メディア操作　歌唱　体験デモ

話声を歌声に自動変換する

齋藤 毅　　後藤 真孝
（後藤グループ：産業技術総合研究所）

概要　歌詞の朗読音声（話声）を歌声に自動変換する歌声合成システムを提案する. このシステムは，音声分析合成系 STRAIGHT による分析 / 合成処理過程において，基本周波数，スペクトル，音韻長を制御するモデルによって歌声特有の音響特徴を操作することで話声を歌声に変換する. 従来の歌声合成システムに見られる歌詞（テキスト情報）からの歌声合成ではなく，話声（波形信号）からの歌声合成という新しいアプローチを採用することで，高品質な歌声合成だけでなく，「歌詞を朗読するだけで自分の歌声を作り・聴くことができる」という新しい歌声情報処理を可能にした. 更に，このシステムによる歌声合成を通じて，歌声を知覚する上で重要な音響特徴を明らかにした.
キーワード：歌声合成, STRAIGHT, F0 制御, 音韻長制御, スペクトル制御, 歌声知覚

1. はじめに

歌を歌うことは，音楽を楽しむ最も手近な手段であると同時に，歌詞である言語情報に加え感情や想いといった非言語情報を表出するための重要な手段である. その為，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 [1] が主流である. 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合成手法に基づいた実用性の高いものが多く，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，我々は，話声から歌声を合成する ***speech-to-singing synthesis*** という新しいアプローチによる歌声合成システムを提案する [3]. このシステムは，基本周波数（以後 F0），音韻長，スペクトルにおける歌声特有の様々な音響特徴に着目し，音声分析・合成系 STRAIGHT[4] による処理過程においてそれらの特徴を制御することによって歌詞の朗読音声を歌声へ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」という新たな歌声情報処理技術が確立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の概要を示す. システムの入力は，合成したい歌の歌詞の朗読音声（話声），その歌の譜面情報（メロディ遷移の概形），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（音韻と音符の同期情報）の 3 つである. 尚，朗読音声の音素境界は強制アライメントによって自動推定される. 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. 朗読音声を STRAIGHT によって F0 変化パターン，スペクトル系列，非周期性指標系列の音響パラメータに分解する
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. 生成・加工した各音響パラメータを用いて STRAIGHT によって歌声を合成する
6. スペクトル制御モデル 2 によって歌声合成音の振幅エンベロープを加工する

3. F0 制御モデル

F0 制御モデルは，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遷移の概形に対して，以下の **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.

1. **オーバーシュート**：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
2. **ヴィブラート**：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
3. **プレパレーション**：音高変化直前に変化とは逆方向振れる瞬時的な変動成分
4. **微細変動**：発声区間全体に観測される不規則で細かい変動成分

4. 音韻長制御モデル

音韻長制御モデルは，強制アライメントによって得られた各音韻の子音ー母音境界を子音部 + 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，結合部分は子音ー母音境界の -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.

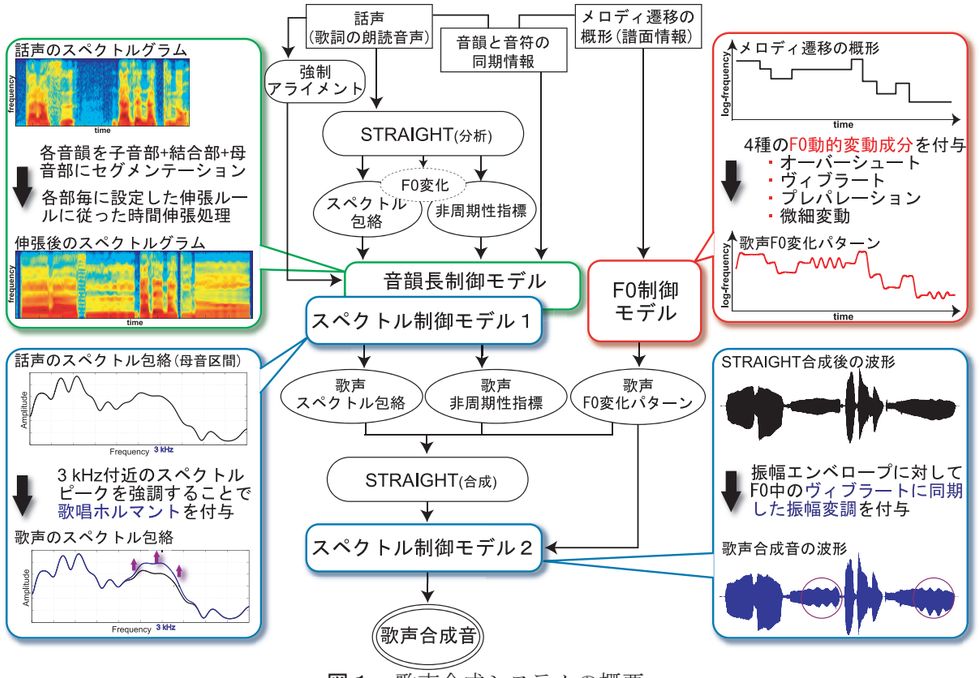

図 1　歌声合成システムの概要.

5. スペクトル制御モデル

スペクトル制御モデルは，2 つの手続きから構成され，制御モデル 1 で**歌唱ホルマント**，制御モデル 2 で**ヴィブラートに同期した音声振幅の変調**が制御される.

歌唱ホルマントは，特に男性のオペラ歌唱おいて観測される 3 kHz 付近の顕著なホルマントピークである. スペクトル制御モデル 1 では，時間伸長した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.

F0 変化中にヴィブラートが存在する場合，音声振幅が変調することが報告されている. スペクトル制御モデル 2 では，F0 制御モデルで付与したヴィブラートに同期した音声エンベロープの振幅変調を付与することで，これらの特徴を制御する.

6. 歌声合成音の評価

聴取実験によって歌声合成音を評価した結果，各種音響特徴を制御することで，話声から歌声に段階的に変化し，すべての音響特徴を制御した合成音の音質は原歌声と同程度であることを確認した. 更に，話声から歌声に変換する上で，F0 動的変動成分が最も重要な役割を果たしていることを明らかにした.

謝 辞

本研究の一部は，北陸先端科学技術大学院大学の赤木正人教授，鵜木祐史准教授の指導によって進められたものである. 強制アライメントの実装に協力頂いた藤原弘将氏 (産業技術総合研究所) に感謝する.

参 考 文 献

[1] 剣持他：歌声合成システム VOCALOID, 情報処理学会研究報告 , 2007-MUS-072, pp.25-28, (2007).
[2] 齋藤他：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 , 20078-MUS-074, pp.25-31, (2008).
[3] 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., Vol. 27, pp. 187-207, (1999).