Mongolian Text-to-Speech System Based on Deep Neural Network

Rui Liu, Feilong Bao*, Guanglai Gao, and Yonghe Wang

College of Computer Science, Inner Mongolia University
010021, Hohhot, China
liurui_imu@163.com;{csfeilong,csggl}@imu.edu.cn;cswyh92@163.com

Abstract. Recently, Deep Neural Network (DNN), which is a feed-forward artificial neural network with many hidden layers, has opened a new research direction for Speech Synthesis. It can represent high dimension and correlated features efficiently and model highly complex mapping function compactly. However, the research on DNN-based Mongolian speech synthesis is still in blank filed. This paper applied the DNN-based acoustic model to Mongolian speech synthesis firstly, and built a Mongolian speech synthesis system according to the Mongolian character and acoustic features. Compared with the conventional HMM-based system under the same corpus, the DNN-based system can synthesize better Mongolian speech than HMM-based system can do. The Mean Opinion Score (MOS) of the synthesized Mongolian speech is 3.83. And it becomes a new state-of-the-art system in this field.

Keywords: Mongolian; Text-to-Speech (TTS); Acoustic Model; Deep Neural Network (DNN).

1 Introduction

There are probably seven thousand languages in the world today [1]. However, the study of the Text-to-Speech (TTS) system only focuses on a few major languages, such as English, Chinese, Japanese, Spanish, Turkey and so on. Mongolian is a widely influential language in the world, with about six million users, but there is less research on Mongolian TTS. Furthermore, Mongolian has its own special characteristics. Its words consist of stem and suffix to form a large number of words. Due to the limited training data of Mongolian TTS, the serious data sparseness problem were caused [2]. These make TTS for Mongolian difficult.

TTS is also called speech synthesis. While for the Mongolian speech synthesis, Ochir et al. proposed a Mongolian speech synthesis system based on waveform concatenation [3]; Monghjaya conducted a research on the Mongolian speech synthesis based on stem and affixes [4]; Aomin carried on a study on Mongolian speech synthesis based on the prosodic [5]; Zhao used the HMM-Based methods in the Mongolian speech synthesis [6]. These studies have made contribution

to the Mongolian speech synthesis, but the naturalness of Mongolian speech synthesis is less than satisfactory.

Recently, Deep Neural Networks (DNNs) [7] have achieved significant improvement in many machine learning areas. Motivated by the success of DNNs in speech recognition [8], DNNs have been introduced to statistical parametric speech synthesis in order to improve the performance of speech synthesis. Zen et al. [9] showed that DNN-based acoustic models offer an efficient and distributed representation of complex dependencies between contextual and acoustic features. However, DNNs can be introduced to components other than acoustic modeling in statistical parametric speech synthesis and it should be further investigated about how DNNs can be used in statistical parametric speech synthesis.

In this paper, we investigate how to use DNNs in Mongolian speech synthesis. We introduce the concept of DNN acoustic model, for the first time, into the Mongolian statistical parametric speech synthesis. By replacing decision-trees with DNN, the effect of DNN acoustic model in statistical parametric speech synthesis is investigated. The rest of this paper is organized as follows. Section 2 describes the Mongolian TTS system based on DNN. The experimental conditions and results are shown in Section 3. Conclusions are presented in Section 4.

2 Mongolian TTS system based on DNN

In this study, we build a Mongolian TTS system based on DNN. Figure 1 illustrates a block diagram of the system. It consists of training part and synthesis part.

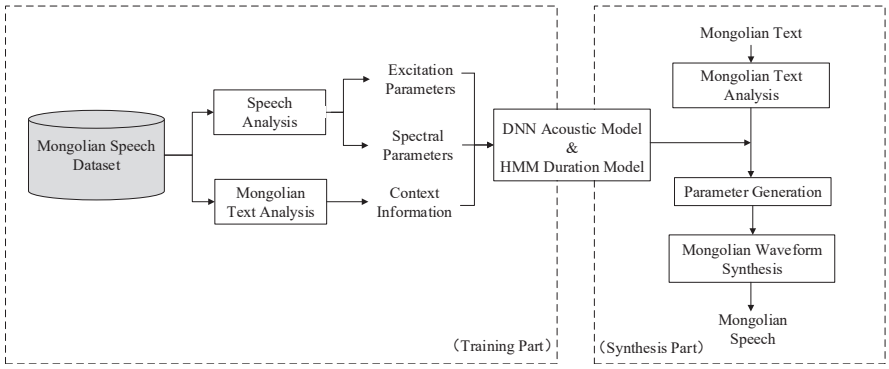


Fig. 1. A Mongolian TTS system framework based on DNN.

2.1 Training Part

At the training part, there are two models are trained in advance, including a HMM-based duration model and a DNN-based acoustic model. In order to complete the model training we need to do the following work:

Mongolian Text Analysis. Firstly, we use PRAAT Toolkit [10] to label the speech data in order to align phoneme boundary, mark the corresponding Mongolian phoneme name (from Mongolian Phoneme Set) and prosodic phrase boundary. Next, both monophone alignment labels and full context labels are extracted from the labeled files generated in the previous step according to the Mongolian Question Set [6] designed expressly.

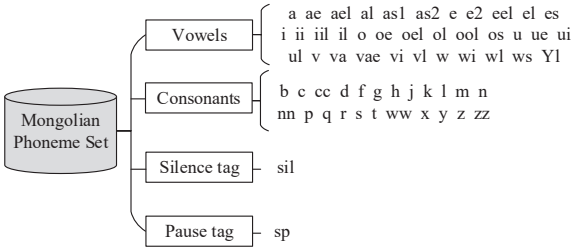


Fig. 2. Mongolian Phoneme Set.

The Mongolian Phoneme Set is described with 60 phonemes which include 35 vowels, 23 consonants, a silence tag and a pause tag as listed in Figure 2.

Speech Analysis. In the following acoustic model training, output vector of DNN consists of excitation part and spectral part. In this work, the frame-level excitation parameter (Log fundamental frequency, LogF0) and the frame-level spectral parameter (Mel-generalized cepstral coefficient, MGC) are extracted by using the Speech Signal Processing Toolkit (SPTK) [11].

HMM-based Duration Model. Using the previous acoustic features and a decision-tree based context clustering technique [12, 13], states of the context dependent HMMs are clustered, and the tied context dependent HMMs are reestimated with the embedded training. Simultaneously, state durations are calculated on the trellis which is obtained in the embedded training stage, and modeled by Gaussian distributions. Finally, context dependent duration models are clustered by using the decision-tree based context clustering technique.

DNN-based Acoustic Model. In the TTS research, DNN is used as an alternative of the HMM shown in Fig. 3. The input linguistic feature vector is converted to an output acoustic vector directly. In this approach, frame-level input linguistic features l_t rather than phoneme-level ones are used. They include binary answers to questions about linguistic contexts (e.g. is-current-phoneme-vowel?), phoneme-level numeric values (e.g. the number of words in the phrase, duration of the current phoneme), and frame-level numeric features

(e.g. the relative position of the current frame in the current phoneme). The target acoustic feature vector o_t includes spectral and excitation parameters and their dynamic features. The weights of DNN are trained using pairs of input and target features extracted from training data at each frame by Back-Propagation.

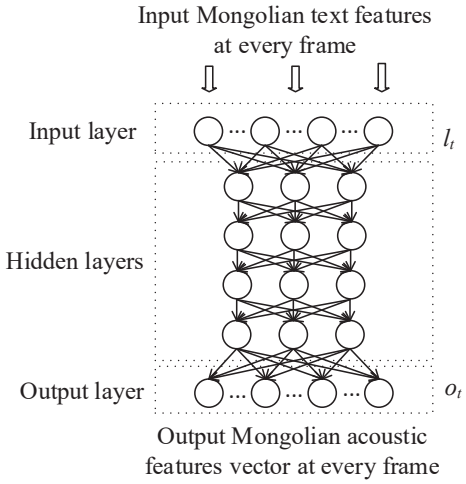


Fig. 3. DNN-based Acoustic Model.

2.2 Synthesis Part

In synthesis part, we first extract the contextual text feature from the given Mongolian text, then use the trained duration model to predict the duration and use the trained acoustic model to generate the acoustic features vector, finally obtain the speech parameter so as to output the synthesized Mongolian speech. We will explain in the following contents.

Mongolian Text Analysis. This part consists of a latin transcriptions module, coding correction module, grapheme to phoneme conversion (G2P) module, syllable segmention module, prosody phrase prediction module and a linguistic features extraction module.

(1) latin transcriptions & coding correction module

In Mongolian language, there is a phenomenon that many words have the same presentation form but represent different words with different codes. Since typists usually input the words according to their representation forms and cannot distinguish the codes sometimes, there are lots of coding errors in Mongolian corpus. For example, the Mongolian word "ᠪᠦᠷᠢᠨ" means "complete"

when its right code is "guiqed", its other wrong code is "guiqad", "huiqed" or "huiqad".

In this work, we transforme the Mongolian characters to their corresponding Latin transcriptions (code), and merge the words with same presentation forms by Intermediate characters [14] to correct the code.

(2) G2P module

Then we use the statistic-based Mongolian grapheme to phoneme (G2P) conversion method [15] to generate Mongolian phoneme sequence.

(3) syllable segmentation module

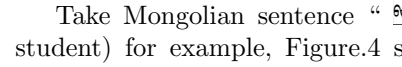
In this step, we mark the syllable boundrary inside Mongolian words according to the rules, and obtain marked phoneme sequence.

(4) prosody phrase prediction module

Mongolian prosodic phrase prediction is essential for generating higher naturalness speech [16]. We predict and mark the prosodic phrase boundary in the marked phoneme sequence by using Conditional Random Field (CRF) Model [16, 17].

(5) linguistic features extraction module

Finally, we extract the linguistic features l_t from the processed phoneme sequences according to the Mongolian Question Set and obtained phone-level frame length.

Take Mongolian sentence "  " (means: I am a college student) for example, Figure.4 shows the above process. The red error code in the text is corrected for the green right code. Each Mongolian word in the Figure is separated by "#" and each Mongolian syllable is separated by "/", the predicted prosodic phrase boundary is marked by "%".

Parameter Generation & Mongolian Waveform Synthesis. We feed the input linguistic features into the trained acoustic model to generate the frame-level acoustic features. Then a sequence of speech parameter including spectral and excitation parameters is determined using the (Maximum Likelihood Parameter Generation) MLPG algorithm [18]. Finally, a speech waveform is resynthesized directly from the generated spectral and excitation parameters by using a source-filter model [19].

In this work, we use *hts_engine* [20] to construct the speech waveforms.

3 Experiments and Results

We implement the Mongolian TTS based on the HTS speech synthesis toolkit [21] and use HMM and DNN to train the Acoustic Model.

3.1 Dataset

We build a phonetically and prosodically rich Mongolian speech synthesis corpus consisting of 2,620 training utterances which contains about 2 hours (90% as training set and the rest as testing set) and 40 extra utterances for evaluation. The material includes Mongolian daily expressions recorded by a female speaker.

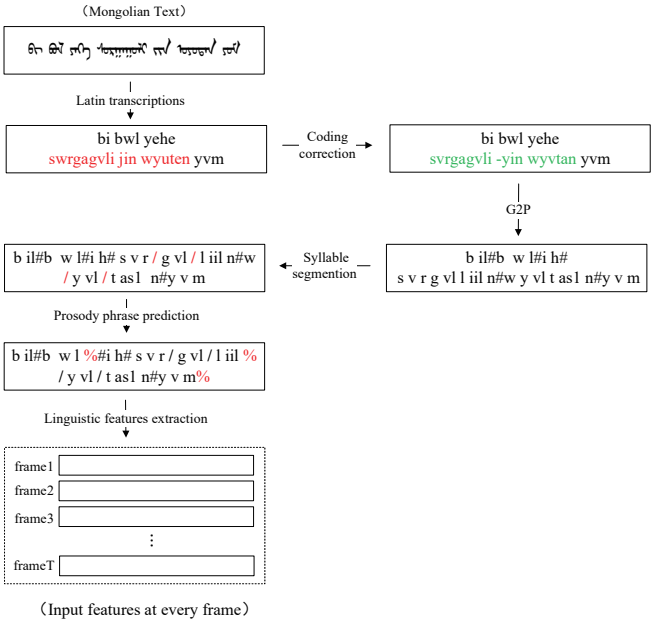


Fig. 4. A sample of Mongolian Text Analysis process.

Speech signals are sampled at 16 kHz, windowed by a 25-ms window shifted every 5-ms. The annotation of the corpus is done by two Mongolian students according to the Mongolian Phoneme Set.

3.2 Experiments Setup

In the baseline HMM-based Mongolian TTS system, five-state, left-to-right HMM phone models, where each state is modeled by a single Gaussian, diagonal covariance output distribution, are adopted. The phonetic and prosodic contexts in Mongolian [6] are used as a Question Set in growing decision trees. To model LogF0 sequences consisting of voiced and unvoiced observations, a multi-space probability distribution (MSD) was used [22]. The number of questions for the Mongolian decision tree-based context clustering was 693. The sizes of decision trees in the HMM-based systems were controlled by changing the scaling factor for the model complexity penalty term of the Minimum description length (MDL) criterion [23, 24].

In the DNN-based Mongolian TTS system, the input linguistic features were automatically extracted from the Mongolian Question Set, the derived context information about the text were further encoded into a vector of 693 dimensions as the input to the neural network. The output feature vector contains 35 MGC, LogF0, their delta and delta-delta features and voiced/unvoiced flag, totally 109 dimensions (3*(35+1)+1=109). Voiced/unvoiced flag is a binary feature that

indicates whether the current frame is voiced or not. To model LogF0 sequences by a DNN, the continuous F0 with explicit voicing modeling approach was used. All silence frames from the training data are adjusted to 0.3 seconds to reduce the computational cost. The sigmoid activation function was used for hidden and output layers. Input features were normalised to the range of [0.01, 0.99] and output features were standardised to have zero mean and unit variance. Both input and output features of training data are trained by back-propagation procedure with a "mini-batch" based stochastic gradient ascent algorithm. The weights of the DNN were initialized randomly and a learning rate of 0.001 was used. A single network which modeled both spectral and excitation parameters was trained.

For the testing utterances, the DNN outputs is firstly fed into a parameter generation module to generate smooth feature parameters with dynamic feature constraints [18], Finally, the Mongolian speech waveforms are synthesized using the Source-Filter Model.

3.3 Evaluation

Objective and subjective measures are used to evaluate the performance of two acoustic model on testing data.

Synthesis quality is measured objectively in terms of distortions between natural test utterances of the original speaker and the synthesized speech. We employ the root mean squared error (RMSE) of LogF0, and the RMSE of MGC as the evaluation metric.

RMSE is commonly used to evaluate the mean error between generated parameter and original parameter. We define it as following function

$$RMSE = \sqrt{\sum_{i}^{N} \left(\log\left(f_o(i)\right) - \log\left(f_e(i)\right)\right)^2 / N} \tag{1}$$

Where N is total frames in all sentence, $f_o(i)$ is original F0 parameter, $f_e(i)$ is estimated F0 parameter.

For HMM trainings, because of space limitations, this article does not show the objective measures of different MDL factors (α =16, 8, 4, 2, 1, 0.5, 0.375, 0.25). Based on these results, we find out that larger MDL factors yield worse objective measures, and the best objective measures emerge from this MDL factors with α =1.

The results of objective measures of different structures (different number of hidden layers: 1, 2, 3, 4, 5 and units per layer: 256, 512, 1024, 2048) in DNN trainings are shown in Table 1. From the experimental results can be seen. For MGC RMSE, the simplest DNN structures (1*256) yields the best results. For LogF0 RMSE, the best performance emerges from the DNN structures which is 2*512.

For all structures, the simple DNN structures are better than the complex structures and the performance of multiple layers can match the performance of more units per layer.

Table 1. The objective measures of different structure in DNN.

DNN Structure	MGC RMSE	LogF0 RMSE	DNN Structure	MGC RMSE	LogF0 RMSE
1*256	**21.643**	3.421	1*1024	22.235	3.435
2*256	22.965	3.350	2*1024	22.156	3.347
3*256	22.449	3.353	3*1024	23.941	3.393
4*256	22.540	3.391	4*1024	23.429	3.390
5*256	22.391	3.371	5*1024	23.476	3.395
1*512	21.816	3.425	1*2048	22.592	3.415
2*512	21.895	**3.341**	2*2048	22.254	3.359
3*512	23.422	3.382	3*2048	23.915	3.399
4*512	23.740	3.389	4*2048	23.490	3.393
5*512	23.055	3.388	5*2048	23.657	3.399

To evaluate the naturalness of the synthesized Mongolian speech by HMM-based TTS system and DNN-based TTS system, a subjective listening test was conducted. The naturalness of the synthesized speech was assessed by the mean opinion score (MOS) test method. In this evaluation, the total number of test utterances was 40, which are synthesized by the best baseline HMM system (α=1), the simplest DNN system (1*256), the most complex DNN system (5*2048), the best DNN system (1*256, 2*512). The subjects were four Mongolian students in our research group. Speech samples were presented in random order for each test sentence. In the MOS test, after listening to each test sample, the subjects were asked to mark the sample a five-point naturalness score (5: natural, -1: bad).

Figure 5 shows the subjective evaluation results. It can be seen from the figure that the TTS system, with 2*512 DNN structures, obtain the highest MOS. Most DNN-based Mongolian TTS systems outperform HMM-based Mongolian TTS system. Too simple or too complex network structure may not be able to achieve good results. This result indicates that replacing the tree-based clustered models into a reasonable DNN-based acoustic model is effective, and the best DNN structure of Mongolian TTS system is 2*512.

4 Conclusions

DNN-based acoustic model has been applied firstly in this study for Mongolian TTS. The results show that DNN performs better than HMM-based baseline does in Mongolian TTS system. The experiment results show that reasonable DNN is efficient and effective in representing high dimensional and correlated features.

In future work, we will investigate the effect of DNNs in statistical parametric speech synthesis on larger Mongolian database. Besides, Recurrent Neural Network (RNN) is used more frequently in TTS now [25–27]. We plan to explore its power in the Mongolian acoustic model.

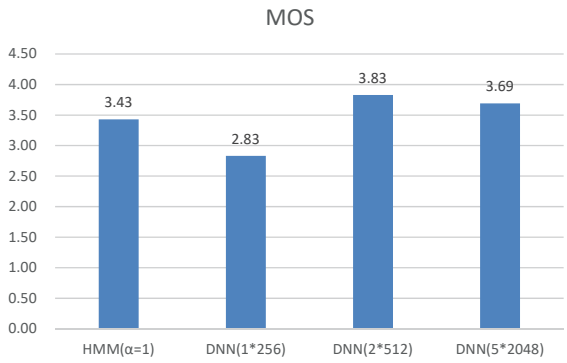


Fig. 5. MOS of the best baseline HMM system and DNN systems.

Acknowledgments. This research was supports in part by the China national natural science foundation (No.61563040, No.61773224) and Inner Mongolian nature science foundation (No. 2016ZD06).

References

1. Ethnologue: Languages of the world, eighteenth edition, `http://www.ethnologue.com`
2. Feilong, B., Guanglai, G., Xueliang, Y., et al.: Segmentation-based Mongolian LVCSR approach. In: 38th IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8136-8139, IEEE Press, Canada (2013)
3. Ochir, Zheng, G.: A Test of The Speech Synthesis With The Waveform Concatenation. In: 3th National Conference on Man-Machine Speech Communication, pp: 408-412, Chongqing (1994)
4. Monghjaya.: A Research on Mongolian Speech Synthesis System Based on Stems and Affixes. Journal of Inner Mongolia University. 39, 693-697 (2008)
5. Aomin., Ziyu, X., He, H., et al.: A Study on the Piano and Rhyme Phrases of Mongolian. In: 10th Phonetic Conference of China Processing, Shanghai (2012)
6. Jiandong, Z., Guanglai, G., Feilong, B.: Research on HMM-based Mongolian Speech Synthesis. Computer Science. 41, 80-104 (2014)
7. Bengio, Y.: Learning Deep Architectures for AI. Foundations & Trends in Machine Learning. 2, 1-55 (2009)
8. Hinton, Li, D., Dong, Y., et al.: Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. IEEE Signal Processing Magazine. 29, 82-97 (2012)
9. Zen, H., Senior, A., Schuster, M.: Statistical parametric speech synthesis using deep neural networks. In: 38th IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7962-7966, IEEE Press, Canada (2013).
10. Praat, `http://www.fon.hum.uva.nl/praat/`
11. SPTK, `http://sp-tk.sourceforge.net/`
12. Yoshimura, T., Tokuda, K., Masuko, T., et al.: State Duration Modeling for HMM-Based Speech Synthesis. IEEE Transactions on Information & Systems. 90, 692-693 (2007)

13. Yoshimura, T., Tokuda, K., Masuko, T., et al.: Simultaneous Modeling of Spectrum, Pitch and Duration in HMM-Based Speech Synthesis. In: 6th European Conference on Speech Communication and Technology, pp. 2099-2107, IEEE Press, Hungary (1999)
14. Xiaofei, Y., Feilong, B., Hongxi, W., et al.: A Novel Approach to Improve the Mongolian Language Model Using Intermediate Characters. In: 15th China National Conference on Chinese Computational Linguistics, pp.103-113, IEEE Press, Shandong (2016)
15. Feilong, B., Guanglai, G., Xueliang, Y.: Research on grapheme to phoneme conversion for Mongolian. Application Research of Computers. 30, 1696-1700 (2013)
16. Rui, L., Feilong, B., Guanglai, G., Hongwei, Z.: Approach to Predicition Mongolian Prosody Phrase Based on CRF Model. In: 13th National Conference on Man-Machine Speech Communication, Tianjin (2015)
17. Rui, L., Feilong, B., Guanglai, G.: Mongolian prosodic phrase prediction using suffix segmentation. In: 20th International Conference on Asian Language Processing, pp. 250-253, IEEE Press, Taiwan (2016)
18. Tokuda, K., Yoshimura, T., Masuko, T., et al.: Speech parameter generation algorithms for HMM-based speech synthesis. In: 25th IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 1315-1318, IEEE Press, Istanbul (2000)
19. Milner B., Shao X.: Speech reconstruction from mel-frequency cepstral coefficients using a source-filter model. In: 7th International Conference on Spoken Language Processing, IEEE Press, Denver (2002)
20. hts-engine, `http://hts-engine.sourceforge.net/`
21. HTS, `http://hts.sp.nitech.ac.jp/`
22. Masuko, T.: Multi-Space Probability Distribution HMM. IEEE Transactions on Information & Systems. 85, 455-464 (2002)
23. Grnwald, P.: The minimum description length principle. Mit Press, 1, 257-268 (2007)
24. Heng, L., Zhenhua, L., Ming, L., et al.: Minimum Generation Error Based Optimization of HMM Model Clustering for Speech Synthesis. Pattern Recognition & Artificial Intelligence. 23, 822-828 (2010)
25. Zen, H., Sak, H.: Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. In: 40th IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4470-4474, IEEE Press, Australia (2015)
26. Achanta, S., Godambe, T., Gangashetty, S.V.: An Investigation of Recurrent Neural Network Architectures for Statistical Parametric Speech Synthesis. In:16th Interspeech, IEEE Press, Germany (2015)
27. Zhizheng, W., S, King.: Investigating gated recurrent networks for speech synthesis. In: 41st IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5140-5144, IEEE Press, Shanghai (2016)