歌い手の声質・歌い方を自動で学習・再現できる統計モデルに基づく歌声合成システム

代表研究者　　徳　田　恵　一　　名古屋工業大学大学院　教授
共同研究者　　南　角　吉　彦　　名古屋工業大学大学院　准教授
共同研究者　　大　浦　圭　一　郎　名古屋工業大学大学院　特任准教授

1 はじめに

近年，歌声合成が幅広い分野で注目を集めている．歌声合成とは音声合成の応用技術のひとつで，コンピュータに楽譜を与えて任意の歌を歌わせる技術である．歌声合成技術を用いることにより，歌手に歌唱の依頼をせずとも歌声を手に入れることが可能となり，ボーカル付きの楽曲制作が容易になった．また，歌声合成技術を用いて有名歌手の歌声を再現することで，有名歌手に直接の依頼が不可能な場合でも歌声を手に入れることが可能である．他にも，カラオケのボーカルアシスト機能やゲームソフト，携帯電話のアプリケーションなどのエンターテイメント分野での応用も見られる．このような歌声合成技術の普及に伴い，音質の向上や歌声の多様性が求められ，より簡単に高品質な歌声を合成できるシステムが必要とされている．

歌声合成の主な手法として，VOCALOID [1] に代表される波形接続型の歌声合成や，統計モデルに基づく歌声合成 [2, 3] が挙げられる．波形接続型の歌声合成では，音声波形の素片を楽譜情報に従ってデータベースから選択し，その素片を接続して歌声を合成する．この手法では音声波形そのものを合成に用いるため，高品質な歌声が合成可能な反面，音声波形の接続部分に歪みが生じやすいという問題もある．また，自然な歌声を合成するために，膨大なデータと人手による細かな調整が必要となる．一方，統計モデルに基づく歌声合成では，歌声に対する音響特徴量を統計的に学習してモデル化し，モデルを基に歌声を合成する．この手法では音声波形そのものではなく統計モデルを用いて歌声を合成するため，波形接続型に比べてデータ量を抑えることが可能である．

統計モデルに基づく歌声合成のなかでも，隠れマルコフモデル (Hidden Markov Model; HMM) に基づく歌声合成 [2, 3] では，あらかじめ用意した歌声データから音高を表す基本周波数や，音色を表すスペクトルなどの歌声の音響特徴を抽出し，HMM を音素単位でモデル化する．合成時には，与えられた楽譜に従って連結した HMM からパラメータを生成し，歌声を合成する．HMM 歌声合成は動的特徴量を考慮して学習するため，滑らかな歌声を合成することが可能である．また，モデルパラメータを適切に変更することで様々な声質の歌声を合成できる．しかし，合成歌声は自然歌声と比べ，自然性の点で未だに人間の歌声に到達できていない．そのため，更なる改善が必要である．

近年，計算機の飛躍的な性能の向上とともに，深層学習 (deep learning) が注目を浴びている．深層学習とは，深層構造をもつニューラルネットワーク (Deep Neural Network; DNN) を用いた機械学習法であり，DNN によってデータからそのデータ特有の特徴を抽出した内部表現を学習し，得られた内部表現を用いることにより音声認識や画像認識の分野において高い性能を示している [4, 5]．DNN は，画像や音声データから特徴量への変換といった直接因果関係を表すことができない非線形写像を比較的容易に表現可能であり，入力から出力への非線形変換を精度良く実現可能である．また，統計モデルに基づく音声合成分野 [6] においても，DNN を用いた手法は大きな成果をあげている [7]．DNN に基づく音声合成では，DNN を言語特徴量から音響特徴量を推定する音響モデルと考えることにより，HMM より精度の高い音響特徴量の推定が可能となり，合成音声の品質が向上した．このことから，音声合成の応用技術である歌声合成においても音響モデルを HMM から DNN に置き換えることで歌声の品質が向上することが期待されるが，DNN を用いた歌声合成の有効性はまだ確認できていない．そこで本研究では，DNN を歌声合成に適応させた DNN に基づく歌声合成を提案し，有効性を評価する．

また，我々はこれまでに HMM 歌声合成システム Sinsy のデモページを公開し運用してきた．Sinsy のデモページでは，ユーザが楽譜をアップロードすることで，誰でも簡単に歌声を合成することができる．本研究の成果についても，Sinsy のデモページにて公開した．Sinsy のデモページ公開・運用により，歌声合成の裾野を広げ Consumer Generated Media の活性化に繋がると期待する．

2 歌声合成システム

2-1 HMM 歌声合成システム

HMM 歌声合成では，歌声から推定したメルスペクトルや対数基本周波数などの音響特徴量と楽譜から推定した言語特徴量の対応関係を HMM によりモデル化する(図 1)．HMM 歌声合成は学習部と合成部に分かれており，学習部では，歌声から音響特徴量を抽出し，HMM を音素単位でモデル化する．その際，決定木に基づくコンテキストクラスタリングを導入し，質問に従って音響特徴が類似するモデルのパラメータを共有化することで，モデルの学習データが少量になることを防ぐことができる．さらに，質問を用いたコンテキストクラスタリングにより，学習データに存在しない未知の言語特徴量にも対応することが可能となる．合成部では，音素単位の HMM を楽譜から得られる言語情報に従って連結することにより，文単位の HMM とし音響特徴量の系列を生成し，合成フィルタに通すことにより歌声を合成する．

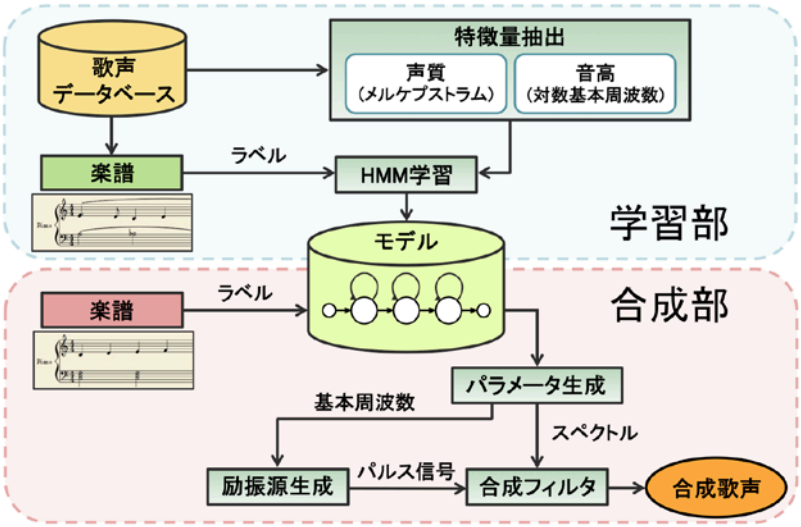

図 1 HMM 歌声合成システム．歌声から推定したメルケプストラムと対数基本周波数を HMM でモデル化する．学習した HMM からパラメータを生成することで任意の歌声が合成できる．

HMM 歌声合成は統計的手法であるため，学習データに含まれない音高は合成することができない．あらゆる音高を合成するためには，あらゆる音高を含む学習データを用意する必要がある．しかし，実際に歌声に含まれる音高には偏りがある．また，歌唱者の歌唱可能な音域外の音高を得ることができない点からも，あらゆる音高を偏りなく十分に含むデータを用意することは困難である．そこで音高を直接モデル化するのではなく，歌声の対数基本周波数と楽譜の音符の音高の差分をモデル化する音高正規化学習［8］が提案されている．

2-2 DNN 歌声合成システム

DNN は，因果関係を直接表すことができない非線形写像を比較的容易に表現でき，入力から出力への非線形変換が可能であるため，音声合成分野など様々な分野において成果をあげている．そこで本研究では，統計的歌声合において用いられてきた HMM を DNN に置き換えることで，歌声の品質向上を目指す．DNN 歌声合成では，入力を言語特徴量，出力を音響特徴量として，その対応関係を DNN によりモデル化する．DNN 歌声合成も HMM 歌声合成と同様に学習部と合成部に分かれており，学習部では，DNN の入力として楽譜情報から推定する言語特徴量，出力として歌声から推定する音響特徴量を用い，これらの対応関係をフレーム単位でモデル化する（図 2)．本研究ではスペクトル特徴量と基本周波数特徴量に着目し，メルケプストラムと対数基本周波数の静的・動的特徴量を音響特徴量として用いて，単一の DNN で学習する．また，楽譜情報から推定される言語特徴量には音符情報や音素情報に関わる質問を数値化して入力とする．これは，HMM 歌声合成におけるコンテキストクラスタリングを，DNN に置き換えたとみなすことができる．合成部では，言語特徴量を DNN に入力することにより音響特徴量の静的・動的特徴量を推定し，静的・動的特徴量の関係を考慮したパラメータ生成を行う．このパラメータを合成フィルタに通すことにより歌声を合成する．

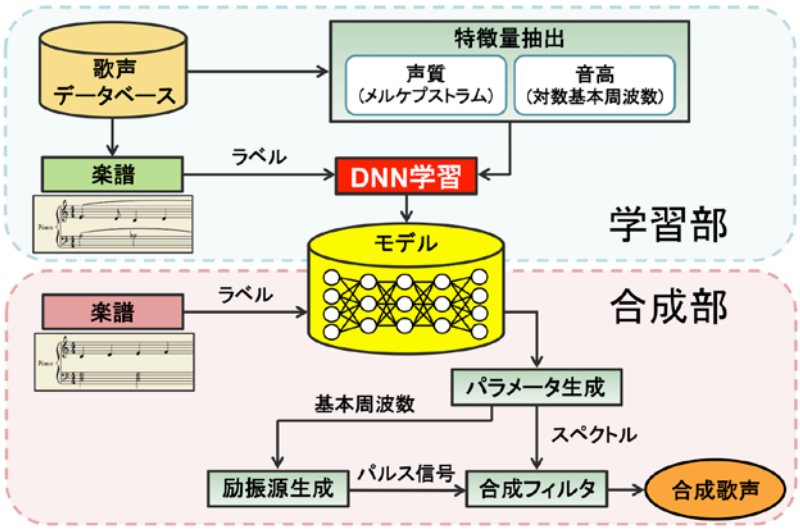

図 2 DNN 歌声合成システム．統計モデルを HMM から DNN に置き換えることで，より自然な歌声を合成可能とする．

DNN 歌声合成も HMM 歌声合成と同じく，学習データに含まれていない音高を合成することができない．そこで本研究では，歌声の対数基本周波数の静的特徴量と音符の音高との差分を求め，DNN で学習する．しかし，歌声の無声部と楽譜の休符部には対数基本周波数の値が存在しない．また，楽譜の音符境界と歌声の音符境界にはズレが生じ，音符上の無声部および休符上の有声部は差分を求めることができない．これらの対処として，差分が計算できない部分を 0 に近似し，無声部や休符部に線形補間を用いる手法を考える．本手法では音高の差分をモデル化するため，対数基本周波数の静的特徴量を線形補間するだけでなく，休符の音高に関しても線形補間をする必要がある．

3 実験

3-1 実験条件

DNN に基づく歌声合成の有効性を示すために，評価実験を行った．本実験では歌声データベースとして，女性 1 名による童謡 70 曲を用いた．学習データとして 60 曲，テストデータとして残りの 10 曲を用いた．サンプリング周波数は 48kHz，量子化ビット数は 16bit である．音響特徴量として，STRAIGHT によって抽出されたスペクトルに，メルケプストラム分析を適用することにより得られた 49 次のメルケプストラム係数，対数基本周波数とそれらの 1 次，2 次の動的特徴量を用いた．DNN の入力は楽譜情報から推定する言語特徴量と継続長情報を表す 650 次元ベクトル，出力は音響特徴量の静的・動的特徴量と有声無声情報を表す 154 次元ベクトルである．ただし，静的・動的特徴量の最小値と最大値を基に，0.01 から 0.99 となるように正規化を行った．DNN の中間層・出力層の活性化関数にはロジスティックシグモイド関数を用いた．客観評価尺度として，メルケプストラム歪み（Mel-cd）と対数基本周波数の平均二乗誤差（FORMSE）を用いた．DNN の中間層（1，2，3，4，5）と各層における隠れユニット数（128，256，512，1024，2048）のそれぞれの組み合わせで実験を行い，最も客観評価尺度の値が低い組み合わせで比較を行った．また，DNN の学習に用いる各音素の継続長情報は，学習済みの HMM を用いて推定した音素アライメント情報から設定しており，学習時も合成時も固定した音素アライメントを用いている．HMM は 5 状態の left-to-right 型 HSMM を用いた．

3-2 評価実験 1

音高の差分の学習において，歌声の対数基本周波数と楽譜の音高の対数基本周波数，また，それぞれの線形補間後の対数基本周波数の組み合わせで，どの手法が最も有効であるかを，客観評価尺度を用いて比較した．歌声の無声部または楽譜の休符部においての線形補間の有無の組み合わせは，図 3～6 に示す 4 通りが考えられる．

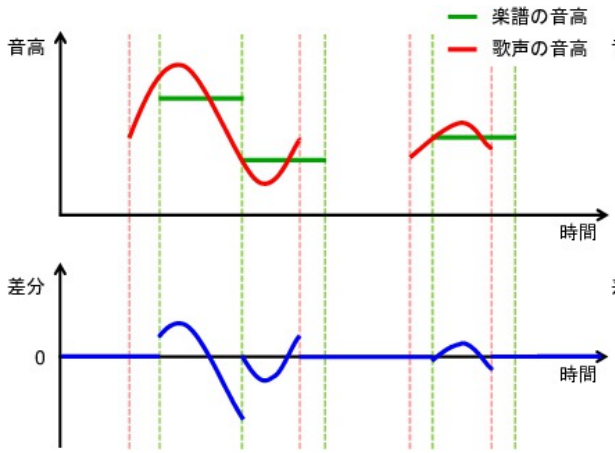

図 3 線形補間なし

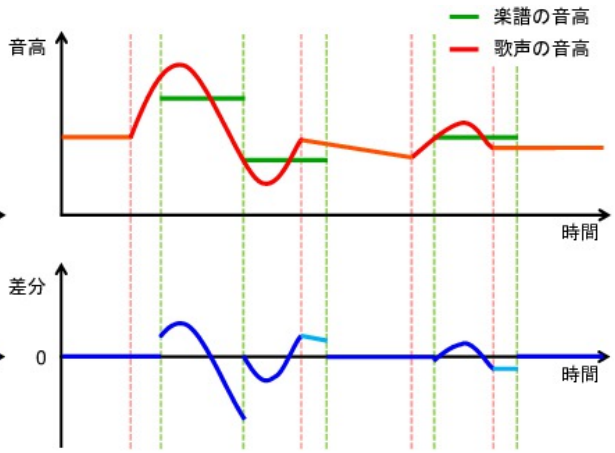

図 4 歌声の音高のみ線形補間

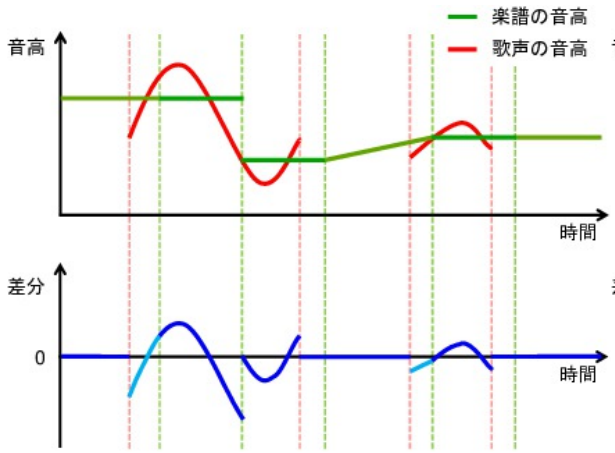

図 5 楽譜の音高のみ線形補間

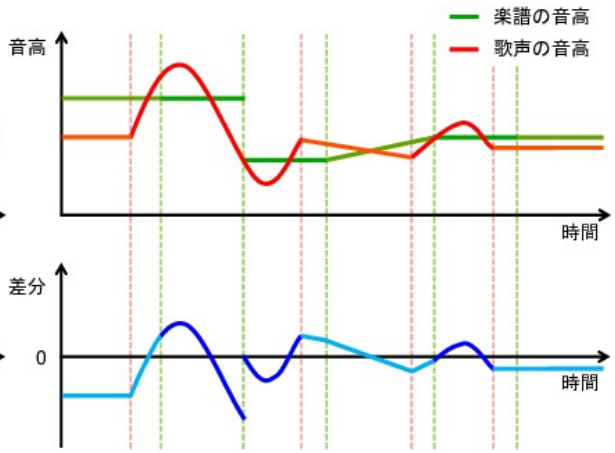

図 6 歌声の音高と楽譜の音高を線形補間

各図における差分の図は，それぞれの手法においての歌声の音高と楽譜の音高との差分の値を示す．また，差分の図における水色の線は，線形補間なしの場合とは異なる差分を取ることを示す．これらについて，FORMSE が最も小さい層数・隠れユニット数の組み合わせにおいて評価した．

表 1 に 4 手法の最も小さい FORMSE の値を示す．全手法において，FORMSE は中間層が 2 層，隠れユニット数が 1024 の組み合わせが最も小さい値を示した．線形補間後の歌声の対数基本周波数と音符の音高との差分を学習した場合が，最も FORMSE が小さかった．また，歌声の対数基本周波数を線形補間した場合に，線形補間をしなかった場合と比較して誤差値が減少する傾向がみられた．これは，対数基本周波数を抽出する際に，有声部を無声部と誤って抽出した部分に対して対数基本周波数を線形補間したことにより，抽出誤りによる影響が減少したためだと考えられる．また，音符の対数基本周波数の線形補間に関しては明確な傾向は見られなかった．

表 1　DNN 歌声合成における対数基本周波数の線形補間手法の比較

歌声の音高の線形補間の有無	×	×	○	○
楽譜の音高の線形補間の有無	×	○	×	○
FORMSE［LogHz］	0.04851	0.04847	0.04777	0.04784

3-3 評価実験 2

HMM 歌声合成システムと DNN 歌声合成システムの性能を比較するために客観・主観評価実験を行った．HMM 歌声合成システムは，HMM を用いてメルケプストラムと対数基本周波数を推定した手法である．HMM は無声部の線形補間を行っていない．DNN における対数基本周波数の線形補間は実験 1 で最も FORMSE が小さくなった

線形補間後の歌声の対数基本周波数と音符の音高との差分を学習する手法を用いた．表 2 に DNN 歌声合成システムの中間層数・隠れユニット数の組み合わせを示す．DNN (mgc)は最も Mel-cd が小さくなった中間層・隠れユニット数の組み合わせであり，DNN (lf0)は最も FORMSE が小さかった中間層・隠れユニット数の組み合わせで DNN の学習を行った手法である．DNN (separated)は，HMM 歌声合成と同様に，メルケプストラムと対数基本周波数を個別の DNN で学習した手法である．

表 2　各手法名と中間層数.隠れユニット数の組み合わせ

		組み合わせ	
		中間層数	隠れユニット数
DNN (mgc)		3	1024
DNN (lf0)		4	1024
DNN (separated)	lf0	1	1024
	mgc	3	1024

表 3 に Mel-cd と FORMSE の結果を示す．また，メルケプストラムと対数基本周波数を個別に DNN で学習した DNN (separated)が最も小さい Mel-cd を示していることから，DNN に基づく歌声合成手法が HMM より高い推定精度を達成できた.一方,DNN を用いた手法は HMM と比較して FORMSE が大きい傾向が見られた.これは，HMM と異なり DNN には対数基本周波数の無声部を含めたモデル化を行う構造が無いことによる影響と考えられる.

表 3　歌声合成における HMM と DNN の比較

	Mel-cd [dB]	FORMSE [LogHz]
HMM	5.162	0.04423
DNN (mgc)	5.027	0.04856
DNN (lf0)	5.054	0.04777
DNN (separated)	4.997	0.04729

次に，歌声の自然性に関する 5 段階平均オピニオン評点(MOS)による主観評価実験を行った．実験結果を図 7 に示す．この実験手法では，MOS の値が大きいほど自然な歌声であることを示す．図 7 より，提案法である DNN 歌声合成システムは，HMM 歌声合成システムより高い自然性のスコアを達成した．この結果からも，統計モデルとして DNN を用いた歌声合成システムの有効性が示された．3 手法の DNN 歌声合成システムのスコアには有意な差は見られなかった．今後は，より適切な DNN の構造を検討することで，より自然な歌声の合成を目指す．

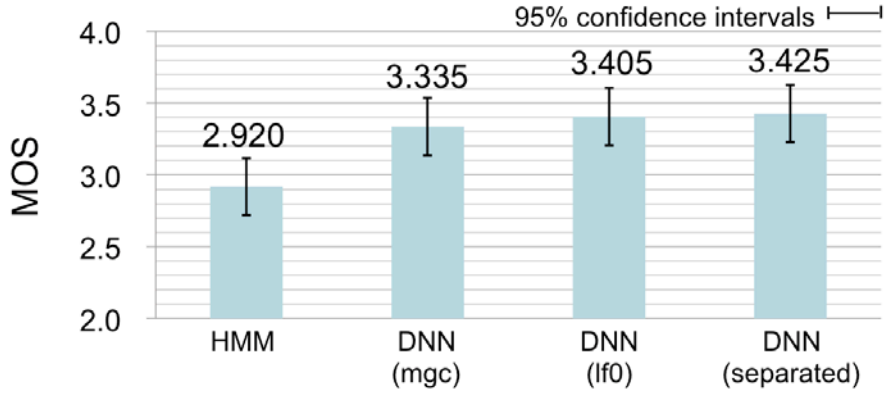

図 7 歌声の自然性に関する主観評価実験．図の縦軸は平均オピニオン評点(MOS)であり，値が大きいほど自然な歌声であることを示す.

4 歌声合成システム Sinsy

我々は，これまでに HMM 歌声合成システム Sinsy のデモページを公開し運用してきた(図 8)．Sinsy のデモページでは，ユーザが楽譜をアップロードすることで，日本語と英語の歌声を合成することができる．そして，言語の違いが歌に与える影響等を調査することで，中国語の歌声合成も実現し Sinsy デモページに公開した．これにより，ユーザは，日本語・英語・中国語の歌声コンテンツを容易に作成できるようになり，コンテンツ作成の幅が広がった．さらに，本研究の成果である DNN 歌声合成システムも Sinsy デモページにて公開した．今後も，歌声合成システムの研究を進め，ユーザが容易に自然な歌声のコンテンツを生成できる枠組みを開発・公開していく予定である．

図 8 Sinsy デモページ．楽譜をアップロードすることで，誰でも歌声を合成することができるシステム．

5 まとめ

本研究では，DNN を用いた歌声合成の有効性を確認するために，DNN に基づく歌声合成を提案した．DNN を楽譜情報から推定する言語特徴量から歌声から推定する音響特徴量の予測をする音響モデルと考え，言語特徴量と音響特徴量の対応関係をフレーム毎に学習した．主観評価実験の結果，提案法である DNN 歌声合成システムは HMM 歌声合成システムより自然な歌声が合成できることが示された．また，本研究の成果である DNN 歌声合成システムを，Sinsy のデモページに公開した． これにより，ユーザは容易に自然な歌声を合成することが可能となった．

【参考文献】

[1] H. Kenmochi and H. Ohshita, “VOCALOID-commercial singing synthesizer based on sample concatenation,” Proc. of Interspeech, 2007.

[2] K. Saino, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, “An HMMbased Singing Voice Synthesis System,” Proc. of ICSLP, pp. 1141– 1144, 2006.

[3] K. Oura, A. Mase, T. Yamada, S. Muto, Y. Nankaku, and K. Tokuda, “Recent Development of the HMM-based Singing Voice Synthesis System - Sinsy,” Proc. of Speech Synthesis Workshop, pp. 211–216, 2010.

[4] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” Proc. of ICASSP 2000, pp. 1315–1318, 2000.

[5] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” Signal Processing Magazine, Proc. of IEEE, vol. 29, no. 6, pp. 82–97, 2012.

[6] A. Krizhevsky, I. Sutskever, and G. Hinton, “Imagenet classification with deep convolutional neural networks,” Advances in neural information processing systems, 2012.

[7] H. Zen, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” Proc. of ICASSP 2013, pp. 7962–7966, 2013.

[8] K. Oura, A. Mase, Y. Nankaku, and K. Tokuda, “Pitch adaptive training for HMM-based singing voice synthesis,” Proc. of ICASSP, pp. 5377–5380, 2012.

〈発　表　資　料〉

題　　名	掲載誌・学会名等	発表年月
Singing voice synthesis based on deep neural networks	Interspeech 2016, pp. 2478-2482	2016 年 9 月
Temporal modeling in neural network based statistical parametric speech synthesis	9th ISCA Speech Synthesis Workshop, pp. 113-118	2016 年 9 月
オーディオブックを用いた表現豊かな音声合成のための言語特徴の検討	音声研究会, vol. 116, no. 414, SP2016-76, pp. 35-50	2017 年 1 月
DNN 音声合成における音響特徴量系列とその時間構造の同時モデル化	音声研究会, vol. 116, no. 414, SP2016-76, pp. 71-76	2017 年 1 月
風雲急を告げる音声合成研究の最新動向	電子情報通信学会情報・システムソサイエティ誌, vol. 21, no. 4, pp. 10-11	2017 年 2 月
ニューラルネットワークに基づく音声合成における音響特徴量抽出条件の検討	日本音響学会 2017 年春季研究発表会, pp. 263-264	2017 年 3 月