Vol.2015-SLP-105 No.2
2015/2/27

Deep Neural Networkに基づく音響特徴抽出・音響モデルを用いた統計的音声合成システムの構築

高木 信二[1,a)]　山岸 順一[1,b)]

概要：近年，Deep Neural Network (DNN) を用いた手法が様々な分野で高い性能を示しており，統計的音声合成においても DNN を用いた手法が注目を集め，盛んに研究されている．従来，統計的音声合成システムでは音声特徴量の 1 つであるスペクトルは，低次元のスペクトルパラメータ (例えば，メルケプストラムや LSP) によって表現され，隠れマルコフモデル (Hidden Markov Model; HMM) や DNN によってモデル化される．本論文では，振幅スペクトルの微細な特徴を捉えるため，DNN の枠組みを用いて振幅スペクトルを直接モデル化することを検討する．本モデル化手法では，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 を連結し，テキストから得られた言語特徴量から振幅スペクトルを直接合成する巨大な DNN を構築する．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．

1. はじめに

コンピュータに自然な音声を発話させる手法として，隠れマルコフモデル (Hidden Markov Model; HMM) に基づく音声合成システムが提案されている [1]．このシステムは非常に柔軟性が高く，HMM のパラメータを操作することで，合成音声の話者性，スタイル，性質等を容易に変更できることが知られている [2], [3], [4]．しかし，コーパスベースのもう 1 つの代表的な手法である素片選択等のサンプルベースの手法と比較して，統計モデルを用いたことで合成音声が機械的になる，過剰に平滑化されるといった問題がある．

近年，深層構造を持つ 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 [5]．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 Mahines (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 [6] 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 [7], [8]．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 [9]．

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．この問題に対して，DNN を用いたポストフィルタが提案されている [10]．この手法では DNN を用い自然音声と合成音声のスペクトル差異の条件付き確率をモデル化している．このポストフィルタを用いることで統計モデリングにより失われたスペクトルの微細な構造が再構築され，合成音声の品質が向上することが報告されている．この実験では，低次元のスペクトルパラメータを用い音響モデルは学習されており，一方で，DNN に基づくポストフィルタは STRAIGHT vocoder から抽出された振幅スペクトルを用いて学習される [11]．このことは，合成音声の品質低下は統計モデリングに伴う平均化だけでなく，低次元のスペクトルパラメータを用いていることに起因していることを示唆している．

本論文では，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 の学習は局所最適や vanishing gradient といった様々な問題が存在することが知られており [12]，また，従来広く用いられるメルケプストラムや LSP といったスペクトルパラメータと比較し

[1] 国立情報学研究所
National Institute of Informatics, Chiyoda, Tokyo 101–8430, Japan
[a)] takaki@nii.ac.jp
[b)] jyamagis@nii.ac.jp

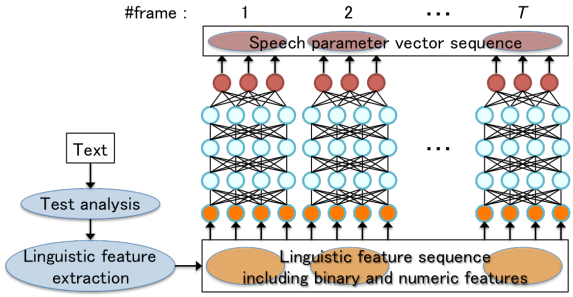


図 1 DNN に基づく音響モデルの枠組み

振幅スペクトルは非常に高次元であることから，DNN の学習の困難性が増すことが予想される．しかし，その一方で例えば音声認識分野において高次元特徴量である FFT スペクトラムを扱う DNN が，Pre-training と呼ばれる効率的な学習手法を用いることで適切に構築できることが報告されている [13]．そこで本論文では，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-traning 手法と見なすことができる．

2. DNN に基づく音響モデル

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 [5], [6], [7], [8]．本セクションでは代表的な DNN に基づく音響モデルの 1 つである [5] について簡潔にレビューする．

図 1 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．[5] 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 [14]．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. DNN に基づく音響特徴抽出

本セクションでは，Deep Auto-encoder (DAE) を用いた，効率的な低次元スペクトルパラメータ抽出法について述べる．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．

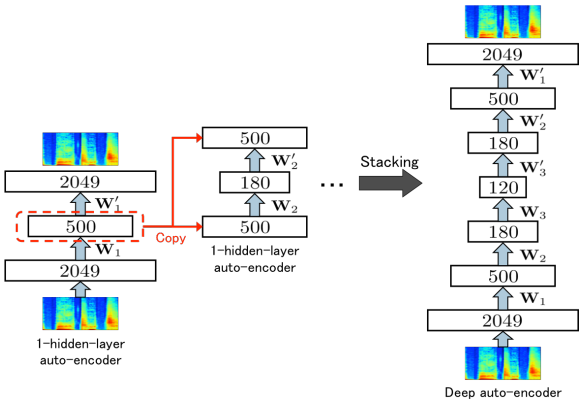


図 2 Deep Auto-encoder 構築のための Pre-training の手順

3.1 Auto-encoder

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．

$$\text{Encoder}: \quad \mathbf{y} = f_\theta(\mathbf{x}) = s(\mathbf{W}\mathbf{x} + \mathbf{b}), \tag{1}$$

$$\text{Decoder}: \quad \mathbf{z} = g_{\theta'}(\mathbf{y}) = t(\mathbf{W}'\mathbf{y} + \mathbf{b}'), \tag{2}$$

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 $m \times n$ の行列，$\mathbf{b}$ は m 次元のベクトル，$\mathbf{W}'$ は $n \times m$ 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．本論文では 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．

3.2 DAE の学習

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 2 に本論文で用いた 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

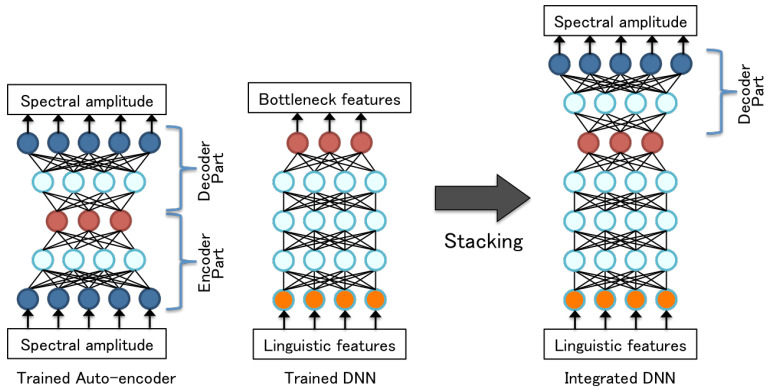
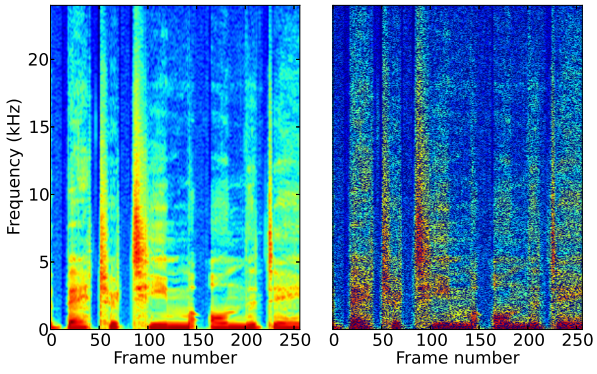


図 4 Deep Auto-encoder と DNN 音響モデルに基づく DNN スペクトルモデルの構築手順



(a) Original (b) Masked

図 3 元スペクトログラムとマスキングノイズを加えたスペクトログラム．右図中の黒点がマスクされている．

Pre-training 済み Auto-encoder の Encoder の出力 (図 2 の赤点線で囲まれたベクトル) 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 [12]．この問題を解決するため，本論文では $\mathbf{W}' = \mathbf{W}^T$ 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 [14]．

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 [15]．本論文では Pre-training 時の各層の入力の値をランダムに 0 にするマスキングノイズの付加を検討した [15]．図 3 に元スペクトログラムとマスキングノイズを加えたスペクトログラムを示す．

3.3 関連研究

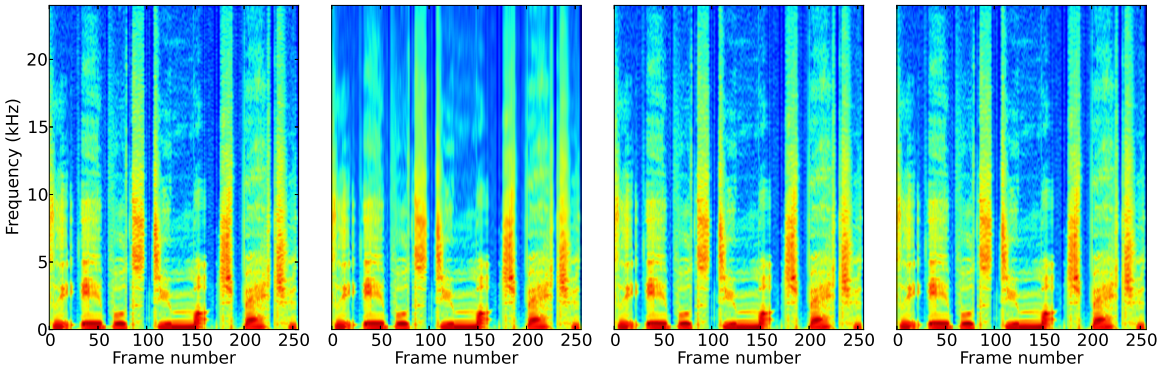
音声認識分野において DAE に基づく Bottleneck 特徴抽出が複数提案され用いられている [16], [17]．また，Deep Denoising Auto-encoder (DDAE) はノイズや反響に頑健な音声認識システム構築に用いられている [18], [19]．本論文と非常に関連強い研究には Deng らの DAE を用いたスペクトルのバイナリコーディング [20] や DDAE を用いた音声強調 [21] が挙げられる．また，Heteroscedatic Linear Discriminant Analysis (HLDA) [22] や Probabilistic Linear Discriminant Analysis (PLDA) [23], [24] とも関係が深い．

音声合成分野においては Auto-encoder を用いた低次元励震源パラメータやスペクトルパラメータ抽出が試みられている [9], [25], [26]．本論文は DNN 音響モデルと Auto-encoder の Decoder 部を積み重ね用いる点で，これらの研究とは異なる．

4. DNN に基づくスペクトルモデリング

本論文では，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．セクション 2 で述べた 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．そこで本論文では，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 4 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．



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

図 6 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

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．

5. 実験

5.1 実験条件

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 [11].

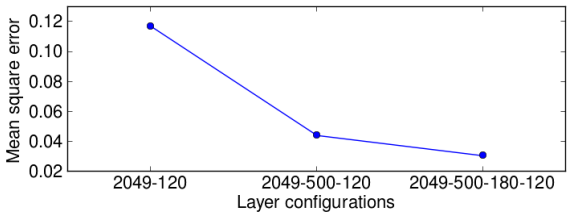


図 5 Deep Auto-encoder の構造の違いによる元対数振幅スペクトルと再構築された対数振幅スペクトルの平均二乗誤差．同次元の Bottleneck 特徴を扱うが隠れ層数が異なる．

主観評価実験にはプリファレンステストを用いた．被験者は 7 名であり，各被験者は被験者ごとにテスト文からランダムに選ばれた 30 文章を比較した．

5.2 分析再合成実験

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 5 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 5 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 6 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 6 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 7 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

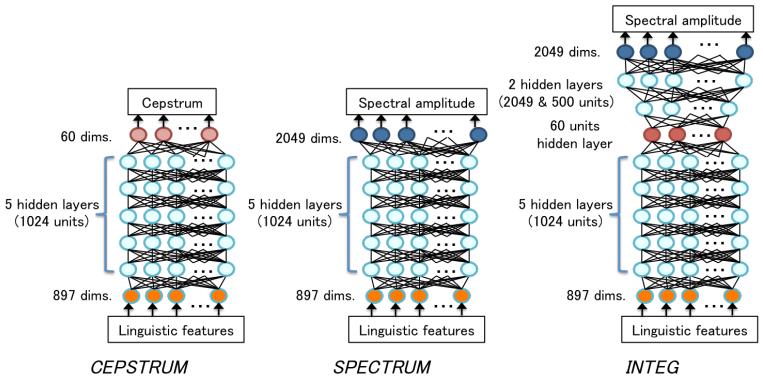


図 9 各手法で構築された DNN の構造

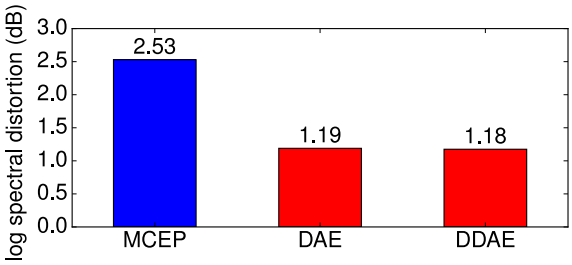


図 7 元振幅スペクトルと各手法により再構築された振幅スペクトルの対数振幅スペクトル距離 (dB)

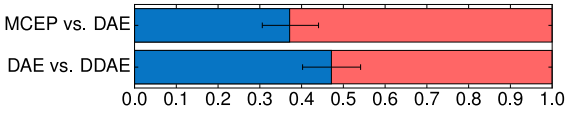


図 8 主観評価実験 (分析再合成)

ことがわかる．次に，図 8 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．

5.3 テキスト音声合成実験

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 Δ，Δ^2 は用いなかった．図 9 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．[5] にならい，DNN 音響モデルは

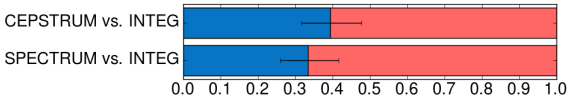


図 10 主観評価実験 (テキスト音声合成)

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 [1]．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 Δ，Δ^2 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 [27]．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 10 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. おわりに

本論文では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．Deep Auto-encoder を用いた分析再合成実験，および，テキスト音声合成実験で改善を確認することができた．今後の課題としては，Pre-training に用いられる Deep Auto-encoder と DNN 音響モデルの構造の影響調査や時間微分特徴量の検討が挙げられる．

謝辞 本研究は，NAVER Lab. の助成を受けた．

参考文献

[1] H. Zen, K. Tokuda, and A. W. Black: Statistical parametric speech synthesis, *Speech Communication*, Vol. 51, pp. 1039–1064 (2009).
[2] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura: Speaker interpolation in HMM-based speech synthesis system, *Proceedings of Eurospeech 1997*, pp. 2523–2526 (1997).
[3] R. Tsuzuki, H. Zen, K. Tokuda, T. Kitamura, M. Bulut, and S. Narayanan: Constructing emotional speech synthesizers with limited speech database, *Proceedings of ICSLP*, Vol. 2, pp. 1185–1188 (2004).
[4] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi: Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis, *IEICE Transactions on Information & Systems*, Vol. E88-D, No. 3, pp. 502–509 (2005).
[5] H. Zen, A. Senior, and M. Schuster: STATISTICAL PARAMETRIC SPEECH SYNTHESIS USING DEEP NEURAL NETWORKS, *Proceedings of ICASSP*, pp. 7962–7966 (2013).
[6] Z.-H. Ling, 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, pp. 2129–2139 (2013).
[7] Y. Fan, Y. Qian, F. Xie, and F. K. Soong: TTS Synthesis with Bidirectional LSTM Based Recurrent Neural Networks, *Proceedings of Interspeech*, pp. 1964–1968 (2014).
[8] R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory: Prosody Contour Prediction with Long Short-Term Memory, Bi-Directional, Deep Recurrent Neural Networks, *Proceedings of Interspeech*, pp. 2268–2272 (2014).
[9] R. Vishnubhotla, S. Fernandez and B. Ramabhadran: An autoencoder neural-network based low-dimensionality approach to excitation modeling for HMM-based text-to-speech, *Proceedings of ICASSP*, pp. 4614–4617 (2010).
[10] L.-H. Chen, T. Raitio, C. Valentini-Botinhao, J. Yamagishi, and Z.-H. Ling: DNN-based stochastic postfilter for HMM-based speech synthesis, *Proceedings of Interspeech*, pp. 1954–1958 (2014).
[11] H. Kawahara, 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, pp. 187–207 (1999).
[12] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber: Gradient Flow in Recurrent Nets: the Difficulty of Learning Long-Term Dependencies, *Citeseer* (2001).
[13] G.E. Hinton: Learning multiple layers of representation, *Trends in Cognitive Sciences*, Vol. 11, pp. 428–434 (2007).
[14] G. E. Hinton and R. Salakhutdinov: Reducing the dimensionality of data with neural networks, *Science 28*, Vol. 313, No. 5786, pp. 504–507 (2006).
[15] P. Vincent, H. Larochelle, Y. Bengio, and P. Manzagol: Extracting and composing robust features with denoising autoencoders, *ICML*, pp. 1096–1103 (2008).
[16] T. N. Sainath, B. Kingsbury, and B. Ramabhadran: AUTO-ENCODER BOTTLENECK FEATURES USING DEEP BELIEF NETWORKS, *Proceedings of ICASSP*, pp. 4153–4156 (2012).
[17] J. Gehring, Y. Miao, F. Metze, and A. Waibel: EXTRACTING DEEP BOTTLENECK FEATURES USING STACKED AUTO-ENCODERS, *Proceedings of ICASSP*, pp. 3377–3381 (2013).
[18] A. L. Maas, Q. V. Le, T. M. O ' Neil, O. Vinyals, P. Nguyen, and A. Ng Andrew：Recurrent Neural Networks for Noise Reduction in Robust ASR, *Proceedings of Interspeech*, pp. 22–25 (2012).
[19] X. Feng, Y. Zhang, and J. Glass: SPEECH FEATURE DENOISING AND DEREVERBERATION VIA DEEP AUTOENCODERS FOR NOISY REVERBERANT SPEECH RECOGNITION, *Proceedings of ICASSP*, pp. 1778–1782 (2014).
[20] L. Deng, M. Seltzer1, D. Yu, A. Acero, A. Mohamed, and G. Hinton: Binary Coding of Speech Spectrograms Using a Deep Auto-encoder, *Proceedings of Interspeech*, pp. 1692–1695 (2010).
[21] X. Lu, Y. Tsao, S. Matsuda1, and C. Hori: Speech Enhancement Based on Deep Denoising Autoencoder, *Proceedings of Interspeech*, pp. 436–440 (2013).
[22] M. J. F. Gales: Maximum likelihood multiple subspace projections for hidden Markov models, *Speech and Audio Processing, IEEE Transactions on*, Vol. 10, pp. 37–47 (2002).
[23] S. J. D. Prince and J. H. Elder: Probabilistic Linear Discriminant Analysis for Inferences About Identity, *ICCV*, pp. 1–8 (2007).
[24] L. Lu and S. Renals: Probabilistic Linear Discriminant Analysis for Acoustic Modelling, *Signal Processing Letters, IEEE*, pp. 702–706 (2014).
[25] T. Raitio, A. Suni, L. Juvela, M. Vainio, and P. Alku: Deep neural network based trainable voice source model for synthesis of speech with varying vocal effort, *Proceedings of Interspeech*, pp. 1969–1973 (2014).
[26] P. K. Muthukumar and Black. A.: A Deep Learning Approach to Data-driven Parameterizations for Statistical Parametric Speech Synthesis, *CoRR*, Vol. abs/1409.8558 (2014).
[27] K. Richmond, R. Clark, and S. Fitt: On generating Combilex pronunciations via morphological analysis, *Proceedings of Interspeech*, pp. 1974–1977 (2010).