母音の発音と歌唱速度の変化を考慮したアカペラオペラ歌声合成*

☆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

歌声合成は任意の歌詞付き楽譜から歌声音声を合成する手法である．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統計的パラメトリック歌声合成 [1] が主に研究されている．近年は深層ニューラルネットワーク (Deep Neural Networks; DNNs) を用いたボコーダや音声パラメータ推定モデル [2] の登場などにより，高品質な歌声音声の生成が容易となり，一般利用者への普及が進んでいる．

また，近年は人間らしい表現をもつ歌声の合成に関する研究が行われている．従来の歌声合成では童謡や J-POP といったジャンルの歌声音声を対象として行っていたが，より自由な表現を含む楽曲を対象とすることで，表情豊かな歌声合成システムの構築を目指している．

本研究ではプロのオペラ歌手によるアカペラ歌唱データを用いてその歌唱表現を分析し，それを考慮した複数のモデルからなる統計的パラメトリック歌声合成システムを提案する．

オペラ歌唱では遠方の聴者へ声を届かせる必要があるため通常の発話より顎を大きく開いて発声を行う．ここで母音の発音は，発声時の顎の開口度と強い関係を有しているため，オペラ歌唱では母音の発音が通常のものから変化する傾向がみられる．歌唱中の母音発音時のケプストラム情報に注目したクラスタリングを行い，その結果を音声パラメータ推定時に考慮させる．

また，アカペラ歌唱の特徴として自由な歌唱速度の変化が挙げられる．伴奏のないアカペラ歌唱では，その歌唱速度は歌手に大きく依存する．よって楽譜で指定されるテンポから大きく逸脱する変化が頻繁にみられる．この歌唱速度の変化を推定するモデルを歌声合成システムに導入する．

2 歌声合成システム

DNN を用いた歌声合成システムの例を Fig. 1 に示す．楽譜から歌詞の音素情報，音符の音高・音価情報，位置情報などを含む楽譜特徴量を抽出し，歌声合成システムの構成モデルの入力とする．

歌声合成システムは大きく分けて音素継続長推定，音響特徴量推定，ボコーダの 3 部分によって構成される．音素継続長推定では歌唱の発声タイミングのずれを再現するために，音符頭の音素を基準とした発声タイミング推定を行い，得られる各タイミング間に含まれる各音素の継続長を混合密度ネットワーク (Mixture Density Networks; MDNs) を用いて推定を行う [1]．音響特徴量推定では，初めに音素継続長情報を用いて楽譜特徴量を音素単位からフレーム単位に引き延ばしを行う．その後加工した楽譜特徴量を音響モデルに入力し，ケプストラムや基本周波数 (F_0)，帯域非周期成分などの音響特徴量を推定する．本研究では音響モデルにトラジェクトリ学習や系列内変動を考慮した GRU ネットワークを用いた [3]．最後に音響特徴量推定によって得られる音響特徴量をボコーダに入力し，最終的な歌声音声波形を生成する．

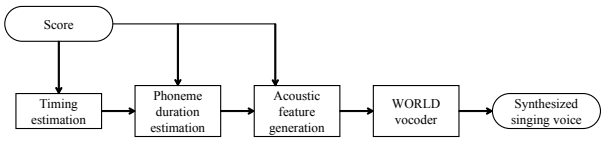


Fig. 1 Example of singing voice synthesis system.

3 母音の発音変化の推定

母音は舌の位置，唇の形，顎の開口度によって発音が決定される．顎の開口度に注目すると，日本語の 5 母音は a，o は開口度が大きく，i，u，e は開口度が小さい．

オペラ歌唱では発声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 [4]，遠方の聴者へ歌声を聴こえやすくさせる．この際，i，u，e の本来顎の開口度が小さい母音が a，o に近い発音へと変化する現象がみられる．

発音変化の例として，Fig. 2 に C5 の音高で発音した u，o，発音変化した u のスペクトルを示す．u，o のスペクトルを比較すると，1600Hz 付近の周波数成分の強度の差異が顕著にみられる．発音変化した u では o に近い概形を示していることから，通常の u の発音から変化していることが確認できる．

これらのオペラ歌唱における母音のスペクトルの変化を歌声合成に考慮するため，その分布を分析し

*Singing voice synthesis for a cappella opera considering variations of vowels and singing speed. by Kenta Katahira (Kobe Univ.), Yuji Adachi (MEC Company Ltd.), Kiyoto Tai (MEC Company Ltd.), Ryoichi Takashima (Kobe Univ.), Tetsuya Takiguchi (Kobe Univ.)

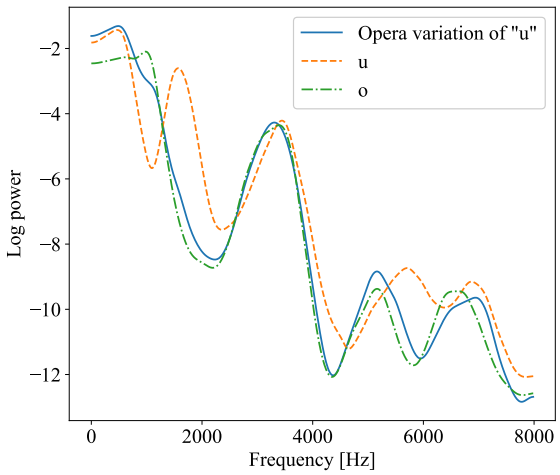


Fig. 2　Comparison of vowel spectrums.

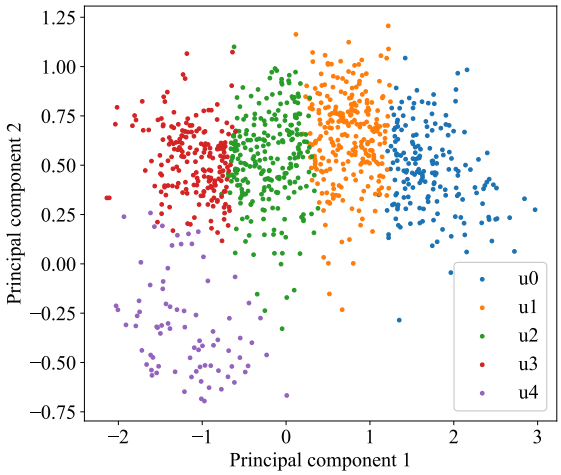


Fig. 3　Clustering results of “u”.

た．プロの女性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に対して WORLD ボコーダ [5] により 1024 次元のスペクトル包絡を得る．その後 60 次元のメルケプストラムに変換し，時間平均を取ったものを 1 発音データとする．発音データを主成分分析したのち，各母音ごとに k 平均法 ($k = 5$) 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，FFT のフレームサイズは 256 サンプルとした．また主成分分析では第 10 主成分まで使用した．

クラスタリングの結果として，u の各クラスタの分布を Fig. 3 に示す．u0 から u3 の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た．ここで各クラスタの母音発音データの平均 F_0 を Table 1 に示す．各クラスタの平均 F_0 は u0，u1，u2，u3，u4 の順に高くなることがわかった．各クラスタの分布と平均 F_0 を比較すると，第 1 主成分が小さくなるほど母音の発音平均 F_0 は高くなることが分かった．この傾向は u に限らず全ての母音において観測された．

また Fig. 3 において，第 1 主成分が 0 以下であり，第 2 主成分が 0.25 以下のものである u4 のクラスタはスペクトルの概形から u の発音が変化したものであることが分かった．他の母音の発音データの分布と比較すると，u4 が分布する領域は o の発音データが分布する領域と重なった．第 1 主成分に注目すると，a，o は負の領域に，それ以外は正の領域に多くが位置しており，これらから第 1 主成分と顎の開口度に相関があると考えられる．また第 2 主成分との関係から，発声時の音高が高いほど母音発音が変化しやすくなると考えられる．

これらのオペラ歌唱における母音の発音の傾向を歌声合成モデルに考慮させるため，音響特徴量推定モデルの入力として各母音発音のクラスタ番号を識別子とした one-hot 形式の特徴量を追加し，音響モデルの学習・生成を行う．また，歌声合成時には音声データのない楽譜から母音クラスタ番号を得る必要があるため，楽譜特徴量を入力として母音のクラスタ番号を推定する決定木モデルを新たに歌声合成システムに追加する．

4　歌唱速度の推定

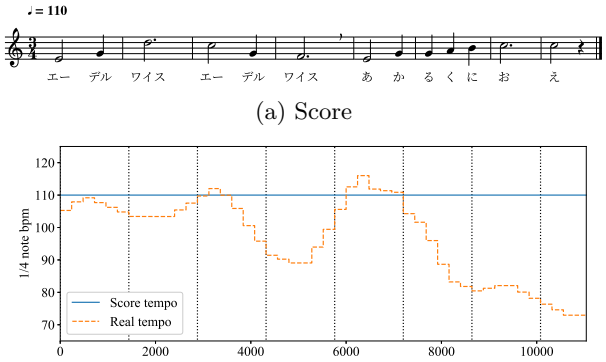
アカペラ歌唱では歌唱速度はその歌手に依存するため，実際の歌唱速度が楽譜で指示されているテンポに厳密でなく，場合によっては大きく逸脱することがみられる．

アカペラ歌唱における歌唱速度の変化には，曲終端部における歌唱速度の鈍化があてはまる．例として Fig. 4 に該当部の楽譜とテンポの関係を示す．以下本稿では 1 分間における 4 分音符の拍数をテンポと呼称する．ここで横軸は全音符の継続長を 1920 とした際の楽譜先頭からの距離を，縦軸はテンポを表す．青の実線が楽譜で指示されるテンポ，オレンジの点線が実際の歌唱のテンポである．この図より曲が終盤に向かうほど実際の歌唱テンポが指定されるテンポより遅くなることが分かる．

アカペラ歌唱を対象とした歌声合成を行うにあたって，アカペラ歌唱における音素継続長を推定することが必要となる．ここで第 2 章で示した発声タイミング推定は，歌唱における「ノリ」や「タメ」といった時間変化を再現するものであり，これらの変化は音符単位の局所的な変化である．対してアカペラ歌唱では複数の音符にわたる中・長期的なテンポ変化が起こるため，発声タイミングのずれのみではこの問題

Table 1　Mean F_0s of clusters of “u”.

Cluster	u0	u1	u2	u3	u4
Mean F_0[Hz]	304.78	355.74	437.19	513.87	608.92



(a) Score

(b) Comparison between score tempo and actual tempo

Fig. 4　Example of tempo changes in a cappella singing.

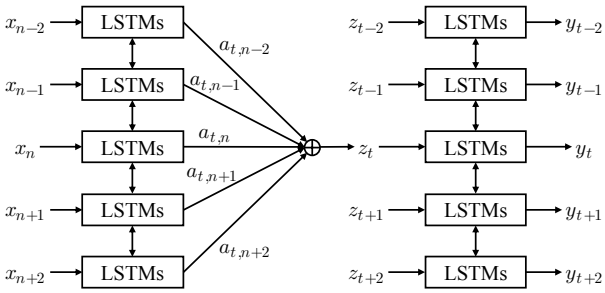


Fig. 5　Model for segment tempo estimation.

を扱うことが難しい．

アカペラ歌唱におけるテンポ変化の推定を行うため，本稿ではセグメントテンポ推定を提案する．楽譜を 8 分音符単位で区切った区間をセグメントと定義し，各セグメントの歌唱経過時間から求められるテンポを推定する [6]．ここで歌唱速度の変化はセグメントテンポの系列によって表現される．推定されたセグメントテンポから音符の継続長を算出し，各音符に含まれる音素の継続長を MDN を用いて推定することで，アカペラ歌唱における音素継続長を推定する．

学習に用いるセグメントテンポ系列は，楽譜上の音符の位置と実際の発声タイミングを用いて算出する．その際より理想的なテンポカーブを得るため，それらの移動平均値を利用する．なお，移動平均値はフレームサイズを 32，フレームシフトを 8 と設定した．

本手法は DNN を用いた統計的手法であるため，学習に含まれない楽譜テンポが指定された際に推定に失敗する恐れがある．そのため，楽譜で指示されるテンポとセグメントテンポの差分をモデル化させることで対処する．

Fig. 5 にセグメントテンポ推定のモデルを示す．セグメントテンポ推定では楽譜特徴量を入力し，セグメントテンポ系列を出力とする．楽譜中のセグメントの個数とそれぞれの位置は事前に計算可能であることを利用し，音素単位である最初の LSTM 層の出力に対して各セグメント毎にセグメントの位置を考慮した重みづけ和を取り，セグメント単位の特徴へと変換する．ここで t 番目のセグメントに対する n 番目の楽譜特徴量への重み $a_{t,n}$ は以下の通りに示される．

$$a_{t,n} = \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left\{-\frac{(n-\mu_t)^2}{2\sigma_t^2}\right\} \tag{1}$$

$$\mu_t = (l+m)/2,\ \sigma_t = \alpha(m-l)/2 \tag{2}$$

なお l と m は楽譜特徴量の添字，α は任意の定数である．l 番目と m 番目の楽譜特徴量はそれぞれ t 番目のセグメントに含まれる最初と最後の音素に対応する特徴量である．

またモデルはセグメントテンポ系列の静的・動的特徴量を推定し，生成時にそれらの特徴量の関係を考慮した合成 [7] を行うことで最終的な出力とした．

5　実験評価

5.1　実験条件

本稿ではアカペラオペラ歌唱音声におけるセグメントテンポ推定と母音発音変化推定の有無による合成音声の品質の比較実験を行う．実験には女性歌手 1 名による日本語アカペラオペラ歌唱音声 48 曲からなる約 93 分の音声データセットを用いた．このうち 43 曲を学習に，5 曲を検証に用いた．また，テストデータとして未知の楽譜データ 5 曲を用いた．

本稿では 4 つの歌声合成システムをそれぞれ比較する．提案手法である **D** の構成を Fig. 6 に示す．セグメントテンポ推定では Fig. 5 で示したネットワーク，音素継続長推定では MDN，母音発音推定では深さ 6 の決定木，音響モデルには GRU で構成されるネットワークを用いた．各モデルには 774 次元の楽譜特徴量を入力する．なおセグメントテンポ推定モデルは最初の LSTM 層は 512 ユニット 3 層，最後の LSTM 層が 128 ユニット 1 層でスカラ値の母音クラスタ番号を出力する．音素継続長推定モデルは 128 ユニッ

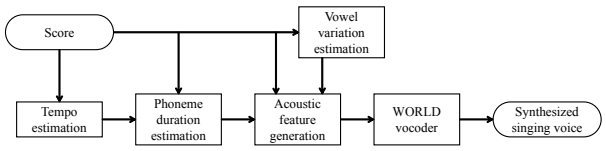
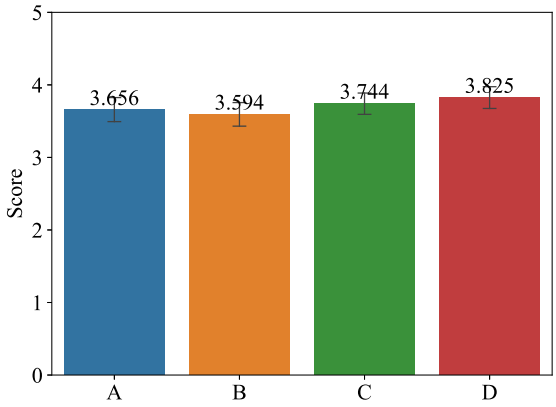


Fig. 6　Proposed system **D**.



Tempo	-	○	-	○
Vowels	-	-	○	○

Fig. 7　MOS evaluation result.

ト 3 層，音響モデルでは 1024 ユニット 3 層の構成である．また，音響モデルで生成される音響特徴量はメルケプストラム 60 次元，対数基本周波数 1 次元，帯域非周期成分 1 次元と有声・無声パラメータ 1 次元である．本稿では式 (2) に関して $\alpha = 0.75$ と設定した．比較システムとして **D** からセグメントテンポ推定モデルを除いたものを **C**，母音発音推定モデルを除いたものを **B**，いずれも除いたものを **A** とする．

5.2　主観評価実験

本稿では主観評価として平均オピニオン指標 (MOS) を用いた．1 が機械的な音声，5 を人間らしいアカペラオペラ歌唱音声として，合成音声がどちらに近いか 5 段階評価を行った．被験者は 8 人でテストデータからランダムに抽出された 20 フレーズに対して評価を行った．Fig. 7 に MOS 評価の結果を示す．図のエラーバーは 95%信頼区間を表す．

提案手法 **D** と **A** との間には，有意な差がみられた．これらからセグメントテンポ推定と母音発音推定がアカペラオペラ歌唱の特徴を考慮した歌声合成において人間らしい表現の補助的な役割を果たしたと考えられる．**A** と比較しての **C** の結果と対照的に，**B** は **A** よりも評価が低かった．これは母音の発音変化推定は通常の発音からオペラにおける発音に一部変更するような付加的な操作であることに対して，セグメントテンポ推定では歌唱速度を一定のものから大きく変化させるものであり，推定に失敗した際の影響がより大きいことが原因と考えられる．しかし，両推定を含んだ **D** が最も高い評価を示すことから，アカペラオペラ歌唱において特異な歌唱速度の変化と母音の発音の変化には関連があると考えられる．

式 (1) で示した重みづけによるセグメントテンポ推定は，セグメント区間内の特定の音符に対する特徴を選択できないため，フェルマータなどの特徴を見逃しやすい．よって，Seq2Seq などに用いられる注意機構などを導入することで，セグメントテンポ推定の精度が向上する可能性がある．

6　おわりに

本稿ではアカペラオペラ歌唱を対象とした歌声合成システムを提案し，従来のモデルとの比較を行った．母音の発音と歌唱速度の変化を考慮したモデルを歌声合成システムに加えることで，アカペラオペラ歌唱における特徴を加味した歌声合成を行うことが示された．

参考文献

[1] Y. Hono *et al.*, “Recent development of the DNN-based singing voice synthesis system—sinsy,” in *APSIPA ASC*, 2018, pp. 1003–1009.

[2] K. Nakamura *et al.*, “Fast and high-quality singing voice synthesis system based on convolutional neural networks,” in *Proc. IEEE ICASSP*, 2020, pp. 7239–7243.

[3] K. Katahira *et al.*, “Opera singing voice synthesis considering vowel variations,” in *Proc. IEEE GCCE*, 2020, pp. 663–664.

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

[5] 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.

[6] A. Maezawa, “Deep linear autoregressive model for interpretable prediction of expressive tempo,” in *Proc. SMC*, 2019, pp. 364–371.

[7] K. Tokuda *et al.*, “Speech parameter generation algorithms for HMM-based speech synthesis,” in *Proc. IEEE ICASSP*, 2000, vol. 3, pp. 1315–1318.