バックコーラス歌唱合成のための DNN を用いた自然性の高い歌声合成方式の検討

Study of natural singing-voice synthesis for backing vocals based on DNN

木岡 智宏

Tomohiro Kioka

岡山大学 阿部研究室

Abe Laboratory, Okayama University

概要 本報告では，DNN を用いて楽譜情報から自然性の高いバックコーラス歌唱を合成する方式を検討する．主旋律を邪魔することなく歌の厚みを持たせることを主な目的とするバックコーラス歌唱を合成する場合であっても，楽譜通りの基本的な音高の制御に加え，歌唱表現としての音高の揺らぎも精度良く制御できる方式が必要となる．そこで，F0 生成部，スペクトル生成部，波形合成部の 3 段階で構成される合成モデルを提案する．

1 はじめに

コーラスは，合唱を意味する音楽用語であり，複数の歌手が声部に分かれて同時に歌うことを指す．ただし，ポピュラー音楽分野においては，歌手の人数によらず，二つ以上の異なるメロディラインを合わせて歌うことを意味する広義な用語として扱われている．バックコーラスとは，コーラスのなかでも特に主旋律を歌う者に対して，その後ろで補助的に歌唱することを言い，メインヴォーカルを邪魔することなく歌の厚みを持たせることを主な目的としている．主旋律を引き立たせるバックコーラスを生成する場合であっても，声の大きさや楽譜通りの基本的な音高への制御はもちろんのこと，歌唱表現としての音高の揺らぎも精度よく制御できる方式が必要となる．

歌声合成は，任意の楽曲の歌声を人工的に作り出す技術である．歌声は，人間が発話する音声でもあり，楽器のような周期的振動を含む楽音でもあるため，歌声合成ではその両方を満足しなければならない．また，音声分析に用いられる特徴量の一つに基本周波数（Fundamental Frequency:F0）があるが，歌声の F0 にはオーバーシュートやビブラートなど，会話音声には見られない歌声特有の表現も含まれており，歌声合成ではこれらの複雑な特徴を捉える必要がある．素片連結型の VOCALOID[1] をはじめ，隠れマルコフモデル（HMM）に基づく歌声合成システム [2] など，楽譜から得られる情報を用いて歌声を合成する様々な手法が提案されている．近年ではこれらの技術とは異なる手法として，Deep Neural Network（DNN）を用いた歌声合成方式が研究されている [3][4][5]．WaveNet[6] は，音声合成の分野で用いられている DNN の一つであり，従来の音声合成方式に比べて，高品質で自然な音声を合成可能である．

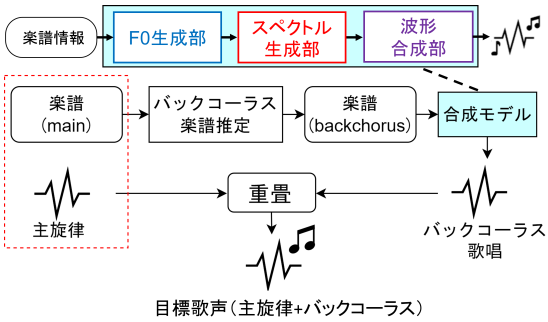

図 1: バックコーラス歌唱合成方式の概要

本研究の目標とする歌声は，既存の主旋律に生成バックコーラス歌唱を重畳したものである．図 1 は，バックコーラス歌唱合成の概要を示している．主旋律に対応する楽譜からバックコーラス歌唱に対応する楽譜を推定し，推定した楽譜に基づいてバックコーラス歌唱を生成する．合成モデルは，F0 生成部，スペクトル生成部，波形合成部の 3 段階で構成され，楽譜から得られる音高情報と音素情報から 1 メロディラインにあたる歌声を合成する．F0 生成部とスペクトル生成部は，時系列データの長期依存関係の学習・予測を可能とする Long short term memory（LSTM）[7] を用いる．一方，波形合成部は，肉声に近い高品質な音声を実現する WaveNet[6] を用いる．本報告では，図 1 で示したバックコーラス歌唱合成の実装前段階として，楽譜から高品質なバックコーラス歌唱を生成する合成モデルの実装を目指す．

2 提案方式

提案方式は，楽譜から得られる音高情報と音素情報を用いて，人間の歌声特有の自然な音高の揺らぎを含むバックコーラスの合成を目指す．提案方式の概要を図 2 に示す．合成モデルは，F0 生成部，スペクトル生成部，波形合成部の 3 段階で構成され，各々は学習ステージと合成ステージに分けられる．合成する歌声はバックコーラスであるため，少なくとも歌声の母音区間の再現の支障がなければバックコーラス歌唱として成立すると考えられる．そこで，提案方式では，母音の

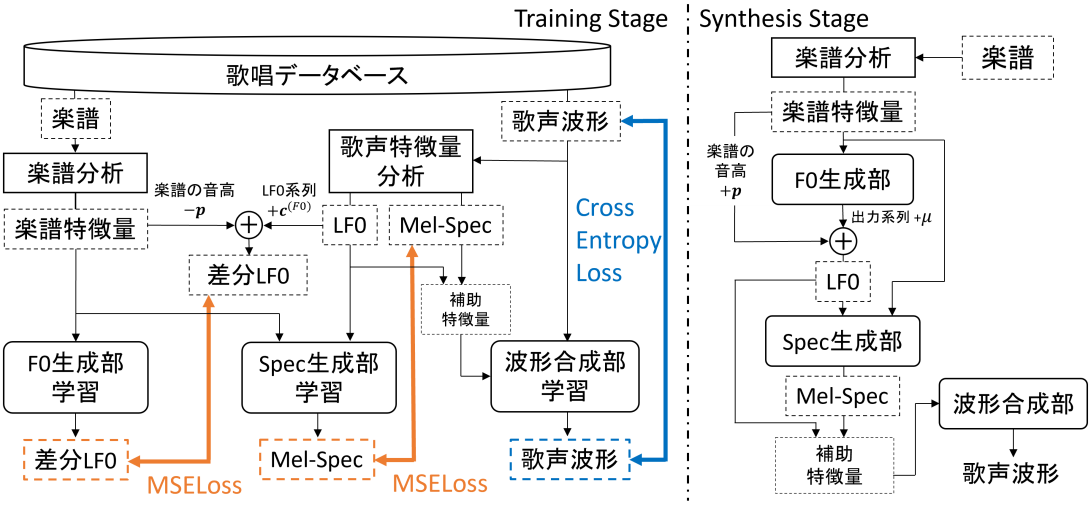

図 2: 合成モデルの概要

みで歌う母音唱のような歌唱を想定する．以下に，F0 生成部，スペクトル生成部，波形合成部の概要を示す．

1. **F0 生成部**：F0 生成部では，楽譜分析によって抽出された楽譜特徴量を基に，歌声の基本周波数（F0）を推定する．歌声の F0 は，ビブラートやポルタメントなど，通常の発話音声とは異なる歌声特有の特徴が表れるため，F0 のモデル化は人間らしい歌声を実現するうえで重要となる．[8] では，歌声の音高を直接モデル化するのではなく，歌声と楽譜の音高の差分をモデル化する手法が提案されている．楽譜の音高に対する揺らぎの傾向のみをモデル化することで，学習データに少ない音高パターンの学習を改善している．楽譜の音高系列を $\boldsymbol{p} = [p_1, p_2, ..., p_T]^\mathsf{T}$，モデルの出力系列を $\boldsymbol{\mu} = [\mu_1, \mu_2, ..., \mu_T]^\mathsf{T}$ とすると，予測対数 F0 系列 $\bar{\boldsymbol{c}}^{(F0)}$ は以下で表される．

$$\bar{\boldsymbol{c}}^{(F0)} = \boldsymbol{p} + \boldsymbol{\mu} \tag{1}$$

提案方式では，底を 2 とする対数 F0 を扱い，楽譜の音高系列は入力楽譜特徴量から入手する．

2. **スペクトル生成部**：スペクトル生成部では，再帰型ニューラルネットワークの一種である LSTM を用いて，時間的に変化する音声の音色特徴を含むメルスペクトログラムを推定する．メルスペクトログラムは人の聴覚特性に調整したメル尺度に基づく特徴量であり，短時間フーリエ変換（Short-time Fourier Transform:STFT）によって算出されたスペクトログラムの周波数方向にメルフィルタバンクを適用することで得られる．学習ステージでは，楽譜特徴量と生波形から分析された歌声の対数 F0 が入力される．一方，予測ステップでは，楽譜特徴量と F0 生成部で予測された歌声の対数 F0 が入力される．

3. **波形合成部**：波形合成部では，メルスペクトログラムと歌声の対数 F0 を WaveNet の補助特徴量とし，歌声の音声波形を予測する．学習ステップでは，生波形から分析されたメルスペクトログラムと歌声の対数 F0 を補助特徴量とする．一方，予測ステップでは，スペクトル生成部で推定されたメルスペクトログラムと，F0 生成部で推定された歌声の対数 F0 を補助特徴量とする．

3 単一の母音のみで歌う歌唱データベース

提案方式の合成モデルを実現するためには，単一母音のみで歌う歌声とその楽譜を有する歌唱データベースが必要となる．本節では，提案方式で使用する歌唱データベースの概要について説明する．

3.1 データベースの内容

本データベースは，著作権の保護期間が終了した童謡 30 曲で構成される．各楽曲について，単一の母音のみで歌った歌唱音声データ（ボーカルのみ），および楽曲の歌唱メロディを MIDI データとして入力した楽譜情報の二組で構成される．歌唱音声データは，日本語発話における 5 母音「あ」「い」「う」「え」「お」のうち，一つの母音のみで一楽曲通して歌ったものである．歌唱者は，合唱経験のある男性 1 名である．データ数は合計 150（30 曲 × 5 母音），収録時間は約 115 分（1 母音あたり 23 分）である．MIDI データは，各楽曲の歌唱メロディを独自で採譜して打ち込み作業を行っている．このとき，MIDI ノートがビートと正確な分数のビートに設定する処理（クオンタイズ処理）を適用している．収録環境は，岡山大学工学部 4 号館の防音室（暗騒音レベル 20 dB 以下）である．収録用のコンデンサマイクには，AKG C414-XLII を利用し，USB オーディオインタフェースの RolandOCTA-CAPTURE UA-1010 を経由して PC で録音した．音声ファイルは，サンプ

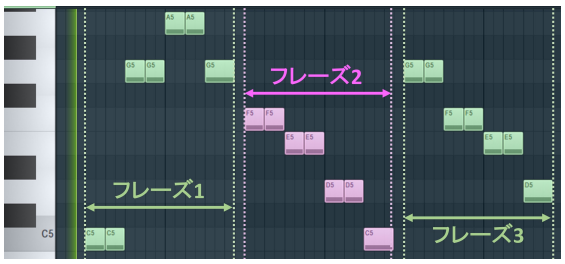

図 3: 主メロディにおけるフレーズ設定．フレーズ内では母音を繋げて歌う

リング周波数 96 kHz，量子化ビット数 24 bit，1 チャネル（モノラル）として記録している．

3.2 収録概要

本収録に関わるすべての歌唱の録音は FL Studio 20 (Image-Line Software — Hookup, Inc.) 上でおこなった．収録を円滑に進めるために FL Studio 20 上のピアノロール画面を補助として利用し，あらかじめ用意された打ち込みによる演奏音に合わせて歌唱の収録をおこなった．歌唱時には図 3 のように主メロディは各楽曲毎に任意長のフレーズを設け，フレーズ内のメロディは母音を繋げて歌うこととした．また，歌唱表現（ビブラートやポルタメントなど）は歌唱者の裁量に委ねた．ただし，基音を大きく逸脱するような過剰な歌唱表現はおこなわないこととした．

4 評価実験

4.1 実験条件

3 章で示した単一母音のみで歌う歌唱データベースのうち，130 曲（1 母音当たり 26 曲）を学習データ，10 曲（1 母音当たり 2 曲）を検証データ，10 曲（1 母音当たり 2 曲）を評価データとした．また，音声データは，事前に 16 bit，24 kHz に変換している．変換処理には，SoX（Sound eXchange）を用いた．STFT，および WORLD 音声分析について，FFT 長を 43 msec，フレームシフト長を 10 msec としたスペクトル生成部，および波形合成部におけるモデルパラメータの詳細を表 1 に示す．本実験で使用する楽譜特徴量の概略は次の通りである．

1. **前，現在，次の音符の音程**：A1（55 Hz）から A5（880 Hz）までの範囲の 49 段階の音程および休符
2. **音符内位置**：現在のフレームの音符内位置
3. **母音 ID**：日本語母音「あ」「い」「う」「え」「お」の識別

3 段階の学習ステップにおいて，最適化手法は共通して Adam[9] を用いた．F0 生成部，スペクトル生成部における LSTM の中間層ユニット数はそれぞれ 64，128 とし，中間層は共通して 3 とした．スペクトル生成部

表 1: モデルの入出力パラメータ

F0 生成部（LSTM）	
入力	楽譜特徴量：156dims
出力	差分対数 F0：1dim
	有声無声フラグ（V/UV）：1dim
スペクトル生成部（LSTM）	
入力	楽譜特徴量：156dims
	対数 F0（One-hot）：482dims
出力	メルスペクトログラム：80dims
波形合成部（WaveNet）	
補助特徴量	メルスペクトログラム：80dims
	対数 F0：1dim
入力 / 出力	音声波形：1dim

の入力である F0 について，フレーム毎の F0 の値を音程 A1（55 Hz）から A5（880 Hz）までの範囲内で 10 cent 間隔で離散化し，音高無しのを含めた 482 次元の one-hot ベクトルに変換したものを用いた．波形合成部における WaveNet の Residual block は 30 層，受容野は 128ms(3070 samples) とした．

4.2 客観評価実験

客観評価実験では，次に示す 2 つのシステムについて比較を行い，提案方式で示した 3 段階のステップからなる合成モデルの有効性を学習に使用していないテストデータを用いて評価する．

- **baseline**：合成モデルは単一の WaveNet で構成される．このとき，WaveNet の補助特徴量は 4.1 節で示した 156 次元の楽譜特徴量である．
- **proposed**：提案方式による合成モデル．

提案方式は F0 生成部とスペクトル生成部を介して段階的に音響特徴量を予測し，それらの音響特徴量を WaveNet の補助特徴量としている．一方で，ベースラインは，提案方式のように音響特徴量を段階的に予測する構造とは異なり，WaveNet の補助特徴量として楽譜特徴量を直接用いて歌声の合成を行う．ベースラインと提案方式を比較することで，提案方式のように段階的に音響特徴量を予測し，歌唱合成する方式の有効性を評価する．

評価尺度は，メルケプストラム歪み（Mel-Cepstral Distortion： MCD）[db]，および F0 の二乗平均平方根誤差（Root Mean Square Error:RMSE）と標準偏差（Correlation coefficient:Corr）を用いる．このとき，評価に用いる合成歌声の F0 は WORLD の Harvest 推定法を用いて分析する．評価尺度の算出結果を表 2 に示す．提案システムは音高の推定精度（F0-RMSE，F0-Corr）の観点でベースラインよりも良い性能を示してい

表 2: 全 5 母音に関する Pitch-Note を基準とした評価尺度の算出結果

	MCD [dB]	F0-RMSE [cent]	F0-Corr
baseline	**8.320**	125.89	0.9063
Proposed	8.694	**39.80**	**0.9842**

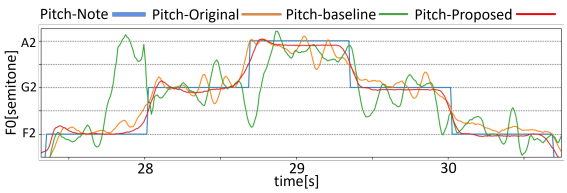

図 4: 対数 F0 軌跡の比較（青 : 楽譜，橙 : Original，緑 : baseline，赤 : proposed）

る．F0-RMSE に関して，ベースラインは 125.89cent であるのに対し，提案方式は 39.80cent を示している．十二平均律の半音は 100cent であり，100cent より大きい誤差は半音より大きい差を意味するため，聴感的に大きく影響する．従って，125.89cent から 39.80cent への F0-RMSE 値の減少は，音高予測精度が大きく改善されたことを示唆する．

一方，提案システムはメルケプストラム歪みの観点でベースラインより劣る性能を示している．以下ではメルケプストラム歪みの値が増加した原因について，対数 F0 軌跡の観点から考察を行う．対数 F0 軌跡の比較を図 4 に示す．Pitch-Original は評価データ（元音声）から抽出された対数 F0 の軌跡，Pitch-baseline，Pitch-Proposed は各々のシステムの合成歌声から抽出された対数 F0 の軌跡をそれぞれ示している．提案方式の F0 に着目すると，楽譜の音高に忠実ではあるが，揺らぎ成分がほとんど消失していることが確認できる．一方，ベースラインの F0 に着目すると，28 秒付近にみられる大きな音高の逸脱はあるものの，揺らぎの成分は保持していることが確認できる．音符内を通して常に平坦な音高が推定された結果，提案方式におけるメルケプストラム歪みの値がベースラインより大きい値を示したと推測される．

客観評価実験の結果から，楽譜通りの基本的な音高制御の観点では提案方式の方が優れているといえるが，音高の自然な揺らぎ成分の表現の観点ではベースラインに劣っているといえる．提案方式の F0 の揺らぎが消失した原因として，F0 生成部，およびスペクトル生成部における音響特徴量の推定結果の過剰な平滑化が考えられ，音響特徴量の推定精度を向上させる改善策が求められる．改善策の例として，新たな入力特徴量の検討や，深層自己回帰モデル（DeepAutoregressive model）の利用などが挙げられる．

5 まとめ

本報告では，DNN を用いて楽譜情報から自然性の高いバックコーラス歌唱を合成する方式を検討した．F0 生成部，スペクトル生成部，波形合成部の 3 段階で構成される合成モデルを提案した．評価実験により，提案方式は楽譜通りの基本的な音高の制御が可能であることを確認した。同時に，合成歌声の音高の揺らぎ成分が消失していることを確認した．今後の課題として，合成歌声の音高の揺らぎ成分の再現精度を向上させる改善策を検討する事が挙げられる．

参考文献

[1] 剣持秀紀，大下隼人，“歌声合成システム VOCALOID-現状と課題，” 情報処理学会研究報告, 音楽情報科学 (MUS)，vol.2008，no.12 (2008-MUS-074)，pp.51–56，2008.

[2] K. Saino, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, “An HMM-based singing voice synthesis system,” Proceedings of Ninth International Conference on Spoken Language Processing, pp.2274–2277, 2006.

[3] M. Nishimura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, “Singing Voice Synthesis Based on Deep Neural Networks.,” Proceedings of Interspeech, pp.2478–2482, 2016.

[4] Y. Hono, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, “Sinsy: A deep neural network-based singing voice synthesis system,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.29, pp.2803–2815, 2021.

[5] M. Blaauw and J. Bonada, “A neural parametric singing synthesizer modeling timbre and expression from natural songs,” Applied Sciences, vol.7, no.12, p.1313, 2017.

[6] A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “WaveNet: A generative model for raw audio,” arXiv preprint arXiv:1609.03499, pp.1–15, 2016.

[7] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural computation, vol.9, no.8, pp.1735–1780, 1997.

[8] M. Nishimura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, “Singing voice synthesis based on deep neural networks.,” Interspeech, pp.2478–2482, 2016.

[9] D.P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, pp.1–15, 2014.