変動長スペクトル特徴量を用いた朗読音声と歌声の判別*

◎橘秀幸, 小野順貴, 嵯峨山茂樹 (東大院・情報理工)

1 はじめに

ミュージカルや音楽番組の分析，音声対話システムのための基礎技術として，朗読音声と歌声の判別技術が重要である．しかし，これは従来多く研究されてきた音声と音楽の識別（例えば文献 [1]）と比較すると困難な問題である．その理由として，音声と音楽の識別では，楽器音の様々な性質を利用できるのに対し，音声と歌声の識別では，両者がいずれも人の声であり，音声と楽器音の間ほど明確な違いが見出されにくいという点が挙げられる．大石ら [2] はこの観点から歌声と朗読音声の違いについて検討し，音色と基本周波数の情報を利用することで，2 秒の朗読音声と歌声を 87%程度の精度で識別する手法を提案している．

本稿では，朗読音声/歌声判別の性能のさらなる向上を目的とし，音色や基本周波数とは独立していると考えられる新しい特徴量「変動長スペクトル」を利用した識別手法を提案する．変動長スペクトルは著者らがこれまでに提案した信号の新しい表現方法であり，信号のパラメータとして新しく「変動長」を導入し，各変動長成分のパワーにより信号の特徴を表現する手法である [3, 4]．朗読音声と歌声では変動長スペクトルの形状が異なっていると考えられるため，変動長スペクトルは両者の判別の特徴量として有効であると期待することができる．

2 変動長スペクトル

本節では，信号を変動長ごとに分離・表示する手法である変動長スペクトル，およびそれに関するいくつかの用語を定義する．まず,「変動長」を次のように定義する．

定義 1 (変動長) 信号をフレーム分析する際に，その信号が非定常的・広帯域な信号と見なされうるような最小のフレーム長 l を，その信号の変動長と呼ぶ．

例えば，打楽器音は瞬間的な信号であり，非常に短いフレーム長で分析してもほとんど非定常的である．すなわち，打楽器は変動長が非常に短い信号である．逆に，長時間持続するような定常的な正弦波は，変動長が非常に長い信号である．これら両極端の信号の中間に位置すると考えられるのが朗読音声や歌声などゆらぎがある信号である．これは両者ともに，ごく短時間で見れば定常的と見なせる一方，長時間で見ると非定常的であるためである．すなわち，朗読音声や歌声は変動長が中程度の信号である．

ところで，著者らはこれまでに，信号のスペクトログラムが時間方向と周波数方向のいずれの方向に滑らかであるかに着目した信号分離手法として**調波打楽器音分離 (HPSS)**[5] を提案している．これは，信号 $s(t)$ が与えられたとき，ある長さ l のフレームを用いてスペクトログラム $\mathbf{S}_l = \mathrm{STFT}_l[s(t)]$ を求め，そのスペクトログラムを時間方向に滑らかな成分 $\mathbf{H}_l = \mathrm{STFT}_l[h_l(t)]$ と周波数方向に滑らかな成分 $\mathbf{P}_l = \mathrm{STFT}_l[p_l(t)]$ に分離することで，信号を $s(t) = h_l(t) + p_l(t)$ と分解する手法である．

HPSS では概ね，変動長が l 以上の成分は $h_l(t)$ に，l 以下の成分は $p_l(t)$ に分離されると考えられる．これは，前者については，変動長が l 以上の成分は，l 程度のフレーム長では定常的・狭帯域な信号と見なされ，スペクトログラム上では時間方向に滑らかに表現されるためである．また後者も同様に，変動長が l 以下の成分は，l 程度のフレーム長では非定常・広帯域な信号と見なされ，スペクトログラム上では周波数方向に滑らかに表現されるためである．

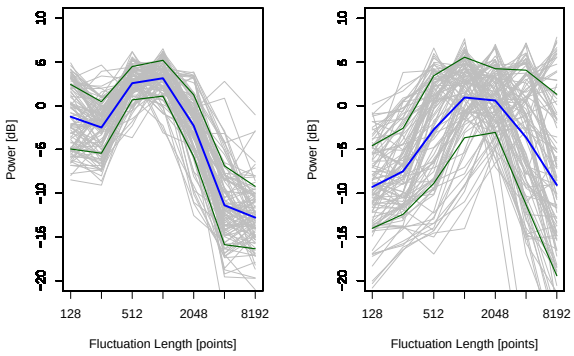

Fig. 1 (左) 朗読音声の変動長スペクトル，(右) 歌声の変動長スペクトル．それぞれ 100 サンプルを重ねてプロット（灰色の線）．青線は平均，緑線は平均 ± 標準偏差．

HPSS のこのような性質を利用すると，複数のフレーム長 l_1, l_2 を利用することにより，

$$s(t) = \overbrace{\underbrace{\alpha(t)}_{h_{l_2}(t)} + \beta(t)}^{h_{l_1}(t)} + \overbrace{\gamma(t)}^{p_{l_1}(t)}_{p_{l_2}(t)} \tag{1}$$

というように信号を 3 つ以上の成分に分解することができる．このときの中間的な成分 $\beta(t) = p_{l_2}(t) - p_{l_1}(t)$ は，変動長が l_1 から l_2 程度の間にある成分にほぼ相当していると考えられる．これを一般化すると，次のような信号の分解方法を考えることができる．

定義 2 (多重 HPSS) $\mathbf{S}_{l_k}$ を入力信号 $s(t)$ をフレーム長 $l_k = l_0 \times 2^k, (1 \le k \le n)$ で分析して得たスペクトログラムとする．ただし l_0 は適当な値である．それぞれのスペクトログラム上にて HPSS を適用することにより得た n 個の信号 $\{p_{l_k}(t)\}_{1 \le k \le n}$ を用いて次のような新しい $n+1$ 個の信号を得る処理を多重 HPSS と呼ぶ．

$$\tilde{s}_{l_k}(t) = \begin{cases} p_{l_1}(t) & k = 1 \\ p_{l_k}(t) - p_{l_{k-1}}(t) & 1 < k \le n \\ s(t) - p_{l_n}(t) & k = n+1. \end{cases} \tag{2}$$

多重 HPSS により得られた各成分 $\tilde{s}_{l_k}(t)$ は，それぞれ変動長が概ね l_{k-1} から l_k 程度までの成分で占められていると考えられる．従って，各成分のパワーを比較することにより，原信号 $s(t)$ はどの変動長の成分が強いかを調べることができると考えられる．これを踏まえ，変動長スペクトルを次のように定義する．

定義 3 (変動長スペクトル) 多重 HPSS により得られた各成分 $\tilde{s}_{l_k}(t)$ のパワーを配列したベクトル $\boldsymbol{S} = [S_{l_1} \cdots S_{l_{n+1}}]^\mathsf{T}$, where $S_{l_k} = \sum_t \tilde{s}_{l_k}(t)^2$ を変動長スペクトルと呼ぶ．

---

*Hideyuki TACHIBANA, Nobutaka ONO, Shigeki SAGAYAMA, (Graduate School of Information Science and Technology, The University of Tokyo) “Discrimination of Speech and Singing Voice Based on Fluctuation-Length Spectral Features.”

信号の変動長スペクトルの例を Fig. 1 に示す. Fig. 1 は，サンプリング周波数 8kHz, $l_0 = 64$ 点, $n = 6$ の条件で，朗読音声および歌声の変動長スペクトルを求め，正規化してから，100 サンプルずつ重ねてプロットしたものである．朗読音声と歌声の変動長スペクトルを比較すると，朗読音声の変動長スペクトルは，歌声の変動長スペクトルに比べて変動長 128 点 (l_1) 成分が大きく，変動長 8192 点 (l_7) 成分が小さい傾向にあり，変動長 512 点 (l_3), 1024 点 (l_4) 付近に強いパワーが現れやすいことが観察できる．一方，歌声の変動長スペクトルは，概ね朗読音声よりも長い変動長 1024 点 (l_4), 2048 点 (l_5) 付近に強いパワーが現れやすく，変動長 128 点 (l_1) 成分が小さい傾向にあることが観察できる．

Fig. 1 に見られるように，朗読音声と歌声の変動長スペクトルの形状が明確に異なっていることから，簡単な識別器により両者を識別することが可能であることが期待される．

3 朗読音声と歌声の判別実験

3.1 実験条件

変動長スペクトルが朗読音声と歌声を識別する際の特徴量として有効であることを検証するため，多数の朗読音声と歌声データに関してそれぞれ変動長スペクトルを求め，それを特徴量として両者を識別する実験を行った．変動長スペクトルはサンプリング周波数 8kHz, $l_0 = 64$ 点，$n = 6$ の条件で求め，それを正規化し，対数値をとったものを使用した．識別器には，簡単な識別器である (1) マハラノビス距離に基づく判別分析，および，強力な識別器として知られる (2) AdaBoost のそれぞれを使用した．AdaBoost の弱識別器には，ある程度の識別能力を有し且つ計算コストが小さい識別器として，朗読音声と歌声の特徴量ベクトルの重心間の垂直二等分面を識別面とするような線形判別器を用いた，なお，本実験では比較対象として，12 次 MFCC + log energy $+\Delta+\Delta\Delta$ の計 39 次元の時間平均を特徴量として使用した場合の性能，および変動長スペクトルと MFCC を組み合わせて使用した場合の性能も評価した．これらの条件で，10-fold の交差検定を行い，朗読音声/歌声判別の正解率を求めた．

3.2 実験データ

実験に用いた歌声と朗読音声データは以下に述べる通りである．歌声データには，RWC データベース [6] の楽器音データベースより歌声 (RWC-MDB-I-2001 No. 45 –50)，市販の模範歌唱付きカラオケデータ[1]からスペクトル減算法により抽出された擬似的なアカペラ，および MIR-1K データベース[2]の歌声のトラックを使用した．これらには，クラシック音楽とポピュラー音楽の両方の歌唱法のデータが含まれており，少数だがラップも含まれている．それぞれのデータは 2 秒ずつに区切って使用した．また，音量が一定値に満たないデータは，無音区間が占める割合が長いデータであると見なし，実験には不適であるため自動的に除去した．これにより得られた歌声データは，全 18715 ファイル（10 時間 23 分 50 秒）である．

朗読音声データには，新聞記事読み上げ音声コーパス (JNAS)[7] を使用した．朗読音声も歌声と同様に 2 秒ずつに区切り，実験に不適なデータを自動的に除去した．これにより得られた朗読音声データのファイル数は 8 万個強で，実験ではその中から歌声と同数の 18715 個のファイルを無作為に抽出して使用した．

3.3 実験結果・考察

実験により得られた歌声/朗読音声判別の正解率を Table 1 に示す．特徴量に変動長スペクトルを用い，識別器に AdaBoost を用いた場合，朗読音声/歌声判別の正解率は 86.7%であった．この値は，MFCC を特徴量とした場合の正解率 78.9%と比較して大きな値であり，本特徴量の有効性を示していると考えられる．

Table 1 朗読音声/歌声判別の正解率

特徴量	識別器	
	判別分析	AdaBoost
変動長スペクトル (7 次元)	**84.7%**	86.7%
MFCC + log energy + Δ + ΔΔ (39 次元)	51.9%	78.9%
変動長スペクトル (7 次元) + MFCC (39 次元)	56.6%	**88.9%**

また，マハラノビス距離に基づく判別分析の場合も 85%程度の正解率が得られた．この結果は，変動長スペクトルの特徴量空間上において，朗読音声，および歌声がそれぞれ特定の領域に集中して分布していることを示唆していると考えられる．このため，他の特徴量を用いた判別手法に比較的容易に変動長スペクトル特徴量を組み合わせることが可能であること，また，この組み合わせにより性能が改善されることを期待することができると考えられる．実際に本実験においても，MFCC のみを特徴量とした場合の AdaBoost による判別の正解率が 78.9%であったのに対し，変動長スペクトルを組み合わせたことによって性能が 10.0 ポイント改善し，88.9%の正解率を示した．

4 まとめ・今後の展望

本稿では，著者らが音響信号の新しい特徴量として検討している変動長スペクトルを朗読音声と歌声の判別に利用することの有効性の検討を行った．変動長スペクトルを特徴量，AdaBoost を識別器として，朗読音声と歌声を識別する実験を行った結果，87%程度の正解率を示した．これは，MFCC を特徴量として用いた場合の正解率を上回っている．また，MFCC と変動長スペクトルを併用した結果，MFCC のみを用いた場合と比較して性能の向上が見られ，89%程度の正解率を示した．これらの結果から，変動長スペクトルの，朗読音声と歌声の判別のための特徴量としての有効性が示された．

本稿では朗読音声と歌声の二値判別問題を扱ったが，より現実的な環境においては，朗読音声と歌声以外にも楽器音や雑音などが現れうるため，多値判別が必要となる．変動長スペクトルのこれらの判別のための特徴量としての有効性については今後の検討課題となる．

謝辞 本研究の一部は日本学術振興会科研費特別研究員奨励費 (22-6961) の助成を受けて行われた．また，模範歌唱付きカラオケデータからのスペクトル減算法による簡易歌声抽出に関して INRIA の Emmanuel Vincent 博士からの助言を受けた．

参考文献

[1] J. Saunders, *Proc. ICASSP*, pp. 993–996, 1996.
[2] 大石 他，情処論, Vol. 47, No. 6, pp. 1822–1830, 2006.
[3] 橘 他，音講論（秋）, pp.607–608, 2010.
[4] 橘 他，信号処理シンポジウム, pp.171–176, 2010.
[5] Ono *et al.*, “Harmonic and Percussive Sound Separation and its Application to MIR-related Tasks,” Springer 274, pp.213-236, 2010.
[6] Goto, *Proc. ICA*, pp. I-553-556, 2004.
[7] 板橋 他，音講論（秋）, pp.187–188, 1997.

[1] http://www.karaokewh.com/keynote-karaoke.cfm
[2] http://sites.google.com/site/unvoicedsoundseparation/mir-1k