Proceedings of ICSP2018

Progressive Neural Networks based Features Prediction for the Target Cost in Unit-Selection Speech Synthesizer

Ruibo Fu [a, b], Jianhua Tao [a,b,c] , Zhengqi Wen [a]

a National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
b School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
c CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China

Email: {ruibo.fu, jhtao, zqwen}@nlpr.ia.ac.cn

***Abstract*—This paper describes a direct acoustic features prediction for calculation of the target cost by progressive neural networks. Compared with conventional methods involving many hand-tuning steps, our method directly predicts the features for calculation of the target cost. By applying the progressive deep neural network (PDNN) to predict these acoustic features, the correlation of these features can be modeled. Each type of the acoustic features and each part of a unit are modeled in different sub-networks with its own cost function and the knowledge transfers through lateral connections. Each sub-network in the PDNN can be trained to reach its own optimum step by step. Extensive comparative evaluations demonstrate the effectiveness of the PDNN in improving the accuracy of predicted acoustic features. The subjective evaluation results demonstrate that the naturalness of synthetic speech has been improved by adopting the proposed method to calculate the target cost.**

Keywords—speech synthesis, unit-selection, target cost, progressive neural networks

I. INTRODUCTION

The unit-selection speech synthesis [1] has been challenged by the statistical parametric speech synthesis (SPSS) [2] and advanced methods (WaveNet, Deep Voice, Tacotron) [3-9] recently. However, the above new advanced methods still need more delicate works in computational efficiency and robustness. And the SPSS based speech synthesis tends to generate "average" speech which would defect the perception of sound. The unit-selection synthesizer is preferred when the speech corpus is highly-curated. And the ability of unit-selection is to yield studio-level quality for limited-domain speech synthesis.

One of the core problem for unit-selection synthesizer is the discontinuousness between the selected adjacent basic units. People would identify that the selected sequence of units are extracted from difference utterances when acoustic clues such as the intonation, the speaking style, and the speed, are unmatched, which would defect the perception of sound.

The target cost and the concatenation cost are defined to decide the best candidate from the corpus database. The target cost is designed to select the proper candidates from database. And the concatenation cost is designed to select adjacent units sound more coherently. The target cost, the metric to the similarities between candidate units and target units, is the foundation to select the proper combination of candidate units. But the target cost is hard to define and predict.

Hunt and Black first presented current form of unit-selection speech synthesis system [1]. The differences of prosodic and phonetic context information are calculated by weighted sum. The system performed best in the situation of the large database and high audio quality. Then the hybrid unit-selection [10], in which the target cost was related to acoustic and prosodic parameters predicted by the statistical model, used more the acoustic features to guide the unit-selection. Several improvements, such as using the Deep Neural Networks (DNN) to generate the guiding parameters, have been made to use more acoustic clues for unit-selection [11-13].

The features that the above methods used were manual well-designed. To make features be extracted automaticly by training, constructing a fixed-size representation of the variable-size audio, which referred as a embedding, were proposed. Approaches that take frame-level embedding of linguistic and acoustic information the intermediate layers of a deep neural network (DNN) [14] or a long short-term memory (LSTM) [15] network. In both cases, the unit-level embedding was constructed heuristically rather than being extracted from the whole unit directly. Then a sequence to sequence LSTM-based auto-encoders method is proposed to encode variable-length audio to a fixed-length vector [16]. The metric of the trained embedding is designed to represent the similarities of sounding units.

The above manual designed features extraction methods might accumulate of errors in separate steps. And embedding methods were hard to train because each unit has text and audio two modalities. It was difficult to generate a uniform fixed-length vector to represent both acoustic and linguistic features. In this paper, we directly predict the acoustic features that is needed for the calculation of the target cost by using PDNN. The features including duration, MFCC and f0 are predicted directly, which avoid the accumulation of errors and is more computing efficient. We divide the unit into four parts.

And the correlation of the acoustic features in the sub-units is modeled by the proposed PDNN framework

Progressive neural networks (ProgNets) [17] was first proposed by Google for the reinforcement learning tasks. ProgNets trained new task by freezing the previous trained. We separate the prediction of acoustic features into several sub-tasks. Each sub-task is modeled in different sub-networks with its own cost function and the knowledge transfers through lateral connections. Each sub-network in the PDNN can be trained to reach its own optimum step by step. The correlation of the acoustic features between the middle and the margin of the units can be modeled and investigated.

An overview of the rest of the paper is as follows: in section II we describe the framework of the unit-selection synthesizer and the PDNN for predicting acoustic features. Section III presents the experiments. And the results and analysis are presented in Section IV. The conclusions and future work are discussed in Section V.

II. METHOD

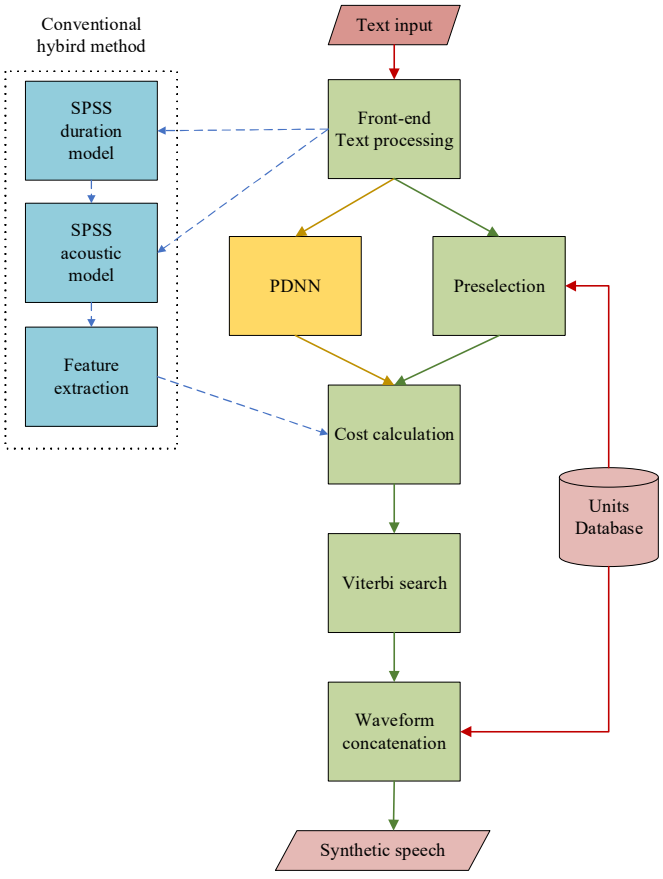

Fig. 1. An overview of the unit-selection speech synthesizer. In our system, the PDNN module (yellow blocks) replace the conventional hybrid method (blue blocks) for generating the designed acoustic features for target cost calculation

As illustrated in the Fig. 1, our speech synthesis system follows the typical unit selection framework, which uses a front-end text processing to produce linguistic features, pre-selection for narrowing down the searching space, statistical model to implement concatenation and target costs for Viterbi search that finds the optimum unit sequence, and waveform concatenation to generate synthetic speech.

The conventional hybrid methods using HMM or DNN based SPSS as guiding system. First, the acoustic model and duration model of SPSS predicts a sequence of acoustic parameters. Second, the generated acoustic parameters are extracted to the manual designed features for calculating the target cost.

In this paper, we replace the above separate steps to a PDNN framework for direct manual designed features predicting. In this section, we will introduce the PDNN framework and designed acoustic features for target cost calculation. Besides, another Multi-task learning method as extensive comparative method is also introduced in this section.

A. *Progressive deep neural networks*

Compared with the conventional transfer learning methods that use the learned parameters as initial parameters, the PDNN use the following strategies:

- Firstly, all the parameters of the old model are frozen when the new task begins.
- Secondly, the new model is initialized randomly.
- Thirdly, lateral connections are built between the new model and the frozen old model.
- Fourthly, the parameters of the new model is learned through backpropagation.

The PDNN framework with 3 columns (3 color blocks) for unit-selection speech synthesis system is shown in Fig. 2 The first task starts with a single column (green in the Fig. 2): A deep neural network having 3 layers with hidden activations $h_i^{(1)} \in R^{n_i}$, with n_i the number of units at layer $i \leq 5$,and parameters $\Theta^{(1)}$ trained to convergence.

When switching to a second task, the parameters $\Theta^{(1)}$ are "frozen" and a new column (yellow in the Figures 1) with parameters $\Theta^{(2)}$ is instantiated with random initialization, where layer $h_i^{(2)}$ receives input from both $h_{i-1}^{(2)}$ and $h_{i-1}^{(1)}$ via lateral connections. This generalizes to K tasks as follows:

$$h_i^{(k)} = f\left(W_i^{(k)} h_{i-1}^{(k)} + \sum_{j<k} U_i^{(k:j)} h_{i-1}^{(j)} + b_i^{(k)}\right) \quad (1)$$

where $W_i^{(k)} \in R^{n_i \times n_{i-1}}$ is the weight matrix of layer i of column k, $U_i^{(k:j)} \in R^{n_i * n_j}$ are the lateral connections from layer $i-1$ of column$j$, to layer i of column k, $b_i^{(k)}$are the

biases and h_0 is the network input. f is the activation function.

In the construction of PDNN, it is important to carefully select a method for combining representations across network and to identify where these representations will be combined. Adaptation layers (as in the Fig. 2) can be included to transform from one task to another. However, due to the limit of the computational complexity in the runtime unit-selection system, a sharing weights strategy of the lateral connections $U_i^{(k:j)}$ is adopted in our framework. All the row of the matrix $U_i^{(k:j)}$ are sharing the same row vector. Except the first layer, each node of in the same layer of the column k would receive the same bias. The real number of parameters in the matrix $U_i^{(k:j)}$ is n_i instead of $n_i \times n_{i-1}$.

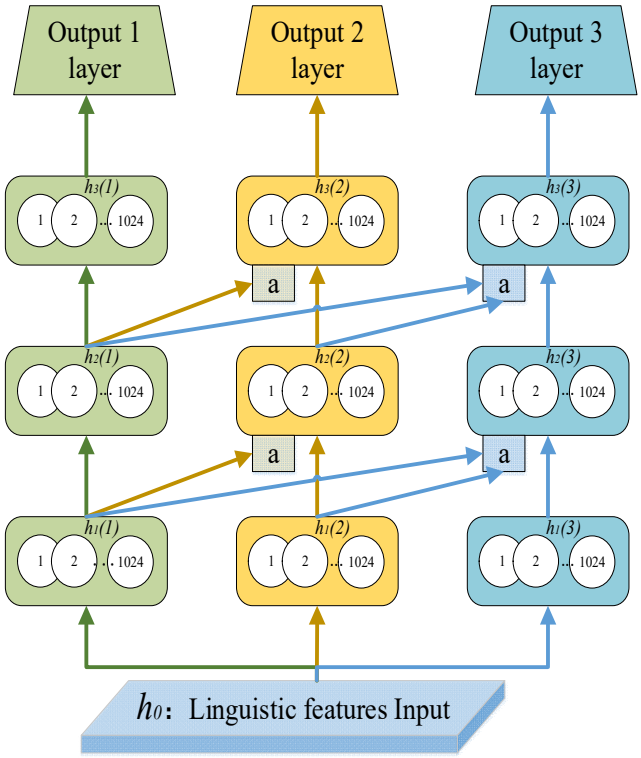

Fig. 2. PDNN framework for predicting the acoustic features in unit-selection speech synthesis system. Blocks a represent the adaption layers as lateral connections.

B. *Designed acoustic features and cost calculation*

The basic unit of our unit-selection synthesizer is vowels or consonants of the Mandarin, which resemble the syllables of the English. One type of the acoustic features is the duration. Instead of duration model which predict continuous value in the SPSS system, we classify the duration into 10 classes. To extract other acoustic features such as MFCC and fundamental frequency f_0, each unit is divided into 4 sections. The whole unit might contain voiced and unvoiced segments, which have different acoustic features distributions. The acoustic features distributions of the whole unit could not represent the articulation of units well. The first and last sections usually contain the unvoiced segments of the utterance. And the middle sections usually contain voiced segments of the utterance. The mean and variance of above are computed as the acoustic features of each unit. We use the above 4 sections of a unit as the metric for the target cost. Besides, the mean and variance of the acoustic features of the whole unit is also extracted as intermediate variables. In total, there are 290 features for predicting. The features are listed in TABLE I.

TABLE I. DESIGNED ACOUSTIC FEATURES FOR THE TARGET COST CALCULATION

Feature	Position	Dimension	Intermediate /final features
Duration	-	10	Final
MFCC-W	Whole	26	Intermediate
ΔMFCC-W	Whole	26	Intermediate
f_0-W	Whole	2	Intermediate
Δf_0-W	Whole	2	Intermediate
MFCC-B	Beginning	26	Final
ΔMFCC-B	Beginning	26	Final
f_0-B	Beginning	2	Final
Δf_0-B	Beginning	2	Final
MFCC-ML	Middle-Left	26	Final
ΔMFCC-ML	Middle-Left	26	Final
f_0- ML	Middle-Left	2	Final
Δf_0- ML	Middle-Left	2	Final
MFCC-MR	Middle-Right	26	Final
ΔMFCC-MR	Middle-Right	26	Final
f_0- MR	Middle-Right	2	Final
Δf_0- MR	Middle-Right	2	Final
MFCC-E	Ending	26	Final
ΔMFCC-E	Ending	26	Final
f_0-E	Ending	2	Final
Δf_0-E	Ending	2	Final
Total	290		

After the prediction of the designed acoustic features, the target cost C_{target} that describe the similarity between the candidate units and the target units is calculated as follow:

$$C_{target} = \alpha_D \|x_{Dc} - x_{Dt}\|_2^2 + \alpha_M \|x_{Mc} - x_{Mt}\|_2^2 + \alpha_F \|x_{Fc} - x_{Ft}\|_2^2 \quad (2)$$

where $\|\cdot\|_2$ is the L_2 norm. α_D, α_M and α_F are pre-defined weights to balance each type of acoustic features x_D, x_M and x_F denote the designed acoustic features of the duration, MFCC and f_0 respectively, in which subscript c represents the candidate unit and subscript t represents the target unit.

The combined cost is defined as:

$$\mathrm{C} = \sigma \mathrm{C}_{target} + \mathrm{C}_{concatenation} \quad (3)$$

where σ denotes the target cost weight, C_{target} denotes the target cost, and $C_{\mathrm{concatenation}}$ denotes the concatenation cost.

The calculation of the concatenation cost for the speech synthesizer is described in [18].

C. Designed Features prediction with multi-task learning method

Multi-task learning (MTL) is also a way to directly predict the designed acoustic features for the three different but related tasks. The prediction of above designed acoustic features can be split into three sub-tasks: Duration, f_0 and MFCC. Square loss is used as the cost function for each sub-task. All the three models are trained together with the global combined cost function:

$$F_g = \alpha F_D + \beta F_F + (1 - \alpha - \beta) F_p \quad (4)$$

where F_D , F_F and F_p are the error costs generated by each sub-task (Duration, f_0 and MFCC, respectively). The coefficient α and β are parameters that need manually adjusted. While in the training process of PDNN, F_p and F_s are cost functions for each task.

III. Experiments

A. Database and features

A Mandarin database, which contains 30,000 phonetically rich sentences from a professional male broadcaster, is adopted in this paper. For the experiments described in this paper, the audio was down-sampled to 16 kHz. In the training process of PDNN, there are 27,000 sentences as training set, 1,500 sentences as validation set, and the rest 1,500 sentences are reserved as test set.

The linguistic features, which contain the phonetic and prosodic contexts of Mandarin in each unit, can be included as follow: The phone identity, the position of a phone, syllable and word in phrase and sentence, POS of word, prosodic phrase, intonational phrase and sentence, the length of prosodic word, prosodic phrase, intonational phrase and sentence, etc. The dimension of the linguistic features is 504.

The continuous acoustic features are normalized to the range of $(0,1]$ and the discrete acoustic features are encoded in One-Hot.

B. Experimental setup

The baseline we select is the BLSTM based hybrid method.

- **Baseline**: It uses BLSTM based SPSS system to generate the acoustic parameters first and then calculate the KL divergence to get the final target cost.

To compare the PDNN model with other models, the conventional DNN based prediction method and the DNN based prediction with MTL method are added as extensive comparative experiments. Besides, we also conduct a series of experiments on the PDNN prediction framework with different predicting sequences and quantity of hierarchies. Four types of systems are implemented for comparison:

- **DNN-C**: Standard DNN-based approach. All the acoustic features are concatenated together and treated as one stream. The intermediate features are not predicted in the model.
- **DNN-I**: Different from DNN-C, each type of the acoustic features (Duration, MFCC, f_0) is trained independently with separate DNN models.
- **MTL-DNN**: DNN approach with MTL method. Different form DNN-C, one DNN model trains the three sub-tasks global combined cost function. Different coefficients α and β are tested.
- **PDNN**: The proposed PDNN method. We first predict the duration and the intermediate features (MFCC and f_0) that represent the whole unit. In the second task, the f_0 in each section of the unit is predicted. In the third task, the MFCC in each section of the unit is predicted. Other different quantities of tasks and different prediction sequences are tried.

After the cost calculation, a Viterbi search is used to find the best sequence that minimizes the combined cost. Except for the calculation of the target cost, the other modules are described in [18]. Our implementation is in TensorFlow [19] and we use the RMSProp optimizer with the global initial learning rate 0.0005 and its Tensorflow defaults parameters. We choose ReLU [20] as the activation function.

C. Objective evaluation

To evaluate the accuracy of the predicted designed acoustic features cost, the root mean square error (RMSE) between the predicting acoustic and the label is chosen as the objective metric for the f_0 and MFCC. For the duration features, we choose F1-score that reveal both precision and recall as the objective metric. TABLE II shows the Objective measures of different models for acoustic features prediction.

D. Subjective evaluation

To evaluate the performance of unit-selection synthesizer with the modification of the target cost, 30 native speakers are arranged to evaluate the synthetic speech based on a 5-point discrete scale Mean Opinion Scores (MOSs) [21] labeled “Bad”, “Poor”, “Fair”, “Good”, and “Excellent”. Each listener listens to 30 pairs random selecting sentences synthesized from five different systems. Different target cost weight σ is tested to investigate the contribution of the target cost to the whole unit selecting procedure.

IV. RESULTS

As illustrated in Table II, the MTL-DNN method achieves a better performance in predicting the f_0 and MFCC compared with standard DNN methods. There are correlations between f_0 and MFCC. The combination of their training can improve the performance. The performance of the duration prediction drops a little because of the weak connection between the duration and the acoustic features (f_0 and MFCC). The cost of the duration is not minimized to its own optimal. The proposed PDNN method has relative improvement in all the three objective measures. The knowledge transfer flow is defined by the lateral connections of PDNN. It illustrates that the goal of the MTL method is to minimize the combined global loss function, which is relevant to the objective measures of duration, f_0 and MFCC. Through epochs of training, it would reach the optimum. But it won't be easy for each sub-target to reach its own optimum because other targets would affect the parameters of the entire networks. For MTL, it is hard to distinguish which parameters to learn the specific. On the contrary, it is easy to distinguish for PDNN because each task is trained by each sub-network. The transfer of memory depends on the lateral connections between these sub-networks.

One thing we need to consider is the sequence of targets. Therefore, we did sets of the experiments on the sequence of predictions. According to the results, we can draw the conclusion that which target is predicted later, the better performance we can get. Predicting the f_0 parameters first has a better overall performance than predicting the MFCC first. We infer that it is more helpful for f_0 to reconstruct the MFCC.

TABLE II OBJECTIVE MEASURES OF DIFFERENT MODELS FOR ACOUSTIC FEATURES PREDICTION

Model		Duration F1-score	$Log f_0$ RMSE	MFCC RMSE
DNN	DNN-C	0.684	0.107	0.231
	DNN-I	0.763	0.098	0.223
	MTL-DNN	0.707	0.092	0.216
PDNN	DF, M	0.735	0.096	0.203
	DM, F	0.727	0.093	0.205
	D, F, M	0.763	0.091	0.199
	D, M, F	0.763	0.087	0.202
	D&FM-i, F, M	0.761	0.083	0.188
	D&FM-i, M, F	0.761	0.081	0.191

[a] D, F, M are short for duration, f_0 and MFCC.

[b] FM-i is short for the intermediate features (MFCC and f_0) that represent the whole unit.

[c] The second column on the left show the predicting sequences.

The introduction of the intermediate features (FM-i) that represent the f_0 and MFCC of the whole unit to the training procedure can improve the performance of predicting f_0 and MFCC in each section of the unit. Although there is a little drop in the F1-score of the duration, the FM-i can help the following details reconstruction of the acoustic features.

Observing the MOS results illustrated in the Fig. 3, the best MOS is 4.14 when the target cost weight is 1.5 in the PDNN experiment. The proposed PDNN method to predict the designed acoustic features for target cost calculation achieve better performance than the baseline. The direct features prediction avoids the accumulation of errors, which could generate more precise target cost to measure the similarity between the target units and candidate units. The MTL-DNN method outperforms the DNN-I and DNN-C methods. We infer that the correlation of these acoustic features could make contribution to the model prediction training. And a better way to joint train all the acoustic features is the key. The MTL-DNN method could balance each type of acoustic features better than the DNN-C method. The PDNN method has improvement in the MOS results than the MTL-DNN method because the PDNN model realize the physical isolation of each sub-task.

The MOS results of BASELINE and DNN-C system both decrease when the weight of the target cost increase. It indicates that target cost calculated by the two methods could not reflect the similarity properly and the systems mainly depend on the concatenation cost to select the candidates. The target cost losses its designed function. Meanwhile, the system with the PDNN method perform better when the weight of the target cost is increasing in certain range. It illustrates that the target cost calculated by the PDNN method is more precise to measure the similarity between the candidates and the targets and can help the unit-selection select the candidates better.

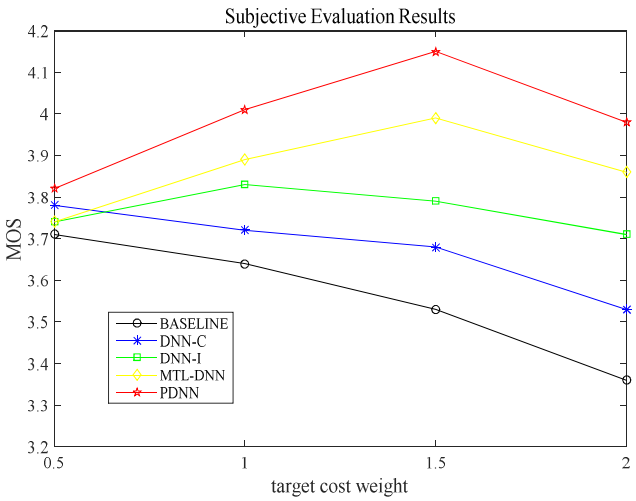

Fig. 3. MOS test for naturalness of synthetic speech using different target cost weight σ and systems.

V. CONCLUSION

In this paper, we present a progressive deep neural networks framework to predict the designed acoustic features directly for target cost calculation, which could predict different types of acoustic features one by one. The PDNN is immune to forget the previous memory on processing the linguistic features. Each training process has its own train criterion. Therefore, each type of acoustic features can be trained to its own optimum. PDNN with different predicting sequences, quantity of hierarchies are compared in the experiments. Compared to the baseline using the BLSTM-guided hybrid method, the MOS results demonstrated the better performance of our proposed PDNN method.

This paper focus on predicting the acoustic features for target cost calculation more directly and reasonably. However, the defined target cost in our method still involves human knowledge. In the future, the similarity between audio still need to be explored by using more delicate semi-supervised methods. Besides, the target cost is only part of the unit-selection synthesizer, the more direct method to choose a proper sequence of candidate units without calculating the target cost and the concatenation cost first is also our research focus.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (NSFC) (No.61425017, No. 61773379, No. 61603390, No. 61771472), the National Key Research & Development Plan of China (No. 2017YFB1002801) and Inria-CAS Joint Research Project (173211KYSB20170061).

REFERENCES

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

[2] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," Speech Communication 51.11: 1039-1064,2009.

[3] A. V. D. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves , et al, "WaveNet: A generative model for raw audio," arXiv preprint arXiv:1609.03499,2016, unpublished.

[4] A. V. D. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals , K. Kavukcuoglu, et al. ,"Parallel WaveNet: Fast High-Fidelity Speech Synthesis," arXiv preprint arXiv:1711.10433 ,2017 , unpublished.

[5] S. O. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang ,et al. "Deep Voice: Real-time Neural Text-to-Speech," in ICML, International Conference on Machine Learning, 2017

[6] S. Arik, G. Diamos, A. Gibiansky, J. Miller, K. Peng, W. Ping, et al. "Deep Voice 2: Multi-Speaker Neural Text-to-Speech," in NIPS-Annual Conference on Neural Information Processing Systems, 2017

[7] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, et al, "Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning," arXiv preprint arXiv:1710.07654 ,2017.

[8] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, et al, "Tacotron: Towards End-to-End Speech Synthesis," in INTERSPEECH 2017–Annual Conference of the International Speech Communication Association,2017,4006-4010.

[9] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang ,et al, "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions," in ICASSP-2018-IEEE International Conference on Acoustics, Speech, and Signal Processing, 2018.

[10] Z.-H. Ling, L. Qin, H. Lu, Y. Gao, L. Dai, R. Wang ,et al, "The USTC and iflytek speech synthesis systems for Blizzard Challenge 2007," in Blizzard Challenge Workshop, 2007.

[11] R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory, "Using deep bidirectional recurrent neural networks for prosodictarget prediction in a unit-selection text-to-speech system," in INTERSPEECH 2015–Annual Conference of the International Speech Communication Association, pp. 1606–1610.

[12] T. Merritt, R. A. Clark, Z. Wu, J. Yamagishi, S. King, "Deep neural network-guided unit selection synthesis," in ICASSP-2016- IEEE International Conference on Acoustics, Speech, and Signal Processing, 2016, pp. 5145–5149.

[13] L.-H. Chen, Y. Jiang, M. Zhou, Z. L., L. D., "The USTC system for Blizzard Challenge 2016," in Blizzard Challenge Workshop, 2016.

[14] T. Merritt, R. A. J. Clark, Z. Wu, J. Yamagishi, S. King, "Deep neural network-guided unit selection synthesis," in ICASSP-2016-IEEE International Conference on Acoustics, Speech, and Signal Processing, March 2016, pp. 5145–5149.

[15] W. Vincent, A. Yannis, S. Hanna, V. Jakub, W. Vincent, A. Yannis,et al, "Google's Next-Generation Real-Time Unit-Selection Synthesizer Using Sequence-to-Sequence LSTM-Based Autoencoders," in NTERSPEECH 2017–Annual Conference of the International Speech Communication Association, 2017:1143-1147.

[16] W.-S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 35, no. 3, pp. 653–668, 2013.

[17] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, "Progressive neural networks," arXiv: 1606.04671, unpublished.

[18] J.Tao, R. Fu, Y. Zheng, Z. Wen, L. Li, B. Liu, "The NLPR Speech Synthesis entry for Blizzard Challenge 2017 " in Blizzard Challenge Workshop, 2017.

[19] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, "Tensorflow: large-scale machine learning on heterogeneous distributed systems", 2016.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in NIPS- Annual Conference on Neural Information Processing Systems, 2012, pp. 1106–1114.

[21] Karimian-Azari, Sam, Nasser Mohammadiha, Jesper R. Jensen, and Mads G. Christensen. "Pitch estimation and tracking with harmonic emphasis on the acoustic spectrum." in ICASSP-2015-IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 4330-4334, 2015.