5R-8

歌声－話声変換における 動的音響特徴量が話声らしさに及ぼす影響

山崎 健史 † 池宮 由楽 ‡ 糸山 克寿 ‡ 奥乃 博 ‡

† 京都大学 工学部情報学科 ‡ 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、VOCALOID を代表とする歌声合成ソフトが普及し，CGM（Consumer Generated Media）において素人による楽曲制作が盛んになっている．それに伴い，より自然で高音質な音声合成技術が求められるようになっている．合成技術の発展には，歌声と話声の識別に関する知見が重要である．すなわち，人間が歌声と話声の聞き分けにどのような音響的特徴を用いているのかが分かれば，その知見を音声合成技術の性能向上に応用することができる．

歌声と話声を扱う先行研究として，阿曽らは歌声から話声を合成する話声合成システム SpeakBySinging [1] を実現し，さらに話声らしさ，歌声らしさに寄与する音響特徴量を調査した [2]．本研究では，各音響特徴量を話声のもの，或いは歌声のものから選択する手法を用いており，歌声－話声間における網羅的な調査は行っていなかった．

また，大石らは F0 と MFCC およびそれらの時間差分成分の利用が，歌声と話声の識別に有効であることを報告している [3]．しかし，この手法は機械上での歌声と話声の識別を対象としたものであり，人間の聴覚的な識別に影響を与える要素に関しては明らかにされていない．

本稿では，聴取実験によって 3 つの音響特徴量が話声らしさにどのような影響を与えるか網羅的に調査する．音響特徴量として F0（基本周波数）・音韻長・パワー (振幅) に着目し，歌声を話声へと変換するシステムを用いて，各音響特徴量を独立して制御する．得られた刺激音を用いて聴取実験を行うことで，各特徴量の有効性を評価する．

2. 歌声－話声変換システム

歌声－話声変換システムは，阿曽らの SpeakBySinging を元に作成した．SpeakBySinging では，音響特徴量として F0・音韻長・パワーに着目し，入力歌声の音響特徴量をサンプル話声の音響特徴量に合わせることで，元の歌声の声質を保持した変換を実現している．これら音響特徴量の制御には，音声分析合成システム STRAIGHT [4] を用いている．STRAIGHT は，音声を F0，非周期性指標、スペクトル情報に分解し，分解された各要素を編集後に再合成することで声色を変化させることができる．

歌声－話声変換システムの全体図を図 1 に示す．本システムの構成は，音素境界アライメント，音韻長制御モデル，F0 制御モデル，パワー制御モデルからなる．それぞれ構成要素については，**2.1 - 2.4** 節に示す．

2.1 音素境界アライメント

入力音声の音素区間のセグメンテーションを行う．音素境界アラインメントは，大語彙連続音声認識エンジン Julius [5] を用いて実装する．Julius では歌詞の音素情報

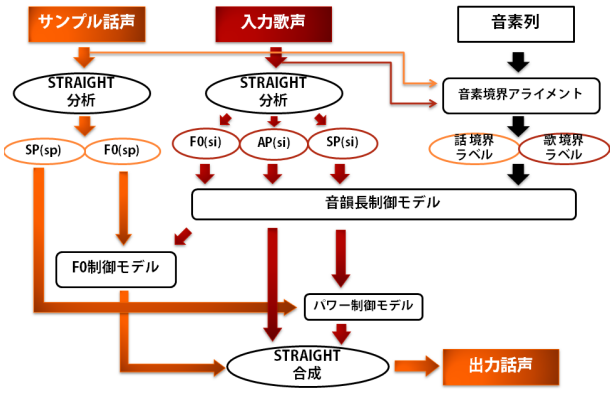

図 1: 歌声－話声変換システムの全体図．F0 は F0 時系列，AP は非周期性指標系列，SP はスペクトル系列を指す．

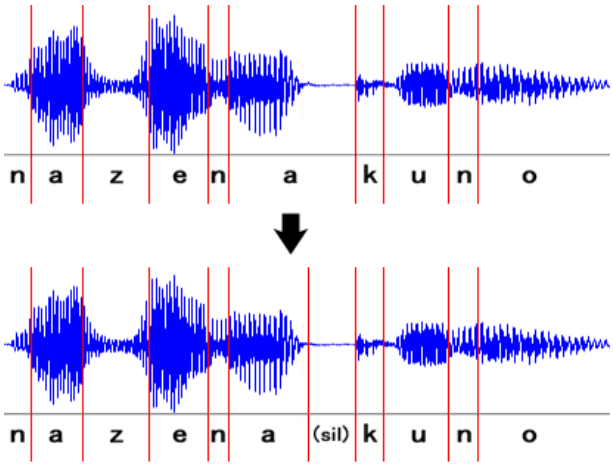

図 2: 音素境界アライメント－無声区間の認識：無声区間を認識し，セグメンテーションに（sil）を追加

を入力として指定することで自動セグメンテーションを行うことができるが，音声時間軸上に対して隙間なくセグメンテーションを行うため，音素間で無音の区間が存在したときに無音区間を含んだセグメンテーションを行ってしまい，音響特徴量を制御する上でノイズの原因となってしまう．

その対処として，STRAIGHT 分析により抽出した有声無声区間に関する情報を用いて，母音区間の内無声区間を無音区間としてセグメンテーションに追加することで，セグメンテーションのずれを防止する（図 2）．

2.2 音韻長制御モデル

音素境界アライメントによって得られた情報に基づき，STRAIGHT 分析で抽出された各要素における音素の音韻長の長さを制御する．この制御は音素毎の音韻長の抽出，所望の音韻長に対する比率計算，時間伸縮処理の 3

Effect of dynamic acoustic features in singing to speaking voice conversion. *Takeshi Yamasaki, Yukara Ikemiya, Katsutoshi Itoyama, and Hiroshi G. Okuno (Kyoto Univ.)*

プロセスからなる．音素間の結合部分における遷移時間は話声，歌声で大きく差がないと考えられるので，子音側 10ms，母音側 30ms は非伸縮区間とする．

音韻長は話声と歌声間の任意の割合で変更可能とする．また，話声の音韻長は音素に依存せずほぼ一定であるという特徴 [1] に基づき，母音区間を固定長で変換できるようにした．

2.3 F0 制御モデル

音韻長制御モデルで長さを制御された F0 時系列の高さを制御する．本稿では，話声－歌声間の任意の割合で F0 の高さを変化できるようにした．

2.4 パワー制御モデル

音韻長制御モデルで長さを制御されたスペクトルをパワー (音量) の指標に従って制御する．スペクトルのパワーを以下の式で定義する．

$$Power(t) = \sum_{f=1}^{F} (N(f,t))^2.$$

ここで F は周波数帯域数，T は時間フレームサイズ，$N(f,t)$ は $F \times T$ のスペクトル系列をそれぞれ表す．本稿では，各時刻の振幅の比率を

$$Rate(t) = rate_{si-sp} \cdot 10\log_{10} \frac{Power_{speak}(t)}{Power_{sing}(t)}$$

と定義することで，話声－歌声間の dB 軸上でスペクトルパワーを任意の割合で変動できるようにしている．ただし，$rate_{si-sp}$ は $0 \leq rate_{si-sp} \leq 1$ を満たす実数で，話声－歌声間のパワー割合を意味する．

3. 聴取実験

本システムで変換された音声を用いて聴取実験を行うことで，各特徴量の話声らしさへの影響度を評価する．入力データには，楽曲は童謡：七つの子の出だし部分とし，男性同一人物が歌唱した歌声 (約 10s)，および朗読したサンプル話声 (約 5s) を用いる．変換音声は **2.2** - **2.4** 節の手法に基づき，音韻長を話声－歌声間で 5 段階と固定長 (150ms) の 6 通り，F0 を話声－歌声間で 5 段階，パワーを話声－歌声間で 5 段階，計 $6 \times 5 \times 5 = 150$ 通り作成した．

評価尺度は SD 法に基づき，以下の 4 項目を選定した．

1. 歌声らしい － 話声らしい
2. 安定した － 不安定な
3. 感じのよい － 感じの悪い
4. きれいな － 汚い

150 通りの変換音声を 50 通りずつの 3 グループに分け，1・2 グループ，2・3 グループ，1・3 グループと 100 通りの 3 組に分けた．被験者 3 人を均等に各組に分けて視聴してもらい，上記の尺度を 7 段階で評価してもらった．各尺度についての結果を下記に示す．

3.1 歌声らしい－話声らしい

各刺激音の平均点を昇順に並べたものを図 3 に示す．図から歌声から話声へと変わる変曲点は見られないことがわかる．歌声らしさへの影響度は，F0 制御，音韻長制御の順で強いことが確認できた．話声らしさへの影響度は，音韻長制御，F0 制御，音韻長固定変換の順で強いことが確認できた．この尺度において，パワーは有意な結果が見られなかった．

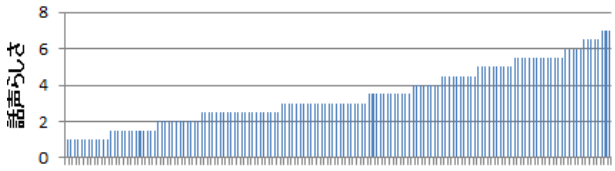

図 3: 評価結果を平均点順に並べたもの（話声らしさ）．

3.2 安定した－不安定な

安定さへの影響度は，F0 の歌声への近さ，音韻長の歌声への近さの順で強いことが確認できた．不安定さへの影響度は，F0 が歌声に近く，また音韻長が話声に近いほど強いことが確認できた．

3.3 感じの良い－感じの悪い

感じのよさへの影響度は，F0 が歌声に近いもの，音韻長が歌声に近いものの順で影響力を確認できた．また F0 と音韻長共に話声に近いものも感じのよいと評価された．感じの悪さへの影響度は，音韻長が歌声に近く，F0 が話声に近いものが最も影響力があることを確認できた．音韻長が固定長のものはどちらでもないと評価される傾向が見られた．この尺度においてパワーは有意な結果が見られなかった．

3.4 きれいな－汚い

きれいさへの影響度は，F0 が歌声に近いもの，音韻長と F0 が共通の割合で変換されたものの順で影響力が確認できた．汚さへの影響度は，F0 が歌声に近く音韻長が話声に近い，あるいは F0 が話声に近く音韻長が歌声に近いものが最も影響力が高かった．音韻長が固定長のものはどちらでもないと評価される傾向が見られた．この尺度において，パワーは有意な結果が見られなかった．

4. おわりに

本稿では，歌声－話声変換システムを用いて歌声らしさ，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．音響特徴量として F0，音韻長，パワーを着目し，各特徴量を網羅的に変化させて 150 通りの刺激音を作成し，それらを用いて聴取実験を行った．その結果，F0，音韻長が話声らしさに影響を与えていることが確認できた．今後は，入力とする曲のジャンルや速さ，また性差によってこの調査に与える影響を調べる必要があると考えている．

なお，本研究は科研費（S）No.24220006 の支援を受けた．

参考文献

[1] 阿曽慎平他：’SpeakBySinging：歌声を話声に変換する話声合成システム’，情報処理学会論文誌 Vol.2010-MUS-86 No.8(July 2010)

[2] 阿曽慎平他：’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’，情報処理学会論文誌，Vol.2011，No.1，pp.255-257(2011)

[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’，情報処理学会論文誌、vol.47， no.6，pp.1822-1830， June 2006

[4] 河原英紀他：’ 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，No7，pp.521-526(1998)

[5] Copyright (c) 1991-2009 京都大学 河原研究室，Copyright (c) 1997-2000 情報処理振興事業協会 (IPA)，Copyright (c) 2000-2005 奈良先端科学技術大学院大学 鹿野研究室，Copyright (c) 2005-2009 名古屋工業大学 Julius 開発チーム