Vol.2019-MUS-123 No.60
Vol.2019-SLP-127 No.60
2019/6/23

変調スペクトルに着眼した DNN音声合成の学習過程に関する分析

川上 大智[1,a)]　齋藤 大輔[1,b)]　峯松 信明[1,c)]

概要：DNN 音声合成等の統計的音声合成では，合成音声のパラメータ系列が自然音声のパラメータ系列と比較して過度に平滑化される傾向にある．そこで，パラメータ系列の変調スペクトルに着眼しながら DNN の学習過程を分析した．その結果，従来手法では DNN の過学習とされる学習回数で，変調スペクトルが自然な音声に近づき，より自然な音声が合成できるという知見が得られた．

1. はじめに

任意のテキストに対応する音声を人工的に生成する技術のことをテキスト音声合成（text-to-speech synthesis; TTS synthesis）又は単に音声合成という．音声合成の代表的な手法には，言語特徴量と音響特徴量の対応付けを，データをもとに学習する統計的音声合成や，大規模な音声データベースから音声素片を接続する素片選択型音声合成等がある [1,2]．

2010 年代に入り，ディープニューラルネットワーク（deep neural network; DNN）等の深層学習を利用した手法が画像処理をはじめとして様々な分野で高い性能を示している．音声合成の分野においても DNN を利用した手法の高い性能が認められ，深層学習に基づく統計的音声合成の利用が盛んになっている [3]．特に音響モデルに DNN を用いた手法は，隠れマルコフモデル（hidden Markov model; HMM）に基づく音声合成と比較して高い性能を示している [3,4]．しかし，HMM や DNN のような統計モデルを用いると，人間が発話する自然な音声に含まれる微細な振動成分が損なわれ，合成音声が過剰に平滑化されてしまうという問題がある．

そこで，合成音声が平滑化されているかどうかの基準として，パラメータ系列の変調スペクトル（modulation spectrum; MS）に着眼する [5]．統計的な音声合成手法では，自然音声のパラメータ系列の変調スペクトルと比較して，合成音声のパラメータ系列の変調スペクトルが高域で減衰してしまうことが示されている [5]．深層学習を利用した音声合成についても，DNN によって出力されるパラメータ系列の変調スペクトルが減衰することになるが，これは中間層における出力パラメータ系列の時系列挙動と深く関連していると考えられる．

しかし，DNN 音声合成に用いられるモデルアーキテクチャの内部で，入力の言語特徴量がどのように中間層を経て出力の音響特徴量まで伝播していくのかということに関しては，未だ十分な議論がなされていない．本研究では中間層の変調スペクトルに着目し，DNN の各中間層において入力パラメータ系列がどのように伝播し出力されていくのかを分析するとともに，変調スペクトルに着眼して DNN の学習を行うことで，より良い音声合成の実現可能性を探る．

[1] 東京大学
The University of Tokyo
[a)] kawakami@gavo.t.u-tokyo.ac.jp
[b)] dsk_saito@gavo.t.u-tokyo.ac.jp
[c)] mine@gavo.t.u-tokyo.ac.jp

2. 変調スペクトル

変調スペクトルの理論的な導出方法とその実装，及び変調スペクトルに関する先行研究について述べる．

2.1 理論

まず，変調スペクトルの理論的な導出方法を述べる．フレーム数 T のパラメータ系列 $\boldsymbol{f}$，及びそれらの離散フーリエ変換 $\boldsymbol{F}$ を，

$$\boldsymbol{f} = [f_0, \cdots, f_t, \cdots, f_{T-1}] \tag{1}$$

$$\boldsymbol{F} = [F_0, \cdots, F_k, \cdots, F_{T-1}] \tag{2}$$

と定める．このとき，

$$F_k = \sum_{t=0}^{T-1} f_t e^{-i\frac{2\pi kt}{T}} \quad (0 \leq k \leq T-1) \tag{3}$$

である．パラメータ系列の振幅スペクトルを $\boldsymbol{A}$ とすると，

$$\boldsymbol{A} = [A_0, \cdots, A_k, \cdots, A_{T-1}] \tag{4}$$

$$A_k = |F_k| = \sqrt{\mathbf{Re}\left(F_k\right)^2 + \mathbf{Im}\left(F_k\right)^2} \tag{5}$$

また，変調スペクトル $\boldsymbol{s}$ を次のように定義する．

$$\boldsymbol{s} = [s_0, \cdots, s_k, \cdots, s_{\frac{T}{2}-1}] \tag{6}$$

$$s_k = \log_{10}\left(A_k\right)^2 \quad \left(0 \leq k \leq \frac{T}{2} - 1\right) \tag{7}$$

変調スペクトルでは，元のパラメータ系列に含まれる位相成分を考慮せず振幅成分のみを考慮しているが，これは元のパラメータ系列にどの程度の微細な時系列変動成分を成分として持っているかということが主な着眼点ゆえである．

2.2 実装

次に，パラメータ系列から変調スペクトルを求める実装について述べる．

まず，パラメータ系列から無音区間を除いた値の平均を 0 にする．次に，FFT を行うため，フレーム数が 2 のべき乗となるようにする．本研究ではフレーム数 N を 4096 とした．このフレーム数 N の値は，理論的にはスペクトルの概形にほとんど影響しない（後述）．また，本研究で設定した N は，用いた発話データセットのパラメータ系列の長さよりも十分大きい．さらに，これらの分散を 1 にする．

ここで，パラメータ系列の平均，分散に関する正規化操作について，妥当性を示す．まず，Parseval の等式より，

$$\sum_{t=0}^{N-1} |f_t|^2 = \frac{1}{N} \sum_{k=0}^{N-1} |F_k|^2 \tag{8}$$

が成り立つ．両辺をフレーム数 N で割るとパラメータ系列の平均が 0，分散が 1 であるから

$$\frac{1}{N} \sum_{t=0}^{N-1} |f_t|^2 = 1 = \frac{1}{N^2} \sum_{k=0}^{N-1} |F_k|^2 \tag{9}$$

したがって，フレーム数 N は 4096 と，一定の値であるから，

$$\sum_{k=0}^{N-1} |F_k|^2 = \sum_{k=0}^{N-1} A_k^2 = N^2 = const. \tag{10}$$

となる．式 (10) はパワースペクトルの総和が一定になることを示している．

パワースペクトルの総和は一定となるため，ある変調周波数でパワースペクトルを大きくすると，別の変調周波数でパワースペクトルを小さくしなければならない．したがって，2 つのパラメータ系列に関して，パワースペクトルを比較した場合，一方の高域が他方と比較して大きくなっているような場合，その分だけ低域は小さくなる．

線形パワースペクトル領域では変調周波数ごとの特徴が十分捉えられないため，対数パワー領域での変調スペクトルを考える．パラメータ系列の平均は 0 であるから，

$$F_0 = \sum_{t=0}^{T-1} f_t = 0 \tag{11}$$

となり，対数を定義することができないため，変調周波数 0 の成分は考えない．

2.3 HMM 音声合成での変調スペクトル

HMM による出力パラメータ系列の変調スペクトルは自然音声の変調スペクトルと比較して高域で大きく減衰していることが示されている [5]．また，出力パラメータ系列の変調スペクトルを，自然音声の変調スペクトルに高域で近付けるため，ポストフィルタを用いることで音声合成の性能が向上することが示されている [5]．ポストフィルタは学習データを用いて事前に設計される．パラメータ系列 $\boldsymbol{x}$ に対する変調スペクトルを，

$$\boldsymbol{s}(\boldsymbol{x}) = [\boldsymbol{s}(1)^\top, \cdots, \boldsymbol{s}(D)^\top]^\top \tag{12}$$

とすると，ポストフィルタの学習では自然音声のパラメータ系列から次の確率密度関数を学習する．

$$P\left(\boldsymbol{s}\left(\boldsymbol{x}\right)|\lambda_{\mathbf{s}}\right) = \mathcal{N}\left(\boldsymbol{s}\left(\boldsymbol{x}\right); \mu^{(N)}, \sum\nolimits^{(N)}\right) \tag{13}$$

ただし，$\mathcal{N}\left(\cdot; \mu^{(N)}, \sum^{(N)}\right)$ は平均 $\mu^{(N)} = [\mu_{1,0}^{(N)}, \cdots, \mu_{D,T/2-1}^{(N)}]^\top$ と対角共分散行列 $\sum^{(N)} = diag\left[\left(\sigma_{1,0}^{(N)}\right)^2, \cdots, \left(\sigma_{D,T/2-1}^{(N)}\right)^2\right]$ の正規分布，$\mu_{d,m}^{(N)}$ と $\left(\sigma_{d,m}^{(N)}\right)$ は $s_d\left(m\right)$ の平均と分散，$\lambda_{\mathbf{s}}$ は変調スペクトルの確率密度関数パラメータセットを表す．同様に，HMM 音声合成で生成されたパラメータ系列から正規分布 $\mathcal{N}\left(\cdot; \mu^{(G)}, \sum^{(G)}\right)$ を学習する．

生成部では，生成されたパラメータ系列 c の変調スペクトルに対して次のポストフィルタを適用する．

$$\begin{aligned} s_d'\left(m\right) &= \left(1-k\right) s_d\left(m\right) \\ &\quad + k\left[\frac{\sigma_{d,m}^{(N)}}{\sigma_{d,m}^{(G)}}\left(s_d\left(m\right) - \mu_{d,m}^{(G)}\right) + \mu_{d,m}^{(N)}\right] \end{aligned} \tag{14}$$

ここで，$\mu_{d,m}^{(G)}$，$\sigma_{d,m}^{(G)}$ は $s_d\left(m\right)$ の平均，標準偏差であり，k はポストフィルタの強度係数である．このようにポストフィルタを用いると平滑された変調スペクトルが補償され，より自然な音声が合成できることが示されている．

3. 中間特徴量の分析

本節では，リカレント構造を持たない単純な 6 層 feed-forward 型 DNN を用いた中間特徴量の分析について述べる．

3.1 構成する DNN アーキテクチャ

使用したデータは，CMU_ARCTIC [6] の女性話者 slt による英語音声 1132 発話である．データのサンプリング周波数は 16kHz である．1132 発話のうち a0001 から b0407 の 1000 発話を学習データ，b0408 から b0473 の 66 発話をバリデーションデータ，b0474 から b0539 の 66 発話をテストデータとした．音声の分析合成には WORLD [7] を用いた．分析のフレームシフトは 5ms とした．DNN の入力には，Festival [8] をフロントエンドとする HTS 形式 [9] のコンテキストラベルをもとに，当該フレームの音素情報等に関する 416 次元，及び継続長に関する 9 次元の計 425 次元を言語特徴量として用いる．DNN の出力には，60 次元のメルケプストラム係数，対数 F_0，及び 1 次元の非周期性成分と，これらの一次差分 Δ，二次差分 ΔΔ，さらに有声/無声フラグ（1 次元）の計 187 次元を用いる．合成を行う際には MLPG [10] により，一次差分 Δ，二次差分 ΔΔ を用いて静的成分を推定する．中間層の層数は 6 層とし，各層ノード数は 1024 とした．各中間層で出力される特徴量には，バッチ正規化を適用した後，活性化関数 tanh を適用した．最終層の活性化関数は線形関数である．DNN の最適化手法は adam [11] を用いた．adam のハイパーパラメータについては論文内の推奨値を採用し，学習率を 0.001，$\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e-8$ とした．損失関数は二乗誤差を用いた．バッチサイズは 64 であり，バッチ正規化における momentum は 0.99，epsilon は 0.001 とした．このモデルの学習エポック数は，validation loss の減少が安定した 40 エポックとした．

3.2 中間層の順方向伝播

ある中間層での入力ベクトルを $\boldsymbol{x}$，重み行列を $\boldsymbol{W}$ とすると，バッチ正規化，活性化をかける前の中間層の出力 $\boldsymbol{y}$ は単純な線形変換

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x} \tag{15}$$

として表される．さらに，バッチ正規化をかけた後の出力を $\boldsymbol{y}_n$ とし，学習された各ノードの平均，標準偏差のベクトルを $\boldsymbol{\mu}, \boldsymbol{\sigma}$ とすると，

$$\boldsymbol{y}_n = \frac{\boldsymbol{y} - \boldsymbol{\mu}}{\boldsymbol{\sigma}} = \frac{\boldsymbol{W}\boldsymbol{x} - \boldsymbol{\mu}}{\boldsymbol{\sigma}} \tag{16}$$

となる．これは平均を 0，分散を 1 に近づける線形変換なので，変調スペクトルはバッチ正規化前後で変化しない．また，活性化関数を適用した後の出力を $\boldsymbol{y}_a$ とすると，

$$\boldsymbol{y}_a = \tanh \boldsymbol{y}_n \tag{17}$$

となるため，活性化前後で変調スペクトルが変化する．

3.3 中間層での出力活性化前後の変調スペクトル

図 1 は，中間層第 1 層目のあるノードについて，活性化関数である tanh の適用前後のパラメータ系列を示している．活性化関数を適用する前は絶対値が 1 より大きくなっている部分が，活性化関数を適用することによって絶対値 1 付近の値に集中している．これにより，活性化関数を適用する前のパラメータ系列の情報が失われてしまい，変調スペクトルは高域で減少すると考えられる．このことを確認するために，このノードに関して，活性化関数前後の変調スペクトルをテストデータ 66 発話分求め，平均したものを図 2 に示す．

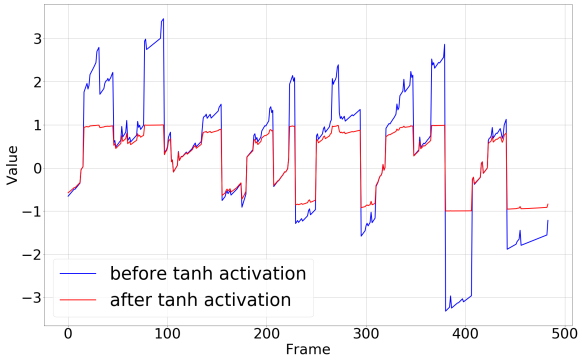

図 1 中間層第 1 層直後の活性化関数前後のパラメータ系列

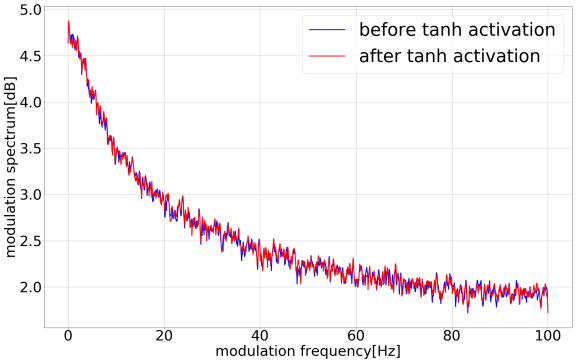

図 2 中間層第 1 層直後の活性化関数前後のパラメータ系列に関する変調スペクトル

図 2 からは，変調スペクトルがほとんど変化していないことが見て取れる．第 6 層までについても同様に活性化関数前後で変調スペクトルがほとんど変化しないことが確認できた．これは，活性化関数を適用する前に，絶対値が 1 より小さくなっている部分が，活性化関数を適用することによってほとんど変化しないことによる影響だと考えられる．パラメータ系列に活性化関数を適用すると，パラメータ系列の分散は小さくなる．変調スペクトルを求める際にはパラメータ系列の分散を 1 に正規化する線形変換を行うので，活性化関数適用後のパラメータ系列において，活性化関数適用前に絶対値が 1 より小さくなっている部分は，

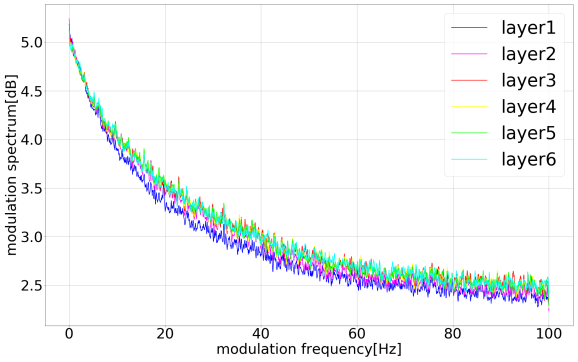

図 3 各層の変調スペクトルについて，変調周波数ごとの最大値

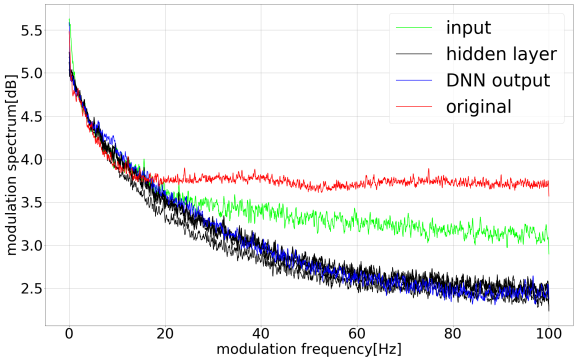

図 4 入力層，出力層，自然音声も含めた変調周波数ごとの最大値

活性化関数適用前と比較して変調スペクトルに与える影響が大きくなる．これにより，活性化関数の適用前後で変調スペクトルがほとんど変化しなかったと考えられる．

3.4 中間層の変調スペクトル

まず，中間特徴量のみを分析した．各層各ノードごとにパラメータ系列の変調スペクトルの 66 発話平均をとり，各層ごとに活性化関数を適用した後の 1024 ノードから得られる変調スペクトルについて変調周波数ごとの最大値をとったものを図 3 に示す．

第一層では他の層と比較して変調スペクトルが高域で小さくなっている．第二層から第六層に関しては，ほとんど変化が見られなかった．

また，中間層に加え，入力層，出力層，自然音声について，同様に変調スペクトルの変調周波数ごとの最大値をとったものを図 4 に示す．

DNN の出力から得られる変調スペクトルは，中間層から得られる変調スペクトルとほぼ同程度の大きさであった．入力層，自然音声から得られる変調スペクトルと比較すると，中間層，及び DNN 出力の変調スペクトルは高域で大きく減衰している．

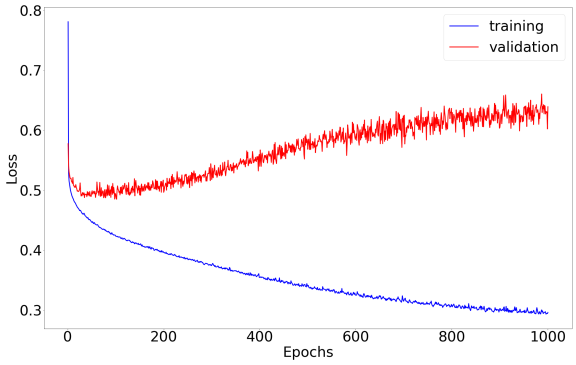

図 5 6 層 feedforward 型 DNN を過学習させたときの loss

3.5 中間特徴量に関する考察

本節では単純な 6 層 feedforward 型 DNN を扱った．DNN の入力である言語特徴量には，音素情報のような 2 値情報であるステップ入力となるものと継続長情報の連続値である三角波入力となるものがあり，後者の変調スペクトルが高域で大きな値をとる．しかし，DNN の中間層では三角波の持つ高域の成分が，ステップ入力による成分と足し合わされることにより，第一層から失われてしまうことになる．出力層までに適切に三角波の持つ高域の成分を伝播させるためには，伝播の過程で三角波の情報が損なわれないようにする必要があると考えられる．

4. 過学習モデル

DNN の出力パラメータ系列が自然音声のパラメータ系列と比較して過度に平滑化されるのは，DNN の汎化作用によるものであると考えられる．そこで，本研究では DNN の過学習について考える．DNN を過学習させた場合，出力パラメータ系列は学習データに特化していく．このとき，出力パラメータ系列の変調スペクトルはどのように変化するのか分析する．

4.1 6 層 feedforward 型 DNN の過学習

まず前節と同様の構成の単純な 6 層 feedforward 型 DNN を 100 エポックずつ学習させていき，得られる出力パラメータ系列の変調スペクトルについて分析した．1000 エポック学習を回して得られた loss を図 5 に示す．validation loss は徐々に増加し，モデルが過学習をしている．エポック数を変化させて学習した各モデルから得られる，学習データ，テストデータそれぞれの出力の変調スペクトルについて，変調周波数ごとの最大値をとり，自然音声のものとユークリッド距離をとって比較したものを図 6 に示す．

学習が進むにつれ，出力の変調スペクトルは自然音声に近づいている．300 エポックから 400 エポック程度学習を回すと，ユークリッド距離はある程度まで小さくなり，これ以降ほとんど減少しなくなる．

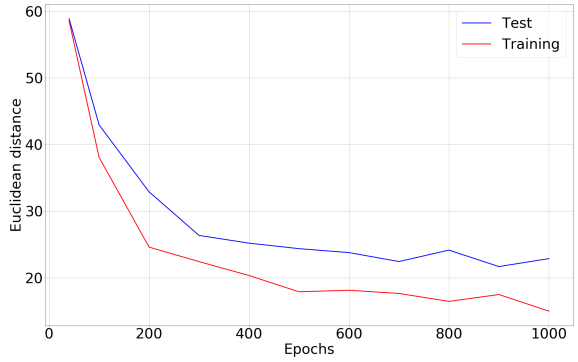

図 6 6 層 feedforward 型 DNN を過学習させたときの出力の変調スペクトルと自然音声の変調スペクトルのユークリッド距離

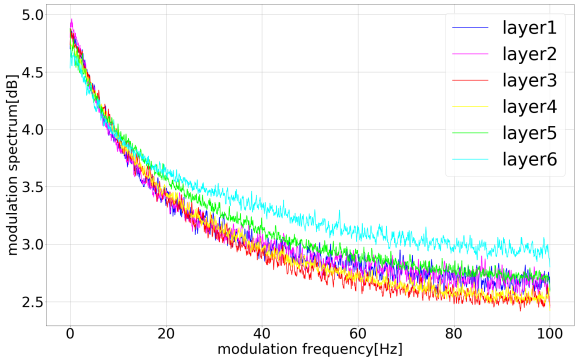

図 8 1000 エポック学習を行った DNN 各層の変調スペクトルについて，変調周波数ごとの最大値

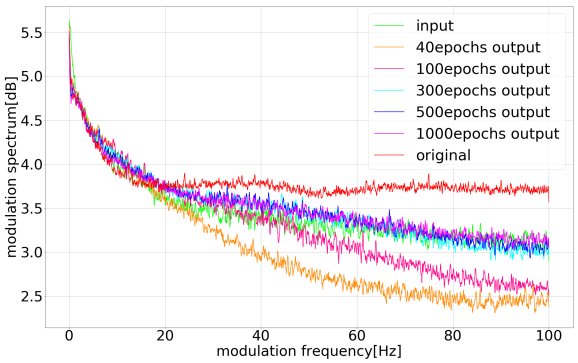

図 7 6 層 feedforward 型 DNN を過学習させた場合の変調スペクトル

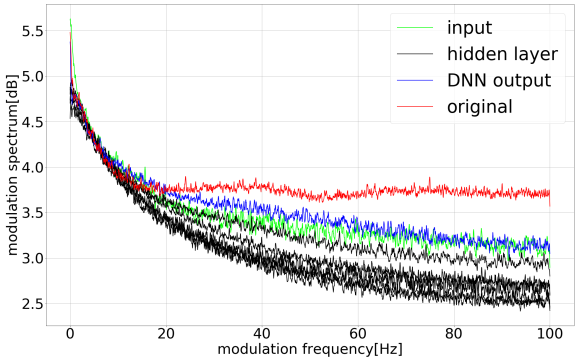

図 9 1000 エポック学習を行った DNN について入力層，出力層，自然音声も含めた変調周波数ごとの最大値

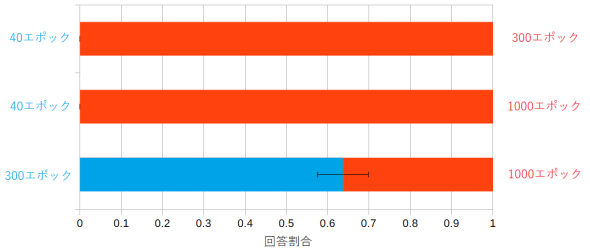

図 10 40 エポック，300 エポック，1000 エポック学習した 3 種類のモデルによる合成音声に関する主観実験結果

変調スペクトルがどのように自然音声のものに近づいていくのかを観察するため， 40 エポック， 100 エポック，300 エポック， 500 エポック， 1000 エポック学習したモデルからそれぞれ得られる出力パラメータ系列の変調スペクトルについて図 7 に示す．

学習エポック数を増やすごとに，変調スペクトルは高域で増加している．300 エポック程度学習したあたりで，出力パラメータ系列から得られる変調スペクトルは高域でほとんど増加が見られなくなる．また， 1000 エポックと十分に過学習をしても，高域では入力パラメータ系列から得られる変調スペクトルより小さい．

過学習を行った DNN から出力される中間特徴量の変調スペクトルについて述べる．図 3 と同様に，1000 エポック学習したモデルから得られる中間特徴量についての変調スペクトルを図 8 に示した．第一層から第三層まで変調スペクトルが高域で減衰し，その後第六層までは変調スペクトルが高域で増加している．また，図 4 と同様に，入力層，出力層，自然音声も含めて比較したものを図 9 に示す．中間特徴量の変調スペクトルはいずれも入力，出力の変調スペクトルと比較して，高域で減衰している．

4.2 主観実験

過学習したモデルによる出力系列から得られる音声が，より自然になっているかどうかを分析するため，主観実験を行った．主観実験では 3 種類のモデルを用いてそれぞれ合成した音声を被験者に提示した．3 種類のモデルについて，学習させたエポック数はそれぞれ，従来通り 2 乗誤差の validation loss が最小になっている 40 エポック，自然音声の変調スペクトルと合成音声の変調スペクトルのユークリッド距離が大きく減少している 300 エポック，ユークリッド距離はさらに減少しているが，さらに過学習している 1000 エポックである．被験者は男性 7 名，女性 5 名の計

12 名であり，全員が大学生である．テストデータ 66 発話を 22 発話ずつ 3 セットに区切り，それぞれのセットで 3 種類の音声の中から 2 種類の音声を選んで分けた．被験者はランダムな順番で提示される 2 種類の音声のうち，どちらの音声がより自然に聴こえるかを選ぶ一対比較を行った．

主観実験の結果，40 エポック学習したモデルによる合成音声と，過学習を行った 2 つのモデルによる合成音声の比較では，全ての被験者が全ての音声の組に対して，過学習を行ったモデルによる合成音声を，より自然に聴こえるとして選択している．また，図 10 のように 300 エポック学習したモデルと 1000 エポック学習したモデルによる合成音声の比較では，前者がより自然に聴こえるという結果が得られた．

4.3 過学習に関する考察

過学習を行うと，出力の変調スペクトルは，入力の変調スペクトル程度まで高域で増加する．しかし，図 8，9 から確認できるとおり，これは入力の変調スペクトルがそのまま出力まで伝播しているわけではない．入力では高域で大きい変調スペクトルが，中間層を経て，一度高域で減衰してから出力にかけて増加している．これは，中間層での特徴量が一度平滑化され，出力に近くなるにつれて微細な振動成分を獲得していくということを意味している．最も変調スペクトルが高域で減衰しているのは第四層であり，40 エポック学習したモデルの中間特徴量の変調スペクトルと比較すると，高域では同程度の値となっている．

単純な 6 層 feedforward 型 DNN を過学習したモデルによる合成音声の主観評価の結果，従来のように 2 乗誤差を最小化するように学習したモデルと比較して，過学習したモデルによる合成音声が，より自然な音声に聴こえるという結果が得られた．これは，変調スペクトルを高域で増加させることが，自然な音声の合成に重要であることを示している．また，自然音声の変調スペクトルと，合成音声の変調スペクトルのユークリッド距離がほとんど減少しなくなった後は，さらに学習を行うと，音声の自然さが低下することが確認できた．validation loss が大きくなると，テストデータに対するパラメータ推定の精度が低下するため，音声の自然さが低下したと考えられる．

5. まとめ

本研究では，DNN 音声合成の出力パラメータ系列から得られる変調スペクトルの高域を自然音声の変調スペクトルに近づけることで，より自然な音声が得られるという視点で DNN の各層における特徴量系列の分析，検討を行った．その結果，合成音声の変調スペクトルを変調スペクトルを自然音声のものに近づけるようにモデルを過学習することで，より自然な音声が合成できることが示された．

本研究では DNN の入力として，自然音声の変調スペクトルと比較して，変調スペクトルが高域で小さい言語特徴量の情報を用いた．また，言語特徴量に含まれる継続長情報である三角波入力の変調スペクトルは，音素情報等であるステップ入力の変調スペクトルと比較して高域で大きい．今回構成した DNN を過学習すると，得られる出力パラメータ系列の変調スペクトルは，入力パラメータ系列から得られる変調スペクトルと比較して，高域で同程度の値をとる．そこで，入力として，三角波入力から得られる変調スペクトルよりも高域でさらに大きい変調スペクトルが得られる入力パラメータ系列を用いることで，出力パラメータ系列の変調スペクトルもさらに高域で大きくなる可能性がある．今回用いた三角波入力を非線形変換してこのような入力パラメータ系列を新たに作成する等，入力パラメータ系列の作成手法を今後検討できるといえる．

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] A. Hunt and A. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 01, pp. 373-376(1996).
[3] H. Zen, A. Senior and M. Schuster, ”Statistical parametric speech synthesis using deep neural networks,” IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, pp. 7962-7966 (2013).
[4] Z.-H. Ling, S.-Y. Kang, H. Zen, A. Senior, M. Schuster, X.-J. Qian, H. Meng and L. Deng, “Deep learning for acoustic modeling in parametric speech generation,” IEEE Signal Process. Mag., 32, pp. 35–52 (2015).
[5] 高道 慎之介, 戸田 智基, G. Neubig, S. Sakti, 中村 哲, “変調スペクトルを考慮した HMM 音声合成,” 日本音響学会講演論文集, pp.307-308 (2013).
[6] J. Kominek and A. W Black, “CMU ARCTIC databases for speech synthesis,” Technical Report CMU-LTI-03-177, Language Technologies Institute, School of Computer Science, Carnegie Mellon University (2013).
[7] M. Morise, F. Yokomori, and K. Ozawa, “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications,” IEICE transactions on information and systems, vol. E99-D, no. 7, pp. 1877-1884 (2016).
[8] A. Black, P. Taylor, and R. Caley, “The Festival speech synthesis system,” http://festvox.org/festival (1999).
[9] J. Yamagishi, H.Zen, Y. J. Wu, T. Toda, and K. Tokuda, “The HTS-2008 System: Yet Another Evaluation of the Speaker-Adaptive HMM-based Speech Synthesis System in The 2008 Blizzard Challenge,” Proc. BLZ4-2008, Sep (2008).
[10] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” IEEE international conference on acoustics, speech and signal processing, vol. 3, pp. 1315-1318 (2000).
[11] D.P. Kingma, and J. Ba, “Adam: A Method for Stochastic Optimization,” arXiv:1412.6980 (2014).