Vol.2012-MUS-94 No.10
Vol.2012-SLP-90 No.10
2012/2/3

v.Connect：ユーザが声色操作可能な歌声合成器

小川　真[†1]　矢崎俊志[†1]　阿部公輝[†1]

VOCALOID「初音ミク」の発売以来，ユーザが自由に歌声ライブラリを制作できるフリーの歌声合成器 UTAU が開発されるなど，歌声合成への関心が高まっている．これら歌声合成器は主にアマチュアの音楽制作に使用されるが，ユーザが声色を任意時刻に混ぜて指定する機能がない．また，声色操作を行うことで処理時間やデータ量が大きくなる．本研究では音声合成分析系 WORLD を用い，メルケプストラムと Vorbis による励起信号からなるコーパスを声色別に収録し，各音素間を時間伸縮関数で接続することで，ユーザがモーフィング率を指定し声色を操作できる歌声合成器 v.Connect を開発した．提案手法を用いて歌声コーパス「波音リツコネクト」を制作した．このコーパスの容量は波形の 2 倍程度であった．合成速度は 1.7〜2.2 倍と改善され，圧縮による劣化は主観的には感じられなかった．

v.Connect : A Singing Synthesis System Enabling Users to Control Vocal Tones

Makoto Ogawa ,[†1] Syunji Yazaki[†1] and Kôki Abe[†1]

Since the release of Hatsune Miku, interets in singing synthesis increase. For example, a singing synthesis system, UTAU, has been developed as a freeware. Most of these systems, however, lack of the function that users can mix vocal tones at any times. Controling tonal changes in singing requires a large amount of time and data for synthesis. We have developed a singing synthesis system, v.Connect, which connects corresponding phonemes with a time-stretching function to enable users to control tonal changes in singing by specifying the rate of voice morphing. The system processes voice signals with WORLD, a voice synthesis and analysis system, and uses corpora of various tonal voices consisting of Mel cepstra and excitation signals compressed by Vorbis. We constructed a corpus, "Namine Ritsu Connect", using the proposed method. It was found that the size of the corpus was two times larger than that of raw waves, and that synthesis from the corpus was 1.7 to 2.2 times faster than that from raw waves. Degradation caused by compression was not sensed subjectively.

1. はじめに

VOCALOID「初音ミク」[1] の発売以来，歌声合成に対する注目が高まっている．VOCALOID はサンプリングされた音声を利用し歌声を合成するソフトウェアであり，ユーザが入力した歌詞，メロディ，演奏情報から歌声を合成する．同一話者における声色の違いを収録した初音ミク Append が発売されるなど，声色に対しても注目がなされている．また，ユーザが自由に歌声ライブラリを作成し歌声を合成できる，フリーの歌声合成ツール UTAU[2] が発表されアマチュアによる歌声ライブラリの制作も盛んに行われており，その中でもピッチの差異や，声色の差異などを収録した歌声ライブラリが制作・発表されている．

声質に注目した研究では任意の話者や感情音声同士を混ぜる技術として，豊田ら[3] による歌声モーフィングが挙げられる．歌声モーフィングでは，時間周波数平面上で自動抽出された候補点から手作業により対応点を選択し，歌声に対してモーフィングを行っている．また中野ら[4] は同一話者における声質の違いとして声色に着目し，VocalListener2 を開発した．VocalListener2 は歌声を入力とし複数のコーパスから合成した音声を用いて，歌声に含まれる声色変化を転写するシステムである．VocalListener2 では入力に歌声が必要であること，転写パラメタが多次元のベクトル量であることから，ユーザが任意時刻での声色を直接指定することが難しい．

本研究では，音声モーフィングを用いユーザが声色を操作できる歌声合成器 v.Connect を提案する．v.Connect は異なる声色で収録された 2 つ以上のコーパスに対して発音が同一の音素片間で時間伸縮関数を設計し，設計した時間伸縮関数を用いてユーザが指定した任意の時刻でのモーフィング率に従って音声を合成する．これにより，ユーザは音楽制作でよく使用される経時パラメタから声色操作を行える．分析および合成には音声分析合成系 WORLD[5] を用い，対数スペクトルと励起信号の線形補間によりモーフィングを実現する．

本研究では実装した音声モーフィングによる声色操作システムを検証するため，インターネット上でフリーの歌声ライブラリを公開している波音リツ Project[⋆1] と共同で，複数声色を収録した歌声ライブラリ「波音リツコネクト」を制作した．また，制作したライブラリを使用し音声モーフィングを使用した声色操作の効果を調べた．

†1 電気通信大学
The University of Electro-Communications

⋆1 波音リツ公式サイト http://ritsu73.net/

Vol.2012-MUS-94 No.10
Vol.2012-SLP-90 No.10
2012/2/3

本論文では第 2 章で v.Connect の概要，第 3 章でコーパスの構築法，第 4 章で合成法，第 5 章で制作したコーパスを用いた本システムの検討，第 6 章でまとめを述べる．

2. v.Connect

本システムは歌声合成用シーケンサ Cadencii[6] 上で動作する歌声合成器である．Cadencii は Cadencii プロジェクト[⋆1] 内で開発されているオープンソース・ソフトウェアである．v.Connect は，Cadencii プロジェクト内でオープンソース・ソフトウェアとして本論文の第一著者により開発された．本章では Cadencii と v.Connect の概要を述べる．

2.1 Cadencii

Cadencii は VOCALOID2 向けに開発されたシーケンサで，現在はさまざまな歌声合成器向けに拡張されており，VOCALOID, VOCALOID2, Sinsy[⋆2]向け MusicXML, UTAU, AquesTone[⋆3] v.Connect など複数の歌声合成器を統一的に扱える GUI アプリケーションである．

Cadencii の操作画面を**図 1** に示す．ユーザは上部のピアノロール内にて音符の高さ，長さ，歌詞などを指定し，下部のコントロール内にて表情パラメータやピッチのコントロールを手書きの線により操作できる．これら操作は歌声合成器によらず，対応する全ての合成器をほぼ同じ操作から制御できる．

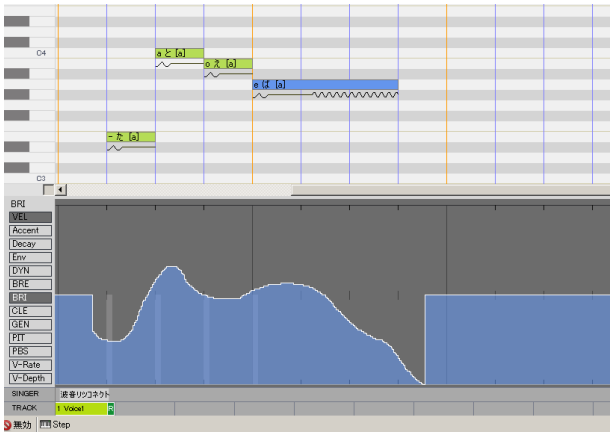

図 1 Cadencii の操作画面
Fig. 1 Screen shot of Cadencii.

⋆1 Cadencii プロジェクト http://sourceforge.jp/projects/cadencii/
⋆2 Sinsy - HMM-based Singing Voice Synthesis System http://www.sinsy.jp/
⋆3 AquesTone - 歌唱合成 VSTi http://www.a-quest.com/products/aquestone.html

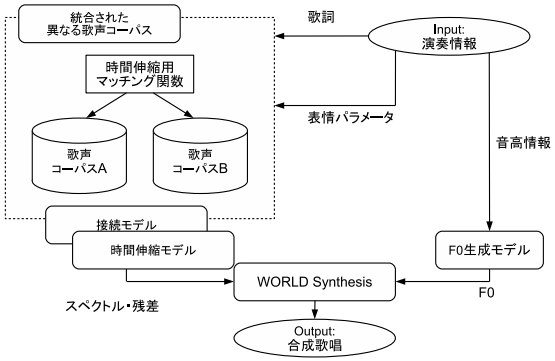

図 2 v.Connect 概要図
Fig. 2 Block diagram of v.Connect.

2.2 v.Connect

v.Connect の概要を**図 2** に示す．

v.Connect が受け取る入力（Input）は VOCALOID2 用の演奏情報に v.Connect 用のコントロールパラメタを追加したシーケンスである．Cadencii はユーザが入力した演奏情報を歌詞，表情パラメタ，音高情報など v.Connect 用のシーケンスに変換して渡す．v.Connect は入力を受け取ると歌声コーパスを参照しながら，音声分析合成系 WORLD を用いて歌唱音声（Output）を合成する．

v.Connect は声色操作のために，異なる声色を持つ歌声コーパスを統合する．本システムが行う声色操作は，参照するコーパス内に複数の声色のデータベースを用意し，それらを時間伸縮関数で接続することで実現される．また声色操作を行うためのコーパスの構築にあたっては，インターネット上での配布できる容量におさえる．また，実行速度をできるだけ速くするために，パワースペクトルには低次メルケプストラム，励起信号にはギャップレス圧縮である OggVorbis[7] を用いる．

Vol.2012-MUS-94 No.10
Vol.2012-SLP-90 No.10
2012/2/3

3. 歌声コーパスの構築

本章では v.Connect が使用する複数の声色を統合したコーパスの構築法について述べる．

各コーパスは VCV 音素単位で収録された音声の低次メルケプストラム，励起信号，および同じ発音で声色の異なる音素片を接続する時間伸縮関数からなる．それぞれの音素片には歌詞として使用される音素名と声色操作用パラメタ，先行発音，音素長が記されている．

音素名はユーザが指定した歌詞とデータを関連付けるために用いられる．声色操作用パラメタは声色操作のモーフィング率を求める際に用いられ，先行発音，音素長は合成時に音素片内で対応する時刻を計算するのに用いられる．

3.1 音声分析合成系 WORLD

本システムでは音声分析合成系 WORLD を使用し分析・合成を行う．音声分析合成系 WORLD は森勢らにより開発された，高速・高品質な Vocoder である．WORLD はピッチ推定法 DIO[8]，スペクトル包絡推定法 STAR[9]，励起信号抽出法 PLATINUM[10] からなる．

PLATINUM は，ピッチマークされた一定区間の信号 $x(t)$ と，区間に対応するパワースペクトルの最小位相応答 $h(t)$ から，以下の式により励起信号スペクトル $R(\omega)$ を求める．

$$R(\omega) = \frac{X(\omega)}{H(\omega)} \ (X(\omega) = \mathrm{FFT}[x(t)w(t)], H(\omega) = \mathrm{FFT}[h(t)]) \tag{1}$$

ここで，$w(t)$ はハニング窓である．窓掛けした波形に対して最小位相応答の逆フィルタをかけることで，WORLD は Vocoder でありながら位相を無視しない．

3.2 分析方法の概要

本システムは WORLD を使用するが，複数の波形を扱うため波形から分析を行いつつ合成を行うと，分析にかかる時間が長くなり実用に適さなくなる．そのため波形をあらかじめ分析することで実行速度の向上を図る．しかし，WORLD による分析で得られるスペクトログラムと励起信号スペクトルは，実波形と比較して非常に大きなサイズとなるためそのままでは実用に適さない．

そこでスペクトログラムは低次メルケプストラムの列，励起信号スペクトルは逆フーリエ変換したのち Vorbis 形式とすることで，品質をあまり下げずに実行速度を改善し，インターネット上の配布に耐えうるサイズまでコーパスを圧縮する．本システムが行う事前分析の概要を**図 3** に示す．

図 3 において，各音素片の音声波形はまず音声分析合成系 WORLD の分析部を用いて，

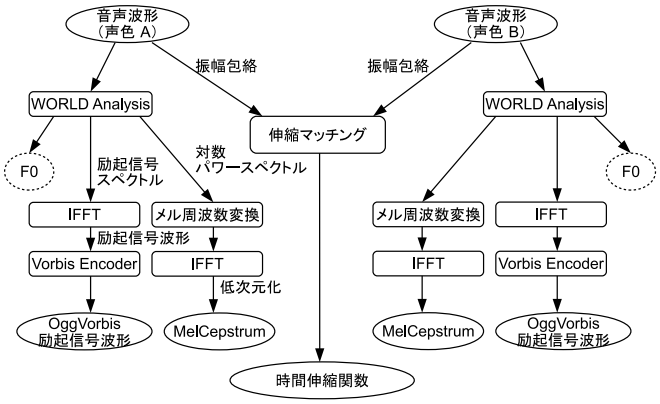

図 3 分析概要図
Fig. 3 Block diagram of v.Connect's analysis.

F0，スペクトログラム，励起信号スペクトルに変換される．その後，スペクトログラムと励起信号スペクトルをそれぞれ低次メルケプストラム，Vorbis による圧縮された励起信号波形へ変換する．さらに 2 つ以上のコーパスを統合するため，発音が同一の音素片ごとに時間伸縮関数を設計する．時間伸縮関数は次節に述べる伸縮マッチングによって求める．

なお，原波形に含まれる F0 情報には有声・無声区間情報も含まれているがすべて破棄される．これは WORLD のピッチ推定の失敗や時間伸縮時に有声・無声区間が重なることによる劣化が，音声モーフィングの際にノイズとして現れる影響が大きいためである．

3.3 時間伸縮関数の設計

本システムが声質変換に使用する時間伸縮関数は，対応する音素片間での時刻の写像関数を用いる．写像関数は発音が同じで声色の異なる音素片 A, B 間での振幅包絡を伸縮マッチングにより求める．

まず時刻 t での振幅包絡 $E(t)$ を音声の信号列 $x(t)$ に対し，

$$E(t) = \sum_{i=-m}^{m} (x(t + \frac{i}{f_s}))^2 \tag{2}$$

として求める．ただし，f_s は標本化周波数，m はサンプリングする点の数を表す定数である．

Vol.2012-MUS-94 No.10
Vol.2012-SLP-90 No.10
2012/2/3

音素片 A, B の振幅包絡 $E_A(t), E_B(t)$ に対し，積分

$$\int_{t=0}^{t=l_A} |E_A(t) - E_B(T(t))| d\sqrt{t^2 + T^2(t)} \quad s.t. \ \frac{dT(t)}{dt} > 0 \tag{3}$$

が最小になるような $T(t)$ を求める．ここで l_A, l_B はそれぞれ音素片 A, B の長さである．ただし，$T(l_A) = l_B$ を満たすとする．

本システムではこの写像関数 $T(t)$ を DP マッチングを用いて近似的に求める．一般的な DP マッチングでは対応付けの重複が許可されており写像として適さないので，DP マッチングの枝を大幅に増やすことで写像関数を近似的に求める．刻み巾 h とし，DP マッチングにおける格子点 $[i, j]$ と $[i+p, j+q]$, $(1 \leq p \leq \frac{l_A}{h} - i, 1 \leq q \leq \frac{l_B}{h} - j)$ を結ぶ枝 d の重みを

$$\begin{cases} d_{ij} < p, q >= \dfrac{h\sqrt{p^2+q^2}}{t_2 - t_1} \displaystyle\int_{t_1}^{t_2} |E_A(t) - E_B(T_{ij} < p, q > (t))| dt \\ t_1 = hi, \quad t_2 = h(i+p) \end{cases} \tag{4}$$

として DP を行う．ここで $T_{ij} < p, q > (t)$ は枝 d に対応する写像関数の一部であり，次式で求める．

$$T_{ij} < p, q > (t) = \begin{cases} \dfrac{q}{p}(t - t_1) + hj & (t_1 < t \leq t_2) \\ 0 & (\text{otherwise}) \end{cases} \tag{5}$$

$\sum_{i=0}^{n} hp_i = l_A, \sum_{i=0}^{n} hq_i = l_B$ を満たす経路 $C = (p_i, q_i)_{0 \leq i \leq n}$ の組み合わせを考える．$p_0 = q_0 = 0$ を満たす任意の C の上で，E_A, E_B の距離 D を次のように定義する．

$$\begin{cases} D = \displaystyle\sum_{i=1}^{n} d^{(i)} \\ d^{(i)} = d_{u_i v_i} < p_i, q_i > \\ u_i = \displaystyle\sum_{j=1}^{i-1} p_j, \quad v_i = \sum_{j=1}^{i-1} q_j \end{cases} \tag{6}$$

距離 D を最小とする経路 C_{min} に対応する時間伸縮関数の列 $T_{min}^{(i)} = T_{u_i v_i} < p_i, q_i >$ から $T = \sum_{i=1}^{n} T_{min}^{(i)}$ と写像関数を求める．しかしこの DP をそのまま解くと計算量が膨大になるため，$1 < N < \frac{1}{h}\max(l_A, l_B)$ の範囲で N を定め，$1 \leq p, q \leq N$ と制限をかけることで枝を減らす．この制限の下で得られた写像関数を時間伸縮関数としてコーパスに保存する．

4. コーパスからの合成

本システムは演奏情報に基づき，事前分析済みのコーパスから各時刻ごとに適した F0，パワースペクトル，励起信号を計算し対応時刻の波形を合成する．本章では演奏情報，F0 の生成，波形の再合成について述べる．

4.1 演奏情報

本システムが入力として Cadencii から受け取る演奏情報は，音符情報と表情パラメタからなる．音符情報は音高，位置，長さ，ビブラート長，ポルタメント長，ポルタメントの深さ，歌詞，ビブラート長を持つ．表情パラメタはビブラートの深さ，ビブラート速度，モーフィング率 BRI からなり，それぞれ時系列 $p(t)$ の形で表される．

4.2 F0 生成モデル

本システムで用いる F0 生成モデルは，前述の演奏情報から音高，長さ，ビブラート情報，ポルタメント長およびポルタメントの深さを参照し F0 列を生成する．一般に歌声中の F0 変動は制動 2 次系のインパルス応答 $H(s) = \dfrac{\Omega}{s^2 + 2\zeta\Omega s + \Omega^2}$ によりオーバーシュート，プリパレーション，ビブラートを制御するモデルとして表される．しかしユーザが音程遷移を操作することを考えると，$H(s)$ による F0 生成モデルは ζ, Ω が音符の長さに直結しておらず，演奏に対応した操作が難しい．そこで以下のように F0 列 $f_0(t)$ を近似する．n 番目の音符の位置，長さ，音高，ポルタメント長，ポルタメントの深さ，ビブラート長を $t^{(n)}, l^{(n)}, f^{(n)}, l_{\text{por}}^{(n)}, d_{\text{por}}^{(n)}, l_{\text{vib}}^{(n)}$，音符の数を n_{note} とすると，$f_0(t)$ は次式で求める．

$$\begin{aligned} \log f_0(t) &= \log f_{\text{vib}}(t) + F_{\text{flu}}(t) + \sum_{i=0}^{n_{\text{note}}} F_{\text{por}}^{(i)}(t) \\ \log f_{\text{vib}}(t) &= \log f_{\text{note}} + \sum_{i=0}^{n_{\text{note}}} F_{\text{vib}}^{(i)}(t) \\ \log f_{\text{note}}(t) &= \begin{cases} f^{(n)} & (t^{(n)} < t \leq t^{(n)} + l^{(n)}) \\ 0 & (\text{otherwise}) \end{cases} \end{aligned} \tag{7}$$

ここに，$F_{\text{por}}^{(n)}(t)$，$F_{\text{vib}}^{(n)}(t)$ はそれぞれ n 番目の音符におけるポルタメントおよびオーバーシュートとプリパレーション，ビブラート， $F_{\text{flu}}(t)$ は微細振動を表し，次式で求める．

Vol.2012-MUS-94 No.10
Vol.2012-SLP-90 No.10
2012/2/3

$$\begin{cases} F_{\mathrm{por}}^{(n)}(t) &= (\log f^{(n+1)} - \log f_{\mathrm{vib}}(t))(\frac{1}{2}(1-\cos\theta_{\mathrm{por}}^{(n)}(t))) \\ & \quad -d_{\mathrm{por}}^{(n)}(\log f^{(n+1)} - \log f^{(n)})(\sin\theta_{\mathrm{pre}}^{(n)}(t)) \\ F_{\mathrm{vib}}^{(n)}(t) &= d(t)\sin\theta_{\mathrm{vib}}^{(n)}(t) \\ F_{\mathrm{flu}}(t) &= \frac{1}{100}(\sin 12.7\pi t + \sin 7.1\pi t + \frac{1}{3}\sin 4.7\pi t) \end{cases} \tag{8}$$

ただし，$\theta_{\mathrm{por}}^{(n)}(t), \theta_{\mathrm{vib}}^{(n)}(t)$ は以下のように定める．

$$\begin{cases} \theta_{\mathrm{por}}^{(n)}(t) &= \begin{cases} \dfrac{t-t_p}{t^{(n)}+l^{(n)}-t_p}\pi & (t_p < t < t^{(n)}+l^{(n)}) \\ 0 & (\text{otherwise}) \end{cases} \\ \theta_{\mathrm{pre}}^{(n)}(t) &= \theta_{\mathrm{por}}^{(n)}(t) + \theta_{\mathrm{por}}^{(n)}(2(t^{(n)}+l^{(n)})-t) \\ \theta_{\mathrm{vib}}^{(n)}(t) &= \begin{cases} \displaystyle\int_{t_v}^{t} s(\tau)d\tau & (t_v < t < t^{(n)}+l^{(n)}) \\ 0 & (\text{otherwise}) \end{cases} \\ t_p = t^{(n)}+l^{(n)}-l_{\mathrm{por}}^{(n)}, & t_v = t^{(n)}+l^{(n)}-l_{\mathrm{vib}}^{(n)} \end{cases} \tag{9}$$

ここで，$d(t), s(t)$ はそれぞれビブラートの深さとビブラートの速度である．なお微細振動は Wen-Hsing Lai による Mandarin Singing Synthesis[11] で用いられているものを使用している．

4.3 声色操作

本システムにおける声色操作は音素片間の音声モーフィングとして実現される．各音符の歌詞に対応する音素片のうち，該当時刻におけるユーザの入力した BRI パラメタと最も近い BRI 表情パラメタを持つ音素片と二番目に近い音素片の 2 つをモーフィングする．モーフィングで使用するモーフィング率は BRI パラメタから求めた値を使用し，写像関数による各音素片の時間伸縮と，パワースペクトルと励起信号の重みつき和で線形補間を行う．

本システムの声色操作の概要を**図 4** に示す． ある音符内での時刻 $t' = t - t^{(n)}$ を元に，モーフィングを行う 2 つの音素片 A, B の対応する時刻 t'_A, t'_B をモーフィング率から求める．音素片 A と音素片 B の BRI 表情パラメタを $b_A, b_B (b_A > b_B)$，先行発音長を p_A, p_B としたとき，モーフィング率 $b(t)$ は入力内の BRI パラメタの値を $p_{\mathrm{bri}}(t)$ として

$$b(t) = \begin{cases} \dfrac{p_{\mathrm{bri}}(t)-b_B}{b_A-b_B} & (b_B \le p_{\mathrm{bri}}(t) \le b_A) \\ 0 & (p_{\mathrm{bri}}(t) < b_B) \\ 1 & (p_{\mathrm{bri}}(t) < b_A) \end{cases} \tag{10}$$

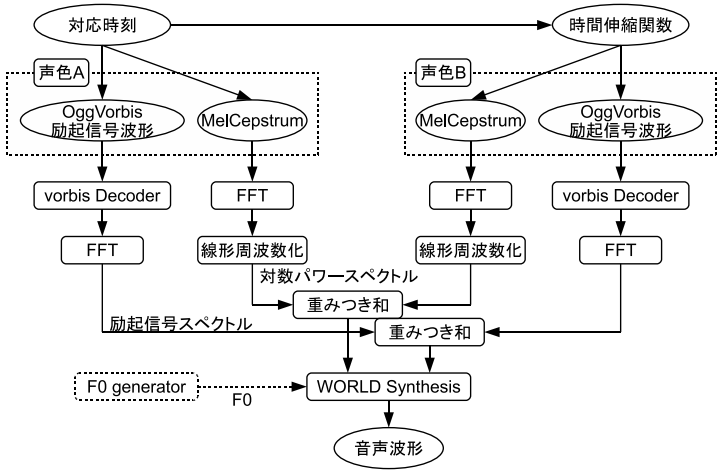

図 4 声色操作概要図
Fig. 4 Block diagram of v.Connect's synthesis.

として求める．次に $A \to B$ の時間伸縮関数を $T_{AB}(t)$ とし，その逆関数 $T_{AB}^{-1}(t)$ を用いて，A, B における音素片内での対応時刻を次式で近似的に求める．

$$\begin{cases} t'_A = & b(t)(t'+p_A) + (1-b(t))T_{AB}^{-1}(t'+p_B) \\ t'_B = & (1-b(t))(t'+p_B) + b(t)T_{AB}(t'+p_A) \end{cases} \tag{11}$$

ここで，p_A, p_B は音素片 A, B に対応する先行発音であり発音位置を補正している．得られた時刻 t'_A, t'_B におけるパワースペクトルをそれぞれ $S_A(\omega), S_B(\omega)$，励起信号波形を $r_A(\tau), r_B(\tau)$ とする．最後に合成に用いるパワースペクトル $S(\omega)$ と励起信号波形 $r(\tau)$ はそれぞれ，

$$\log S(\omega) = b(t)\log S_A + (1-b(t))\log S_B \tag{12}$$

$$r(\tau) = b(t)r_A(\tau) + (1-b(t))r_B(\tau) \tag{13}$$

と重みつき和により求める．モーフィングされたパワースペクトル，励起信号列 $S(\omega,t), r(\tau,t)$ を一つの音素片として扱う．

Vol.2012-MUS-94 No.10
Vol.2012-SLP-90 No.10
2012/2/3

表 1 波音リツコネクトの収録内容
Table 1 Recording environment and contents of Namine Ritsu Connect.

声色指定	強い	中間	弱い
収録単位	VCV 音素片		
収録語数	955 語		
マイク	Audio-Technica AT-4040		
Audio I/F	Roland UA-25EX		
収録場所	ソーゴ製業務用冷凍庫 SG 2000 RT		
メルケプストラムの次元数	32 次元		
OggVorbis	64kbits / 44100 samples		
収録ピッチ	B4(493.9Hz)	F4(349.2Hz)	

4.4 波形の再合成

前節で求めたモーフィング後の音素片は接続部分でも線形補間を用いて接続される．音素片の先行発音長 $l_{\mathrm{pre}} = \min(p_A, p_B)$ に対して，n 番目の音符に対する音素片の配置時刻 $t^{(n)}_{\mathrm{begin}} = t^{(n)} - l_{\mathrm{pre}}$ が，先行音の終了時刻 $t^{(n-1)} + l^{(n-1)}$ より前である場合重なっている部分を線形補間することで音素同士を接続する．これにより求めたパワースペクトルと励起信号スペクトル，および生成された F0 を WORLD の合成部に与え波形を再合成する．

5. システムの初期的検討

本システムの声色操作機能を検討するため，女性アマチュア歌手の発話を収録し合成を行った．実験用に作成した歌声コーパスは波音リツコネクト[⋆1]として公開されている．本章では収録内容，コーパスの容量，合成速度，出力の主観的評価について述べる．

5.1 収 録 内 容

波音リツコネクトは,「強い」「中間」「弱い」の三種類の声色を指定して録音したコーパスである．収録内容を**表 1** に示す．なお，励起信号波形を圧縮する際，OggVorbis は標本化周波数 44.1kHz の音声における 64kbps の設定を用いたため，44100 サンプルに対して 64kbits と表記している．

収録は 6〜7 モーラの無意味語約 150 語を 3 種類の声色を指定し，一定テンポ（2mora/sec），声色内で一定ピッチ，約 60 坪の業務用冷凍庫の中で行った．音声は標本化周波数 48kHz，量子化ビット数 24bit で収録したのち，標本化周波数 44.1kHz，量子化ビット数 16bit へダウンサンプリングを行った．収録した音声に対し手作業で先行発音長 250ms，音素長 500ms となるようアライメントをし事前分析を行った．マッチングにおける振幅包絡を決めるパラメタ $m = 1024$，枝の数を決めるパラメタ $N = 8$ とした．

⋆1 http://hal-the-cat.music.coocan.jp/ritsu.html

表 2 2ms あたりのデータ量の比較（単位は byte）
Table 2 Comparision of data amount per 2ms (in bytes).

	元データ	WORLD 分析後	v.Connect 分析後
波形	176.4	-	-
スペクトル	-	4096	128
励起信号	-	8192	約 200
合計	176.4	12288	約 328

表 3 実行時間比較
Table 3 Comparision of synthesis time with WORLD and v.Connect.

CPU	WORLD(sec)	提案手法 (sec)	スレッド数
Celeron 1.73Ghz	89.1	40.4	1
Core2Quad 2.80Ghz	39.6	20.7	1
Core2Quad 2.80Ghz	22.3	10.5	2
Corei7 3.50Ghz	22.9	13.1	1
Corei7 3.50Ghz	11.6	6.6	2

5.2 コーパスの容量

事前分析前の波形のコーパスでは約 210MB だったのに対し，事前分析後のコーパスでは約 430MB と約 2 倍におさえられた．WORLD による分析で得られるスペクトログラムと励起信号によるコーパスの容量は約 14.2GB であった．

2ms 分のデータを波形，WORLD の分析で得られる形式，v.Connect の分析で得られる形式で保持した場合のデータ量の比較を**表 2** に示す．ただし表の単位は byte であり，WORLD と v.Connect における分析シフト長は 2ms とした．

5.3 実 行 速 度

本システムを用い童謡「ふるさと」の一番，約 32 秒のシーケンスを作成し合成する時間を測定した．測定の結果を**表 3** に示す．

本システムのコーパス部分を波形に置き換え WORLD の分析を行いながら合成した場合と提案手法で合成時間を比較したところ，波形を使用した場合と比較して提案手法では 1.7 倍〜2.2 倍程度高速に合成された．

5.4 主観的評価

童謡「ふるさと」の一番に対して，すべての音符のポルタメント長を音符長の 20%，ポルタメントの深さを 8%としたデータに，BRI パラメタを「中間」から「弱い」の範囲で合成されるよう手作業により付加したデータ，常に「中間」の声が合成されるデータ，常に「弱い」の声が合成されるデータの 3 つを，圧縮したコーパスから合成したものと波形から直接合成したものの 2 つずつ計 6 個を作成した．

著者の主観によれば圧縮したコーパスから合成された音声と，波形から合成された音声との差は認められなかった．また，手書きした BRI パラメタに沿ってモーフィングを行われていることも確認ができた．しかしモーフィングを使用した音声では適切に操作されたフレーズについてはより自然に聞こえるが不自然に聞こえる箇所もあり，またモーフィングにより音声が劣化したように聞こえる箇所もあった．前者はモーフィング率の指定が手書きによるもので適切なモデルが無いこと，後者はモーフィングが線形補間によるものであることが原因として考えられるため，声色変化を扱うモデルとモーフィング法について検討する必要があるだろう．

6. おわりに

ユーザが声色操作を可能にする歌声合成器 v.Connect を提案し開発した．アマチュア女性歌手により収録された声色別の録音から，音声モーフィングによりユーザ指定のモーフィング率から指定の声色を合成するのに適したコーパスを構築した．歌声を合成する際の音声モーフィングによる容量，処理時間の増加に対処するため，コーパスには WORLD による分析データから低次メルケプストラムと OggVorbis による励起信号波形に変換したものを使用した．

圧縮されたコーパスは容量を波形の 2 倍に押さえながら，合成速度はシステムにより 1.7〜2.2 倍程度となった．この圧縮による劣化は主観的には感じられなかった．また，本システムから得られた合成音では，音声モーフィングを用いた声色変化により表現力の向上は認められるが，モーフィング率によっては不自然な音声が合成されることがある．これはモーフィングが単純な線形補間であることから，フォルマントの山や谷が潰れることによる影響と考えられる．

今後は音声モーフィングの精度の向上を目指すとともに，経時データによる声色操作だけでなくより手軽に表情付けが行える規則の作成が必要だろう．また，合成品質の評価も併せて行ってゆきたい．

謝辞 本システムの開発に御助力いただいた Cadencii プロジェクトの kbinani 様，コーパスの制作に御協力いただいた波音リツプロジェクトの皆様，WORLD 開発者の森勢様に感謝いたします．

参 考 文 献

1) 佐々木渉：仮想楽器をリアルにする「未来（ミク）の記号」と，VOCALOID で注目される「人の形」「声の形」について，音楽情報科学研究会， Vol.2008, No.50, pp.57–60 (2008).
2) 飴屋／菖蒲：歌声合成ツール ＵＴＡＵサポートページ，（オンライン），入手先⟨http://utau2008.web.fc2.com/⟩（参照 2011-11-28）.
3) 豊田健一，片寄晴弘，河原英紀：STRAIGHT による歌声モーフィングの初期的検討，音楽情報科学研究会，Vol.2006-MUS-64, No.19, pp.59–64 (2006).
4) 中野倫靖，後藤真孝：VocaListener2: ユーザ歌唱の音高と音量だけでなく声色変化も真似る歌声合成システムの提案，音楽情報科学研究会，Vol.2010-MUS-86, No.3, pp. 1–10 (2010).
5) 森勢将雅，西浦敬信，河原英紀：高品質音声分析変換合成システム WORLD の提案と基礎的評価〜基本周波数・スペクトル包絡制御が品質の知覚に与える影響〜，日本音響学会聴覚研究会，Vol.41, No.7, pp.555–560 (2011).
6) kbinani: cadencii_jp @ wiki - Cadencii, Cadencii Project (online), available from ⟨http://www9.atwiki.jp/boare/pages/18.html⟩ (accessed 2011-11-28).
7) XIPH.ORG: Vorbis.com, XIPH.ORG (online), available from ⟨http://www.vorbis.com/⟩ (accessed 2011-11-28).
8) 森勢将雅，西浦敬信，河原英紀：基本波検出に基づく高 SNR の音声を対象とした高速な F0 推定法，電子情報通信学会論文誌 D， Vol.93-D, No.2, pp.109–117 (2010).
9) 森勢将雅，西浦敬信，河原英紀：高品質音声合成を目的とした母音の高速スペクトル包絡推定法，電子情報通信学会論文誌 (D)， Vol.94-D, No.7, pp.1079–1087 (2011).
10) 森勢将雅，松原貴司，中野皓太，西浦敬信：高品質音声合成を目的とした励起信号抽出法の検討，日本音響学会 2011 年春季研究発表会，pp.325–326 (2011).
11) Lai, W.-H.: F0 Control Model for Mandarin Singing Voice Synthesis, *Second International Conference on Digital Telecommunications*, Vol.ICDT2, p.12 (2007).