アカペラ歌唱における歌唱速度の変化を考慮した歌声合成に関する検討*

☆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

歌声合成の研究における発声タイミングの推定に関して，楽譜に記されているテンポを基準として，歌唱におけるノリやタメなどの発声タイミングのずれを再現する手法が提案されている．

アカペラ歌唱音声を対象とした場合，歌唱速度が頻繁に変化し楽譜上のテンポから大きく逸脱する特徴を捉えるため，音符単位のテンポを推定し発声タイミングを求める手法 [1] を提案したが，不自然なテンポ変化が現れることがあった．

本稿ではテンポの変化を楽譜上の各セグメントの経過時間の変化と捉え，より自然なテンポ系列を推定する手法を検討する．

2　アカペラ歌唱音声

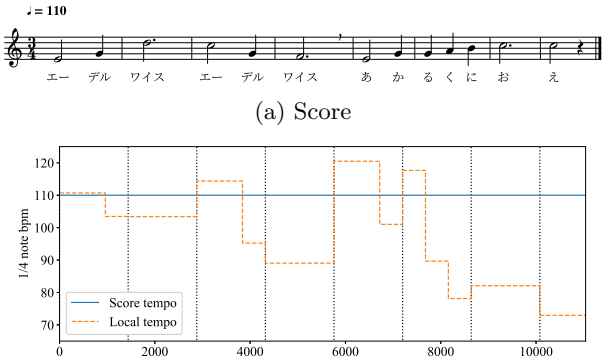
主に歌声合成で用いられる歌唱は楽譜に指定されたテンポに従った歌唱であることが多い．ノリやタメなどといった歌唱タイミングの変化が見られることがあるが，これらは音符単位の局所的な変化とみなすことができる．

一方アカペラ歌唱では，歌唱速度はその歌手に依存し，楽譜上のテンポをから大きく逸脱することが頻繁に見られる．これは小節を超えるような長期的の変化となることがある．

アカペラ歌唱における歌唱速度の変化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．なお，本稿では 1 分間における 4 分音符の拍数をテンポと表現する．ここで横軸は全音符の長さを 1920 とした時の楽譜の先頭からの距離であり，縦軸はテンポを表す．青の実線が楽譜で指定されたテンポ，オレンジの点線が音価と実際の音符継続長から計算される音符ごとのテンポである．この図から曲の終盤になるほど歌唱テンポが楽譜の指定よりも遅くなることが分かる．

3　音符継続長推定

深層学習を用いたパラメトリック歌声合成では楽譜より音素単位で抽出される楽譜特徴量からボコーダによって得られる歌声音声のスペクトル等のフレーム単位の音響特徴量を推定する．ここで推定モデルの入力と出力に単位の不一致があるため，楽譜特徴量を音素の継続長情報をもとに引き延ばしてフレーム単位に変換する．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必要がある．音素の継続長はその音素が含まれる音符の継続長を推定したのちに，その値を用いて推測される．



(a) Score

(b) Comparison between global and segment tempo

Fig. 1　Tempo changes.

以前の研究 [1] でアカペラ歌唱音声合成における音符継続長推定を提案した．これは楽譜上の音符継続長に対する実際の音符継続長の伸縮率を音符ごとに推定する．アカペラ歌唱においては従来の音符の発声タイミングのずれを推定する手法と比較してより精度の高い推定が行えること示した．しかし音符単位の継続長の伸縮率を推定するため，絶対時間でのずれが音価に依存するため不自然なテンポ変化が現れることがあった．

4　セグメントテンポ推定

本稿では楽譜を等間隔で区切った区間をセグメントと定義し，各セグメントの経過時間から求められるテンポを推定する [2]．セグメントは基準の長さが全て等しいため，絶対時間でのずれがセグメントのテンポに直接反映される．なお実際には楽譜から計算されるセグメントテンポと実際の歌唱のセグメントテンポの差分を推定する．

Fig. 2 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音素単位の楽譜特徴量を入

*Segment tempo prediction for a cappella singing voice synthesis. by Kenta Katahira (Kobe Univ.), Yuji Adachi (MEC Company Ltd.), Kiyoto Tai (MEC Company Ltd.), Ryoichi Takashima (Kobe Univ.), Tetsuya Takiguchi (Kobe Univ.)

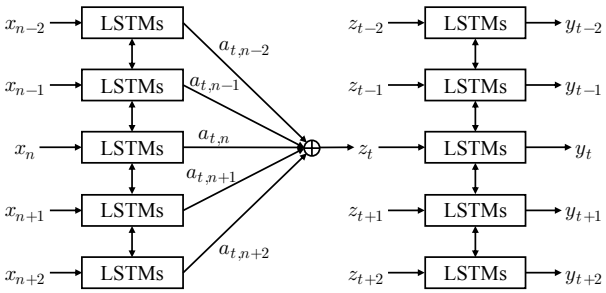


Fig. 2　Model for segment tempo prediction.

力，セグメント単位のテンポの差分系列を出力とするため，入出力間に単位の不一致が存在する．しかしセグメントの個数，位置は楽譜から算出可能なため，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ることで単位の変換を行う．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，

$$a_{t,n} = \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left(-\frac{(n-\mu_t)^2}{2\sigma_t^2}\right)$$
$$\mu_t = (l+m)/2,\ \sigma_t = \alpha(m-l)/2$$

となる．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特徴量である．

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セグメントテンポ推定と音符継続長推定の比較実験を行う．

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．

セグメントテンポ推定では 4 章で示したモデルを用いた．なお中間層は最初の LSTM 層が 512 ユニット 3 層，最後の LSTM 層が 128 ユニット 1 層である．音符継続長推定には中間層 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，モデルの入力は，セグメントテンポ推定では楽譜から得られる音素単位の楽譜特徴量 609 次元，音符継続長推定では同様の入力のうち各音符の先頭母音に対応するものである．なおセグメントは 8 分音符の長さであり，$\alpha = 0.75$ とする．

本実験の客観評価指標として，セグメント継続長の RMSE(SL-RMSE)，音符継続長の RMSE(NL-RMSE) と音符の発声タイミングのずれの RMSE(NT-RMSE) を用いた．

Table 1　Objective evaluation results.

Model	Segment tempo	Note length
SL-RMSE (ms)	95.972	98.902
NL-RMSE (ms)	187.53	165.86
NT-RMSE (ms)	756.53	832.03

5.2　実験結果

Table 1 に客観評価の結果を示す．セグメント継続長の RMSE について 2 つの手法に大きな差は生じなかった．音符継続長の RMSE では音符継続長推定の精度が高く，音符発声タイミングの RMSE はセグメントテンポ推定の精度が高かった．

音符継続長推定では音符の長さに注目した学習を行っている．またモデルが扱うデータは音符の伸縮率であり，基準の音符が各データ間で異なるため絶対時間での継続長の変化が考慮されにくい．よって不自然なテンポ変化が起こり，発音タイミングの誤差が大きくなると考えられる．対してセグメントテンポ推定では楽譜を等間隔に区切ったセグメント単位で推定するため，絶対時間での継続長の変化を捉え易い．よって音符の継続長の推定精度は甘くなるが，滑らかなテンポ変化系列を推定することが可能で音符発音タイミングのずれは小さくなると考えられる．

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セグメントテンポ推定と音符継続長推定の比較を行った．セグメント単位でのテンポ推定により滑らかなテンポの変化系列を推定することができた．

今後はセグメントテンポ推定における推定エラーを防ぐため，テンポ変化系列の包絡線の推定やテンポ変化の周期などを考慮したスペクトルの推定を検討する．

参考文献

[1] 片平健太 *et al.*，“自由な歌唱速度の歌声の合成に関する検討”，**日本音響学会春季研究発表会 講演論文集**，2020，pp. 1125–1126.

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