THE BLIZZARD MACHINE LEARNING CHALLENGE 2017

Kei Sawada[1]*, Keiichi Tokuda*[1]*, Simon King*[2]*, Alan W Black*[3]

[1]Department of Computer Science, Nagoya Institute of Technology, Nagoya, JAPAN
[2]Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK
[3]Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA
blizzard@festvox.org www.synsig.org/index.php/Blizzard_Challenge

ABSTRACT

This paper describes the Blizzard Machine Learning Challenge (BMLC) 2017, which is a spin-off of the Blizzard Challenge. The annual Blizzard Challenges 2005–2017 were held to better understand and compare research techniques in building corpus-based text-to-speech (TTS) systems on the same data. The series of Blizzard Challenges has helped us measure progress in TTS technology. However, to get competitive performance, a lot time has to be spent on skilled tasks. This may make the Blizzard Challenge unattractive to machine learning researchers from other fields. Therefore, we recommend that the BMLC not involve these speech-specific tasks and that it allow participants to concentrate on the acoustic modeling task, framed as a straightforward machine learning problem, with a fixed dataset. In the BMLC 2017, two types of datasets consisting of four hours of speech data suitable for machine learning problems were distributed. This paper summarizes the purpose, design, and whole process of the challenge and its results.

Index Terms— Speech synthesis, machine learning, evaluation, listening test, Blizzard Challenge

1. INTRODUCTION

A text-to-speech (TTS) system generates intelligible, natural-sounding artificial speech for a given input text. Because computing performance has steadily improved, TTS systems have evolved from "rule-based" systems, which connect speech units adjusted manually, to "corpus-based" systems like ones for unit-selection synthesis, which selectively connect suitable speech units extracted from a large-scale speech database [1]. However, the unit-selection-based speech synthesis system restricts the output speech to the same style as that in the original recordings because no modifications to the selected pieces of recorded speech are normally done. Therefore, a unit-selection-based speech synthesis system that can generate a huge variety of high-quality speech can only be constructed if a large-scale speech database containing high-quality speech is built. Statistical speech synthesis based on machine learning has been drawing attention as a means of achieving it [2].

Statistical speech synthesis is a means of "mapping" (i.e., representing a map) of speech waveforms from text on the basis of a statistical model. However, a statistical model for directly predicting a speech waveform from text is difficult to construct. Accordingly, for conventional statistical speech synthesis, mapping a speech waveform from text can be divided into three steps: (i) estimating linguistic features expressed as phonemes, parts of speech, words, etc. from text (called "text analysis"); (ii) estimating acoustic features, which express characteristics of a speech waveform, from linguistic features; and (iii) generating a speech waveform from acoustic features. For step (ii), the process is referred to as "acoustic modeling," namely, predicting acoustic features from linguistic features. Good examples of architectures suitable for modeling time-series data are available for acoustic modeling, and efficient training algorithms have been developed. For those reasons, hidden Markov models (HMMs) are widely utilized, and statistical speech synthesis based on HMMs (called "HMM-based speech synthesis") have become widely used as a standard speech-synthesis technique [3].

In various fields, methods utilizing "deep learning" have demonstrated high performance. In the field of statistical speech synthesis, the quality of synthesized speech can reportedly be improved by switching the acoustic model from a HMM to a "deep neural network" (DNN) [4, 5, 6, 7]. Acoustic models that can directly generate speech waveforms from linguistic features (such as WaveNet [8] and SampleRNN [9]) have been proposed—in other words—an integration of steps (ii) and (iii). A method for acoustic modeling a speech waveform without using acoustic features is available; accordingly, degradation in speech quality can be avoided using a vocoder, which represents one of the bottlenecks in the flow of HMM-based speech synthesis. Furthermore, investigations on "end-to-end styles," which predict speech waveforms directly from text, have started [10, 11, 12, 13]. These methods based on deep learning have been implemented by taking advantage of the knowledge of not only researchers in the field of speech synthesis but also researchers in the field of machine learning. From now on, the quality of synthesized speech could conceivably be improved by applying technologies in the field of deep learning. However, constructing a TTS system requires having speech-specific tasks (such as updating the lexicon, removing inappropriate audio files, segmenting and aligning audio files, detecting alignment errors, etc.), and that necessity might discourage machine learning researchers from moving into this field.

The annual Blizzard Challenges 2005–2017 were held to better understand and compare research techniques in building corpus-based TTS systems on the same data [14]. At annual Blizzard Challenges held up until now, challenges using English, Mandarin, Indian languages, and audiobooks were set as training data. The series of Blizzard Challenges has helped us measure progress in TTS technology [15]. However, participation in Blizzard Challenges up until now has required a working knowledge of the speech-synthesis field. Therefore, we recommend that the Blizzard Machine Learning Challenge (BMLC) not involve these speech-specific tasks and that it allow participants to concentrate on the acoustic modeling task, framed as a straightforward machine learning problem, with a fixed dataset. The BMLC aims to encourage development in the speech-synthesis field and to promote entry of new researchers from the field of machine learning into that field. At BMLC 2017 [16], data concerning children's audiobooks distributed at Blizzard Chal-

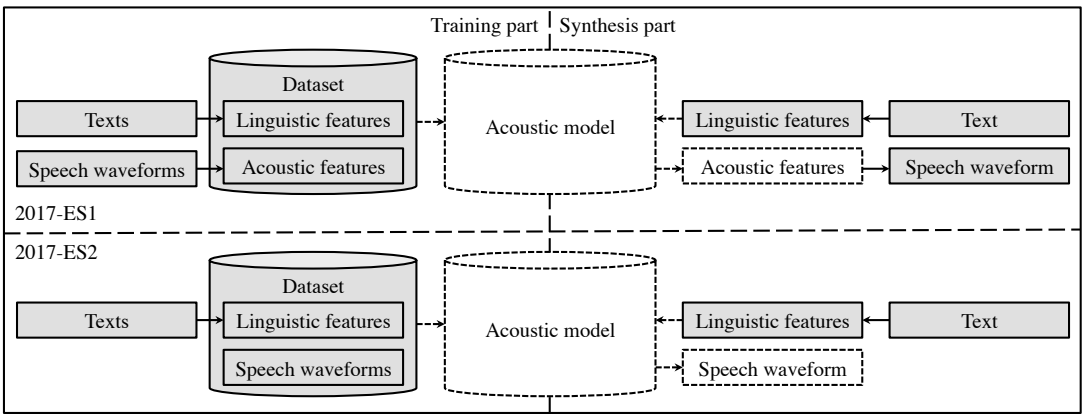

Fig. 1. Tasks of the BMLC 2017. Solid boxes and arrows represent the processes of the organizers, and dashed boxes and arrows represent the processes of participants. The speech-specific tasks were done by the organizers. Participants could concentrate on the acoustic modeling task.

lenge 2016 [17] were pre-processed so as to be suitable for machine learning problems and for distribution to participants. The participants constructed acoustic models based on rules that were judged to be impartial from the viewpoints of both speech-synthesis researchers and machine-learning researchers. The quality of synthesized speech was also evaluated by the organizers of the challenge using subjective-evaluation tests.

The rest of this paper is organized as follows. Section 2 describes the tasks and datasets for the BMLC 2017. Section 3 presents the benchmark and submitted systems. The listening tests conditions and results are given in Section 4, followed by discussion and future plans described in Section 5. Concluding remarks are presented in Section 6.

2. BLIZZARD MACHINE LEARNING CHALLENGE

2.1. Tasks

This first version of the BMLC is posed as a pure and simple machine learning problem. The rules are intended to create a level playing field between entries for both speech synthesis experts and machine learning experts alike. Therefore, expert interventions into the provided data are not allowed. In the BMLC 2017, the two English spoke (ES) tasks were organized in the form of 2017-ES1 and 2017-ES2.

2017-ES1: prediction of acoustic features from linguistic features. Frame-level sequence pairs of acoustic features and linguistic features are distributed by the organizers. Participants must train a model to predict acoustic features from linguistic features.

2017-ES2: prediction of speech waveforms from linguistic features. Pairs of speech waveforms and linguistic features are distributed by the organizers. Participants must train a model to predict speech waveforms directly from linguistic features.

Figure 1 summarizes the tasks of the BMLC 2017. The main guidelines to participate with an entry were as follows:

- Participants are not allowed to submit both 2017-ES1 and 2017-ES2 tasks, thereby enabling the number of participants to be controlled.
- Participants are not allowed to use external data in any way. External data are defined as data of any type that are not part of the provided dataset.
- Participants may automatically apply transforms such as normalization or quantization to any of the provided data.
- Participants are not allowed to add and remove the linguistic features.
- In the 2017-ES2 task, participants must not extract additional features, e.g, F_0 and cepstrum, from the speech waveforms.
- In the 2017-ES2 task, quantizing and downsampling the speech waveforms are permitted.
- Participants need to predict acoustic features (2017-ES1) or speech waveforms (2017-ES2) for a test set of previously-unseen linguistic features.

2.2. Datasets

The data were provided by Usborne Publishing Ltd. and are from a commercial product range of children's audiobooks. All speech data were recorded by one native British English female professional speaker. Around five hours of material was made for last year's Blizzard Challenge 2016. Each of the 50 books is rated by Usborne for reading ages (mainly for children 4, 5, or 6 years old, with a handful of books rated as "18 months+"). Genres include classic children's stories (e.g., The Three Little Pigs), simplified and abridged versions of Shakespeare (e.g., Romeo and Juliet), and factual books (e.g., Knights and Castles). A sentence-level segmentation was created by Toshiba's Cambridge Research Laboratory and Innoetics. At BMLC 2017, datasets 2017-ES1 and 2017-ES2, which are suited to machine learning for statistical speech synthesis using the Blizzard Challenge 2016 data, were prepared.

2.2.1. Waveforms

Speech data were provided as files of sentence units under a supposed sampling frequency of 44.1 kHz, quantization of 16 bits, and monaural Waveform Audio File Format (WAVE). The lengths of silence at the beginning and end of the speech waveform were adjusted

to suitable values to model silence of the appropriate length. Moreover, small noise values were added in the case of a great number of successive "0"s to enable estimation of acoustic features.

2.2.2. Acoustic features

For general speech synthesis based on DNNs, DNNs are applied to regression problems (namely, predicting acoustic features from linguistic features). However, estimating acoustic features from a speech waveform requires specialist knowledge of the speech-synthesis field. Accordingly, acoustic features were prepared in advance by the organizers for the 2017-ES1 task. Acoustic features were assumed using log fundamental frequency (F_0), mel-cepstral coefficients [18], and mel-cepstral analysis aperiodicity measures extracted from a speech waveform every 4.989 ms (220 samples / 44,100 samples). Voting results concerning F_0 (estimated using RAPT [19], SWIPE' [20], and REAPER [21]) were taken as F_0 of acoustic features. Values of voiced parts were linearly interpolated for unvoiced parts of F_0. Information concerning voiced and unvoiced parts was respectively expressed as "1" and "0." A spectral envelope and aperiodicity were estimated using WORLD [22]. The voting F_0 was used in that estimation. The spectral envelope and aperiodicity were converted to 50- and 25-order mel-cepstral coefficients using SPTK [23]. The acoustic features were composed of a total of 77 dimensions, namely, log F_0 (acquired by linearly interpolating values in unvoiced parts), the voiced and unvoiced information, and the 50- and 25-order mel-cepstral coefficients. Binary files for Linux and macOS were distributed to allow the participants to generate speech waveforms from the acoustic features.

2.2.3. Linguistic features

Specialist knowledge in the speech-synthesis field is needed for both linguistic features and acoustic features. Accordingly, linguistic features estimated by the organizers were provided. First, text was analyzed by Festival [24] using the CMU Pronouncing Dictionary [25]. The results of the text analysis were converted to linguistic features of HTS-2.3.1 demo script format [26]. The HTS-2.3.1 demo script format includes linguistic features that depend on the state of a hidden Markov model (HMM), i.e., the position of the current HMM state in the phoneme and the position of the current frame in the HMM state. Linguistic features depending on the state of a HMM were removed to reduce HMM dependency. The resulting linguistic features were composed of a total of 687 dimensions. The linguistic features were normalized to be within 0.0–1.0 based on their minimum and maximum values in the training data. The order of features was randomized to prevent expert TTS researchers from trying to reverse engineer them.

The speech waveforms, acoustic features, and linguistic features consisted of sample-level, frame-level, and phoneme-level, respectively, with different time-levels. A general DNN-based speech-synthesis system is unable to train a DNN using different time-level feature sequences. Accordingly, in 2017-ES1, acoustic feature sequences in the frame level must have a corresponding relationship with a linguistic feature sequence in the phoneme level. The correspondence between the acoustic feature sequence in the frame level and the linguistic feature sequence in the phoneme level (i.e., phoneme alignment) was estimated, and linguistic features in the frame level were thereby provided for both tasks. In 2017-ES2, because the frame shift had a fixed length (4.989 ms), the linguistic feature sequence could be easily converted from the frame level to the sample level. The HMM-based acoustic model was used to estimate phoneme alignment of acoustic features. A multi-stream multi-space probability distribution hidden semi-Markov model (MSD-HSMM) [27, 28, 29, 30, 31] with five states and left-to-right context dependency and 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 was used for constructing the HMM-based acoustic model [26]. The linguistic and acoustic features were time-aligned frame-by-frame by using the trained full-context MSD-HSMM. Moreover, the trained MSD-HSMM was used to predict the phoneme duration of the development set and the test set.

2.2.4. Data pruning

The children's audiobooks used as training data were created for commercial purposes. The data prepared for statistical-model training were not ideal, e.g., the training data contained mismatches between speech waveform and text. These mismatches were caused by the misreading of a text or words that do not exist in the text, i.e., description of a book or onomatopoeia. Furthermore, because the children's audiobooks were read emphatically, emotionally, etc., a lot of expressive speech data were included. Phoneme-alignment errors are easily generated for this kind of speech data. Moreover, such errors have a negative effect on statistical-model training. Accordingly, phoneme alignment was estimated using a monophone MSD-HSMM, and specific speech data were removed from the dataset: speech data in which the number of phonemes with a significant difference between the estimated duration of phonemes and the average duration of phonemes is greater than or equal to a threshold value. About four hours of speech data (4651 files) were used as the datasets for BMLC 2017.

3. SYSTEMS

Seven teams registered, and three teams (CMU [32], iFLYTEK Research [33], and USTC-NELSLIP [34]) submitted systems. Accordingly, the BMLC 2017 had three submitted systems along with three benchmark systems. Each system is briefly explained as follows.

3.1. 2017-ES1 systems

3.1.1. Benchmark system 1

The first benchmark system used a feed-forward neural network (FFNN) using simple frame-by-frame training [4]. The delta features for the acoustic feature are not used because they require knowledge of the speech-synthesis field. The HTS-2.3.1 demo script was used, and the FFNN was trained. The architecture of the FFNN was three hidden layers with 1024 units per layer. The sigmoid activation function was used in the hidden layers, and the linear activation function was used in the output layer. The Adam algorithm [35] was repeatedly executed as the training algorithm for 50 epochs, and the dropout ratio was taken as 0.5. The architecture and training algorithm was set on the basis of default values of the HTS-2.3.1 demo script.

3.1.2. Benchmark system 2

The second benchmark system used a single-mixture density network consisting of a feed-forward neural network incorporating trajectory training with global variance (GV) [36, 37, 38]. The architecture of the single-mixture density network was three hidden layers with 2048 units per layer. The sigmoid activation function was used in the hidden layers, and the linear activation function was used in the output layer. The AdaGrad algorithm [39] was repeatedly executed as the training algorithm for 300 epochs, and the dropout ratio and GV weight were assumed to be 0.5 and 0.001, respectively. The output features were normalized to have zero-mean unit-variance.

3.1.3. System H

System H used a long short term memory (LSTM) embedding layer feeding a recurrent highway network [40] with a recurrence of six, followed by a linear output layer. Each hidden layer had 128 units. The Adam optimizer, L2 weight regularization, and Glorot weight initialization were used for training.

3.1.4. System I

System I was composed of two modules, an LSTM recurrent neural network (RNN)-based acoustic model and a generative adversarial network (GAN) [41] based post-filter for mel-cepstral. The first part had four hidden layers in this architecture stacked with a feed-forward layer and three bidirectional LSTM-RNN layers with 1024 units in each layer. The network was trained under a minimum mean square error criterion using the stochastic gradient descent (SGD) algorithm. The second part had a principal component analysis (PCA) based GAN post-filter. The details of mel-cepstral were removed using dimension reduction with PCA and recovered using a GAN.

3.2. 2017-ES2 systems

3.2.1. Benchmark system 3

The third benchmark system used the WaveNet [8]. The waveforms were down-sampled to 16 kHz and quantized to 8 bits using the μ-law algorithm. For the architecture of WaveNet, we repeated 10-layer dilation three times as $1, 2, 4, \ldots, 512$ (forming a dilated causal convolution layer with a total of 30 layers). The number of causal-convolution channels, residual channels, and skip channels were taken as 128, 256, and 256, respectively. The Adam algorithm was used as the training algorithm.

3.2.2. System G

System G used the WaveNet and bidirectional LSTM. The waveforms were down-sampled to 24 kHz and quantized to 10 bits using the μ-law algorithm. WaveNet with three blocks, each with dilations $1, 2, 4, \ldots, 512$ was used. The Adam algorithm was used as the training algorithm. WaveNet conditioned on the short-time Fourier transform (STFT) amplitude spectra of waveforms was submitted. A bidirectional LSTM-based network was also trained to predict STFT amplitude spectra from linguistic features.

4. EVALUATION

4.1. Listening test

A subjective listening test was conducted by the BMLC organizers. Participants were asked to synthesize many hundreds of test sentences, of which only a small subset were used in the listening test. This provided a large amount of material that might be used in future listening tests and also prevented participants from manually intervening in synthesis. The evaluation combined the entries for 2017-ES1 and 2017-ES2 into the single listening test. The listening test had the following structure, comprising 11 sections each with seven stimuli—or six in the case of intelligibility—because no natural recorded semantically unpredictable sentences (SUS) were available.

Section1–2: A 5-point mean opinion score (MOS) test was conducted to evaluate speaker similarity. In each test, listeners could play four reference samples of the original speaker and one synthetic sample. They chose a response that represented how similar the synthetic speech sounded to the speech in the reference samples on a scale from 1 "sounds like a totally different person" to 5 "sounds like exactly the same person."

Section3–7: A 5-point MOS test was conducted to evaluate naturalness. In each test, listeners heard one sample and chose a score that represented how natural or unnatural the sentence sounded on a scale of 1 "completely unnatural" to 5 "completely natural."

Section8–11: To evaluate intelligibility, the listeners were asked to transcribe SUS by typing in the sentence they heard. Listeners were allowed to listen to each sentence only once. The average word error rate (WER) was calculated from these transcripts.

The organizers used a total of 50 paid native listeners, and each listener heard one sentence from each system per section. The listeners took the test in soundproof listening booths using high-quality headphones.

4.2. Results

Figures 2, 3, and 4 show the MOS for naturalness, the MOS for speaker similarity, and the WER of SUS, respectively. In these figures, the identifying letters represent the following systems.

A: natural speech.

X: 2017-ES1 benchmark system 1.

Y: 2017-ES1 benchmark system 2.

Z: 2017-ES2 benchmark system 3.

H and I: 2017-ES1 submitted systems.

G: 2017-ES2 submitted system.

The plots are color-coded: blue for natural speech, green for 2017-ES1 systems, and red for 2017-ES2 systems. In Figures 2 and 3, standard boxplots are presented for the ordinal data, where the median is represented by a solid bar across a box showing the quartiles; whiskers extend to 1.5 times the inter-quartile range, and outliers beyond this are represented as circles. In Figure 4, bar charts are presented for the WER interval data. A single ordering of the systems is used in all plots. This ordering is in descending order of MOS for naturalness. Note that the ordering is intended only to make the plots more readable using the same system ordering across all plots for both tasks and cannot be interpreted as a ranking. In other words, the ordering does not tell us which systems are significantly better than others. The numbers n at the bottom of the figures represent the number of evaluated sentences.

According to Figure 2, system G, which directly predicts the speech waveform, achieved a higher MOS (naturalness) than that

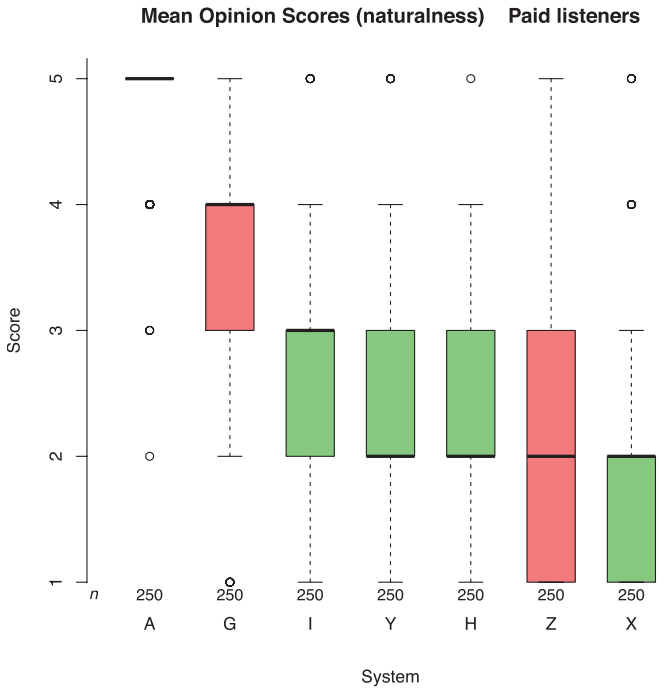

Fig. 2. MOS for naturalness

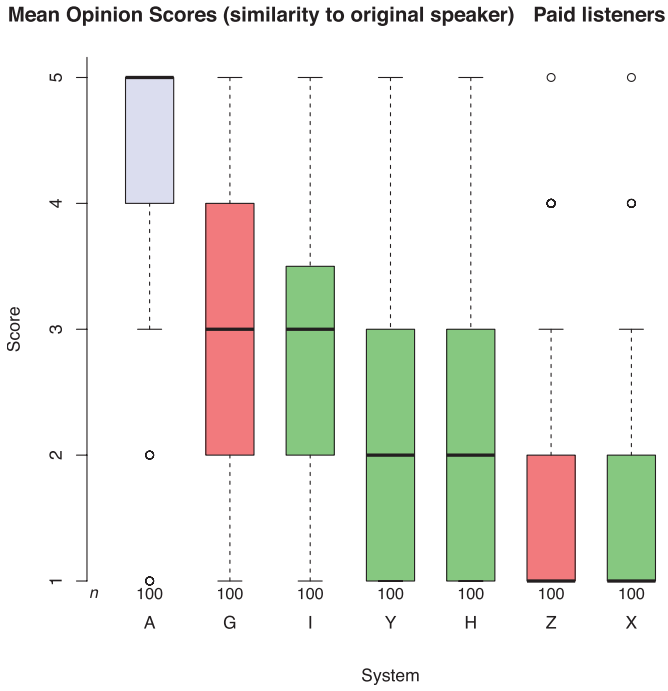

Fig. 3. MOS for speaker similarity

achieved by the other speech-synthesis systems. The method for directly predicting the speech waveform avoids degradation in speech quality by utilizing a vocoder; thus, that is likely to be the reason it achieves a high MOS for naturalness. However, benchmark system

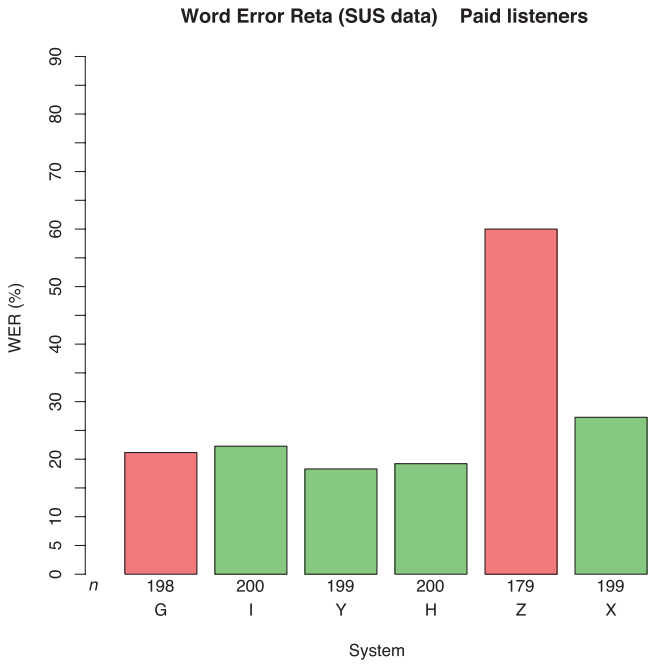

Fig. 4. WER of SUS

Z, which also directly predicts the speech waveform, does not attain a high MOS. The main difference between system G and system Z is whether or not the system uses acoustic information as intermediate representation. A combined training for not passing through acoustic intermediate representation is necessary to follow guidelines. This result indicates that training WaveNet with only linguistic features is more difficult than training with acoustic information. In addition, sufficient GPU resources and training time could not be secured for configuring benchmark system Z; consequently, the prediction accuracy of the model used was inadequate. For training with a model for directly predicting a speech waveform with high accuracy, secure abundant GPU resources, an efficient training algorithm, and adequate tuning are necessary.

According to Figure 3, compared to MOS for natural speech, MOS for "similarity with original speaker" attained by the speech-synthesis system is lower. Up until now, speech with low similarity with that of the original speaker has been cited as a weak point of statistical speech synthesis. Accordingly, a key challenge is to improve the similarity of speech with the original speaker attained using statistical speech synthesis. System I, which uses a GAN-based post-filter, achieved a higher MOS for speaker similarity than other 2017-ES1 systems (system H, X, and Y). This result suggests that the GAN-based post-filter probably improves speaker similarity.

According to Figure 4, compared to other systems, benchmark system Z obtained significantly low intelligibility (i.e., high WER). That is because the training of the statistical model is inadequate using only linguistic features and because the synthesized speech by that system was ambiguous. The low intelligibility conceivably influences the low "naturalness" and "similarity with original speaker" scores attained by benchmark system Z. Meanwhile, benchmark system Y, which uses trajectory training considering GV, achieved a lower WER than benchmark system X, which uses frame-by-frame training. This result demonstrates that trajectory training consider-

ing GV is useful for training the statistical models used in the DNN-based speech synthesis.

5. DISCUSSION AND FUTURE PLAN

At BMLC 2017, a challenge focused on a machine-learning problem was set for statistical speech synthesis. Seven teams registered, and three teams submitted systems. Although researchers in the machine-learning field registered for the challenge, they could not submit a system. Researchers in the machine-learning field might have discovered that speech synthesis is harder than they thought. Also, rules to which knowledge of the speech-synthesis field cannot be applied were imposed; subsequently, the speech-synthesis researchers expressed the opinion that the challenge imposed many restrictions. It means that the tasks were difficult due to limited training data (four hours), noisy training data, and limitations on knowledge of the speech-synthesis field. Some systems could not follow the rules. From now on, we plan to devise rules and tasks that allow researchers from both fields to participate more easily. Furthermore, we will endeavor to recruit machine learning researchers and to release benchmark systems in advance to easily participate. Statistical speech synthesis research is quickly moving onto an end-to-end style. Therefore, tasks for the end-to-end style are also planned.

The purpose of speech synthesis research is not only to synthesize high-quality speech. Building a framework for freely modeling and controlling speaking style, expression of emotion, and language is an important challenge facing research on speech synthesis. To achieve these purposes, we expect to develop speech-synthesis technology further by exploiting machine-learning techniques.

6. CONCLUSIONS

This paper presented Blizzard Machine Learning Challenge (BMLC) 2017. At the event, two types of dataset suitable for machine learning were provided for training acoustic models. According to the results of the subjective evaluations by the challenge organizers, the systems using state-of-the-art machine learning approaches achieved a higher "naturalness" score than that achieved by simple benchmark systems. In the future, rules will be devised to allow researchers from the fields of both machine learning and speech synthesis to participate more easily, and the BMLC will be further incorporated into the annual Blizzard Challenge.

7. ACKNOWLEDGEMENTS

We wish to thank a number of contributors without whom running the challenge would not have been possible. Rob Clark designed and implemented the statistical analysis; Dong Wang wrote the WER program. Tim Bunnell of the University of Delaware provided the tool to generate SUS sentences for English. The data were provided by Usborne Publishing Ltd. and are from their commercial product range of children's audiobooks. A sentence-level segmentation was created by Toshiba's Cambridge Research Laboratory and Innoetics. The authors would like to thank Keiichiro Oura, Kei Hashimoto, and Takenori Yoshimura for constructive discussions. The listening test scripts are based on earlier versions provided by previous organizers of the Blizzard Challenge. We offer thanks to all participants and listeners.

8. REFERENCES

[1] A. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 373–376, 1996.

[2] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, pp. 1039–1064, 2009.

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

[4] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," *2013 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 7962–7966, 2013.

[5] Z.-H. Ling, L. Deng, and D. Yu, "Modeling spectral envelopes using restricted Boltzmann machines for statistical parametric speech synthesis," *2013 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 7825–7829, 2013.

[6] S. Kang, X. Qian, and H. Meng, "Multi-distribution deep belief network for speech synthesis," *2013 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 8012–8016, 2013.

[7] R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory, "F_0 contour prediction with a deep belief network-Gaussian process hybrid model," *2013 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 6885–6889, 2013.

[8] 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.

[9] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, "SampleRNN: An unconditional end-to-end neural audio generation model," *arXiv:1612.07837*, 2016.

[10] W. Wang, S. Xu, and B. Xu, "First step towards end-to-end parametric TTS synthesis: Generating spectral parameters with neural attention," *Interspeech 2016*, pp. 2243–2247, 2016.

[11] J. Sotelo, S. Mehri, K. Kumar, J.F. Santos, K. Kastner, A. Courville, and Y. Bengio, "Char2Wav: End-to-end speech synthesis," *5th International Conference on Learning Representations*, 2017.

[12] S. O. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman, S. Sengupta, and M. Shoeybi, "Deep Voice: Real-time neural text-to-speech," *arXiv:1702.07825*, 2017.

[13] Y. Wang, RJ Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," *Interspeech 2017*, pp. 4006–4010, 2017.

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

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

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

[17] S. King and V. Karaiskos, "The Blizzard Challenge 2016," *Blizzard Challenge 2016 Workshop*, 2016.

[18] F. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," *1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 137–140, 1992.

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

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

[21] "REAPER," https://github.com/google/REAPER.

[22] 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, pp. 1877–1884, 2016.

[23] "SPTK," http://sp-tk.sourceforge.net/.

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

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

[26] "HTS," http://hts.sp.nitech.ac.jp/.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *Acoustical Science and Technology*, vol. 21, no. 2, pp. 76–86, 2000.

[32] P. Baljekar, S. K. Rallabandi, and A. W. Black, "The CMU entry to Blizzard Machine Learning Challenge," *2017 IEEE Automatic Speech Recognition and Understanding Workshop*, 2017.

[33] L.-J. Liu, C. Ding, Y.-J. Hu, Z.-H. Ling, Y. Jiang, M. Zhou, and S. Wei, "The USTC system for Blizzard Machine Learning Challenge 2017-ES2," *2017 IEEE Automatic Speech Recognition and Understanding Workshop*, 2017.

[34] Y.-J. Hu, L.-J. Liu, C. Ding, Z.-H. Ling, and L.-R. Dai, "The iFLYTEK system for Blizzard Machine Learning Challenge 2017-ES1," *2017 IEEE Automatic Speech Recognition and Understanding Workshop*, 2017.

[35] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.

[36] 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.

[37] 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.

[38] 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.

[39] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *The Journal of Machine Learning Research*, , no. 12, pp. 2121–2159, 2011.

[40] J. G. Zilly, R. K. Srivastava, J. Koutnik, and J. Schmidhuber, "Recurrent highway networks," *arXiv:1607.03474*, 2016.

[41] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems 27*, pp. 2672–2680, 2014.