テキスト音声合成のための Auto-encoder を用いた音響モデル化に頑健な音響特徴量抽出

代表研究者　高木信二　国立情報学研究所　コンテンツ科学研究系　特任助教
共同研究者　山岸順一　国立情報学研究所　コンテンツ科学研究系　准教授

1 はじめに

近年，Deep Neural Network (DNN)に基づく統計的音声合成システムが高い性能を示している．例えば，DNN は音響モデルに用いられており，全らはテキストと音響特徴量との関係を学習するのに DNN を用いている．この手法で DNN は，HMM 音声合成システムにおける決定木に基づくコンテキストクラスタリングの代わりとして用いられる．また，Restricted Bolzmann Machines (RBMs)や Deep Belief Networks (DBNs)を GMM の代わりに HMM の出力分布として用いる手法や，Recurrent Neural Network や Long-short Term Memory をプロソディや音響特徴のトラジェクトリのモデル化に用いる手法が提案されている．その他，低次元の励震源パラメータ抽出のための Auto-encoder が提案されている．しかし，統計的音声合成システムから出力される合成音声は依然として統計モデルによる平均化に伴い過剰に平滑化されており，自然音声で観測される微細な構造を持つスペクトルを表現ができていないという問題がある．

本研究では，DNN の 1 つである Deep Auto-encoder，もしくは Deep Denoising Auto-encoder を用いた，振幅スペクトルからの効率的な低次元スペクトルパラメータの抽出法を検討する．従来広く用いられている低次元スペクトルパラメータ抽出法であるメルケプストラム分析は，対数スペクトルの線形変換(Discrete Cosine Transform)に基づいているが，DAE を用いることで非線形変換を内包でき，また，データドリブンに低次元特徴量を抽出できる．また，振幅スペクトルの微細な特徴を捉えるため，テキストから得られた言語特徴量から直接振幅スペクトルを合成する DNN の構築を行う．本研究では，入力テキストから直接高次元の振幅スペクトルを合成する DNN を構築するための効率的な学習法を検討する．提案法ではスペクトルパラメータ抽出器である Deep Auto-encoder (DAE)と音響モデルのための DNN を連結することで，直接振幅スペクトルを合成する DNN の初期化を行う．この提案法は DNN に基づく音声合成システムにおける Function-wise な Pre-training 手法と見なすことができる．分析再合成実験による Deep Auto-encoder を用いて抽出された低次元特徴量の評価，及び，テキスト音声合成実験による提案スペクトルモデリングの評価を行った．

2 Deep Auto-encoder に基づく低次元スペクトルパラメータの抽出

Auto-encoder は学習データの効率的な次元圧縮に広く用いられる Neural Network であり，入力データを隠れ層の空間へ写像する Encoder と元の信号へ復元する Decoder で構成される．入力データを $\mathbf{x}$, Bottleneck 特徴と呼ばれる圧縮された低次元表現を $\mathbf{y}$，復元されたデータを $\mathbf{z}$ とすると，隠れ層が 1 つの単純な Auto-encoder の Encoder，Decoder はそれぞれ次のように表現される．

Encoder: $\mathbf{y} = f_{\theta}(\mathbf{x}) = s(\mathbf{W}\mathbf{x}+\mathbf{b})$,

Decoder: $\mathbf{z} = g_{\theta'}(\mathbf{y}) = t(\mathbf{W}'\mathbf{y}+\mathbf{b}')$,

ここで，$\theta=\{\mathbf{W},\mathbf{b}\}$，$\theta'=\{\mathbf{W}',\mathbf{b}'\}$，はそれぞれ Encoder，Decoder のモデルパラメータを表す．入力データ，低次元表現の次元数をそれぞれ n，m とすると，$\mathbf{W}$ は m × n の行列，$\mathbf{b}$ は m 次元のベクトル，$\mathbf{W}'$ は n × m の行列，$\mathbf{b}'$ は n 次元のベクトルを表す．また，s，t は非線形変換を表現する．Decoder では非線形変換を用いず線形変換のみが用いられる場合もある．深層構造を持つ Auto-encoder は Deep Auto-encoder (DAE)と呼ばれる．本研究では DAE を用いることで振幅スペクトルからの効率的な低次元スペクトルパラメータの抽出を行う．

2-1 Deep Auto-encoder の学習

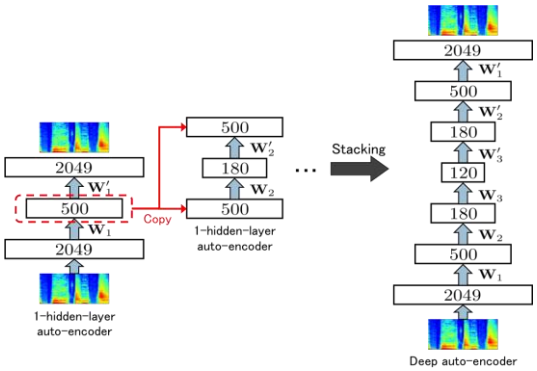


図 1: Deep Auto-encoder 構築のための Pre-training の手順

深層構造を持つ Neural Network を効果的に学習するには，Pre-training と呼ばれる初期値設定手法が用いられることが多い．図 1 に本研究で用いた DAE の Pre-training の手順を示す．Pre-training では隠れ層が 1 つの Auto-encoder を学習し，その Encoder 部，Decoder 部をそれぞれ積み重ねることで DAE を構築する．学習は Layer-wise に行われ，中間層の Pre-training では，入力データとして 1 つ下層の Pre-training 済み Auto-encoder の Encoder の出力が用いられる．Pre-training 後には，バックプロパゲーションを用いた Fine-tuning を行う．しかし，バックプロパゲーションを用いた Fine-tuning では下層において vanishing gradients の問題が発生することが知られている．この問題を解決するため，本研究では $\mathbf{W}' = \mathbf{W}^{T}$ とし，Encoder と Decoder の重み行列を共有することとした．ここで $\cdot^{T}$ は転置を表す．学習には確率的勾配降下法を用いた．

また，よりロバストに低次元特徴量を抽出するため，入力データにノイズを加えて Pre-training 学習を行う Denoising Auto-encoder が提案されている．本研究では，Pre-training 時の各層の入力の値に対して，ノイズの付与を検討した．

3 Deep Neural Network に基づく音響モデル

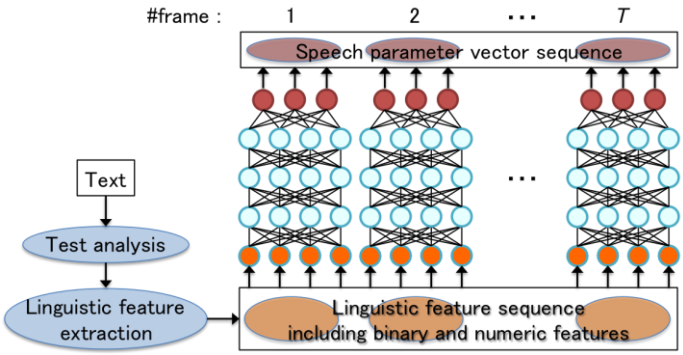


図 2: DNN に基づく音響モデルの枠組み

従来，HMM が音響モデルとして広く用いられているが，近年，DNN に基づく音響モデル(以降，DNN 音響モデル)が提案されている．ここでは代表的な DNN に基づく音響モデルの 1 つについて簡潔にレビューする．

図 2 に DNN 音響モデルの枠組みを示す．本手法は HMM 音声合成におけるコンテキストクラスタリングに用いられる決定木と同様の役割を持ち，DNN を用いることでテキストから抽出された言語特徴が音声から抽出された音声パラメータに写像される．入力データである言語特徴にはバイナリデータ(例えば，コンテキストに関する質問の答え)と数値データ(例えば，フレーズ内の単語の数，単語内のシラブルの位置，音素継続長)を用いることができる．では，音声パラメータには音源，スペクトルを表現する特徴量とそれらの時間微分が用いられている．DNN は学習データから抽出された言語特徴と対応する音声特徴を用いて確率的勾配降下法により学習することができる．また，任意テキストの音声パラメータは学習された DNN からフォワードプロパゲーションを用いることで予測できる．

3-1 Deep Neural Network に基づく言語特徴からのスペクトル予測

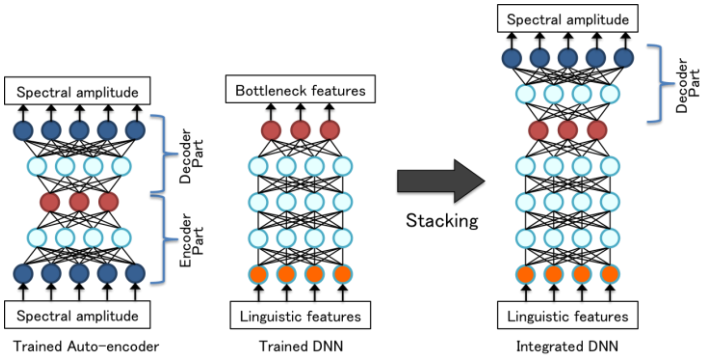


図 3: Deep Auto-encoder と DNN 音響モデルを用いた DNN スペクトルモデルの構築手順

本研究では，振幅スペクトルの微細な特徴を捉えるため，テキストから得られた言語特徴量から直接振幅スペクトルを合成する DNN の構築を行う．DNN 音響モデルにおいて音声パラメータに振幅スペクトルを用いることで，言語特徴から直接振幅スペクトルを合成する DNN を構築することは可能である．しかし，振幅スペクトルは従来スペクトルパラメータとして用いられるメルケプストラムや LSP と比較し非常に高次元である．例えば，サンプリング周波数 48kHz の音声データの場合，40〜60 次程度のメルケプストラムが用いられることが多いが，振幅スペクトルの次元数は FFT 長に依存し 2049 次程度が用いられる．言語特徴量とこのような高次元振幅スペクトルを直接関連付ける DNN を適切に構築するためには，より効率的な学習が必要であると考えられる．そこで本研究では，一般的に用いられている統計的音声合成システムの構築手順に基づき，直接スペクトルを合成する DNN の Function-wise な Pre-training 手法を提案する．つまり，DNN を用い音響特徴量抽出器と音響モデルをそれぞれ構築し，それらを積み重ね統合することで最終的な DNN の初期化を行う．

図 3 に提案法による DNN に基づくスペクトルモデル構築手順を示す．手順は次の通りである．

[Step 1.] 振幅スペクトルを用いた Deep Auto-encoder の学習を行い，Step 2. での DNN 音響モデル学習のため bottleneck 特徴を抽出する．Deep auto-encoder の学習では Layer-wise な Pre-training 等の初期化手法を用いることができる．
[Step 2.] Step 1. で抽出された bottleneck 特徴を用い DNN 音響モデルを学習する．DNN 音響モデルの学習においても Layer-wise な Pre-training 等の初期化手法を用いることができる．
[Step 3.] 学習された DNN 音響モデルと Deep Auto-encoder の Decoder 部を積み重ね，所望の構造を持つ DNN を構築する．その後，全ネットワークの最適化を行う．

このように，一般的な統計的音声合成システムの構築手順に基づき，DAE の Decoder 部，及び，DNN 音響モデルを用いることで，言語特徴と振幅スペクトルを直接関連付ける DNN を明示的に初期化する．初期化後には，全ネットワークに対して学習データを用い確率的勾配降下法により Fine-tuning を行う．

4 実験

Deep Auto-encoder を用いた低次元スペクトルパラメータ抽出の有効性を示すため，まず振幅スペクトル再構築による分析再合成実験を行った．次に，提案法による DNN に基づくスペクトルモデリングの有効性を示すため，テキスト音声合成実験を行った．実験データには女性プロナレータにより発話された英語 4,558 文を用いた．分析再合成実験では 4,558 文中の 3,676 文を学習データとし，441 文をテストデータとした．テキスト音声合成実験では 4,558 文全てを学習データとし，テスト文として異なる 180 文を用いた．また，サンプリング周波数は 48kHz である．FFT 長を 2049 ポイントとし，STRAIGHT を用いてスペクトルを抽出し，対数振幅スペクトルを用いた．

4-1 分析再合成実験

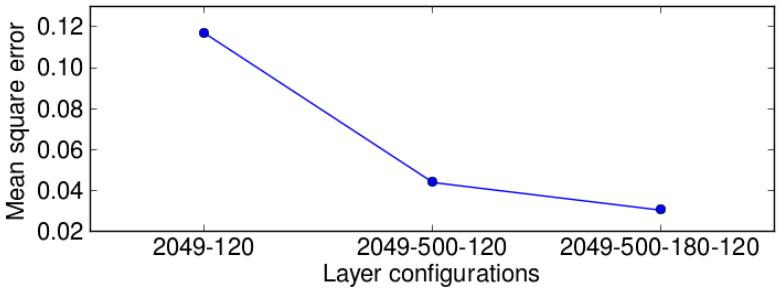
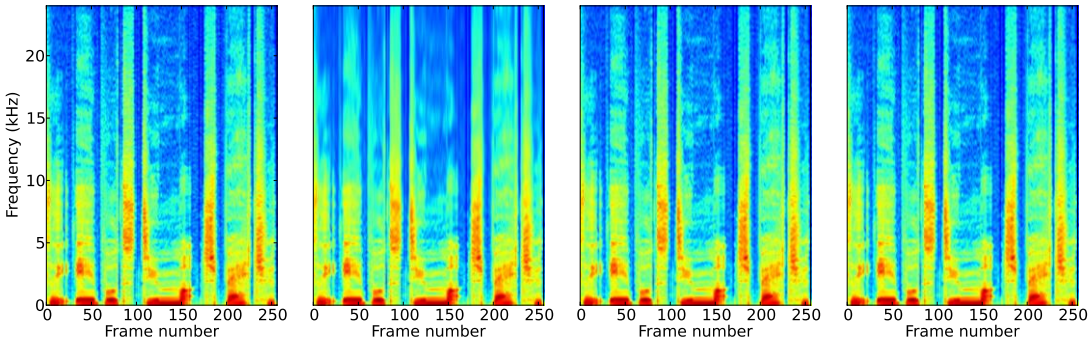


図 4: Deep Auto-encoder の構造の違いによる元対数振幅スペクトルと再構築された対数振幅スペクトルの平均二乗誤差．同次元の Bottleneck 特徴を扱うが隠れ層数が異なる．



(a)Original (b)MCEP (c)DAE (d)DDAE

図 5: 元スペクトルと各手法により再構築されたスペクトログラム

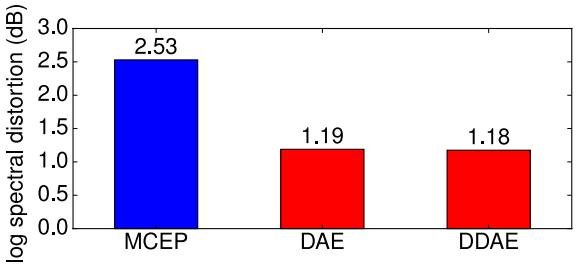


図 6: 元振幅スペクトルと各手法により再構築された振幅スペクトルの対数振幅スペクトル距離

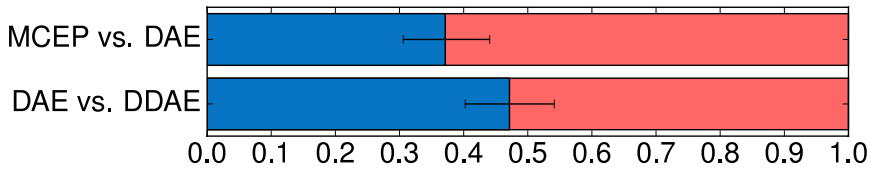


図 7: 主観評価実験(分析再合成)

振幅スペクトル再構築による分析再合成実験では，メルケプストラム分析 (MCEP)，Deep Auto-encoder (DAE)，Deep Denoising Auto-encoder (DDAE)の 3 手法を比較した．Auto-encoder で用いる際には対数振幅スペクトルを 0.0--1.0 の範囲へ正規化した．まず，図 4 にテストデータを用いた Deep Auto-encoder の構造の違いによる，元対数振幅スペクトルと再構築された対数振幅スペクトルの平均二乗誤差を示す．図 4 から分かるように平均二乗誤差は隠れ層が多いほど減少していることが分かる．この結果を踏まえ，以降の実験では，DAE と DDAE の Auto-encoder の構造を，隠れ層は 7，各隠れ層の素子数は 2049，500，180，120，180，

500, 2049 とした．そのため，120 次元のスペクトルパラメータが抽出される．MCEP においても同次元の 119 次メルケプストラム（0 次含む）を抽出した．

図 5 に元スペクトログラムと各手法により再構築されたスペクトログラムを示す．図 5 から Deep Auto-encoder を用いることで精度よく再構築されていることがわかる．また，図 6 に元振幅スペクトルと再構築された振幅スペクトルの対数振幅スペクトル距離を示す．この図から MCEP と比較して，DAE，DDAE は距離が大幅に減少していることがわかる．次に，図 7 に主観評価実験結果を示す．この実験ではスペクトル以外の要因を統一するため，全ての手法において，音声サンプルは再構築された振幅スペクトル，及び，音声分析時に得た基本周波数，非周期成分を用い STRAIGHT Vocoder を用いて合成した．主観評価実験では MCEP と DAE の比較，及び，DAE と DDAE の比較を対比較で行った．この実験結果より DAE は MCEP よりも自然性の高い音声が合成できていることがわかる．しかし，客観評価実験，主観評価実験共に DAE と DDAE の結果には大きな差はなかった．

4-2 テキスト音声合成実験

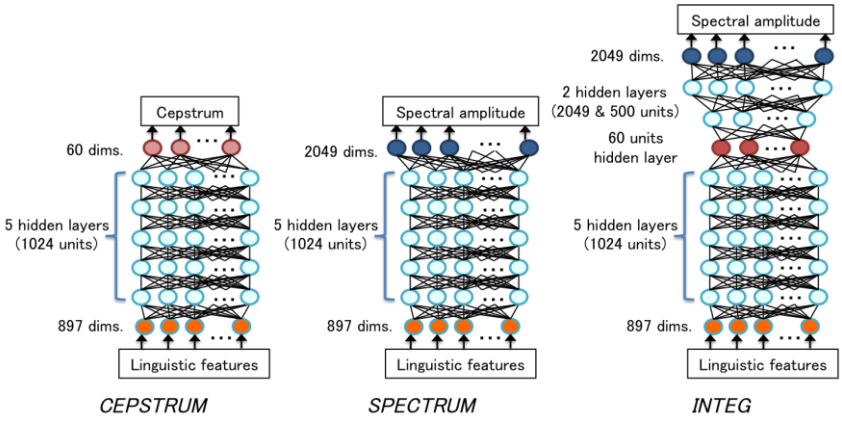


図 8：各手法で構築された DNN の構造

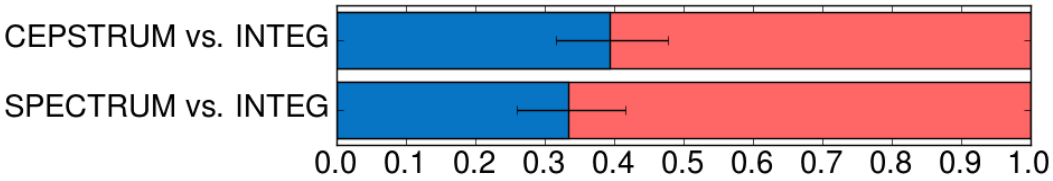


図 9：主観評価実験(テキスト音声合成
)

テキスト音声合成実験では，メルケプストラムを出力する DNN（以降，CEPSTRUM と呼ぶ），CEPSTRUM と同様の構造を持つが振幅スペクトルを出力する DNN（以降，SPECTRUM と呼ぶ），提案 Pre-training 手法を用いて初期化した振幅スペクトルを出力する DNN(以降，INTEG と呼ぶ)の 3 手法を比較した．テキスト音声合成実験では全ての手法で音響特徴量に 1 階微分，2 階微分は用いなかった．図 8 に各手法で構築された DNN の構造を示す．全手法において DNN 音響モデルの構造は隠れ層数 5，全ての隠れ層の素子数を 1024 とした． DNN 音響モデルは Pre-training を行わず，モデルパラメータはランダム値で初期化した．一般的に統計的音声合成システムにおいて用いられるスペクトルパラメータの次元数を考慮し，INTEG において DNN の初期化に用いられる Deep Auto-encoder の構造は隠れ層数 5，各隠れ層の素子数は 2049，500，60，500，2049 とし，60 次元の bottleneck 特徴を抽出した．そのため INTEG では，最終的に隠れ層数 8，各隠れ層の素子数は 1024，1024，1024，1024，1024，60，500，2049 の DNN が構築される．CEPSTURM では bottleneck 特徴と同次元の 59 次メルケプストラム(0 次含む)を用いた．本実験では全手法で DNN は出力として振幅スペクトル，または，スペクトルパラメータのみを扱い，音声の合成に必要となるその他の特徴量(基本周波数，非周期成分)は HMM 音声合成システムにより合成した．HMM 音声合成システム構築には 60 次メルケプストラム，基本周波数，25 次非周期成分とそれらの 1 階微分，2 階微分を用いた．コンテキストラベルは発音辞書 Combilex を用いて作

成された．DNN 音響モデルの入力として用いられる言語特徴は 897 次元であり，858 次のバイナリデータ，39 次の数値データから構成される．DNN 音響モデルの入力データとして用いられる音素継続長は HMM 音声合成システムを用いて推定した．言語特徴，スペクトルパラメータ，対数振幅スペクトルは，DNN で用いる際正規化を行った．INTEG では bottleneck 特徴の正規化は行わず，そのため，統合された DNN では隠れ層において正規化処理は行われない．言語特徴は平均 0 分散 1 に，スペクトルパラメータ，対数振幅スペクトルは 0.0-1.0 の範囲への正規化を行った．

テキスト音声合成実験の結果を示す．図 9 に主観評価実験結果を示す．主観評価実験では CEPSTRUM と INTEG の比較，及び，SPECTRUM と INTEG の比較を対比較で行った．この実験結果より INTEG は CEPSTRUM，SPECTRUM よりも自然性の高い音声が合成できていることがわかる．提案法により言語特徴と振幅スペクトルを直接関連付ける DNN が適切に学習されたためだと考えられる．

6 まとめ

本研究では，統計的音声合成システムの品質向上のため，Deep Auto-encoder を用いた効率的な低次元スペクトルパラメータ抽出を行った．分析再合成において提案手法は高い評価値を得た．さらに，入力テキストから得られた言語特徴から直接振幅スペクトルを合成する DNN の構築手法を提案した．一般的な統計的音声合成システム構築手順に基づき，スペクトルパラメータ抽出器である Deep Auto-encoder と音響モデルのための DNN を用い，効果的に Pre-training を行った．テキスト音声合成実験で合成音声の品質改善を確認することができた

【参考文献】

[1] Y. Fan and Y. Qian and F. Xie and F. K. Soong, "TTS Synthesis with Bidirectional LSTM Based Recurrent Neural Networks", Proceedings of Interspeech, 1964-1968, 2014.

[2] R. Fernandez and A. Rendel and B. Ramabhadran and R. Hoory, "Prosody Contour Prediction with Long Short-Term Memory, Bi-Directional, Deep Recurrent Neural Networks", Proceedings of Interspeech, 2268-2272, 2014.

[3] G. E. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks", Science 28, vol. 313, num. 5786, 504-507, 2006

[4] H. Kawahara and I. Masuda-Katsuse and A. Cheveigne, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds", Speech Communication, vol. 27, 187-207, 1999.

[5] Z.-H. Ling and L. Deng and D. Yu, “Modeling Spectral Envelopes Using Restricted {Boltzmann} Machines and Deep Belief Networks for Statistical Parametric Speech Synthesis”, Audio, Speech, and Language Processing, IEEE Transactions on, vol. 21, 2129-2139, 2013.

[6] K. Richmond and R. Clark and S. Fitt, "On generating Combilex pronunciations via morphological analysis", Proceedings of Interspeech, 1974-1977, 2010.

[7] P. Vincent and H. Larochelle and Y. Bengio and P. Manzagol, “Extracting and composing robust features with denoising autoencoders", ICML, 1096-1103, 2008.

[8] S. Vishnubhotla and R. Fernandez and B. Ramabhadran, "An autoencoder neural-network based low-dimensionality approach to excitation modeling for HMM-based text-to-speech", "Proceedings of ICASSP", 4614-4617, 2010.

[9] H. Zen and A. Senior and M. Schuster, "STATISTICAL PARAMETRIC SPEECH SYNTHESIS USING DEEP NEURAL NETWORKS", Proceedings of ICASSP, 7962-7966, 2013.

[10] H. Zen and K. Tokuda and A. W. Black, “Statistical parametric speech synthesis", Speech Communication, vol. 51, 1039-1064, 2009.

〈発　表　資　料〉

題　　名	掲載誌・学会名等	発表年月
Constructing a Deep Neural Network based Spectral Model for Statistical Speech Synthesis	NOLISP	2015 年 5 月
複数の Feed-forward Deep Neural Network に基づく統計的パラメトリック音声合成	信学技報　115(146) 49-54	2015 年 7 月
Multiple Feed-forward Deep Neural Networks for Statistical Parametric Speech Synthesis	INTERSPEECH, 2242-2246	2015 年 9 月
統計的パラメトリック音声合成のための DNN を用いた特徴抽出・音響モデル・ポストフィルタ	日本音響学会秋季研究発表会, 239-240	2015 年 9 月
A Function-wise Pre-training Technique for Constructing a Deep Neural Network based Spectral Model in Statistical Parametric Speech Synthesis	MLSLP	2015 年 9 月
統計的パラメトリック音声合成のための FFT スペクトルからの Deep Auto-encoder に基づく低次元音響特徴量抽出	信学技報　115(346) 99-104	2015 年 12 月