FOCUSING ON ATTENTION: PROSODY TRANSFER AND ADAPTATIVE OPTIMIZATION STRATEGY FOR MULTI-SPEAKER END-TO-END SPEECH SYNTHESIS

Ruibo Fu [1,2], Jianhua Tao [1,2,3], Zhengqi Wen [1], Jiangyan Yi [1], Tao Wang [1,2]

[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, 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
{ruibo.fu, jhtao, zqwen, jiangyan.yi,tao.wang}@nlpr.ia.ac.cn

ABSTRACT

End-to-end speech synthesis can generate high-quality synthetic speech and achieve high similarity scores with low-resource adaptation data. However, the generalization of out-domain texts is still a challenging task. The limited adaptation data leads to unacceptable errors and the poor prosody performance of the synthetic speech. In this paper, we present two novel methods to handle the above problems by focusing on the attention. Firstly, compared with the conventional methods that extract prosody embeddings for conditioning input, a duration controller with feedback mechanism is proposed, which can control the states transition in the sequence-to-sequence model more directly and precisely. Secondly, to alleviate the unmatching text-audio pairs' impact on model, an adaptative optimization strategy which would consider the matching degree of the training sample is also proposed. Experimental results on Mandarin dataset show that proposed methods lead to an improvement on both robustness and overall naturalness.

Index Terms— prosody transfer, optimization strategy, speaker adaptation, attention, speech synthesis

1. INTRODUCTION

End-to-end speech synthesis, such as Tacotron, can achieve the state-of-art performance, and even close to human recording for in-domain text [1-4]. However, the generalization to out-domain is still a challenge, especially in the circumstance that the target speaker data for adaptation training is very limited. A lot of unacceptable errors could occur, including skipping, repeating, mispronunciation, and etc. Besides, the prosody performance is usually poor due to the large range of unsimilar domain text for adaptation model.

Generally, the above low performance on generalization and prosody can be ascribed to two aspects: attention mechanism and corpus. One aspect is the misalignments from attention. The attention is expected to predict the alignments between states from encoder and decoder [5], which has been widely applied in different fields [6-8]. Content-based [5] and location-based [9] attention have been applied in the speech synthesis [10,11]. To mitigate unacceptable errors, Tacotron2 [1] deploys local sensitive attention [12] to encourages the model to move forward consistently through the input. Furthermore, forward attention [13] and stepwise monotonic attention [14] are proposed to improve the robustness on out-of-domain scenarios for phoneme-based models. The above two methods assume that the alignments between encoder and decoder states is monotonically and continuously without skipping any encoder states, which may have a performance degradation when some inputs scripts have no mapping to the audio. For instance, there are some tokens representing the prosodic boundaries or richer tokens to describe a phone. Besides, all the above attention mechanisms are designed for mono speaker acoustic model, which ignores that diversity of prosody styles may result in different alignment patterns. The state-of-art adaptive frameworks can generate high similarity speech in low-resource speaker adaption task [15-17]. These researches mainly focus on the extraction of prosody or speaker embeddings, which would be the conditioning input. However, the diversity of states transition is not directly controlled and the prosody performance needs to be improved.

In this paper, we focus on attention to realize the robustness and prosody controllability for multi-speaker end-to-end speech synthesis system. We propose an improved attention based on the learned alignment. A duration controller structure is embedded in the seq-to-seq model to control the states transition. Inspired by the PID controller in the control theory [18], a feedback mechanism is proposed, which monitor the states hold and transition. On another aspect, the corpus is the foundation of a well-trained speech synthesis model. The unmatching text-audio training sample would cause errors that effects the naturalness. Conventional methods usually use manual checking or forced alignment by ASR technologies, which is time-consuming and low-accuracy respectively. We expect the corpus checking could be conducted with model training simultaneously and automatically. Therefore, an adaptative optimization strategy which would consider the matching degree of the training sample based on the attention mechanism is also proposed.

Overall, the contributions of this paper are two-fold. First, a duration controller with feedback mechanism is proposed. Second, an adaptative optimization strategy is proposed to immune to the unmatching training sample. Experiments demonstrate the better prosody, more robustness, and higher naturalness by applying proposed methods.

The rest of the paper is organized as follows. Section 2 describes methods. Experiments and results are analyzed in section 3 and 4. The conclusions are discussed in Section 5.

978-1-5090-6631-5/20/$31.00 ©2020 IEEE

2. METHOD

Fig. 1 shows the architecture of the Tacotron based multi-speaker end-to-end speech synthesis model. The whole network consists of two components. In the acoustic model part, Tacotron as proposed in [1] does not include explicit modeling of speaker identity; however, due to the flexibility of all neural sequence-to-sequence models, learning multi-speaker models via conditioning on speaker identity is straightforward. Besides, we add an extra prosody embedding to focus on modeling the variety of the duration distribution conditioned on the text information. In the training process, the prosody style could be identical to the speaker id, which means different people have different prosody styles. For the same speaker, the prosody styles could also be different if the types of recording text change from news to novel or something else. By this way, the well-trained prosody embedding with sufficient corpus can be used in the circumstance of limited adaption corpus to improve the prosody performance. The whole seq-to-seq acoustic model consists of three parts: Encoder mainly processes text information. The attention mechanism connects the encoder and decoder and controls the prosody. The decoder generates acoustic features conditioned on speaker embeddings to ensure high similarity of the target speaker. In the neural vocoder part, we deploy the LPCNet [19], which significantly improve the efficiency of speech synthesis and remain high quality. In the following sections, the duration controller and adaptive optimization strategy focusing on the attention mechanism would be introduced.

2.1. Duration controller with feedback mechanism

Generally, the attention-based encoder-decoder model is deployed for acoustic modeling in end-to-end speech synthesis. Therefore, it is difficult to control the duration just like conventional statistical parametric speech synthesis [20].

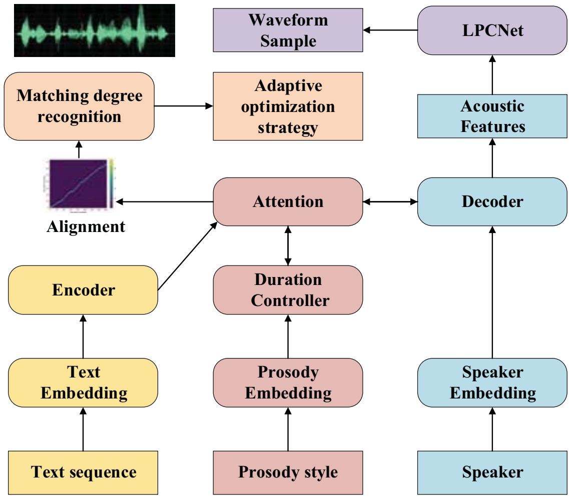


Fig. 1 System architecture of the Tacotron based multi-speaker end-to-end framework.

The alignment path in the attention mechanism indicates the mapping relation between text information and corresponding acoustic features. Inspired by the forward attention [11], we propose a more robust attention mechanism by adding the duration controller with feedback mechanism.

We assume that the alignment paths do not move strictly monotonically, which means there are several small mismatches between text and acoustic features due to the unsilenced inputs text or special language phenomena. More concretely, the attended phone should move forward to the following one, remain motionless or move backward to the previous one at the decoder timestep t. Given the encoding results x and query q_t, let $\beta_t(n)$ denote alignment results from local sensitive attention for the index n of x at the timestep t.The forward variable $\alpha_t(n)$ is defined as the new alignment results reweighted from $\beta_t(n)$. The $\alpha_t(n)$ can be calculated recursively from $\alpha_{t-1}(n)$, $\alpha_{t-1}(n-1)$ and $\alpha_{t-1}(n+1)$ as

$$\alpha_t(n) = (\gamma_{t-1}(0) \cdot \alpha_{t-1}(n) + \gamma_{t-1}(1) \cdot \alpha_{t-1}(n-1) + \gamma_{t-1}(2) \cdot \alpha_{t-1}(n+1)) \cdot \beta_t(n) \qquad (1)$$

where $\gamma_{t-1}(0), \gamma_{t-1}(1), \gamma_{t-1}(2)$ are the outputs of duration controller structure.

Then we define

$$\hat{\alpha}_t(n) = \alpha_t(n) / \sum_n \alpha_t(n) \qquad (2)$$

to normalize forward variable $\alpha_t(n)$. The reweighted context vector can be computed as

$$c_t = \sum_n^N \hat{\alpha}_t(n) \, x_n \qquad (3)$$

The duration controller is an DNN with two hidden layers and sigmoid output layer, which predicts the probability of the attended phone' next move (remain motionless: $\gamma_t(0)$, move forward: $\gamma_t(1)$, move backward: $\gamma_t(2)$). The inputs of duration controller consist of three parts: linguistic part, embedding part and duration feedback part. The linguistic part contains c_t and q_t , which tells the information of attended phone and utterance level context information. For each prosody style, an embedding vector $E_{prosody}$ is initialized with Glorot [21] initialization like speaker embedding, which could model the discrepancy.

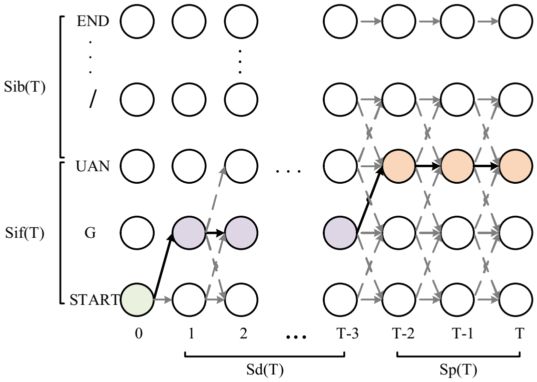


Fig. 2 Colored circles represent a possible alignment path. Same color means the current phone remain motionless. The feedback parameters $S_p(T)$, $S_{if}(T)$, $S_{ib}(T)$, $S_d(T)$ are illustrated by brackets.

Inspired by the PID control theory, we also design several feedback variables to observe current seq-to-seq model generation status for improving robustness. As shown in the Fig. 2, all the feedback variables at the decoder timestep T are illustrated. The steps for the current phone $S_p(T)$ is the direct control variable, which can been as the proportion part. $S_d(T)$ is the total steps of the last phone, which can be seen as the differentiation part. It can be interpreted as the estimated current generation velocity since the end of current phone is unknown. The forward position of current phone $S_{if}(T)$ and the backward position of current phone $S_{ib}(T)$ can be seen as the integration part, which accumulate the entire decoder processing. The complete duration controller with feedback mechanism is described in Algorithm 1.

2.2. Adaptative optimization strategy

Although the attention mechanism we proposed in the previous section loosens up the restriction on strictly monotonically assumption, the unmatching text-audio samples would cause the unacceptable errors. Therefore, we propose an adaptative optimization strategy to decrease the unmatching training samples' influence. During the training process, the learning rate of one step would be adjusted according to the matching degree of text-audio training sample in this batch. As shown in the Fig. 1, the matching degree recognition is based on the alignments results. The matching degree M for one sample σ is defined as

$$M(\sigma) = \left(\sum_{t=0}^{t\leq T_E} \max_{0\leq n\leq N}\{\hat{\alpha}_t(n)\}\right)/T_E \tag{4}$$

where T_E is length of the decoder step. A learning decaying strategy regarding to the training step is also deployed in our method. Let $L(i)$ denote the learning rate with decaying strategy at the training step i . The propose adaptative optimization strategy $\tilde{L}(i)$ is defined as

$$\tilde{L}(i) = L(i) \cdot \left(\sum_{\sigma\epsilon\Omega} M(\sigma)\right)/B \tag{5}$$

Algorithm 1 Duration Controller with Feedback Mechanism

```
Initialize:
  $\hat{\alpha}_0(1) \leftarrow 1$
  $\hat{\alpha}_0(n) \leftarrow 0, n = 2, \ldots, N$
  $\gamma_0(i) \leftarrow 1/3$ ,  $i = 0,1,2$
  $S_p \leftarrow 0, S_{if} \leftarrow 0, S_{ib} \leftarrow N, S_d \leftarrow 0, S_{lif} \leftarrow 0$
for t=1 to T do
  $\beta_t(n) \leftarrow Atten(x, q_t)$
  $S_{if} = \underset{0\leq n\leq N}{argmax}\{\beta_t(n)\}$
  $S_{ib} = N - S_{if}$
  if $S_{if} == S_{lif}$
    then
    $S_p = S_p + 1$
    else
    $S_d = S_p$
    $S_p = 0$
  $S_{lif} = S_{if}$
  $\hat{\alpha}'_t(n) \leftarrow (\gamma_{t-1}(0) \cdot \hat{\alpha}_{t-1}(n) + \gamma_{t-1}(1) \cdot \hat{\alpha}_{t-1}(n-1)$
          $+ \gamma_{t-1}(2) \cdot \hat{\alpha}_{t-1}(n+1)) \cdot \beta_t(n)$
  $\hat{\alpha}_t(n) \leftarrow \hat{\alpha}'_t(n) / \sum_{m=1}^{N} \hat{\alpha}'_t(n)$
  $c_t \leftarrow \sum_{m=1}^{N} \hat{\alpha}_t(n) x_n$
  $\gamma_t(0), \gamma_t(1), \gamma_t(2) \leftarrow \boldsymbol{DNN}(c_t, q_t, S_p, S_{if}, S_{ib}, S_d, E_{prosody})$
end for
```

where Ω denotes the set of all samples in a batch, and B denotes the batch size. For each batch of training sample, if some wrong labels occur, the model would be updated less in this step by adopting this adaptative optimization strategy.

3. EXPERIMENTAL SETUP

We use the Blizzard Challenge 2019 dataset and our own internal dataset to conduct the experiments. Our internal dataset consists of 25 different professional Mandarin speakers with about 200 hours. The Blizzard Challenge dataset is an estimated 8 hours of speech from one native Mandarin Chinese speaker collected from talk shows. All the wav files are sampled at 16kHz. In this work, we limit the input of the synthesis to 32 features: The 30-dim Bark-scale [22] cepstral coefficients, and 2 pitch parameters (period, correlation) are extracted directly from recorded speech samples. The input text is processed by our G2P frontend and transformed to the phone sequences, which also include tone information of vowels.

For the Tacotron training, we set output layer reduction factor $r = 2$. We use the Adam optimizer [23] with adaptative learning rate decay, which starts from 0.001 and decay as introduced in §2.2. The training batch size is 16, where all sequences are padded to a max length. For the low-resource adaptation task, after about 600K global steps, there are about 2-3K global steps for adaptative training.

For the LPCNet training, the network is 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 $\alpha = \alpha_0/(1 + \delta \cdot b)$ 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.

The models on which we conduct experiments include:
1. Baseline: Tacotron2 with location sensitive attention
2. Forward attention, with or without transition agent, which is denoted as FA+TA and FA w/o TA [13].
3. Stepwise monotonic attention: soft inference, which is denoted as SMA soft [14].
4. Our proposed method: To do ablation studies, we make several models. Duration Controller is denoted as DC. And fm, pe and aos are short for feedback mechanism, prosody embedding and adaptative optimization strategy respectively. For instance, DC-fm-pe-aos is our final proposed method.

We evaluate the performance of our models in terms of both intelligibility and naturalness. We conduct an automatic objective intelligibility evaluation by ASR model. The test sets are about 40,000 utterances, containing the in-domain and out-domain text, which involve news, encyclopedias, story and poetry. To evaluate naturalness, a subset is selected by sorting high frequency errors from ASR model based intelligibility evaluation. 30 listeners conducted crowd-sourcing ABX preference tests. In each experimental group, 30 parallel sentences are selected randomly from test subset for each system.

4. EVALUATION AND DISCUSSION

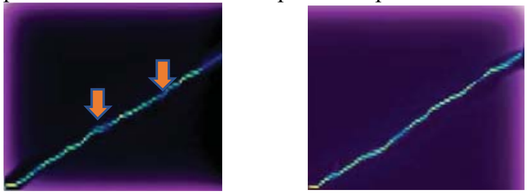
4.1. Complexity and convergence speed

The duration controller we proposed is only a DNN structure without any recurrent structure. The size of the model does not increase significantly, which is about 0.3%. During the training processing, we find that our proposed method DC and FA, SMA can achieve faster convergence speed. Besides, by adding feedback mechanism, the convergence speed is further improved. And a better alignment results can be observed in the Fig.3. We infer that the introduction of S_p in the feedback can help model distribute a more reasonable duration, which reduce the occurrence of vague alignments.

4.2. Intelligibility evaluation

Intelligibility tests are performed with metrics of case level unintelligible rate and number of words errors to evaluate the robustness of models. All the speakers are tested and three types of well-trained prosody are chosen in the DC ablation test. As shown in the Fig.4, baseline fails to produce intelligible results, while all the improved attention mechanism (FA, SMA, DC) can achieve better performance. By observing ablation test, the proposed prosody embedding, feedback mechanism and adaptative optimization can further improve the intelligibility. Besides, our submitted system for Blizzard Challenge 2019 rank fifth of 24 teams in two intelligibility evaluations, which also demonstrates the effectiveness of our proposed DC-fm-pe-aos methods. By sorting results from different speakers, the data size is also the key factor that effects the final performance.

Robustness of different attention mechanism are further demonstrated in Fig.4 and Table 1. The monotonicity assumption considering the speech characteristic which avoid the skipping. It can be seen the feedback in the DC can further reduce the occurrence of skipping and repeating. It can be interpreted that the adding of $S_{if}(T)$ and $S_{ib}(T)$ enhance the position information and improve the performance.



(a) DC without feedback (b) DC with feedback

Fig. 3 Attention alignments with the same text on a test utterance. Note that (a) has two points vague alignments.

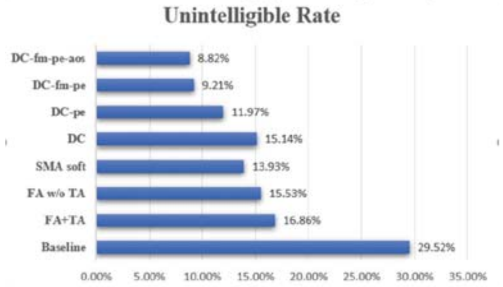


Fig. 4 Intelligibility results based on the ASR system

4.3. Naturalness evaluation

In this part, we evaluate the naturalness of the synthetic speech from different models for low-resource speaker adaptation. One male and one female target speaker is selected. Both speakers have recordings for about 15 minutes. One of well-trained prosody embedding by large dataset in multi-speaker model is selected in the speech generation process. By observing the preference scores results from Table 2, we can find that the proposed DC-fm-pe-aos method can have better naturalness. And the prosody embedding can transfer prosody styles to the low-resourced situation, where is no enough prosody information for adaption. The synthesized speech sounds closer to human. We infer that the proposed the duration controller play an important role.

Table 1 : Number of errors results

(Blizzard Challenge 2019 speaker, total 38360 words)

Model	Skipping	Repeating	Mispronunciation
Baseline	196	372	8962
FA+TA	149	265	5836
FA w/o TA	163	248	5925
SMA soft	126	132	4862
DC	139	148	5358
DC-pe	94	125	3852
DC-fm-pe	68	59	3294
DC-fm-pe-aos	**61**	**50**	**2861**

Table 2: Preference scores on naturalness of speech

System	Scores	Scores	Scores	System
A	**A (%)**	**Neutral (%)**	**B (%)**	**B**
DC-PE	69.23	17.69	13.08	**Baseline**
	56.81	10.24	32.95	**FA+TA**
	52.36	16.73	30.91	**SMA soft**
	41.47	26.15	32.38	**DC-ownPE**

Note: DC-PE represents the proposed DC-fm-pe-aos method choosing well-trained prosody embedding, while DC-ownPE represents that only its own prosody embedding is selected.

5. CONCLUSION

In this paper, we propose a duration controller with feedback mechanism and adaptative optimization strategy in the neural end-to-end speech synthesis system. Experimental results demonstrate that both the methods improve robustness, especially in the circumstance of low-resource speaker adaptation task. The introduction of prosody embedding in duration controller provides more and better prosody styles. These methods could be further applied to other sequence-to-sequence tasks similar to speech synthesis. In the future, we will investigate faster neural end-to-end speech synthesis framework.

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.61831022, No.61901473, No.61771472, No.61773379) and Inria-CAS Joint Research Project (No.173211KYSB20170061 and No.173211KYSB20190049)。

REFERENCES

[1] J. Shen, R. Pang, R. J. Weiss, et al, "Natural TTS synthesis by conditioningWaveNet on mel spectrogram predictions," in Proceedings ICASSP-2018-IEEE International Conference on Acoustics, Speech, and Signal Processing,2018, pp. 4779-4783.

[2] W. Ping, K. Peng, A. Gibiansky, et al, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning" , in Proceedings ICLR 2018-International Conference on Learning Representations, 2018.

[3] N.Li, S. Liu, Y. Liu, et al, "Close to human quality TTS with transformer," arXiv preprint arXiv:1809.08895, 2018.

[4] A. V. D. Oord, S. Dieleman, H. Zen, et al. "WaveNet: A Generative Model for Raw Audio," in Proceedings INTERSPEECH 2017 –Annual Conference of the International Speech Communication Association,2017.

[5] D. Bahdanau, K. Cho, Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," arXiv preprint arXiv:1409.0473, 2014.

[6] M. Luong, H. Pham, C. D. Manning, "Effective approaches to attention-based neural machine translation,"arXiv preprint arXiv:1508.04025, 2015.

[7] K. Xu, J. Ba, R. Kiros, et al. "Show, attend and tell: Neural image caption generation with visual attention," in Proceedings ICML 2015- International Conference on Machine Learning, 2015, pp. 2048–2057.

[8] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in Proceedings ICASSP 2016-IEEE International Conference on Acoustics, Speech, and Signal Processing, 2016, pp. 4960–4964.

[9] A. Graves, " Generating sequences with recurrent neural networks," arXiv preprint arXiv:1308.0850, 2013.

[10] W. Wang, S. Xu, B. Xu, "First step towards end-to-end parametric TTS synthesis: Generating spectral parameters with neural attention.," in Proceedings INTERSPEECH 2016 – Annual Conference of the International Speech Communication Association, 2016, pp. 2243–2247.

[11] J. Sotelo, S. Mehri, K. Kumar, " Char2Wav: End-to-end speech synthesis," in Proceedings ICLR2017 workshop submission, 2017

[12] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in Proc. NIPS, 2015, pp. 577–585.

[13] J. X. Zhang, Z. H. Ling and L. R. Dai, "Forward Attention in Sequence-to-Sequence Acoustic Modeling for Speech Synthesis," in Proceedings ICASSP 2018- IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 4789–4793, 2018.

[14] M He, Y Deng , L He, "Robust Sequence-to-Sequence Acoustic Modeling with Stepwise Monotonic Attention for Neural TTS," in Proceedings INTERSPEECH 2019 –Annual Conference of the International Speech Communication Association, 2019.

[15] RJ Skerry-Ryan, E. Battenberg, Y. Xiao, et al, "Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron,"in Proceedings ICML 2018-Proceedings of the 35th International Conference on Machine Learning, 2018, pp. 4700–4709.

[16] Y. Wang, D. Stanton, Y. Zhang, et al, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in Proceedings ICML 2018-Proceedings of the 35th International Conference on Machine Learning, 2018.

[17] D. Stanton, Y. Wang, and RJ Skerry-Ryan, " Predicting expressive speaking style from text in end-to-end speech synthesis," arXiv preprint arXiv:1808.01410, 2018.

[18] Rivera D E , Morari M , Skogestad S . Internal model control: PID controller design. Industrial & Engineering Chemistry Process Design & Development, 1986, 25(1):2163-2163.

[19] J. M. Valin , J. Skoglund, "LPCNet: Improving Neural Speech Synthesis Through Linear Prediction". in Proceedings ICASSP 2019- IEEE International Conference on Acoustics, Speech, and Signal Processing,2019.

[20] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," Speech Communication, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[21] X. Glorot, Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 249–256, 2010.

[22] B.C.J. Moore, An introduction to the psychology of hearing, Brill, fifth edition, 2012.

[23] D. Kingma, J. Ba , " Adam: A method for stochastic optimization," in Proceedings ICLR 2015-International Conference on Learning Representations, 2015.

[24] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of adam and beyond," in Proceedings ICLR 2018-International Conference on Learning Representations, 2018.