歌声の母音変化を考慮した歌声合成の検討*

☆片平健太 (神戸大), 足立優司 (メック株式会社), 田井清登 (メック株式会社),
高島遼一 (神戸大), 滝口哲也 (神戸大)

1 はじめに

歌声合成システムは，任意に与えられた楽譜の歌詞や音符の音高，長さなどの情報から歌声を合成する音声合成システムの一つである．現在，歌声合成システムは主に娯楽分野において広く普及しつつある．またこのシステムは故人の歌声の再現や病気等で声を失った患者の歌声を再現するなどの応用手法としての利用も考えられる．

現在主に研究されている歌声合成手法として，波形接続歌声合成と統計的パラメトリック歌声合成が挙げられる．波形接続音声合成 [1] は細かな音声波形を組み合わせることにより歌声を合成する手法である．この手法では自然な音声を生成することが可能であるが，そのためには多量の合成元音声が必要となりモデルが肥大化する問題が存在する．また歌声データベースの中に滑らかに接続できる音声単位が存在しない場合は不連続を伴う合成音となり品質が劣化する．これに対し統計的パラメトリック歌声合成は，歌声データベースから統計モデルを構築し，音声パラメータを生成する．構築されたモデルは波形接続型モデルに対して比較的小さいものとなる．これまで主に隠れマルコフモデル (HMM) を用いた手法 [2] が広く研究されてきたが，近年ではより高品質な歌声を合成できる深層学習を用いた統計的パラメトリック歌声合成手法 [3, 4] が提案されている．

これらの歌声合成では主に一般的な歌唱音声を用いてモデル学習を行うが，本研究では表情豊かな歌声の合成を目的とし，より特殊性，専門性の高い歌唱音声であるオペラ歌唱音声を対象歌声音声として用いる．オペラ歌唱では一般的な歌唱と比較して周波数分布が異なるなど，様々な要因によって聴者の聴こえ易さなどに影響を与え，歌声のオペラらしさを特徴づけている．この一般的な歌唱とオペラ歌唱の相違点を上手く捉えることは，オペラという枠組みを越え音声合成や声質変換などに応用が可能である．例として雑音状況下においても聞き取り易いスピーチ音声の生成や音声に抑揚などの表現を付加することで感情を表現したり，説得力のある音声を生成することなどが考えられる．

日本語のオペラ歌唱では，歌唱音声に通常の a, i, u, e, o の他に曖昧化した母音が現れることがあり，オペラ歌唱が一般歌唱と異なるものである要因の一つとなっている．本研究ではこの母音の変化に着目した歌声合成を検討する．はじめに歌声における母音の発音の分布から，母音の曖昧化の傾向を分析する．そしてこの傾向を考慮した DNN 歌声合成手法を検討する．

2 母音の変化の分析

2.1 母音の曖昧化

母音は舌の位置，唇の形，顎の開きの度合いによって発音が決定される．日本語の a, i, u, e, o の 5 母音に関して顎の開口度をみると，a, o は開口度が大きく，i, e, u は開口度が小さい．

日本語のオペラ歌唱では，母音の発声において曖昧化した母音が現れることがある．オペラ歌唱では母音の発音時に顎の開口度を大きくすることで第 1 フォルマントを上昇させ [5]，歌声を聴者に聴こえ易くする．よって，i, e, u の発音時でも顎を大きく開いたまま発音することにより a, o に近い曖昧化した母音が出現する．なおそれぞれの曖昧化した母音は，舌の位置や唇の形の関係より，i, e は a に，u は o に近い発音となる．

Fig. 1 に u, o, そして曖昧化した u のスペクトルを示す．なお，いずれの母音も C5 の音程で発声したものである．u のスペクトルでは 100 次元目付近の周波数成分が強く山が見られるが，o のスペクトルではこの山は存在しない．曖昧化した u のスペクトルの 100 次元目付近ではこのような山は見られず， u, o のスペクトルと比較すると，その概形は o に近いことが分かる．

2.2 曖昧化の傾向

本研究では曖昧化した母音を考慮した歌声合成を行うため，初めに曖昧化した母音の出現の傾向について分析を行った．

プロのオペラ歌手による日本語オペラ歌唱音声 48 曲 (93 分) の音声データより母音の発声区間のみを抽出し，WORLD ボコーダ [6] を用いてスペクトル包絡を得る．更にメルケプストラムに変換し，時間平均を

*Singing Voice Synthesis Considering Vowel Variations. by Kenta Katahira (Kobe Univ.), Yuji Adachi (MEC Company Ltd.), Kiyoto Tai (MEC Company Ltd.), Ryoichi Takashima (Kobe Univ.), Tetsuya Takiguchi (Kobe Univ.)

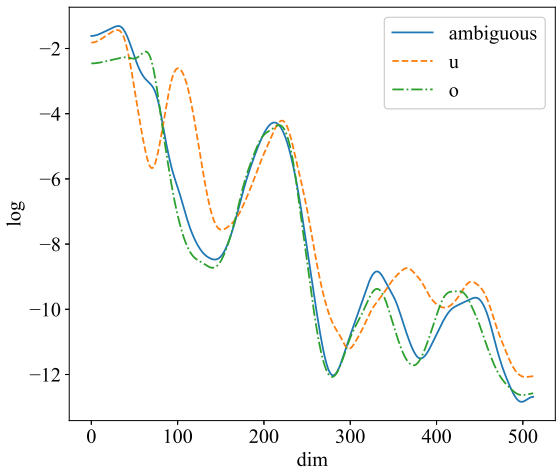

Fig. 1　Comparison of vowel spectrums.

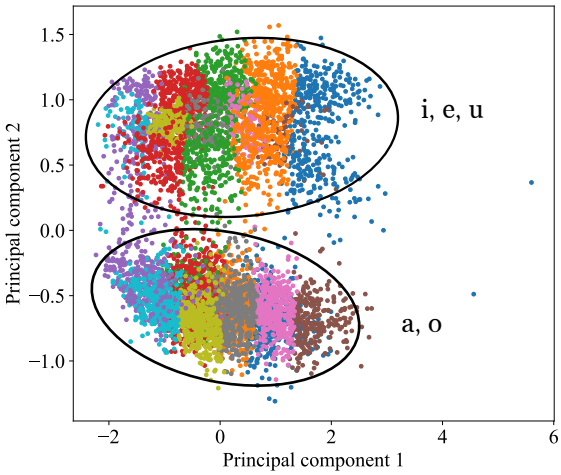

Fig. 2　Clustering results of vowels.

取ったものを母音発声データとする．このデータを主成分分析したのち，各母音 (a, i, u, e, o) ごとに k 平均法による非階層型クラスタリングを行い，それぞれのクラスタの分布を分析した．音声のサンプリングレートは 16kHz，フレームサイズは 256 サンプル，FFT は 1024 次元，メルケプストラムは 60 次元とした．また主成分分析では第 10 主成分まで使用し，寄与率は 95.63%であった．k 平均法では $k = 5$ とした．

はじめに全ての母音発声データに対して分布の分析を行った．Fig. 2 に全母音のクラスタリング結果を示す．Fig. 2 では第 2 主成分の 0 付近で母音発声データが大きく 2 つに分かれて分布しており，0 以上のものが i, e, u の顎の開口度が小さい母音であり，0 以下のものが a, o の顎の開口度が大きい母音であった．これから第 1 主成分は母音発音時の顎の開口度と相関があると考えられる．また顎の開口度との関係により，i, e, u の曖昧化した母音は第 2 主成分の 0 付近または 0 以下に分布していると推測できる．

次にそれぞれの母音におけるクラスタリングの結果について記述する．Fig. 3 は母音 i におけるクラスタリングの結果である．各クラスタは垂直方向に切り分けられる形で境界が設定されており，クラスタ分類が第 1 主成分に強く依存していることが分かる．

ここで各クラスタに属する母音発声データの平均発音周波数を求めた．Table 1 に各クラスタとその平均周波数を示す．各クラスタの平均周波数は i0, i1, i2, i3, i4 の順に高くなることがわかった．これは Fig. 3 の各クラスタの分布と比較すると，第 1 主成分が小さくなるほど母音の発音周波数は高くなっており，第 1 主成分と母音の発音周波数には負の相関があると考えられる．この傾向は i に限らずすべての母音において観測された．

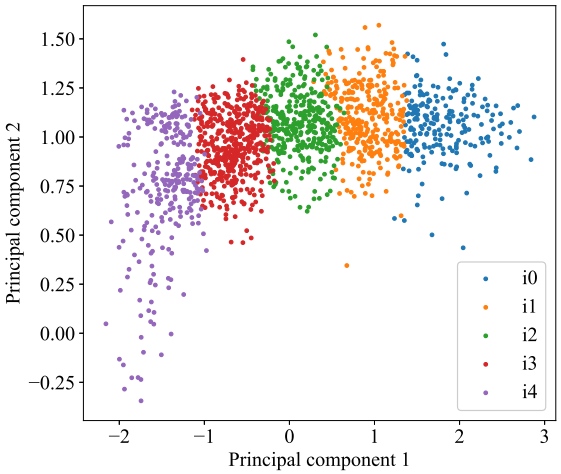

Fig. 3　Clustering results of ”i”.

次に母音 u に関してクラスタリングを行った．Fig. 4 にその結果を示す．u に関しても周波数帯域別によるクラスタリングの傾向が見られたが，第 1 主成分が 0 以下であり，第 2 主成分が 0.25 以下のものは u2, u3 とは異なるクラスタ u4 として分類された．u4 の発音分布は Fig. 2 と比較すると a, o の発音分布と重なっており，またその発音は u よりも o に近く，曖昧化した母音 u のクラスタであることが分かった．

ここで，第 1 主成分における発音分布が重なるクラスタ u3, u4 について，発音時の音程とその出現数の関係を Fig. 5 に示す．なお横軸は音程番号であり，60 が C5 の音程に対応し，1 増加すると音程は半音上昇する．Fig. 5 より音程番号 62，つまり D5 よりも高い音程では u4 のみ存在する．逆に D5 以下では u4 である発音の個数は減少し，u3 の個数は増加している．これらより u の発音では母音発音時の音程が高くなるほど曖昧化しやすいことが分かった．

Table 1　Mean frequencies of clusters of ”i”.

Cluster	i0	i1	i2	i3	i4
Mean Frequency [Hz]	288.49	343.19	403.99	482.88	587.21

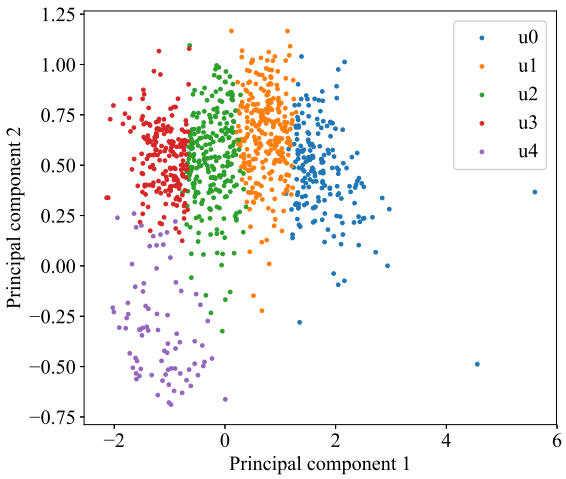

Fig. 4　Clustering results of ”u”.

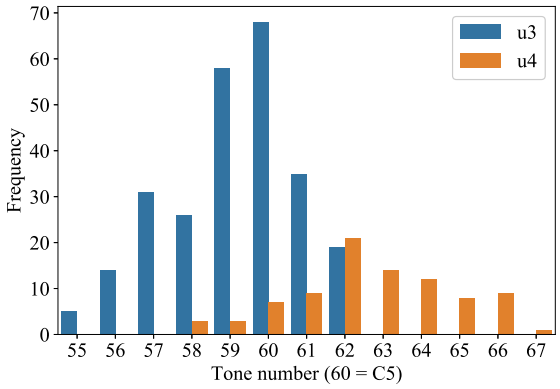

Fig. 5　Frequency distribution of ”u3” and ”u4”.

Fig. 3 の i の分布においても Fig. 4 の u の分布と同様に，第 1 主成分の値が負になるほど，つまり母音発音時の音程が高くなるほど第 2 主成分が負であり a の分布に近づくため，高音での母音発音が曖昧化する傾向があると考えられる．

3　母音の曖昧化を考慮したオペラ歌唱合成

3.1　楽譜特徴量

楽譜特徴量は，MusicXML 形式の楽譜データを解析して抽出した．楽譜特徴量には，当該音素に前後２つの音素を加味した音素情報や音素の位置情報を含んでいる．また，音節における音符位置，音高，音価，小節やフレーズ中の位置情報などが含まれる．音符情報にはタイ・スラーの有無，強弱などが含まれている．これらの特徴量をバイナリや連続値，one-hot 形式で表現し，音響モデルの入力とする．各音素長は HMM を用いた音素アライメントにより推定された値を修正して求めた．

本研究では母音の変化を考慮した音響モデル学習を行うため，母音の発声時に 2.2 節で割り当てられた母音クラスタ計 25 個を one-hot 形式で表現した 25 次元のベクトルを付加した．

3.2　深層学習を用いた音響特徴量推定

学習時，音響モデルの入力として楽譜特徴量，教師としてオペラ歌唱音声から抽出された音響特徴量を用いる．音響特徴量の推定には系列内変動 (GV) を考慮した Bidirectional Gated Recurrent Units (Bi-GRU) を用いた深層学習モデル [7] を使用する．Bi-GRU により過去と未来の時間変化を考慮した学習を行うことが可能である．またトラジェクトリ学習によりモデルより推測された静的特徴量と動的特徴量を用いて尤もらしい特徴量系列 (トラジェクトリ) を推定する．その後，特徴量系列から時系列における静的特徴量ベクトルの分散で表現される GV を算出し，特徴量系列と GV の誤差をそれぞれ求め，モデルの重み更新に用いる．

4　実験評価

4.1　実験条件

実験には，女性オペラ歌手 1 名による日本語オペラ歌唱音声 48 曲からなる約 93 分の音声データセットを用いる．このうち 43 曲を音響モデルの学習に，5 曲をテストに用いた．音声のサンプリング周波数は 16kHz, フレーム長は 256 サンプルとした．

この実験では音響モデルの入力特徴量に 3.1 節で述べた従来と同じ楽譜特徴量のみ 534 次元と，母音クラスタ情報を付加したもの 559 次元の 2 つを用いてモデルを比較する．音響特徴量には，WORLD[6] によって抽出されるスペクトル包絡から計算したメルケプストラム 60 次元，対数基本周波数 1 次元，帯域非周期成分 1 次元とこれらの 2 次までの動的特徴量に加えて有声・無声パラメータ 1 次元を用いた．楽譜特徴量，音響特徴量は共に平均 0，分散 1 になるよう事前に正規化を行った．

Table 2　Objective evaluation results.

	MGC (dB)	GVD
Conventional	5.378	1.955×10^{-1}
Proposed	5.258	1.567×10^{-1}

音響モデルは 3.2 節で述べたものを用いた．GRU ネットワークの構造は 1024 ユニットを 3 層重ねたものとした．音響モデルの学習は，GRU ネットワークのみを学習し，その学習済みの重みを用いてトラジェクトリ学習を行い，最後にその重みを用いて GV を考慮した学習を行う 3 段階に分けて行った．GV 尤度の重みは $w = 1.0 \times 10^{-6}$ とした．歌声波形はモデルによって推測される音響特徴量を WORLD による合成を行うことで得られる．

本研究では客観評価としてメルケプストラム歪み (MCD)，系列内変動距離 (GVD) を用いた．いずれも数値が低くなるほど精度は高くなる．

4.2　実験結果と考察

Table 2 に客観評価実験の結果を示す．MGC，GVD どちらの指標においても母音クラスタリングの結果を用いて学習したモデルの方が高い合成精度を示した．クラスタリングは，母音の発音の曖昧化や発音音程の違いによって区切られており，クラスタ番号情報が楽譜特徴量からのケプストラムの変化の学習を容易にさせたと考えられる．

Fig. 6 に目標歌声音声の u の曖昧化が見られる箇所に対応する，それぞれのモデルで合成された音声のスペクトルを示す．クラスタ番号情報を含むデータで学習したモデルは合成音声も曖昧化しているが，従来のデータを用いたモデルでは異なるスペクトルを示す．Fig. 1 と比較すると，このスペクトルは u のものであることが分かる．これらより，クラスタリングの結果を用いることで母音の曖昧化を再現することが可能であることが確認できた．

5　おわりに

本稿では母音の変化を考慮した歌声合成手法を検討した．モデル学習時に母音の周波数帯域や曖昧化による変化を捉えたクラスタリング結果を楽譜特徴量に加えることで，より高品質な歌声の合成を行った．

今回は楽譜データと対応する音声データのスペクトルから母音のクラスタリングを行ったが，実際の歌声合成では楽譜情報のみ与えられるためスペクトルによる母音のクラスタリングが不可能である．今後は楽譜情報から母音のクラスタを推測するモデルの構築を検討する．

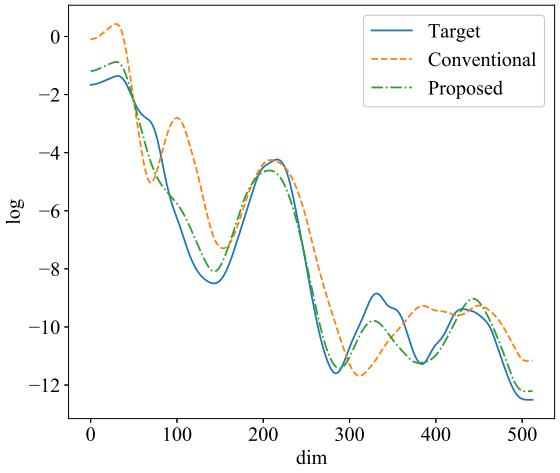

Fig. 6　Comparison of synthesized vowel spectrums.

謝辞　本研究の一部は，JSPS 科研費 JP17H01995 の支援を受けたものである．

参考文献

[1] J. Bonada, M. Umbert, and M. Blaauw, “Expressive singing synthesis based on unit selection for the singing synthesis challenge 2016,” in *Proc. Interspeech*, 2016, pp. 1230–1234.

[2] K. Saino *et al.*, “An HMM-based singing voice synthesis system,” in *Ninth International Conference on Spoken Language Processing*, 2006.

[3] 法野行哉 *et al.*，“Deep neural network に基づく歌声合成システム - sinsy”，**日本音響学会 秋季研究発表会 講演論文集**，2018.

[4] 片平健太 *et al.*，“深層学習を用いた歌声合成の検討”，**日本音響学会春季研究発表会 講演論文集**，2019，pp. 1091–1092.

[5] Johan Sundberg *et al.*，**歌声の科学**．　東京電機大学出版局，2007，pp. 124–130.

[6] M. Morise, F. Yokomori, and K. Ozawa, “World: a vocoder-based high-quality speech synthesis system for real-time applications,” *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[7] 片平健太 *et al.*，“Bidirectional gated recurrent units を用いた歌声合成に関する検討”，**音学シンポジウム**，2019.