6U-1

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

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 ‡‡

† 京都大学 工学部情報学科 ‡ 産業技術総合研究所 ‡‡ 京都大学 大学院情報学研究科 知能情報学専攻

1. はじめに

近年，コンピュータによる合成音声の利用機会が増加してきている．それに伴い，従来のテキスト読み上げの話声を合成することに特化した技術ではなく，感情音声や歌声といった表情豊かな音声を自由に合成可能な技術の構築が求められている．

齋藤らは話声から歌声を合成する歌声合成システム SingBySpeaking を実現した [1]．入力の話声に対して歌声固有の音響特徴を付与することで，自然な歌声合成を行う．これは，話声の声質を歌声合成音に反映できるので，声質を変えた話声を入力すれば表情豊かな歌声が自由に合成できる利点もある．しかし，歌唱時特有の音色を制御することは困難であり，歌声から話声への逆変換はできない．

本研究では，表現力豊かな話声合成の実現方法として，歌声を話声に変換する新しい音声合成アプローチに基づいた音声合成システム SpeakBySinging を提案する．話声には無い歌声固有の声質を保持した話声の合成方法を開発し，従来の音声合成では実現困難な表現力豊かな話声合成が実現可能になると考えられる．齋藤らは，F0，スペクトル，音韻長に話声と歌声の違いを規定する各種音響特徴の存在を指摘している [2]．そこで SpeakBySinging では，歌声と話声の音響構造の違いに着目し，歌声の F0，振幅，音韻長の各種音響パラメータを話声特有の特性に変換することで話声への変換合成を実現する．歌声のスペクトル特性を出来るだけ保持した状態で話声へ変換することで，表情豊かな話声の合成を試みる．

2. 歌声と話声の音響特徴の違い

歌声と話声の F0・振幅・音韻長は，以下の特徴をもつ．これらの特徴の違いを図 1 に示す．

F0 歌声は楽曲のメロディに対応した階段型に，話声は文頭と文末で F0 が低い「への字」型に遷移する．平均的に歌声の F0 は話声よりも高い傾向がある．

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

音韻長 歌声では歌の譜面に依存して長さが大きく異なり，多くの場合，母音が持続するので子音よりも長い．話声では音素に非依存にほぼ一定である．

3. 歌声から話声への変換

図 2 に SpeakBySinging の概要を示す．SpeakBySinging は，STRAIGHT [5] の分析・合成処理体系に F0，振幅，音韻長の制御を組み込んだ構成となっている．システム

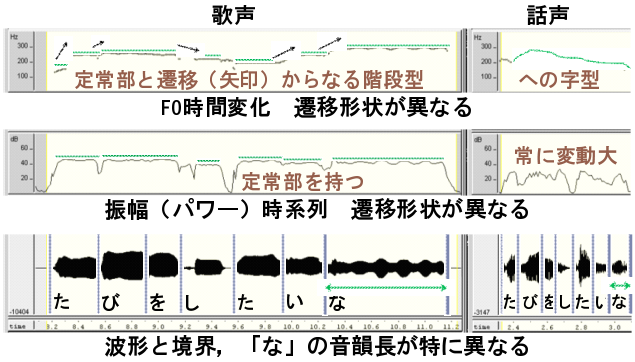

図 1: 歌詞「たびをしたいな」の歌唱音声（左）と同一歌詞の朗読音声（右）の，上から順に音声信号波形と音節境界，F0 時系列，振幅（パワー）時系列．ただし，本システムでは音節境界ではなく，より細かい音素境界を用いる．

への入力は単独歌唱音声（歌声），朗読音声（話声），歌詞テキストの 3 つであり，出力は歌唱音声の声質および朗読音声の音韻長・F0・振幅をもつ合成音声である．処理の流れを以下に示す．

1. 歌声に対して STRAIGHT 分析を行い，F0 軌跡 (**歌 F0**) と非周期指標系列 (**歌 AP**) とスペクトル系列 (**歌 N3S**) を抽出する．
2. 隠れマルコフモデルを用いた強制アライメントで歌声と話声それぞれの音素境界ラベルを求める．
3. 2 で求めた音素境界に基づいて歌声の各種音響パラメータ (**歌 F0**, **歌 AP**, **歌 N3S**) 中の音素境界時刻が，話声の音素境界時刻と同期するように伸縮する．
4. 話声に対し，STRAIGHT 分析によって F0 軌跡 (**話 F0-FULL**) を抽出する．ここでは，無声区間においても強制的に F0 を推定することで，発声区間すべてにおいて F0 値を持つ F0 軌跡を抽出する．この**話 F0-FULL** と 3 で伸縮した**歌 F0** との間で有声・無声区間が同一となるように**話 F0-FULL** を調整し，合成に用いる**話 F0** を得る．また，**話 F0** を用いて再度 STRAIGHT 分析を行うことで，話声のスペクトル系列**話 N3S** を抽出する．
5. **歌 N3S** の各時刻における振幅（パワー）が，**話 N3S** と同じになるように調整する．
6. 上記の処理で生成された**話 F0**，**歌 AP**，**歌 N3S** を用いて STRAIGHT 合成を行い，出力音声を生成する．

3.1 音素境界情報に基づく歌声音韻長制御

歌声と話声の音素境界に基づき**歌 AP** と**歌 N3S** を時間方向に線形補間伸縮する．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．このため，各音素の全区間を均一に伸縮すると音素遷移時間も変化してしまい，合成音の音質に悪影響を与える可能性が高

SpeakBySinging: A Speaking Voice Synthesis System Converting Singing Voices to Speaking Voices By Controlling F0, Amplitude, and Duration
Shinpei Aso (Kyoto Univ.), Takeshi Saitou, Masataka Goto (AIST), Katsutoshi Itoyama, Toru Takahashi, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto Univ.)

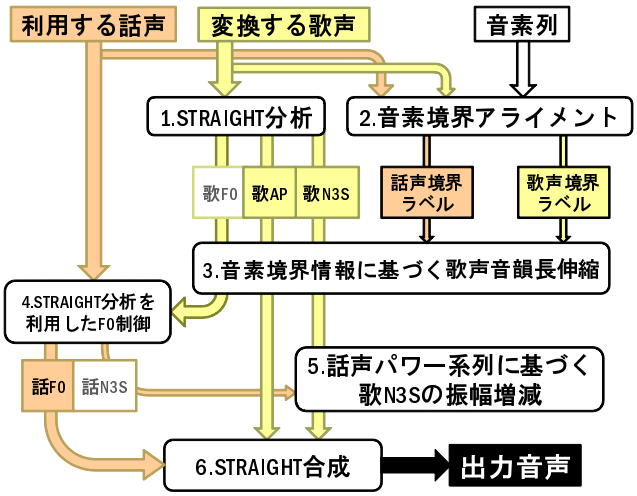

図 2: SpeakBySinging のブロックダイアグラム．**F0** は F0 軌跡，**AP** は非周期性指標系列，**N3S** はスペクトル系列を表す．図中白抜きの箱・点線で示した特徴量・矢印は，音響特徴の制御時のみに使用し，合成時には用いない特徴量を表す．

い．そこで，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，それ以外の区間を伸縮することで**歌 AP** と**歌 N3S** 中の各音素長が話声の音素長となるように時間伸長している [1].

また，歌声のスペクトルの時間変化にはビブラート等の動的な変動の影響が含まれており，時間伸長した**歌 N3S** 中にもそれらの変動が残る可能性が高いと考えられる．そこで，時間伸長した**歌 N3S** において，各周波数毎に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している．

3.2 有声・無声情報を利用した F0 制御

STRAIGHT 合成は，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．しかし，歌声と話声で有声区間が異なる場合があり，この不一致によって無声音区間に F0 割り当てた状態で合成が行われた場合，音質の劣化が生じる可能性がある．この不一致を避けるため，話声から全時刻において強制的に F0（**話 F0-ALL**）を抽出し，音韻長伸縮後の**歌 F0** において無声区間（F0 が 0 の区間）の**話 F0-ALL** の値を 0 とすることで，時間伸長した歌声スペクトル**歌 N3S** と有声・無声区間の対応がとれた話声の F0 軌跡 **(話 F0)** を生成する．

4. 本システムの評価と考察

本システムで合成される音声の評価を行う．大石らによって提案されている話声と歌声の自動識別システムでは，MFCC を識別パラメータとして用いることで高精度な識別を実現している [3]．そこで，同一人物による歌唱音声と歌詞の朗読音声を入力に与えて得られる合成音と，入力に用いた朗読音声との MFCC 距離を算出することで，合成された話声の評価を行う．評価用の入力には歌声研究用音楽データベース「AIST ハミングデータベース」[4] を用いる．ハミングデータベースには日本人歌唱者 75 名が「RWC 研究用音楽データベース：ポピュラー音楽」[6] の一部を単独で歌唱した音声と歌詞を朗読した音声が含まれている．評価には，女性 3 名（J002, J003, J014）と男性 2 名（J052, J054）の計 5 名による楽曲 RWC-MDB-P-2001 No.78 の出だし部分の単独歌唱

表 1: 各発声者の合成音と実話声の MFCC 距離．

各人各合成音と実話声との MFCC 距離，値が小さいほど良い					
入力データ	J002	J003	J014	J052	J054
SING-BASE	42.0	14.4	14.8	27.7	63.5
SING-F0	29.7	15.8	15.4	28.1	70.1
SING-AMP	28.2	27.6	30.6	18.5	69.5
SING-ALL	26.1	14.7	15.4	27.5	37.4
SPEAK-straight	5.3	4.8	7.7	8.3	7.0

音声と朗読音声データを使用し，各人の歌唱音声（入力歌声）と J002 の朗読音声（入力話声）を対としてシステムの入力に与えて作成した以下の合成音を用いた．

SING-BASE 歌声に音韻長制御を行った合成音（F0 軌跡は歌声の F0 軌跡を時間伸縮したものを用いる）

SING-F0 歌声に音韻長制御と F0 置換制御を行った合成音

SING-AMP 歌声に音韻長制御と振幅制御を行った合成音

SING-ALL 歌声にすべての制御を行った合成音

SPEAK-straight 話声を STRAIGHT によって分析・再合成した合成音

SPEAK-straight は STRAIGHT 分析合成による音質変化を知るために合成し，本システムの限界（最良値）を求めるために用いる．各発声者の 5 種の合成音と朗読音声（サンプリング周波数はいずれも 16kHz）に対して，12 次元の MFCC を抽出し，同一フレーム時刻の合成音 MFCC と朗読音声 MFCC とでユークリッド距離を求め，時間平均（フレーム数で割り算）した値を表 1 に示す．結果より，F0 や振幅を制御することで本人の話声に近くなり，両方の制御を行うことでより近くなることが確認できた．一方で，F0，振幅，音韻長の各種特徴の制御による MFCC 距離の変化は，発声者によって異なる結果となった．また，SING-ALL と SPEAK-straight との MFCC 距離平均が大きく異なる原因としては，スペクトル情報を話声の特徴に変換していない事が大きく影響していると考えられる．

5. おわりに

本研究では，単独歌唱音声の F0・振幅・音韻長を制御し話声に変換する話声合成システム SpeakBySinging を開発した．各音響特徴を話声の特性に制御・置換することにより，実際の話声に近く音声を合成可能なことを確認した．今後は，入力として話声の情報を用いずに歌声から話声に合成可能なシステムに発展させる予定である．

謝辞 強制アライメントをはじめ多くの助言を頂いた産業技術総合研究所の中野倫靖氏・藤原弘将氏に感謝する．本研究の一部は，科研費，GCOE，CrestMuse の支援を受けた．

参考文献

[1] 齋藤 毅他: “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,” 情処研報 MUS, No.12, pp.25-32, 2008.
[2] 齋藤 毅他: “歌声らしさの知覚モデルに基づいた歌声特有の音響特徴量の分析,” 日本音響学会誌, Vol.64, No.5, pp.267-277, 2008.
[3] 大石 康智他: “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,” 情処学論, Vol.47, No.6, pp.1822-1830, 2006.
[4] 後藤 真孝他; “AIST ハミングデータベース : 歌声研究用音楽データベース,” 情処研報 MUS, No.82, pp.7-12, 2005.
[5] 河原 英紀; “Vocoder のもう一つの可能性を探る-音声分析変換合成システム STRAIGHT の背景と展開- ,” 日本音響学会誌, Vol.63, No.8, pp.442-449, 2007.
[6] 後藤 真孝他; “RWC 研究用音楽データベース: 研究目的で利用可能な著作権処理済み楽曲・楽器音データベース,” 情処学論, Vol.45, No.3, pp.728-738, 2004.