Vol.2019-SLP-128 No.1
2019/7/19

非負値行列因子分解のアクティベーションに着目したDNN音声合成

後藤 駿介[1,a)]　齋藤 大輔[1,b)]　峯松 信明[1,c)]

概要：本稿では，非負値行列因子分解のアクティベーションを音響特徴量としたDNN統計的パラメトリック音声合成を提案する．それに加えて，提案手法において帯域拡張の実現が可能であることを示す．実験的評価では，特にサンプリング周波数48 kHzの合成音声において，他手法に比べて自然な音声が生成が可能であり，16 kHzから48 kHzへの帯域拡張においては48 kHzの合成音声と比較し同程度に自然の音声が生成可能であることを示す．

1. はじめに

テキスト音声合成はテキストからそれに対応した音声を合成する取り組みである．従来ではテキスト音声合成は言語特徴量からボコーダパラメータを推定し，ボコーダにより波形生成する統計的パラメトリック音声合成が一般的であった [1], [2] が，近年のDeep Neural Network（DNN）の発展によりボコーダを用いずに，テキストや言語特徴量から直接振幅スペクトルや波形を推定することが可能になってきた [3], [4], [5], [6]．これらの手法の中には肉声に近い自然な音声を生成できるものもあるが，多くのデータを要する場合が多く学習も難しい問題がある．

一方ボコーダを用いた既存の統計的パラメトリック音声合成では，人間の発声をモデル化したソースフィルタモデルに基づき，励振源と声道フィルタに対応する音響特徴量を用いて音声を生成するが，生成される音声の自然性はあまり高くない．しかし，少量のデータでも動作しやすく，また人間の発声に基づいて得られた音響特徴量を用いるため直感的に発話を制御しやすい利点もある．統計的パラメトリック音声合成における生成音声の自然性向上は未だなお重要な課題である．

統計的パラメトリック音声合成の品質を向上させる取り組みとして，これまで言語特徴量から音響特徴量を推定する音響モデルに単純なfeed-forward型DNNではなく時間構造を考慮したRecurrent Neural Network（RNN）やLong-Short Term Memory（LSTM）を用いることでより正確に特徴量を推測しようとする手法 [7] や，励振源をより正確にモデル化しようとする手法 [8], [9]，またスペクトル包絡を直接推定しようとする手法 [10], [11] も提案されている．しかし，声道フィルタの特性を表す特徴量としては未だにメルケプストラムが使われることが一般的であり，より適切な特徴量について検討すべきである．

メル尺度によって周波数伸縮された対数スペクトルの逆フーリエ変換であるメルケプストラムは，スペクトル包絡を表す特徴量として用いられる．メルケプストラムは，スペクトル包絡をsin（あるいはcos）カーブの重ね合わせで表現でき，少数のパラメータでスペクトル包絡をモデル化できるが，微細なスペクトル構造を失ってしまうという欠点がある．一方，中間的な特徴量を用いずにスペクトル包絡を直接推定しようとする手法もあるが，次元数が高い特徴量を精緻に推定することは難しい．

これらの問題を解決する為に，非負値行列因子分解（Non-negative Matrix Factorization; NMF）[12] を用いたスペクトル包絡のモデリングに着目する．NMFは，非負値の行列を2つの非負値行列に分解する計算手法であり，音声信号処理の分野ではスペクトルを基底スペクトルの重ね合わせで表現できる．この重み付けのことをアクティベーションと呼び，基底の非負制約によりアクティベーションはスパースな傾向を持つことが知られている．これにより，得られるスペクトル基底は微細構造を保持したものになり，それらの足し合わせでスペクトルが表現される．本稿では，NMFによって得られるアクティベーションを音響特徴量とした音声合成を提案する．スパースであるアクティベーションの推定は，適切な基底を選択するクラス認識問題のように解釈することがふさわしいと考えられるため，

[1] 東京大学 大学院工学系研究科電気系工学専攻,
7-3-1 Hongo,Bunkyo-ku, Tokyo 113-8656, Japan.
[a)] goto@gavo.t.u-tokyo.ac.jp
[b)] dsk_saito@gavo.t.u-tokyo.ac.jp
[c)] mine@gavo.t.u-tokyo.ac.jp

アクティベーションをカテゴリカル分布とみなして，カルバック・ライブラー情報量（Kullback-Leibler Divergence; KLD）から導かれる誤差関数を用いた．

また，NMF には様々な応用が提案されており，話者ごとのスペクトル基底を用いた声質変換 [13] や，サンプリング周波数ごとの基底を用いた帯域拡張 [14]，また各楽器のスペクトル基底を学習することで行う音源分離 [15] などがその例として挙げられる．提案手法は言語特徴量とアクティベーションの関係を学習するものであり，テキスト音声合成にこれらの NMF の応用例を適用することが期待できる．本稿ではその中でも帯域拡張に注目した．本手法では，少数の広帯域の音声を用意すれば，狭帯域の音声によって作られた音響モデルを用いて広帯域の音声を生成できることを示す．

2. テキスト音声合成におけるスペクトルモデリング

2.1 中間特徴量としてのメルケプストラム

統計的パラメトリック音声合成において，DNN は言語特徴量と音響特徴量の関係を学習する音響モデルとして用いられる [2]．言語特徴量はテキスト解析器によって得られる二値あるいは連続値のコンテキストに関する質問の回答を連結したものであり，音響特徴量は声帯振動と声道フィルタに対応する特徴量である．声帯振動を表す特徴量としては基本周波数が用いられ，声道フィルタを表す特徴量としてはスペクトル包絡を表現するメルケプストラムが用いられることが多い．メルケプストラムはスペクトル包絡を表現するための低次元の中間的な特徴量である．

一般的に DNN による音響モデルを学習するために平均二乗誤差（Mean Squared Error; MSE）が最小化するべき指標として採用されることが多い．MSE は式（1）のように表される．ここで，T はフレーム数，D は特徴量の次元を表している．

$$\mathcal{L}_{\mathrm{MSE}}\left(\mathbf{y} \mid \hat{\mathbf{y}}\right) = \frac{1}{2}\sum_{t=1}^{T}\sum_{d=1}^{D}\left(y_{t,d} - \hat{y}_{t,d}\right)^2 \qquad (1)$$

DNN 音響モデルの学習においては，$\mathbf{y}$ は自然音声から得られる音響特徴量系列であり，$\hat{\mathbf{y}}$ は DNN によって推測された音響特徴量系列である．

2.2 中間特徴量を介さないスペクトルモデリング

メルケプストラムのような中間的な低次元の特徴量を介すことで言語特徴量からの推測は行いやすいが，このような圧縮された表現では精緻にスペクトル包絡を再現することはできない．これを避ける為にスペクトル包絡を言語特徴量から直接推定する手法が提案されている [10], [11]．スペクトル包絡を表す係数の次元数は窓長に依存し，普通は 513 や 1025 といった大きな次元を持つ為，適切に高次元の特徴量を扱う工夫が必要である．制限付きボルツマンマシン（Restricted Boltzmann Machines; RBM）を用いた手法 [11] では，隠れマルコフモデル（Hidden Markov Model; HMM）において高次元の特徴量の出力確率が RBM によってモデル化されている．またオートエンコーダ（Auto Encoder; AE）を用いた手法 [10] ではスペクトル包絡に適した特徴量を AE を用いて抽出し，その特徴量を DNN で推定している．

一方で，統計的パラメトリック音声合成のようにスペクトル包絡と基本周波数に分解せず，調波構造を持つ振幅スペクトルを言語特徴量やテキストから推定する試みもある [3], [4]．これらの手法では振幅スペクトルに対して位相復元を行うことで音声を得ることができる為，ボコーダを用いる必要がない特徴がある．[3] では言語特徴量から調波構造を持つ振幅スペクトルを推定する場合，誤差関数として KLD が MSE よりも適していることが主張されている．KLD の定義は式（2）に示されている．

$$\mathcal{L}_{\mathrm{KLD}}\left(\mathbf{y} \mid \hat{\mathbf{y}}\right) = \sum_{t=1}^{T}\sum_{d=1}^{D} y_{t,d}\log\frac{y_{t,d}}{\hat{y}_{t,d}} - y_{t,d} + \hat{y}_{t,d} \qquad (2)$$

3. NMF のアクティベーションに着目したスペクトルモデリング

3.1 NMF

NMF は，ある行列 $\mathbf{Y} = (y_{k,n})_{K\times N}$ を $\mathbf{H} = (h_{k,m})_{K\times M}$ と $\mathbf{U} = (u_{m,n})_{M\times N}$ の 2 つの行列の積として表現するアルゴリズムである．この時全ての行列は非負になる特徴があり，また $K \gg M$ と $N \gg M$ を満たすように M を設定する場合が多い．NMF による行列の分解は式（3）のように表される．

$$\mathbf{Y} \simeq \mathbf{HU} \qquad (3)$$

行列 $\mathbf{H}$ は基底行列や辞書行列と呼ばれ，行列 $\mathbf{U}$ はアクティベーションと呼ばれる．スペクトルを NMF を用いてモデリングする場合，式（4）で表されるように n 番目のフレームのスペクトル $\mathbf{y}_n$ は，それぞれの基底スペクトル $\mathbf{h}_1, \ldots, \mathbf{h}_M$ に対して $u_{1,n}, \ldots, u_{M,n}$ で重み付けを行うことによる線形の足し合わせで表現される．

$$\mathbf{y}_n \simeq \sum_{m=1}^{M}\mathbf{h}_m u_{m,n} = \mathbf{H}\mathbf{u}_n \qquad (4)$$

NMF はその非負制約により得られるアクティベーションはスパースになる傾向があるが，よりスパースな解を得る為にスパース制約を導入することも可能である [16]．提案手法では，スペクトル包絡 $\mathbf{y}_n$ から得られるスパースな特徴量であるアクティベーション $\mathbf{u}_n$ を音響特徴量とする．

$\mathbf{Y} \simeq \mathbf{HU}$ を満たすような $\mathbf{H}$, $\mathbf{U}$ は，誤差関数に基づいて反復的に更新することによって求まる．NMF を用いた声質変換では振幅スペクトログラムに対する誤差関数とし

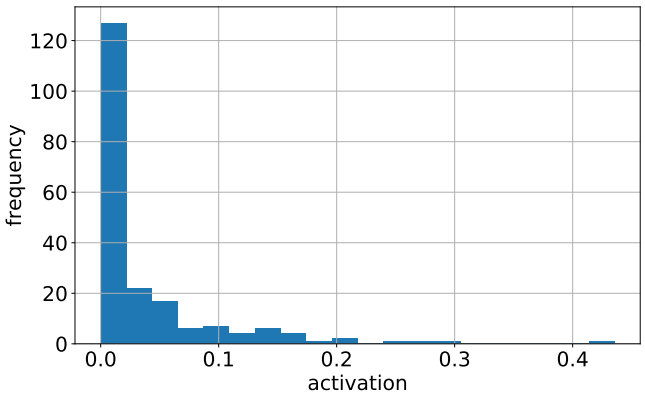

図 1　ある 1 フレームにおけるアクティベーションのヒストグラム
Fig. 1　Histogram of activation at one frame.

て KLD が用いられ [17]，本稿でも KLD に基づいて NMF を行うこととする．KLD を最小化する為に，**H**, **U** は式 (5)，(6) に従って反復的に更新される [12]．

$$h'_{k,m} \leftarrow h_{k,m} \frac{\sum_n y_{k,n} u_{m,n} / x_{k,n}}{\sum_n u_{m,n}} \quad (5)$$

$$u'_{m,n} \leftarrow u_{m,n} \frac{\sum_n y_{k,n} h_{k,m} / x_{k,n}}{\sum_k h_{k,m}} \quad (6)$$

$$x_{k,n} = \sum_m h_{k,m} u_{m,n}. \quad (7)$$

3.2　中間特徴量としてのアクティベーション

メルケプストラムでは，スペクトル包絡は異なる周波数の sin（あるいは cos）カーブの重ね合わせで表現される．同様に NMF でも，スペクトルは基底の重ね合わせで表現される．それゆえ，どちらにおいてもスペクトル包絡はそれ自身よりも低次元の特徴量で効率的に表現される．

しかし，メルケプストラムの場合ではどの音声データに対しても同じ基底が用いられ，係数（メルケプストラム）の違いのみによってスペクトルが表現される．一方で NMF を用いた場合は音声データ毎に基底を導く為に，異なる話者や音源の基底や異なるサンプリング周波数の基底を用意することができ，声質変換 [13] や帯域拡張 [14] の技術との親和性が高い．さらに各基底がスペクトルの微細構造を保持し，またその重みがスパースである為に，得られるスペクトル包絡の微細な構造を保持できることが期待できる．

3.3　アクティベーションの推定

NMF のアクティベーションに着目した統計的パラメトリックテキスト音声合成の構成を図 2 に示す．まず訓練用の音声データにより得られた振幅スペクトログラム **Y** を，NMF によって基底行列 **H** とアクティベーション **U** に分解し，その後言語特徴量と音響特徴量（**U**）の関係を

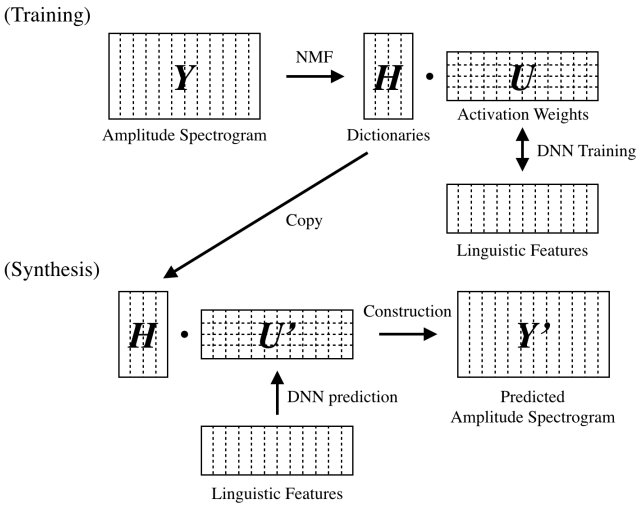

図 2　NMF を取り入れた統計的パラメトリックテキスト音声合成の構成
Fig. 2　Overview of staticstical parametric text-to-speech synthesis incorporating NMF.

表す DNN 音響モデルの学習を行う．そして，基底行列 **H** と DNN 音響モデルによって推測されたアクティベーション $\mathbf{U}'$ を掛け合わせることでテキストと対応した振幅スペクトログラム $\mathbf{Y}'$ を得る．

$$\mathbf{Y} \simeq \mathbf{HU} \quad (8)$$

$$\mathbf{Y}' = \mathbf{HU}' \quad (9)$$

アクティベーションの推定については，アクティベーションの持つスパース性が故に MSE は適切な誤差関数ではないことが考えられる．また，スパースであるアクティベーションを学習するということはつまり，所望のスペクトルに近い基底スペクトルを重み付けていくつか選び出す事に相当する為，クラス認識に近い問題であると考えるのが適当であると考えられる．その為，本稿では KLD から導かれるアクティベーションに適した誤差関数を提案する．

まず，$\mathbf{u}' = [u'_1, u'_2, \ldots, u'_M]$ をあるフレームにおけるアクティベーションであるとし，その和を c とする．つまり，$c = \sum_{m=1}^{M} u'_m$ を満たす．式 (10) で示されるように，$\mathbf{u}'$ の各要素をその和である c で割ることで正規化されたアクティベーション $\mathbf{u} = [u_1, u_2, \ldots, u_M]$ が得られる．

$$\mathbf{u} = \frac{\mathbf{u}'}{c} \quad (10)$$

u の和は 1 であり，**u** はカテゴリカル分布とみなすことができる．DNN 音響モデルで推定されたアクティベーション $\hat{\mathbf{u}}'$ と観測されたアクティベーション $\mathbf{u}'$ の KLD は式 (11) のように展開できる．

表 1　実験条件
Table 1　Experimental conditions.

Model	Output	Dimension	Normalization	Loss Function	Output Layer
MCEP	Mel-cepstrum	180 (mcep+Δ+ΔΔ)	Mean: 0, Var: 1	MSE	linear
SP	spectrum	513 (16 kHz), 1025 (48 kHz)	Min: 0.01, Max: 0.99	KLD	sigmoid
LogSP	log spectrum	513 (16 kHz), 1025 (48 kHz)	Mean: 0, Var: 1	MSE	linear
ACT	activation	201 (norm-act: 200 + power: 1)	Sum:1 + None	CE + D-ISD	softmax + softplus

$$\mathcal{D}_{KL}(\mathbf{u}' \mid \hat{\mathbf{u}}') = \sum_m \left(c u_m \log \frac{c u_m}{\hat{c}\hat{u}_m} - c u_m + \hat{c}\hat{u}_m \right)$$
$$= c \left\{ -\sum_m u_m \log \hat{u}_m + \left(\frac{\hat{c}}{c} - \log \frac{\hat{c}}{c} - 1 \right) + \sum_m u_m \log u_m \right\}$$
$$= c \{ \mathcal{D}_{CE}(\mathbf{u} \mid \hat{\mathbf{u}}) + \mathcal{D}_{IS}(\hat{c} \mid c) - \mathcal{D}_{CE}(\mathbf{u} \mid \mathbf{u}) \} \qquad (11)$$

ここで，$\mathcal{D}_{CE}$ と $\mathcal{D}_{IS}$ はそれぞれクロスエントロピーと板倉斎藤擬距離（Itakura Saito Divergence; ISD）を表す．つまり，KLD の最小化は正規化アクティベーション $\hat{\mathbf{u}}$，$\mathbf{u}$ 間のクロスエントロピーと，パワー $\hat{c}$，c 間の双対板倉斎藤擬距離（Dual Itakura Saito Divergence; D-ISD）[*1]の和と等価である．

$$\mathcal{L}_{KL} = -\sum_m u_m \log \hat{u}_m + \left(\frac{\hat{c}}{c} - \log \frac{\hat{c}}{c} - 1 \right) \qquad (12)$$

よって提案手法では，音響モデルの出力を $\hat{\mathbf{u}}$ と $\hat{c}$，観測値を $\mathbf{u}$ と c として，上式の $\mathcal{L}_{KL}$ を誤差関数とする．

3.4　提案手法における帯域拡張

3.2 節で述べた通り，NMF では同一のアクティベーションを持つ仮定のもと，話者や音源，あるいはサンプリング周波数毎の基底行列を得ることができる為，声質変換，雑音除去，帯域拡張などの応用が提案されている．その応用の 1 つとして，提案手法では，狭帯域の音源から得られた基底行列を広帯域の音源から得られた基底行列に置き換えることで帯域拡張 [14] を実現している．狭帯域と広帯域のパラレルな発声データを用いて，これらが同一のアクティベーションを持つ仮定のもとパラレルな基底行列を構成した．最終的に得られる音声は，広帯域の基底と，狭帯域の音声で学習した音響モデルによって出力されたアクティベーションを掛け合わせることで得られる．

4.　実験的評価

4.1　実験条件

実験的評価は，ATR 音素バランス 503 文を利用し，A–I セット 450 文を学習に，J セット 53 文を評価に用いた [18]．音声サンプルは HTS[*2]のデモスクリプトにある男性の発声データを利用した．本実験では異なるサンプリング周波数（16 kHz と 48 kHz）の発声データを用いた．サンプリング周波数 16 kHz の発声データは元の 48 kHz の発声データをダウンサンプリングすることによって得た．WORLD 分析を用いて，スペクトル包絡，基本周波数（F_0），非周期性指標を抽出し，また音響モデルで推測された特徴量から波形を生成する際は WORLD ボコーダを用いた [19]．

まず，従来手法（MCEP, SP, LogSP）と，提案手法（ACT）の比較を行った．実験条件は表 1 に示している．全ての手法について，DNN の入力特徴量として 0.01 から 0.99 の値を持つように正規化された言語特徴量を用い，DNN のアーキテクチャは Feed-Forward 型であり，6 層・1024 ユニット，活性化関数 *tanh* の隠れ層を持つようにした．MCEP の出力は 60 次元のメルケプストラムと，その動的特徴量 120 次元を合わせた 180 次元の特徴量である．この次元数は DNN 音声合成ツール Merlin[*3]のデフォルト値である．SP と LogSP の出力は共に 513 次元（16 kHz）と 1025 次元（48 kHz）のスペクトル包絡であろ，ACT の出力は 200 次元の正規化アクティベーションと 1 次元のパワーからなる 201 次元の特徴量である．またアクティベーションは振幅スペクトログラムに対して NMF を行い，基底行列は更新の各基底の L2 ノルムが 1 になるように正規化し，反復回数は 1000 とした．アクティベーションの次元数は再構成誤差が小さい上に，スペクトル包絡と比較して十分に圧縮した次元数となるような値を選択した．

MCEP と LogSP については，出力特徴量は平均 0 分散 1 になるように正規化を行い，誤差関数は MSE を用いた．これらのモデルでは線形の出力層を用いた．SP については出力特徴量は 0.01 から 0.99 の値を持つように正規化を行い，誤差関数は KLD を用いた．このモデルでは [3] と同様に出力層は sigmoid 関数を用いた．ACT では，正規化アクティベーションは既に和が 1 になるように正規化されており，パワーについては正規化しなかった．誤差関数は，正規化アクティベーションについてはクロスエントロピーを用い，パワーについては D-ISD を用いた．出力層は正規化アクティベーションには softmax，パワーには softplus を用いた．

加えて，帯域拡張実験も行った．まずサンプリング周波

[*1] 双対板倉斎藤擬距離は，板倉斎藤擬距離において観測値と推測値を逆にした距離尺度である
[*2] http://hts.sp.nitech.ac.jp/
[*3] http://www.cstr.ed.ac.uk/projects/merlin/

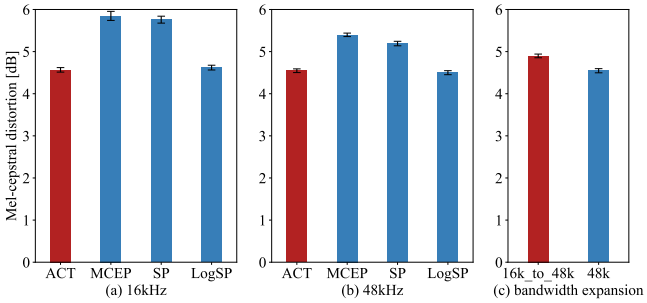

図 3 メルケプストラム歪み測定の結果．（a）はサンプリング周波数 16 kHz，（b）は 48 kHz の場合の 4 手法の比較結果であり，（c）は帯域拡張実験の場合の 2 手法の比較結果である．エラーバーは 95%信頼区間を示す．

Fig. 3 Mel-cepstral distortion; (a) in 16 kHz, (b) in 48 kHz sampling rates among the four models, and (c) of the experiment of bandwidth expansion. Error bars indicate 95 % confidence intervals.

数 16 kHz と 48 kHz の 50 文のパラレルデータを用いてパラレルな基底を構成した．この 50 文は訓練用の 450 文に含まれているものである．そして 48 kHz の基底行列と，16 kHz の音声で学習した DNN によって出力されるアクティベーションを掛け合わせることで 48 kHz の合成音声を得た（*16k_to_48k*）．この合成音声と，48 kHz の音声のみで DNN を学習したものとを比較した（*48k*）．

全ての手法について，公平にスペクトル包絡モデリングの比較を行う為に，音声合成時の F_0 や非周期性指標は観測された音声から抽出されたものを用い，スペクトル包絡パラメータのみを DNN から推定した．そして全手法において音声の品質を向上させるポストフィルタは用いなかった．また MCEP では動的特徴量を考慮する為に最尤パラメータ推定（Maximum Likelihood Parameter Generation; MLPG）を適用した [20]．

本実験では，客観的に評価を行う為に，メルケプストラム歪みを用いた [20]．また主観的な評価を行う為に，音声の品質に関するプリファレンス AB テストを行った．各テストにつき，比較する 2 手法の 10 ペアの音声がランダムに選択され，評価者が自然であると感じた方を選択するようにした．各評価について，評価者は 25 人である．

4.2 実験結果

図 3（a）（b）は異なるサンプリング周波数におけるテスト音声のメルケプストラム歪みの平均を示している．（a）は 16 kHz，（b）は 48 kHz の場合の 4 手法の比較である．図 3 より，16 kHz と 48 kHz のどちらにおいても提案手法は MCEP や SP に比べて低い歪みとなっていることが分かる．また，メルケプストラム歪みという尺度で計測しているにも関わらず提案手法は MCEP よりも小さい歪みとなっている．これは ACT の方がより微細な構造を保持できている為と考えられる．また LogSP と比較すると，ACT

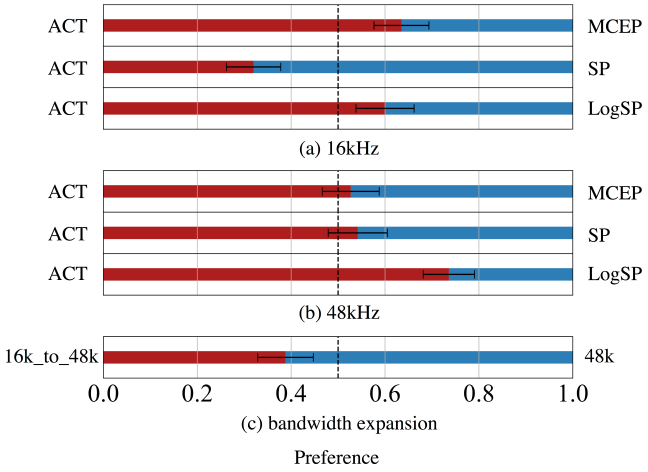

図 4 主観評価の結果．（a）はサンプリング周波数 16 kHz，（b）は 48 kHz の場合の 4 手法の比較結果であり，（c）は帯域拡張実験の場合の 2 手法の比較結果である．エラーバーは 95%信頼区間を示す．

Fig. 4 Subjective results; (a) in 16 kHz, (b) in 48 kHz sampling rates among the four models, and (c) of the experiment of bandwidth expansion. Error bars indicate 95 % confidence intervals.

はほぼ同等の結果となった．

図 4（a）（b）は異なるサンプリング周波数における主観結果を示している．サンプリング周波数 16 kHz での SP を除いて，ACT は他の手法に比べて良い自然性を得ていることが示された．この実験結果は提案手法は広帯域の条件において特に自然な音声を生成できることを示している．

また，帯域拡張実験の結果を図 3（c）と図 4（c）に示す．どちらの評価においても，*48k* は *16k_to_48k* に対して僅かに良い結果となっていることが分かる．*48k* では 450 文の広帯域の音声が学習に使われているが，*16k_to_48k* では広帯域の音声は 50 文しか用いられていない．*16k_to_48k* の音響モデルが狭帯域の音声のみで学習されているにも関わらず，提案手法は *48k* と大きく自然性が変わらないことが示された．

5. まとめ

本稿では，NMF におけるアクティベーションを音響特徴量とした統計的パラメトリック音声合成を提案した．スパースな特徴量であるアクティベーションを適切に推定する為に KLD から導かれる，クロスエントロピーと双対板倉斎藤擬距離の和を誤差関数として設定した．アクティベーションを音響特徴量とした場合，スペクトル包絡に比べて少量のパラメータでスペクトル包絡を表現できる上に，同様の低次表現であるメルケプストラムと異なりデータに合わせた基底を得ることができ，また基底が微細なスペクトル構造を保持できるという特徴がある．

加えて，提案手法の拡張性の 1 つとして帯域拡張を取り込むことができることが示された．少量の広帯域の音声を

用意すれば，提案手法は音響モデルが狭帯域の音声のみで学習された場合でも広帯域の音声を生成可能である．

実験的評価では，提案手法は特にサンプリング周波数 48 kHz の場合に他手法に勝る自然性の音声が生成できることが示された．また 16 kHz から 48 kHz への帯域拡張実験では，狭帯域の音声のみで学習した音響モデルを用いて相応の品質の広帯域音声が生成できることが示された．

今後の展望として，NMF を用いた声質変換を応用した複数話者音声合成が期待できる．アクティベーションの話者非依存性を仮定することで，異なる話者の音声を使ったデータ拡張の可能性も考えられる．また NMF を用いた雑音除去を用いて雑音入りの音声でも学習できる枠組みを取り込むことも期待できる．帯域拡張における品質の改善や NMF におけるスパース制約の導入による効果についても検討の余地がある．

参考文献

[1] Keiichi Tokuda, Yoshihiko Nankaku, Tomoki Toda, Heiga Zen, Junichi Yamagishi, and Keiichiro Oura. Speech synthesis based on hidden markov models. *Proceedings of the IEEE*, Vol. 101, No. 5, pp. 1234–1252, 2013.

[2] Heiga Zen, Andrew Senior, and Mike Schuster. Statistical parametric speech synthesis using deep neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 7962–7966, 2013.

[3] Shinji Takaki, Hirokazu Kameoka, and Junichi Yamagishi. Direct modeling of frequency spectra and waveform generation based on phase recovery for dnn-based speech synthesis. In *INTERSPEECH*, pp. 1128–1132, 2017.

[4] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017.

[5] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[6] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779–4783. IEEE, 2018.

[7] Yuchen Fan, Yao Qian, Feng-Long Xie, and Frank K Soong. Tts synthesis with bidirectional lstm based recurrent neural networks. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[8] Manu Airaksinen, Bajibabu Bollepalli, Lauri Juvela, Zhizheng Wu, Simon King, and Paavo Alku. Glottdnn-a full-band glottal vocoder for statistical parametric speech synthesis. In *Interspeech*, pp. 2473–2477, 2016.

[9] Eunwoo Song, Kyungguen Byun, and Hong-Goo Kang. Excitnet vocoder: A neural excitation model for parametric speech synthesis systems. *arXiv preprint arXiv:1811.04769*, 2018.

[10] Shinji Takaki, SangJin Kim, Junichi Yamagishi, and JongJin Kim. Multiple feed-forward deep neural networks for statistical parametric speech synthesis. In *Interspeech*, pp. 2242–2246, 2015.

[11] Zhen-Hua Ling, Li Deng, and Dong Yu. Modeling spectral envelopes using restricted boltzmann machines and deep belief networks for statistical parametric speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 21, No. 10, pp. 2129–2139, 2013.

[12] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pp. 556–562, 2001.

[13] Zhizheng Wu, Tuomas Virtanen, Tomi Kinnunen, Eng Siong Chng, and Haizhou Li. Exemplar-based voice conversion using non-negative spectrogram deconvolution. In *ISCA Workshop on Speech Synthesis, SSW8*, pp. 201–206, 2013.

[14] Paris Smaragdis and Bhiksha Raj. Example-driven bandwidth expansion. In *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 135–138, 2007.

[15] Paris Smaragdis and Judith C Brown. Non-negative matrix factorization for polyphonic music transcription. In *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No. 03TH8684)*, pp. 177–180. IEEE, 2003.

[16] 亀岡弘和. 非負値行列因子分解とその音響信号処理への応用. 日本統計学会誌, Vol. 44, No. 2, pp. 383–407, 2015.

[17] Zhizheng Wu, Tuomas Virtanen, Eng Siong Chng, and Haizhou Li. Exemplar-based sparse representation with residual compensation for voice conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 22, No. 10, pp. 1506–1521, 2014.

[18] Akira Kurematsu, Kazuya Takeda, Yoshinori Sagisaka, Shigeru Katagiri, Hisao Kuwabara, and Kiyohiro Shikano. Atr japanese speech database as a tool of speech recognition and synthesis. *Speech communication*, Vol. 9, No. 4, pp. 357–363, 1990.

[19] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*, Vol. 99, No. 7, pp. 1877–1884, 2016.

[20] Tomoki Toda, Alan W Black, and Keiichi Tokuda. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, No. 8, pp. 2222–2235, 2007.