無伴奏歌唱におけるブレスの音響特性と それに基づく自動ブレス検出

中野 倫靖 † 後藤 真孝 † 緒方 淳 † 平賀 譲 ‡

† 産業技術総合研究所 ‡ 筑波大学

{t.nakano, m.goto, jun.ogata}@aist.go.jp hiraga@slis.tsukuba.ac.jp

あらまし 本研究では、歌唱音声のブレス (吸気、息継ぎ) 音を対象として、音響特性の分析とその結果に基づく自動ブレス検出法の実現を目的とする。従来、朗読音声や歌唱音声のブレスを検出する研究はあったが、その音響特性は比較的単純な指標しか示されていなかった。そこで本研究では、まずブレス音の音響特性を詳細に検討するために、18 人の歌手が歌った 128 分の歌唱データ (1488 箇所のブレス) を用いて分析を行った。その結果、ブレス音のスペクトル包絡形状が同一曲中で類似していること、その長時間平均には 1.6kHz (男性), 1.7kHz(女性) 付近に顕著なピークが認められることを示す。また、音響特性に基づく自動ブレス検出の第一段階として、MFCC, ΔMFCC, Δ パワーを特徴量とした HMM を用いた検出実験を行った。27 曲分の歌唱音声を対象に実験した結果、本手法は再現率と精度それぞれについて 97.5%, 77.7%を得た。

Acoustic Characteristics of Breath Sounds in Solo Vocal and Their Application to Automatic Breath Detection

Tomoyasu Nakano† Masataka Goto† Jun Ogata† Yuzuru Hiraga‡

†National Institute of Advanced Industrial Science and Technology (AIST)

‡University of Tsukuba

Abstract The aim of this study is to develop an automatic breath detection method based on an acoustic analysis of breath sounds in singing. Although there are previous works on automatic breath detection, the reported characteristics of breath sounds were relatively simple. In contrast, this study starts with a detailed acoustic analysis of breath sounds, with the aim to explore novel characteristics. The acoustic analysis used singing voice recordings of 18 singers with a total length of 128 mins (1488 breath events). The results of the analysis suggest that the spectral envelopes of breath sounds remain similar within the same song, and their long-term average have a notable spectral peak at about 1.6 kHz (male) and 1.7 kHz (female). A preliminary approach presented in the experiment used HMM based on MFCC, ΔMFCC, and Δpower as acoustic features. In the evaluation experiment with 27 unaccompanied song samples, our method achieved an overall recall/precision rate of 97.5%/77.7% for breath sound detection.

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。また、そのために、ブレスの音響特性を詳細に分析する。

歌唱音声中のブレス位置を自動的に検出できれば、様々な場面で応用できる。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、そのような機能は実際にオーディオ編集ソフトに導入されている [1]。また、歌声合成において人間の歌唱を真似るシステムが提案されているが [2]、その際にブレスを自然な位置に挿入できる。

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、自動検出は長い歌唱音声データを適切な長さへ自動的に切り分ける際にも有用である。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 [3]、句読点の挿入に息継ぎ位置を利用する研究もあった [4]。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 [5]。さらに、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 [6] への応用も考えられる。

一方、ブレス音の音響特性の解明は、自動検出の精度向上への寄与が考えられるだけでなく、歌声合成分野においても重要である。歌声合成において、歌唱中のブレスは、その歌唱をより人間らしく聞かせるための重要な要素の一つである。したがって、音響特性を詳細に分析することは、高品質なブレス音合成、高品質な歌声合成につながると言える。

従来、歌唱音声を対象としたブレスの自動検出では、

Ruinskiy *et al.* [7] による研究があった。Ruinskiy *et al.* は、ブレスの音響特性として「ブレス前後における無音の存在」「子音/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[1] 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。

また、話し声を対象としたブレスの自動検出では、Price *et al.* [8]、Wightman *et al.* [9] の研究がある。Price *et al.* は、プロのアナウンサーが発声した音声に対して、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 (83 箇所のブレス) を得た [8]。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、112 箇所のブレスに対して Open 実験で 73.2%、Closed 実験で 91.3% の検出性能を得ている [9]。

歌唱や話し声以外では、フルート音 (楽器音) とブレス音の識別 [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。堀内 他は、「調波構造をほとんど持たない」「パワーがある程度の時間長持続する」という特性を利用して、フルート音とブレス音の識別を行った [10]。

このように、ブレスに関する研究は従来行われてきたが、自動検出が主な研究対象とされてきた。しかし、その音響的な特性としては比較的単純な指標しか示されておらず、分析結果として具体的な数値も示されていなかった。そこで本稿では、歌唱音声のブレスに対する分析を詳細に行い、音響特性に関する新たな知見を示す。また、ブレスの音響特性に基づいた自動検出の実現を目指し、その第一段階として HMM (Hidden Markov Model) による自動検出法を提案する。

本論文は以降、第 **2** 章でブレスの音響特性の分析を行い、その結果について述べる。続いて、第 **3** 章でブレスの自動検出について述べた後、第 **4** 章でまとめる。

2 ブレスの音響特性

本研究では、ブレス音としてポピュラー音楽の歌唱を対象として、その音響特性を分析する。まずは、最も基本的な分析として、ブレス音の継続時間長を調査する。また、ブレス音の自動検出に関する従来研究 [7–9] の結果から、ケプストラムのようなスペクトル包絡に関する音響特徴量が、ブレス検出に有効であると考えられるため、ブレス音のスペクトル包絡にも着目する。

[1] 文献 [7] 中では、再現率と精度をそれぞれ Sensitivity と Specificity と呼んでいる。

2.1 分析対象の歌唱データ

分析対象の歌唱音声として、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。

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。**表 1** に分析に用いた曲を示す (計 128 分)。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、英語歌詞の曲も含めること
- 歌唱力の差も考察できること

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]) に対応している。AIST-HDB のデータは 50 フレーズを繋げて 1 曲として扱い、歌唱力の「うまい/へた」は文献 [6] のラベルを利用して決定した。

これ以降、歌唱音声信号は全て、16kHz/16bit サンプリングのモノラル音声信号を扱う。

2.2 分析の方法及び条件

ブレス位置は、全て手作業でラベル付けを行い、分析にはその結果を利用して以下のように行った。

時間長に関する分析 ブレス音の時間長の統計量を算出する。算出した統計量は、ブレス長の平均、標準偏差、最小長さ及び最大長さである。

スペクトル包絡に関する分析 ケプストラム及びフォルマント周波数の分析を行う。ケプストラムは、歌唱音声信号に対して短時間フーリエ変換 (STFT) を行い、窓幅 1024 点 (64msec) のハニング窓を 160 点 (10msec) ずつシフトさせて計算した。スペクトル包絡は、ケプストラムの低次 23 項を用いて (0 次項を除く)、逆フーリエ変換によって算出した。また、フォルマント周波数の算出には、音声分析ツール WaveSurfer [14] を用いた。WaveSurfer での分析条件を**表 2** に示す。

表 2: WaveSurfer のフォルマント周波数分析条件

Number of formants	4
Analysis window length	64 msec
Pre-emphasis factor	0.7
Frame interval	10 msec
LPC order	12
LPC type	0
Down-sampling frequency	10 kHz
Nominal F1 frequency	-10.0 Hz

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

RWC-MDB

曲番号	歌唱者名	歌唱者の性別	歌詞の言語	曲の長さ (sec)	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
No.001	西 一男	男性	日本語	207.2	54	9.7	179.7	47.7	81.7	311.6
No.009	西 一男	男性	日本語	275.0	54	13.7	253.1	85.6	100.0	432.5
No.012	西 一男	男性	日本語	202.8	45	10.3	228.6	107.7	55.0	530.0
No.004	風戸 ヒサヨシ	男性	日本語	240.5	10	2.0	199.5	79.1	105.0	325.0
No.011	風戸 ヒサヨシ	男性	日本語	265.9	23	7.0	304.0	134.7	137.5	535.0
No.019	風戸 ヒサヨシ	男性	日本語	287.0	31	5.4	174.0	63.4	87.5	412.5
No.006	オリケン	男性	日本語	204.5	43	16.4	380.6	187.5	105.0	975.0
No.015	小澤 克之	男性	日本語	160.9	9	2.5	277.8	222.2	102.5	832.5
No.037	波多江 良徳	男性	日本語	237.1	59	15.2	257.5	68.7	157.3	532.8
No.038	森元 康介	男性	日本語	273.3	49	13.5	275.6	96.8	95.0	527.5
No.048	関谷 洋	男性	日本語	269.6	48	22.0	457.4	208.4	190.0	1225.0
No.057	橋本 まさし	男性	日本語	265.1	62	23.0	371.2	157.8	175.0	845.0
No.085	Jeff Manning	男性	英語	208.5	53	11.2	212.2	60.2	95.0	375.0
No.095	Jeff Manning	男性	英語	230.8	70	20.8	296.8	185.8	100.9	1074.0
No.100	Shinya Iguchi	男性	英語	293.3	39	19.0	487.4	168.0	199.6	932.5
No.007	緒方 智美	女性	日本語	296.7	12	3.1	262.2	94.2	124.3	442.8
No.018	緒方 智美	女性	日本語	252.1	48	9.3	193.8	73.2	50.0	347.5
No.014	凛	女性	日本語	232.9	60	14.5	242.2	88.6	75.0	585.0
No.021	凛	女性	日本語	266.9	60	20.5	342.0	128.2	107.5	590.0
No.016	吉井 弘美	女性	日本語	262.1	43	10.9	254.0	108.5	85.0	497.5
No.017	吉井 弘美	女性	日本語	239.4	55	16.1	292.3	110.8	122.5	687.5
No.075	吉井 弘美	女性	日本語	199.9	33	10.7	323.7	98.5	123.2	542.7
No.077	服部 まきこ	女性	日本語	234.1	51	15.4	301.3	126.6	135.0	757.5
No.092	Betty	女性	英語	218.4	37	11.3	304.4	108.3	122.5	690.0
No.094	Betty	女性	英語	221.9	44	11.5	261.1	92.5	112.5	452.5
No.091	Donna Burke	女性	英語	221.8	48	15.1	315.0	109.2	117.5	690.0
No.097	Donna Burke	女性	英語	241.1	63	18.1	287.1	158.0	66.1	786.1
合計（RWC-MDB）				6508.8	1203	348.3	289.5	143.2	50.0	1225.0

※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	歌唱者名	歌唱者の性別	歌詞の言語	曲の長さ (sec)	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	595.9	144	45.9	318.5	143.5	87.5	807.5
うまい	E008	女性	英語	558.6	141	44.5	315.3	160.6	102.5	1069.3
合計（AIST-HDB）				1154.5	285	90.3	316.8	141.6	87.5	807.5

※歌唱力の「うまい/へた」は文献 [6] のラベルを利用

RWC-MDB と AIST-HDB

合計				7663.3	1488	438.6	294.7	143.3	50.0	1225.0

2.3 分析結果

ブレス音の時間長及びスペクトル包絡に関して、前節で述べた方法で分析を行った結果を述べる。

2.3.1 時間長に関する分析結果

表 1 に、分析に用いた全曲に対して、時間長に関する次のような分析結果を示す。

- 曲の長さ
- ブレスの総数
- 総ブレス長
- ブレス長の統計量 (平均、標準偏差、最小、最大)

ラベル付けされたブレスの総数は 1488 箇所で、1 曲に付き平均 51.3 箇所あった。このうち、RWC-MDB では平均 44.6 箇所、AIST-HDB では平均 142.5 箇所であった。総ブレス長の合計は 438.5 秒であり、これは曲の総長さ 128 分 の 5.7 % に相当する。また、各ブレス長は、50 msec (最小) から 1225 msec(最大) まで大きな変動を持ち、その平均と標準偏差は 294.7 msec と 143.3 msec であった。

2.3.2 スペクトル包絡に関する分析結果

歌唱データに周波数解析を行って全時刻のスペクトル包絡を得た後で、そこからブレス区間だけを切り出して並べると[2] 、およそ同じ周波数帯域にパワーのピークの存在が確認できる。**図 1** に、No.038(日本語男性) と No.097(英語女性) のブレス区間の全フレームにおける、スペクトル包絡とその長時間平均を示す。また、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。

図 1 からは以下のことが分かる。

- スペクトル包絡形状が 1 曲中で類似していること
- スペクトル包絡に顕著なピークが認められること
- 長時間平均をとってもピークが残り、その周波数帯域は第 2 フォルマント周波数付近であること

さらに、表 1 の全歌唱データについて、ブレス区間のスペクトル包絡の長時間平均を**図 2** に示す。また、**表 3** と**表 4** に、それぞれ全歌唱者のブレスの F1〜F3 の

[2] 歌声区間と無音区間を除いて表示している。

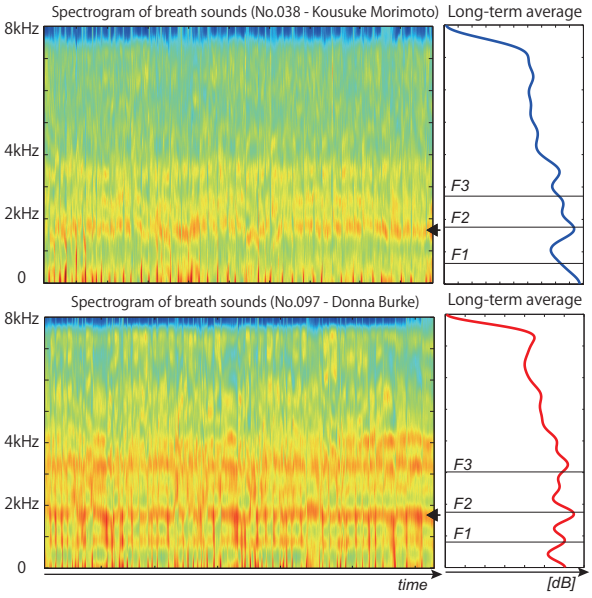

図 1: No.038 と No.097 における全ブレス区間のスペクトル包絡 (左図) 及びその長時間平均 (右図)。右図にはフォルマント周波数 ($F1$, $F2$, and $F3$) の平均値も示した。

平均と、男女 6 人分の歌唱の冒頭のフレーズ (平均 6.1 秒) から日本語 5 母音の F1 と F2 の平均を算出した結果を示す。これらの図表からは以下のことが分かる。

- 同一歌唱者のブレスは、そのスペクトル包絡の長時間平均が類似した形状を持つこと
- 歌唱者・曲・言語・歌唱力・収録条件が異なっても 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
- 上述のピークに加えて、850Hz〜1kHz 付近にもピークが存在する場合があること
- ブレスの F1 の平均は 704Hz (男性), 775Hz (女性) であり、F2 の平均は 1.72kHz (男性), 1.83kHz (女性) であること

ここで、1.7kHz 付近のピークについては、そ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 No.048、女性であれば No.007, No.014, No.021 の歌唱音声では、ピーク周波数が 1.5〜3kHz の範囲で大きく変動していた。ただし、No.048, No.014, No.021 については、850Hz〜1kHz 付近に安定してピークがあった。またこれらのピークは、ブレス前後の音韻によって変動 (調音結合) する場合があった。

2.4 考察

前章の分析結果から、ブレスのスペクトル包絡は、1.7kHz 付近に顕著なピークを持つことを示した。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。このようなピークは、声道の共鳴特性に起因すると考えられるが、表 3 に示したブレスのフォルマント周波数は、表 4 や文献 [15]（p.117, 図 5.17）に示されている母音の F1, F2 と比較すると、/a/や/e/もしくはその中間の/æ/に近い。すなわち、ブレスでは/a/や/e/に近い声道形状をとっていることが多いと考えられる。

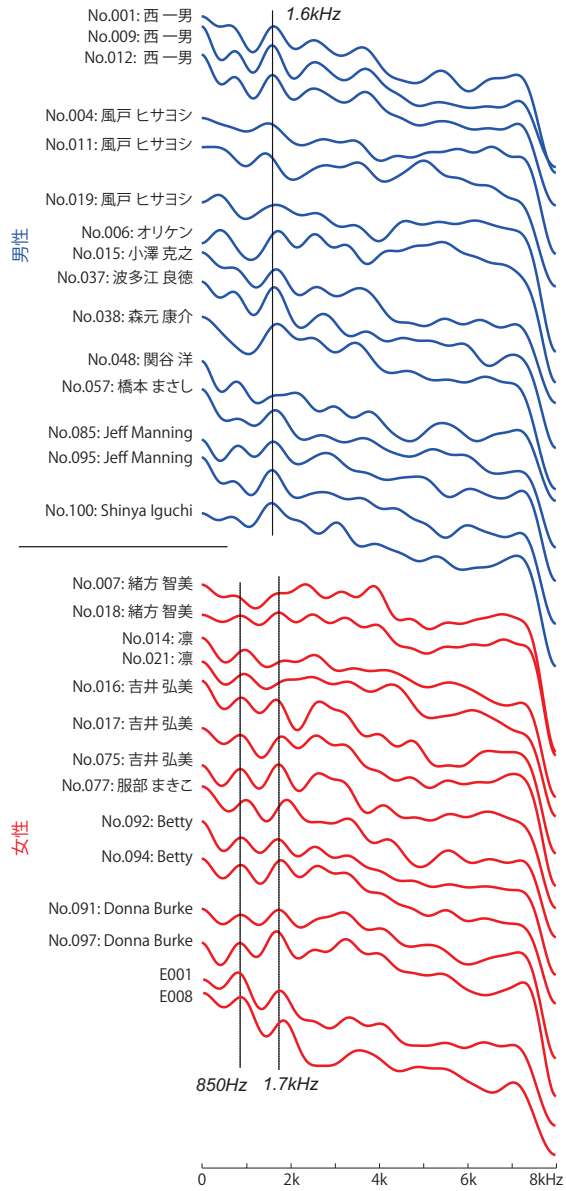

図 2: 各曲におけるブレスのスペクトル包絡の長時間平均と、特徴的なピークが存在する周波数。

また、同一歌唱者のブレスのスペクトル包絡形状は、曲が変わっても類似していた[3] 。すなわちブレスは歌詞、メロディ、テンポへの依存が少ない音響特性を持っている可能性がある。

従来、スペクトル包絡に基づく音響特徴量がブレス検出に有効であった [7–9] が、ここで示したピークの存在が一つの寄与であったと考えられる。自動検出においては、スペクトル包絡を表す音響特徴量を用い、また、その変動を吸収できる必要がある。また、ブレスはその継続時間長は変動が非常に大きく (表 1)、自動検出ではこのような変動にも対処する必要がある。

[3] 例えば、No.001, No.009, No.012 や No.016, No.017, No.075。

表 3: 各曲におけるブレス区間のフォルマント周波数 ($F1, F2, F3$) の平均と標準偏差 (括弧内)。

男性

曲番号	F1 [Hz]	F2 [Hz]	F3 [Hz]
No.001	608.7 (272.0)	1726.9 (201.9)	2711.0 (301.2)
No.009	789.5 (296.6)	1684.3 (197.7)	2673.5 (219.4)
No.012	742.1 (271.4)	1720.4 (214.5)	2716.7 (273.0)
No.004	651.8 (299.2)	1661.4 (383.7)	2908.6 (405.3)
No.011	576.8 (272.8)	1620.5 (383.0)	2852.0 (393.6)
No.019	436.0 (225.3)	1606.1 (288.0)	2686.2 (449.0)
No.006	495.6 (192.0)	1708.8 (202.4)	2621.6 (293.3)
No.015	611.1 (238.6)	1652.1 (156.2)	2632.8 (209.9)
No.037	633.3 (227.0)	1650.7 (156.7)	2740.9 (228.9)
No.038	646.4 (340.4)	1759.2 (248.2)	2715.3 (415.9)
No.048	791.2 (228.9)	1902.6 (265.9)	2848.6 (347.1)
No.057	780.9 (303.8)	1736.7 (179.2)	2835.0 (324.0)
No.085	847.4 (232.3)	1704.9 (253.6)	2777.6 (243.3)
No.095	759.2 (322.0)	1674.0 (267.1)	2754.4 (354.8)
No.100	760.2 (292.6)	1718.8 (278.4)	2817.1 (393.2)
全体	704.9 (294.5)	1722.4 (251.4)	2758.6 (336.7)

女性

曲番号	F1 [Hz]	F2 [Hz]	F3 [Hz]
No.007	724.0 (264.5)	1985.6 (301.8)	2931.2 (381.2)
No.018	805.2 (264.6)	1930.9 (353.9)	2947.2 (391.2)
No.014	896.5 (286.0)	2114.9 (396.4)	3079.4 (467.4)
No.021	928.4 (248.7)	2182.6 (339.8)	3214.3 (348.1)
No.016	750.2 (313.5)	1719.5 (342.2)	2806.6 (269.9)
No.017	740.9 (253.3)	1900.8 (281.6)	2796.7 (304.7)
No.075	832.3 (205.8)	1776.3 (183.7)	2773.1 (254.7)
No.077	869.4 (271.3)	1931.5 (242.7)	3135.0 (338.3)
No.092	790.3 (288.2)	1793.5 (243.9)	2788.3 (342.5)
No.094	778.5 (245.0)	1846.6 (222.5)	2839.8 (333.8)
No.091	776.5 (278.8)	1786.9 (269.4)	2980.9 (300.4)
No.097	796.1 (243.6)	1744.2 (196.7)	2998.8 (350.9)
E001	757.8 (155.5)	1765.1 (246.0)	2895.7 (448.0)
E008	636.8 (193.2)	1634.6 (246.9)	2750.0 (415.4)
全体	775.0 (250.1)	1828.8 (317.0)	2913.7 (405.3)

表 4: 5 母音の F1, F2 周波数の平均（それぞれ数秒の 1 フレーズから算出)。

曲番号	性別	/a/		/e/		/i/		/u/		/o/	
		F1 [Hz]	F2 [Hz]	F1 [Hz]	F2 [Hz]	F1 [Hz]	F2 [Hz]	F1 [Hz]	F2 [Hz]	F1 [Hz]	F2 [Hz]
No.001	男性	749.7	1605.5	538.7	2015.7	276.2	2055.4	244.1	924.6	405.2	1059.0
No.004	男性	672.6	1699.9	499.8	1779.8	286.5	1968.7	267.5	1339.3	430.0	1183.6
No.006	男性	650.9	1488.6	560.6	1737.0	359.7	2017.7	403.6	1313.6	596.5	1294.7
No.016	女性	797.3	1828.6	654.7	2264.8	310.2	2383.5	358.9	1384.6	585.2	1281.8
No.018	女性	747.5	1455.0	684.8	2161.2	416.0	2382.8	442.2	1852.5	620.4	1368.7
No.021	女性	558.9	1194.0	551.1	1236.0	381.8	1218.0	384.3	1287.0	445.5	1705.4

3 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（例えば [2] へも応用しやすい)。

3.1 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。前章での分析結果から、これらの特徴量はブレスのスペクトル包絡をよく表し、検出に有効であると考えられる。具体的な分析条件を**表 5**、アラインメント用文法を**図 3** に示す。

実験では、表 1 における RWC-MDB の 27 曲分の歌唱を用い、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 を学習した。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、Hidden Markov Model Toolkit (HTK) [16] を用いた。

3.2 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、それぞれ以下のように定義する。

表 5: ブレス検出のための歌唱音声分析条件

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

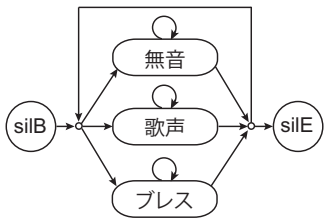

図 3: ブレス検出におけるアラインメント用文法

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

$$P = \frac{\text{正しく検出されたブレス数}}{\text{検出されたブレスの総数}} \times 100 \quad (2)$$

ここで、検出結果が正解のブレス位置と時間的に重なりがあれば正解とした。**図 4** に 27 曲の R と P を示す。それぞれ、全曲での平均は 97.5%, 77.7%であった[4] 。

図 5 に、ブレスが正しく検出された区間に対して、その始端と終端の推定時刻のずれの頻度分布を示す。横軸は、推定時刻からラベル付けした正解時刻とのず

[4] 正解判定を 1 フレーム毎に行って R と P を計算した結果は、それぞれ 95.3%, 71.2%であり、大きな性能低下はなかった。

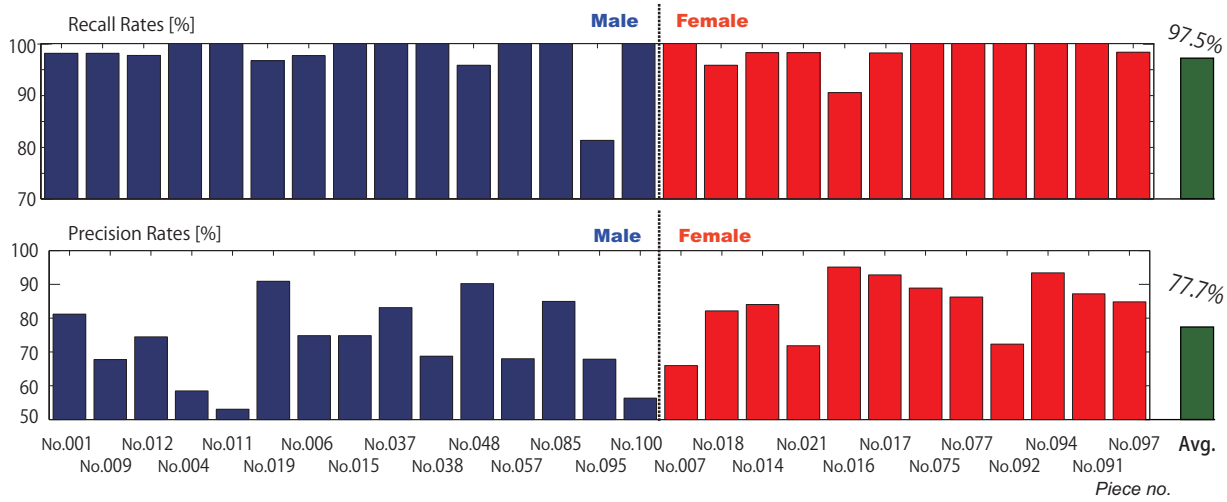

図 4: ブレス検出の再現率 (recall rate) と精度 (precison rate)

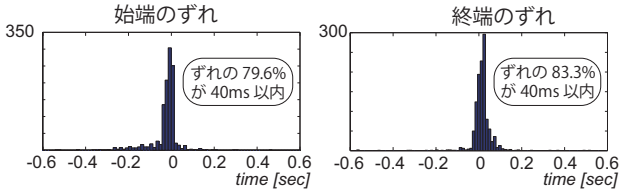

図 5: ブレス検出で推定された始端終端のずれの頻度分布

れを引いた値である。正解時刻から 40 msec 以内に収まった割合は、始端で 79.6%、終端で 83.3%であった。

3.3 考察

現在の実装では、ブレス以外の箇所の誤検出によって、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検出されることが多く、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[5] 。また、/h/などの子音を誤検出することもあり、/h/ はスペクトル包絡の形状がブレスに良く似ていた。精度が特に低かった No.011, No.100 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。

本手法は現時点では精度が比較的低いが、全ての曲に対して高い再現率が得られている。すなわち、得られた結果からブレスを選別する方法を導入できれば、このような誤検出へ対処できる可能性がある。

4 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、そのスペクトル包絡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詳細に調査し、自動検出法の精度向上を検討していく予定である。

[5] 335 箇所の誤り中 175 箇所。ブレスとして検出されたのは計 1509 箇所であった。

謝辞

本研究に対し有益な議論をして頂いた、藤原 弘将 氏 (産総研)、齋藤 毅 氏 (産総研) に感謝致します。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)、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] 中野, 後藤：VocaListener: ユーザ歌唱を真似る歌声合成パラメータを自動推定するシステムの提案. 情処研報 2008-MUS-75, pp.49-56, 2008.

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

[4] 西村, 伊東, 山崎: 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.

[5] 中村: 音楽における「間」と呼吸について, 音響学会 音楽音響研資 MA94-16, pp.19–26, 1994.

[6] 中野, 後藤, 平賀: 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.

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

[8] Price, Ostendorf, and Wightman: "Prosody and Parsing", In *Proc. DARPA Workshop on Speech and Natural Language*, pp.5–11, 1989.

[9] Wightman and Ostendorf: "Automatic Recognition of Prosodic Phrases", In *Proc. ICASSP 91*, pp.321–324, 1991.

[10] 堀内, 増田, 西田, 市川: 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 2007-MUS-71, pp. 1–6, 2007.

[11] 伊賀., 伊藤, 安村: 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 2000-HI-87, pp. 49–54, 2000.

[12] 後藤, 橋口, 西村, 岡: 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, 西村: 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 2005-MUS-61, pp.7–12, 2005.

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

[15] Sundberg: *The Science of the Singing Voice*, Northern Illinois Univ Pr, 226p., 1987.

[16] Young, Evermann, Hain, Kershaw, Moore, Odell, Ollason, Povey, Valtchev, and Woodland: *The HTK Book*, Version 3.2.1, 346p., 2002.