Blizzard Challenge 2018 のための NITech テキスト音声合成システム *

◎沢田慶 (名工大/MSD), 吉村建慶, 橋本佳, 大浦圭一郎, 南角吉彦, 徳田恵一 (名工大)

1 はじめに

本稿では，Blizzard Challenge 2018 [1] のために名古屋工業大学 (NITech) チームが開発したテキスト音声合成 (text-to-speech; TTS) システムの構成と評価結果について述べる．Blizzard Challenge 2018 では，Blizzard Challenge 2016, 2017 [2, 3] と同様に，表現豊かな音声が数多く含まれる児童書のオーディオブックを学習コーパスとして TTS システムを構築するタスクが行われた．我々は，これまでのチャレンジのために開発してきた学習コーパスの自動構築法 [4] と表現豊かな音声合成を実現する手法 [5] に加え，本チャンレンジではディープニューラルネットワーク (deep neural network; DNN) に基づくポーズ挿入モデルと WaveNet ボコーダを導入した．主催者による大規模主観評価実験の結果，NITech TTS システムは高いスコアを示した．

2 Blizzard Challenge 2018

TTS システムとは，任意のテキストに対応する音声を生成する技術のことである．近年では深層学習の導入により，合成音声の品質は劇的に向上している．例えば，2016 年に提案された WaveNet [6] は，音声波形の直接モデル化・生成を実現し，フレーム単位の処理に基づくボコーダによる音声品質劣化の問題を解決した．さらに，2018 年に提案された Tacotron2 [7] は，自然音声の品質に近い音声の合成を実現した．しかし，人のような様々な発話スタイルを再現する TTS システムを実現するためには改善の余地がある．

TTS システムを構築するための手法は数多く提案されているが，それらの手法を比較・評価することは容易ではない．これは，TTS システムの比較・評価に広く用いられている，受聴試験による主観評価結果は学習コーパス・タスク・受聴試験が異なる場合に，直接比較することが難しいからである．この問題への取り組みのひとつとして，共通の学習コーパスを用いて TTS システムを構築して受聴試験を行う Blizzard Challenge と題する評価会が 2005 年から毎年開催されている [8].

Blizzard Challenge 2018 では，Blizzard Challenge 2016, 2017 と同様に，英語の児童書のオーディオブック 56 冊 (約 7 時間) が学習コーパスとして配布された [1]．オーディオブックは，書籍を朗読した音声を収録したコンテンツであり，音声とテキストの対を大量に収集することができることや，同一話者による比較的クリーンかつ同一環境で収録された音声が多く存在することから，TTS システムのための学習コーパスとして注目を集めている．しかし，本チャレンジで配布された児童書のオーディオブックには，TTS システムを構築することを難しくするいくつかの挑戦的な点がある．まず，配布されたオーディオブックには言い間違いや言い淀み，擬音語等を含むため，音声とテキストは完全には一致するわけではない．また，TTS システム用の音声コーパスを想定した条件で収録されていないため，従来の TTS システムのための学習コーパスの品質には及ばない．さらに，複数のコーディックにより音声データが圧縮されている．このような問題に取り組むことは，インターネット上に存在するデータ (found data) を用いて TTS システムを構築するための知見につながると考えられる．また，児童書のオーディオブックは，複数のキャラクターが登場することや子供を楽しませるために表現豊かに読み上げていることから様々な発話スタイルの音声が含まれている．そのため，様々な発話スタイルの学習・合成を実現するための枠組みを検討する必要がある．以上のことから，Blizzard Challenge 2018 の学習コーパスを利用した TTS システム構築は挑戦的なタスクであると言える．

3 NITech テキスト音声合成システム

3.1 Blizzard Challenge 2016, 2017 のための NITech テキスト音声合成システム

NITech チームは，Blizzard Challenge 2016, 2017 にも参加してきた．これまでのチャレンジでは，配布された音声とテキストの不一致の問題に対して，音声認識に基づく TTS システムのための学習コーパス自動構築法を提案した [4]．この手法により，十分に整備されていないデータから，高品質な学習コーパスの構築を実現した．また，児童書のオーディオブックのために，言語特徴量の設計を行った [5]．テキストのみから発話スタイルを予測し再現する言語特徴量の導入により，表現豊かな音声合成を可能とした．さらに，系列内変動を考慮したトラジェクトリ学習 [9] に

* Overview of the NITech text-to-speech system for the Blizzard Challenge 2018. by SAWADA, Kei (Nagoya Institute of Technology / Microsoft Development Co., Ltd.), YOSHIMURA, Takenori, HASHIMOTO, Kei, OURA, Keiichiro, NANKAKU, Yoshihiko, and TOKUDA, Keiichi (Nagoya Institute of Technology)

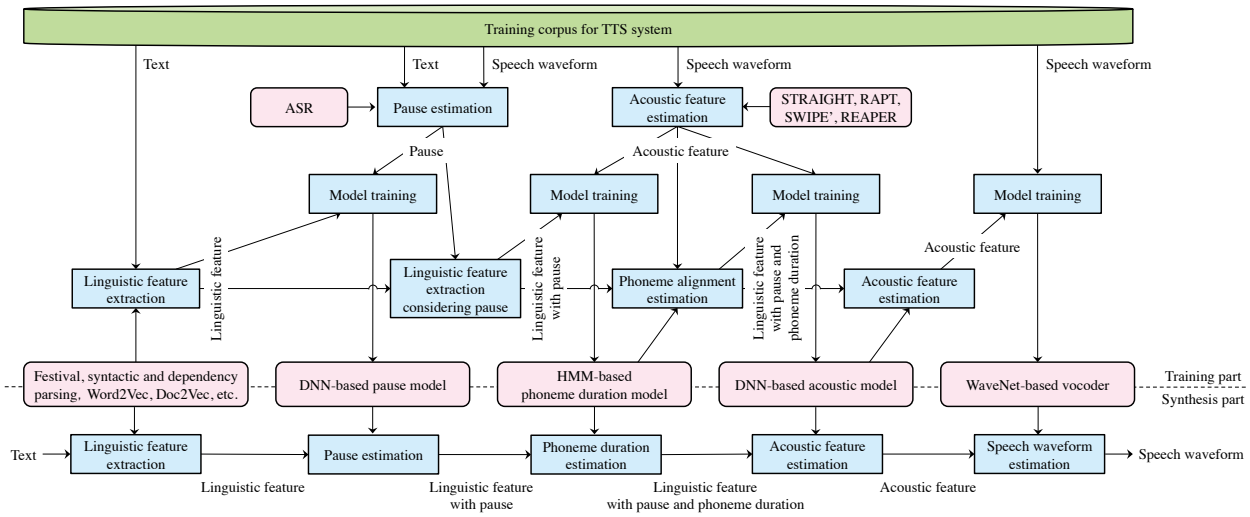

Fig. 1　Overview of the NITech TTS system

よる混合密度ネットワーク (mixture density network; MDN) [10] を音響モデルに用いた．これらの手法により，Blizzard Challenge 2017 では，NITech TTS システムは高い自然性と明瞭性を示した [3, 5].

3.2　Blizzard Challenge 2018 のための NITech テキスト音声合成システム

Blizzard Challenge 2018 のための NITech TTS システムの概要を Fig. 1 に示す.

3.2.1　ポーズ挿入モデル

自然な音声を合成するためには，適切な位置にポーズを挿入する必要がある．特にオーディオブックでは，登場人物の感情表現の一つとしてポーズを利用することもあるため，ポーズ挿入は重要な課題である．本システムでは，TTS システムのための学習コーパスのポーズ挿入スタイルを再現するポーズ挿入モデルを導入する.

まず，学習コーパスに含まれるポーズを音声認識 (音素アライメント推定) により予測する．音素アライメントは，全ての単語境界にショートポーズを持つ構造で推定する．ショートポーズモデルには，隠れマルコフモデル (hidden Markov model; HMM) の状態遷移において，状態のスキップを許可するモデル構造を用いた．そして，推定されたショートポーズの継続長が閾値以上の場合に，その単語境界にポーズが含まれるとした.

ポーズ挿入モデルには，音響モデルの入力にも利用する言語特徴量を入力とし，単語の後にポーズが挿入されるか否かを出力とする DNN を用いた．ただし，入力に用いる言語特徴量は，Blizzard Challenge 2017 のための NITech TTS システム [5] で設計した単語・文単位の言語特徴量のみとした．このポーズ挿入モデルにより，学習コーパスのポーズ挿入スタイルを再現することができると考えられる.

3.2.2　WaveNet ボコーダ

これまでの NITech TTS システムでは STRAIGHT を利用して音声を合成していたが，フレーム単位の処理に基づくボコーダの利用による音声品質の劣化が問題となっていた．そこで，音声波形を直接モデル化・生成可能な WaveNet [6] をボコーダとして利用する [11]．言語特徴量から音響特徴量への変換はこれまでの NITech TTS システムで開発してきたトラジェクトリ学習に基づく MDN [5, 9] を利用し，音響特徴量から音声波形への変換に WaveNet を利用する．WaveNet を学習する際には，自然音声から抽出した音響特徴量ではなく，学習済みの MDN の音響モデルから出力された音響特徴量を，WaveNet の補助特徴量として用いる．これは，学習時と合成時の補助特徴量の不一致による合成音声の品質劣化を防ぐためである.

WaveNet では，音声波形を精度良くモデル化するために，音声波形に対して量子化を適用する．このとき生じる量子化ノイズは時間に依存しないため白色となる．音声は高周波数領域のエネルギーが少ないため，特に量子化ノイズの高域の周波数成分が知覚されやすく，これが合成音声の品質を低下させる要因になると考えられる．十分な量子化ビット数により量子化すれば量子化ノイズは減少するものの，WaveNet の最適化が難しくなるという問題がある．本システムでは，量子化ビット数は増加させず，メルケプストラムに基づくノイズシェーピング量子化法およびプリフィルタリング [12] を利用することで合成音声の品

質を向上させる．この手法の基本的なアイディアは，量子化ノイズが知覚されにくくなるように，メルケプストラムを用いて音声波形を変形させることにある．これは WaveNet の学習における前処理であると見なせ，合成時には新たに追加となる処理を必要としない．詳しくは文献 [12] を参照されたい．

4 Blizzard Challenge 2018 評価実験

本稿で述べた TTS システムを用いて，Blizzard Challenge 2018 に参加した．Blizzard Challenge 2018 では，同一女性話者により収録された英語の児童書 56 冊の音声とテキストが学習コーパスとして配布された．

4.1 実験条件

音声認識に基づく TTS システムのための学習コーパス自動構築法を用いて学習コーパスを構築した [4]．学習データの取捨選択により，配布された 1258 ページの中，924 ページを学習コーパスとした．

言語特徴量は Festival, Stanford Parser, SyntaxNet, gensim を用いて抽出した．ポーズ挿入モデルには，テキストから推定した 251 次元の 0–1 に正規化した言語特徴量を入力とする bi-directional gated recurrent unit (GRU) を用いた．Bi-directional GRU の中間層は 3 層の 128 ユニット，ドロップアウト率は 0.2 とした．音響モデルのための言語特徴量には，925 次元のテキストから推定した特徴，10 次元の継続長特徴，150 次元の単語コード，600 次元の句コードから構成される計 1685 次元を用いた [5]．ここで，単語・句コードは固定次元の平均 0 分散 1 の正規分布に基づく乱数ベクトルとした．音響特徴量には，配布されたサンプリング周波数 44.1kHz からダウンサンプリングした 32kHz の音声に対してフレームシフト 5ms で推定した 64 次の STRAIGHT メルケプストラム，無声部を有声部で線形補間した F_0 と有声無声情報，32 次の非周期成分のメルケプストラムを用いた．精度の高い F_0 を抽出するために，RAPT, SWIPE', REAPER を用いて抽出した F_0 の多数決の結果を利用した．音響モデル学習のために，言語特徴量は 0–1 で正規化し，音響特徴量は平均 0 分散 1 の正規分布となるように正規化した．MDN の中間層は 3 層の 8000 ユニット，ドロップアウト率は 0.6，GV 重みは 0.001 である [9]．活性化関数には隠れ層にシグモイド関数，出力層に線形関数を用いた．WaveNet の学習のために，32kHz の音声波形を μ-law アルゴリズムを用いて 8bit に量子化した．メルケプストラムに基づくノイズシェーピングおよびプリフィルタリングの強度を調整するパラメータは，それぞれ $\gamma = 0.1$, $\beta = 0.1$ とした [12]．WaveNet の構造には，dilation を $1, 2, 4, \ldots, 512$ とした 10 層からなるブロックを 3 段重ねて構造を用いた．また，dilation, residual-connection, skip-connection [6] における畳み込み層のチャンネル数は 256 とした．Blizzard Challenge 2018 のルールを満たすために，32kHz の合成音声を 48kHz にアップサンプリングした音声を提出した．

4.2 評価結果

Blizzard Challenge 2018 の主催者により大規模主観評価実験が行われた．ページ単位の合成音声を評価するために 7 つの基準 (overall impression; OI, pleasantness; PL, speech pause; SP, stress; ST, intonation; IN, emotion; EM, listening effort; LE) の 60 段階 MOS 試験，文単位の合成音声を評価するために 2 つの基準 (naturalness; NAT, similarity; SIM) の 5 段階 MOS 試験，合成音声の明瞭性を評価するために無意味文 (semantically unpredictable sentences; SUS) の書き取りテストが実施された．全被験者 (有償・無償被験者，非英語ネイティブを含む) による各評価基準の MOS と，英語ネイティブの有償被験者による書き取りテストの単語誤り率 (word error rate; WER) を表 1 に示す．表 1 において，システム A は自然音声，システム B から E はベンチマークシステム，システム F から O は参加者のシステムを表し，NITech TTS システムは **I** である．システム順は，NAT の MOS を基準に表記されている．また，表 1 の * は NITech TTS システム **I** と有意な差があるシステムを意味する．

表 1 より，システム **I** は，ページ・文単位ともに高い MOS を示した．さらに，参加システムの中で最も低い WER を示したことから，高い明瞭性の合成音声であることが示された．このことから，NITech TTS システムは自然性・話者類似性・明瞭性が高い音声を合成できたと言える．

次に NITech 2017[1] と NITech 2018[2] の合成音声の違いについて考察する．WaveNet ボコーダの導入により，フレーム単位の処理に基づくボコーダによる音質劣化を回避することができ，自然・話者性が改善した．しかし，WaveNet ボコーダの予測精度は十分でなく，発音が不安定になることがあった．学習コーパスとして配布された音声データは複数のコーディックで圧縮されており，このことが波形を直接モデル化する WaveNet の学習を難しくしたと考えられる．さらに，学習コーパスには多くの表現豊かな音声が含まれるが，様々な発話スタイルをモデル化するための

[1] http://www.sp.nitech.ac.jp/~swdkei/syn/Blizzard_2017/index.html

[2] http://www.sp.nitech.ac.jp/~swdkei/syn/Blizzard_2018/index.html

Table 1　Evaluation results

System	Page domain							Sentence domain		SUS
	OI	PL	SP	ST	IN	EM	LE	NAT	SIM	WER
A	48 ± 7*	48 ± 7*	48 ± 7*	48 ± 8*	48 ± 8*	48 ± 8*	49 ± 7*	4.8 ± 0.5*	4.5 ± 0.9*	–
K	38 ± 10*	37 ± 10*	36 ± 11*	36 ± 12*	37 ± 10*	38 ± 11*	37 ± 10*	4.0 ± 0.9*	3.9 ± 1.0	16
J	34 ± 10	33 ± 11	36 ± 11*	35 ± 11*	35 ± 11	35 ± 11	34 ± 10	3.7 ± 0.9	3.6 ± 0.9	18
I	**34 ± 10**	**33 ± 10**	**32 ± 11**	**33 ± 11**	**33 ± 11**	**35 ± 11**	**33 ± 10**	**3.5 ± 1.0**	**3.5 ± 1.1**	**11**
L	28 ± 11*	28 ± 11*	25 ± 12*	26 ± 12*	26 ± 12*	29 ± 11*	25 ± 11*	3.0 ± 1.2*	3.4 ± 1.1	24*
M	27 ± 11*	26 ± 11*	25 ± 13*	25 ± 12*	25 ± 12*	30 ± 11*	23 ± 11*	3.0 ± 1.2*	3.0 ± 1.2*	22*
B	29 ± 11*	28 ± 12*	27 ± 13*	27 ± 13*	27 ± 13*	31 ± 12*	25 ± 12*	2.9 ± 1.1*	3.2 ± 1.1	29*
D	27 ± 10*	25 ± 10*	31 ± 11	30 ± 12*	28 ± 11*	27 ± 12*	28 ± 10*	2.8 ± 1.0*	2.5 ± 1.0*	15
E	25 ± 9*	23 ± 9*	30 ± 11*	28 ± 11*	26 ± 11*	27 ± 11*	25 ± 9*	2.6 ± 1.0*	2.5 ± 1.0*	14
G	24 ± 10*	23 ± 10*	31 ± 11	30 ± 11*	27 ± 11*	25 ± 12*	27 ± 10*	2.6 ± 1.0*	2.2 ± 1.0*	15
F	19 ± 10*	20 ± 10*	25 ± 11*	24 ± 12*	21 ± 11*	19 ± 11*	21 ± 9*	2.4 ± 0.9*	1.9 ± 0.9*	20*
O	22 ± 10*	21 ± 9*	28 ± 12*	27 ± 12*	25 ± 11*	23 ± 12*	24 ± 10*	2.3 ± 1.0*	1.7 ± 0.8*	14
C	20 ± 9*	19 ± 9*	27 ± 12*	24 ± 11*	22 ± 10*	22 ± 11*	21 ± 9*	2.2 ± 0.9*	2.1 ± 1.0*	15
N	17 ± 9*	16 ± 8*	29 ± 11*	26 ± 12*	23 ± 11*	21 ± 11*	20 ± 9*	1.8 ± 0.9*	1.5 ± 0.7*	17
H	13 ± 7*	13 ± 7*	23 ± 12*	20 ± 11*	17 ± 11*	18 ± 11*	14 ± 8*	1.6 ± 0.8*	1.5 ± 0.8*	37*

学習コーパスとしては十分な量ではなかったと考えられる．実際に，表現豊かな音声を合成することができる MDN 音響モデルから生成した音響特徴量を，WaveNet ボコーダに入力すると発音が不安定になることが多くあった．このことから，表現豊かな合成音声という点では NITech 2017 TTS システムの方が優れていると思われる．そのため，WaveNet ボコーダにおいても表現豊かな音声を合成することが今後の課題である．

5　むすび

本稿では，Blizzard Challenge 2018 のための NITech TTS システムの構成と評価結果を述べた．本チャレンジにおいて NITech チームは，ポーズ挿入モデルと WaveNet ボコーダを TTS システムに導入した．大規模主観評価実験の結果，NITech TTS システムは高い自然・話者類似性と最も高い明瞭性を示した．今後の課題として，WaveNet ボコーダにおいて表現豊かな音声合成を安定して実現することや，複雑な TTS システムの構造を簡略化する End-to-End モデルの導入が挙げられる．

謝辞　SCOPE (受付番号 162106106)，JSPS 基盤研究 (A) 18H04128.

参考文献

[1] Blizzard Challenge 2018, http://www.synsig.org/index.php/Blizzard_Challenge_2018.

[2] S. King *et al.*, “The Blizzard Challenge 2016,” Blizzard Challenge 2016 Workshop, 2016.

[3] S. King *et al.*, “The Blizzard Challenge 2017,” Blizzard Challenge 2017 Workshop, 2017.

[4] K. Sawada *et al.*, “The NITech text-to-speech system for the Blizzard Challenge 2016,” Blizzard Challenge 2016 Workshop, 2016.

[5] K. Sawada *et al.*, “The NITech text-to-speech system for the Blizzard Challenge 2017,” Blizzard Challenge 2017 Workshop, 2017.

[6] A. van den Oord *et al.*, “WaveNet: A generative model for raw audio,” arXiv:1609.03499, 2016.

[7] J. Shen *et al.*, “Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,” ICASSP 2018, pp. 4779–4783, 2018.

[8] A.W. Black *et al.*, “The Blizzard Challenge – 2005: Evaluating corpus-based speech synthesis on common datasets,” Interspeech 2005, pp. 77–80, 2015.

[9] K. Hashimoto *et al.*, “Trajectory training considering global variance for speech synthesis based on neural networks,” ICASSP 2016, pp. 5600–5604, 2016.

[10] H. Zen *et al.*, “Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis,” ICASSP 2014, pp. 3844–3848, 2014.

[11] A. Tamamori *et al.*, ”Speaker-dependent WaveNet vocoder, ” Interspeech 2017, pp. 1118–1122, 2017.

[12] T. Yoshimura *et al.*, ”Mel-cepstrum-based quantization noise shaping applied to neural-network-based speech waveform synthesis,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 7, pp. 1173-1180, 2018