Vol.2016-SLP-114 No.21
2016/12/20

DNNに基づくテキスト音声合成のための FFTスペクトルを用いた位相復元に基づく音声波形生成

高木 信二[1,a)]　SangJin Kim[2,b)]　亀岡 弘和[3,c)]　山岸 順一[1,d)]

概要：多くの統計的パラメトリック音声合成システムでは，高品質ボコーダを用い，音声波形を構築する．その際，ソース・フィルタモデルに基づくボコーダが利用されることが多く，言語特徴量からメルケプストラム，F0 といった音響特徴量を予測し，ボコーダによる音声波形生成が行われる．しかし，ボコーダを用いたことに起因する合成音声の自然性の低下が常に問題となっており、これまで様々な研究が報告されている．しかし，ソース・フィルタモデルに基づいている限り，この問題を完全に解決することは容易ではない．そこで本研究では，ボコーダを用いない音声合成システム構築することを考える．具体的には、統計的パラメトリック音声合成において，振幅スペクトルからの位相復元，逆短時間フーリエ変換，および重加算法 (OLA) に基づき波形を生成することについて検討する．今回提案する音声合成の枠組みでは，まず，調波構造を含む振幅スペクトルの予測を DNN 音響モデルにより行い，次に，予測された振幅スペクトルから Griffin/Lim 法により位相を復元することで，音声波形の生成を行う．主観評価実験により，高品質ボコーダを用いた DNN 音声合成システムと提案システムの比較を行った結果、提案法ではボコーダに基づく合成音声特有のバジー感が無い合成音声の生成が可能であることを確認できた．

キーワード：統計的パラメトリック音声合成，DNN，FFT スペクトル，位相復元，ボコーダ

1. はじめに

統計的パラメトリック音声合成を実現する代表的な手法として，隠れマルコフモデル (Hidden Markov Model; HMM) に基づく枠組みが挙げられる [1]．HMM に基づく音声合成を用いることである程度高品質な音声の合成を実現できるが，決定木に基づくコンテキストクラスタリングにより学習データが分割されてしまうことや，出力分布として単純なガウス分布が状態単位で割り当てられるといった問題が存在する．近年では，このような問題に対してニューラルネットワークを用いることが検討されており，例えば，HMM 音響モデルとニューラルネットワークを組み合わせる手法 [2] や，HMM 音響モデルの枠組み全体をニューラルネットワークに置き換える手法 [3], [4], [5] が提案され，ニューラルネットワークに基づく音声合成システムは高い性能を持つことが報告されている．

統計的パラメトリック音声合成システムは，STRAIGHT[6] や WORLD[7], [8] 等の高品質ボコーダを用い構築されることが多い．これらボコーダを利用し音響特徴量の抽出，及び，音響特徴量から音声波形の生成が行われる．統計的パラメトリック音声合成では，音響モデルを用いテキストから音響特徴量の予測を行い，ボコーダを用いて音声波形の生成を行う．近年，DNN を用いることで統計的パラメトリック音声合成システムの性能は改善されているが，DNN に基づく統計的パラメトリック音声合成システムにおいても，ボコーダを用いることで音声の劣化が生じてしまう問題がある．この問題に対して様々な研究が報告されており，例えば，励振源モデルの改良 [9], [10]，Sinusoidal ボコーダの利用 [11]，複素スペクトルのモデル化 [12]，音声波形そのものの利用 [13] が挙げられる．しかし，統計的パラメトリック音声合成システムおいて，ボコーダを用いたことによる品質劣化を回避することは依然として問題である．

我々はオートエンコーダを用いることで，単純な FFT を用いて得られた調波構造を含む振幅スペクトル (以下，FFT スペクトルと呼ぶ) から，低次元特徴量を抽出する

[1] 国立情報学研究所
National Institute of Informatics, Chiyoda, Tokyo 101–8430, Japan
[2] Naver Labs, Naver Corporation, Korea
[3] 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
Morinosato Wakamiya, Atsugi, Kanagawa, 243-0198, Japan
[a)] takaki@nii.ac.jp
[b)] sangjin.kim@navercorp.com
[c)] kameoka.hirokazu@lab.ntt.co.jp
[d)] jyamagis@nii.ac.jp

ⓒ 2016 Information Processing Society of Japan

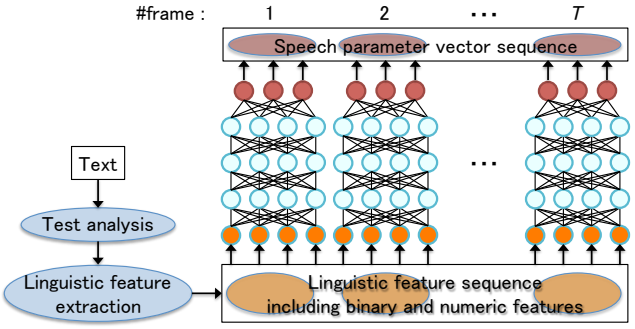


図 1　A framework for the DNN-based acoustic model.

枠組みを提案した [14]．[14] では，FFT スペクトルを学習データとして利用したが，DNN に基づく音響モデル (以降，DNN 音響モデルと呼ぶ) とオートエンコーダにより，自動的に調波構造の取り除かれたスペクトルの予測が行われ，WORLD ボコーダを用い高品質な音声の合成が可能であることを示した．この自動的に調波構造が取り除かれ，ボコーダに利用可能なスペクトル包絡が予測されたことは非常に興味深い．しかし一方で，調波構造を持つ FFT スペクトルを予測することにより，位相復元，逆短時間フーリエ変換に基づく音声波形生成 [15] といった，ボコーダを用いない音声波形生成アルゴリズムの利用が可能となることから，調波構造を持つ FFT スペクトルのモデル化も非常に興味深い．

そこで本研究では，統計的パラメトリック音声合成において，FFT スペクトルからの Griffin/Lim 法による位相復元，逆短時間フーリエ変換に基づく音声波形生成を検討する．FFT スペクトルに基づく提案システムの構築には少なくとも，1) FFT スペクトルの調波構造の予測，2) 適切な位相復元を可能とする高精度な FFT スペクトルの生成が必要となる．本研究では，調波構造を含む FFT スペクトルの高精度な生成を行うため，1) DNN 音響モデルの入力 (言語特徴量に加え F0 に関する特徴量の利用)，2) DNN 音響モデルの学習基準 (Kullback-Leibler divergence の利用)，3) DNN 音響モデルにより予測された FFT スペクトルのピーク強調 (信号処理に基づくポストフィルタの利用) の検討を行う．実験では，FFT スペクトルに基づく合成音声と高品質ボコーダ WORLD に基づく合成音声の比較を行った．

2. DNN に基づく音響モデル

2.1 概要

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が提案されている [2], [3], [4], [16]．本セクションでは代表的な DNN 音響モデルの 1 つである [3] について簡潔にレビューし，FFT スペクトルモデル化のための DNN 音響モデルの学習基準について述べる．

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．DNN 音響モデルの利点の一つとして，例えば i-vector による話者情報 [17] といった言語特徴量以外の特徴量を入力として容易に利用できる点が挙げられる．[3] 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．本研究では，単純な FFT により得られた高次元振幅スペクトルを音響特徴量として扱う．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 [18]．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．

2.2 学習基準

テキスト音声合成のための DNN 音響モデルの構築では，学習基準として最小二乗誤差 (SE) が用いられることが多い．最小二乗誤差に基づく学習基準は，以下のように表される．

$$\begin{aligned}\hat{\boldsymbol{\lambda}}_{SE} &= \arg\min_{\boldsymbol{\lambda}} E_{SE} \\ &= \arg\min_{\boldsymbol{\lambda}} \frac{1}{2}\sum_{t=1}^{T}\sum_{d=1}^{D}(o_{t,d}-y_{t,d})^2, \end{aligned} \quad (1)$$

ここで，$y_{t,d} = g_d^{(\boldsymbol{\lambda})}(\boldsymbol{l}_t)$ と $o_{t,d}$, $\boldsymbol{l}_t$, t, d, $\boldsymbol{\lambda}$ はそれぞれ観測 (音響特徴量)，DNN 音響モデルの入力 (言語特徴量)，フレームインデックス，次元，DNN のモデルパラメータを表す．また，関数 $g^{(\boldsymbol{\lambda})}(\cdot)$ は DNN によって表現される非線形変換である．

[3] で用いられた音響特徴量と異なり，本研究では高次元 FFT スペクトルを学習データとして用いる．本研究では FFT スペクトルを直接学習データとして用いる利点を活かし，適切な音響モデルの構築を行うため，Kullback-Leibler divergence (KLD) に基づく評価基準を用い，DNN 音響モデルの学習を行う．KLD 基準は非負値行列因子分解に基づく音源分離において広く利用されている [19], [20]．本研究で用いた KLD に基づく学習基準を以下に示す．

$$\begin{aligned}\hat{\boldsymbol{\lambda}}_{KL} &= \arg\min_{\boldsymbol{\lambda}} E_{KL} \\ &= \arg\min_{\boldsymbol{\lambda}} \sum_{t=1}^{T}\sum_{d=1}^{D} o_{t,d}\log\frac{o_{t,d}}{\tilde{y}_{t,d}} - o_{t,d} + \tilde{y}_{t,d}. \end{aligned} \quad (2)$$

ここで，$\tilde{y}_{t,d} = s_d y_{t,d} + b_d$ であり，s_d と b_d は学習データ

ⓒ 2016 Information Processing Society of Japan

表 1　データベースの詳細.

Speaker	Professional female
#Utterance (Train)	12,085
#Sentence (Test)	200
Sampling rate	48kHz

から前もって計算され，正規化された値を元に戻す処理に用いる値である．KLD に基づく学習基準を適用するためには観測と $\tilde{y}_{t,d}$ は正の数である必要がある．本研究では，出力層にシグモイド関数を用い正規化された 0 から 1 の間の値を出力することで，$\tilde{y}_{t,d}$ が取り得る値の範囲の制限を行い，KLD に基づく学習基準を適用する．

また，以下の通り表現される $y_{t,d}$ に関する偏微分を用い，確率的勾配降下法 [18] により DNN のパラメータは効率良く学習できる．

$$\frac{\partial E_{SE}}{\partial y_{t,d}} = y_{t,d} - o_{t,d} \tag{3}$$

$$\frac{\partial E_{KL}}{\partial y_{t,d}} = s_d \left(1 - \frac{o_{t,d}}{s_d y_{t,d} + b_d}\right). \tag{4}$$

3. 位相復元に基づく波形生成

本セクションでは，提案システムで用いられる音声波形生成アルゴリズムについて述べる．本研究では，DNN 音響モデルから調波構造を含む振幅スペクトルが予測されると仮定し，FFT スペクトルからの位相復元，逆短時間フーリエ変換，および重加算法 (OLA) に基づく音声波形生成を用いる．FFT スペクトルからの位相復元として Griffin/Lim 法による位相復元 [15] を用いる．この位相復元アルゴリズムは，1) 逆短時間フーリエ変換，重加算法による波形生成， 2) 窓掛け処理，短時間フーリエ変換によるスペクトル分析の繰り返し処理に基づく．このアルゴリズムでは振幅スペクトルの値は更新されず固定されるが，位相情報は繰り返し毎に短時間フーリエ変換を行った際に得られる位相情報により更新を行う．

提案システムでは，以下の手順により音声波形生成を行う．

(1) DNN 音響モデルにより FFT スペクトルを予測する．
(2) Griffin/Lim 法による位相復元を行う．本研究では，位相の初期値にはランダム値を用いた．
(3) 音響モデルにより予測された FFT スペクトルと復元された位相情報を用い，逆短時間フーリエ変換，および重加算法により音声波形生成を行う．

4. 実験

4.1 実験条件

Blizzard Challenge 2011 において配布された約 17 時間の英語データを実験に用いた [21], [22]．表 1 にデータベースの詳細を示す．

実験では，5 種類の音響モデル (WORLD-SE，FFT-SE，FFT-KLD，FFT-SE+F0，FFT-KLD+F0) の構築を行った．表 2 にこれら音響モデル構築に用いた入力特徴量，出力特徴量，学習基準を示す．FFT スペクトルのための DNN 音響モデル構築では，学習基準だけでなく入力特徴量について検討を行っており，FFT-SE+F0 と FFT-KLD+F0 では F0 情報 (log F0，有声/無声パラメータ) を音響モデルの入力として用いている．WORLD-SE の構築には，WORLD スペクトル包絡から抽出したメルケプストラムを用いた．その他の音響モデル構築には，高次元 FFT スペクトルが音響特徴量として用いたが，学習基準として二乗誤差に基づく学習基準を用いた音響モデル (FFT-SE，FFT-SE+F0) では log スケールの FFT スペクトル，KLD に基づく学習基準を用いた音響モデル (FFT-KLD，FFT-KLD+F0) では linear スケールの FFT スペクトルをそれぞれ学習に用いた．音声波形生成として，WOLRD-SE では WORLD ボコーダを用い，その他のシステムでは，Griffin/Lim 法による位相復元に基づく音声波形生成アルゴリズムを用いた．

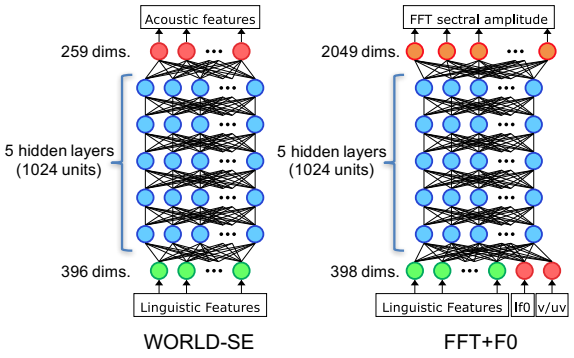


図 2　Configurations of neural networks used for acoustic models. FFT-SE, FFT-KLD, FFT-SE+F0, FFT-KLD+F0 use right side configuration (FFT+F0), though log F0 and voiced/unvoiced values were not used for constructing FFT-SE and FFT-KLD.

FFT 長 4096 で FFT スペクトル，WORLD スペクトルを得た．WORLD-SE 構築に用いた特徴量は 259 次元であり，59 次 WORLD メルケプストラム，対数基本周波数，25 次非周期成分とそれらの Δ，Δ^2，及び，1 次元有声/無声パラメータである．英語コンテキストラベルは発音辞書 Combilex を用いて作成された [23]．DNN 音響モデルの入力として用いられる言語特徴量は 396 次元である．また，言語特徴に含まれる音素継続長は HMM を用いて推定した．DNN 音響モデルの入力として用いられる言語特徴量，F0 情報は平均 0 分散 1 に正規化を行った．WORLD-SE，FFT-SE，FFT-SE+F0 で用いられる音響特徴量は 0.0–1.0 の範囲へ正規化を行った．FFT-KLD，FFT-KLD+F0 の学習に用いる FFT スペクトルについては正規化処理は行わないが，式 (2)，(4) の通り 0.0–1.0 の範囲への正規化を元に戻す値は用いられる．図 2 に実験で用いた音響モデルのネットワーク構造を示す．全ての DNN の全ての隠れ層，

ⓒ 2016 Information Processing Society of Japan

表 2 Inputs, output references and objective criteria for training each acoustic model are listed in this table. Here, v/uv and bap represent voiced/unvoiced values and band aperiodicity measures, respectively.

Model name	Input	Output reference	Objective criterion	Waveform generation
WORLD-SE	linguistic features	mel-cep. log F0, v/uv, bap	square error	vocoder
FFT-SE	linguistic features	log FFT spectral amplitude	square error	phase reconstruction + iFFT
FFT-KLD	linguistic features	FFT spectral amplitude	KL-divergence	phase reconstruction + iFFT
FFT-SE+F0	linguistic features, log F0, v/uv	log FFT spectral amplitude	square error	phase reconstruction + iFFT
FFT-KLD+F0	linguistic features, log F0, v/uv	FFT spectral amplitude	KL-divergence	phase reconstruction + iFFT

出力層のユニットでシグモイド関数を用いた．

WORLD-SE，FFT-SE+F0，FFT-KLD+F0 に対して信号処理に基づくケプストラムのためのポストフィルタ [24] を適用した．WORLD-SE では，予測されたメルケプストラムに対してポストフィルタを適用した．FFT-SE+F0 と FFT-KLD+F0 では，1) DNN 音響モデルにより予測された 2049 次 FFT スペクトルを 2049 次ケプストラムに変換，2) 2049 次ケプストラムに対してポストフィルタを適用，3) ポストフィルタが適用されたケプストラムを 2049 次振幅スペクトルに変換し，ケプストラムのためのポストフィルタを適用した．

FFT-SE+F0 と FFT-KLD+F0 の学習には学習データから得られた言語特徴量，有声/ 無声パラメータ，log F0，FFT スペクトルを用いた．FFT-SE+F0 と FFT-KLD+F0 を用いた音声の合成時には，テキストから得られた言語特徴量と WORLD-SE を用い得られた log F0，有声/無声パラメータを入力として利用した．

主観評価実験には MUSHRA 法を用い，自然音声を隠れアンカーとして使用した．被験者数は 9 人である．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 20 文章を比較した．

4.2 実験結果

スペクトログラム

図 3 に各システムにおいて生成されたスペクトログラムの一部を示す．図 3 より，明示的に F0 情報を入力として用いた音響モデル (FFT-SE+F0，FFT-KLD+F0) は，他の音響モデルと比較して調波構造の予測が行えていることがわかる．また，入力に F0 情報を利用していないシステム (FFT-SE，FFT-KLD，WORLD-SE) の結果を比較すると，FFT-SE と FFT-KLD は WORLD-SE と比較して微かに調波構造の一部が予測されているが，F0 を入力として利用したシステムと比較して予測精度は低い．

次に，学習基準の違いに注目すると，二乗誤差基準を用い構築された音響モデル (FFT-SE，FFT-SE+F0) と比較し，KLD 基準を用い構築された音響モデル (FFT-KLD，FFT-KLD+F0) では調波構造のピークがより強調されたスペクトルの予測が行われている．また，FFT-SE+F0 では十分に予測できていない 3.0kHz から 4.0kHz 付近の調波構造が，FFT-KLD+F0 では予測されている．このことから，KLD 基準による音響モデル学習が，調波構造を含む FFT スペクトルのモデル化に有効であるとわかる．

最後に，図 3 よりポストフィルタを適用することで，調波構造のピーク強調が行われていることがわかる．これらの結果から，DNN 音響モデルへの入力としての F0 情報の利用，KLD 基準による音響モデル学習，信号処理に基づくポストフィルタの利用が，調波構造を含む FFT スペクトルの生成に有効であることがわかる．

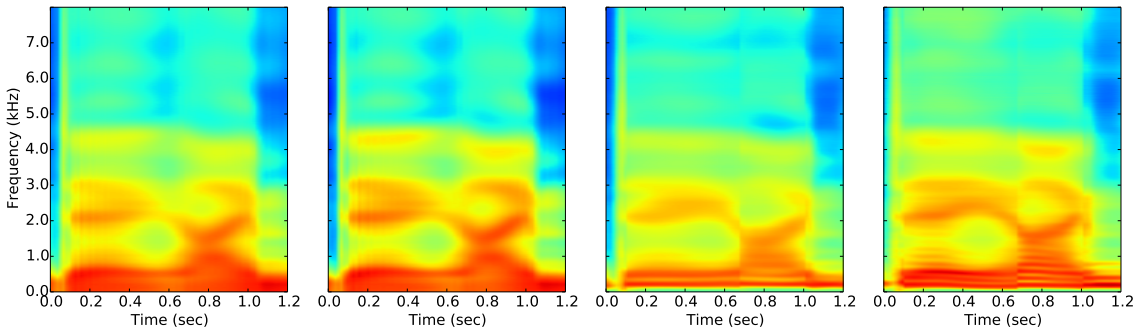
主観評価実験結果

図 4 に主観評価実験結果を示す．隠れアンカーの結果は図から除いている．主観評価実験では 32kHz にダウンサンプリングした音声を用い音響モデル (FFT-KLD+F0 (PF, 32kHz)) を新たに構築した．音響モデルの構築方法，音声波形生成の手順は FFT-KLD+F0(PF) と同様であるが，FFT 長は 2048 とした．32kHz と 48kHz の自然音声の品質は同等であると考えられるが，32kHz とすることで大幅に FFT スペクトルの次元数の削減が行われ，DNN 音響モデルの学習が容易になることが期待される．主観評価実験では WORLD-SE，WORLD-SE (PF)，FFT-SE+F0, FFT-KLD+F0, FFT-KLD+F0 (PF), FFT-KLD+F0 (PF，32kHz) による 6 種類のシステムを用いた．

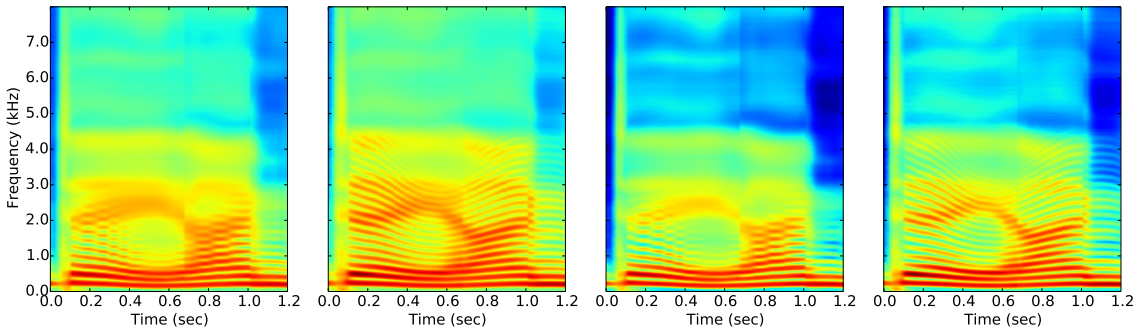
まず，ポストフィルタを適用していないシステム間で比較を行うと，図 4 より KLD 基準を用いたシステム (FFT-KLD+F0) が二乗誤差基準を用いたシステム (FFT-SE+F0) より評価が良いことがわかる．このことは KLD 基準が FFT スペクトルのための音響モデル構築に有効であることを示している．しかし，ポストフィルタの適用を行っていない FFT スペクトルに基づくシステム (FFT-SE+F0, FFT-KLD+F0) の性能は，WORLD ボコーダに基づくシステム (WORLD-SE) より低い結果となった．

次に，FFT スペクトルに基づくシステムにおいて，ポストフィルタの適用の有無について結果を比較すると，図 4 よりポストフィルタを適用したシステム (FFT-KLD+F0 (PF)) は適用していないシステム (FFT-KLD+F0) と比較し，大幅に性能が向上していることがわかる．ポストフィルタの適用を行っていないシステム (FFT-KLD+F0) では

ⓒ 2016 Information Processing Society of Japan



(a) WORLD-SE (b) WORLD-SE (PF) (c) FFT-SE (d) FFT-KLD



(e) FFT-SE+F0 (f) FFT-KLD+F0 (g) FFT-SE+F0 (PF) (h) FFT-KLD+F0 (PF)

図 3 Low-frequency parts (8 kHz) of synthetic spectral amplitudes in each system. PF means the post-filter.

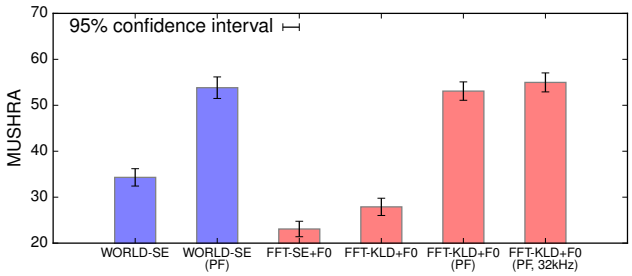


図 4 Subjective results.

復元された位相が適切ではなく，ノイズを多く含む音声が合成されたが，ポストフィルタの適用によりノイズが低減された．この結果よりポストフィルタによる FFT スペクトルのピーク強調が，Griffin/Lim 法による位相復元，音声波形生成に有効であるとわかる．

最後に，ポストフィルタを適用した FFT スペクトルに基づくシステム (FFT-KLD+F0 (PF)，FFT-KLD+F0 (PF，32kHz)) と高品質ボコーダ WORLD に基づくシステム (WORLD-SE (PF)) を比較すると，図 4 よりほぼ同程度の性能となっている．ポストフィルタを適用した場合でも FFT スペクトルに基づくシステムは Griffin/Lim 法による位相復元，波形生成に伴いノイズが生じているが，ボコーダを用いた合成で生じるバジー感は無い音声の合成が行われた．

5. おわりに

本研究では，統計的パラメトリック音声合成において，FFT スペクトルから Griffin/Lim 法により位相復元，短時間フーリエ変換，および重加算法 (OLA) に基づく音声波形生成を検討した．提案システムでは，STRAIGHT や WORLD といった高性能ボコーダを用いず音声波形の生成が行われる．音声合成実験により，調波構造を含む FFT スペクトルの DNN 音響モデル構築には，明示的な F0 情報の DNN 音響モデルへの入力としての利用，KLD に基づく学習基準の利用が有効であることがわかった．また，主観評価実験の結果より，音響モデルによる FFT スペクトルの予測精度は十分に高いとはいえず，ポストフィルタの適用が高品質な音声の合成に必要であることがわかった．ポストフィルタを適用した FFT スペクトルに基づく提案システムの性能は，高性能ボコーダ WORLD に基づく音声合成システムの性能と同程度であった．

今後の課題として，オートエンコーダを用いた調波構造を含む FFT スペクトルを表現する低次元特徴量抽出，FFT スペクトルのための DNN に基づくポストフィルタの構築が挙げられる．その他，複素スペクトル，音声波形そのものの利用も今後の課題として挙げられる．

6. 謝辞

本研究の一部は MEXT 科研費 JP16K16096，電気通信

ⓒ 2016 Information Processing Society of Japan

普及財団の助成を受けた.

参考文献

[1] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[2] Z.-H. Ling, L. Deng, and D. 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, pp. 2129–2139, 2013.

[3] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," *Proceedings of ICASSP*, pp. 7962–7966, 2013.

[4] 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.

[5] S. Kang and H. M. Meng, "Statistical parametric speech synthesis using weighted multi-distribution deep belief network," pp. 1959–1963, 2014.

[6] 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.

[7] M. Morise, "An attempt to develop a singing synthesizer by collaborative creation," *the Stockholm Music Acoustics Conference 2013 (SMAC2013)*, pp. 289–292, 2013.

[8] ——, "Cheaptrick, a spectral envelope estimator for high-quality speech synthesis,," *Speech Communication*, vol. 67, pp. 1–7, 2015.

[9] L. Juvela, B. Bollepalli, M. Airaksinen, and P. Alku, "High-pitched excitation generation for glottal vocoding in statistical parametric speech synthesis using a deep neural network," *Proceedings of ICASSP*, pp. 5120–5124, 2016.

[10] L. Juvela, X. Wang, S. Takaki, M. Airaksinen, J. Yamagishi, and P. Alku, "Using text and acoustic features in predicting glottal excitation waveforms for parametric speech synthesis with recurrent neural networks," *Proceedings of Interspeech*, pp. 2283–2287, 2016.

[11] Q. Hu, Z. Wu, K. Richmond, J. Yamagishi, Y. Stylianou, and R. Maia, "Fusion of multiple parameterisations for DNN-based sinusoidal speech synthesis with multi-task learning," *Proceedings of Interspeech*, pp. 854–858, 2015.

[12] Q. Hu, J. Yamagishi, K. Richmond, K. Subramanian, and Y. Stylianou, "Initial investigation of speech synthesis based on complex-valued neural networks," *Proceedings of ICASSP*, pp. 5630–5634, 2016.

[13] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016.

[14] S. Takaki and J. Yamagishi, "A deep auto-encoder based low-dimensional feature extraction from FFT spectral envelopes for statistical parametric speech synthesis," *Proceedings of ICASSP*, pp. 5535–5539, 2016.

[15] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 32, pp. 236–243, 1984.

[16] 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.

[17] Z. Wu, P. Swietojanski, C. Veaux, R. Renals, and S. King, "A study of speaker adaptation for dnn-based speech synthesis," *Proceedings of Interspeech*, pp. 879–883, 2015.

[18] G. E. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science 28*, vol. 313, no. 5786, pp. 504–507, 2006.

[19] H. S. S. D. D. Lee, "Algorithms for nonnegative matrix factorization," *Proceedings of Adv. Neural Inform. Process. Syst.*, pp. 556–562, 2001.

[20] B. R. P. Smaragdis and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," *Proceedings of 7th Int. Conf. Ind. Compon. Anal. Signal Separat.*, pp. 414–421, 2007.

[21] S. King and V. Karaiskos, "The blizzard challenge 2011," *Blizzard Challenge 2011 Workshp*, 2011. [Online]. Available: http://festvox.org/blizzard/bc2011/summary_Blizzard2011.pdf

[22] "Data and tools related to the blizzard challenge." [Online]. Available: http://www.cstr.ed.ac.uk/projects/blizzard/

[23] K. Richmond, R. Clark, and S. Fitt, "On generating combilex pronunciations via morphological analysis," *Proceedings of Interspeech*, pp. 1974–1977, 2010.

[24] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Incorporation of mixed excitation model and postfilter into HMM-based text-to-speech synthesis," *IEICE*, vol. J87-D-II, no. 8, pp. 1565–1571, 2004.

ⓒ 2016 Information Processing Society of Japan