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 *

◎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

我々は歌声と朗読音声の自動識別システムの構築を進めている [1] が, その性能を評価して改善を図る上で, 人間が音声信号のどのような特徴を手がかりとして識別しているかを知ることは重要である. そこで, 本研究では聴取実験を実施し, その結果とシステムの性能との比較を通じて考察を行う.

2 歌声と朗読音声の人間の識別能力の調査

2.1 歌声データベース

本研究では, 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」[2] 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"* (RWC-MDB-P-2001) [3] 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. 音声サンプルの長さの平均は歌声で約 8s 程度, 朗読音声で約 5s 程度であった.

2.2 識別に必要な音声信号長の調査

歌声データベースから女性 25 名, 男性 25 名を選び, 歌声 2500 サンプル, 朗読音声 2500 サンプルを発声開始から 10 段階の異なる長さで切り出したもの 50000 サンプルを利用する. この中から, 音声信号 430 サンプルを表 1 のように切り出した長さごとにランダムに選んだ評価セットを 10 種類作る. 10 人の被験者ごとに異なる評価セットを割り当て, その全サンプルをランダムな順番で 1 回だけ聴取させ, "歌声", "朗読音声", もしくは"識別不可能"かの 3 通りで回答させた.

図 1 より, およそ発声開始から 1s 程度の聴取により, 人間は識別が可能であることがわかる. 200ms の時点で既に正答率は 70.0%を超えており, 特に短時間の場合, 朗読音声の正答率が高い. これは音声信号のリズムやメロディ, イントネーションというような大局的な特徴だけでなく, スペクトル包絡のような局所的な特徴も, 識別の手がかりとしているのではないかと考えられる.

2.3 識別に必要な音声信号の特徴の調査

2.2 節では 1s の音声信号に対して人間は 99.7%で識別が可能であることを確認した. そこで本節では識別の手がかりとなる特徴を検討するために, 1s の音声信号の言語, 非言語情報をマスクするように加工した 2 種類の音声信号を用いて聴取実験を行う.

Random Splicing 手法 [4, 5] 音声区間をある長さの断片に分割し, ランダムに接合することによって, 非言語情報のうちイントネーション, テンポ, リズムをマスクする手法である. 従来は音声信号に含まれる感情, 個人性を調査するために利用されていた. 今回は, 歌声のメロディとリズム, 朗読音声のイントネーション, またそれぞれのテンポは失われるが, 音色 (響き) は保持される音声信号を聴取したときの識別能力の調査を行う.

Filtering 手法 低域通過フィルタにより音声信号の高調波成分を除去し, 音色, 音質を低下させる. カットオフ周波数は 800Hz とした. すなわち, 基本周波数とその 2 倍音, 3 倍音程度までを含む音声信号を聴取したときの識別能力の調査を行う.

*Investigation of Human Performance for Building a System Discriminating between Singing and Speaking Voices. by Y. Ohishi (Nagoya Univ.), M. Goto (AIST), K. Itou, and K. Takeda (Nagoya Univ.)

表 1: 評価セットの構成

時間長	歌声	朗読音声
100, 150, 200, 250, 500, 750, 1000ms	25 サンプル	25 サンプル
1250ms	20 サンプル	20 サンプル
1500, 2000ms	10 サンプル	10 サンプル
合計	215 サンプル	215 サンプル

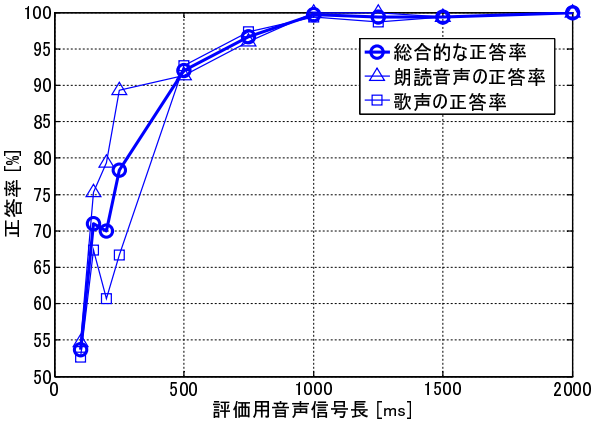

図 1: 歌声と朗読音声を人間が聴取して判断する場合の識別率

女性 25 名, 男性 25 名を選び, 発声開始から 1s の音声信号に対して 125ms, 200ms, 250ms の分割長で Random Splicing した音声 15000 サンプル, Filtering した音声 5000 サンプルを用意した. この加工した 2 つの音声サンプル群の中から, 表 2 のようにそれぞれ 200 サンプルをランダムに選んだ評価セットを 10 種類作る. 10 人の被験者ごとに異なる評価セットを割り当て, その全サンプルをランダムな順番で 1 回だけ聴取させ, "歌声", "朗読音声"の 2 通りで回答させた. また, その回答の確信度を 5〜1 で評価させた. つまり回答に自信があればを, 自信がなければ 1 を評定することになる.

3 聴取実験の結果と考察

図 2 は横軸を朗読音声, 歌声としたときのそれぞれの評価音声信号に対する正答率を示す. 一方, 図 3 は横軸を音声サンプルの性別としたときの正答率を示す.

3.1 Random Splicing による聴取実験の考察

図 2 より 1s の原音の朗読音声, 歌声の正答率は 100%, 99.3%に対して, Random Splicing し, さらにその分割長を短くするにつれて正答率は低下した. 特に歌声の正答率の低下は著しく, 分割長 125ms の場合, 正答率は 70.6%であった. このとき正答, 誤答の平均確信度はそれぞれ 3.66, 2.70 であった. 一方で分割長 125ms の朗読音声の正答率

表 2: 加工した音声の評価セットの構成

Random Splicing 手法		
分割する長さ	歌声	朗読音声
125ms	40 サンプル	40 サンプル
200ms	40 サンプル	40 サンプル
250ms	20 サンプル	20 サンプル
合計	100 サンプル	100 サンプル

Filtering 手法		
	歌声	朗読音声
合計	100 サンプル	100 サンプル

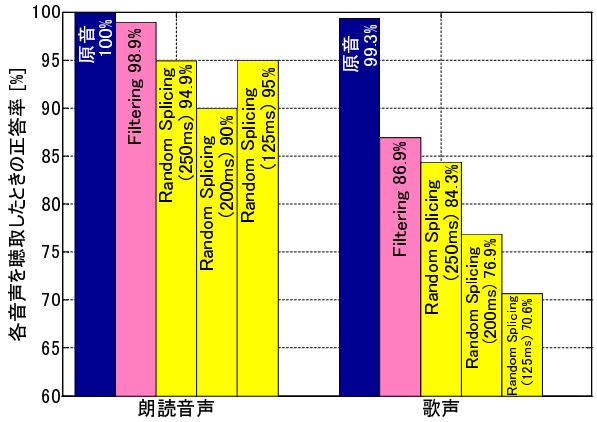

図 2: Random Splicing，Filtering した音声信号の正答率

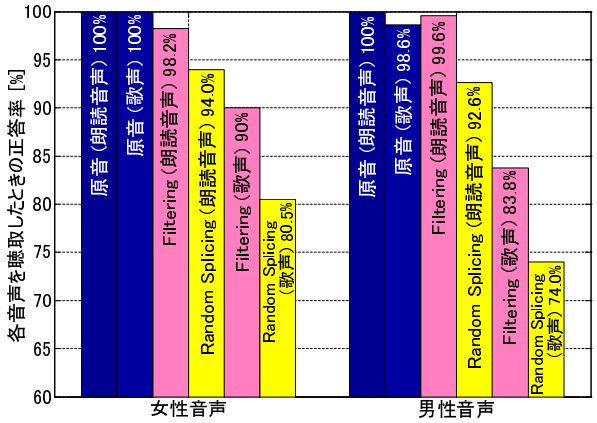

図 3: 性別ごとの音声信号の正答率

は 95%であり，正答，誤答の平均確信度はそれぞれ 4.06, 2.92 であった．ともに歌声の確信度を上回った．図 3 より Random Splicing した女性の歌声は 80.5%(分割長 125, 200，250ms の正答率の平均値) に対して，男性は 74.0%であり，6.5%の差が見られる．実験後の被験者の感想によると，「歌声の伸ばす発声に着目」「声の大きさの変動が大きければ歌声」「女性音声の方が朗読音声と歌声の音高差が大きく識別しやすい」「音声信号内の F0 の変動が大きければ歌声」という意見を得た．

以上より，Random Splicing 手法は F0 の軌跡，テンポ，リズムをマスクするが，局所的な F0 の変動を観測することにより識別可能である．しかし，分割長が 200ms，125ms と短くなるにつれて，母音を伸ばす発声が破壊されてしまうため，朗読音声と誤識別してしまう．また，女性の朗読音声と歌声は音色の違いにより，男性よりも識別が容易である．これは女性の裏声による歌い方と男性の歌い方の差であると考えられる．

3.2 Filtering による聴取実験の考察

図 2 より Filtering した朗読音声の正答率は 98.9%である．一方，歌声の正答率は 86.9%であり，12.0%の差が生じた．このとき歌声の正答，誤答の平均確信度はそれぞれ 3.58，2.88 であった．一方，朗読音声の正答，誤答の平均確信度はそれぞれお 4.56，4.07 であり，歌声を上回った．図 3 より男女ともに朗読音声の正答率は 98%を越える一方で，男性の歌声の正答率は 83.8%であった．実験後の被験者の感想によると，「テンポ，発声速度，リズムの違いに着目」「音高が持続する箇所があれば歌声」「イントネーションの違いに着目」という意見を得た．

以上より音色，音質をマスクしても，朗読音声はイントネーションやテンポの違いから識別が可能である．しかし，歌声の識別は容易でなく，その高調波構造も識別に必要なのではないかと考えられる．

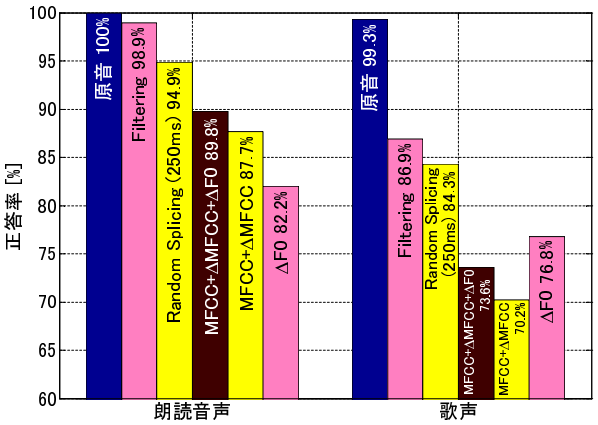

図 4: 自動識別システムによる正答率との比較

4 自動識別システムの性能との比較

聴取結果と我々の提案する歌声と朗読音声の自動識別システムによる 1s の音声信号に対する識別結果 [1] との比較を行う．音声の局所的な特徴量として MFCC と ΔMFCC，大局的な特徴量として ΔF0 を利用した．Δ 算出の幅は 50ms である．識別器は GMM を利用した．図 4 より，特に歌声の正答率の差が著しい．Filtering，Random Splicing した歌声の正答率は，それぞれ 87.9%，84.3%であるのに対して自動識別手法はすべて 80%を下回っている．これは，現在利用している特徴量が，歌声と朗読音声の音色，イントネーションの違いを的確に抽出できていないためであると考えられる．すなわち，聴取実験結果を踏まえて，周波数の高域におけるスペクトル包絡の違い，250ms 以上の F0 の時間変化構造の算出手法について再検討する必要がある．また，Random Splicing により，マスクされたテンポ (発声速度)，リズムを識別に利用することについても検討する必要がある．

5 まとめと今後の展開

本報告では，人間の音声信号の識別能力を調査するために，聴取実験を実施し，その結果とシステムの性能との比較を行った．その結果，人間は 250ms, 1s の音声信号に対して，それぞれ 78.3%, 99.7%で識別が可能であることを確認した．次に，言語，非言語情報をマスクするために 2 種類の加工を 1s の音声信号に施した．一つは Random Splicing 手法により F0 の軌跡，テンポ，リズムをマスクした音声信号，もう一つは Filtering をすることにより音色をマスクした音声信号である．これらを聴取したところ，特に歌声の正答率が約 15%程度低下した．このことから歌声の音色，F0 の軌跡，テンポ，リズム，それぞれが相補的に識別の手がかりになると考えられる．しかし，人間の聴取能力と自動識別システムの正答率の差は 20%以上である．今後は音声波形の時間領域に着目し，振幅，パワー，発声速度による識別を検討する予定である．

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, 2005.

[2] 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, 2005.

[3] 後藤真孝, 橋口博樹, 西村拓一, 岡隆一, “RWC 研究用音楽データベース: 研究目的で利用可能な著作権処理済み楽曲・楽器音データベース,” 情報処理学会論文誌, Vol.45, No.3, pp.728-738, 2004.

[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] M. Friend and M. J. Farrar, “A comparison of content-masking procedures for obtaining judgments of discrete affective states,” *Journal of Acoustical Society of America*, Vol. 96, No. 3, pp. 1283-1290, 1996.