Vol.2010-MUS-86 No.8
2010/7/28

SpeakBySinging: 歌声を話声に変換する話声合成システム

阿曽慎平[†1] 齋藤毅[†2,*1] 後藤真孝[†2]
糸山克寿[†1] 高橋徹[†1] 駒谷和範[†1,*2]
尾形哲也[†1] 奥乃博[†1]

本報告では歌声を話声に変換する話声合成システム SpeakBySinging について述べる．システムの入力は無伴奏単独歌唱とその歌詞，出力は歌詞の朗読音声である．この変換は歌声の音韻長，F0，パワーを操作することで実現される．声質を保つためにスペクトル包絡は維持する．合成された音声を心理実験で評価した結果声質が概ね保たれていることを確認した．

SpeakBySinging: A Speaking Voice Synthesis System Converting Singing Voices to Speaking Voices

Shimpei Aso,[†1] Takeshi Saitou,[†2,*1]
Masataka Goto,[†2] Katsutoshi Itoyama,[†1]
Toru Takahashi,[†1] Kazunori Komatani,[†1,*2]
Tetsuya Ogata[†1] and Hiroshi G. Okuno[†1]

This report describes a singing-to-speaking synthesis system called “SpeakBySinging” that can synthesize a speaking voice from an input singing voice and the song lyrics. The system controls the phoneme duration, fundamental frequency (F0), and power (volume). To retain the timbre of the input singing voice, the system does not control the spectral envelope. Experimental results show that SpeakBySinging can convert singing voices into speaking voices whose timbre is almost the same as the original singing voices.

1. はじめに

近年，初音ミクなどの歌声作成ソフト VOCALOID[1] が普及し，ユーザ自身の手でメディアコンテンツを制作する CGM (Consumer Generated Media) が注目を集めている．それとともに，より魅力的な音声の合成技術が求められるようになってきた．これまでの研究では，音声を素材として，話声を合成する技術（ボコーダ，波形合成[2),3)]，隠れマルコフモデル（HMM）[4)–6)] 等）や話声から歌声を合成する SingBySpeaking[7)]，さらには，歌声から歌声を合成する VocaListener[8)] が開発されてきたものの，逆に歌声から話声を合成する技術については，ほとんど研究がされてこなかった．

歌声から話声を合成する SpeakBySinging は，日頃よく経験する歌手の歌声と話声の音色の大きな違いを軽減し，歌声の音色をできるだけ保持したままで表情豊かな話声を合成することを目的としている．さらに，話声と歌声間の 4 通りの変換手法の研究を通じて，話声と歌声との本質的な違いの解明にも貢献したいと考えている．最初に，SpeakBySinging は入力した歌声から基本周波数 (F0) 時系列，音素の継続時間長（音韻長），パワー時系列の 3 つの音響特徴を抽出する．次に，テキスト読み上げシステム (TTS) に歌詞を入力することで，話声らしい F0 時系列，音韻長，パワー時系列（ターゲット[⋆1]の音響特徴）を得る．最後に，入力歌声の音響特徴をターゲットの音響特徴に近づくよう操作し，音響信号を再合成することで話声を出力する．本報告の構成は以下の通りである．第 2 章で歌声と話声の声質の違いについて考察する．第 3 章で，歌声から話声を合成する SpeakBySinging について，その課題と解決策について述べる．なお，以前の報告[9)] では，歌声と同じ歌詞を朗読した実話声をアライメントのために用いていたが，本報告では，そのようなデータは不要である．第 4 章で，評価実験を行い，第 5 章で本論文をまとめる．

†1 京都大学 大学院情報学研究科
Graduate School of Informatics, Kyoto University
†2 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)
∗1 現在，金沢大学 理工学域 電子情報学類
Presently with School of Electrical and Computer Engineering, College of Science and Engineering, Kanazawa University
∗2 現在，名古屋大学 大学院工学研究科
Presently with Graduate School of Engineering, Nagoya University
⋆1 話声らしい音響特徴を持つ，仮想的な話声をターゲットと呼ぶ．

Vol.2010-MUS-86 No.8
2010/7/28

2. 歌声と話声

本章では歌声と話声の音響的な違いについて述べる．

2.1 音響特徴量の違い

従来研究[10),11)]では歌声と話声を区別するために音韻長，F0 時系列，パワー時系列に着目してきた．これら特徴の比較例を**図 1** に示す．

音韻長の違い

歌声では，楽譜上の音価（全音符や八分音符など）や演奏表現（スタッカートやレガートなど）に依存し，個々の音節の長さが大きく異なる．特に音節の時間長を伸ばす時には母音部分が子音部分より長く持続する．これは，実音声（歌声，話声問わず）で一般に音韻長を変化させる場合には，母音部分を伸縮することが多いためである．

歌声に対して，話声は各音節ごとの時間長に大きな違いは見られず，一般に一定のテンポで発音される[12)]．

F0 時系列の違い

歌声は楽曲の楽譜音高情報に依存し，ほぼ一定の F0 を保つ定常部と，1 つの定常部から次の定常部へと移る遷移部からなり，図 1 に示すような階段構造を持つ[13)]．定常部ではヴィブラートに代表される動的変動成分が含まれる場合がある．平均的に歌声の F0 は話声よりも高い傾向がある．

パワー時系列の違い

歌声は F0 遷移と同期し変動の少ない定常部を持つ．対して話声は定常部をほとんど持たず，常に大きく変動する．

2.2 話声から歌声を合成するシステム

齋藤らは，F0 時系列，スペクトル（包絡や時間変動），音韻長を制御して話声を歌声の変換するシステム SingBySpeaking を開発した[7)]．システムの入力を以下に述べる．

- 合成したい歌の歌詞朗読音声（話声）
- 歌の譜面情報（メロディ遷移の概形）
- 朗読音声中の音韻（または単語）と譜面中の音符の対応関係を記述した情報（音韻と音符の同期情報）

このシステムは音韻長，F0 時系列，スペクトルを制御する 3 つのモデルから構成される．音韻長制御モデルは楽曲のテンポに応じて各音韻長を伸縮する．F0 制御モデルは楽譜から生成したメロディの概形にヴィブラート等の歌声特有の時間変動成分を付与する[14)]．スペクトル制御モデルは，“singer’s formant”[15)] と呼ばれる周波数軸上 3kHz 付近に現れるフォルマントピークを付与し，ヴィブラートに同期してパワー変調を付与する．このシステムで合成された歌声を聴取実験により評価した結果，実際の歌声と同程度の音質で，かつ話声の声質を崩すことなく自然な歌声を合成可能であることを確認している．

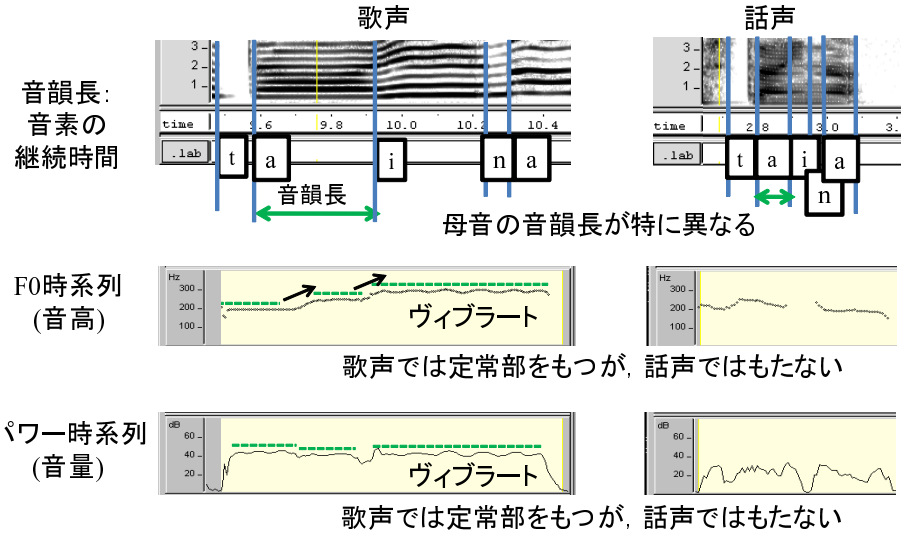

図 1 歌声と話声とで音韻長，F0 時系列，パワー時系列を比較．発話内容は「たいな」（歌詞の一部）である．

3. SpeakBySinging

本章では我々が開発した歌声を話声に変換する話声合成システム，SpeakBySinging について述べる．本システムは SingBySpeaking[7)] を受けて，音韻長，F0，パワーに着目し，テキスト読み上げを用いた処理を行う．本システムの入出力は次のとおりである．

入力 単独歌唱音声（歌声）とその歌詞

出力 合成された歌詞朗読音声（話声）

この音声変換は歌声の音韻長，F0 時系列，パワー時系列の 3 つの音響特徴量を話声らしいもの（ターゲット）に変換することで実現できる．ターゲットとなる特徴量は TTS に歌詞を入力して得る．これら 3 つの音響特徴量は 2 章で述べたように歌声と話声の違いを決定す

る主要な要素である．本システムでは音声分析合成システム STRAIGHT[16] と歌詞（音素列）と歌声のアライメントを行うビタビアライメント[17),18)] を用いてこれら 3 つの音響特徴量を抽出する．歌声の声質を保つために，2.2 節で述べたスペクトル包絡に含まれる singer's formant は保存する．

SpeakBySinging は STRAIGHT の分析・合成過程に音韻長，F0 時系列，パワー時系列を制御する 3 つのモジュールと TTS を組み込んだ構成となっている．STRAIGHT 分析では F0 時系列と，F0 による変動が除去されたスペクトル包絡時系列が抽出できる．また，これらのデータを加工した後に音声を合成可能である．これにより声質を保ったまま F0 の操作が可能となる．**図 2** に本システムの概要を示す．変換の流れはまず，前処理として入力歌声から音韻長，F0 時系列，パワー時系列を抽出する．入力歌声を STRAIGHT 分析して F0 時系列，スペクトル包絡時系列，非周期性指標系列を得る．ここで得られるデータについて，F0 時系列は声帯の振動に該当する．有声区間とは，声帯の振動伴う発声の時間上の区間であり，F0 時系列上では正の値をもつ区間として表現される．無声区間とは，声帯の振動を伴わない発声の区間であり，0 の値をもつ区間として表現される．スペクトル包絡時系列とは声道特性に該当する．また，パワー時系列はスペクトル包絡時系列から算出できる．非周期性指標は非周期成分の強さを表す．また，音韻長を抽出するために入力歌声に対して歌詞（音素列）のアライメントを行う．次に，以下の処理を行う．各項目番号は図 2 に記された番号と対応する．

(1) TTS に歌詞を入力してターゲットとなる音韻長，F0 時系列，パワー時系列を得る．
(2) 音素アライメントで得た音韻長とターゲットの音韻長との比率に基づいて歌声の F0 時系列，非周期性系列，スペクトル包絡時系列の時間伸縮を行う．
(3) (1) で得られたターゲットの F0 時系列に対し，有声/無声区間（F0 の存在区間）を調整する．また，平均 F0 も操作する．
(4) スペクトル包絡時系列のパワーを，ターゲットのパワー時系列に合わせて調整する．

以上の処理で得られる (3) の F0 時系列，(4) のスペクトル包絡時系列，(2) の非周期性指標系列を用いて STRAIGHT 合成を行い，話声を得る．

3.1 TTS を用いたターゲットの音響特徴量生成

TTS を用いてターゲットとなる話声らしい音韻長・F0 時系列・パワー時系列を生成する．これは歌声を話声に変換するには歌声の音韻長・F0 時系列・パワー時系列を話声のものに

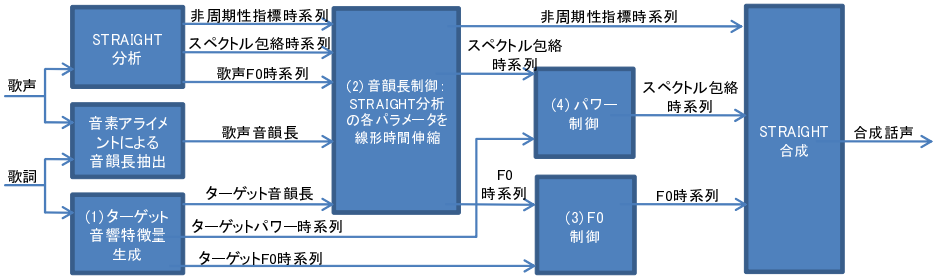

図 2 SpeakBySinging のブロックダイアグラム．

変える必要があるからである．本報告では TTS として OpenJtalk[⋆1] を利用する．音声信号の合成・STRAIGHT 再分析に伴うノイズ発生を避けて，ターゲットの音韻長・F0 時系列・パワー時系列を直接得られるためである．

3.2 音韻長制御モジュール

ここでは STRAIGHT 分析により得られた F0 時系列，スペクトル包絡時系列，非周期性指標系列をターゲットの音韻長にあわせて音素ごとに時間方向に伸縮する．このモジュールの入出力は次のとおりである．

入力 F0 時系列，スペクトル包絡時系列，非周期性指標系列，音韻長制御前の音韻長，ターゲットの音韻長．

出力 ターゲットの音韻長をもつ F0 時系列，スペクトル包絡系列，非周期性指標系列．

まず，各音韻長の伸縮率を求める．n 番目の音素について，歌声の音韻長を $D_{\mathrm{sing}}(n)$，ターゲットの音韻長を $D_{\mathrm{target}}(n)$ とおくと，伸縮率 $S(n)$ は

$$S(n) = \frac{D_{\mathrm{target}}(n)}{D_{\mathrm{sing}}(n)} \tag{1}$$

として求まる．この伸縮率に基づき，F0 時系列，スペクトル包絡時系列，非周期性指標系列を区分線形で時間伸縮を行う．ここで，音素境界前後における子音部 10ms と母音部 30ms の区間は伸縮させず，それ以外の区間を線形補完伸縮させる．これは，音素間の遷移時間は歌声と話声で大きく変化しないと考えられるためであり，全区間を均一に伸縮することによって合成音声の音質が低下することを避ける[7)]．また，スペクトル包絡時系列の母音区間

⋆1 http://open-jtalk.sourceforge.net/

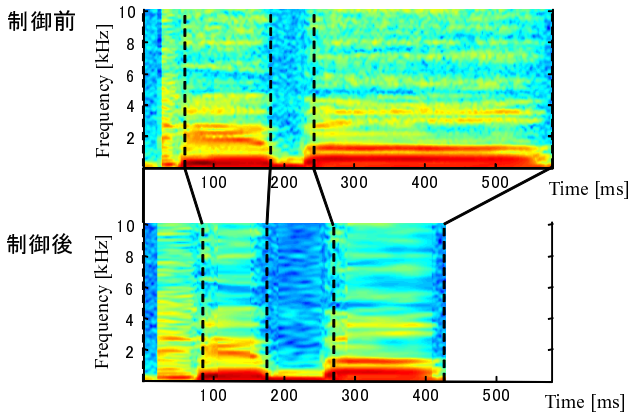

図 3 音韻長制御前後のスペクトル包絡時系列．ある音素区間のみ表示しており上が制御前で下が制御後である．

にはヴィブラート等の動的な時間変動の影響が含まれており，時間伸縮した後にもそれらの変動が残っており，高周波化している可能性が高い．そこで，母音区間の時間方向に対してローパスフィルタ（カットオフ周波数：20 Hz）をかけることで，歌声固有のスペクトルの時間変動成分を除去する．ローパスフィルタの適応範囲も境界前後における子音部 10ms と母音部 30ms を除く．音韻長制御モジュールでスペクトル包絡時系列を伸縮させ，動的変動成分を除去した例を**図 3** に示す．

3.3 F0 制御モジュール

F0 制御モジュールでは，音韻長を制御された F0 時系列とターゲットの F0 時系列から合成に用いる F0 時系列を生成する．このモジュールの入出力は次のとおりである．

入力 音韻長を制御された F0 時系列，ターゲットの F0 時系列．（音韻長制御により，この２つの F0 時系列は時間長が同じで同一時刻で音素が対応していることに注意する．）

出力 F0 時系列．（概形はターゲットの F0 時系列であり，有声/無声区間は音韻長を制御された F0 時系列と一致し．平均 F0 が操作されたもの）

このモジュールは，有声/無声区間の操作と平均 F0 の操作を行う．

有声/無声の操作

我々は，ターゲットの F0 時系列の有声/無声区間を修正してから合成用のパラメータとする．なぜなら，我々は音韻長を操作した F0 時系列をターゲット F0 時系列に置き換えたいのだが，後に行う STRAIGHT 合成では，各時刻での声帯振動の有無（有声/無声）に応じて合成方法を切り替えるため，F0 時系列の置換前後で有声/無声区間が保たれていないと，スペクトル包絡時系列との不一致が起こり合成後の音声が不自然になる．この不一致を避けるため次の操作を行う．

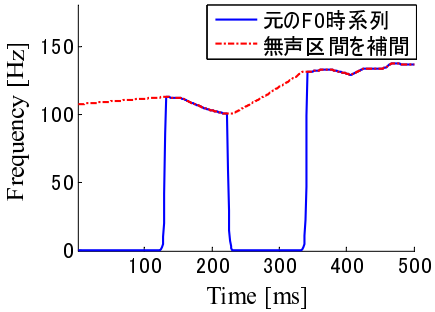

図 4 線形補完により，F0 時系列の無声区間を有声区間にし，仮想的に全区間を有声にする．

（1） 音韻長を制御された F0 時系列から各時間が有声であるか無声であるか（有声ならば 1，無声であれば 0）の時系列データを抽出する．

（2） **図 4** のようにターゲットの F0 時系列の無声区間部分を線形補完し，仮想的に全時刻で有声となる F0 時系列を得る．

（3） 2 で得られた F0 時系列に対し（1）で得られた時系列データをかけて有声/無声区間が音韻長を制御したものであり，概形がターゲットのものとなる F0 時系列を得る．

周波数方向への平行移動による平均 F0 操作

次に，有声/無声区間を操作した F0 時系列を周波数方向に移動する．これは，有声/無声区間を操作した F0 時系列が元の歌声 F0 時系列と比べ低い値を持つことがあり，合成される音声が不自然に感じられるためである．不自然になる原因は F0 とスペクトル包絡とで不一致が起きるためであると考えられる．どの程度平行移動を行うかは後述の評価実感を通して決定する以上から次のように平均 F0 を操作する．有声/無声区間を操作した F0 時系列 $F_{\mathrm{vuv}}(t)$ の時間平均 M_{vuv} と，音韻長制御モジュールで伸縮した有声区間のみの F0 時系列 $F_{\mathrm{vuv}}(t)$ の時間平均 M_{sing} を求める．出力される F0 時系列 F_{output} は次式で表される．

$$F_{\mathrm{output}}(n) = F_{\mathrm{vuv}}(n) \times \left(1 + \alpha \left(\frac{M_{\mathrm{sing}}}{M_{\mathrm{vuv}}} - 1\right)\right) \tag{2}$$

ここで α は平均 F0 を調整するパラメータである．$\alpha = 0$ ならば F0 時系列に変化はなく，

Vol.2010-MUS-86 No.8
2010/7/28

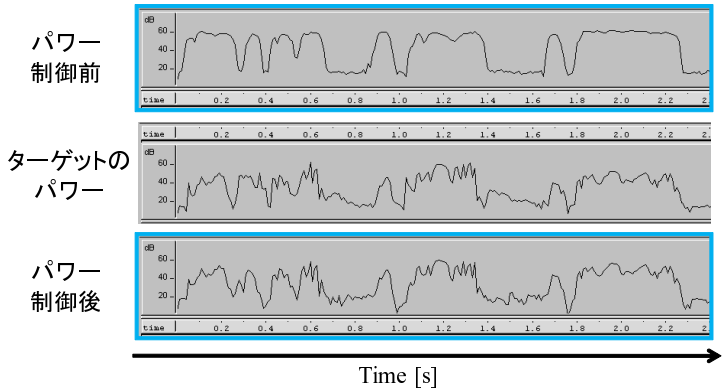

図 5 パワー制御の入出力例．上図が入力スペクトル包絡時系列から抽出したパワー時系列，中図がターゲットのもの，下図が出力スペクトル包絡時系列のもの．

$\alpha = 1$ ならば平均 F0 が音韻長を制御されたものと一致する．

3.4 パワー制御

パワー制御モジュールでは，3.2 節で音韻長を制御されたスペクトル包絡時系列のパワー時系列を各時刻で増減してターゲットのパワー時系列に変える．このモジュールの入出力は次のとおりである

入力 スペクトル包絡時系列，ターゲットのパワー時系列．

出力 パワー時系列をターゲットのものに修正されたスペクトル包絡時系列

入出力の例を**図 5** に示す． まず，スペクトル包絡のパワー時系列 P_{se} を以下のように定義する．

$$P_{\mathrm{se}}(t) = \sum_{f=1}^{F} (N_{\mathrm{se}}(f,t))^2 \tag{3}$$

ここで，N_{se} は周波数帯域数 G 時間フレームサイズ T の $G{\times}T$ 行列でありスペクトル包絡時系列を表す，ターゲットのパワー時系列を P_{tg} とおくと，各時刻のパワーの比率は，

$$\mathrm{Ratio}(t) = 10\log_{10}\frac{P_{\mathrm{t}}(t)}{P_{\mathrm{se}}(t)}\ [\mathrm{dB}] \tag{4}$$

として求まる．次に，得られた Ratio に対し，**図 6** のような非線形な変換を行う．これは，Ratio でそのままスペクトル包絡時系列を増減すると，増幅率が高くなってしまう場合（例えば +15dB）に，音声録音時の雑音成分や子音の非周期成分が過剰に増幅されて不自然な音声が合成されてしまう．これを防ぐため，以下のように Rate を調整して増幅率とする．

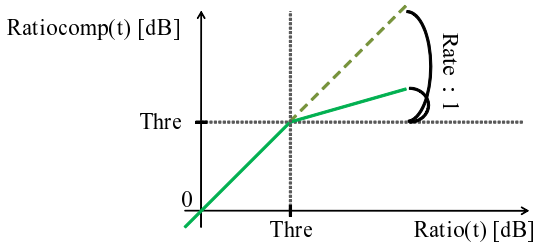

図 6 Ratio と Ratiocomp の関係．

$$\mathrm{Ratiocomp}(t) = \begin{cases} \mathrm{Ratio}(t) & (\mathrm{Ratio}(t) \le \mathrm{Thre}) \\ \mathrm{Thre} + \frac{\mathrm{Ratio}(t)-\mathrm{Thre}}{\mathrm{Rate}} & (\mathrm{Ratio}(t) > \mathrm{Thre}) \end{cases} \tag{5}$$

ここで Thre，Rate はそれぞれ定数で，Thre は圧縮をかける閾値，Ratio は圧縮率を表す定数である．パワー制御モジュールの出力 N_o は次式で得られる．

$$N_o(f,t) = N_{\mathrm{se}}(f,t){\times}10^{\frac{\mathrm{Ratiocomp}(t)}{20}} \tag{6}$$

4. 評価実験

本システムで合成される音声を二つの被験者実験で評価する．一つめは，3.3 章で述べた平均 F0 の操作パラメーター α が合成音の自然さに与える影響を評価する．二つめは，合成された話声が歌声の声質を保っているか評価する．これらの評価では，ビタビアライメントのアライメント誤差による影響を避けるため，音素をまたぐような誤りは手動で修正する．また，TTS の音響モデルの違いが評価に影響を与えないよう，入力の歌声が女声の場合でも単一の男性音響モデルを用いる．

4.1 平均 F0 操作の評価

この実験では平均 F0 だけが異なる音声を複数用意し，各音声の自然さを Mean Opinion Score（MOS）により評価する．評価する合成話声を作るため，研究用音楽データベース *AIST Humming Database*[19] を用いる．このデータベースから *P078_DK* の一部分（約 12 秒）を歌っている男声 5 名，女声 5 名の単独歌唱音声を本システム入力用歌声とする．一

表 1 *Opinion Score* の評価カテゴリと対応する点数.

評価カテゴリ	評価点
とても自然	5
自然	4
普通	3
不自然	2
とても不自然	1

表 2 各 α 値, 歌手の性別ごとにグループ化した *MOS* 評価値. $\alpha = 0.0$ は *F0* 平均移動をしないことを意味する.

α	女声	男声	両方
0.0	2.10	**2.65**	2.38
0.2	**2.43**	2.60	**2.43**
0.4	2.31	2.33	2.31
0.6	2.29	2.40	2.29
0.8	2.23	2.30	2.22
1.0	2.18	2.20	2.18

表 3 被験者が作った対の正解率を音声の性別ごと，または両方まとめてグループ化し平均値を表示したもの.

女声	男声	両方
85.0%	72.5%	78.8%

つの歌声に対し，3.3 で述べた α を 0, 0.2, 0.4, 0.6, 0.8, 1.0 の 6 段階で変えながら音声を合成する．結果，10 名分 $\times$ 6 段階 $=$ 60 個の評価用音声が合成される．各被験者は各評価用音声をヘッドホン（SONY MDR-CD900ST）を介して聴取し，音声の自然さを**表 1** のように 5 段階で評価する．全評価結果を集計し，歌手の性別ごと・各 α 値ごとに平均を算出した結果を**表 2** に示す.

4.2 声質が保存されているかの評価

この実験では，被験者に複数の歌声と合成話声を聞いてもらい，声質が最も近いと感じられるもので対を作ってもらう．被験者の平均正解率を求めることで声質が保存されているかどうかの評価とする．ここで使う話声は 4.1 と同じもので，男声 5 名，女声 5 名の単独歌唱音声である．これらの 10 個の歌声から**図 7** のように本システムを用いて 10 個の話声を合成する．この実験では，F0 平均を操作するパラメータ α を，$\alpha = 0.6$ に設定した．これは，TTS の F0 平均が元の歌声に比べ低すぎるのを補正するためである．また，STRAIGHT の分析・合成過程による音質変化が評価に影響を与えないようにするため，歌声も STRAIGHT で分析し，(何も操作を加えず）合成する．被験者はまず女声の合成話声と合成歌声を聞き，**図 8** のように声質が最も近いもの同士で対を作る．次に男声の音声に対しても同様に行う．ここで被験者に対し，どの話声がどの歌声をもとに合成されたかについての情報は与えない．評価の結果，全被験者の平均正解率を音声の性別ごとにグループ化して算出した結果を**表 3** に示す.

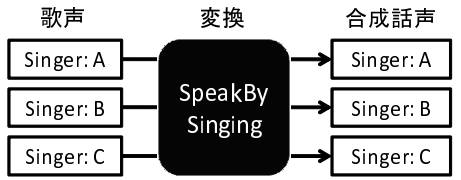

図 7 評価用の音声として，複数の歌声を話声に変換する.

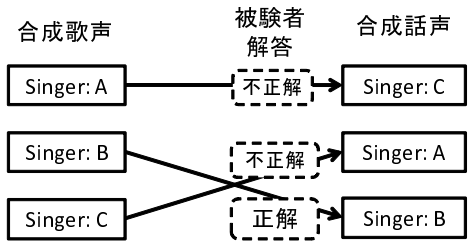

図 8 被験者は歌声と話声間で声質が最も近いと思われるもの同士の対を作る.

4.3 結果の考察

4.1 節の結果表 2 から，女声に対して F0 平均操作を行うことで合成話声がより自然になることが示された．この理由の一つに，合成に用いた音響モデルが男性のものだったことが挙げられる．また，男声に対しては平均 F0 を操作しない時が一番良い結果となった．これは，男声の歌声と TTS の出力とで平均 F0 が異なっていても TTS の出力をそのまま使うほうが良いことを示している。

4.2 節の結果表 3 から，本システムは歌声の声質を保ったまま話声に変換できることを示している．また，女声の平均正解率が男声のものより高いことから，女声に対してより良い結果が得られることを示唆している.

我々は別のフィードバックを得るため，実験後，被験者に意見を記述してもらった．数名の被験者から，音のつなぎ目が不自然であるとの意見が得られた．これは，合成話声の音素遷移部分に問題があり，高度な音韻長制御処理の必要性を示唆している．例えば Dynamic Time Warping[20] に基づいて歌声と TTS から生成されたターゲットの音声とでオーディオアライメントを行い，アライメント結果に基づいて時間伸縮を行うことが挙げられる.

5. 結　　論

我々は歌声を話声に変換する話声合成システム SpeakBySinging を開発した．本システムは F0 時系列，音韻長，パワー時系列を制御するモジュールから構成される．評価の結果，本システムは歌声の声質を保ったまま話声に変換できることを確認した．

今後の課題には，音韻長制御の改善による音質向上を目指すことと，歌声から話声へのモーフィングができるようなシステムへと拡張することが挙げられる．また，ターゲットの特徴量を TTS から生成したものを使う場合と，実話声から抽出したものを使う場合とで合成される話声を比較してみることも重要である．

謝辞　産業技術総合研究所の中野倫靖氏には有声/無声区間処理について助言を頂いた．産業技術総合研究所の藤原弘将氏にはビタビアライメントについて助言を頂いた．本研究の一部は，科研費，GCOE，CrestMuse の支援を受けた．

参 考 文 献

1) 剣持秀紀，大下隼人：歌声合成システム VOCALOID，情報処理学会研究報告，Vol.2007, No.102, pp.25–28 (2007).
2) Black, A. W. and Campbell, N.: Optimising Selection Of Units From Speech Databases For Concatenative Synthesis, *Proc. Eurospeech*, pp.581–584 (1995).
3) Schwarz, D.: A system for data-driven concatenative sound synthesis, *Proc. Digital Audio Effects*, pp.97–102 (2000).
4) 吉村貴克，徳田恵一，益子貴史，小林隆夫，北村　正：HMM に基づく音声合成におけるスペクトル・ピッチ・継続長の同時モデル化，電子情報通信学会論文誌 (D-II), Vol.J383-D-II, No.11, pp.2099–2107 (2000-11-25).
5) Saino, K., Zen, H., Nankaku, Y., Lee, A. and Tokuda, K.: An HMM-Based Singing Voice Synthesis System, *Proc. International Conference on Spoken Language Processing*, pp.1141–1144 (2006).
6) Zen, H., Tokuda, K., Masuko, T., Kobayashi, T. and Kitamura, T.: A Hidden Semi-Markov Model-Based Speech Synthesis System, *IEICE Transactions on Information and Systems*, Vol.E90-D, No.5, pp.825–834 (2007).
7) 齋藤　毅，後藤真孝，鵜木祐史，赤木正人：SingBySpeaking：歌声知覚に重要な音響特徴を制御して話声を歌声に変換するシステム，情報処理学会研究報告， Vol.2008, No.12, pp.25–32 (2008).
8) 中野倫靖，後藤真孝：VocaListener: ユーザ歌唱を真似る歌声合成パラメータを自動推定するシステムの提案，情報処理学会研究報告， Vol.2008, No.50, pp.49–56 (2008).
9) 阿曽慎平，齋藤　毅，後藤真孝，糸山克寿，高橋　徹，駒谷和範，尾形哲也，奥乃　博：F0・振幅・音韻長の制御により歌声を話声に変換する話声合成システム SpeakBySinging，情報処理学会 第 72 回全国大会, 6U-1 (2010).
10) Ohishi, Y., Goto, M., Itou, K. and Takeda, K.: Discrimination between Singing and Speaking Voices, *Proc. Eurospeech*, pp.1141–1144 (2005).
11) Sundberg, J.: *The Science of the Singing Voice*, Northern Illinois University Press (1987).
12) 阿部匡伸：発話様式のバリエーション，日本音響学会誌， Vol.51, No.11, pp.882–886 (1995).
13) 大石康智，後藤真孝，伊藤克亘，武田一哉：スペクトル包絡と基本周波数の時間変化を利用した歌声と朗読音声の識別，情報処理学会論文誌， Vol.47, No.6, pp.1822–1830 (2006).
14) Saitou, T.: Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis, *Speech Communication*, Vol.5, No.3-4, pp.267–277 (2005).
15) Sundberg, J.: Articulatory interpretation of the “singing formant”, *The Journal of the Acoustical Society of America*, Vol.55, pp.838–844 (1974).
16) 河原英紀：聴覚の情景分析が生み出した高品質 VOCODER:STRAIGHT，日本音響学会誌， Vol.54, No.7, pp.521–526 (1998).
17) Fujihara, H., Goto, M., Ogata, J., Komatani, K., Ogata, T. and Okuno, H.G.: Automatic synchronization between lyrics and music CD recordings based on Viterbi alignment of segregated vocal signals, *Proc. IEEE International Symposium on Multimedia*, pp.257–264 (2006).
18) Odell, J., Ollason, D., Woodland, P., Young, S. and Jansen, J.: *The HTK Book for HTK V2.0*, Cambridge University Press, Cambridge, UK (1995).
19) 後藤真孝，西村拓一：AIST ハミングデータベース: 歌声研究用音楽データベース，情報処理学会研究報告， Vol.2005, No.82, pp.7–12 (2005).
20) Keogh, E. and Ratanamahatana, C.A.: Exact indexing of dynamic time warping, *Knowledge and Information Systems*, Vol.7, No.3, pp.358–386 (2005).