The UTokyo speech synthesis system for Blizzard Challenge 2019

Shunsuke Goto, Yuma Shirahata, Gaku Kotani, Hitoshi Suda, Daisuke Saito, Nobuaki Minematsu

The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8656 Japan

{goto,shirahta,kotani,hitoshi,dsk_saito,mine}@gavo.t.u-tokyo.ac.jp

Abstract

This paper presents a speech synthesis system developed at the University of Tokyo (UTokyo) for the Blizzard Challenge 2019. The task of the challenge in 2019 is to build a voice using 8 hours of Mandarin Chinese speech data from an internet talk show by a well-known Chinese character. In this challenge, we have developed a statistical parametric speech synthesis system based on deep neural networks (DNN) incorporating non-negative matrix factorization (NMF). The developed system has been submitted, and the results of the large-scale subjective evaluation demonstrated the performance of our system.

Index Terms: deep neural network, DNN-based speech synthesis, non-negative matrix factorization, speaker adaptation

1. Introduction

To compare different speech synthesis techniques to develop a corpus-based speech synthesis system using shared data sets, Blizzard Challenge was devised in January 2005 [1], and has been held every year. This year's Blizzard Challenge has a single task that a voice of a well-known Chinese character is built from about 8 hours of speech data originally from an internet talk show by the target speaker.

Developing a speech synthesis system for an unknown or unfamiliar language is one of the challenging tasks in speech synthesis studies. Although Mandarin Chinese is not an unfamiliar language, treatment of Mandarin Chinese for speech synthesis generally requires expert knowledge, such as Chinese characters, syllabic tones and so on. Since we have no native speaker of Mandarin Chinese in our developing team this year, lack of expert knowledge for the target language is a big problem to be solved. A possible way to develop such a system is to adopt an end-to-end approach [2]. However, released data in this year are encoded audio files by MP3 format and they would have been recorded in various and bad environments. End-to-end systems could be directly affected by the quality of the prepared data. To treat such kinds of data, we have adopted an adaptation approach to develop the speech synthesis system to be submitted to the challenge.

In this challenge, we have developed a statistical parametric speech synthesis system based on deep neural networks (DNN). To develop acoustic models, we utilized external data from multiple speakers, which are 48 kHz-sampled RIFF WAV files. On the other hand, released data of the target speaker have been used for adaptation. To bridge the gap between different audio formats (or codec), and that between different speaker identities, we integrated an approach based on non-negative matrix factorization (NMF) with DNN-based statistical parametric speech synthesis, as one of the key techniques in the developed system. The developed system has been submitted, and the results of the large-scale subjective evaluation demonstrated the performance of our system.

2. Data and task

The task of this year's Blizzard Challenge is to produce a set of voices given Mandarin Chinese corpora. The database includes 480 samples, each of which is a one-minute talk spoken by a well-known Chinese character. Totally, the database has approximately 8 hours' speech data. The sampling rate of speech data is 24 kHz. In this database, speech data are recorded in MP3 audio format with relatively low bit rate. Text information are also given to each one-minute talk, but they are not aligned to audio utterances frame-by-frame. The testing transcriptions includes Chinese texts collected from an internet talk, news, and poems. Therefore, different sets of audio are required to be synthesized.

3. Voice building for UTokyo speech synthesis system

In this section, the voice building process of the UTokyo speech synthesis system for Blizzard Challenge 2019 is described.

3.1. Overview

A basic framework of UTokyo speech synthesis system is a statistical parametric speech synthesis, in which linguistic features derived from raw text transcriptions are mapped to the corresponding acoustic features. The mapped acoustic features are finally converted into waveforms.

In the training of the acoustic models, external data from multiple speakers are used. First, NMF is applied to spectral envelopes. Then, activity patterns are used as acoustic features for regression from linguistic features. Note that each speaker has a dictionary matrix and their index are correctly aligned.

To construct the dictionary matrix for the target speaker of the challenge, we estimate the mixing weights which are used to weight multiple base matrices of the external data. The released data of the target speaker are used for this purpose. By combining the estimated weights and the dictionary matrices from 48 kHz-sampled data, we can prepare the dictionary matrix for the target speaker in 48 kHz WAV domain. Finally, by multiply the generated activity patterns from the acoustic models and the constructed dictionary, we can obtain spectral envelopes of the target speaker. Note that F_0 and aperiodicity parameters are derived from an ordinary process of DNN-based speech synthesis.

3.2. External data

For constructing acoustic models by high quality speech data, we utilize ATR Chinese speech database[1]. This database includes Mandarin Chinese speech data of native speakers who are from 4 areas. The samples are 48 kHz-sampled RIFF WAV files. From the database, we picked up 3 speakers (Speaker A, B, and C, henceforth), who are similar to the target speaker of

[1] https://www.atr-p.com/products/sdb.html#REGCHINESE

this challenge in the viewpoint of spectra and F_0. Utterances of each speaker are about 30 minutes. Totally, 1.5 hours of speech data were utilized for training of acoustic models.

3.3. Data pruning of the released data

Since the released data includes the samples recorded on bad environments, some samples are not suitable for building voices. To prune the data, first we prepare the reference speech by Siri's default TTS in Chinese. Next, dynamic time warping (DTW) between the released data and the prepared reference were applied. In parallel, we obtained the forced alignments between the text transcriptions and the released data. Based on the DTW cost and the result of the forced alignment, the released data were pruned, and about 0.5 hours of data were left. The remaining data were utilized for speaker adaptation.

3.4. Text processing module

Since Chinese is a tonal language, both pronunciation and tone should be estimated from raw text. To extract linguistic features from raw text, we have adopted Kytea as a front-end module [3]. To estimate the correct pronunciation for Arabic numbers, text normalization where the numbers are converted to their corresponding Chinese characters is applied before input to Kytea. As a model for pronunciation estimation, the model released in the Kytea webpage was used [4]. Finally, the module makes a contextual vector for conditioning acoustic models.

3.5. Speech processing module

This module performs audio data preprocessing. As a preprocessing, first, MP3-formatted data (the released data) was converted to the RIFF WAV format. On the other hand, WAV-formatted external data were also converted to the MP3-formatted data and reconverted to WAV-formatted one again, because they were also used for estimating the mixing weights from the released data.

After the preprocessing, speech parameters are extracted from the audio data, using the WORLD analysis-synthesis system [5, 6]. Spectral envelopes, 5-band band aperiodicity, continuous log-scaled F_0, and unvoiced/voiced labels were extracted by the WORLD analysis.

3.6. DNN-based TTS incorporating NMF

We built DNN-based acoustic models for the developed system. As acoustic features, the activity patterns derived from NMF are employed for acoustic features.

3.6.1. Non-negative Matrix Factorization

NNF is a group of algorithms where matrix $\boldsymbol{Y} = (y_{k,n})_{K\times N}$ is factorized into two matrices $\boldsymbol{H} = (h_{k,m})_{K\times M}$, $\boldsymbol{U} = (u_{m,n})_{M\times N}$, with the property that all the matrices have no negative elements. That is,

$$\boldsymbol{Y} \simeq \boldsymbol{H}\boldsymbol{U}. \tag{1}$$

Usually, $K \gg M$, and $N \gg M$. The matrix $\boldsymbol{H}$ is called *exemplar* or *dictionary*, and $\boldsymbol{U}$ is called *activation*. In the modeling of spectra with NMF, a spectrum at the n-th frame is represented as a linear combination of basis spectra $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_M$ as follows;

$$\boldsymbol{y}_n \simeq \sum_{m=1}^{M} \boldsymbol{h}_m u_{m,n} = \boldsymbol{H}\boldsymbol{u}_n. \tag{2}$$

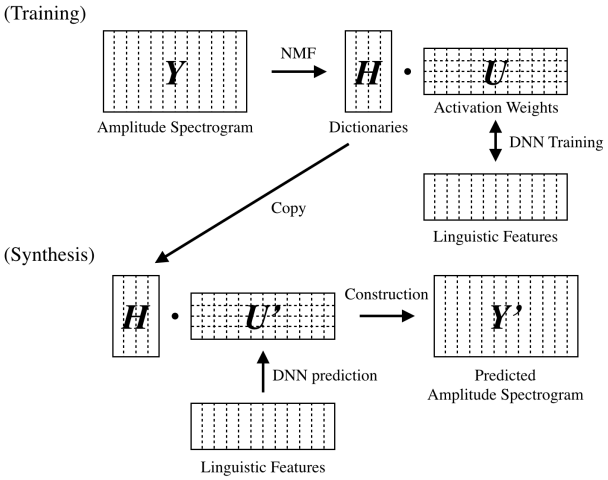

Figure 1: *Overview of speech synthesis incorporating NMF.*

To find an approximate factorization $\boldsymbol{Y} \simeq \boldsymbol{H}\boldsymbol{U}$, the elements of two matrices $\boldsymbol{H}$, $\boldsymbol{U}$ are iteratively updated based on a cost function.

3.6.2. Activation as spectral parameters

In speech synthesis, Mel-frequency cepstrum (MCEP) is often employed for acoustic features. In MCEP, a spectral envelope is represented as a linear combination of fixed envelope curves (sines and cosines). Similarly, in NMF, a spectrum is also represented as a linear combination of basis spectra. Therefore, in both of them, spectral envelopes are represented efficiently. However, in NMF, spectral bases are obtained flexibly by the decomposition of the spectrogram and each basis spectrum has fine structure, while bases in MCEP would lose the details of spectral envelopes.

The overview of the proposed TTS scheme incorporating NMF is shown in Figure 1. $\boldsymbol{Y}$, the amplitude spectrogram obtained by training speech data, is factorized into two matrices $\boldsymbol{H}$ and $\boldsymbol{U}$, and then a DNN-based acoustic model representing the relationship between the linguistic and acoustic features ($\boldsymbol{U}$) is trained. By multiplying the dictionaries $\boldsymbol{H}$ and the predicted activation $\boldsymbol{U'}$, the predicted amplitude spectrogram $\boldsymbol{Y'}$ is obtained. Detailed implementation is described in [7].

3.7. Multiple speaker training

NMF is widely used for voice conversion [8], noise reduction [9], etc. by operating the constructed spectral bases while keeping the activity patterns. By adopting a strategy of voice conversion based on NMF, multiple speaker training can be achieved. In the developed system, the following process was carried out.

1. NMF is applied to speech data of Speaker A. Acoustic models which map linguistic features to activation patterns are trained using the Speaker A's data.
2. Using the trained acoustic models and the transcription of the data for Speaker B and C, the corresponding activation patterns to the data of Speaker B and C are generated.
3. NMF is applied to speech data of Speaker B and C. Initial values of activation patterns are the generated activation patterns in the previous step. By this step, parallel

dictionaries of Speaker A, B, and C can be obtained. Activation patterns are also updated in the iteration of NMF.

4. Using all the activation patterns from Speaker A, B, and C, we can update the acoustic models.

The above process simultaneously achieves data augmentation and construction of parallel dictionaries. In this training phase, 48 kHz-sampled RIFF WAV data were utilized.

3.8. Construction of dictionaries for the target speaker

Since the released data are 24 kHz-sampled MP3 data, 48 kHz-sampled dictionaries cannot be constructed directly from the data. In the developed system, we assume that the dictionary of the target speaker can be represented by the weighted sum of the multiple dictionaries from the other speakers, i.e. Speaker A, B, and C. The mixing weights are inferred by the following process.

1. Using the trained acoustic models and the transcription of the prestored data, activation patterns of the data or Speaker A, B, and C are generated.
2. NMF is applied to MP3-formatted speech data of the prestored speakers. Initial values of activation patterns are the generated ones in the previous step. 24 kHz-sampled dictionaries are obtained in this step.
3. NMF is applied to MP3-formatted speech data of the target speakers. Initial values of activation patterns are generated from the transcription and the acoustic models. 24 kHz-sampled dictionary of the target speaker is obtained in this step.
4. For each spectral slice of the dictionary in the previous step, NMF is applied with the fixed dictionaries from the prestored speakers. The mixing weights can be inferred in this step.
5. Using the obtained mixing weights and the dictionaries in 48 kHz-sampled data, the dictionary of the target speaker in 48 kHz-sampled RIFF WAV format is obtained.

4. Experimental evaluation

4.1. Experimental settings

Our designated system identification letter is 'L.' System A is natural speech. System B is a DNN benchmark using Merlin toolkit. Others are participants' systems. The subjects who are involved in the listening test are paid listeners, speech experts, and online volunteers.

4.2. Results and analysis

We show 1) naturalness in Figure 2, 2) speaker similarity in Figure 3, 3) Pinyin error rates without tone (PER) in Figure 4, and 4) Pinyin error rates with tone (PTER) in Figure 5.

Totally, our results were not good, in naturalness, speaker similarity, and intelligibility. The main reason of the results would be in the adaptation phase of the proposed method. In the proposed adaptation method, only 3 speakers were used as a basis dictionary set. In addition, adaptation weights were inferred based on an NMF manner for each spectral slice of the dictionary. In that case, the estimated speaker weights tend to be changed frame-by-frame. This inconsistency in dynamics of the parameters would degraded both the quality and speaker similarity of the synthesized speech. Effects of the number of basis speakers for speaker adaptation in the proposed framework should be investigated in further works.

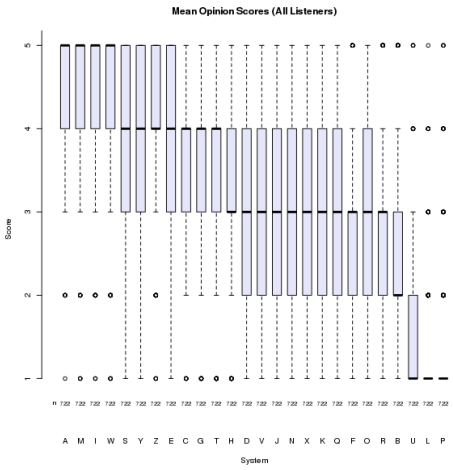

Figure 2: *Mean opinion scores (naturalness of synthetic speech).*

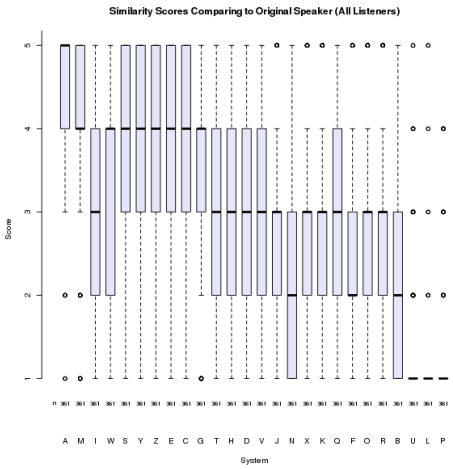

Figure 3: *Mean opinion scores (similarity of synthetic speech to original speaker).*

5. Conclusions

We introduced the UTokyo speech synthesis system for Blizzard Challenge 2019. The results of the listening test for our system were not good, but we have found many interesting problems that we should have attacked.

6. Acknowledgements

This research and development work was supported by the MIC/SCOPE #182103104.

7. References

[1] A. W. Black and K. Tokuda, "The blizzard challenge-2005: Evaluating corpus-based speech synthesis on common datasets," in *Proc. INTERSPEECH*, Lisbon, Portugal, Sep. 2005.

[2] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly,

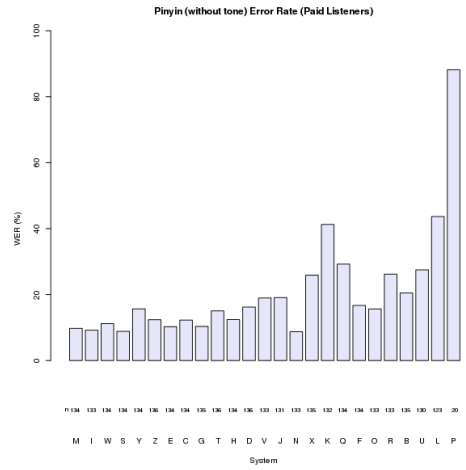

Figure 4: *Pinyin error rates (without tones).*

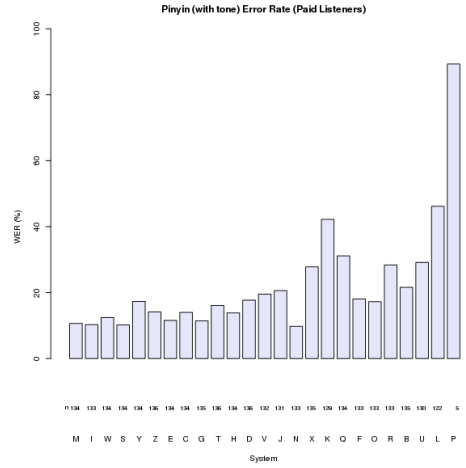

Figure 5: *Pinyin error rates (with tones).*

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," vol. abs/1609.03499, 2017. [Online]. Available: https://arxiv.org/abs/1703.10135

[3] "Kytea: the Kyoto text anlysis toolkit http://www.phontron.com/kytea/index.html."

[4] "Kytea Models: the Kyoto text anlysis toolkit http://www.phontron.com/kytea/model.html."

[5] 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, no. 7, pp. 1877–1884, 2016.

[6] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, vol. 84, pp. 57–65, 2016.

[7] S. Goto, D. Saito, and N. Minematsu, "DNN-based statistical parametric speech synthesis incorporating non-negative matrix factorization," in *Proc. APSIPA ASC*, Lanzhou, Chaina, Nov. 2019.

[8] Z. Wu, T. Virtanen, T. Kinnunen, E. S. Chng, and H. Li, "Exemplar-based voice conversion using non-negative spectrogram deconvolution," in *ISCA Workshop on Speech Synthesis, SSW8*, 2013, pp. 201–206.

[9] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 4029–4032.