6U-9

アクセント特徴量を用いた歌声と朗読音声の識別システム

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

歌声と話声の自動識別技術は，音声メディア処理の拡大に重要である．人は，音声を歌声と話声を聞き分け，それに応じた応答をする．計算機システムも同様のことができれば，様々な可能性が広がる．例えば，ユーザが計算機に対し，明日の天気等を尋ねれば計算機は情報を返し，ユーザが歌を歌うと計算機はそのカラオケトラックを再生する，という機能が実現できる可能性がある．従来の歌声と話声の自動識別研究は，ノイズの無いクリーンな歌声と朗読音声を用いて識別問題を考えている．

より現実的な場面に沿った自動識別システムの性能を知るために，雑音環境での識別を実現する必要がある．エアコンの動作音等の環境雑音が混入すると，識別精度が低下することが予想される．従来研究 [1, 2] では，識別に，アクセントピーク間隔（以後ピーク間隔と略記），メル周波数ケプストラム係数（MFCC），ΔMFCC，基本周波数（F0）が有効であるとされている．本稿では，これらの特徴が，ホワイトノイズ環境でも識別に有効であるかどうかを検討する．雑音を含む歌声と朗読音声の識別を行い，単独の特徴量で構成された識別器と，複数特徴を重み付け和により統合した識別器 [1] の識別精度を比較する．

2. 識別手法

阿曽らの手法を用いる [1]．識別に用いる特徴量は，ピーク間隔・MFCC・ΔMFCC・F0 の 4 つである．特徴量毎に，その分布から単独特徴量識別器と呼ばれる，歌声と話声の尤度差を出力（値が大きいほど歌声らしい）する識別器を構成する．単独特徴量識別器の重み付けを和を取ることで，最終的な尤度差（重み付け統合法）とする．

2.1 識別特徴量

先行研究において，人間は 1 秒を聴取すれば 99.7%の精度で歌声か朗読音声か識別できると報告されており [2]，またその識別には音素継続時間・短時間のスペクトル・韻律の特徴が手がかりとなっていることが確認されている [2, 3]．この結果を基に基づき，以下を識別特徴量に用いる [1].

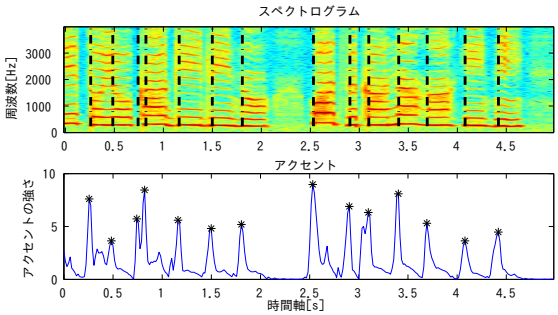

図 1: 入力音声のスペクトログラムを表示したもの（上図）とアクセント抽出結果（下図）．下図*印はピーク，上図破線は，ピーク位置に合わせて引いた補助線．発声開始や音素が変わる時刻でピークとなることがわかる．

アクセントピーク間隔　Klapuri らの手法を用いてアクセントを抽出し [4]，そのピーク間隔をピークピッキング法を適用して得られたピーク時刻に対し，隣り合うピーク間隔を算出することで求められる．アクセントとピークの抽出例を図 1 に示す．

このピーク間隔は，音響的な変化が顕著な箇所を境界として区切った区間の長さであり，音素継続時間と必ずしも同一ではないが関連した値を持つ．例えば，音素継続時間が短ければ，ピーク間隔も短くなる．音素継続時間に関連した特徴量を用いた自動識別手法については，従来十分に調査されてこなかった。

ピーク間隔は他の特徴量のようにフレーム（時間）毎に値が求められるものではなく，ピークが検出される度にその間隔の値が求まる．そのため，他の特徴量がフレーム数分得られるのに対し，ピーク間隔の特徴量が得られる個数は少なくなるが，個数にかかわらず得られた特徴量の分布を学習して識別器を構成する．

F0・MFCC　韻律に関連する特徴量として F0 を用いる [1, 2]．F0 は，yegnanarayana の提案した手法を利用して 10 ミリ秒ごとに推定した [5]．短時間のスペクトル特徴として MFCC・ΔMFCC を 10 ミリ秒ごとに抽出して利用した [1, 2].

2.2 単独特徴量に基づく識別器

単独特徴量識別機の構成には，阿曽らの手法 [1] を用いる．まず特徴量の分布を 16 混合ガウス分布（GMM）でモデル化し，歌声・朗読音声それぞれでパラメータを学習することで，歌声・朗読音声それぞれの尤度を出す GMM を構築する．次に平均歌声尤度から平均朗読音声

An Automatic Discrimination System between Singing and Speaking Voices Using accent feature. *Shinpei Aso (Kyoto Univ.), Takeshi Saitou (Kanazawa Univ.), Masataka Goto (AIST), Katsutoshi Itoyama, Toru Takahashi, Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto Univ.)*

尤度を減算した尤度差を出力するように構成することで得られる．つまり，この識別器の出力は，歌声尤度のほうが高い場合は正の値，朗読音声尤度のほうが高い場合は負の値である．精度評価のための2値判定は符号の正負（但し，0の時は朗読音声とする）により決定する．

2.3 識別器の重み付け統合法

各単独特徴量識別器から出力される尤度差に対して重み付けをした和を，統合尤度差とする [1]．最終的な識別結果は，単独特徴量識別器同様，符号の正負により決定する．重みは学習データから事前に決定する．すなわち，識別器の学習に用いるデータに対する識別精度が高くなるように選ぶ．各識別器に対する重みを0から10までの11段階で変えながら（11^M 通り，M は単独特徴量識別器の数），最も識別精度（2値判定したときの正解精度）が高くなる組み合わせを選び，総和が1となるよう正規化して使用する．

3. 評価実験

各単独特徴量識別器・重み付け統合法それぞれの，雑音を含む歌声と朗読音声の識別性能を比較評価する．手法毎に，3.1節のデータベースに雑音信号としてホワイトノイズを加えたものを利用してクロスバリデーションにより学習・識別を行い，2値識別結果（歌声または朗読音声）と正解ラベルを比較して別精度を算出する．手法一覧を表1に示す．

3.1 評価用音声データベース

評価実験には研究用音楽データベース「AIST ハミングデータベース」[6] 中の，日本人による歌声（3750音）と歌詞朗読音声（3750音）を使用する．内訳は男性が37名，女性が38名で，「RWC 研究用音楽データベース: ポピュラー音楽」25曲の出だしとサビの部分を歌った音声と歌詞を朗読した音声である．音声の平均長は，歌声が12秒，朗読音声が7秒である．

各音声を発声開始時刻から一定時刻（1秒）切り出し，ホワイトノイズを加えたものを用いて，話者を3グループ，楽曲を5グループに分けた15回のクロスバリデーションで評価を行う．SNR を 20dB（雑音小）から −15dB（雑音大）までの8段階の範囲で変化させながら，識別精度の変化を比較する．

3.2 実験結果

図2から，重み付け統合法による識別は単独特徴量識別器に比べ，雑音を含む歌声・朗読音声の識別精度が高くなることが示された．図2重み付け統合法が常に最も高い精度となっていることがわかる．

特徴量毎の比較を行うと，雑音が小さい時は ΔMFCC，雑音が大きい時は MFCC による識別精度が高いことがわかる．単独特徴量識別器の精度を比較すると，SNR が

表 1: 手法一覧．呼称は実験結果図に使用する．

呼称	識別方法説明
ピーク間隔	ピーク間隔に基づく単独特徴量識別器
ΔF0	ΔF0 に基づく単独特徴量識別器
MFCC	MFCC に基づく単独特徴量識別器
ΔMFCC	ΔMFCC に基づく単独特徴量識別器
本統合法	上記4つの識別器の重み付け和

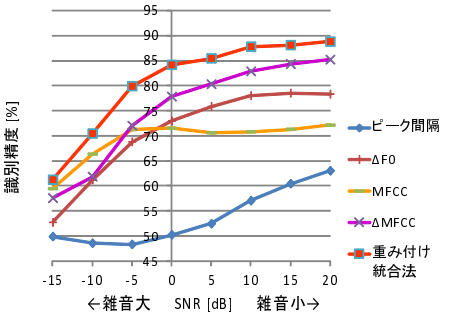

図 2: SNR と識別精度の関係図. 重み付け統合法 [1] が最も高い精度となっていることが分かる．

−5以上の時は ΔMFCC が，それ未満の時は MFCC の精度が最も高くなっている．

SNR が −5・0 の音声に対する識別精度はそれぞれ，アクセントピーク間隔が 48.3%・50.2%・，ΔF0 が 68.8%・73.0%，MFCC が 71.3%・71.5%，ΔMFCC が 72.0%・77.8%，重み付け統合法が 79.9%・84.2%であった．

4. おわりに

本稿では雑音を含む歌声と朗読音声の識別手法の精度比較を目的として，従来提案された歌声と朗読音声の自動識別手法を用いて，雑音信号としてホワイトノイズを含む音声に対する識別実験を行い，その精度を比較した。その結果，SNR が0で，長さが1秒の音声に対し，阿曽らの重み付け統合法は 84.2%となり，従来提案された識別特徴量を単独で用いた場合に比較し，6.4%高い結果となった．今後は，より現実的な場面を想定して，様々な雑音を含む自動識別にも取り組む必要がある．

謝辞本研究の一部は科研費, CrestMuse，GCOE の支援を受けた．

参考文献

[1] 阿曽他：スペクトル変化量のピーク間隔・F0・MFCC を用いた歌声と朗読音声の自動識別システム，情処全大，2R-U, 2011.
[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] 阿曽他：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，情処全大，2R-U, 2011.
[4] A.P. Klapuri *et al.*: Analysis of the meter of acoustic musical signals, *IEEE TASLP*, Vol. 14, No.1, pp. 342–355, 2006.
[5] B. Yegnanarayana *et al.*: Analysis of stop consonants in Indian languages using excitation source information in speech signal, *ISCA ITRW SPKD-2008*, 2008
[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.