自由な歌唱速度の歌声の合成に関する検討*

☆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，任意に与えられた歌詞付きの楽譜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．主に娯楽分野において広く普及しつつあり，近年では故人の歌声の再現や病気等で声を失った患者の歌声の再現などの利用も考えられている．

現在主に研究されている歌声合成の対象歌唱は童謡や J-POP などの音声であり，これらの歌声の合成における音素継続長推定では歌唱におけるノリやタメなどの特徴を再現するような手法が提案されている．

一方で伴奏がないアカペラ歌唱では歌唱速度が頻繁に変化し，楽譜上のテンポから逸脱することは珍しくない．表現力豊かな歌声を合成するために，自由な歌唱速度に対応した音素発音時間の推定が必要である．本稿ではテンポの変化を音符継続長の変化と捉え，アカペラ歌唱音声の合成のための音符継続長推定手法を検討する．

2　アカペラ歌唱音声

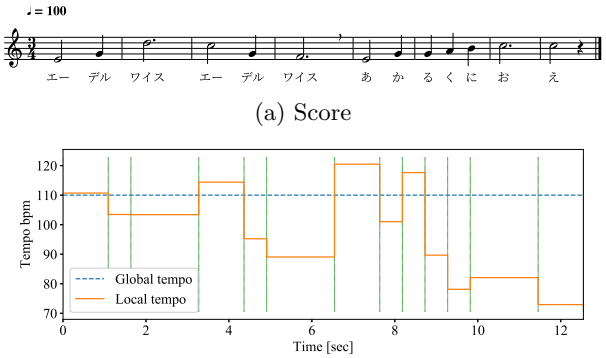
一般に歌声合成で用いられる音声は童謡や J-POP などの伴奏が存在する歌唱である．この伴奏は楽譜で指定されている一定なテンポで演奏されるため，その歌声もテンポに忠実な歌唱であることが多い．

一方アカペラ歌唱では伴奏が存在しないため歌唱速度はその歌手に依存し，楽譜上のテンポを基準としながら局所的な歌唱速度の変化が頻繁に見られる．

例として曲終端部における歌唱速度の鈍化が挙げられる．Fig. 1 に楽譜とテンポの関係を示す．ここで横軸は楽譜上のテンポ指定と音価から求められる経過時間，縦軸はテンポを表す．青の実線が楽譜上のテンポ指定であり，オレンジの点線が音価と実際の音符発声時間から計算される局所テンポである．この図から曲の終盤になるほどテンポが鈍化し，音符の実継続長が楽譜上の長さよりも長くなることが分かる．

3　音素継続長推定

深層学習を用いたパラメトリック歌声合成では楽譜から抽出される音素，音程，位置情報などの楽譜特徴量から歌声音声のスペクトル，基本周波数，非周



(a) Score

(b) Comparison between global and local tempo

Fig. 1　Tempo changes.

期成分を推定する [1]．ここで楽譜特徴量は音素単位で抽出されるのに対して，音響特徴量はボコーダによる解析から得られるフレーム単位の特徴量であり，単位が不一致である．音響特徴量の推定では音素の継続長情報をもとに楽譜特徴量を引き延ばしてフレーム単位に変換し，推定モデルの入力と出力を対応させる．そのため予め与えられた楽譜から音素の継続長を推定することが必要となる．

従来手法 [2] では音素継続長推定を 2 段階に分割して行っている．始めに音符に対応する歌詞の最初の母音の実発声タイミングと楽譜から計算される音符の発声タイミングのずれを予測する音符発声タイミング推定を行う．次に推定された音符発声タイミングから計算される音符継続長を用いて音符内の各音素の継続長を混合ガウス分布を用いた予測により計算する．これらの手法によって，歌唱における「ノリ」や「タメ」の表現を考慮した音素継続長の推定を行う．

4　音符継続長推定

アカペラ歌唱音声合成を行うには，その歌唱を対象とした音素継続長を推定することが必要となる．ここで 3 章で示した音符発声タイミング推定において，従来の推定で扱う歌唱はテンポに比較的忠実であり，音符の発声タイミングのずれは音符継続長に対して微小なものであった．しかし本研究で用いるアカペラ歌唱では局所的にテンポが変化するため音符の継続長が著しく変化し，音符の発声タイミングのずれは過

*Note length prediction for singing voice synthesis with free 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.)

去のずれが積算されるため変化の大きいものとなり，推定が困難になることが考えられる．

ここで単なる音符発声のタイミングでなく過去のずれを除いた音符発声タイミングのずれを考える．これは楽譜から計算される音符継続長と実際の音符継続長の差分とみることができる．また差分の推定では推定された差分が楽譜から算出される音符継続長を打ち消す恐れがあるため，実際の音符継続長の楽譜での音符継続長に対する伸縮率を考え，これを推定する音符継続長推定を行う．

5　実験評価

5.1　実験条件

本稿ではアカペラ歌唱音声を対象とした音符発声タイミング推定と音符継続長推定の比較実験を深層学習を用いて行う．

実験には女性歌手 1 名による日本語アカペラ歌唱音声 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 曲をテストに用いた．

音符発声タイミング推定には中間層 512 ユニットの LSTM3 層で構成されるネットワーク (**LSTM**) を用いた．音符継続長推定では **LSTM** と過去の推定結果を次の推定の入力に結合する中間層 512 ユニットの LSTM3 層で構成される自己回帰ネットワーク (**AR**) の 2 つのモデルを用いた．各モデルの出力はそれぞれの手法で推定するデータのスカラ値であり，**LSTM** の入力は楽譜から得られる楽譜特徴量 609 次元の各音符の先頭母音に対応するもの，**AR** はそれに過去 10 回の推定結果を結合した 619 次元のデータである．

本実験の客観評価指標として，音符の発声タイミングのずれの RMSE(T-RMSE) と音符継続長の差分の RMSE(L-RMSE) を用いた．

5.2　実験結果

Table 1 に客観評価の結果を示す．いずれの客観評価においても音符継続長推定が音符タイミング推定より高い推定精度を示した．ずれの表現に音符継続長伸縮率を用いることで音符発声タイミングを用いる場合より変化の幅が小さくなることが影響したと考えられる．また微小であるが **AR** が **LSTM** よりも推定誤差が小さくなることが分かった．この結果からテンポ変化は以前の変化の影響を受けており，その情報を考慮して推定することで精度が向上したと考えられる．

また音符継続長推定による曲終盤の局所テンポの変化を Fig. 2 に示す．5 秒から 8 秒付近では曲終盤のテンポの鈍化が見られるが，実験より得られたものはこの傾向に従った継続長を推定している．一方で目標データのテンポとの差が 10 以上離れている箇所が存在する．音符継続長推定ではアカペラ歌唱におけるテンポの揺らぎを音符単位の局所的なテンポ変化と定義し，その結果である音符継続長の変化を推定する．しかし楽譜の音符の音価は一定でなく，音価に大きく依存する音符単位の局所テンポでは実時間での長さのずれの局所テンポへの影響の大きさが音符によって異なるためこれらの推定ミスにつながると考えられる．

Table 1　Objective evaluation results.

	Timing	Note length	
	LSTM	LSTM	AR
T-RMSE (ms)	1941.3	1864.2	1710.5
L-RMSE (ms)	232.93	196.08	188.71

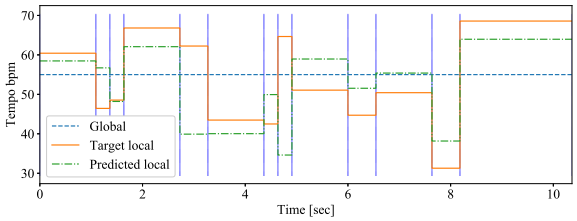


Fig. 2　Tempo changes by note length prediction.

6　おわりに

本稿ではアカペラ歌唱音声の歌声合成における音符発声タイミング推定と音符継続長推定の比較を行った．音符伸縮率を用いることでより誤差の小さい音符継続長の推定を行うことができるが，自然なテンポの変化を推定するには改善の余地が見られる．

今後は局所テンポを楽譜上の一定単位の経過時間から計算されるものと定義したうえでテンポの揺らぎの推定を行う手法を検討する．

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] 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.

[2] 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，pp. 1099–1102.