歌声と朗読音声の自動識別のための基本周波数の時間構造のモデル化に関する検討 *

◎大石康智 (名大・情報科学), 後藤真孝 (産総研), 伊藤克亘, 武田一哉 (名大・情報科学)

1 はじめに

人間は，様々な発声のスタイルを上手に使い分けることによって複数の相手とのコミュニケーションを成り立たせている．そこで我々は，まず歌声と朗読音声に着目して，どこに発声のスタイルの違いがあるのかを自動識別実験を行って検証している．その識別性能の改善を図るために，本報告では，基本周波数 (以後，F0 と呼ぶ) の時間構造に着目する．従来よりも長時間のフレームごとに算出される ΔF0 に符号語を割り当て，ΔF0 の系列を離散的な符号語列で表現する．その並びを Ngram モデルで学習することにより，歌声，朗読音声の ΔF0 が時間的にどのように変化していくかをモデル化した．その結果，1s，2s の音声信号に対して，それぞれ識別率が 82.2%，89.5%であり，従来手法に対して 13.2%，27.6%の誤りが改善された．

2 局所的な F0 の時間変化による識別

2.1 局所的な F0 の時間変化に基づく識別尺度 [1]

歌声は楽曲のメロディとリズムパターンの制約を受けるため，F0 の遷移が音符に従った階段構造を形成する．一方，日本語の朗読音声の韻律は，下降する F0 の軌跡によって特徴づけられる．そこで，音声信号から抽出される F0 の時間変化の違いが識別の手がかりになると考え，後藤らの提案した F0 推定手法 [2] を利用して，F0 を 10ms ごとに推定し，連続点から計算される回帰係数を ΔF0 とした．各音声の ΔF0 の頻度分布を GMM で学習し，フレームごとにその事後確率を求め，平均事後確率の大きい方を識別結果とした．結果的に，Δ 算出の時間幅は 50ms と非常に短時間の場合が最適であり，局所的な F0 の変化を累積して観測することにより，2 つの音声の長時間の F0 の軌跡の違いを捉えることができたと考えられる [1]．

2.2 Random Splicing を施した音声信号の聴取実験 [3]

人間の音声の識別に影響する音響的特徴を調査するために，Random Splicing[4] を施した音声信号を利用して聴取実験を行った．この手法は，音声区間を短い断片に分割し，ランダムに接合することによって，韻律を変形させることが可能となる．

図 1 より，Random Splicing した朗読音声の識別率には大きな低下がみられないが，歌声の識別率は原音に比べて大幅に低下した．しかも，断片長を短くするにつれて，さらに人間の聴取能力は低下することがわかる．このことから，人間の音声の識別には，250ms 以上の音声信号に含まれる時間的に変化する特徴が非常に重要であることが確認できた．しかし，従来の ΔF0 による自動識別手法の性能と比較すると，図 1 では，原音の識別率よりも，20%以上識別率が低い．この原因としては，ΔF0 の時間幅が 50ms と非常に短く，その値が時間的にどのように変化していくかをモデル化していなかったためであると考えられる．次節でこれらを改善するための新しい手法を提案する．

3 F0 の時間構造のモデル化による歌声と朗読音声の自動識別

従来よりも長い区間における ΔF0 を計算し，その時系列を符号語列に変換することにより，時間構造をモデル化する手法を提案する．

*Fundamental Frequency Temporal Structure Modeling Approach for Discriminating the Singing and the Speaking Voices by Y. Ohishi (Nagoya Univ.), M. Goto (AIST), K. Itou, and K. Takeda (Nagoya Univ.)

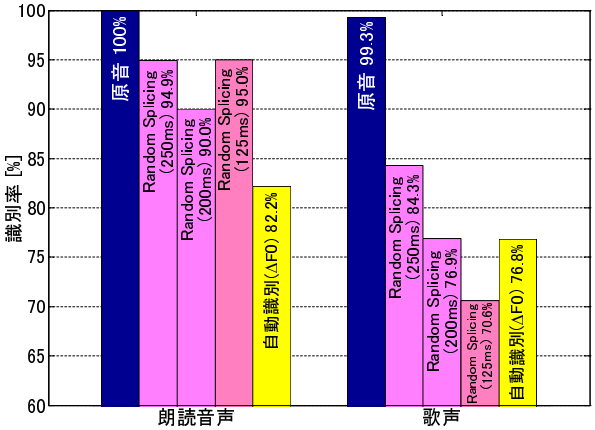

図 1: 1s の音声信号を Random Splicing したときの聴取実験結果と局所的な ΔF0 による自動識別結果との比較: 括弧は Random Splicing の分割した長さを指す

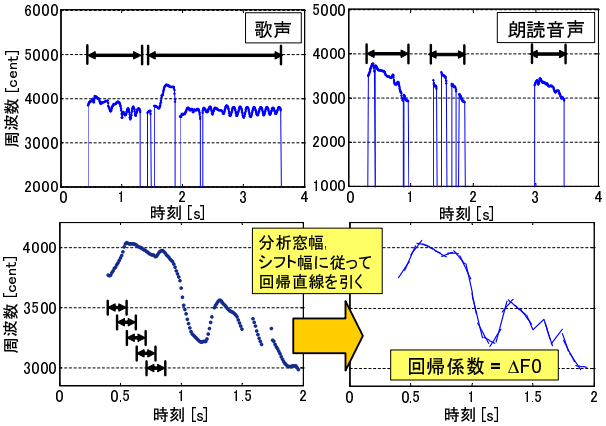

図 2: 発声区間の検出と ΔF0 の算出方法: 200ms 以上の休止を除いた区間で，分析窓幅，シフト幅に従って回帰係数を算出する

3.1 長時間の ΔF0 の算出方法

まず，図 2 の上図のように 200ms 以上 F0 が推定されていない部分をを休止区間とする．それ以外の矢印の区間を発声区間と考える．従来はこの発声区間で連続した点から ΔF0 を計算していたが，無声音の部分で F0 が不連続になり，ΔF0 を計算できない部分があった．本手法では，分析窓幅とシフト幅に従って，フレーム内の不連続に推定された F0 に対しても回帰直線を引き，回帰係数を ΔF0 とする (図 2 の下図)．休止区間では，回帰係数は算出しない．

3.2 ΔF0 の符号化

ΔF0 の相対頻度分布が図 3 である．また，各音声の頻度分布を足し合わせたものが，太線である．これを面積が等しくなるように M 等分し，符号語を割り当てる．図 3 の下図は，例として $M = 4$ の場合であり，3.1 節でフレームごとに算出された ΔF0 に対して A，B，C，D の 4 つの符号語を割り当てる．休止区間については符号語”R”を割り当て，結果的に ΔF0 の系列が，A，B，C，D，R の 5 つの符号語からなる符号語列に変換される．

3.3 符号語列の Ngram モデルによる学習

符号語列を Ngram モデルで学習する．Ngram モデルとは統計的言語モデルに広く使われもので，符号語列の i 番目の符号語 w_i の生成確率が，直前の $N-1$ 個の符号

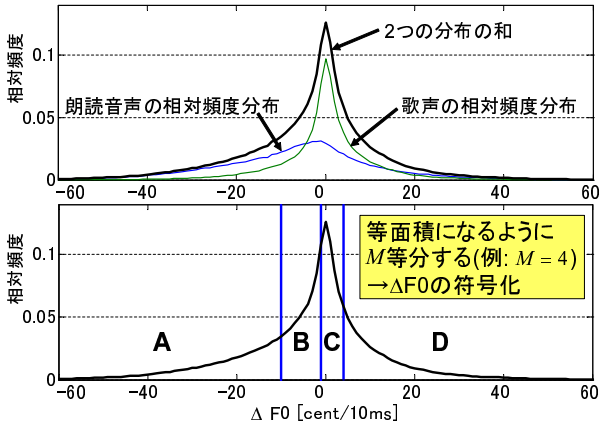

図 3: ΔF0 の相対頻度分布とその符号化方法

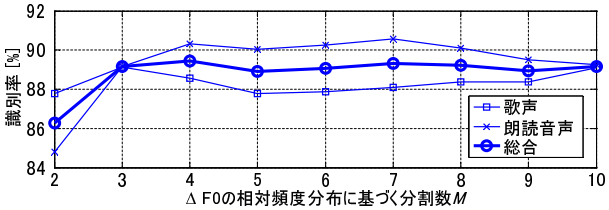

図 4: ΔF0 の相対頻度分布に基づく分割数 M を変化させたときの識別率 (3gram モデルを利用した場合)

語 $w_{i-N+1} \ldots w_{i-2} w_{i-1}$ だけに依存すると考える．すなわち ΔF0 の系列を表す $w_1 w_2 \ldots w_n$ の符号語列の生成確率 $P(w_1 w_2 \ldots w_n)$ の推定をする場合に，

$$P(w_1 w_2 \ldots w_n) = \prod_{i=1}^{n} P(w_i | w_{i-N+1} \ldots w_{i-1}) \qquad (1)$$

と近似を行うモデルである．これにより，歌声と朗読音声の ΔF0 の値がどのように変化していくかをモデル化できる．今回は $N = 1, 2, 3$ のモデルを作成し，識別方法は，従来手法と同様に事後確率の大きい方を識別結果とする．

4 評価実験

前節で提案した手法による識別性能を検証し，従来手法との比較を行う．提案手法の ΔF0 は，分析窓幅を 100ms，シフト幅を 50ms と実験的に決定して算出した．

4.1 歌声データベース

産業技術総合研究所 (AIST) によって収録された歌声研究用音楽データベース「AIST ハミングデータベース」[5] の一部である，日本人歌唱者 75 名分（男性 37 名，女性 38 名）の音声データを使用した．各歌唱者が，*"RWC Music Database: Popular Music"* から抜粋した合計 25 曲の歌の出だしの部分とサビの部分を歌い，またその歌詞を朗読した音声を用いた．つまり 1 名あたり計 100 サンプル (歌声: 50 サンプル，朗読音声: 50 サンプル) となり，75 名全員で 7500 サンプルとなる．話者，楽曲に対してオープンデータで評価するために，話者を 3 グループ，楽曲を 5 グループに分け，15 回のクロスバリデーションを行った．

4.2 実験結果

図 4 は，符号語の割り当てにおける分割数 M を変化させたときの 2s の音声信号の識別率である．3gram モデルを使用した．$M = 4$ のときがもっとも識別率が高く，総合して 89.5%であった．図 5 に，$M = 4$ のときの 1s，2s の音声信号の識別結果を示す．$N = 3$，すなわち 3gram モデルを使用したとき，総合的に 1s，2s の音声信号に対して識別率が 82.2%，89.5%であり，従来手法に対して誤り改善率が 13.2%，27.6%であった．

モデルの適合性を評価するためにテストセットパープレキシティを算出した (表 1)．パープレキシティが低いということは，テストセットの符号語の出現確率が高いということになり，認識したい符号語とそうでない符号語を峻別する能力が高いモデルであることを意味する．$N = 3$ のときテストセットパープレキシティは，歌声，朗読音声，それぞれ 1.48，1.51 となり，$N = 1, 2, 3$ の中で最も小さくなった．すなわち，$N = 1, 2, 3$ の中で 3gram モデルが最も識別率が高く，かつテストセットに適合したモデルであることが明らかとなった．また，$M = 4$ のときの 2gram モデルを確認したところ，A となる事後確率 ($P(A|\bigcirc)$) は朗読音声が大きく，C となる事後確率 ($P(C|\bigcirc)$) は，歌声が大きくなった．すなわち，F0 が朗読音声では下降する傾向と歌声では音符に従った平坦な変化になる傾向をモデル化できていると考えられる．

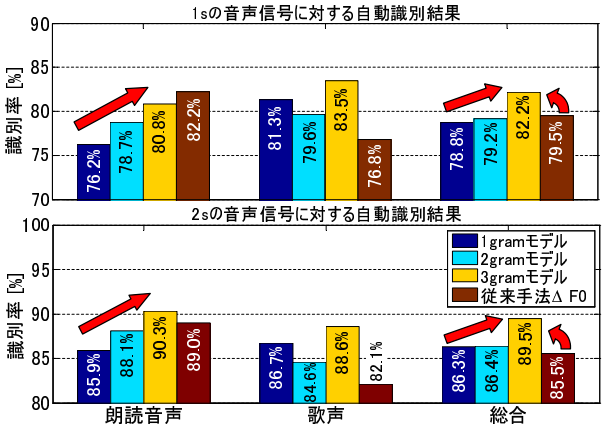

図 5: 評価実験結果

表 1: Ngram モデルに対するテストセットパープレキシティ

テストセット	歌声 (2s)		朗読音声 (2s)	
Ngram モデル	歌声	朗読音声	歌声	朗読音声
$N = 1$	1.54	1.65	1.67	1.57
$N = 2$	1.49	1.55	1.57	1.52
$N = 3$	1.48	1.54	1.58	1.51

5 まとめと今後の展開

本報告では，歌声と朗読音声の自動識別性能を向上させるために，聴取実験結果を分析し，長時間の F0 の時間構造をモデル化する手法を提案した．従来よりも長時間の ΔF0 を算出して，それを符号化し，Ngram モデルで学習した．これにより，歌声と朗読音声の F0 の時間構造の違いをモデル化することができた．2s の音声信号に対して，歌声と朗読音声の識別率は 89.5%で，従来手法に比べて 27.6%の誤りが改善された．今回，分布が等面積になるように符号語を割り当てたが，横軸に等間隔に割り当てたり，k-means でクラスタリングして符号語を割り当てる方法も考えられる．また，ΔF0 だけでなく，音声信号のパワーの時間変化特徴を利用することも今後検討していく．

参考文献

[1] 大石 康智, 後藤 真孝, 伊藤 克亘, 武田 一哉, "局所的・大局的な特徴を利用した歌声と朗読音声の識別," 情処研報音楽情報科学, Vol.2005, No.82, pp.1-6, 2005.

[2] 後藤 真孝, 伊藤 克亘, 速水 悟, "自然発話中の有声休止箇所のリアルタイム検出システム," 電子情報通信学会論文誌, Vol.J83-D-II, No.11, pp.2330-2340, 2000.

[3] 大石 康智, 後藤 真孝, 伊藤 克亘, 武田 一哉, "歌声と朗読音声の識別システム構築のための人間の識別能力の調査と考察," 日本音響学会 2005 年秋季研究発表会講演論文集, 2-7-10, pp.77-78, 2005.

[4] K. R. Scherer *et al.*, "Vocal cues to deception: A comparative channel approach," *Journal of Psycholinguistic Research*, Vol. 14, No. 4, pp. 409-425, 1985.

[5] 後藤 真孝, 西村 拓一, "AIST ハミングデータベース: 歌声研究用音楽データベース," 情処研報音楽情報科学, Vol.2005, No.82, pp.7-12, 2005.