SingBySpeaking：話声を歌声に変換する歌声合成システム*

◎齋藤毅, 後藤真孝 (産総研), 鵜木祐史, 赤木正人 (北陸先端大)

1　はじめに

歌声合成システムの構築は，計算機による音楽の新たな楽しみ方を創造するだけでなく，人間の音声コミュニケーションを理解する上でも重要な取り組みである．現在の歌声合成の研究は，テキスト又は歌詞から歌声を合成する *text-to-singing synthesis* のアプローチによる取り組みが主流である [1, 2, 3, 4]．これらは，話声を対象とした *text-to-speech synthesis* で用いられる波形接続合成や HMM 合成といったコーパスベースの手法に基づいたものが多く，特に YAMAHA の VOCALOID[2] は市販の歌声合成ソフトウェアとして計算機音楽の新しい可能性を示している．

それに対して，本稿では，話声から歌声を合成する *speech-to-singing synthesis*[5, 6] という新しいアプローチで歌声合成システムを構築することを目的とする．このシステムは，音声分析系 STRAIGHT[7] の処理過程において歌声特有の音響特徴を制御・付与することで，歌詞の朗読音声を歌声に変換するものである．我々は，この方法を用いることで，従来の歌声合成システム以上の自然な歌声の合成と，「歌詞を朗読さえすれば元の声質を保持した歌声を生成できる」という新たな歌声アプリケーションが実現できると考えている．更には，歌声特有の各種音響特徴を操作・変換した歌声を合成できるシステムの枠組み自体が，歌声の知覚・生成機構を解明する有効な手法になり得ると考えている．

2　歌声合成システムの処理体系

図 1 に歌声合成システム SingBySpeaking の概要を示す．このシステムは，STRAIGHT の分析・合成処理体系に，F0，スペクトル，音韻長の各音響パラメータにおける歌声特有の音響特徴を制御するモデルを組み込むことで構成される．

システムの入力は，合成したい歌の歌詞の朗読音声（speaking voice），その歌の譜面情報（musical score），そして朗読音声中の音韻（または単語）と譜面中の音符の対応関係を記述した情報（synchronization information）の 3 つである．尚，朗読音声のセグメンテーションと，セグメンテーションされた各音韻と音符との対応付けは手動で行う必要がある．このシステムでは，以下の 6 つの手続きによって歌声合成音が生成される．

1. STRAIGHT によって朗読音声の音響パラメータ（F0 変化パターン，スペクトル系列，非周期性指標系列）を抽出する
2. 歌声 F0 制御モデルによって譜面情報から歌声の F0 変化パターンを生成する
3. 音韻長制御モデルによって朗読音声中の各音韻のスペクトルと非周期性指標の時間系列を伸長する
4. スペクトル制御モデル 1 によって時間伸長後の母音区間のスペクトル包絡と非周期性指標を加工する
5. STRAIGHT によって生成・加工した各音響パラメータから歌声を合成する
6. スペクトル制御モデル 2 によって合成歌声の振幅エンベロープを加工する

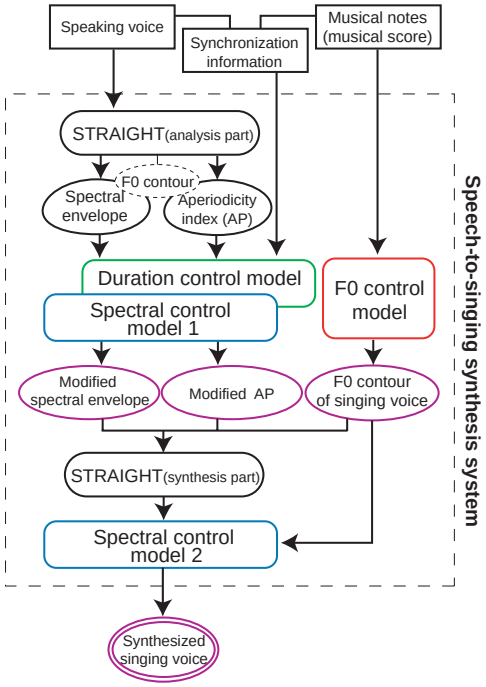

Fig. 1　*Block diagram of our speech-to-singing synthesis system "SingBySpeaking".*

3　F0 制御モデル

本章では，歌声の F0 変化特有の音響特徴である F0 動的変動成分について述べ，その特徴を制御可能な歌声 F0 制御モデルについて概説する．

3.1　歌声特有の F0 動的変動成分

筆者らの先行研究 [8] において様々な歌声データの F0 を分析した結果，以下の 4 種の F0 動的変動成分が歌唱スタイルや歌唱者に関係なく存在することが明らかとなっている．

オーバーシュート (Overshoot)：滑らかな音高の変化，およびその直後に目的音高を越える瞬時的な変動成分

ヴィブラート (Vibrato)：同一音高区間で観測される 4~8 Hz の準周期的な変動成分

プレパレーション (Preparation)：音高変化直前に変化とは逆方向に振れる瞬時的な変動成分

*SingBySpeaking: A Singing Voice Synthesis System Converting Speaking Voices to Singing Voices. by SAITOU, Takeshi, GOTO, Masataka (AIST), UNOKI, Masashi and AKAGI, Masato (JAIST)

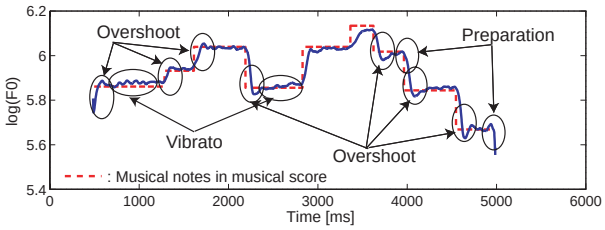

Fig. 2 *Examples of F0 fluctuations in the singing voice of an amateur singer.*

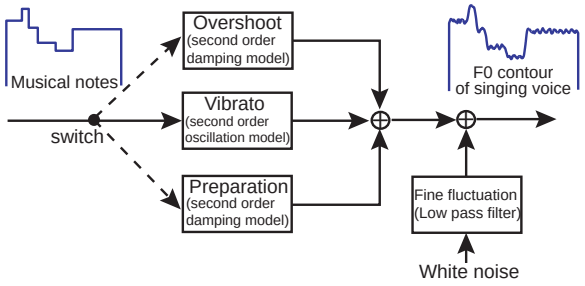

Fig. 3 *Block diagram of the F0 control model for singing voices.*

微細変動 (Fine fluctuation) ：発声区間全体に観測される不規則で細かい変動成分

図 2 に，アマチュア歌手による日本童謡「七つの子」歌唱時の F0 変化パターンと，そこに含まれる F0 動的変動成分を示す．

3.2 F0 制御モデル

図 3 に歌声の F0 制御モデルの概要を示す．このモデルは，譜面中の各音符をステップ関数で記述し，それらを重ね合わせることで生成したメロディの遷移の概形（Melody contour）に対して，4 種の F0 動的変動成分を制御・付与することで歌声の F0 変化パターンを生成する．

オーバーシュート，ヴィブラート，プレパレーションは，メロディの遷移の概形を複数のフィルタに通すことで制御される．フィルタは，次式の二次系伝達関数のインパルス応答で与えられる．

$$H(s) = \frac{k}{s^2 + 2\zeta\omega s + \omega^2}, \qquad (1)$$

ここで，ω は固有角周波数，ζ は減衰項，k は振幅項である．インパルス応答 $h(t)$ は，ζ の値に従って以下のように与えられる．

$$h(t) = \begin{cases} \frac{k}{2\sqrt{\zeta^2-1}}(\exp(\lambda_1\omega t) - \exp(\lambda_2\omega t)), & |\zeta| > 1 \\ \frac{k}{\sqrt{1-\zeta^2}}\exp(-\zeta\omega t)\sin(\sqrt{1-\zeta^2}\omega t), & 0 < |\zeta| < 1 \\ kt\exp(-\omega t), & |\zeta| = 1 \\ \frac{k}{\omega}\sin(\omega t), & |\zeta| = 0 \end{cases} \qquad (2)$$

ここで，$\lambda_1 = -\zeta + \sqrt{\zeta^2 - 1}$，$\lambda_2 = -\zeta - \sqrt{\zeta^2 - 1}$ である．そして，オーバーシュートとプレパレーションは減衰振動モデル $(0 < |\zeta| < 1)$，ヴィブラートは定常振動モデル $(|\zeta| = 0)$ で記述される．また，各 F0 動的変動成分の特性は，パラメータ ω，ζ，k によって制御される．微細変動は，白色雑音をカットオフ周波数 10 Hz の低域通過フィルタに通した後，最大振幅が 5 Hz になるように正規化することで生成し，F0 変化パターン全体に付与することで制御している．尚，F0 制御モデルの詳細に関しては文献 [8] を参照されたい．

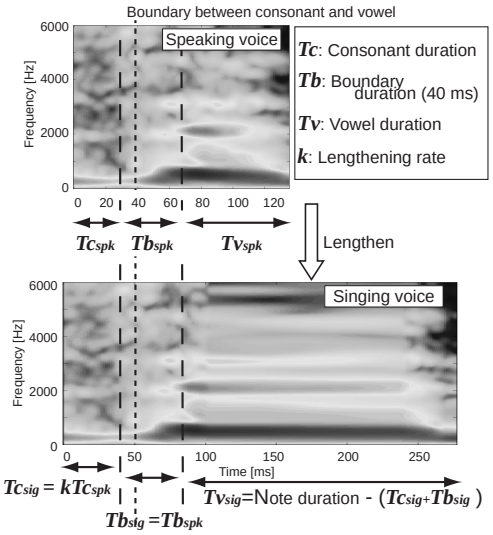

Fig. 4 *Schema of the duration control model.*

4 音韻長制御モデル

本章では，楽曲のテンポに従って，個々の音符が割り当てられた音韻（もしくは単語）ごとに時間伸長処理を行い，話声の時間構造を歌声に変換する音韻長制御モデルについて概説する．

図 4 に音韻長制御モデルの概要を示す．このモデルは，手動で与えられた各音韻の子音–母音境界を子音部 + 結合部 + 母音部に自動セグメンテーションし，各部のスペクトルと非周期性指標の時間系列を線形補間によって時間伸長する．尚，結合部分は子音–母音境界の-10 ∼ 30 ms までの計 40 ms としている．

子音部の時間長は，予め設定した朗読音声中の子音長に対する歌声中の子音長の比率（摩擦音 1.58，破裂音 1.13，半母音 2.07，鼻音 1.77，/y/1.13）に従って伸長処理を行う．

結合部は，時間伸長をせずに 40 ms で固定とする．

母音部は，伸長の対象としている音韻に割り当てられた音符長から，伸長した子音部の長さと結合部の 40 ms を差し引いた時間長に伸長する（図 4 を参照）．

5 スペクトル制御モデル

本章では，2 種の歌声特有のスペクトル特性を示し，それらを制御可能なスペクトル制御モデルについて概説する．

5.1 歌声特有のスペクトル特性

代表的な歌声特有のスペクトル特性の一つに，Sundberg[9] が“ 歌唱ホルマント (singer's formant) ”と命名した男性のオペラ歌唱の 3 kHz 付近において観測される顕著なホルマントピークがある．この特徴は，声に響きや明瞭さを与えると言われている [10]．

図 5 は，オペラ歌唱と，長唄歌唱における歌唱ホルマントの一例である．また，歌声特有の F0 変化に

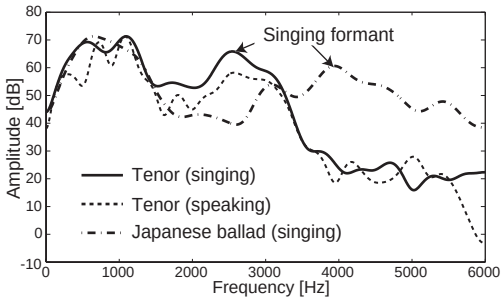

Fig. 5 *Examples of singing formant near 3 kHz.*

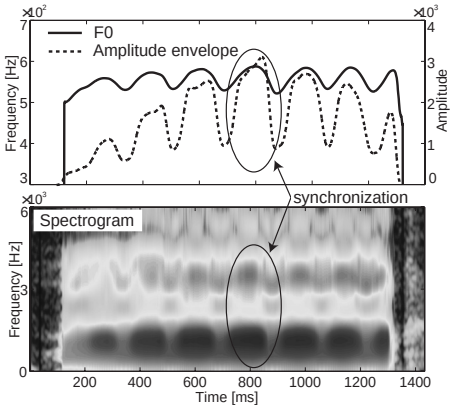

Fig. 6 *Example of formant amplitude modulation (AM) in synchronization with vibrato of the F0.*

連動したスペクトルの動的変動として，音声振幅全体がヴィブラートによって振幅変調し，それによって個々のホルマントも振幅変調することが報告されている [11]．図 6 に，オペラ歌唱による母音/a/発声時のサウンドスペクトログラム，F0 変化パターン，振幅エンベロープを示す．振幅エンベロープとホルマントが，F0 中のヴィブラートに同期して振幅変調していることが確認できる．

5.2 スペクトル制御モデル

図 1 に示すように，スペクトル制御モデルは，2 つの手続きから構成され，1 番目のモデルで歌唱ホルマントが，2 番目のモデルでヴィブラートに同期した音声振幅及びホルマントの振幅変調が制御される．

図 7 に，スペクトル制御モデル 1 の概要を示す．このモデルは，次式に従って，話声の母音区間のスペクトルにおいて 3 kHz 付近に存在するピークを強調することで歌唱ホルマントを制御・付与する．

$$S_{sg}(f) = W_{sf}(f) S_{sp}(f), \tag{3}$$

ここで，$S_{sp}(f)$ と $S_{sg}(f)$ は，それぞれ話声と歌声のスペクトルである．$W_{sf}(f)$ は，話声のスペクトルピークを強調させる荷重関数で次式で表わされる．

$$W_{sf}(f) = \begin{cases} (1+k_{sf})(1-\cos(2\pi \frac{f}{F_b+1})), & |f-F_s| \le \frac{F_b}{2} \\ 1, & \text{otherwise} \end{cases} \tag{4}$$

ここで，F_s は，$S_{sp}(f)$ の 3 kHz 付近のピーク周波数で，F_b は強調させる帯域幅を，そして k_{sf} は強調させる割合をそれぞれ調整するパラメータである．また，非周期性指標に関しては，3 kHz 付近に存在する

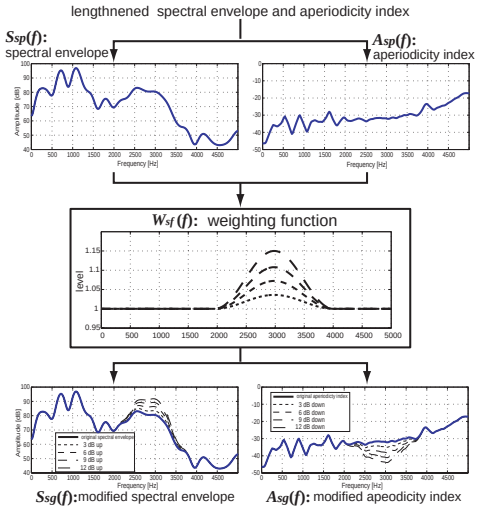

Fig. 7 *Schema of the spectral control model 1.*

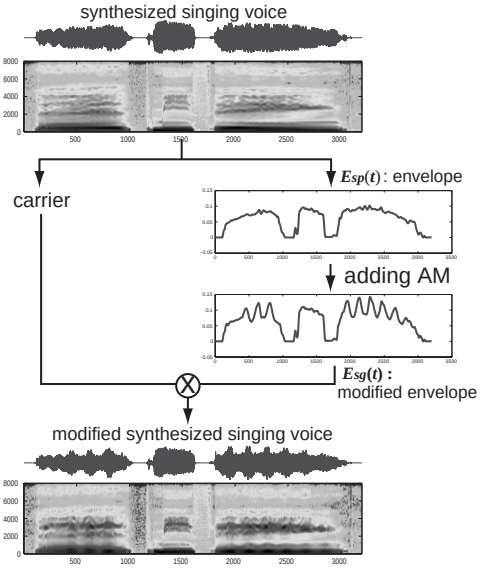

Fig. 8 *Schema of the spectral control model 2.*

ディップを関数 $W_{sf}(f)$ によって強める（顕著な谷を付与する）処理を行う．

図 8 にスペクトル制御モデル 2 の概要を示す．このモデルは，スペクトル制御モデル 1 の処理後に合成された歌声の振幅エンベロープに対して振幅変調を付与することで，音声振幅とホルマントの振幅変調を制御する．この振幅変調は，F0 制御モデルによってヴィブラートが付与された区間において，次式によって付与される．

$$E_{sg}(t) = (1 + k_{am} \sin(2\pi f_{am} t)) E_{sp}(t), \tag{5}$$

ここで，$E_{sp}(f)$ と $E_{sg}(f)$ は，それぞれ話声と歌声の振幅エンベロープである．f_{am} と k_{am} は，振幅変調の速さ（rate）の大きさ（extent）をそれぞれ制御するパラメータである．

6 歌声合成音の評価

本章では，提案した歌声合成システムを用いて，前章までで述べてきた F0，スペクトルにおける各種音響特徴を個々に制御した歌声合成音を作成し，それらを

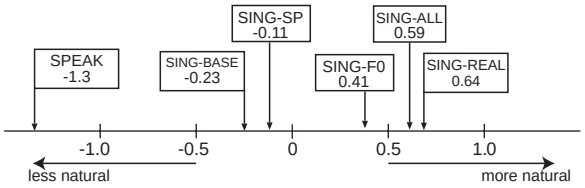

Fig. 9 *Perceptual effects of acoustic features unique to singing voices.*

聴取実験によって評価する．これにより，提案している歌声合成システムの評価を行うと同時に，各種特徴の歌声知覚への影響を比較する．

6.1 歌声合成

合成音は，男女各 1 名が日本童謡「七つの子」の歌いだし“ からすなぜなくの ”を朗読した音声を対象に作成した以下の 6 種である．

SPEAK：歌詞の朗読音声
SING-BASE：音韻長制御と F0 制御（但し F0 動的変動成分は付与していない）を行った合成音
SING-F0：音韻長制御と F0 制御（すべての F0 動的変動成分を付与）を行った合成音
SING-SP：音韻長制御とスペクトル制御 1，2 を行った合成音
SING-ALL：すべての制御を行った合成音
SING-REAL：**SPEAK** と同じ人間による歌声

尚，**SPEAK** と **SING-REAL** は STRAIGHT によって分析・再合成されたものを使用した．

6.2 聴取実験

上記合成音を実験刺激として，シェッフェの一対比較法 [12] によって，歌声の自然性に関する間隔尺度を求めた．実験に用いた評価尺度は，歌声の自然性に関する 7 段階評価（+3:先の刺激がとても自然，+2:自然，+1:やや自然，0:どちらとも言えない，-3:後の刺激がとても自然，-2:自然，-1:やや自然）である．被験者は，正常な聴力を有した大学院生 6 名（男性 5 名，女性 1 名）である．実験環境は防音室内で行い，刺激音はヘッドホン (STAX SR-404) を介して呈示した．

実験結果を図 9 に示す．この結果から，各特徴を付与することで自然な歌声として知覚されるようになり，すべての特徴を付与した合成音の自然性は，原歌声と同程度であることが確認された．また，各種特徴の影響を比較した結果，音韻長だけを制御しても（合成音 **SING-BASE**）自然な歌声とは知覚されず，それに加えて F0 とスペクトルの音響特徴が歌声を知覚する上で必要であることが明らかとなった．その中で，F0 動的変動成分の影響はとりわけ大きく，スペクトル特性は F0 動的変動成分と共存することで強い影響を与えていることが示された．

7 まとめ

本稿では，歌詞の朗読音声を歌声に変換する歌声合成システム：SingBySpeaking を提案した．このシステムは，F0，音韻長，スペクトルをそれぞれ制御するモデルで構成され，各モデルにおいて歌声特有の音響特徴を朗読音声に制御・付与することで歌声を合成する．F0 制御モデルは，譜面情報から得られるメロディの遷移の概形に対して，4 種類の動的変動成分を付与することで歌声の F0 変化パターンを生成する．音韻長制御モデルは，楽曲のテンポに基づいて，話声中の各音韻長を伸長する．スペクトル制御モデルは，話声のスペクトルに対して，2 種の歌声特有のスペクトル特性を付与することで歌声のスペクトルを生成する．聴取実験によって合成音声を評価した結果，原歌声と同程度の自然な歌声を合成できることを示した．また，話声から歌声の変換には，F0 動的変動成分の制御が最も重要であることが明らかとなった．

今後は，朗読音声のセグメンテーションを自動化することで，誰でも簡単に利用できる歌声合成システムへ発展させることが必要である．また，本システムを用いた歌声合成・知覚実験によって，新たな歌声特有の音響特徴や，歌唱者や歌唱スタイルの違いを規定する音響特徴の抽出を行う予定である．

謝辞 本研究の一部は，科学技術振興機構 CrestMuse プロジェクトによる支援を受けた．

参考文献

[1] J. Bonada *et al.*, “Synthesis of the Singing Voice by Performance Sampling and Spectral Models,” IEEE Sig. Proc. Mag., 24, 2, 67-79, 2007.
[2] 剣持他, “歌声合成システム VOCALOID,” 情処研報, 2007-MUS-072, 25-28, 2007.
[3] 酒向他, “隠れマルコフモデルに基づいた歌声合成システム,” 情処学論, 45, 3, 719-727, 2004.
[4] 吉田他, “歌声合成システム：CyberSingers ,” 情処研報, 1998-SLP-025, 35-40, 1998.
[5] T. Saitou, *et al.*, “Vocal Conversion from Speaking Voice to Singing Voice Using STRAIGHT,” Proc. Interspeech07, TuC.SS-2, 2007.
[6] T. Saitou, *et al.*, “Speech-To-Singing Synthesis: Converting Speaking Voices to Singing Voices by Controlling Acoustic Features Unique to Singing Voices,” Proc. WASPAA07, pp. 215-218, 2007.
[7] H. Kawahara, *et al.*, “Restructuring speech representations using a pitch adaptive time-frequency smoothing and an instantaneous- frequency based on F0 extraction: Possible role of a repetitive structure in sounds,” Speech Commun., 27, 187-207, 1999.
[8] T. Saitou, *et al.*, “Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis,” Speech Commun., 46, 405-417, 2005.
[9] J. Sundberg, “Articulatory Interpretation of the ’Singing Formant’,” J. Acoust. Soc. Am, 55, 838-844, 1974.
[10] J. Sundberg, “Singing and timbre,” Music room acoustic, Stockholm: Royal Swedish Academy of Music Publications, Vol. 17, pp. 57-81, 1977.
[11] Y. Horii, “Acoustic analysis of vocal vibrato: a theoretical intepretation of data,” J. Voice 3, 36-43, 1989.
[12] 天坂他, 官能評価の基礎と応用：自動車における感性エンジニアリングのために, 日本規格協会, 2000.