狭帯域信号について行い, それらの和を出力信号とする.

2.2 提案方式による時間長/ピッチ制御法

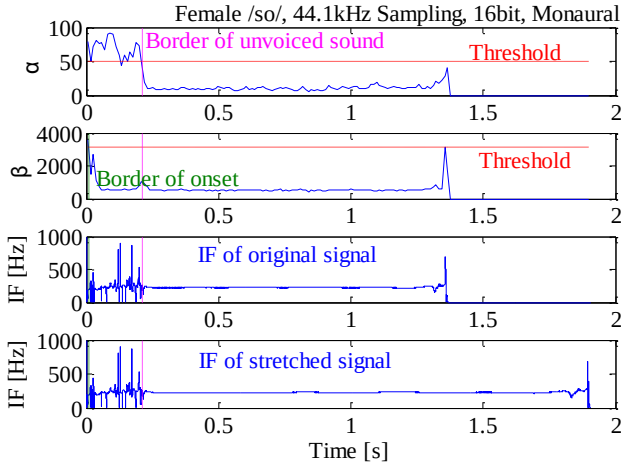
2.2.1 時間長伸縮

時間長伸縮は, リサンプリングと補間により, 瞬時振幅及び瞬時周波数を伸縮することによって行う. なお, 現フレームの初期位相は, 前フレームの位相から求める. また, 自然な時間長伸縮を可能とするため, 子音部および音の立ち上がりは伸縮を行わない. そのため, 子音部の判定に(5)式を, 音の立ち上がりの判定に(6)式を用いた. なお, この操作は単音を対象にしている.

$$\alpha = \sum_{n=0}^{N-1} X(n) \Big/ X_{MAX} \tag{5}$$

$$\beta = \sum_{k=0}^{K-1} \sum_{i=0}^{I-2} \left| p_k(i+1) - p_k(i) \right| \tag{6}$$

ここで, $X(n)$ は入力信号 $x(n)$ のパワースペクトル, X_{MAX} はその最大値, N は FFT ビンの数である. また, K はチャネル数, $p(i)$ は瞬時周波数(IF)のピークあるいは谷であり,



IF: instantaneous frequency

図 1　女声の α 値, β 値, 瞬時周波数とこれを 1.47 倍に伸張した瞬時周波数

Figure 1 An example of temporal stretching

I はその個数である. それぞれ, 閾値よりも大きい場合に, その信号が子音部あるいは音の立ち上がりであるとする. 図 1 に実際の女声に適用した例を示す. 雑音部は周波数上の広域でエネルギーがあるため, α 値は大きくなる. また, 有ピッチ区間では, IF は安定した値で定常的であり, 立ち上がり, 立下りなどの過渡部では振動形をとるため, β の値が大きくなる.

2.2.2 ピッチ変換

ピッチ変換は瞬時周波数を定数倍することによって行う. このとき, スペクトル包絡を保持するため, 各狭帯域信号の二乗平均値を求め, その比によって, (7)式のように狭帯域信号の振幅を増幅させる.

$$\tilde{s}_k(t) = \begin{cases} s_k(t)\dfrac{R'}{R_k} & : \dfrac{R'}{R_k} < 1 \\ s_k(t)\log_{10}\left(\dfrac{R'}{R_k}\right) & : \dfrac{R'}{R_k} \geq 1 \end{cases} \tag{7}$$

ここで, $s_k(t)$ はチャネル k の狭帯域信号, R_k はその二乗平均値, R' はその狭帯域信号がピッチ変換によって移行するチャネルの二乗平均値, $\tilde{s}_k(t)$ は振幅を増減した後の狭帯域信号であり, この和によって出力信号を合成する.

3. PSOLA（ピッチ同期オーバーラップ加算）

PSOLA とはピッチ同期オーバーラップ加算の略であり, Moulines らによって提案された[3]. PSOLA には時間領域 PSOLA（TD-PSOLA）, 周波数領域 PSOLA（FD-PSOLA）などが存在するが, 今回は TD-PSOLA を用いることとした. TD-PSOLA の分析合成は以下の 3 つの手順で行われる.

- ローカルピッチ周期と同期して信号を短時間信号の系列に分割
- この中間表現をもたらす変換
- 変換された中間表現から変換された合成信号を出力

これらの 3 つの手順を以下に詳説し, その後 TD-PSOLA を用いたタイムスケーリング, ピッチスケーリングについて述べる.

また, TD-PSOLA で分析合成を行うために, 短時間波形のピッチ抽出を行う. その際にケプストラムによるピッチ抽出を用いたが, 手動のピッチ修正は行っていない.

3.1 TD-PSOLA の分析合成法

3.1.1 ピッチ同期分析

デジタル音声波形の表現 $x(n)$ は, (8)式のように, ピッチ同期の分析窓を信号に掛け合わせることによって得られる短時間信号の系列 $x_m(n)$ から構成される.

$$x_m(n) = h_m(t_m - n)x(n) \quad (8)$$

窓 $h_m(n)$ の長さは信号の有声部分ではピッチ同期に比例して，無声部分では一定のレートで設定される．また，この窓はピッチマークと呼ばれる分析瞬間 t_m を中心として信号 $x(n)$ に窓かけを行う．窓 $h_m(n)$ はハニング形で，隣接する短時間信号が，常にオーバーラップするように設定される．この際，窓 $h_m(n)$ の長さは比例定数 μ でローカルピッチ周期の何倍長とするかを設定する． $\mu = 2$ の場合，現在の窓 $h_m(n)$ の長さの右半分だけ，次の窓 $h_{m+1}(n)$ がオーバーラップを伴うように設定され， $\mu = 4$ の場合，現在の窓 $h_m(n)$ の長さの右 3/4 長分だけ，次の窓 $h_{m+1}(n)$ がオーバーラップを伴うように設定される．以下の(9)式は，窓の長さに関する比例規則である．

$$h_m(n) = h(n/\mu P) \quad (9)$$

$h(n)$ は長さが 1 で正規化されている窓で， P はローカルのピッチ周期のサンプル数である．

3.1.2 ピッチ同期を変更する方法

分析時の短時間信号 $x_m(n)$ を用いて時間伸縮およびピッチ変換を行うために, 以下の 3 つの基本的操作が行われる．

- 短時間信号の数の変更
- 短時間信号の遅延の変更
- 個々の短時間信号の波形の変更

まず，分析時に μ =2 とおく．すなわち，ピッチ周期波形の２波分長さの窓で切り出す．ピッチ波形は，最大振幅を生じさせる立ち上がり部から始まるようにする．その結果，窓の中心にピッチ周期波形の冒頭部が来ることになり，最大振幅部がほぼ中心位置となる．

μ =2 の条件は，分析合成時には，窓長の 1/2 でオーバーラップ加算をするため, 振幅にＡＭ（振幅変調）は起きず，原音がそのまま再合成される．

3.2 TD-PSOLA による時間伸縮/ピッチ変換

3.2.1 時間伸縮

時間伸縮を行う具体的な方法は以下のようにする. まず，分析時に切り出された $x_m(n)$ はそのまま用いる．これを合成時に再配置する問題となる．これは図 2 に示すように，t_m を分析の，また $\tilde{t}_q$ を合成時のピッチマーク(ピッチ波形の開始点，窓の開始点)としたとき，$\tilde{t}_q \to t_m$ の対応関係は伸縮率を γ としたとき， γt_m に最も近い整数を $\tilde{t}_q$ とする．伸縮を行うためには，この $\tilde{t}_q$ に沿って $x_m(n)$ を配置していく．その間は $x_m(n)$ を用いて等間隔でオーバーラップ加算をするため，TD-PSOLA ではＡＭは起こらない．

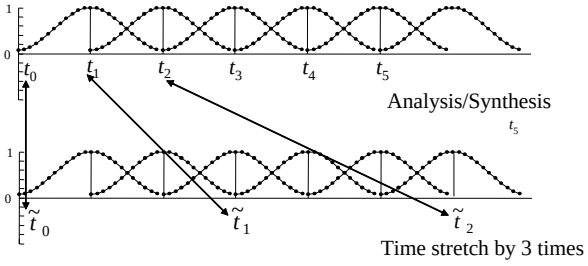


図 2 時間伸縮を行うときの窓の配置
Fig. 2 Window displacement at Time stretch

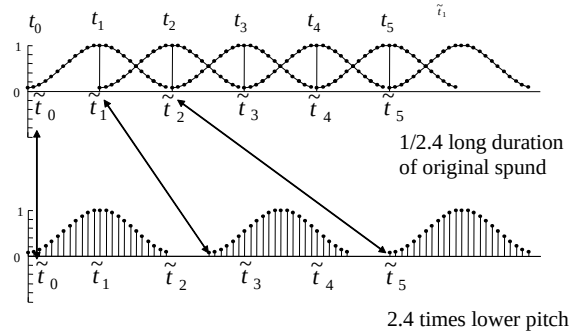


図 3 ピッチ変換時の窓の配置(2.4 倍低い場合)
Fig. 3 Window displacement at pitch conversion

3.2.2 ピッチ変換

ピッチ変換は時間伸縮より複雑である．ピッチを変換することにより，時間の長さも変わってしまう．そこで，これを補正するよう, 逆算して, 先取りして時間伸縮を行う．例えばピッチを２倍に変換するとき, ピッチ周期波形は 1/2 になるため，何も時間伸縮を行わないと，合成音長は原音の 1/2 になってしまう．そこで，前節で述べた方法で，先取りして 2 倍に伸張する．次に，オーバーラップ加算は小ピッチ変換に応じた比率で行う．

図 3 はピッチを 2.4 倍低くする場合の例で，一旦時間伸張した後の窓の配置との関係を示している．この例は極端ではあるが，ＡＭが発生することがわかる．

4. 評価実験

4.1 実験の条件

本実験では，原音，および 2 節で紹介したフェーズボコーダ，および 3 節で紹介した PSOLA を用いて時間伸縮，ピッチ変換を行った合成音の品質を評価し，比較する．表 1 に，フェーズボコーダの分析条件を示す．また，表 2 に使用する実験音と時間伸縮，ピッチ変換の係数の一覧を示す. これらの 39 条件をそれぞれ 3 回ずつランダムに再生し，

音質を以下の 5 段階で評価する.

- 5. 良い
- 4. やや良い
- 3. どちらでもない
- 2. やや悪い
- 1. 悪い

被験者は成人男性 8 名，成人女性 2 名の計 10 名で行った. 評価は試験音の範囲で相対評価となるよう, 評価前に, 実験の中で代表的な試験音を聴取させた.

表 1　フェーズボコーダの分析条件

Table 1　Condition of Phase-Vocoder.

フレーム長	23 [ms]
フレームシフト幅	12 [ms]
フィルタのタップ数	1023
サンプリング周波数	44.1 [kHz]
チャネル数	150

表 2　実験条件

Table 2　List of sound of experiment.

要因	楽音／音声試料	適用方式	伸縮倍率／ピッチ高低
時間伸縮	クラシックギター	フェーズボコーダ，PSOLA	原音，0.625 倍，1.5 倍，2.0 倍
	女声短文章		
	女声歌唱「あ」		
ピッチ変換	クラシックギター	フェーズボコーダ，PSOLA	原音，1/3oct 低，1/3oct 高，2/3oct 高
	女声短文章		
	女声歌唱「あ」		

4.2 実験結果

図 4 に時間伸縮実験の評価値の平均値(MOS)と 95%信頼区間を，図 5 にピッチ変換実験の MOS と 95%信頼区間を示す．図 4，5 では，上部は 2 節で述べたフェーズボコーダを，下部は 3 節で紹介した TD-PSOLA を用いて変換した音に対する評価結果である．サンプル番号 1～4 は A4 のクラシックギター音，5～8 は短文章女声，9～12 はピッチが一定の女声歌唱「あ」に適用した結果である．時間伸縮，およびピッチ変化の 2 要因とも 4 水準あり，図 4 では，4 サンプル単位で，左から原音，0.625 倍長，1.5 倍長，2 倍長の時間伸縮音，また，図 5 も同様に，4 サンプル単位で左から原音，1/3oc 低，1/3oct 高，2/3oct 高，のピッチ変換音の順にまとめて示した.

4.3 考察

図 4 の時間伸縮音に関する評価では, フェーズボコーダ, PSOLA の両方とも伸縮度が大きいほど評価が下がる傾向にあることが示された．また，上部のフェーズボコーダを見ると全ての実験データが標準以上の品質を保つという結果が示されていることが分かる．一方，下部の TD-PSOLA では，ギターや女声の歌唱に関してはフェーズボコーダよりも高い評価であったが，短文章女声に関しては評価が著しく悪い.

図 5 のピッチ変換音に関する評価では，ピッチ変換のオクターブが上がるほど評価が下がる傾向にある．上部のフェーズボコーダを見ると，全体として評価がやや悪いことが分かる．また，下部の PSOLA では，ギターや女性歌唱に関してはピッチ変換音の評価が標準以上であるが，短文章女声に関しては評価が著しく悪い．これは，連続音声についてまだ音質向上の余地があることを意味している.

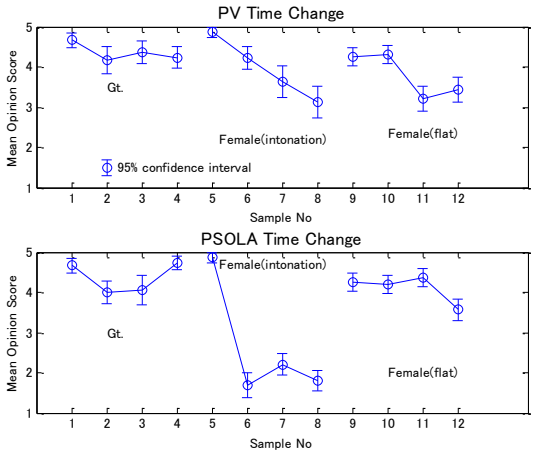


図 4　時間伸縮の評価

Fig. 4　Evaluation of Time-scale modification .

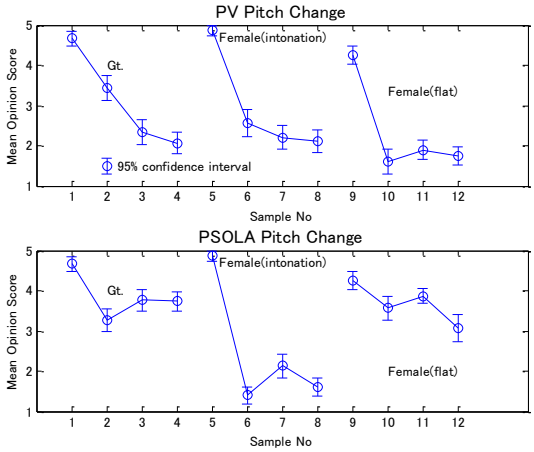


図 5　ピッチ変換の評価

Fig. 5　Evaluation of pitch modification

5. おわりに

本論文では，三田らが提案した狭帯域 AM/FM 信号による新しいフェーズボコーダの分析合成方式，時間伸縮，ピッチ変換方式と, Moulines らが提案した TD-PSOLA 法の分析合成方式，時間伸縮，ピッチ変換の概要について述べ,

それぞれの方式を用いて合成した時間伸縮音，ピッチ変換音を比較し，音質の評価実験を行った．

この評価実験から，フェーズボコーダに関しては，時間伸縮の際には伸縮が大きくない場合は高品質を保つことが分かった．また，ピッチ変換については少しのピッチ増減でも評価が悪くなることが分かった．また，PSOLA に関しては，ピッチの変化しないギターや女声に関して時間伸縮やピッチ変換を行なっても，高品質を保つが，短文章女声に関しては，時間伸縮，ピッチ変換ともに著しく悪くなることが分かった．

今後の課題としては，評価が低いピッチ変換に関しての原因や，どのような種類の音が品質を保ちやすいかについて，比較する音を増やし，品質の評価実験を行うことにより,それぞれの性能を明らかにすることなどが考えられる．

謝辞 本研究を進めるにあたり，有用な議論をして頂いた三田篤志氏と貴重な御助言，御助力を頂きました音メディア表現研究室の諸氏に感謝致します．

参考文献

[1] 三田篤志，小坂直敏，“狭帯域 AM-FM 信号による時間長/ピッチ制御方式，”*日本音響学会，2010 年春季研究発表会*, 2010.
[2] N. Osaka, and A. Mita, “Time/pitch modification using narrowband AM-FM signals,” *The proc. of ICMC 2012*,Ljubljana, Slovenia, Sept. 2012.
[3] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech communication*, vol. 9, pp. 453-467, 1990.
[4] J. Flanagan and R. Golden, "Phase vocoder," *Bell System Technical Journal*, vol. 45, pp. 1493-1509, 1966.
[5] M. Portnoff, "Implementation of the digital phase vocoder using the fast Fourier transform," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 24, pp. 243-248, 1976.