# 深層学習を用いた歌声合成の検討*

☆片平健太 (神戸大), 北村毅 (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), 滝口哲也 (神戸大)

## 1 はじめに

本研究では，深層学習を用いたオペラ歌唱音声を生成する歌声合成システムを提案する．歌声合成システムは，任意に与えられた楽譜の言語や音高などの表現から歌声を合成するシステムである．これまで HMM を用いた手法 [1] 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 DNN を用いた手法 [2] が提案されている．さらに，テキスト音声合成分野において高い性能を示している WaveNet に基づいたボコーダ [3] など従来の品質を上回る手法が提案されており，歌声合成においても音響特徴量の推定 [4] に用いられている．

本稿では，オペラ歌唱を生成する歌声合成システムを実現するため，オペラ歌唱のスペクトル特徴について調査し，楽譜からオペラ歌唱に対応する楽譜特徴量の抽出を行った．客観評価実験により，従来手法と比較して提案手法が優れた品質のオペラ歌唱を生成できることを示す．

## 2 オペラ歌唱の合成

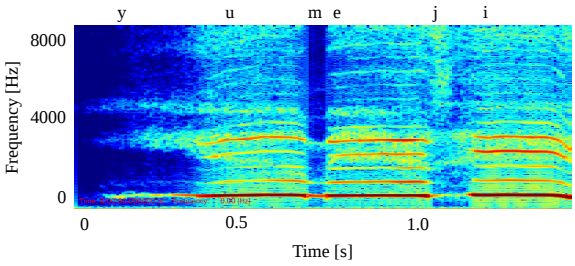
本研究ではオペラ歌唱音声を DNN 歌声合成の手法を用いて生成する．

### 2.1 オペラ歌唱音声

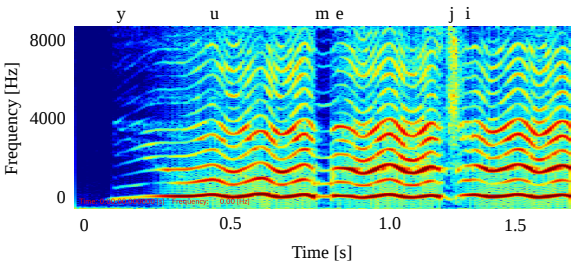
オペラ歌唱音声のスペクトログラムを Fig.1 に示す．Fig.1 から，オペラ歌唱では 3000 Hz から 4000 Hz の中高音域の周波数成分が多く含まれる．この帯域は第 3 フォルマントから第 4 フォルマントなど響きや聞こえ度に関する成分が多く含まれており，オーケストラの伴奏の中でも聴衆に演奏との聞き分けを容易にさせる効果がある．また，オペラ歌唱は母音の発声時にビブラートが顕著に現れることが確認できる．

### 2.2 楽譜特徴量

楽譜特徴量は，MusicXML 形式の楽譜データを解析することで抽出した．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む．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ている．音符情報にはタイ・スラーの有無，強弱などが含




(a) Amateur




(b) Opera

Fig. 1: Singing spectrograms.

まれている．これらの特徴量データをバイナリや連続値，one-hot vector で表現し，DNN の入力データとして用いる．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．

### 2.3 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．音響モデルには Bidirectional LSTM を用いる．合成時は音響モデルに楽譜特徴量を入力して得られる音響特徴量からボコーダを用いて音声を合成する．

## 3 実験評価

### 3.1 実験条件

実験には，オペラ歌手 1 名について歌唱音声 29 曲からなる約 45 分の音声を用いた．このうち 26 曲を音響モデルの学習に，3 曲をテストに用いた．サンプリング周波数は 16 kHz ，フレーム長は 256 サンプルとした．

音響特徴量には，WORLD[5] を用いて抽出したスペクトル包絡から計算したメルケプストラムを 24 次

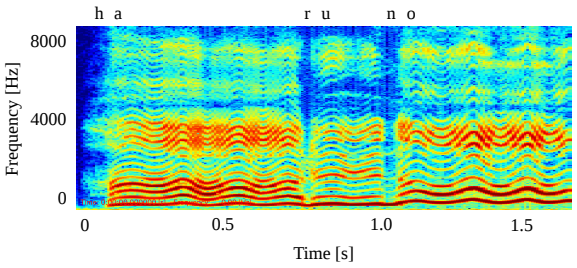
*Singing Voice Synthesis System Using Deep Learning. by Kenta Katahira (Kobe Univ.), Tsuyoshi Kitamura (Kobe Univ.), Yuji Adachi (MEC Company Ltd.), Kiyoto Tai (MEC Company Ltd.), Tetsuya Takiguchi (Kobe Univ.)

元，基本周波数を $F_0$1 次元，帯域非周期成分を 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．また，学習時は音響特徴量は平均 0 分散 1 となるよう正規化を行った．楽譜特徴量には，2.2 節で示した特徴を含めた 544 次元を用いた．

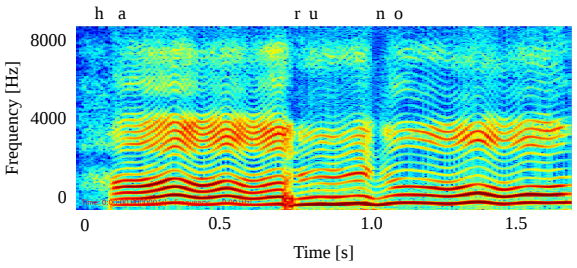
音響モデルの隠れ層には，ユニット数 256 からなる Bidirectional LSTM3 層とした．比較音声として HMM 歌声合成システムを用いて生成した音声を使用した．また客観評価として，メルケプストラム歪み (MCD), 基本周波数の 2 乗誤差 ($F_0$ RMSE)，帯域非周期成分歪み (BAPD)，有声・無声判断の偽陽性率 (V/UV FPR)，偽陰性率 (V/UV FNR) を用いた．

### 3.2 実験結果と考察

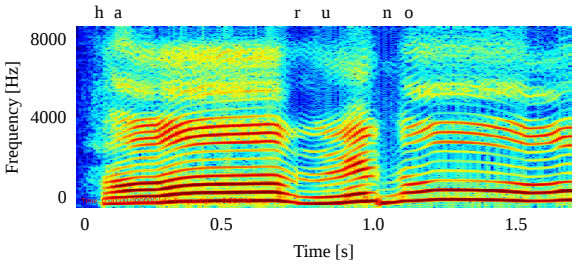
Fig. 2 に提案手法により生成された音声と目標音声のスペクトログラムを示す．Fig. 2 から，提案手法では低域だけではなく中高域のエネルギー成分の推定も可能である．また，ビブラートの推定も行われていることがわかる．




(a) Target




(b) DNN




(c) HMM

Fig. 2: Sample spectrograms.

Table 1 に客観評価実験の結果を示す．有声・無声判断の偽陽性率を除いて，提案手法による結果が HMM の結果と比較して低い値を示している．

Table 1: Objective evaluation results.

| | **HMM** | **DNN** |
|---|---|---|
| MCD (dB) | 6.141 | 5.552 |
| $F_0$ RMSE (cent) | 97.68 | 93.70 |
| V/UV FPR | $1.776 \times 10^{-1}$ | $1.037 \times 10^{-1}$ |
| V/UV FNR | $1.194 \times 10^{-2}$ | $1.995 \times 10^{-2}$ |
| BAPD (dB) | 28.22 | 19.29 |

## 4 おわりに

本稿では，Bidirectional LSTM を用いたオペラ歌唱の歌声合成システムを提案した．Table 1 より HMM によるシステムと比較して，提案手法は優れた品質の音声を生成できた．本研究では，1 つの音響モデルで全ての音響特徴量の推定を行ったが，各特徴量ごとに最適なモデルは異なると考えられるため，それぞれに最適なモデルを学習し音声を合成する手法を検討する．また，音響特徴量から波形への合成部に WaveNet ボコーダを用いることで合成音質の向上を検討する．

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

[2] M. Nishimura *et al.*, “Singing voice synthesis based on deep neural networks,” in *Proc. Interspeech*, 2016, pp. 2478–2482.

[3] A. Tamamori *et al.*, “Speaker-dependent wavenet vocoder,” in *Proc. Interspeech*, 2017, pp. 1118–1122.

[4] M. Blaauw and J. Bonada, “A neural parametric singing synthesizer,” *arXiv preprint arXiv:1704.03809*, 2017.

[5] M. Morise *et al.*, “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.