Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020
30 October 2020, Shanghai, China

The NLPR Speech Synthesis entry for Blizzard Challenge 2020

Tao Wang[1,2], *Jianhua Tao*[1,2,3], *Ruibo Fu*[1,2], *Zhengqi Wen*[1], *Chunyu Qiang*[1,2]

[1]National Laboratory of Pattern Recognition, CASIA, Beijing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3]CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China

{tao.wang, jhtao, ruibo.fu, zqwen, chunyu.qiang}@nlpr.ia.ac.cn

Abstract

The paper describes the NLPR speech synthesis system entry for Blizzard Challenge 2020. More than 9 hours of speech data from an news anchor and 3 hours of speech from one native Shanghainese speaker are adopted as training data for building system this year. Our speech synthesis system is built based on the multi-speaker end-to-end speech synthesis system. LPCNet based neural vocoder is adapted to improve the quality. Different from our previous system, some improvements about data pruning and speaker adaptation strategies were made to improve the stability of our system. In this paper, the whole system structure, data pruning method, and the duration control will be introduced and discussed. In addition, this competition includes two tasks of Mandarin and Shanghainese, and we will introduce the important parts of each topic respectively. Finally, the results of listening test are presented.

Index Terms: speech synthesis, LPCNet, end-to-end, Blizzard Challenge 2020

1. Instruction

This paper describes details about our seventh entry speech synthesis for Blizzard Challenge. The task of this year is to build a voice from the provided data, suitable for expressive text-to-speech (TTS) from plain text input. Mandarin Chinese speaker collected from news anchor is full of expressiveness. The corpus contains some colloquial scenes, which brings some difficulties to the model training task. In addition, there is no mature front-end tool in Shanghai dialect, which is also a difficulty in the task of synthesizing Shanghainese.

End-to-end speech synthesis has made great progress in recent years and achieved state-of-art performance [1, 2, 3]. The single-speaker speech synthesis system has achieved results comparable to human pronunciation. The interests in style control and speaker adaptation speech synthesis with limited corpus keep rising. Recently, there have been many studies in this topic, such as transferring prosody and speaking style within or cross speakers based on end-to-end speech synthesis model [4, 5, 6].

In general, there are mainly two aspects on speaker adaptive training methods. One aspect is the speaking style representations, which is the one of inputs in the system. The methods mainly used fixed global speaker style representations for speaker recognition, such as i-vectors [7], d-vector [8]. It is not optimal for the multi-speaker speech synthesis and adaptation task. Therefore, methods [9, 10] that extracted trainable speaker representations from waveform were proposed in the statistic parametric TTS framework. In the end-to-end TTS framework, [2, 11] use trainable speaker embeddings for multi-speaker speech synthesis. Besides, in [5], a method that learning a latent embedding space of prosody is proposed to be used as as extension to the Tacotron-based TTS architecture. Another aspect is vocoder, which is the output of the system. Speaker dependent layers with vocoder STRAIGHT [12], WORLD [13] is applied in the conventional statistical parametric speech synthesis method. And the speaker embedding information was also applied in the neural network based vocoder like WaveNet [14], SampleRNN [15] in the voice conversion tasks [15, 16, 17, 18].

To select a vocoder which is suitable for adaptative training and online system. Recent research work [19] reported that LPCNet could achieved higher quality than WaveRNN [20] with the same network size. LPCNet speech synthesis is achievable with a complexity under 3 GFLOPS. Therefore, we use LPCNet neural vocoder as a replacement for Griffin-Lim audio generation. And we used the trainable speaker embedding from Tacotron to do an adaptative training on LPCNET.

Considering the low quality of provided recording, we use the following strategies to improve the performance of our system. Firstly, we use a data pruning model to automatically select the unmatching data pairs, which improves the efficiency of manual checking process and shorts the time on the database building. Combining with manual checking, the performance of our submitted final system is improved significantly. Secondly, we add phone duration predicting model to control the speech generation by Tacotron, which prevents the speed is too high. It improves the intelligibility and rhythm of synthetic speech. Thirdly, we use bandwith expansion technologies by adding different style embeddings to further improve the quality of synthetic speech based on our own external data. Fourthly, the adaptative LPCNet is adopted to further improve the quality of synthetic speech.

The rest of the paper is organized as follows. Section 2 describes the overview of our methods. Section 3 presents details about the database processing and phone duration model. The results and analysis are presented in section 4. The conclusions and future work are presented in section 5.

2. System Overview

The whole network consists of two components, which is shown in the Figure 1. We add speaker identity information on the basis of Tacotron proposed in [20] to train a multi-speaker speech synthesis system. In the acoustic model part, we follow a modified scheme based on [11] to model multiple speakers. For each speaker in the dataset, a speaker embedding vector is initialized with Glorot [21] initialization. We form the separate speaker embeddings. For each step of the training process, the speaker embeddings would be updated. For each example, we concatenate the d-dimensional speaker embedding to encoder, decoder and attention of the Tacotron. In the neural vocoder part, we deploy the LPCNet, which significantly improve the efficiency of speech synthesis and remain high quality. In the frame rate

10.21437/VCC_BC.2020-12

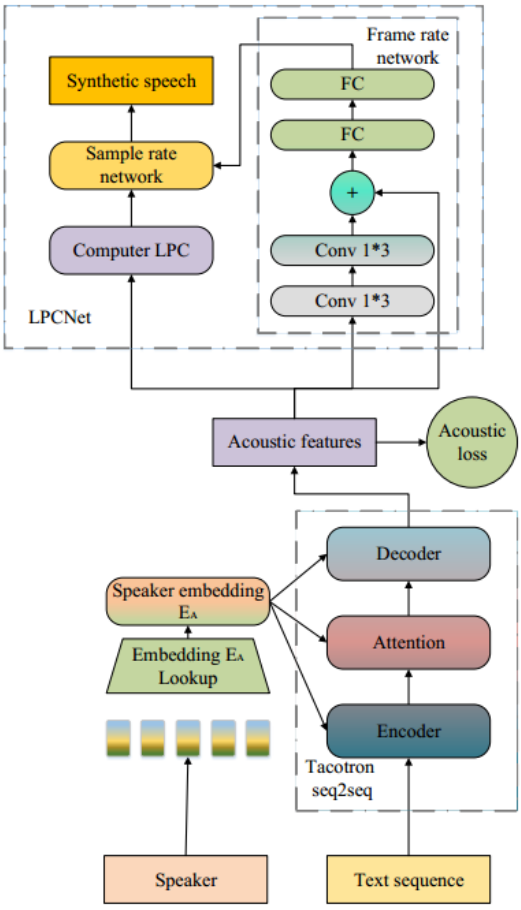

Figure 1: *An overview of our system.*

network of LPCNet, we combine the trainable speaker embeddings from Tacotron with the output of convolution layer and the acoustic features that Tacotron predict.

2.1. Mandarin Database Processing

The database of this year task is a challenge. Firstly, the speed of the talking speaker is high, which makes it hard for manual segmentation. We use an ASR model and silence detection model to automatic segmentation of the provided data. But there are still a lot of mistakes in the corpus. We use the forced alignment by the ASR technologies to further check the matching between the audio and the text. Furthermore, we also use a trained model to eliminate the unmatching audio-text data pairs. The model can the find the mistakes of the text more thoroughly than the forced alignment by ASR. Due to the low quality of the provided corpus, we also use external data to improve the performance of submitted system. The external data could improve the stability and quality of our system. All the data is processed by the same technologies and checked by annotator.

2.2. Shanghainese Database Processing

In the Shanghaiese task, we build a front-end system suitable for Shanghainese based on Wuyu phonemes in Table 2. The phoneme labeling constructed by this method is more in line with the actual pronunciation rules of Shanghainese.

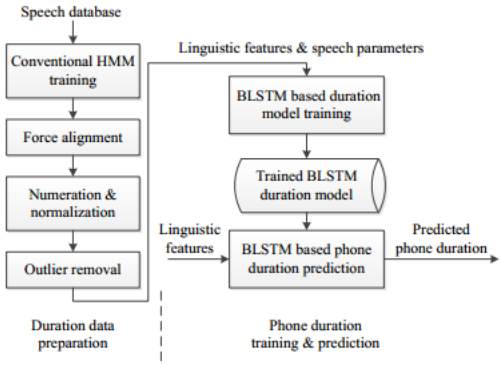

Figure 2: *BLSTM based phone duration predition in our system.*

Table 1: *Composition of external Mandarin database (M=Male; F= Female)*

Sentence number	Training	Validation	Test	Speaker
Large set	9,000×8	500×8	500×8	2 M 6 F
Small set	900×4	50×4	50×4	1 M 3 F
Total	75,600	4200	4200	3 M 9 F

2.3. Phone Duration Modeling

The prosody in the provided database is not very well, thus we use external data to train a duration model separately. The duration model is added to the input of end-to-end system by combining the word embedding and duration embedding together to control the prosody output to adapt different styles. In these tasks, its powerful sequence modeling has been proved. Here, we consider duration modeling at the phone level, for the audiobook data. BLSTM based duration method with outlier removal is shown in Figure 2. It is a general framework, which is easy to be replaced by other neural network modules. We will present some important details of duration model below.

In duration data preparation part, force alignment is used at the phone level after conventional HMM training. This step is to segment each utterance into a sequence of shorter and simple speech units, thus each unit can be modelled independently in subsequent steps. Unlike the decision tree, BLSTM can only handle numeric features, thus it is necessary to encode all nominal features to be numeric values. Normalization is immediately carried out to transfer feature values into a limited interval.

3. System Building

In the inference, we just need to specify the initial length which is usually the max length of frames. Then after given a text sequence and mask sequences of frames, the decoder1 module will predict the coarse acoustic features and stop tokens. We mask the additional information of coarse acoustic features with the information of the stop token to input to decoder2. After decoder2, we can get final predicted features.

3.1. Speech database

The mandarin database is an estimated 9.5 hours of speech from one native Mandarin Chinese speaker collected from the news program for the Blizzard Challenge 2020. Although it is the voice of the news broadcast, but the host's speaking style is very distinctive. The Shanghai database is about 3 hours of speech

Table 2: *Wuyu phonemes list.*

a	iU	ts\
A	i@U	ts
o	y{	tsh
@U	yE	ts‘
@	ua	ts‘h
e	uA	ts\h
7	u@	x
U	ue	f
u	uo	s
i	:\i	s‘
i\	r\‘	s\
i‘	:n	m
y	N	n
AU	p	l
ia	ph	z‘
ia	t	w
iAU	th	j
ie	k	
iE	kh	

from one native Shanghainese speaker.

We also build a multi-speaker speech synthesis system by adding external database. The composition of the external Mandarin database is shown in the Table 1. All the wav files are sampled at 24kHz. In this work, we limit the input of the synthesis to just 22 features: The 20-dims Bark-scale cepstral coefficients, and 2 pitch parameters (period, correlation) are extracted directly from recorded speech samples. The cepstrum uses the same band layout as [22] and the pitch estimator is based on an open-loop cross-correlation search. The input text is processed by our G2P frontend to transform to the phone sequences with tone in vowel.

3.2. System Building

For the Tacotron training, we set output layer reduction factor r=2 . We use the Adam optimizer [23] with learning rate decay, which starts from 0.001 and is reduced to 0.0005, 0.0003, and 0.0001 after 500K, 1M and 2M global steps, respectively. The post-processing net is discarded. We use a simple loss for seq2seq decoder, which is the acoustic loss. Besides, the "stop token" prediction [1] is used to allow the model to stop dynamically, which is the stop token loss. The combined cost is the sum of PIP loss, acoustic loss and stop token loss with equal weights. For the Tacotron adaptation, the learning rate is set to 0.0001. After about 600K global steps, there are about 2-3K global steps for adaptative training. For the LPCNet training, the network was trained for 120 epochs, with a batch size of 64, each sequence consisting of 15 10-ms frames. We use the AMSGrad [24] optimization method (Adam variant) with a step size $\alpha = \alpha_0/(1 + \delta \times b)$ where $\alpha_0 = 0.001$, $\delta = 5 \times 10^{(-5)}$, and b is the batch number. For the LPCNet adaptation, there are about 10 epochs for adaptative training. Our proposed system consists two parts: Acoustic model based on Tacotron, neural vocoder based on LPCNet. Therefore, we have following groups of baseline systems.

- **Taco Baseline1 (Mono-Taco)** The basic structure is similar as Tacotron 2 , only provided data is used to bulid a mono speaker speech synthesis system.
- **Taco Baseline2 (Multi-Taco-1)** The basic structure is similar as Tacotron 1, which has a post-processing net after decoder layer. We replace the spectral magnitude to the vocoder parameters for WORLD or LPCNet. And we add stop prediction and speaker embeddings.
- **Taco Baseline3 (Multi-Taco-2)** Compared with Taco Baseline 1, The post-processing net is discarded just like our proposed method. Only one speaker embedding is deployed to train together with Tacotron.

Table 3: *Abbreviations for Acoustic model + Vocoder*

	WORLD	LPCNet
Mono-Taco	M-T-W	M-T-L
Multi-Taco-1	T-1-W	T-1-L
Multi-Taco-2	T-2-W	T-2-L
Multi-Taco-Cleaned	T-C-W	T-C-L
Multi-Taco-Duration	T-D-W	T-D-L

Table 4: *Preference scores on quality of synthetic speech*

System A	Scores A(%)	Scores Neutral(%)	Scores B(%)	System B
T-1-W	14.67	16.98	68.35	T-2-L
T-2-W	11.67	13.73	74.6	T-1-L
T-1-W	40.45	25.82	33.73	T-2-W
T-1-L	36.42	29.94	33.64	T-2-L

To compare the performance of vocoders, we set WORLD and LPCNet as two baselines.The following Table 3 is the abbreviations for several acoustic model + vocoder combinations.

4. Experiments and evaluation

To evaluate our proposed method, we conduct an evaluation to validate the effectiveness of the our submitted system.

4.1. Quality

We first compare different vocoders and different Tacotron. By observing the preference scores of subjective evaluations in the Table 4, it is worth noticing that the LPCNet can significantly improve the quality of synthetic speech. Besides, the Tacotron with post-processing net can not improve the quality of synthetic speech significantly. The dimensionality of acoustic features for LPCNet is 22, while the dimensionality of acoustic features for WORLD is 187. The discard of post-processing net and decreasing the dimensionality of predicting acoustic features can reduce the complexity of Tacotron model. Therefore, in the following sections we mainly concentrate on Tacotron 2 structure.

4.2. Naturalness

By observing the preference scores of subjective evaluations in the Table 5, it is worth noticing that by using model to clean the database and adding the duration control can improve the performance of the submitted system.

4.3. Similarity

By observing the preference scores of subjective evaluations in the Table 6, it is worth noticing that the deployment of LPCNet can improve the similarity of synthetic speech. By using adaptative training on the LPCNet, the neural vocoder can be more

Table 5: *Preference scores on naturalness of synthetic speech*

System A	Scores A(%)	Scores Neutral(%)	Scores B(%)	System B
T-C-L	48.93	13.26	37.81	T-2-L
T-D-L	62.48	17.14	20.38	T-2-L
T-D-L	52.69	8.94	38.37	T-C-L

Table 6: *Preference scores on similarity of synthetic speech*

System A	Scores A(%)	Scores Neutral(%)	Scores B(%)	System B
T-C-L	53.62	6.94	39.44	T-C-W
T-D-L	58.73	12.75	28.52	T-D-W

suitable for the target speaker.

4.4. Evaluation results

We conduct Mean Opinion Score(MOS) listening test for naturalness and similarity of speech. 20 listeners participated in the evaluation. In each experimental group, 20 parallel sentences are selected randomly from testing sets of each system. The results are shown in Figure 3 and indentity of our system is C. For all these evaluation (naturalness, similarity) results, our system only ranks average level.

4.5. Discussion of the results

From the result of evaluation, there is still a gap between our system to the top one. There are many reasons leading to this results. And the mainly one is that LPCNet neural vocoder we used had a large gap with WaveNet on the quality. The big background noise produced by the LPCNet lead to a significant influence on the perception, which leads to a low scores on the MOS results. These results reminder us there are still many works need to be done, especially on improving the quality of the synthetic speech.

5. Conclusion

In this paper, the multi-speaker end-to-end speech synthesis system built for Blizzard Challenge 2020 by NLPR is introduced. There are several improvements from our previous Challenge system. The first one is the use of end-to-end system. The second one is the use of data pruning and speaker adaptation strategies. The final one is the duration control approaches. The internal evaluation results show that the effectiveness of these three techniques. Also, the evaluation results from the Blizzard Challenge committee shows that, the naturalness, similarity of our system is of average level. Many works need to be done, especially on improving the quality of the synthetic speech.

6. Acknowledgements

This work is supported by the National Key Research Development Plan of China (No.2017YFC0820602), the National Natural Science Foundation of China (NSFC) (No.61425017, No.61831022, No.61773379, No.61603390), the Strategic Priority Research Program of Chinese Academy of Sciences (No.XDC02050100), and Inria-CAS Joint Research Project (No.173211KYSB20170061).

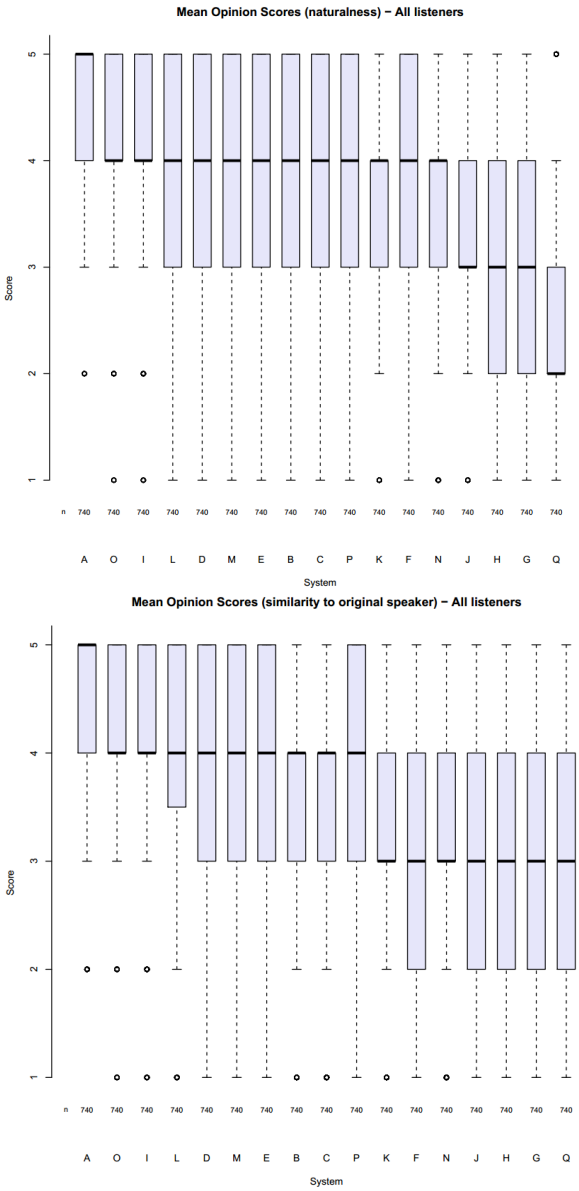

Figure 3: *Mean Opinion Scores (ALL Listeners).*

7. References

[1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[2] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," *arXiv preprint arXiv:1710.07654*, 2017.

[3] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 6706–6713.

[4] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," *arXiv preprint arXiv:1803.09017*, 2018.

[5] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," *arXiv preprint arXiv:1803.09047*, 2018.

[6] D. Stanton, Y. Wang, and R. Skerry-Ryan, "Predicting expressive speaking style from text in end-to-end speech synthesis," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 595–602.

[7] Z. Wu, P. Swietojanski, C. Veaux, S. Renals, and S. King, "A study of speaker adaptation for dnn-based speech synthesis," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[8] Y. Zhao, D. Saito, and N. Minematsu, "Speaker representations for speaker adaptation in multiple speakers blstm-rnn-based speech synthesis," *space*, vol. 5, no. 6, p. 7, 2016.

[9] M. Wan, G. Degottex, and M. Gales, "Waveform-based speaker representations for speech synthesis," 2018.

[10] R. Fu, J. Tao, Z. Wen, and Y. Zheng, "Phoneme dependent speaker embedding and model factorization for multi-speaker speech synthesis and adaptation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6930–6934.

[11] A. Gibiansky, S. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," in *Advances in neural information processing systems*, 2017, pp. 2962–2970.

[12] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[13] 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. 99, no. 7, pp. 1877–1884, 2016.

[14] A. v. d. 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 preprint arXiv:1609.03499*, 2016.

[15] K. Chen, B. Chen, J. Lai, and K. Yu, "High-quality voice conversion using spectrogram-based wavenet vocoder." in *Interspeech*, 2018, pp. 1993–1997.

[16] L.-J. Liu, Z.-H. Ling, Y. Jiang, M. Zhou, and L.-R. Dai, "Wavenet vocoder with limited training data for voice conversion." in *Interspeech*, 2018, pp. 1983–1987.

[17] B. Sisman, M. Zhang, and H. Li, "A voice conversion framework with tandem feature sparse representation and speaker-adapted wavenet vocoder." in *Interspeech*, 2018, pp. 1978–1982.

[18] C. Zhou, M. Horgan, V. Kumar, C. Vasco, and D. Darcy, "Voice conversion with conditional samplernn," *arXiv preprint arXiv:1808.08311*, 2018.

[19] J.-M. Valin and J. Skoglund, "Lpcnet: Improving neural speech synthesis through linear prediction," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5891–5895.

[20] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[21] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.

[22] J.-M. Valin, "A hybrid dsp/deep learning approach to real-time full-band speech enhancement," in *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2018, pp. 1–5.

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

[24] T. Tan, S. Yin, K. Liu, and M. Wan, "On the convergence speed of amsgrad and beyond," in *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, 2019, pp. 464–470.