無伴奏歌唱におけるブレスの音響特性と自動検出*

◎中野倫靖 (筑波大), 緒方淳 (産総研), 後藤真孝 (産総研), △平賀譲 (筑波大)

1 はじめに

本研究では、ポピュラー音楽の歌唱における、マイク収録されたブレス (吸気、息継ぎ) 音について、その位置を自動検出する手法の実現を目的とする。歌唱中のブレスを自動的に検出することができれば、様々な応用が考えられる。まず、歌唱音声の収録において、ブレスを消したり強調したりする場面で有用であり、実際にオーディオ編集ソフトに導入されている [1]。さらに、ブレスはフレーズ (音楽的なまとまり) 境界に位置する可能性が高いため、長い歌唱音声データを適切な長さに自動的に切り分ける場面で有用である。音声認識分野においては、ブレス位置が自然な区切り箇所であることが指摘されており [2]、句読点の挿入に息継ぎ位置を利用する研究もあった [3]。また、実際の歌唱においては、フレーズ境界では深くブレスを行い、そのブレスの終端が次フレーズの直前であることが報告されている [4]。ブレス位置は、歌唱者のリズム感や「間」の取り方のうまさに関係している可能性があり、歌唱力の自動評価 [5] への応用も考えられる。

従来、マイク入力されたブレスの自動検出は研究されてきたが、その音響的な特性としては「調波構造を持たない」「継続時間長が長い」「ブレスの前後には無音が存在する」「/s/に比べて零交差が小さな値を取る」等の単純な指標しか示されていなかった。本論文では、実際の伴奏なしの歌唱を分析し、ブレスが特徴的なスペクトル形状を持つことを示す。また、HMM (Hidden Markov Model) によってブレスの音響モデルを構築して、実際の歌唱に対して検出実験を行う。

2 関連研究

歌唱音声を対象としたブレスの自動検出では、Ruinskiy *et al.* [6] による研究がある。Ruinskiy *et al.* は、「ブレスの前後には無音が存在する」「/s/との零交差値の比較 (ブレスの方が小さい)」「母音とのパワーの比較 (ブレスの方が小さい)」等の知見を示している。自動検出では、MFCC, 零交差, Spectral Slope (11 – 22kHz の帯域の傾斜) 及びパワーを特徴量としたテンプレートマッチングを行い、マッチング結果からさらにブレスの始端終端を探索していた。手法の有効性は、20 人の歌手と 2 人のナレーターの計 24 分の音声データ (332 箇所のブレス) に対して評価された。男女で同一のテンプレートを用意した場合は、再現率と精度はそれぞれ 94.4%, 96.5%、男女を分けてテンプレートを作った場合は、それぞれ 97.6%, 95.7%と報告されている。

話し声を対象としたブレスの自動検出では、Price *et al.* [7]、Wightman *et al.* [8] の研究がある。Price *et al.* は特徴量をケプストラムとした GMM (混合ガウス分布モデル) ベースの識別器で識別を行い、93%の検出率を得ている [7]。ここで、自動検出したブレスには、4 人の聴取者が聞き取れなかったこともあることを述べている。Wightman *et al.* は、特徴量をケプストラムとしたベイズ識別によって最大で 91.3% の検出性能を得ている [8]。

また、歌唱や話し声以外では、フルート音 (楽器音) とブレスの識別 [9, 10]、呼気音・吸気音の識別を利用したインタフェース [11] に関する研究がある。

3 ブレスの音響特性

本研究では、ブレスの音響特性を分析するために、ポピュラー音楽の歌唱を対象とする。そこで、分析対象の歌唱音声には、RWC 研究用音楽データベースのポピュラー音楽 RWC-MDB-P-2001 [12] の伴奏なしのデータ (以下、RWC-MDB) を用いた。また、収録条件の異なる歌唱データとして、AIST ハミングデータベース [13] 中の歌唱データ (以下、AIST-HDB) も利用した。AIST-HDB は、RWC-MDB の曲を初めて聴く歌唱者が 5 回聴いた後に、思い出しながら歌う音声を収録したものである。

3.1 分析対象の歌唱データ

RWC-MDB からは 16 人が歌った 27 曲を、AIST-HDB からは 2 人が歌った各 50 フレーズ (フレーズ長の平均は約 11.4 秒) を用いた。これらの曲は次の 5 点を考慮してデータベースから選別した:

- 複数の歌唱者間の比較ができること
- 同一歌唱者の比較もできること
- 男女比がおよそ同じとなること
- 日本語だけでなく、英語の曲も含めること
- 歌唱力の差も考察できること

表 1 に、本実験で使用した曲に対して手作業でラベリングしたブレスの個数、総ブレス長、及びブレス長の統計量 (平均、標準偏差、最小、最大) を示す。

---

*Analysis and Automatic Detection of Breath Sounds in Solo Vocal. by NAKANO, Tomoyasu†, OGATA, Jun‡, GOTO, Masataka‡, HIRAGA, Yuzuru† († University of Tsukuba, ‡ AIST)

ここで、RWC-MDB における曲番号は、RWC 研究用音楽データベース (ポピュラー音楽 RWC-MDB-P-2001 [12]) に対応しており、AIST-HDB における「うまい/へた」は文献 [5] のラベルを利用して決定した。

3.2 ブレスのスペクトル形状

歌唱データに周波数解析を行ってスペクトログラムを得た後で、そこからブレス区間だけを切り出して並べると、およそ同じ周波数帯域にパワーのピークの存在が視認できる。例として、p038(日本語男性) と p097(英語女性) の全ブレス区間のスペクトログラムとその長時間平均を図 1 に示す。ここでは、各フレーム毎にスペクトル包絡 (ケプストラムから算出) を表示している。参考のために音声分析ツール WaveSurfer [14] を用いて算出した第 1〜第 3 フォルマント周波数 (それぞれ、$F1, F2$, and $F3$) の平均も示した。

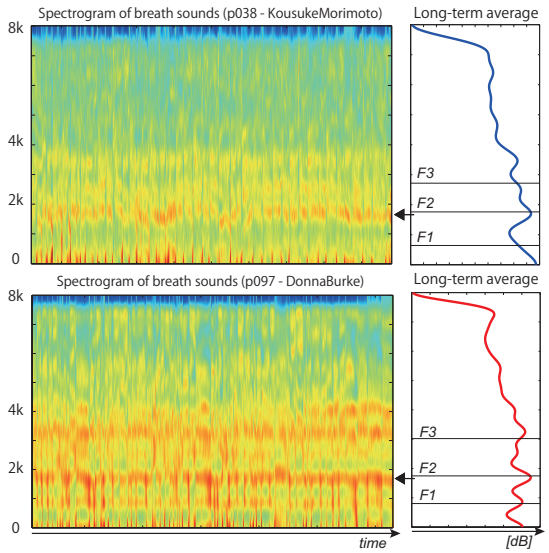

Fig. 1　p038 と p097 のブレスのスペクトログラム (左図) 及びその長時間平均 (右図)。右図にはフォルマント周波数 ($F1$, $F2$, and $F3$) の平均値も示した。

表 1 の全歌唱データについての、ブレス区間の長時間平均スペクトルを図 2 に示す。

3.3 考察

前章の解析結果からは、ブレスのスペクトル形状が 1 曲中で類似していること (図 1)、同一歌唱者のブレスの長時間平均スペクトルの形状が類似していること (図 2)、歌唱者・曲・言語・歌唱力が異なっても 1.6kHz(男性)〜1.7kHz(女性) 付近にピークが存在することが多いことが分かる。また、850Hz〜1kHz 付近にピークが存在することも多くあった。RWC-MDB の歌唱音声と AIST-HDB の歌唱音声のスペクトル上のピークが、ほぼ同じ周波数帯域に存在していることは、これがブレスの音響特性である可能性が高い。

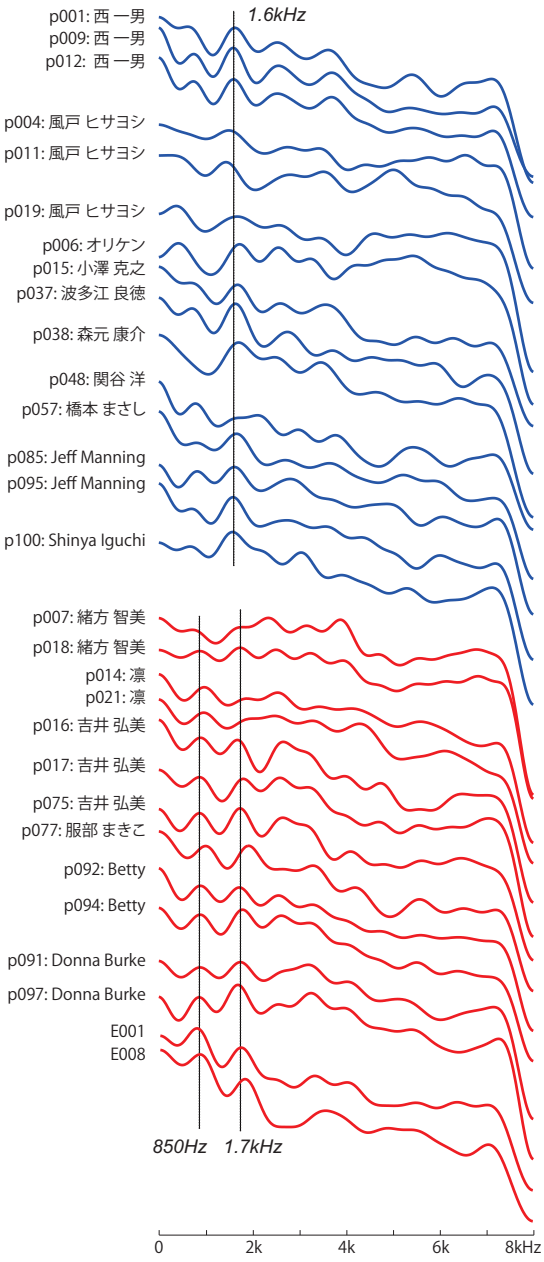

Fig. 2　各曲におけるブレスの長時間平均スペクトルと、その特徴的なピークの周波数。

ただし、1.7kHz 付近のピークについては、そのピークの周波数が大きく変動する歌唱音声もあった。男性であれば p048、女性であれば p007, p014, p021 の歌唱音声では、ピーク位置が 1.5〜3kHz の範囲で大きく変動していた。ただし、p048, p014, p021 については、850Hz〜1kHz 付近に安定してピークがあった。

以上の考察を踏まえると、スペクトル包絡に基づく音響特徴量は、ブレス検出に有効であると考えられる。実際、先行研究でもケプストラム、MFCC といったスペクトル包絡に基づく特徴量が用いられ、その有効性が確認されている [7, 8, 6]。また、表 1 に示すように、ブレスはその継続時間長は短くて 50msec、最大で 1 秒を超えるなど変動が非常に大きい。すな

Table 1　RWC 研究用音楽データベース中の 27 曲 (歌唱者 16 人) と、AIST ハミングデータベース中の 2 人の歌手による歌唱音声 (それぞれ 50 フレーズ)、におけるブレスの個数、総ブレス長、ブレス長の統計量。

RWC-MDB

曲番号	歌手名	歌唱者の性別	歌詞の言語	ブレスの個数	総ブレス長 (sec)	ブレス長の統計量 (msec)			
						平均	標準偏差	最小	最大
p001	西 一男	男性	日本語	54	9.7	179.7	47.7	81.7	311.6
p009	西 一男	男性	日本語	54	13.7	253.1	85.6	100.0	432.5
p012	西 一男	男性	日本語	45	10.3	228.6	107.7	55.0	530.0
p004	風戸 ヒサヨシ	男性	日本語	10	2.0	199.5	79.1	105.0	325.0
p011	風戸 ヒサヨシ	男性	日本語	23	7.0	304.0	134.7	137.5	535.0
p019	風戸 ヒサヨシ	男性	日本語	31	5.4	174.0	63.4	87.5	412.5
p006	オリケン	男性	日本語	43	16.4	380.6	187.5	105.0	975.0
p015	小澤 克之	男性	日本語	9	2.5	277.8	222.2	102.5	832.5
p037	波多江 良徳	男性	日本語	59	15.2	257.5	68.7	157.3	532.8
p038	森元 康介	男性	日本語	49	13.5	275.6	96.8	95.0	527.5
p048	関谷 洋	男性	日本語	48	22.0	457.4	208.4	190.0	1225.0
p057	橋本 まさし	男性	日本語	62	23.0	371.2	157.8	175.0	845.0
p085	Jeff Manning	男性	英語	53	11.2	212.2	60.2	95.0	375.0
p095	Jeff Manning	男性	英語	70	20.8	296.8	185.8	100.9	1074.0
p100	Shinya Iguchi	男性	英語	39	19.0	487.4	168.0	199.6	932.5
p007	緒方 智美	女性	日本語	12	3.1	262.2	94.2	124.3	442.8
p018	緒方 智美	女性	日本語	48	9.3	193.8	73.2	50.0	347.5
p014	凛	女性	日本語	60	14.5	242.2	88.6	75.0	585.0
p021	凛	女性	日本語	60	20.5	342.0	128.2	107.5	590.0
p016	吉井 弘美	女性	日本語	43	10.9	254.0	108.5	85.0	497.5
p017	吉井 弘美	女性	日本語	55	16.1	292.3	110.8	122.5	687.5
p075	吉井 弘美	女性	日本語	33	10.7	323.7	98.5	123.2	542.7
p077	服部 まきこ	女性	日本語	51	15.4	301.3	126.6	135.0	757.5
p092	Betty	女性	英語	37	11.3	304.4	108.3	122.5	690.0
p094	Betty	女性	英語	44	11.5	261.1	92.5	112.5	452.5
p091	DonnaBurke	女性	英語	48	15.1	315.0	109.2	117.5	690.0
p097	DonnaBurke	女性	英語	63	18.1	287.1	158.0	66.1	786.1

※楽曲の曲番号は RWC 研究用音楽データベース (ポピュラー音楽 RWC-MDB-P-2001 [12]) に対応

AIST-HDB

歌唱力	歌手名	歌唱者の性別	歌詞の言語	ブレスの個数	総ブレス長 (sec)	ブレス長の統計量 (msec)			
						平均	標準偏差	最小	最大
へた	E001	女性	英語	144	45.9	318.5	143.5	87.5	807.5
うまい	E008	女性	英語	141	44.5	315.3	160.6	102.5	1069.3

※「うまい/へた」は文献 [5] のラベルを利用

わち、ブレス検出ではそのような時間変動に対処する必要があると言える。

4　ブレスの自動検出

本研究では、HMM (Hidden Markov Model) による歌唱音声中のブレス検出法を提案する。本稿では、その第一段階として、ブレス/歌声/無音 (sp) の 3 種の HMM を構築して検出実験を行った結果を報告する。HMM によるブレス検出は、継続時間長や特徴量の変動に対処できる利点がある。また、ブレス以外のイベント検出 (例えばビブラートなど) や歌詞認識なども、同様の枠組みで行える可能性がある点で、拡張性が高いといえる。

4.1　実験条件

HMM の構築には、特徴量として音声認識で広く用いられている MFCC (Mel-Frequency Cepstrum Coefficient)、ΔMFCC, ΔPower を利用した。具体的な分析条件を表 2 に示す。

実験では、表 1 に示される RWC-MDB の楽曲 27 曲を用い、評価データを歌唱者毎 (16 人) に分けて、16 回のクロスバリデーションで評価を行った。つまり、ある歌唱者によって歌われている楽曲を評価する際は、その歌唱者以外に歌われているデータ全てを用いて HMM を学習する。

Table 2　歌唱音声の分析条件

サンプリング周波数	16kHz
分析窓	ハミング窓
フレーム幅	25ms
フレームシフト	10ms
特徴量	12th order MFCC 12th order ΔMFCC Δ Power
音響モデル	状態数 3、混合数 16

ここで、特徴抽出、音響モデルの学習と Viterbi アラインメントには、HTK Speech Recognition Toolkit [15] を用いた。

4.2　検出結果

提案手法の有効性を評価するために、ブレス検出の再現率 (R) と精度 (P) を算出した。R と P は、それ

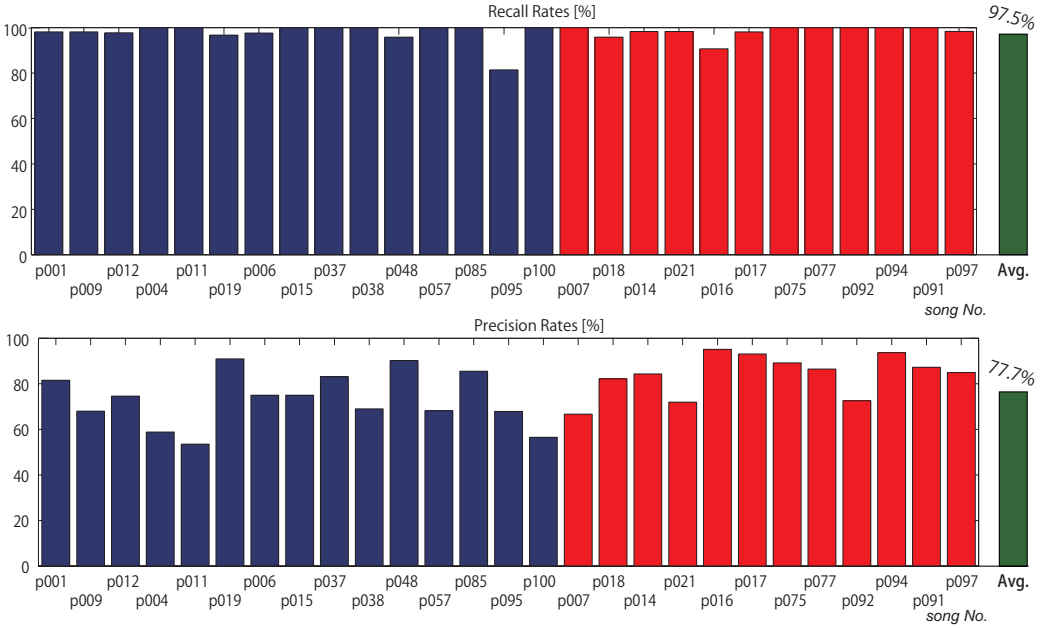

Fig. 3　ブレス検出の再現率 (recall rate) と精度 (precison rate)

ぞれ式 (1), (2) のように定義する。実験に用いた 27 曲の R と P は、図 3 に示した。

$$R = \frac{\text{正しく検出されたブレス数}}{\text{正解ブレスの総数}} \times 100 \quad (1)$$

$$P = \frac{\text{正しく検出されたブレス数}}{\text{ブレスとして検出された区間の総数}} \times 100 \quad (2)$$

4.3　考察

現在の実装では、ブレス以外の箇所をブレスとして誤検出してしまうために、検出精度が低くなってしまうことが多かった。ここで、誤検出の原因は吐く息に起因することがほとんどであった。例えば、フレーズの終わりがブレスとしてラベル付けされることが多く、フレーズの終わりでは息を吐くように歌うことが多かった。また、子音を伸ばして歌うような箇所を誤検出することもあった。精度が特に低かった p011, p100 ではこれらの要因に加えて、わずかに背景音楽 (歌唱音声の収録時にマイクに混入した音) が入っている区間を誤検出してしまうことがあった。

しかし、本手法は全ての曲に対して高い再現率が得られており、得られた結果からブレスを選別するような方法を導入すれば、このような誤検出へ対処できる可能性がある。

5　おわりに

本論文では、無伴奏歌唱におけるブレスの音響特性について、そのスペクトルに特徴的なピークが存在する場合があることを述べた。また、ブレスの自動検出として HMM を用いたブレス検出法を提案した。今後は、ブレスの音響特性をより生かした音響特徴量及び、背景音楽に対して頑健なブレス検出法を検討していく予定である。

謝辞　本研究に対し有益な議論をして頂きました齋藤 毅 氏、藤原 弘将 氏 (産総研) に感謝いたします。また、本研究では、RWC 研究用音楽データベース (ポピュラー音楽 RWC-MDB-P-2001)、及び AIST ハミングデータベースを使用しました。

参考文献

[1] Waves, "Waves | プラグイン | DeBreath," <http://www.waves.com/content.aspx?id=2173>

[2] Wightman *et al.*: Automatic Labeling of Prosodic Patterns, In *IEEE Trans. on Speech and Audio Processing*, Vol. 2, No. 4, pp.469–481, 1994.

[3] 西村 他: 単語を認識単位とした日本語の大語彙連続音声認識, 情処学論, Vol.40, No.4, pp.1395–1403, 1999.

[4] 中村敏江: 音楽における「間」と呼吸について, 日本音響学会音楽音響研究会資料, MA94-16, pp.19–26, 1994.

[5] 中野 他: 楽譜情報を用いない歌唱力自動評価手法. 情処学論. Vol.48, No.1, pp.227–236, 2007.

[6] Ruinskiy *et al.*: An Effective Algorithm for Automatic Detection and Exact Demarcation of Breath Sounds in Speech and Song Signals, In *IEEE Trans. on the Audio, Speech and Language Processing*, Vol. 15, Issue 3, pp.838–850, 2007.

[7] Price *et al.*: Prosody and Parsing, In *Proc. Workshop on Speech and Natural Language*, pp.5–11, 1989.

[8] Wightman *et al.*: Automatic Recognition of Prosodic Phrases, In *Proc. ICASSP 91*, pp.321–324, 1991.

[9] 堀内 他: 伴奏システムでのブレス情報利用に関する検討, 情処研報 MUS, Vol.2005, No.45, pp. 13–18, 2005.

[10] 堀内 他: ブレスの合図を認識する伴奏システムの実装と評価, 情処研報 MUS, Vol.2007 No.81, pp. 1–6, 2007.

[11] 伊賀 他: Kirifuki: 呼気・吸気を利用した計算機とのインタラクション, 情処研報 HI, Vol.2000, No.12, pp. 49–54, 2000.

[12] 後藤 他: 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. 情処学論. vol.45, no.3, p.728–738, 2004.

[13] 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 MUS, Vol.2005, No.82, pp.7–12, 2005.

[14] Sjolander *et al.* WaveSurfer – An Open Source Speech Tool, In *Proc. ICSLP-2000*, Vol.4, pp.464-467, 2000.

[15] Young *et al.*: *The HTK Book*, Ver. 3.2.1, 346p., 2002.