サンプル値制御理論にもとづく信号長不変なピッチシフト処理

○埴淵千誉　永原正章　山本裕 (京都大学)

Duration-Invariant Pitch-Shifting Method Based on Sampled-Data Control Theory

∗ K. Hanibuchi, M. Nagahara and Y. Yamamoto (Kyoto University)

Abstract− In this article, we propose a new pitch-shifting method for stringed instruments by using sampled-data H^∞ optimal fractional delay filters. Since the duration of a pitch-shifted signal is in general different from that of the original signal, we also propose a new method for controlling the duration by dividing the signal into frames on pre-specified intervals and pitch-shifting the signal on each frame.

Key Words: Digital signal processing, sampled-data control, sampling rate conversion, pitch shifting.

1　はじめに

音楽および音楽コンテンツの制作においては，そのすべてをコンピュータ上で行う DTM (Desktop Music) が現在主流になりつつある [6]．DTM はコンピュータ上にソフトウェアとして実装され，楽譜の入力やそれにもとづく楽器の自動演奏，音の加工やミキシングなどの機能がコンピュータ 1 台で可能となる．本稿では，そのなかでも重要な処理である楽器音合成について考察する．

楽器音合成とは，コンピュータ上でアルゴリズムとして楽器音を生成する手法である．楽器音合成には，加算合成や波形テーブル合成，変調合成等があるが，近年では，楽器音をディジタル録音し，それを加工して再生する PCM (Pulse Code Modulation) 方式が主流である．この方式により，元の楽器音をある程度忠実に再現することが可能となるが，一方サンプリングされたディジタル信号系列をすべてを記録するためデータベースの容量が膨大になる．例えば CD の音質である 16bit/44.1kHz の精度でサンプリングされた PCM 音源の 1 秒間の波形を記録するのにおよそ 86 キロバイトの容量を必要とする．この波形は 1 つの楽器，例えばピアノの 1 つの音階のデータであり，88 鍵のピアノを再現する場合は，約 7.6 メガバイトの容量が必要となる．

この問題に対処するために，ある一つの音源に対して，その音高を再生時にシフトすることにより，すべての音階を記憶せずに基準音のみを記憶する方法がある [6]．この方法によって音源としての記憶容量を大幅に削減できることが期待できる．音高をシフトするには時間領域で時間軸の伸長を行う必要があり，このためにディジタル信号のサンプル点間を補間する処理が不可欠となる．本稿では，楽器音のモデルを有限次元の線形システムと仮定し，その仮定のもとでアナログの誤差系の H^∞ ノルムを最小化する補間方式を採用する．具体的にはサンプル値制御理論にもとづいて設計された非整数遅延フィルタ [3, 4] を用いることによってサンプル点間補間を行う．これにより，音源のアナログ特性を反映した時間軸の伸長が可能になる．

しかし時間軸を単に伸長する方法では，例えば音階を高くしたとき音の長さも短くなるという問題がある．本稿では対象を弦楽器の音信号に絞り，その音高のシフトに対して，その信号長の調整を考慮に入れた処理を提案する．具体的には楽器音のディジタル信号を立上がり区間と減衰区間に分割し，さらに減衰区間の信号を小区間のフレームに分割する．そのフレームを適切に反復することにより信号長を保存する音高シフトを実現する．

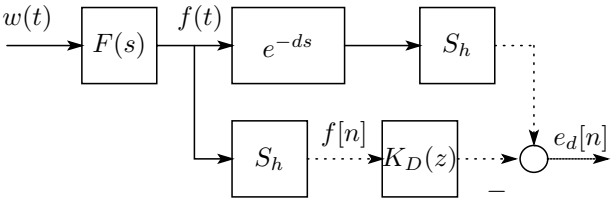


Fig. 1: Error system for fractional delay filter design

2　信号長を保存するピッチシフト処理

2.1　サンプル点間の補間

音はそのアナログ信号を標本化し，ディジタル化することで計算機上の処理が可能となる．基本的にこのようにして得られたデータ系列を用いて信号処理を行う．その際，データ系列のサンプル点間を考慮する必要が生じる場合がある．標本化前の信号のアナログ特性を反映させるために，その特性を設計問題に取り込んだ非整数遅延フィルタによるサンプル点間補間を考える．

信号 f を一定の標本化間隔 h で標本化した系列 $\{f[n]\}_{n=0}^{\infty}$ を $f[n] := f(nh)$ で定義する．以下では標本化間隔について特に指定のない場合は一定の値 h と定める．Fig. 1 に非整数遅延フィルタの設計のための誤差系を示す．ここで $F(s)$ は元の連続時間信号の特性を反映したフィルタである．連続時間信号を D だけ時間軸方向に遅延させてから標本化した信号系列に対して，元の連続時間信号を標本化してからフィルタを適用することで得られる信号系列との偏差 e_d の w からの拡大率を表す L^2 誘導ノルムを最小化することを目標として設計されるのが非整数遅延フィルタである．遅延を $0 < d \leq h$ とし，伝達関数を $F(s) = \omega_c/(s+\omega_c)$

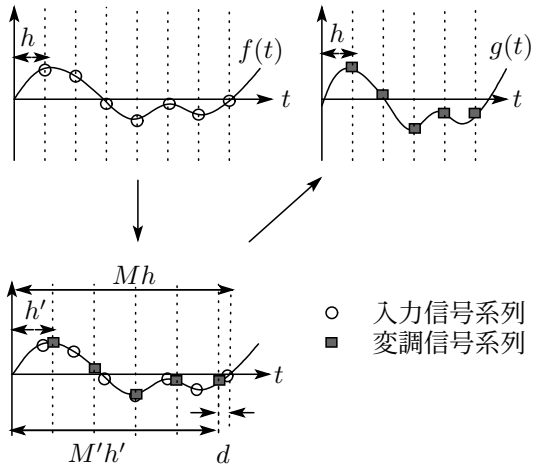


Fig. 2: Original signal f and modulated signal g

と仮定すると，最適フィルタは次式で与えられる [4]．

$$K(z) = a_0(d) + a_1(d)z^{-1},$$

$$a_0(d) = \frac{\sinh \omega_c(h-d)}{\sqrt{\omega_c}\sinh \omega_c h}, \quad a_1(d) = e^{-\omega_c h}(e^{\omega_c d} - a_0).$$

これより，サンプル点間の点 $f(nh-d)$ $(0 < d < h)$ はその前後のサンプル値 $f[n-1]$ と $f[n]$ とを用いて

$$f(nh-d) = a_0(d)f[n] + a_1(d)f[n-1], \quad n = 1, 2, \ldots \tag{1}$$

と表される．すなわちサンプル点間の値 $f(nh-d)$ は $f[n]$ と $f[n-1]$ の線形結合により推定することができ，サンプル点間上の任意の負でない時刻の点の値を補間できる．

2.2 サンプル点間補間によるピッチシフト

基音の周波数が ω のディジタル信号を周波数 ω' のディジタル信号に変調する，すなわち音高を $R = \omega'/\omega$ の倍率でシフトすることを考える．原信号を f，変調信号を g とする．変調信号は原信号を R^{-1} の倍率で時間軸方向に伸長することで得られる．対象はディジタル信号であるから問題は原信号 f を標本化した系列をもとに，変調信号 g の系列を出力する問題に帰着される．変調信号は原信号を伸長したものであることに注意すると，変調信号系列は原信号 f について，サンプリング周期を $h' := Rh$ として標本化して出力した系列に相当する (Fig. 2)．これらのデータ系列をサンプリング周期 h に戻して再び系列化することで g を構成できる．原信号，変調信号のデータ系列の総数をそれぞれ $M+1$, $M'+1$ とおくと，変調信号系列 $\{g[i]\}_{i=0}^{M'}$ は

$$g[i] = f(ih') = f(Mh-(M'-i)h'-d),\ i = 0, 1, \ldots, M'$$

で与えられる．ただし $d = Mh - M'h'$ である．上式の右辺の値は，(1) により得ることができる．すなわち，$f(Mh - mh' - d)$, $m = M', M'-1, \ldots, 0$ の値は，l_m を $(mh'+d)/h$ を超えない最大の整数とし，$\delta_m := mh' + d - l_m h$ とすると，(1) より

$$f(Mh - mh' - d) = a_0(\delta_m)f[M - l_m] + a_1(\delta_m)f[M - l_m - 1]$$

で求められ，これらから目的の変調信号 y のデータ系列が得られる．

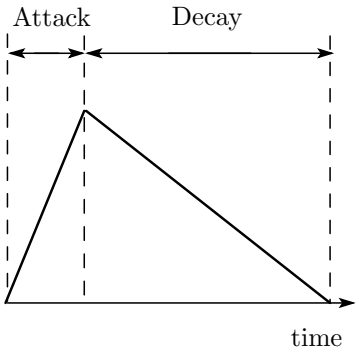


Fig. 3: Envelope of stringed-instruments signal

2.3 連続時間信号における音高シフトと信号長調整

楽音の音高シフトは，信号の局所的な周期特性に着目すれば，時間軸方向に伸長を行うことで実現できる．しかしその場合，信号の長さが同時に変化してしまう．以下では，連続時間信号について信号長を維持したまま音高を変える手法を時間領域において検討する．

Fig. 3 は弦楽器信号のエンベロープを模式化したものであり，信号の立ち上がりの区間とその後減衰していく区間で構成される．一般に減衰区間では波形の局所的な周期特性が見られるが，立ち上がりの区間では減衰区間ほどの周期特性は見られず，比較的複雑な構造をもつ．この現象を踏まえて，複雑な構造を持つ立ち上がりの区間の特性を残し，減衰区間の周期特性を活かした信号長の調整をすることで，信号長を維持した音高変換を実現する．

2.3.1 立ち上がりの区間の処理

複雑な構造を持つこの立ち上がりの区間は楽器の特徴を決定づける重要な特性を持つ．したがってこの区間の信号はそのまま音高を変換するために信号を時間軸方向に伸長させる．信号処理前後でこの区間の信号長は変化する．信号全体としてその長さを維持するための処理は減衰区間での処理で行う．

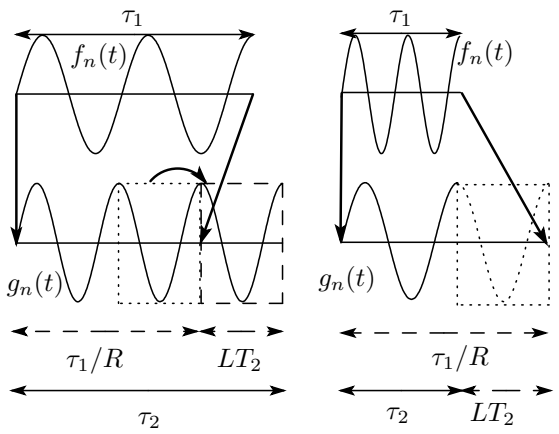
2.3.2 減衰区間における信号長調整

減衰区間ではその周期的構造に着目した信号長調整を踏まえた音高シフト処理を行う．この区間の信号を分割しその各々について独立に処理する．原信号の分割区間の信号長を τ_1 として，これを信号処理によって音高を R 倍し，減衰区間でのその信号長を $\tau_2 = a\tau_1$ に調整する．

音高を上げる場合は，原信号を伸長することで音高を R 倍にする．その際の信号の信号長は τ_1/R となる．次に区間 $[\tau_1/R - T_2, \tau_1/R]$ から基本波形 (周期 T_2) を参照する．変調信号の区間 $[\tau_1/R, \tau_2]$ は $t = \tau_1/R$ と $t = \tau_2$ の位相が等しくなるようにし，Fig. 4(a) のように，参照した基本周期の整数倍でこの区間を補う．L 個の基本周期で補う場合，$a > 1$ とすると，原信号の分割区間の信号長は

$$\tau_1 \quad = \quad \frac{RLT_2}{aR-1} \tag{2}$$

と決定できる．また基本信号を参照するための条件 $\tau_1/R \geq T_2$ に注意すれば，(2) より，L の条件 $L \geq aR - 1$ が得られる．L はこれを満たす最小の整数と定める．



(a) 音高を上げる場合　　(b) 音高を下げる場合

Fig. 4: Processing on each frame

音高を下げる場合は，上げる場合と同様，原信号を伸長させることで音高を R 倍にする．Fig. 4(b) のように区間 $[t_2, t_1/R]$ が基本周期の整数倍となるように調整を行い．最終的にこの区間は削除する．ここで $a < 1$ とすると原信号の分割区間の信号長は

$$\tau_1 \quad = \quad \frac{RLT_2}{1-aR} \tag{3}$$

と決定される．制約条件を $\tau_2 \geq T_2$ とすると，(3) より，L の条件 $L \geq (aR)^{-1} - 1$ が得られ，L はこれを満たす最小の整数と定める．減衰区間の分割時間は音高を上げる場合も，下げる場合もすべて τ_1 と定める．

2.4　非整数遅延フィルタによる出力信号系列

2.3 節の手法はアナログ信号モデルをベースとした手法である．一方，実際の原信号，出力信号ともに，ある一定の標本化時間をもつディジタル信号であるが，サンプル点間のアナログ特性をもつ非整数遅延フィルタを応用することでアナログ信号モデルにおける調整手法を実現する．(1) で示す通り，サンプル点間の信号値はその前後のサンプル値の線形結合で表すことができる．$f(t)$，$g(t)$ をそれぞれ原信号，出力信号とする．分割区間の総数を N とし，各分割区間の原信号，出力信号をそれぞれ $f_n(t)$，$g_n(t)$ とする $(n = 0, 1, 2, \cdots, N-1)$．ただし $f_0(t)$，$g_0(t)$ はそれぞれ原信号，出力信号の立ち上がり分割区間を表す．f_n, g_n はそれぞれ

$$\begin{aligned} f_n(t) &= f(t + t_{1,n}),\ 0 \leq t \leq t_{1,n+1} - t_{1,n}, \\ g_n(t) &= g(t + t_{2,n}),\ 0 \leq t \leq t_{2,n+1} - t_{2,n}, \end{aligned}$$

と定義する．ここで $t_{1,n}$ と $t_{2,n}$ は，t_a を立ち上がり区間の終端時刻として，$t_{1,0} = t_{2,0} = 0$, $t_{1,1} = t_a$, $t_{2,1} = t_a/R$, $t_{1,n+1} = t_{1,n} + \tau_1$, $t_{2,n+1} = t_{2,n} + \tau_2$ で定義する $(n = 1, 2, \ldots, N-1)$．

出力信号系列 $g[k]\,(k = 0, 1, \ldots, K)$ について，$g[k] = g_n(ph + \theta)$ $(0 < \theta < h)$ と表せる場合，M_n を g_n 上のサンプル点数として，$p = 0, 1, \ldots M_n - 2$ のとき，$g_n[p]$ と h だけ時刻が進んだ $g_n[p+1]$ の 2 点の情報から $g_n(ph+\theta)$ の値を得ることができる．具体的には (1) から，$p = 0, 1, \ldots, M_n - 2$ のときは

$$g_n(ph + \theta) = a_0(h - \theta)g_n[p+1] + a_1(h - \theta)g_n[p],$$

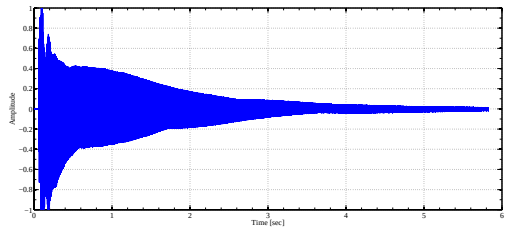


Fig. 5: Original signal

で，$p = M_n - 1$ のときは

$$g_n(ph+\theta) = a_0(h-\theta)g_{n+1}(M_n h - \tau_2) + a_1(h-\theta)g_n[p],$$

で表される．不等式 $0 \leq M_n h - \tau_2 < h$ に注意すると $g_{n+1}(M_n h - \tau_2)$ の値は $g_{n+1}[0]$ と $g_{n+1}[1]$ の 2 点の情報から得ることができる．音高をあげる場合は $h' = Rh$ として，$g_0[p] = f_0(ph')$ とし，また $0 \leq p < \tau_1/h'$ のときは

$$g_n[p] = f_n(ph'),\ n = 1, \ldots, N-1$$

で，$\tau_1/h' + lT_2/h \leq p < \tau_1/(Rh) + (l+1)T_2/h$, $l = 0, 1, \ldots, L-1$ のときは

$$g_n[p] = g_n(ph - (l+1)T_2),\ n = 1, \ldots, N-1$$

で表される．音高を下げる場合は $g_0[p] = f_0(ph')$, $g_n[p] = f_n(ph')$ となり，いずれの場合も $f(t)$ の上の点を参照することで $g[k]$ が得られることがわかる．

3　数値例

ギターの音を原信号として録音した楽音データをもとにシミュレーションを行った結果を以下に記す．原信号の波形を Fig. 5 に示す．基音の周波数は 110.4032Hz であった．

3.1　弦楽器音のピッチシフトの効果

原信号に対して $2^{20/12} = 3.1748$ 倍のピッチシフトを試みる．このとき立ち上がり区間終端時刻 t_a の設定は，原信号系列 $f[k]$ のうち音圧の絶対値が最大となる k について $t_a = kh$ とする．h と t_a の値はそれぞれ $h = 1/44100 = 2.2676 \times 10^{-5}$(s)，$t_a = 93.7$(ms) となる．Fig. 6，7 はそれぞれ原信号の区間 $I_8 := [t_{1,8},\ t_{1,9}] = [179.9,\ 192.2]$ (ms)，および出力信号の区間 $J_8 := [t_{2,8},\ t_{2,9}] = [116.6,\ 129.0]$ (ms) 上の波形を表す．原信号の基本周期は $T_1 = 9.1$(ms)，出力信号の基本周期は $T_2 = 2.9$(ms)，原信号の分割フレーム長 (減衰区間) は $\tau_1 = 12.3$(ms)，並べる基本波形の数は $L = 3$ である．区間 J_8 のフレーム g_8 は区間 I_8 のフレーム f_8 上の信号にもとづいて構成される．まずフレーム f_8 の全体を時間軸方向に伸長させた信号がフレーム g_8 上の区間 $[t_{2,8},\ t_{2,8} + \tau_1/R] = [116.6,\ 120.5]$(ms) 上に構成され，次に後続の区間は区間 $[t_{2,8} + \tau_1/R - T_2,\ t_{2,8} + \tau_1/R] = [117.6,\ 120.5]$(ms) の信号を基本波形をみなし，これを参照して，後続の区間に基本波形を $L = 3$ 個並べている様子がわかり，2.3.2 節のアルゴリズムにもとづいて信号が合成されていることが確認できる．また Fig. 8 は出力信号を $t = t_{2,10} = 141.4$(ms) まで拡張した区間の波形を表す．区間 $J_8 = [t_{2,8},\ t_{2,9}]$ のフレームと区

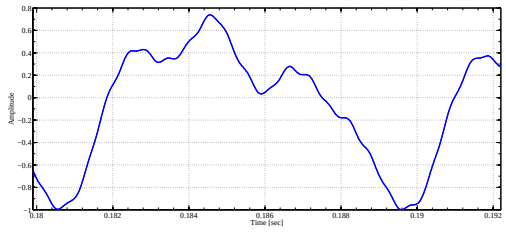


Fig. 6: Original signal on interval $[t_{1,8},\ t_{1,9}]$

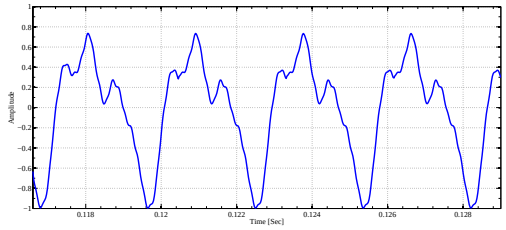


Fig. 7: Modulated signal on interval $[t_{2,8},\ t_{2,9}]$

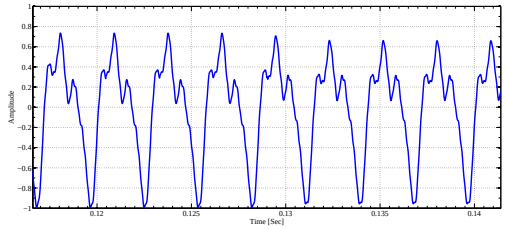


Fig. 8: Modulated signal on interval $[t_{2,8},\ t_{2,10}]$

間 $J_9 := [t_{2,9},\ t_{2,10}]$ のフレームとの間にクリックノイズを生じさせることなく接続されていることが読み取れる．

3.2　従来ピッチシフト手法との比較

従来よりピッチシフト手法としてよく用いられるフェーズボコーダ [2, 5, 1] による手法と提案法について数値例による比較検討を行う．原信号をそれぞれの手法で変調させる．さらに逆数倍率で再度変調を行った信号と原信号との誤差をとる．提案法，従来法について倍率 2 倍，1/2 倍の順にピッチシフトさせた復元信号と原信号との系列ごとの誤差をそれぞれ Fig. 9，Fig. 10 に示す．両者の比較から提案法の方が信号のもつ特徴を劣化させずに変調できていることが確認できる．原信号および従来法による復元信号を拡大した Fig. 11 からも，従来法による信号劣化が確認できる．提案法，従来法ともに音の局所的特性にもとづいた手法である．提案法は楽器音の基本波形に着目した手法であり，その形は保存される．それに対して従来法は各フレームの周波数分布に着目した手法であるが，信号の基本波形が保存される保障はない．また従来法は，各分割フレームに対するフーリエ変換前，逆フーリエ変換後の段階で窓関数を掛ける操作を行うが，この操作も信号劣化の一因になっていると考えられる．

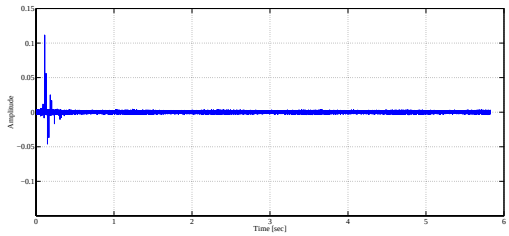


Fig. 9: Reconstruction error of proposed method

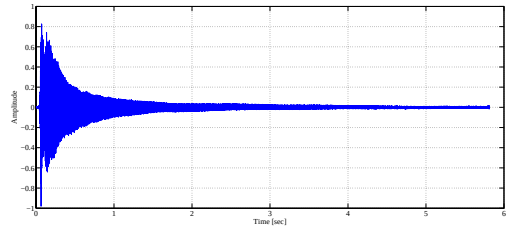


Fig. 10: Reconstruction error of phase vocoder

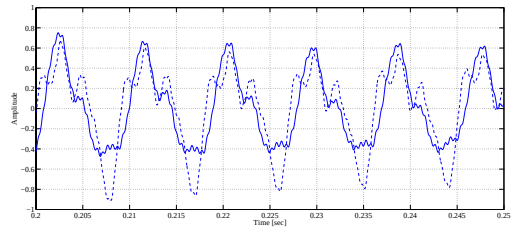


Fig. 11: Reconstructed signal of phase vocoder (solid) and original signal (dash)

4　おわりに

本稿では，ピッチシフトにおける補間処理に，サンプル値制御理論にもとづき設計された非整数遅延フィルタを用いることを提案した．このフィルタは 1 次の FIR フィルタであり，非常に高速な処理が可能である．さらに，変換後の信号長を保存するために，楽器の信号の特性を考慮して，ブロック処理により信号長を自由に変化させる方法についても提案した．実際の弦楽器の音源に対して提案法を適用し，従来法であるフェーズボコーダ法よりも性能の優れた音高シフトが可能であることを例題により示した．

参考文献

1) D. P. W. Ellis: A phase vocoder in matlab, (2002) `http://www.ee.columbia.edu/~dpwe/resources/matlab/pvoc/`

2) J. L. Flanagan and R. M. Golden: Phase vocoder, *Bell System Technical Journal*, 1493/1504 (1966)

3) M. Nagahara and Y. Yamamoto: Optimal design of fractional delay filters, *Proc. of 42nd IEEE Conf. on Decision and Control*, 6539/6544 (2003)

4) M. Nagahara and Y. Yamamoto: Optimal design of fractional delay FIR filters without band-limiting assumption, *Proc. of ICASSP*, 221/224 (2005)

5) M. R. Portoff: Implementation of the digital phase vocoder using the fast fourier transform, *IEEE Trans. Acous. Speech Sig. Proc.*, 24(3), 243/248 (1976)

6) C. Roads: *The Computer Music Tutorial*, MIT Press (1996)