歌声の歌詞認識における音高の影響について*

◎尾関弘尚，△鎌田貴幸，後藤真孝†，速水悟
岐阜大学　†産業技術総合研究所／科技団さきがけ研究21

１．はじめに

本研究では，楽曲中の歌声（ボーカル）の歌詞を自動認識することを目指し，その第一段階として，伴奏のない歌声単独（独唱）の音響信号を対象とした歌詞認識に取り組む．従来，歌詞付きの楽譜が事前に用意されたときに，楽譜中のどこを歌っているかを音高と歌詞に基づいて追跡し，自動伴奏する研究がなされてきた[1][2]．しかし，文献[1]の対象は母音に限定され，文献[2]も個々の母音と子音をモデル化していなかったため，歌詞の自動認識の目的には利用できなかった．

そこで本研究では，連続音声認識技術を歌声に適用して，歌詞の自動認識を実現していく．一般に歌声では，通常の音声と異なり，歌手による意図的な調音器官の制御によって，音高（基本周波数）や，音韻の継続時間が大きく変化する．そこで本稿では，まず，一般的な音声認識エンジンで歌声の歌詞認識をおこない，その結果を，同じ歌詞を読み上げた音声の認識結果と比較する．次に，歌声の歌詞認識における誤認識がどのような場合に起きているかを，基本周波数の観点と音韻の継続時間の観点から分析する．これらの要因の分析結果は，歌詞固有の認識手法の研究に役立つと考えられる．

２．実験方法

以下の手順で実験をおこなった．

1. 歌声と読み上げ音声のデータの用意

歌声の音響信号として，「RWC研究用音楽データベース：ポピュラー音楽」[3] に収録されている12曲 (RWC-MDB-P-2001 No.3, 4, 7, 11, 21, 27, 34, 37, 41, 44, 55, 74)の「歌のみ」(伴奏なし) のデータを使用した（歌手が単独でグループでなく，歌詞中に英語表現が比較的少ない12曲を選んだ)．12曲のうち男性歌手は7名，女性歌手は5名である．一方，これらと比較実験をする読み上げ音声として，各楽曲の歌詞のテキストを，普通に読み上げた音声と，意図的に高く裏声で読み上げた音声を新たに収録した．読み上げは，成人男性1名，成人女性1名がおこなった．

2. 個々の音声区間（フレーズ）への分割

上記の歌声と読み上げ音声の各データをフレーズに分割し，各フレーズを音声認識の対象とする．具体的には，音響信号のパワーを用いて，ある閾値よりも小さい無音区間が一定時間（70フレーム，フレームシフト5msec）連続する箇所で分割し，無音部分を除去して音声区間のみを切り出した．

3. 各フレーズに対する音声認識

各フレーズに対し，同一の音声認識エンジン，言語モデルを用いて音声認識する．音声認識エンジンには，CSRCの日本語ディクテーション基本ソフトウェアJulius3.3[4]を利用した．言語モデルと辞書には，上記の１２曲の歌詞のテキストを，奈良先端大のChasen-2.2.9[5]を使って形態素解析したものを利用した．

4. 認識性能（正解率と誤り率）の算出

音声認識結果を正解率と誤り率の二つの尺度で評価した．両者の計算方法を以下に示す．ここでは，形態素解析結果の各要素を単語とした．

$$\text{正解率}(\%) = \frac{(\text{正しく認識した単語数})}{(\text{元の歌詞の単語数})} \times 100$$

$$\text{誤り率}(\%) = \frac{(\text{誤り単語数})+(\text{脱落単語数})+(\text{挿入単語数})}{(\text{元の歌詞の単語数})} \times 100$$

正解率によって，歌詞テキストの単語を，どの程度正しく認識できたかがわかる．一方，誤り率によって，別の単語と誤認識した語数（誤り単語数）や，単語が抜けてしまった語数（脱落単語数），誤って単語が挿入された語数（挿入単語数）がどの程度あったかがわかる．

３．結果及び考察

歌声と読み上げ音声に対する認識性能を比較した後に，基本周波数や音韻の継続時間の違いによる性能の変化を調査する．

3.1 歌声と読み上げの違い

同一の歌詞内容を歌った場合（歌声）と読み上げた場合（読み上げ）の音声認識性能を調査した結果を図1に示す．これは，発声スタイルの違いが認識結果に及ぼす影響を示している．

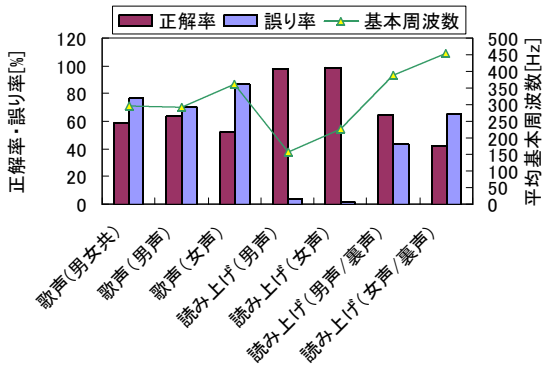

図1. 歌声と読み上げの認識性能の比較

「歌声（男声)」「歌声（女声)」の結果を，「読み上げ（男声)」「読み上げ（女声)」の結果と比較するとわかるように，歌声の場合には正解率が低下し，誤り率が大きくなることがわかる．この一つの要因として，歌声では多様な音高（特に高域）の音声が出現する

*The influence of vocal pitch on lyrics recognition of sung melodies, by Hironao Ozeki, Takayuki Kamata, Masataka Goto†, Satoru Hayamizu (Gifu University, †PRESTO, JST/AIST)

のに対し，読み上げ音声では音響モデル作成時に近い範囲の音高しか出現しないことが考えられる．そこで，同じ読み上げでも，意図的に高く裏声で読み上げた「読み上げ（男声/裏声）」「読み上げ（女声/裏声）」の結果と比較すると，普通に読み上げた場合より大きく性能が低下していることがわかる．このことは，高域の発声では音響モデルとの不一致が起きて，認識性能が低下していることを示唆している．

3.2 基本周波数の違いによる性能の変化

3.1節において高域における認識性能の低下が示唆されたが，実際に基本周波数が高くなると性能がどう低下していくかを調査する．そのための準備として，歌声（メロディー）の基本周波数（音高）を，文献[6]の実験用に開発された音高情報エディタを用いて，人間が手作業で10msecごとに指定した．これから，楽曲全体や，フレーズごとの平均基本周波数が求まる．

まず，各楽曲ごとの正解率と誤り率の結果を，楽曲全体の平均基本周波数と共に図2に示す．横軸の楽曲は，平均基本周波数の小さい順に並べた．これから，平均基本周波数の特に高い右の三曲（いずれも女声）では，他と比較して性能が低いことがわかる．

次に，全12曲を対象に，各周波数帯域（20 Hzごと）内の平均基本周波数を持つフレーズに関して，正解率と誤り率を平均した結果を図3に示す．これから，基本周波数が高くなるにつれて，実際に性能が低下していく傾向があることが確認された．

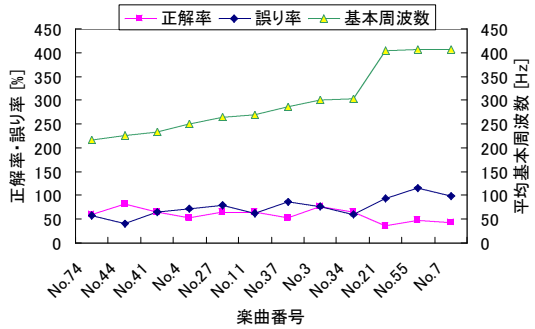

図2. 各楽曲ごとの平均基本周波数と正解率・誤り率

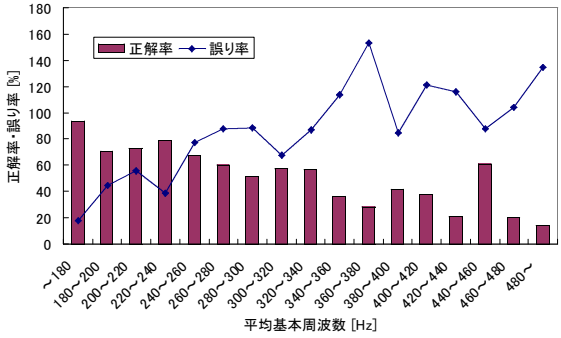

図3. 平均基本周波数の各帯域ごとの正解率と誤り率

3.3 音韻継続時間長の違いによる性能の変化

性能低下に関する基本周波数以外の要因として，歌声特有の長音化（音符に応じた音韻の引き延ばし）が考えられるため，実際にどの程度長音化すると性能が低下するかを調査する．そこで，4分音符以上の長さに引き延ばされた音を末尾以外に含む単語について，認識性能を求めた．該当する単語数は271個あるが，そのうち正しく認識した単語数（正解単語数）は92個で，正解率は33.95%であった．これは全単語を対象とした正解率58.76%と比較すると，かなり低い．

さらに，単語内での最長引き延ばし音の長さを横軸として，各長さのグループにおける正解単語数と誤り単語数の関係を図4に示す．引き延ばし音の長さは，テンポを用いて楽譜から算出した．これから，200～300msec程度引き延ばされる単語では正解単語数が多いのに比べ，400msec以上の引き延ばし音を含むと誤認識の割合が増加していることがわかる．これは長音化が認識性能の低下に影響していることを示している．

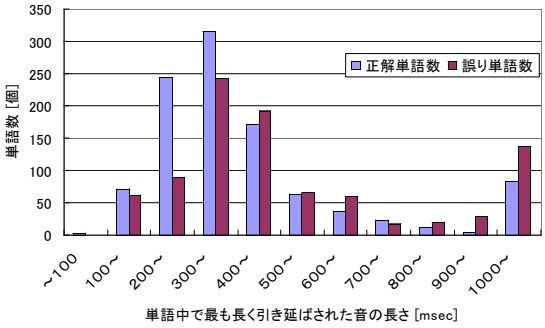

図4. 長音化の度合いに応じた正解単語数と誤り単語数

4. まとめ

本稿では，歌唱を歌詞認識する際に性能低下を招く要因として，高い音高（基本周波数）と長音化（音韻の引き延ばし）に着目し，具体的な性能低下を調査した結果を述べた．実際に，音高が高い場合や音韻の継続時間が長い場合に，正解率が低下することを確認した．これらが，通常の読み上げ音声の認識よりも，歌詞認識が難しい原因の一部となっていると考えられる．

今後は，音高や音韻の継続時間の変化に対応した認識手法の研究に取り組んでいく予定である．また，より多くの楽曲の調査やポピュラー音楽以外の楽曲の調査も検討している．

参考文献

[1]東,橋本:"音声認識とピッチ検出を併用した歌声の自動伴奏",情報処理学会 音楽情報科学研究会 研究報告 97-MUS-22-1, pp.1-5, 1997.

[2]L. Grubb, R. Dannenberg:"Enhanced Vocal Performance Tracking Using Multiple Information Sources",Proc.ICMC98,pp37-44, 1998.

[3]後藤,橋口,西村,岡:"RWC研究用音楽データベース:ポピュラー音楽データベースと著作権切れ音楽データベース",情報処理学会 音楽情報科学研究会 研究報告 2001-MUS-42-6, pp.35-42, 2001.

[4]河原,住吉,李他:"連続音声認識コンソーシアム 2001年度版ソフトウエアの概要",情報処理学会 音声言語情報処理研究会 研究報告 2002-SLP-43-3,2002.

[5]松本,北内,山下他:"日本語形態素解析システム『茶筌』 version 2.2.1 使用説明書", http://chasen.aist-nara.ac.jp/ , Dec.2000.

[6]後藤:"音楽音響信号を対象としたメロディとベースの音高推定",電子情報通信学会論文誌 D-II, Vol.J84-D-II,No.1, pp.12-22, Jan.2001.