2R-6

F0・音韻長・パワー制御による歌声らしさ・話声らしさの変化の評価

阿曽 慎平 † 齋藤 毅 ‡ 後藤 真孝 ‡‡ 糸山 克寿 † 高橋 徹 † 尾形 哲也 † 奥乃 博 †

† 京都大学 大学院情報学研究科 知能情報学専攻 ‡ 金沢大学 理工学域電子情報学系 ‡‡ 産業技術総合研究所

1. はじめに

歌声合成技術の発展と動画コミュニケーションサイトの普及により，歌声の表現やその自然さに注目して歌声を聴き、またそれについて議論・批評する機会が増えてきた．歌声は話声より非言語的な意味も含めて伝達するという側面が大きいので，表現としての歌声に関する議論は、話声に関する議論よりも興味を持つ人達がいる。人は歌声か話声かで異なるとらえ方をするが、聞き手は何を基準に歌声と話声を聞き分けているのだろうか．

歌声と話声の両方を扱った研究に，我々の先行研究 SpeakBySinging [1] がある，SpeakBySinging では，歌声のスペクトル包絡 (主観量の声質に相当) を保ちながら基本周波数 (F0，主観量の音高に相当) ・音韻長 (各音素の継続時間，主観量の話速に相当) ・パワーの 3 つを話声のものへと制御することで歌声から話声への変換を実現している．SpeakBySinging では F0・音韻長・パワーの 3 つの特徴量を一度に制御していたが，どの要素が歌声らしさ・話声らしさに重要であるのかまでは調査していなかった. それに対して，大石ら [2] は F0，MFCC（音色特徴量）とそれらの時間差分が歌声と話声の識別に有効であることを報告している．しかし，これらは比較的大局的な特徴であり，歌声と話声を聞き分ける知覚機構を明らかにするためには，より詳細な音響特徴量に着目した取組みが必要になってくる．

本稿では，どのような音響特徴量が歌声らしさ・話声らしさの知覚に重要であるかを調査する．SpeakBySinging で制御した 3 つの特徴量に加えて，F0 のゆらぎ成分である Jitter という音響特徴 [3] をそれぞれ独立に制御した音声を合成し，聴取実験によって話声らしさ，歌声らしさの知覚に寄与する音響特徴量を調査する．

2. 歌声と話声を扱う研究

我々の SpeakBySinging (歌声を話声に変換) [1]，齋藤らの SingBySpeaking (話声を歌声に変換) [4] では、いずれも F0・音韻長・パワー (振幅) を積極的に制御する一方、スペクトル包絡は歌唱フォルマントの操作のような最小限の制御にすることで、元の音声の声質を保ったままの変換を実現していた．これらの音響特徴の制御には音声分析合成系 STRAIGHT[5] が用いられている．STRAIGHT の分析過程では，音声から F0・STRAIGHT スペクトル・非周期性指標を抽出する．以後の説明で STRAIGHT スペクトル・非周期性指標の 2 つをまとめてスペクトルパラメータと表現する．STRAIGHT によって，各種音響パラメータが自由に加工でき，それら加工が反映された音声を合成できる．また，2 つの音声のパラメータを組み合わせた音声を合成することもできる．この方法を用いて歌声と話声から刺激音を合成する．

Evaluation of Singing-ness and Speaking-ness Conversion by Controlling the F0, Phoneme Duration, and Power *Shinpei Aso (Kyoto Univ.), Takeshi Saitou (Kanazawa Univ.), Masataka Goto (AIST), Katsutoshi Itoyama, Toru Takahashi, Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto Univ.)*

3. 刺激音の合成方法

本章では，聴取実験に用いる刺激音の合成方法を説明する．刺激音は，ある歌詞を朗読した音声 (話声) と歌った音声 (歌声) の対の音声に対して音響特徴量の置換を行うことで得られる．本稿で行う置換方法には多数あり，その一覧を表 1 に示す.

表より，異なる方法の数は $2\times3\times2\times3\times2=72$ 通りある．歌声・話声の F0 及びスペクトルパラメータは，STRAIGHT によって抽出し，両者の間で各パラメータを置換することで，表 1 に示すような刺激音を作成した．また，F0 に含まれる様々な時間変動の影響を調べるために，発声区間全体で F0 を一定値 150[Hz] に変換した刺激音も作成した．

以下、各制御の方法を述べる．

3.1 Jitter 制御

Jitter とは一定周期をもつパルス列が，局所的にタイミングずれが生じることを示す言葉である．声においてパルス列は声帯の振動間隔に相当し，声の Jitter は，歌声における感情知覚に関係があると報告されている [3]. 歌声らしさ・話声らしさに与える影響を調べるために，選択した F0 に対して Jitter の制御を行う．本稿では Jitter を数値化するために次式で定義する．

$$\mathrm{Jitter}(T(n)) = |2T(n+1) - T(n) - T(n+2)| \quad (1)$$

ここで $T(n)$ は n 番目のパルス（声帯振動）列から $n+1$ 番目までの時間間隔であり，1 番目の発生時刻を 0 とおくと，時刻 t の F0 データ $F(t)$ との関係は $T(n) = \frac{1}{F(\sum_{i=1}^{n} T(i))}$ で記述できる．以下の式で $n=1$ から $n=N-2$ (N はパルス列の総数) まで T を T' に更新することで Jitter を制御する．

$$T'(n+2) = (1-r)(2T(n+1) - T(n) - T(n+2)) \quad (2)$$

ここで r は元の Jitter 値を何倍にするかを決定するパラメータであり，r=1 (制御しない), 2 の 2 つの選択肢がある．

3.2 音韻長制御

各音韻の長さは、歌声における音韻長，話声における音韻長，一定のモーラ長（子音＋母音の長さ）のいずれかになるように制御する．

表 1: 合成する音声の選択肢一覧．各特徴量の制御方法を 1 つずつ選んで合成音を決定する．タグは表 2 用である．平均点はその制御が用いられている刺激音の評価値（4 章で説明）平均であり，5 に近いほど歌声らしさへの，1 に近いほど話声らしさへの寄与が大きいことを示す．

特徴量	タグ	制御方法の選択肢	平均点
スペクトル	sg	歌声のものを使用	2.97
パラメータ	sp	話声のものを使用	2.76
F0 (時系列)	sg	歌声の F0 使用	**4.06**
	sp	話声の F0 使用	**1.80**
	fix	F0 を 150[Hz] 固定	2.73
Jitter	2	F0 の Jitter 値 2 倍	2.84
	1	何もしない	2.88
音韻長	sg	歌声の音韻長を使用	3.70
	sp	話声の音韻長を使用	**1.61**
	fix	モーラ長を 150[ms] 固定	2.41
パワー	sg	歌声のパワー包絡使用	2.89
	sp	話声のパワー包絡使用	2.85

この制御は音韻長の抽出，所望の音韻長に対する音韻長の比率計算，時間伸縮処理の 3 プロセスからなる [1]. まず，音素既知とした歌声と話声に対し隠れマルコフモデルを用いたビタビアライメントを行った後，アライメント誤りを手動で修正して音素境界を得る．次に，各音素ごとに元の音韻長と所望の音韻長との比率を計算する．例えば，ある音素の元の音韻長が 150ms., 所望の音韻長が 50ms. の場合は 1/3 となる．最後に比率を基にスペクトルパラメータ，F0 を音素区分の線形時間伸縮する．

3.3 パワー制御

パワー制御は，振幅包絡を話声，又は歌声のものに置換することで実現する．例えば，歌声のパワー制御を行う場合は，次の処理を行う．パワーは各時刻ごとに，スペクトルパラメータの各周波数 bin の総和として計算されるため，スペクトルパラメータを加工することになる．パワー包絡の制御には各時刻の歌声のスペクトルパラメータに対し，歌声のパワーに対する話声のパワーの比率をかければよい [1].

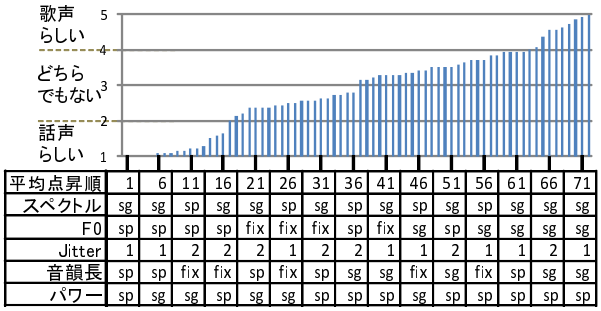
4. 聴取実験

3 章で述べた方法に基づいて刺激音を作成し，聴取実験で評価を行う．合成の素材となる歌声と話声には，研究用データベース (AIST ハミングデータベース) [6] の歌唱者 J048 による楽曲 P078 の一部を歌唱した音声（約 12[ms]）と歌詞朗読した音声（約 5[ms]）を用いる．被験者は 14 名（共著者 2 名を含む）名で，刺激音はヘッドホン (MDR-900ST) で再生される．被験者は全 72 個の刺激音を 1 回だけ聞き，それぞれに対し 5 段階 (1. 話声である　2. どちらかというと話声である 3. どちらでもない 4. どちらかというと歌声である 5. 歌声である) で評価する．

各刺激音の平均点（評価平均）を昇順に並べたものを図 1，各制御選択肢ごとの平均点を表 1，音韻長と F0 制御時の平均点を表 2 に示す．これらから今回の実験条件下では F0・音韻長が歌声らしさ・話声らしさへ大きく寄与し，また表 1 から両者を同時に制御すれば歌声と話声の相互変換が可能であると言える．他方でスペクトルパラメータ・Jitter・パワーは制御による影響が少ない．F0 や音韻長を単純に一定値にしただけでは歌声らしく，または話声らしくすることができないこともわかる．

表 2: 音韻長制御と F0 制御を組み合わせた時の平均点．5 に近いほど歌声らしく，1 に近いほど話声らしい．括弧内は表 1 のタグを表す．

	音韻長 (sg)	音韻長 (sp)	音韻長 (fix)
F0 (sg)	**4.71**	3.79	3.69
F0 (sp)	3.05	**1.05**	**1.31**
F0 (fix)	3.34	2.38	2.46



平均点昇順	1	6	11	16	21	26	31	36	41	46	51	56	61	66	71
スペクトル	sg	sg	sp	sg	sg	sp	sg	sp	sg	sp	sg	sp	sg	sg	sg
F0	sp	sp	sp	sp	fix	fix	fix	sp	fix	sg	sp	sg	sg	sg	sg
Jitter	1	1	2	2	2	1	2	2	1	1	2	1	1	2	1
音韻長	sp	sp	fix	fix	sp	fix	sp	sg	sg	fix	sg	fix	sp	sg	sg
パワー	sp	sg	sg	sp	sg	sg	sp	sp	sp	sg	sp	sp	sp	sp	sp

図 1: 評価結果を平均点順に並べたもの．縦軸は平均点，横軸は順位を表す．いくつかの刺激音について，それぞれ使用された制御方法を表に示す．音韻長・F0 が平均点に影響しているのがわかる．余白の都合スペクトルパラメータをスペクトルと略記している．

5. おわりに

歌声らしさ・話声らしさに寄与する音響特徴量を聴取実験により調査した．着目した特徴量はスペクトルパラメータ・F0・Jitter・音韻長・パワーであり，これらの特徴量を独立に制御した 72 個の刺激音を 14 名の被験者に提示し，各音を歌声らしいか話声らしいかで 5 段階評価した．聴取実験の結果，F0 と音韻長が，歌声らしさ・話声らしさの聴感的違いを規定している可能性が示された．他方，スペクトルパラメータ・Jitter・パワーについては今回の実験での影響は確認できなかった．今後は，もっと多くの音声を評価用の刺激音に用いた実験を行い，更には F0・音韻長のどのような違いが，歌声らしさ・話声らしさに寄与しているか調査する必要があると考えている．

謝辞本研究は科研費, CrestMuse，GCOE の支援を受けた．

参考文献

[1] S. Aso *et al.*: SpeakBySinging: Converting Singing Voices to Speaking Voices While Retaining Voice Timbre, *DAFx*, pp. 143–150, 2010.
[2] 大石他：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，情処論，Vol. 47, No. 6, pp. 1822–1830, 2006.
[3] D. Erickson *et al.*: Ah, how sweet the sound: Some acoustic characteristics of emotionally sung /ah/, *Intersinging*, 2010.
[4] 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，情処研報，Vol. 2008, No. 12, pp. 25–32, 2008.
[5] 河原：聴覚の情景分析が生み出した高品質 VOCODER:STRAIGHT, 音響誌，Vol. 54, No. 7, pp. 521–526, 1998.
[6] 後藤他：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，No. 82, pp. 7–12, 2005.