Vol.2021-MUS-131 No.20
Vol.2021-SLP-137 No.20
2021/6/18

歌声の高音・長音の分析に特化した音響モデルの構築

内藤 悟嗣[2,a)]　齋藤 康之[1,b)]

概要：歌声合成技術やカラオケの採点技術などのシステム構築には，歌声の特徴を音素単位で解析する必要があり，ある音素と音源の時間的な対応付け作業をアノテーションという．手動によるアノテーションは，精密な解析結果が得られる一方で，甚大な時間と労力を要する．そのため自動でアノテーションするツールが開発された．しかし，ツールに付属している音響モデルは話声や読み上げ音声に対して学習されており，高音や長音の特性を持つ歌声に対しては認識精度が低下するという問題がある．以上の問題点を解決するために，本研究は隠れマルコフモデルを用いて歌声に対して学習した歌声特化の音響モデルを構築する．そして，高音・長音を含む歌声における音素の発声時刻の推定精度について，先行研究で構築された話声音響モデルによる推定結果と比較・評価を行う．

Construction of an acoustic model specialized for analysis of high pitchs and long tones in singing voice.

Satoshi Naito[2,a)]　Yasuyuki Saito[1,b)]

Abstract: The purpose of this study is to construct an acoustic model trained on singing voices. For the construction of systems such as singing voice synthesis technology and scoring technology for karaoke, it is necessary to analyze the characteristics of singing voices. In manual phonetic analysis, it takes a great deal of time and effort to correspond sentences to speech sounds. For this reason, automatic annotation tools have been developed. However, the acoustic model attached to the tools is trained on spoken voices, and the recognition accuracy of singing voices is reduced. In order to solve the above problems, we construct an acoustic model trained on singing voices using hidden Markov model, and compare and evaluate it with conventional methods.

1. はじめに

現代では歌声合成技術やカラオケの自動採点技術など，歌とコンピュータ技術が融合した娯楽システムが社会に浸透している．こうしたシステムの構築には，音素単位で歌声の特性を解析する必要があり，ある音素と音源の時間的な対応付け作業をアノテーションという．しかしながら，例として歌詞が 400〜500 文字程度の楽曲には，800〜1000 個程度の音素が含まれているため，全ての音素を手動でアノテーションするには甚大な時間と労力を要する．そこで，発声文と音源から自動でアノテーションをするツールが存在する．しかし，当ツールに付属する音響モデル（単語と音の関係を表す確率モデル）は，話声（朗読音声）に対して学習されているため，高音や長音を含む歌声に対しては認識精度が低下するという問題点がある [1]．

上記の問題を解決するために，本研究は歌声をもとに学習した「歌声に特化した音響モデル」を構築する．そして，高音や長音の特性を含む歌声に対し，音素の発声時刻の推定精度について従来の話声音響モデルと比較・検証を行う．

2. 音響モデルの構築

2.1 学習・評価用データ作成

まず男性と女性が歌唱する楽曲をそれぞれ複数用意する．ただし，日本語の歌唱部が含まれており，複数人の歌唱音声が重なっていない楽曲を対象とする．そして，これ

[1] 木更津工業高等専門学校 情報工学科, 2-11-1 Kiyomidai-Higashi, Kisarazu-shi, Chiba 292-0041, Japan.
[2] 千葉大学 工学部 情報工学コース, 1-33 Yayoi-chou Inage-ku, Chiba-shi, Chiba 263-8522, Japan.
[a)] 19t1692w@student.gs.chiba-u.jp
[b)] saito@j.kisarazu.ac.jp

表 1　音素の種類
Table 1　Types of phonemes

発音の種類	音素
母音	a, i, u, e, o
清音	k, s, t, ts, n, h, f, m, y, r, w
濁音	g, z, d, b
半濁音	p
拗音	ky, sh, ch, ny, hy, my, ry
拗濁音	gy, j, by, dy
拗半濁音	py
促音	q
撥音	N
長音	a:, i:, u:, e:, o:
無音	silS, silE, sp

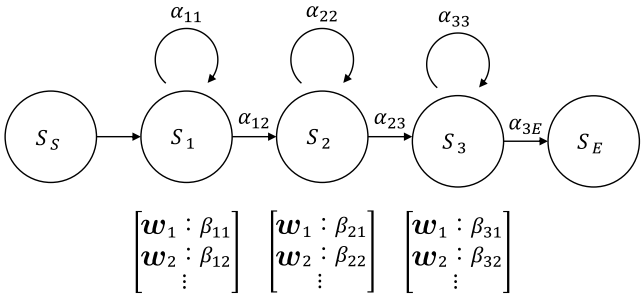

図 1　3 状態の HMM の構造
Fig. 1　Structure of 3-state HMM

らの楽曲からボーカル音のみを抽出する．

次に上記の楽曲の歌詞を用意し，（あらゆる→ arayuru）のように音素へ変換する．音素は表 1 の 43 種類存在する．このとき，「母音:」は発声長音，「silS」は歌唱開始前の無音区間，「silE」は歌唱後の無音区間，「sp」は歌唱中に存在する無音区間と定義する．また，助詞として扱う場合の「へ」や「は」は「e」や「wa」，「を」は「wo」のように，実際の発声音に合わせた音素に置き換える必要がある．

最後に楽曲を聴きつつスペクトログラムを確認しながら全て手動で音素の発声時刻を付与していく．ただし，用意した全楽曲のうち，出現頻度が低い（5 個以下の）音素については，自身の音声で補填を行う．

2.2　HMM の構築

ある単語 $l_1, \cdots, l_n$ を発声したとき，音声から抽出された特徴ベクトル系列 $\boldsymbol{w}_1, \cdots, \boldsymbol{w}_t$ が観測される確率を音響モデル $P(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_t | l_1, \cdots, l_n)$ で表す．従来の音響モデルの構築には HMM（hidden Markov model : 隠れマルコフモデル）が使われてきた [2]．HMM とは，状態を外部から直接観測できない（隠れている）マルコフモデルのことであり，状態に応じた確率分布に従って信号を出力する．したがって，各音素に対応した HMM を構築することで，特定の音素を発声したときの音声特徴ベクトル系列の観測される確率が得られる．特に，一方向にしか遷移しない HMM を left-to-right モデルと呼び，状態数 3 の HMM は図 1 のようにして表される．S_S を初期状態，S_E を終了状態とするとき，i 番目の状態 S_i は，パラメータとして，状態 S_i から S_j への遷移確率 α_{ij}

$$\alpha_{ij} = P(S_i \to S_j) \tag{1}$$

と，音声から抽出された特徴ベクトル $\boldsymbol{w}_k$ の出力確率 β_{ik}

$$\beta_{ik} = P(\boldsymbol{w}_k | S_i) \tag{2}$$

を持っている．また，HMM の各状態が持つ音響特徴量のパターンに GMM（Gaussian Mixture Model : 混合正規分布）を組み込んだものが GMM-HMM である．

上記のような HMM が持つ状態遷移確率および信号出力確率などのパラメータは，Baum-Welch アルゴリズム [3] により学習される．Baum-Welch アルゴリズムとは，反復的に尤度を求め，統計的にパラメータを推定していく EM アルゴリズムの一種である．

本研究では歌声からの特徴抽出値を学習データとして扱う left-to-right の GMM-HMM で音響モデルを構築する．

表 2　実験に使用するツール
Table 2　Tools used in the experiments

機能	ツール名
ボーカル音抽出	spleeter [6]
ラベル付与・調節	WaveSurfer [7]
特徴抽出，HMM 作成	HTK [8]
自動アノテーション	Julius segmentation kit [9]

表 3　特徴抽出の条件
Table 3　Conditions for feature extraction

条件種類	条件の値
標本化周期	16 [kHz]
ハイパスフィルタ	0.97
分析窓	ハミング窓
フレーム長	25 [ms]
フレーム周期	10 [ms]
フィルタバンク	24 チャネル

3.　音響モデルの実験的評価の手法

3.1　音響モデルの構築条件

本実験で用いるツールを表 2 に示す．また，音声特徴の抽出は表 3 の条件で行う．音響モデルの学習および評価には，RWC 研究用音楽データベース（ポピュラー音楽） [4] の楽曲を使用する．音響モデルの汎用性を高めるために，様々な曲調の男性歌唱楽曲と女性歌唱楽曲を各々 10 曲ずつ用意し，楽曲をさらにフレーズごとに分割したものを用いる．使用する楽曲は表 4 の通りである．

3.2　音響モデルの評価

音響モデルの高音・長音の推定性能は，式 (3) に示す，

表 4　実験に使用する楽曲
Table 4　Musics used in the experiments

男性歌唱楽曲		女性歌唱楽曲	
No.	楽曲名	No.	楽曲名
1	永遠のレプリカ	3	HORO（女性歌唱部のみ）
6	Funky Life	7	PROLOGUE
9	慟哭	13	キャッチボール
11	言えない	16	Game of Love
12	KAGE-ROU	17	あなたと逢えて
15	old fashioned	18	True Heart
22	恋におちる時間に関する考察	20	ときめきの瞬間
24	it's all right	21	Feeling In My Heart
29	One Two STEP	25	tell me
32	what could I do for you	28	Fly away

N 個の音素の発声開始時刻の推定結果と正解時刻との総誤差時間 L で評価する．ここで，y_k は k 番目の音素発声の開始推定時刻，t_k は正解時刻を示す．

$$L = \sum_{k=0}^{N-1} |y_k - t_k| \tag{3}$$

また，音響モデルの構築条件において，不適切な音響モデルの構造により，探索に失敗して音素の位置が定まらず，誤差時間が得られない場合が考えられる．そのため音響モデル同士の性能の比較については，式 (4) に示す，N 個の音素の発声開始時刻の推定結果と正解時刻との平均誤差時間 L_{ave} で，1 音素あたりの平均誤差時間で評価する．

$$L_{ave} = \frac{1}{N} L \tag{4}$$

HMM の状態数を M としたとき，音響モデルの比較条件として，(1) M = 3 の場合，(2) M = 4 の場合，(3) M = 5 の場合の各々に対し，表 5 に示す 3 つの音響特徴パラメータ条件の全組み合わせで学習された音響モデルを用意し，従来の話声（朗読音声）で学習された音響モデル「hmmdefs_monof_mix16_gid.binhmm」 [5] との歌声におけるアノテーション精度の比較・評価を行う．

4. 実験結果

今回の実験で用いた PC は，CPU Intel(R) Core(TM) i7-8700K，メモリ 16GB，Windows 10 Home である．楽曲の時間長が 15〜30 [sec] 程度で 100〜150 個ほどの音素が含まれるとき，アノテーション処理時間は 0.5 [sec] 前後であった．

4.1 音響モデルによる音素推定の精度の比較

1 曲単位で交差的にモデルの性能を測る leave one out 法により得られた各モデルの平均誤差時間を表 6 に示す．表 6 より，特徴次元数が多いほど平均誤差時間が短くなっていることがわかる．また，状態数が多いほど平均誤差時間が短くなっていくことがわかる．しかし，HMM の状態数 M = 4 における音響特徴パラメータ条件が c_1 のときに楽曲 No. 13 と No. 15， HMM の状態数 M = 5 における音響特徴パラメータ条件が c_1 のときに楽曲 No. 13 と No. 15， c_2 のときに楽曲 No. 13 の探索に失敗し，アノテーション位置が定まらなかった．そして，歌声に対して学習をした音響モデルは，どれも従来の音響モデルよりも平均誤差時間が短いという結果が得られた．

表 6 より HMM 状態数 M = 5 で音響特徴パラメータ条件 c_3 の音響モデルは，推定性能を維持しつつ，汎用性を保っているため，以後この条件で学習された音響モデルを「歌声特化の音響モデル」と呼ぶ．

表 5　音響特徴パラメータ条件
Table 5　Acoustic feature parameter conditions

パラメータ名	パラメータの次元
c_1	MFCC の 12 次元と平均パワーの計 13 次元
c_2	c_1 に 1 次差分（Δ MFCC）を加えた 26 次元
c_3	c_2 に 2 次差分（ΔΔ MFCC）を加えた 39 次元

表 6　モデルの性能比較
Table 6　Model performance comparison

HMM の状態数 M	音響特徴パラメータ条件	平均誤差時間 [sec]	アノテーション位置の探索に失敗した楽曲
3	c_1	0.173	–
	c_2	0.088	–
	c_3	0.075	–
4	c_1	0.145	13, 15
	c_2	0.071	–
	c_3	0.050	–
5	c_1	0.110	13, 15
	c_2	0.060	13
	c_3	0.043	–
先行研究の話声音響モデル		0.308	–

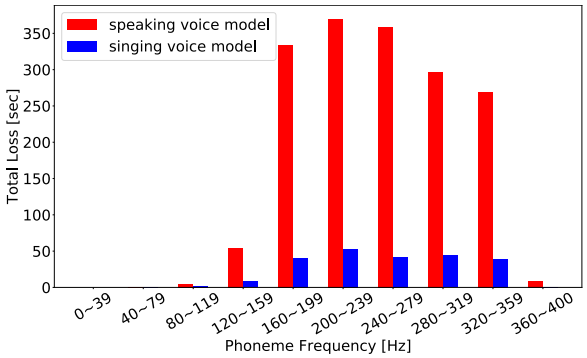

図 2　高音による総誤差時間 L の比較
Fig. 2　Comparison of total error time L by treble

4.2 歌声特化の音響モデルの高音に対する推定性能

歌声特化の音響モデルと従来の話声音響モデルとの音素の発声高による総誤差時間の比較を図 2 に示す．ただし，

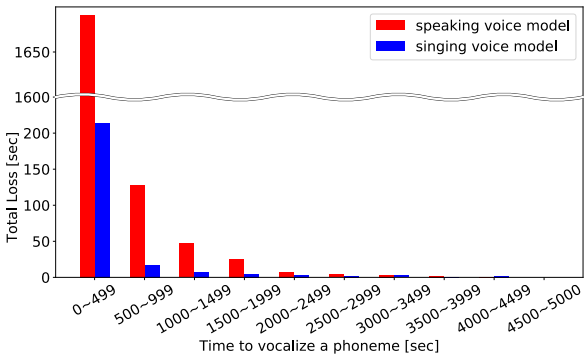

図 3　長音による総誤差時間 L の比較
Fig. 3　Comparison of total error time L by long vowel

音素のうち「silS」,「silE」,「sp」は，実際に発声しない音素であるため，総誤差時間からは除外するものとする．この図 2 から，従来の話声音響モデルと歌声特化の音響モデルは，ともに音素の発声高が 200〜239 [Hz] で総誤差時間がピークとなり，話声音響モデルの総誤差時間が 370.19 [sec] に対して歌声特化の音響モデルの総誤差時間は 53.19 [sec] である．以上のことから，ピーク時の歌声特化の音響モデルによる総誤差時間は従来の話声音響モデルの総誤差時間より約 14 %に抑えられていることがわかる．また，音素の発声高が 240 [Hz] 以降も従来の話声音響モデルより歌声特化の音響モデルの方が総誤差時間が下回ることから，高音に関する推定性能が向上したといえる．

4.3　歌声特化の音響モデルの長音に対する推定性能

歌声特化の音響モデルと従来の話声音響モデルとの音素の発声長による総誤差時間の比較を図 3 に示す．こちらも高音性能の比較時と同様に，音素のうち「silS」,「silE」,「sp」に関しては，実際に発声しないため除外する．この図 3 から，従来の話声音響モデルと歌声特化の音響モデルはともに音素の発声長が 0〜499 [msec] で総誤差時間がピークとなり，話声音響モデルの総誤差時間が 1960.53 [sec] に対して歌声特化の音響モデルの総誤差時間は 214.31 [sec] である．以上のことから，歌声特化の音響モデルによる総誤差時間は従来の話声音響モデルの総誤差時間より約 12 %に抑えられていることがわかる．また，発声長が 500 [msec] 以降でも話声音響モデルより歌声特化の音響モデルの方が総誤差時間が下回ることから，長音に関する推定性能が向上したといえる．

4.4　歌声特化の音響モデルの音素別の推定性能

歌声特化の音響モデルの各音素について，推定の平均誤差時間が短い順に並べた音素を表 7 に示す．表 7 より，発声の種類が清音であり学習音素数が 100 を超えている音素が平均誤差時間が短い．また，濁音や拗濁音，拗音は学習音素数が 100 未満であるにもかかわらず，平均誤差時間が短い．そして，母音は学習音素数が数百以上であるが，平均誤差時間が他の子音よりも長い傾向がある．

表 7　平均誤差時間が短い音素
Table 7　Phonemes with short mean error time

音素	学習音素数	平均誤差時間 [s]	発音の種類
sh	121	0.012	清音
t	392	0.017	清音
b	64	0.019	濁音
j	29	0.019	拗濁音
d	148	0.019	濁音
k	431	0.023	清音
ch	33	0.023	拗音
s	150	0.028	清音
a	781	0.030	母音
ky	10	0.031	拗音

表 8　平均誤差時間が長い音素
Table 8　Phonemes with long mean error time

音素	学習音素数	平均誤差時間 [s]	発音の種類
p	9	0.196	半濁音
i:	188	0.089	長音
u:	116	0.088	長音
e:	165	0.084	長音
r	261	0.064	清音
q	61	0.062	促音
gy	12	0.057	拗濁音
o:	241	0.054	長音
f	43	0.050	清音
g	77	0.050	濁音

次に，歌声特化の音響モデルの各音素について，推定誤差時間が長い順に並べた音素を表 8 に示す．表 8 より，学習音素数が少なく，半濁音である「p」は他の音素よりも大きく離れて平均誤差時間が長い．また，発声の種類が母音である長音が，平均誤差時間が長くなる傾向にある．

5.　検討事項

5.1　認識結果の視覚化と実用性の検討

認識結果の例として，RWC 研究用音楽データベース（ポピュラー音楽）の楽曲番号 No. 6 の一部の正解ラベル位置と話声音響モデルによる推定ラベル位置との関係を図 4 ，正解ラベルと歌声に特化した音響モデル（HMM の状態数 M = 5，音響特徴パラメータ条件 c_3）による推定ラベル位置との関係を図 5 に示す．

従来の話声音響モデルのアノテーションは，図 4 のように探索位置を誤って推定し，一部の時刻にラベルが集中していることがわかる．その一方，図 5 は複数の音素を跨ぐほどの誤差が解消されていることがわかる．

しかしながら，正解ラベルと完全に一致しているとはいえない．加えて，実験結果の表 6 から，1 つの音素あたり

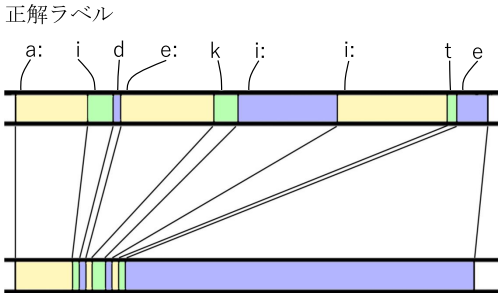

図 4 先行研究の話声音響モデルによるラベル位置（RWC 研究用音楽データベース（ポピュラー音楽）の楽曲 No.6）

Fig. 4 Label position by the speaking voice acoustic model in the previous study

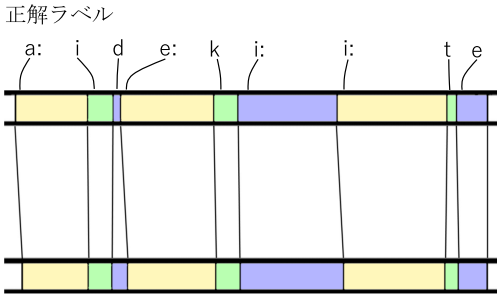

図 5 提案手法の歌声音響モデル (M=5, c_3) によるラベル位置（RWC 研究用音楽データベース（ポピュラー音楽）の楽曲 No. 6）

Fig. 5 Label positions by the proposed method's singing voice acoustic model (M=5, c_3)

0.0043 [sec] のアノテーション誤差があるといえる．このことから，アノテーションを全て自動化するのではなく，自動でのアノテーション後に大きく誤差のある部分を手動で修正する半自動方式が，本研究で構築した歌声に特化した音響モデルの現実的な実用法であると考えられる．

5.2 認識精度の改善

5.2.1 アノテーション位置の探索成功率の改善

実験結果の表 6 より，状態数が多いほど平均誤差時間が短くなる半面，特徴次元数が少ないとアノテーション位置の探索に失敗してしまう問題が起きた．これは，状態数が多くなるほど学習データに適合しすぎて，未知のデータに対して判別性能が落ちてしまったことが原因ではないかと考えられる．この問題は，学習に使用するデータ数を増やして汎用性を高めることで解決できると考えられる．

5.2.2 音響モデルの改善

本研究では，前後の音素を考慮しないモノフォン形式で各 HMM を定義・学習した．その一方，たとえば，注目音素とその前後の音素を考慮したトライフォン形式の HMM が存在する．声は連続で発声する場合，たとえ同じ音素であっても，続く前後の音素によっては特性が変わる．例として発音「byu」のスペクトログラムを図 6，発音「kyu」のスペクトログラムを図 7 にそれぞれ示す．「byu」と「kyu」にはどちらも音素「u」が含まれているが，子音と隣接する境界では，音の特性が異なることがわかる．以上より，トライフォンで細かくクラス分けすることで認識表現の曖昧さを回避できる．しかし，モノフォンモデルに比べ膨大な量の HMM を学習する必要がある点と，用意できるサンプルに限りがあるため，今回は実現できなかった．上記の問題を解決できれば，さらなる精度の向上が期待される．

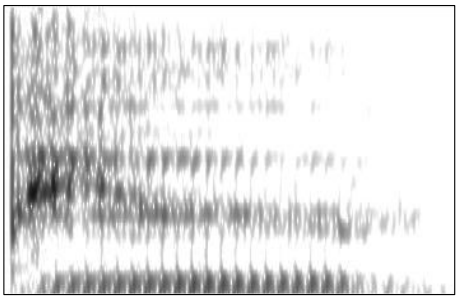

図 6 発音「byu」のスペクトログラムの例

Fig. 6 Example spectrogram of the pronunciation "byu"

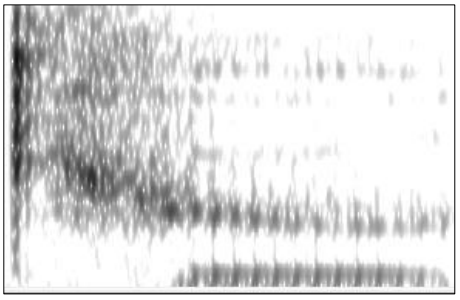

図 7 発音「kyu」のスペクトログラムの例

Fig. 7 Example spectrogram of the pronunciation "kyu"

また，表 8 から，母音の長音は平均誤差時間が大きくなる傾向が確認できた．これは，数ミリ秒ほどの長音から，数秒にわたるビブラートを含む長音を一つの音素として学習させたことが原因ではないか考えられる．たとえ同じ音素であっても，ビブラートを含む長音は細かく音程が上下するため，一定の音程で発声する歌声とは特徴が異なる．そのため，ビブラートの細かい音程の上下に対応できる専用の HMM を別途構築することで，母音長音の推定精度の改善が期待できる．

そして，2010 年以降から GMM の確率計算部分を DNN(Deep Neural Network) に置き換えた DNN-HMM が登場した [2]．この DNN-HMM は，多層のネットワークを築くことで，GMM よりも認識表現の幅が広がる．したがって，DNN-HMM を用いて音響モデルを構築することでも，精度の向上が期待される．

6. まとめ

音声解析において音素と音声の時間的対応付けには，手動では甚大な時間と労力を要する．自動でアノテーションをするツールが存在するが，従来の話声（読み上げ音声）で学習された音響モデルは，歌声に対しては精度が低下してしまうという問題点があった．そこで，本研究は歌声を中心に学習させた歌声に特化した音響モデルを GMM-HMM を用いて構築した．HMM の状態数 M = 5 で表 5 の音響特徴パラメータ条件 c_3 で歌声に対して学習したとき，音素の発声時刻の推定結果と正解時刻との平均誤差時間は 0.043 [s] という結果が得られ，話声音響モデル（平均誤差時間 0.308 [s]）よりも精度の良いモデルが構築できたといえる．また，発声高で総誤差時間を比較したとき，歌声特化の音響モデルは従来の話声音響モデルより総誤差時間が下回ったことから高音の推定性能が向上したといえる．そして，発声長で総誤差時間を比較したとき，歌声特化の音響モデルは従来の話声音響モデルより総誤差時間が下回ったことから長音の推定性能が向上したといえる．

今後の課題として， 4 点の改善案を示す．1 点目は，HMM を学習させるデータ数を増やすことである．様々な状況の音素データを学習させることで汎用性を向上させることにより，探索失敗の回避が期待される．2 点目は，トライフォン形式で音響モデルを構築することである．本研究において，注目音素のみ考慮したモノフォン形式の音響モデルを構築したが，さらに前後の音素にも注目するトライフォン形式の音響モデルを構築することで，発声特徴の曖昧さ回避が期待される．3 点目は，母音長音に対して，ビブラートを含む音声に対応する HMM を別途定義することである．一定の音程の長音とビブラートによる長音の特徴について曖昧さ回避が期待される．4 点目は， GMM の確率計算部分を DNN に置き換えた，DNN-HMM を用いて音響モデルを構築することである．多層にわたるネットワークにより，認識表現の幅が広がることが期待される．

本研究の成果は，歌声特化の音声認識や歌声解析などに実用的に応用できると考えられる．ただし，本研究で構築した音響モデルによるアノテーション位置は，正解ラベルの位置と完全に一致するとは限らず，1 つの音素あたり 0.0043 [sec] のアノテーション誤差が生じている．このことから，アノテーションが必要な場合は，アノテーションツールの音響モデルを本研究で構築した歌声特化の音響モデルに組み替え，自動でのアノテーション後に，大きく異なるラベル位置を手動で修正をする半自動方式による活用が現実的であると考えられる．今後，上記の応用分野に実際に適用して，効果を検証したい．

謝辞 本研究は JSPS 科研費 21H03462 の助成を受けたものです．

参考文献

[1] 尾関 弘尚, 鎌田 貴幸，後藤 真考，速水 悟，“歌声の歌詞認識における音高の影響について”, 日本音響学会 2003 年秋季講演論文集，1-1-1，pp. 637-638，Sep. 2003.
[2] 河原 達也, “音声認識技術の変遷と最先端:——深層学習による End-to-End モデル——”, 日本音響学会誌, vol. 74, no. 7, pp. 381-386, 2018.
[3] 村上 仁一, “Baum-Welch アルゴリズムの動作と応用例”, 電子情報通信学会 基礎・境界ソサイエティ Fundamentals Review, vol. 4, no. 21, pp. 48-56, 2010.
[4] 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：ポピュラー音楽データベースと著作権切れ音楽データベース”, 情報処理学会 音楽情報科学研究会 研究報告, 2001-MUS-42-6，vol．2001，no．103，pp. 35-42, Oct. 2001.
[5] “Model trained on conventional reading voice”, 入手先 ⟨http://sap.ist.i.kyoto-u.ac.jp/dictation/doc/phone_m.pdf⟩
[6] “A tool for extracting vocal sounds”, 入手先 ⟨https://github.com/deezer/spleeter⟩
[7] “Tool to add labels and adjust their position”, 入手先 ⟨https://sourceforge.netprojects/wavesurfer⟩
[8] “Tool for feature extraction and HMM training”, 入手先 ⟨http://htk.eng.cam.ac.uk⟩
[9] “Speech Segmentation Toolkit using Julius”, 入手先 ⟨https://github.com/julius-speech/segmentation-kit⟩