The NITech text-to-speech system for the Blizzard Challenge 2018

Kei Sawada[1,2]*, Takenori Yoshimura*[1]*, Kei Hashimoto*[1]*,*
Keiichiro Oura[1]*, Yoshihiko Nankaku*[1]*, Keiichi Tokuda*[1]

[1]Nagoya Institute of Technology, Nagoya, JAPAN
[2]Microsoft Development Co., Ltd., Tokyo, JAPAN
{swdkei, takenori, bonanza, uratec, nankaku, tokuda}@sp.nitech.ac.jp

Abstract

This paper describes a text-to-speech (TTS) system developed at the Nagoya Institute of Technology (NITech) for the Blizzard Challenge 2018. In the challenge, about seven hours of highly expressive speech data from English children's audiobooks were provided as training data. For this challenge, we introduced deep neural network (DNN)-based pause insertion model and WaveNet-based neural vocoder. Large-scale subjective evaluation results show that the NITech TTS system achieved high score in various evaluation criteria.
Index Terms: text-to-speech system, deep neural network, WaveNet neural vocoder, Blizzard Challenge, audiobook

1. Introduction

A number of studies on text-to-speech (TTS) systems have been conducted. Consequently, the quality of synthetic speech has improved, and such systems are now used in various applications, such as for smartphones and smart speakers. Accordingly, the demand for TTS systems offering high-quality synthetic speech, various speaking styles, and various languages is increasing.

Although many TTS systems have been proposed, comparisons of such systems are difficult when the corpus, task, and listening test is different. The Blizzard Challenge was started in order to better understand and compare research techniques in constructing corpus-based speech synthesizers with the same data in 2005 [1]. This challenge has so far provided English, Mandarin, some Indian languages, English audiobooks, etc. as training data. The series of Blizzard Challenges has helped us measure progress in TTS technology [2].

As computer processing power increased, approaches based on big data have been successful in various research fields. In corpus-based speech synthesis, a quality of synthesized speech was improved by using a large amount of training data. Therefore, a TTS system based on big data is important in speech synthesis research. Speech data recorded with less noise and under the same recording conditions are suitable for training TTS systems. A large amount of training data is also necessary to synthesize various speaking styles. For this reason, recording a large amount of speech data for a TTS system requires a huge cost. Therefore, TTS system construction method based on audiobooks has received considerable attention. Audiobooks can be relatively easily collected as a large amount of speech data and text pairs. In the Blizzard Challenge 2013, around 300 hours of audiobooks were provided as training data [3]. In the Blizzard Challenge 2016 [4] and 2017 [5], highly expressive speech data from professionally produced English children's audiobooks were provided as training data. In the Blizzard Challenge 2018, about seven hours of speech data from children's audiobooks, which is identical to the Blizzard Challenge 2017 and includes the five hours released in Blizzard Challenge 2016, were provided as training data [6]. All 56 books were recorded by one native British English female professional speaker. Speech data were sampled at a rate of 44.1 kHz and coded in the MP3, M4A, and WMA formats. Texts corresponding to speech data were also provided. The task was to construct a speech from this data that is suitable for reading audiobooks to children.

The Nagoya Institute of Technology (NITech) have been submitting statistical parametric speech synthesis (SPSS) system for the Blizzard Challenge. Typical SPSS systems have three main components: linguistic features estimation, acoustic features estimation, and speech waveform generation. In the linguistic features estimation component, linguistic features, e.g., phonemes, syllables, accents, and parts-of-speech, of an input text is estimated. In the acoustic features estimation component, acoustic features, which express characteristics of a speech waveform, is estimated with the linguistic features. In the speech waveform generation component, a speech waveform is generated from the acoustic features by using a vocoder.

We focused on three approaches for Blizzard Challenge 2016 [7] and 2017 [8]: 1) automatic construction of a training corpus for SPSS systems from audiobooks; 2) design of linguistic features for SPSS based on audiobooks; and 3) mixture density network acoustic model [9, 10] incorporating trajectory training [11]. The last year's NITech system showed good performance in terms of naturalness and intelligibility. For this year's challenge, we introduced deep neural network (DNN)-based pause insertion model and WaveNet-based neural vocoder.

The rest of this paper is organized as follows. Section 2 describes the NITech TTS system for the Blizzard Challenge 2018. Subjective listening test results are given in Section 3 and concluding remarks and an outline for future work are presented in the final section.

2. NITech TTS system

2.1. NITech TTS system for Blizzard Challenge 2016, 2017

The Nagoya Institute of Technology (NITech) team submitted a text-to-speech (TTS) system for the Blizzard Challenge 2016 and 2017. In these challenges, the provided audiobooks contained mismatches between speech data and text. To overcome this problem, we investigated the automatic construction of a training corpus from audiobooks using a speech recognizer [7]. This method realized construction of high quality training corpus from audiobooks. Moreover, we redesigned linguistic features for statistical parametric speech synthesis (SPSS) based on audiobooks [8]. Introduction of linguistic features which can predict and reproduce speaking style from text, enabled expressive speech synthesis. In addition, we introduced the parame-

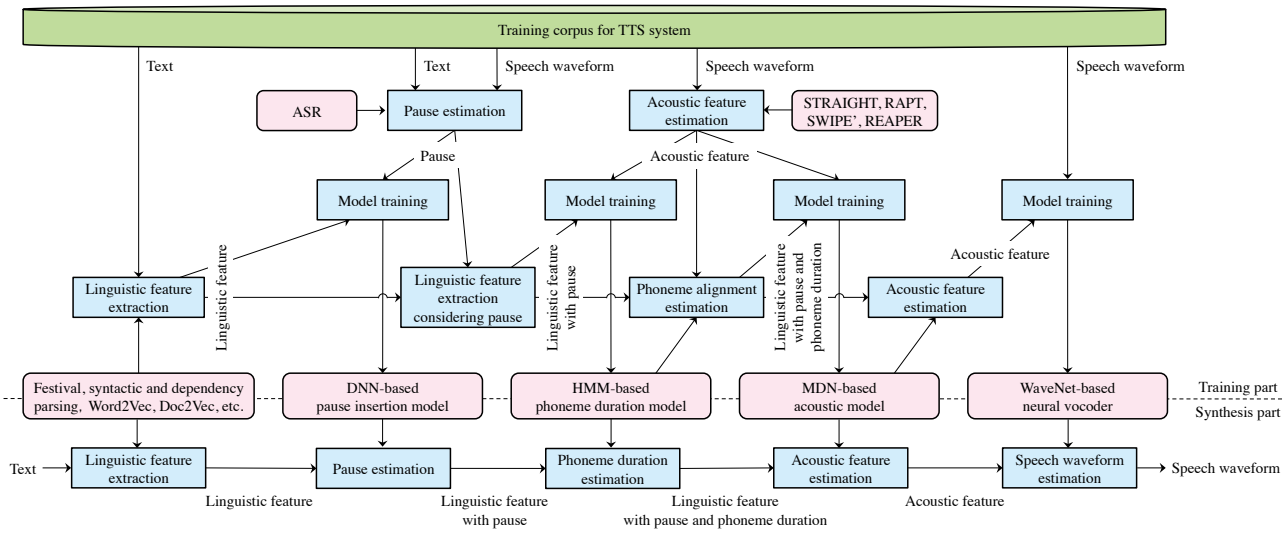

Figure 1: *Overview of the NITech TTS system*

ter trajectory generation process considering the global variance (GV) into the training of mixture density network (MDN)-based acoustic models [8]. The last year's NITech system showed good performance in terms of naturalness and intelligibility.

2.2. NITech TTS system for Blizzard Challenge 2018

Figure 1 shows an overview of the NITech TTS system for the Blizzard Challenge 2018. In addition to the NITech 2016 and 2017 TTS systems, the NITech 2018 TTS system is introduced pause insertion model and WaveNet-based neural vocoder.

2.2.1. Pause insertion model

Pause insertion to the proper position is necessary required to natural synthesized speech. Especially in audiobooks, since pause is used as one of emotional expressions, pause insertion is an important subject. In the NITech 2018 TTS system, introduce a pause insertion model to reproduce the pause insertion style of the training corpus.

Pauses included in the training corpus are predicted by using automatic speech recognition (ASR), i.e., phoneme alignment estimation. Phoneme alignments are estimated with a structure with a short pause at all word boundaries. A hidden Markov model (HMM) with state skip transitions is used as the short pause model. If the duration of an estimated short pause is equal to or greater than a threshold value, we assumed that the word boundary contains a pause.

Deep neural network (DNN) is used as pause insertion model. The input of DNN is linguistic features also used for the input of the MDN-based acoustic model. For linguistic features of pause insertion model, linguistic features of word- and sentence-level designed in the NITech 2017 TTS system [8] are used. The output of DNN is whether or not a pause is inserted after the word. Pause insertion style of the training corpus can be reproduced by using the pause insertion model.

2.2.2. WaveNet-based neural vocoder

Although vocoders such as STRAIGHT [12] can easily produce synthetic speech from acoustic features, they inevitably introduce degradation in speech quality. To solve this problem, the WaveNet generative model [13] is used as the vocoder [14] of the NITech 2018 TTS system. The relationship between the acoustic features predicted by the MDN-based acoustic model and the corresponding waveform samples is modeled by the WaveNet. The two models are trained independently.

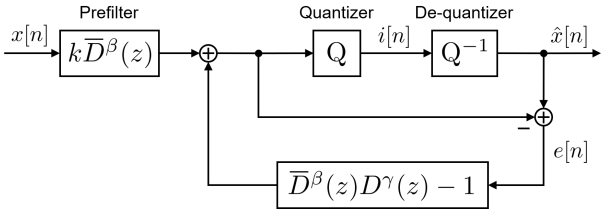

Figure 2: *Block diagram of quantization noise shaping with prefiltering*

One of the key techniques of WaveNet is modeling speech signals composed of a set of discrete values instead of continuous ones. This enables flexible waveform modeling because a categorical distribution has no assumptions about the shape. A nonlinear quantization with μ-law companding is typically used to obtain the discrete valued speech signals. However, the quantization scheme introduces white noise into the original signals, resulting in the degradation of speech quality. To overcome this problem, the mel-cepstrum-based quantization noise shaping [15] is used. The basic idea is to apply time-variant mask derived from mel-cepstrum to the quantization noise. Since mel-cepstrum can be based on the human auditory system, some of the quantization noise should be difficult for a human listener to perceive. Furthermore, the mel-cepstrum-based prefilter that emphasizes formants is applied to speech signals. This is known as the postfilter in speech coding [16].

Figure 2 shows the block diagram of the mel-cepstrum-based quantization noise shaping with prefiltering. In the figure, $D(z)$ is a minimum phase transfer function derived from the following spectral envelope model using M^{th} order mel-

cepstral coefficients $\{\tilde{c}(m)\}$:

$$H(z) = \exp \sum_{m=0}^{M} \tilde{c}(m)\, \tilde{z}^{-m} = K \cdot D(z), \quad (1)$$

where

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \quad (2)$$

and K is the gain factor. The phase characteristic of the all-pass function $\tilde{z}^{-1}$ can approximate the mel-frequency scale by tuning α. In the figure, $D^{\gamma}(z)$ and $\bar{D}^{\beta}(z)$ are the filters for noise shaping and prefiltering, respectively. A time-variant variable k normalizes the power of the prefilter output signal. The z-transform of the reconstructed speech sample $\hat{x}[n]$ is represented as

$$\hat{X}(z) = \{kX(z) + D^{\gamma}(z)E(z)\}\bar{D}^{\beta}(z), \quad (3)$$

where $X(z)$ and $E(z)$ are the z-transforms of $x[n]$ and $e[n] = \hat{x}[n] - x[n]$, respectively. It can be seen from Eq. (3) that the noise spectrum $E(z)$ is shaped by the noise shaping filter $D^{\gamma}(z)$. Tunable parameters γ and β ($0 \leq \gamma, \beta \leq 1$) control the effects of noise shaping and prefiltering. The case in which $\gamma = 0$ and $\beta = 0$ corresponds to conventional quantization.

By following one of the ways in proposed in [17], we can derive

$$K = \exp b(0), \quad (4)$$

$$D(z) = \exp \sum_{m=1}^{M} b(m)\, \Phi_m(z), \quad (5)$$

where

$$b(m) = \begin{cases} \tilde{c}(M), & (m = M) \\ \tilde{c}(m) - \alpha b(m+1), & (m < M) \end{cases} \quad (6)$$

$$\Phi_m(z) = \begin{cases} \dfrac{(1-\alpha^2)z^{-1}}{1 - \alpha z^{-1}} \tilde{z}^{-(m-1)}, & (m > 0) \\ 1. & (m = 0) \end{cases} \quad (7)$$

The prefiltering filter $\bar{D}(z)$ is represented as

$$\bar{D}(z) = \exp \sum_{m=1}^{M} \bar{b}(m)\, \Phi_m(z), \quad (8)$$

where

$$\bar{b}(m) = \begin{cases} b(m), & (m > 1) \\ -\alpha b(2). & (m = 1) \end{cases} \quad (9)$$

The reason why $\tilde{c}(1)$ is set to zero in Eq. (9) is to avoid the change of overall slope of spectral envelope, i.e., voice characteristics. It should be note that the mel-cepstrum-based quantization noise shaping with prefiltering can be viewed as the preprocessing of WaveNet, i.e., there is no additional computational cost in the synthesis stage.

3. Blizzard Challenge 2018 evaluation

3.1. Training corpus construction conditions

The collection of provided children's audiobooks consisted of 56 books with a total 1258 pages. An ASR was trained to construct a training corpus for SPSS [7]. The CMU Pronouncing Dictionary [18] and the WSJ0, WSJ1 [19], and TIMIT [20] databases were used to train the ASR. Speech signals were sampled at a rate of 16 kHz and windowed by a 25-ms hamming window with a 10-ms shift. The acoustic-feature vector consisted of 39 components composed of 12-dimensional mel-frequency cepstral coefficients (MFCCs) including the energy with the first- and second-order derivatives. A three-state left-to-right GMM-HMM without skip transitions was used. The trained GMMs had 32 mixtures for pause and 16 mixtures for the other phonemes. A tri-gram LM was created based on the text of the provided children's audiobooks. The HTK [21] and SRILM [22] were used to construct the ASR. The training recipe was the same as that of the HTK Wall Street Journal Training Recipe [23]. Thresholds of word-match accuracy for adaptation and training corpora were set to 90% [7]. After pruning, the training corpus for SPSS consisted of 924 pages.

3.2. TTS system construction conditions

Linguistic features [8] were extracted using Festival [24], Stanford Parser [25], SyntaxNet [26], and gensim [27]. The speech signals were sampled at a rate of 32 kHz and windowed with a fundamental frequency (F_0)-adaptive Gaussian window with a 5-ms shift. Acoustic features were composed of 64-dimension STRAIGHT [28] mel-cepstral coefficients including the 0th coefficient, F_0, and 32-dimension mel-cepstral analysis aperiodicity measures. Voting results concerning F_0 (estimated by using RAPT [29], SWIPE' [30], and REAPER [31] tools) were taken as F_0 of acoustic features.

In the pause insertion model, the input feature was a 251-dimensional linguistic feature vector extracted from text which were normalized to be within 0.0–1.0 based on their minimum and maximum values in the training data. The architecture of the DNN was bi-directional gated recurrent unit (GRU) with three hidden layers which had 128 units per layer. For training the bi-directional GRU, a adaptive moment estimation (Adam) algorithm and dropout with a probability of 0.2 were used.

The HMM-based phoneme duration model was constructed to estimate phoneme-level alignments for training and phoneme duration for synthesis. In addition to static features, first- and second-order derivatives of static features were used for acoustic features. A five-state left-to-right context-dependent multi-stream multi-space probability distribution hidden semi-Markov model (MSD-HSMM) [32, 33, 34, 35] without skip transitions was used as the acoustic model. Each state output probability distribution was composed of a spectrum, F_0, and aperiodicity streams. The spectrum and aperiodicity streams were modeled using single multi-variate Gaussian distributions with diagonal covariance matrices. The F_0 stream was modeled using an MSD consisting of a Gaussian distribution for voiced frames and a discrete distribution for unvoiced frames. State durations were modeled using a Gaussian distribution. The HTS [36] and SPTK [37] were used for constructing the HMM-based phoneme duration model.

In the MDN-based acoustic model, the input feature was a 1685-dimensional feature vector consisting of 925 linguistic features including binary features and numerical features for contexts, 10 duration features, 150-dimensional word code, and 600-dimensional phrase code. Fix-dimensional normally distributed random vector was used as word and phrase codes, and pre-trained word2vec and doc2vec were used to measure word and phrase similarity. The output feature was a 98-dimensional feature vector consisting of STRAIGHT mel-cepstral coefficients, F_0 acquired by linearly interpolating values in unvoiced parts, voiced/unvoiced binary value, and mel-cepstral analysis aperiodicity measures. The input features were normalized to be within 0.0–1.0 based on their minimum and maximum values in the training data, and the output features were normalized to have zero-mean unit-variance. The input and output

Table 1: *Evaluation results*

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

features were time-aligned frame-by-frame by using the trained MSD-HSMM. A single MDN, which models spectral, excitation, and aperiodicity parameters, was trained. The architecture of the MDNs was three hidden layers with 8000 units per layer. The sigmoid activation function was used in the hidden layers and the linear activation function was used in the output layer. For training the MDNs, a mini-batch stochastic gradient descent (SGD)-based back-propagation algorithm and dropout with a probability of 0.6 were used. The GV weight w was set to 0.001 [11].

Speech signals at 32 kHz were quantized 8-bit using μ-law compression for WaveNet neural vocoder training. The tunable parameters of the mel-cepstrum-based quantization noise shaping were set to $\gamma = 0.1$ and $\beta = 0.1$. The dilations of the WaveNet model were set to $1, 2, 4, \ldots, 512$. The 10 dilation layers were stacked three times, resulting in a receptive field with a size of 3072. The channel size for dilation, residual block, and skip connection were 256, respectively [13]. The auxiliary features were used 98-dimensional acoustic features generated the MDN-based acoustic model. In order to satisfy the rule of Blizzard Challenge 2018, synthesized speech waveforms at 32 kHz were upsampled to 48 kHz.

3.3. Experimental conditions of listening test

Large-scale subjective listening tests were conducted by the Blizzard Challenge 2018 organization. The listeners included paid participants, speech experts, and volunteers. The paid participants (native speakers of English) took the test in soundproof listening booths using high-quality headphones. The speech experts and volunteers included non-native speakers of English.

To evaluate the page domain of a children's book, 7-page-domain-criteria 60-point mean opinion score (MOS) tests were conducted. The terms in the parentheses were used to label the points 10 for "bad" and 50 for "excellent" on the scale. Listeners listened to one whole page from a children's book and chose a scored from 1 to 60 based on the following 7-page-domain-criteria.

- overall impression (OI): "bad" to "excellent"
- pleasantness (PL): "very unpleasant" to "very pleasant"
- speech pauses (SP): "speech pauses confusing/unpleasant" to "speech pauses appropriate/pleasant"
- stress (ST): "stress unnatural/confusing" to "stress natural"
- intonation (IN): "melody did not fit the sentence type" to "melody fitted the sentence type"
- emotion (EM): "no expression of emotions" to "authentic expression of emotions"
- listening effort (LE): "very exhausting" to "very easy"

To evaluate the sentence domain of children's book, 2-sentence-domain-criteria 5-point MOS tests were conducted. Listeners listened to one sample and chose a scored from 1 to 5 based on the following 2-sentence-domain-criteria.

- naturalness (NAT): "completely unnatural" to "completely natural"
- similarity (SIM): "sounds like a totally different person" to "sounds like exactly the same person"

To evaluate intelligibility, the participants were asked to transcribe semantically unpredictable sentences (SUS) by typing in the sentence they heard. The average word error rate (WER) was calculated from these transcripts.

3.4. Experimental results

Table 1 lists the MOSs (means and standard deviations) of the listening test results from the all listeners and WER of the listening test from paid listeners. Systems A, B, C, D, E and ***I*** represent the following systems.

- A: natural speech
- B: unit-selection benchmark system
- C: HMM benchmark system
- D: DNN benchmark system
- E: DNN benchmark system 2
- ***I***: NITech system

The ordering of systems is in descending order of NAT. Wilcoxon's signed rank tests were used to determine significance difference [38]. In Table 1, asterisk * means a statistically significant difference between system ***I*** and other systems.

From Table 1, our system ***I*** achieved good performance for page-domain-criteria and sentence-domain-criteria. Moreover, our system ***I*** achieved the lowest WER. These results suggest that, the NITech 2018 TTS system was able to synthesize speech waveform with high naturalness, speaker similarity, and intelligibility.

We consider the difference in synthesized speech between the NITech 2017[1] and 2018[2] TTS systems. The introduction of WaveNet neural vocoder was able to improve naturalness and speaker similarity by avoiding degradation of speech quality accompanying use of a frame-level processing vocoder. However, the prediction accuracy of the WaveNet neural vocoder was not sufficient, and the pronunciation of synthesized speech was sometimes ambiguous. Speech data provided as the training corpus was compressed by multiple codecs, and this complicated the training WaveNet which models speech waveform directly. Additionally, although the training corpus contains various expressive speech, it was not enough for the amount of training data to model various speaking styles. In actuality, when inputting acoustic features generated from the MDN acoustic model capable of synthesizing expressive speech into the WaveNet neural vocoder, pronunciation of synthesized speech became ambiguous in many cases. Therefore, the future work is to synthesize expressive speech by WaveNet neural vocoder.

4. Conclusion

We described the Nagoya Institute of Technology (NITech) text-to-speech (TTS) system for the Blizzard Challenge 2018. Deep neural network (DNN)-based pause insertion model and WaveNet-based neural vocoder were introduced the NITech 2018 TTS system. Large-scale subjective evaluation results show that the NITech 2018 TTS system was able to synthesize speech waveform with high naturalness, speaker similarity, and intelligibility. Future work includes generating expressive synthesized speech in WaveNet neural vocoder and introducing end-to-end approach to simplify the structure of complex TTS system.

5. Acknowledgements

The MIC/SCOPE #162106106 and JSPS Grant-in-Aid for Scientific Research (A) 18H04128.

6. References

[1] A. W. Black and K. Tokuda, "The Blizzard Challenge – 2005: Evaluating corpus-based speech synthesis on common datasets," *Interspeech 2005*, pp. 77–80, 2005.

[2] S. King, "Measuring a decade of progress in text-to-speech," *Loquens*, vol. 1, no. 1, 2014.

[3] S. King and V. Karaiskos, "The blizzard challenge 2013," *Blizzard Challenge 2013 Workshop*, 2013.

[4] ——, "The blizzard challenge 2016," *Blizzard Challenge 2016 Workshop*, 2016.

[5] S. King, L. Wihlborg, and W. Guo, "The blizzard challenge 2017," *Blizzard Challenge 2017 Workshop*, 2017.

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

[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

[7] K. Sawada, C. Asai, K. Hashimoto, K. Oura, and K. Tokuda, "The NITech text-to-speech system for the Blizzard Challenge 2016," *Blizzard Challenge 2016 Workshop*, 2016.

[8] K. Sawada, K. Hashimoto, K. Oura, and K. Tokuda, "The NITech text-to-speech system for the Blizzard Challenge 2017," *Blizzard Challenge 2017 Workshop*, 2017.

[9] C. M. Bishop, "Mixture density networks," *Aston University*, 1994.

[10] H. Zen and A. Senior, "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," *2014 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3844–3848, 2014.

[11] K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Trajectory training considering global variance for speech synthesis based on neural networks," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5600–5604, 2016.

[12] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight," *MAVEBA 2001*, pp. 13–15, 2001.

[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," *arXiv:1609.03499*, 2016.

[14] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder," *Interspeech 2017*, pp. 1118–1122, 2017.

[15] T. Yoshimura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "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.

[16] K. Tokuda, H. Matsumura, T. Kobayashi, and S. Imai, "Speech coding based on adaptive mel-cepstral analysis," *1994 IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 197–200, 1994.

[17] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectrum approximation (MLSA) filter for speech synthesis," *Electronics and Communications in Japan*, vol. 66, no. 2, pp. 11–18, 1983.

[18] "CMU Pronouncing Dictionary," http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[19] D. B. Paul and J. M. Baker, "The design for the wall street journal-based CSR corpus," *The workshop on Speech and Natural Language*, pp. 357–362, 1992.

[20] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT: acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[21] "HTK," http://htk.eng.cam.ac.uk/.

[22] "SRILM," http://www.speech.sri.com/projects/srilm.

[23] K. Vertanen, "Baseline WSJ acoustic models for HTK and Sphinx: Training recipes and recognition experiments," *Cavendish Laboratory*, 2006.

[24] "Festival," http://www.festvox.org/festival/.

[25] "Stanford Parser," http://nlp.stanford.edu/software/lex-parser.shtml.

[26] D. Andor, C. Alberti, D. Weiss, A. Severyn, A. Presta, K. Ganchev, S. Petrov, and M. Collins, "Globally normalized transition-based neural networks," *arXiv:1603.06042*, 2016.

[27] R. Rehurek and P. Sojka, "Software framework for topic modelling with large corpora," *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50, 2010.

[28] 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.

[29] D. Talkin, “A robust algorithm for pitch tracking (RAPT),” *Speech Coding and Synthesis*, pp. 495–518, 1995.

[30] A. Camacho, “SWIPE: a sawtooth waveform inspired pitch estimator for speech and music,” *Ph.D. Thesis, University of Florida*, 2007.

[31] “REAPER,” https://github.com/google/REAPER.

[32] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Hidden semi-Markov model based speech synthesis,” *8th International Conference on Spoken Language Processing*, pp. 1185–1180, 2004.

[33] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis,” *Eurospeech 1999*, pp. 2347–2350, 1999.

[34] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” *2000 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 936–939, 2000.

[35] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, “Multi-space probability distribution HMM,” *IEICE Transactions on Information & Systems*, vol. E85-D, no. 3, pp. 455–464, 2002.

[36] “HTS,” http://hts.sp.nitech.ac.jp/.

[37] “SPTK,” http://sp-tk.sourceforge.net/.

[38] R. A. J. Clark, M. Podsiadlo, M. Fraser, C. Mayo, and S. King, “Statistical analysis of the blizzard challenge 2007 listening test results,” *Blizzard Challenge 2007 Workshop*, 2007.