音楽情報科学 24－3
(1998. 2. 19)

KL展開を用いた歌唱音源システム

新井清嗣
イクスクラ
〒339-0001 埼玉県岩槻市鹿室354
048-794-4193
arai@excla.com
http://www.excla.com/

足立雅人
コルグ
〒168-0073 東京都杉並区下高井戸1-15-12
03-5376-5231
adachi@korg.co.jp
http://www.korg.co.jp/

本研究は音程と音量をパラメーターとして音色を精密に制御するシンセサイザー方式に関するものである。音響系統はKL基底波形による残差音源と復色化のためのラティス・フィルターおよび低次フィルターからなり、それらの係数を時々刻々のピッチとレベルに応じて供給する2次元係数テーブルを制御系統に持つ点に特徴がある。本方式によれば、微かにビブラートを掛けただけでも音色が自然に変化するので、ループ特有の表情の貧弱さや人工感がなく、1オクターブに及ぶ自然なポルタメントも可能で、これまで合成が困難であった歌声合成(母音)に本方式を適用して良好な結果を得ることができた。

歌声, ミュージック・シンセサイザー, 加算合成,KL展開, 母音, ビブラート

A singing synthesis system based on the Karhunen-Loève transform

Arai,Kiyotsugu
Excla Inc.
354,Kanamuro,Iwatsuki-si,
Saitama-ken,Japan,339-0001
81-48-794-4193
arai@excla.com
http://www.excla.com/

Adachi,Masato
KORG Inc.
1-15-12,Shimotakaido,Suginami-ku,
Tokyo,Japan,168-0073
81-3-5376-5231
adachi@korg.co.jp
http://www.korg.co.jp/

A new music synthesizer which has a KL expanded residuals, a latice filter and a low order filter is presented. Existence of 2-dimensional coefficient tables which control timbre accurately according to instantaneous pitch and level distinguishes this system. As the timbre varies even by a very weak vibrato or fluctuation, this synthesizer can overcome poorness and artificiality of short looped sounds. And the single KL based oscillator covers wide range of the scale so that a natual portamento over 1 octave is also achieved. Singing voice synthesis has been difficult for conventional PCM-type synthesizers, but this new system can "sing" vowels very naturally.

Singing voice, Music synthesizer, Additive synthesis, Karhunen-Loève transform, Vowel, Vibrato

1 はじめに

ミュージック・シンセサイザー(以下シンセサイザー)において、リアルな楽器音を容易に再生できるPCM音源方式は、ここ10年間の主要な音源技術であった。しかし多様な楽器音を収録して現実的なコストで提供するには、1音色あたりの使用メモリー量を抑えねばならず、いくつかのメモリー削減方式が常用され、音質劣化の原因となっている。

収録する音程数を減らした場合には、収録しなかった音程を上下音程サンプルから補間再生するために音質が劣化してしまい、スケール演奏時に音色の不連続性が目立ってしまう問題があった。

定常区間を反復再生することで時間方向にメモリー量を削減する「ループ」方式は、ループ区間が短いと音の表情が極端に貧しくなり、人工的な感じになってしまう問題があった。

また、共鳴体が複雑な伝達関数を有するチェロ等の低音弦楽器では、微かなビブラートでもスペクトラムに大きな変調を生じるが、従来PCM方式は単一波形の再生速度を上下していただけであり、ビブラートに伴うスペクトラム変化を表現できなかった。

2 合成システムの構成

今回報告するシンセサイザー方式は、Karhunen-Loève (以下KL) 展開 [1]-[3] された残差音源と線形時変フィルターを組み合わせた音響系統を持ち、その係数すべてが2次元係数テーブルを介してピッチとレベルに応じて精密に制御される点に特徴があり (**図1**)、微かに揺らいだピッチとレベルを与えることで、前述の問題を大幅に改善できる。

以下、説明に当たり、鍵盤等によって与えられるMIDIノートナンバー、ピッチ・エンベロープ、ビブラート変調LFO、ピッチベンダー位置などによって定まる対数周波数を「ピッチ」と呼び、その楽器の発音可能な最低周波数との比をとって [cent] で表わすものとする[1]。

[1] 100[cent] が半音に相当する。

また、打鍵速度、アンプ・エンベロープ、エクスプレッション・ペダル、アフタータッチなどによって与えられるデシベル値を「レベル」と呼ぶことにする。

なお時間はすべて離散時間とし、n でサンプル番号を表わし、サンプリング周波数は F_s[Hz] とする。

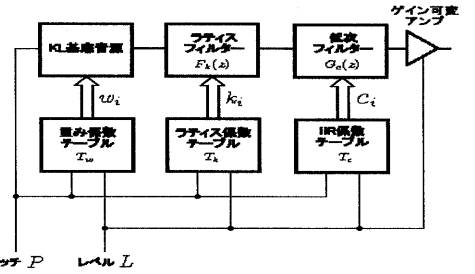


図1:合成システム構成

2.1 KL基底音源

KL基底音源は同期した K個のKL基底波形 b_i を指定ピッチで反復再生しながら、各々に時々刻々のKL重み係数 $w_i(n)$ を掛けて加算する (**図2**)。

$$y_0(n) = \sum_{i=1}^{K} w_i(n) b_i(n) \qquad (1)$$

KL基底波形 b_iは、そのフーリエ係数 d_{im}で定まる固定PCM周期波形である[2]。

$$b_i(n) = \sum_{m=1}^{M} d_{im} sin(m\theta(n)) \qquad (2)$$

ここに$\theta(n)$ は位相加算部でピッチを指数変換し、累算することで得られる読み出しアドレスで、全基底波形共通である。

一方、KL重み係数 $w_i(n)$ は後述の重み係数テーブルT_wから時々刻々のピッチとレベルに応じて供給されるものである。

[2] 演算量およびメモリー量削減のため、現在は位相を無視し、sin 合成している。

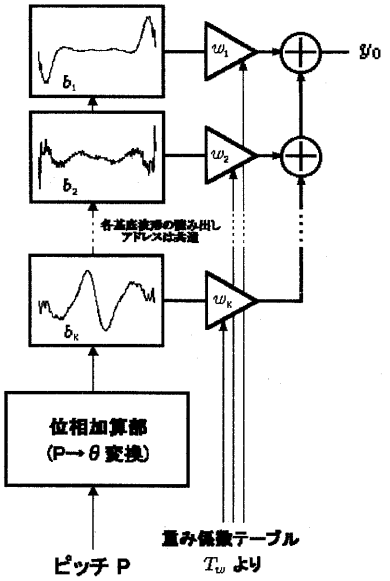


図 2:KL 基底音源

2.2 ラティス・フィルター

ラティス・フィルター $F_k(z)$ は全極型 IIR で、フォルマント特性を付加するものである。

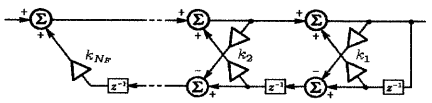


図 3:ラティス・フィルター $F_k(z)$

ただし係数 $k_i(n)$ は時々刻々のピッチとレベルに応じて後述のラティス係数テーブル T_k から供給される。今後はこれを「ラティス係数」と略称する。

2.3 低次フィルター

低次フィルター $G_c(z)$ は低次の IIR で、大まかなスペクトラム傾斜特性を付加するものである。現在は静特性が式 (3) の直流ゲイン 1 の高域減衰型 1 次 IIR を使用している[3]。

$$G_c(z) = \frac{1+c_1}{1+c_2} \cdot \frac{1+c_2 z^{-1}}{1+c_1 z^{-1}} \qquad (3)$$

[3]歌声では $c_2(n) \simeq 1$ となるので、ゲイン項を $(1+c_1(n))/2$ としてしまっても差し支えない。

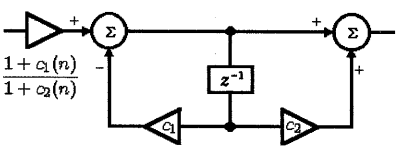


図 4:低次フィルター $G_c(z)$

ただし係数 $c_i(n)$ は時々刻々のピッチとレベルに応じて後述の低次フィルター係数テーブル T_c から供給される。

2.4 2 次元係数テーブル

時々刻々のピッチ $P(n)$ とレベル $L(n)$ にふさわしい音色が得られるように KL 重み係数 $w(n)$、ラティス係数 $k(n)$、低次フィルター係数 $c(n)$ を供給する 3 種類の 2 次元テーブル T_w, T_k, T_c は本方式を特徴づける構成要素である。

$$w(n) = \tilde{T}_w(P(n), L(n)) \qquad (4)$$
$$k(n) = \tilde{T}_k(P(n), L(n)) \qquad (5)$$
$$c(n) = \tilde{T}_c(P(n), L(n)) \qquad (6)$$

ここで$\tilde{T}$は補間値であることを示す

いずれもピッチとレベルから 2 次元テーブルを参照するものであるが、3 種類ともハード/ソフト的に同一構成で実現できるので、総じて「2 次元係数テーブル」または単に「係数テーブル」と呼ぶことにする。

各 2 次元係数テーブルのピッチ方向は共通に数 10[cent] 刻みで $\{P_0, P_1, P_2, \cdots, P_{Np}\}$ と等分割されており、半音 (100[cent]) 以下の細かい分解能を持つ事で、微かなビブラートやベンダー操作にも微妙な音色変化が得られるようになっている。

同様にレベル方向も共通に数 [dB] 刻みで $\{L_0, L_1, L_2, \cdots, L_{Nl}\}$ と等分割されており、十分な分解能を持つ事で、ゆっくりとデクレッシェンドしてゆく場合にも自然な音色変化が得られるようになっている。

鍵盤その他コントローラーからピッチとレベル $(P(n), L(n))$ が与えられると、2 次元テーブル内で

$$P_p \le P(n) < P_{p+1}, L_l \le L(n) < L_{l+1} \qquad (7)$$

$$P(n) = P_p + (P_{p+1} - P_p)\Delta p \quad (8)$$
$$L(n) = L_l + (L_{l+1} - L_l)\Delta l$$

となる (p,l) の近傍メッシュ4 点の係数を双 1 次補間して音響系統各部に供給する。

$$\tilde{T}(P(n), L(n)) = (1-\Delta l)\tilde{T}_0 + \Delta l \tilde{T}_1 \quad (9)$$
$$\tilde{T}_0 = \tilde{T}(P(n), L_l) \quad (10)$$
$$= (1-\Delta p)T(P_p, L_l) + \Delta p T(P_{p+1}, L_l)$$
$$\tilde{T}_1 = \tilde{T}(P(n), L_{l+1}) \quad (11)$$
$$= (1-\Delta p)T(P_p, L_{l+1}) + \Delta p T(P_{p+1}, L_{l+1})$$

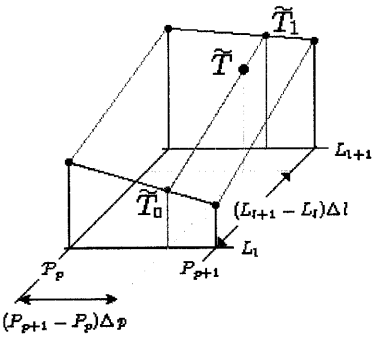


図 5:双 1 次補間

2.5 ゲイン可変アンプ

出力パワーを制御するもので、

$$g(n) = 10^{\frac{L(n)}{20}} \quad (12)$$

なる時々刻々のゲインを掛ける乗算器である。

3 係数設計手順

本方式の各係数を設計する手順をフローチャートにして**図 6** に示す。分析は合成システムの逆を辿って進む。

3.1 楽器音収録

本シンセサイザー方式は、ピッチとレベルに応じて時々刻々の音色が制御されるので、実際の楽器が様々なピッチ、レベルでどのような音色 (すなわちフーリエ係数) を持つかを網羅的に測定する必要がある。そこで、± 50〜100[cent] 程度のビブラートを掛けながら、フォルテシモからピアニシモまでデクレッシェンドする単音サンプルを、演奏可能な範囲で半音毎にサンプリングする[4]。

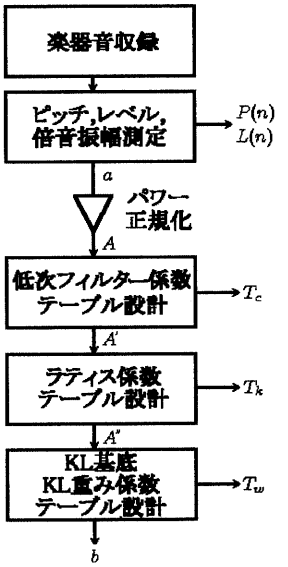


図 6:係数設計フローチャート

3.2 ピッチ, レベル, 倍音振幅測定

楽器によっては高速にスペクトラムが変化するものがあるので、分析には時間分解能の高い解析信号による方法[5]を使用し、収録したすべての楽音サンプルの各倍音振幅 $a_m(n)$ および基音周波数 $f_1(n)$ を測定する[6]。

ピッチ $P(n)$ は得られた $f_1(n)$ から

$$P(n) = 1200 log_2 \frac{f_1(n)}{f_{min}} \quad (13)$$
$$f_{min} = \min_n f_1(n) \quad (14)$$

[4]半音毎に網羅的にサンプリングすることが望ましいが、メッシュの「抜け」は上下左右の係数から補間できるので、楽器によっては全音 (200[cent]) かそれ以上の間引きも可能と思われる。

[5]サンプリングされた楽器音の倍音 1 本ごとに解析信号を求めて、その絶対値によって瞬時振幅を、位相差分によって瞬時周波数を測定する方法。

[6]ノイズの影響を低減するため、数 [msec] から数 10[msec] ごとに時間方向のメディアンをとる。

によって計算する。f_{min}は得られた$f_1(n)$の全サンプルにおける最低値(最低周波数)である。

レベル$L(n)$は得られた$a_m(n)$から

$$L(n) = 10log_{10} \sum_{m=1}^{M} |a_m(n)|^2 \qquad (15)$$

によって計算する。

3.3 パワー正規化

合成システムはゲイン可変アンプを持っており、ここで最終的にレベルを制御する。従って以降の分析に使用する倍音振幅はパワーで正規化しておく。

$$A_m(n) = a_m(n)10^{-\frac{L(n)}{20}} \qquad (16)$$

$$\sum_{m=1}^{M} |A_m(n)|^2 = 1 \qquad (17)$$

以下、これを正規化倍音振幅と呼ぶ。

3.4 低次フィルター係数テーブル設計

まず、係数$c = \{c_1, c_2\}$によって定まる低次フィルターG_cの振幅伝達関数

$$G_c(f) = G_c(z)|_{e^{j2\pi \frac{f}{F_s}}} \qquad (18)$$

が時々刻々の各正規化倍音振幅$A_m(n)$にデシベルで誤差最小になるよう、最適な低次フィルター係数$c(n)$をPowel法で求める。

$$c(n) = \{c_1, c_2 | \min_c \varepsilon_c(n)\} \qquad (19)$$

$$\varepsilon_c(n) = \sum_{m=1}^{M} 10log_{10} \left| \frac{G_c(mf_1(n))}{A_m(n)} \right|^2 \qquad (20)$$

次に$c(n)$を各時点の$(P(n), L(n))$に基づきピッチ-レベル平面上にプロットすると**図7**のような分布が得られる(図はc_1の分布例)。

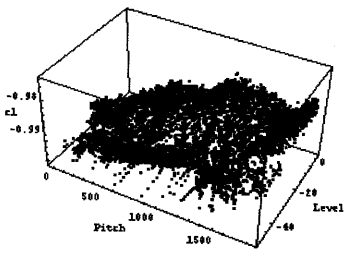


図7:低次フィルター係数c_1分布

得られた分布に対し、2次元係数テーブルの各メッシュ$[P_p, P_{p+1}) \times [L_l, L_{l+1})$に属する点を抜き出し、メディアンをとって低次フィルター係数テーブルT_cのメッシュ点(P_p, L_l)における値とする[7]。

さらに以降の分析のため、得られた係数テーブルT_cを使って、各正規化倍音振幅$A_m(n)$を白色化し、第1残差倍音振幅$A'_m(n)$を求めておく[8]。

$$A'_m(n) = \frac{1}{|G_{\tilde{c}(n)}(mf_1(n))|} A_m(n) \qquad (21)$$

$$\tilde{c}(n) = \tilde{T}_c(P(n), L(n)) \qquad (22)$$

3.5 ラティス係数テーブル設計

まず、係数$k = \{k_1, k_2, \cdots, k_{N_F}\}$によって定まるラティス・フィルター$F_k$の振幅伝達関数

$$F_k(f) = F_k(z)|_{e^{j2\pi \frac{f}{F_s}}} \qquad (23)$$

と時々刻々の第1残差倍音振幅$A'_m(n)$の比の絶対値が1に近づくよう、Hooke-Jeeves法で最適なラティス係数$k(n)$を求める[9]。

$$k(n) = \{k_1, k_2, \cdots, k_{N_F} | \min_k \varepsilon_k(n)\} \qquad (24)$$

---

[7] ひとつもデータが存在しないメッシュは適当に補間・補外して埋める。

[8] これはパワーが一定になるようゲイン調整された楽器音を、ピッチとレベルに応じて周波数特性が変化する1次FIR(1次IIRの逆フィルター)に通して残差信号を求める処理を周波数領域で行うことと考えられる。

[9] 初期値は$A'_m(n)$に基づいてPARCOR分析によって得た値を使用している。

$$\varepsilon_k(n) = \sum_{m=1}^{M} \omega(mf_1(n)) \left(\left| \frac{F_k(mf_1(n))}{A'_m(n)} \right| - 1 \right)^2 \quad (25)$$

$$\omega(f) = \left(\frac{f}{F_s} \right)^r \quad (26)$$

なお、ここでは人間の聴覚特性を考慮して、低域を重視する重みづけ$\omega(f)$も施している[10]。

同様に$k(n)$を各時点の$(P(n), L(n))$に基づきピッチ-レベル平面上にプロットすると**図8**のような分布が得られる(図はk_1の分布例)。

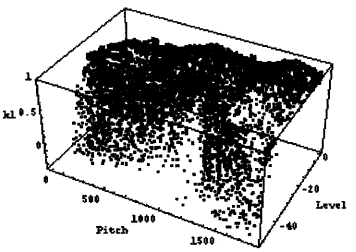


図8:ラティス係数k_1分布

こうして得られた分布に対しても、メッシュごとにメディアンをとって、ラティス係数テーブルT_kの値とする。

さらに以降の分析のため、得られた係数テーブルT_kを使って、第1残差倍音振幅$A'_m(n)$を白色化し、第2残差倍音振幅$A''_m(n)$を求めておく[11]。

$$A''_m(n) = \frac{1}{|F_{\tilde{k}(n)}(mf_1(n))|} A'_m(n) \quad (27)$$

$$\tilde{k}(n) = \tilde{T}_k(P(n), L(n)) \quad (28)$$

3.6 KL基底設計

$A''_m(n)$をそのままKL展開したのでは誤差の2乗和が最小化され、レベルの小さな倍音のデシベル近似誤差が大きくなって、人間の聴覚との不整合が生じてしまう。そこで各$A''_m(n)$の時間方向の最大値を求め、予め最大値が1に近づくように弱い正規化[12]を施した$Q(n)$を求めておく。

$$Q(n) = \{ \frac{A''_1(n)}{\alpha_1}, \frac{A''_2(n)}{\alpha_2}, \cdots, \frac{A''_M(n)}{\alpha_M} \}^T \quad (29)$$

$$\alpha_m = \left(\max_n A''_m(n) \right)^{\mu(m)} \quad (30)$$

この$Q(n)$をKL展開して基底V_iを得る。

$$R = E\left[QQ^T\right] \quad (31)$$

$$RV_i = \lambda_i V_i \quad (32)$$

ただし

$$\lambda_1 \geq \lambda_2 \geq \cdots \lambda_K \cdots \geq \lambda_M \quad (33)$$

$$|V_i| = 1 \quad (34)$$

得られた基底V_iにかかっている先ほどの弱い正規化を取り除き、フーリエ逆変換することでKL基底波形bを求める。

$$d_i = \begin{pmatrix} \alpha_1 & & 0 \\ & \alpha_2 & \\ & \cdots & \\ 0 & & \alpha_M \end{pmatrix} V_i \quad (35)$$

$$b_i(n) = \sum_{m=1}^{M} d_{im} sin(2\pi m \frac{n}{N_b}) \quad (36)$$

N_bは基底波形の1周期のサンプル数である。

3.7 KL重み係数テーブル設計

式(31〜34)より基底V_iは正規直交しているので、重み係数は内積によって求められる。

$$w_i(n) = \langle Q(n), V_i \rangle \quad (37)$$

$w(n)$を各時点の$(P(n), L(n))$に基づきピッチ-レベル平面上にプロットすると**図9**のような分布が得られる(図はw_1の分布例)。

[10]指数rは聴感上の歪みが少なくなる値を実験的に求め$r = -1.5$としたが、重みの周波数特性は今後改良の余地が大きいと思われる

[11]これはピッチとレベルに応じて周波数特性が変化するラティスFIR(ラティスIIRの逆フィルター)に第1残差信号を通して第2残差信号を求める操作に当たる。

[12]完全に1に正規化してしまうとノイズ・フロアーに埋もれている弱い倍音成分が強調され過ぎてしまう。ここでは実験的に$\mu(m) = 0.05 + 0.95/m$とし、やはり低次倍音を重視するようにした。

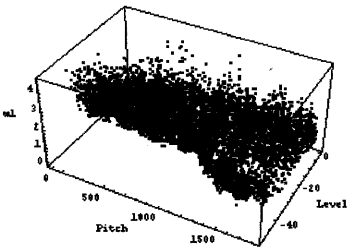


図 9:KL 重み係数 w_1 分布

ここで得られた分布に対しても、メッシュごとにメディアンをとって、KL 重み係数テーブル T_w の値とする。

4 分析例

日本人女性プロ歌手 A の歌声を分析した結果の一部を示す。使用したのは E_3〜A_4 の 1 オクターブ半の音域を半音ごとにビブラートをかけながらフォルテシモからピアニシモまでデクレッシェンドした 18 個の母音「ア」のサンプルである。

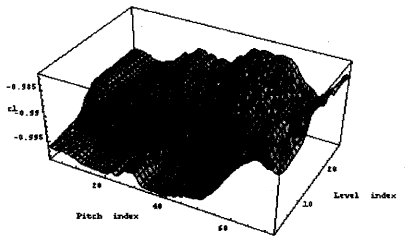


図 10:低次フィルター係数テーブル T_{c1}
(図 7 の分布を近似したもの)

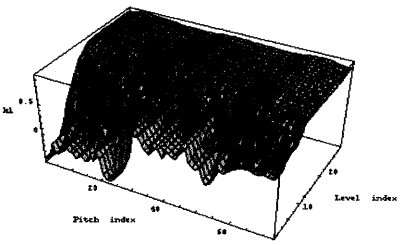


図 11:ラティス係数テーブル T_{k1}
(図 8 の分布を近似したもの)

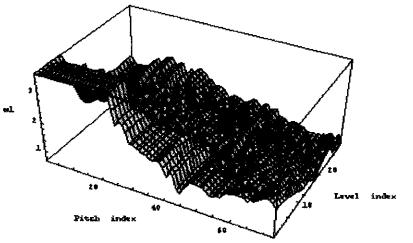
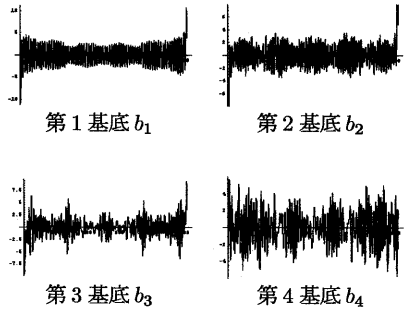


図 12:重み係数テーブル T_{w1}
(図 9 の分布を近似したもの)



第 1 基底 b_1　　第 2 基底 b_2

第 3 基底 b_3　　第 4 基底 b_4

図 13:KL 基底波形

5 合成条件

外国人女性プロ歌手 B が歌う一節から、そのピッチとレベルを解析信号による方法で求め[13]、先に得た日本人女性プロ歌手 A の基底波形と 2 次元係数テーブルを用いて合成した。

合成時のサンプリング周波数 F_s は 48[kHz]、KL 基底は 1 周期 1024 サンプルのものを 24〜48 個、ラティス・フィルターは 32 次、低次フィルターは 1 次 (係数は 2 個)、各 2 次元係数テーブルのメッシュ分割は、ピッチ方向を E_3〜A_4 の範囲で 25[cent] 刻み、レベル方向を 50[dB] の範囲で約 2[dB] 刻みである。

ただし歌手 B は一部ファルセット (裏声) を使用して歌手 A の声域より 1 オクターブ高く歌っているため、1 オクターブ下げて歌手 A の音域内に移動した。

[13] ピッチ情報はノート・ナンバーとピッチベンド、レベル情報はベロシティーとエクスプレッションで表現し、Standard MIDI file (SMF) として記録できるようになっている。

6 試聴結果

合成音を試聴した結果、歌手Aが歌手Bの節回しで1オクターブ下げて歌っているような自然な歌声が、広い音域にわたって得られることが確認できた。

なお、今回はハードウェア試作に当たってラティス・フィルター次数を32次と多めに確保したが、KL基底24個でも十分に歌手Aらしい声が得られているので、もっと次数を下げてもよいかもしれない。

7 今後の課題

本シンセサイザー方式は歌声だけを対象としたものではないので、他楽器との兼ね合いを考慮しなければならないが、現実的な価格レンジ内で十分なクオリティーの子音をも合成できるなら大きな付加価値となる。歌声以外にもノイズ成分が重要な楽器は多数存在するので、今後ノイズ(子音)合成についても検討したい。

また、本方式は「ピッチとレベルが決まれば音色も一意に決定される」という前提に立っているため、今回報告した範囲内では同ピッチ、同レベルで口の形や舌の位置だけを変えて音色を変化させる「わたり」のようなことができない。この点についても拡張を検討したい。

8 まとめ

ピッチとレベルによって精密に音色を制御するKL基底波形を残差音源としたシンセサイザー方式を考案し、従来PCM方式シンセサイザーで合成が困難であった歌声(母音)に適用して、その有効性を確認した。

本方式は発音域を単一のKL基底セットでカバーするので、スケール演奏時にも音色の不連続が生じず、従来のPCM方式シンセサイザーでは不可能であった1オクターブに及ぶ自然なポルタメント演奏も可能である。

しかも、ピッチ、レベルを微かに変動させるだけで音色が自然に変化するので、ループ特有の表情の貧弱さや人工感がなく、複雑な共鳴特性を有する楽器のビブラートもリアルに合成できる長所を持つ。

さらに、音作りにおいても、一部のシンセサイザー方式では膨大な試行錯誤(手作業)を要することがあるが、本方式は測定によって得られるピッチとレベルに基づくパラメトリックな圧縮-伸長技術であり、試行錯誤はほとんど必要ない。基底の数やフィルター次数等が決まれば、コンピューターに任せきりで原音に非常に近い音が最初から得られる。

9 謝辞

本研究の機会を与えてくださった(株)コルグ加藤孟会長、加藤世紀副社長、三枝文夫取締役、池内順一商品開発部長をはじめ、ハードウェア試作にご協力いただいた皆様に感謝致します。

参考文献

[1] J.C.Stapleton, S.C.Bass "Synthesis of Musical Tones Based on the Karhunen Loève Transform" IEEE Trans. on acoustics,speech,and signal processing. Vol.36, No.3

[2] W.Brent Weeks, W.Andrew Shloss, R.Lynn Kirlin "Implementation of the KL Synthesis Algorithm under Real-Time Control" ICMC 1991

[3] G.Berndtsson "The KTH Rule System for Singing Synthesis" Computer Music Journal 20:1, Spring 1996

[4] G.Bennet, X.Rodet "Synthesis of the Singing Voice" in M.V.Mathews and J.R.Pierce, 1989, Current Directions in Computer Music Research. MIT press, pp. 19-44.