FORWARD ATTENTION IN SEQUENCE-TO-SEQUENCE ACOUSTIC MODELING FOR SPEECH SYNTHESIS

Jing-Xuan Zhang, Zhen-Hua Ling, Li-Rong Dai

National Engineering Laboratory for Speech and Language Information Processing
University of Science and Technology of China, Hefei, P.R.China
nosisi@mail.ustc.edu.cn, {zhling,lrdai}@ustc.edu.cn

ABSTRACT

This paper proposes a forward attention method for the sequence-to-sequence acoustic modeling of speech synthesis. This method is motivated by the nature of the monotonic alignment from phone sequences to acoustic sequences. Only the alignment paths that satisfy the monotonic condition are taken into consideration at each decoder timestep. The modified attention probabilities at each timestep are computed recursively using a forward algorithm. A transition agent for forward attention is further proposed, which helps the attention mechanism to make decisions whether to move forward or stay at each decoder timestep. Experimental results show that the proposed forward attention method achieves faster convergence speed and higher stability than the baseline attention method. Besides, the method of forward attention with transition agent can also help improve the naturalness of synthetic speech and control the speed of synthetic speech effectively.

Index Terms— sequence-to-sequence model, encoder-decoder, attention, speech synthesis

1. INTRODUCTION

A statistical parametric speech synthesis (SPSS) [1–3] system typically consists of a text analysis frontend, an acoustic model, a duration model and a vocoder for waveform reconstruction. The task of the acoustic model is to convert linguistic input into acoustic output. In conventional neural-network-based acoustic modeling [4–8], we usually align a linguistic feature sequence and the corresponding acoustic trajectory by a hidden Markov model (HMM) at first due to the different lengths of these two feature sequences. Then, a deep neural network (DNN) or long short-term memory (LSTM)-based [9] acoustic model can be built using the aligned frame-level input-output pairs. Besides, a separate duration model is always necessary to predict the duration of HMM states or phones at synthesis time.

On the other hand, sequence-to-sequence (seq2seq) neural networks [10, 11] have been proposed recently, which can transduce an input sequence directly into an output sequence that may have different length. Encoder-decoder with attention is the most popular architecture to achieve seq2seq modeling at current stage. It has been successfully applied to various tasks, such as machine translation [12, 13], image caption generation [14] and speech recognition [15–17].

This work was partially funded by the Fundamental Research Funds for the Central Universities (Grant No. WK2350000001), the National Nature Science Foundation of China (Grant No. U1636201), National Key R&D Program of China (Grant No. 2017YFB1002202) and the Key Science and Technology Project of Anhui Province (Grant No. 17030901005).

The seq2seq modeling techniques have also been applied to speech synthesis in the last two years [18–20]. To our knowledge, the first work among them [18] adopted content-based attention [12] to build the encoder-decoder acoustic model for speech synthesis. The windowing technique and convolutional features [15] were also used to stabilize the attention alignment. Char2Wav [19] employed location-based attention [21]. Tacotron [20] improved the network architecture of encoder and decoder, and adopted a reduction trick to help the attention moving forward without getting stuck. There are several advantages of these seq2seq models for speech synthesis. First, we can train acoustic models from scratch data conveniently, which helps to build end-to-end systems without explicit text analysis modules. Second, the separate duration model is not necessary any more. To predict acoustic features with appropriate durations from a unified model may lead to better naturalness of synthetic speech.

Speech synthesis can be considered as a *decompressing* process, i.e., one input phone should be translated into tens of acoustic frames. Therefore, it is a challenge for the attention mechanism to keep focus on one phone for many decoder timesteps and go forward step by step. Current seq2seq models for speech synthesis still suffer from the issue of instability, such as missing phones and repeating phones in the synthetic speech or even failing to generate intelligible speech. Besides, without a separate duration model, it is difficult to control the speed of synthetic speech using seq2seq acoustic models.

Therefore, this paper proposes a forward attention method for the seq2seq acoustic modeling of speech synthesis. This method is motivated by the nature of the monotonic alignment from phone sequences to acoustic sequences. Only the alignment paths that satisfy the monotonic condition are taken into consideration at each decoder timestep. The modified attention probability at each timestep can be computed recursively using a forward algorithm. Furthermore, a transition agent for forward attention is proposed, which helps the attention mechanism to make decisions whether to move forward or stay at each decoder timestep.

Overall, the contributions of this paper are two-fold. First, we propose a new forward attention method, which achieves faster convergence speed, better stability of acoustic feature generation, and higher naturalness of synthetic speech than baseline attention method. Second, we can control the speed of synthesized speech based on the proposed forward attention method, which is difficult for the original content-based attention method.

2. PREVIOUS WORK

A model of encoder-decoder with attention [12, 13] converts an input sequence into an output target sequence with different length.

Encoders and decoders are usually recurrent neural networks (RNN). The encoder first processes the input sequence $\boldsymbol{t} = [\boldsymbol{t}_1, \boldsymbol{t}_2, ..., \boldsymbol{t}_N]$ to produce a sequence of hidden representations $\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N]$ which are more suitable for the attention mechanism to work with. The decoder then generates the output sequences $\boldsymbol{o} = [\boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_T]$, conditioning on $\boldsymbol{x}$.

At each decoder timestep t, the attention mechanism uses an internal inference step to perform a soft-selection over these representations [22]. Let $\boldsymbol{q}_t$ denote the query of the output sequence at the t-th timestep which is usually the hidden state of the decoder RNN, and $\pi_t \in \{1, 2, ..., N\}$ be a categorical latent variable that represents the selection among hidden representatioins according to the conditional distribution $p(\pi_t|\boldsymbol{x}, \boldsymbol{q}_t)$. The context vector derived from the input is defined as

$$\boldsymbol{c}_t = \sum_{n=1}^{N} y_t(n)\boldsymbol{x}_n, \tag{1}$$

where $y_t(n) = p(\pi_t = n|\boldsymbol{x}, \boldsymbol{q}_t)$. Finally, the output vector $\boldsymbol{o}_t$ can be computed conditioning on the context $\boldsymbol{c}_t$. In the widely-used content-based attention mechanism [12], $p(\pi_t|\boldsymbol{x}, \boldsymbol{q}_t)$ is calculated as

$$e_{t,n} = \boldsymbol{v}^T \tanh(\boldsymbol{W}\boldsymbol{q}_t + \boldsymbol{V}\boldsymbol{x}_n + \boldsymbol{b}), \tag{2}$$

$$y_t(n) = \exp(e_{t,n}) \Big/ \sum_{m=1}^{N} \exp(e_{t,m}). \tag{3}$$

Some techniques have been proposed to improve the performance of original attention mechanism. One is *adding convolutional features* [15] for stabilizing the attention alignment. In detail, k filters with kernel size l are employed to convolve the alignment of previous decoder timestep. Let $\boldsymbol{F} \in \mathbb{R}^{k \times l}$ represent the convolution matrix. Then it is used as an extra term for calculating the attention probabilities and we have

$$\boldsymbol{f}_t = \boldsymbol{F} * \boldsymbol{y}_{t-1}, \tag{4}$$

$$e_{t,n} = \boldsymbol{v}^T \tanh(\boldsymbol{W}\boldsymbol{q}_t + \boldsymbol{V}\boldsymbol{x}_n + \boldsymbol{U}\boldsymbol{f}_{t,n} + \boldsymbol{b}), \tag{5}$$

where $*$ denotes convolution and $\boldsymbol{y}_t = [y_t(1), \ldots, y_t(N)]^\top$.

Another technique is *windowing* [15]. Only a subset of the encoding results $\hat{\boldsymbol{x}} = [\boldsymbol{x}_{p-w}, ..., \boldsymbol{x}_{p+w}]$ are considered at each decoder timestep when using the windowing technique. Here, w is the window width and p is the middle position of the window, e.g., the mode of the alignment probability of previous decoder timestep. This technique can not only stabilize the attention alignment but also reduce the computational complexity.

In the application of speech synthesis, the alignment path $\{\pi_1, \pi_2, ..., \pi_T\}$ between $\boldsymbol{x}$ and $\boldsymbol{o}$ indicates how input linguistic features are mapped to their corresponding acoustic features. When phone sequences are used as the input, we expect that the attention should focus on one phone to generate context vectors for tens of acoustic frames, and then move forward to the next phone along a monotonic direction. Therefore, we will propose a new forward attention method for the seq2seq acoustic modeling of speech synthesis in the next section.

3. FORWARD ATTENTION FOR SEQUENCE-TO-SEQUENCE MODELING

3.1. Forward Attention

Assuming π_t at different decoder timesteps are conditionally independent given encoding results $\boldsymbol{x}$ and query $\boldsymbol{q}_t$, we can write the probability of an alignment path $\pi_{1:t} = \{\pi_1, \ldots, \pi_t\}$ as

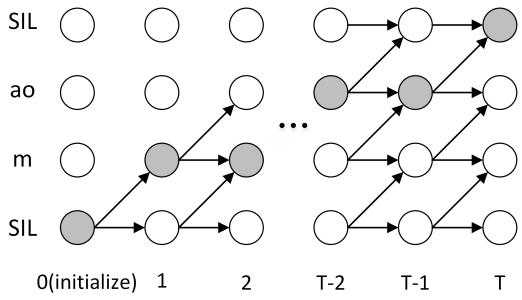

Fig. 1. Grey circles represent a possible alignment path. The alignment paths composed of arrows satisfy $\{\pi_0, \pi_1, ..., \pi_t\} \in \mathcal{P}$.

$$p(\pi_{1:t}|\boldsymbol{x}, \boldsymbol{q}_{1:t}) = \prod_{t'=1}^{t} p(\pi_{t'}|\boldsymbol{x}, \boldsymbol{q}_{t'}) = \prod_{t'=1}^{t} y_{t'}(\pi_{t'}). \tag{6}$$

We introduce a constant $\pi_0 = 1$ for initialization and the probability of the alignment path $\{\pi_0, \pi_1, ..., \pi_t\}$ can also be defined using Equation (6). Let $\mathcal{P}$ denote the space of alignment paths in which each path moves monotonically and continuously without skipping any encoder states. Fig. 1 gives an illustration of the alignment path when decoding acoustic features from an input phone sequence /SIL m ao SIL/ for speech synthesis.

Similar to the connectionist temporal classification (CTC) model [23], a forward variable $\alpha_t(n)$ is defined here to be the total probability of $\{\pi_0, \pi_1, ..., \pi_t\} \in \mathcal{P}$ and $\pi_t = n$ as

$$\alpha_t(n) \overset{def}{=} \sum_{\substack{\pi_{0:t} \in \mathcal{P} \\ \pi_t = n}} \prod_{t'=1}^{t} y_{t'}(\pi_{t'}). \tag{7}$$

Notice that $\alpha_t(n)$ can be calculated recursively from $\alpha_{t-1}(n)$ and $\alpha_{t-1}(n-1)$ as

$$\alpha_t(n) = (\alpha_{t-1}(n) + \alpha_{t-1}(n-1))y_t(n). \tag{8}$$

Then we define

$$\hat{\alpha}_t(n) \overset{def}{=} \alpha_t(n) \Big/ \sum_{n} \alpha_t(n) \tag{9}$$

to make sure the sum of $\hat{\alpha}_t(n)$ for the t-th timestep to be 1 and substitute $\hat{\alpha}_t(n)$ for $y_t(n)$ in Equation (1) to calculate the context vector as

$$\boldsymbol{c}_t = \sum_{n=1}^{N} \hat{\alpha}_t(n)\boldsymbol{x}_n. \tag{10}$$

The complete forward attention method is described in Algorithm 1.

Algorithm 1 Forward Attention

Initialize:
  $\hat{\alpha}_0(1) \leftarrow 1$
  $\hat{\alpha}_0(n) \leftarrow 0, n = 2, ..., N$
for $t = 1$ to T **do**
  $y_t(n) \leftarrow Attend(\boldsymbol{x}, \boldsymbol{q}_t)$
  $\hat{\alpha}'_t(n) \leftarrow (\hat{\alpha}_{t-1}(n) + \hat{\alpha}_{t-1}(n-1))y_t(n)$
  $\hat{\alpha}_t(n) \leftarrow \hat{\alpha}'_t(n) \Big/ \sum_{m=1}^{N} \hat{\alpha}'_t(m)$
  $\boldsymbol{c}_t \leftarrow \sum_{n=1}^{N} \hat{\alpha}_t(n)\boldsymbol{x}_n$
end for

3.2. Forward Attention with Transition Agent

A strategy of transition agent (TA) is further designed to help forward attention control the action of moving forward or staying during alignment flexibly. Specifically, a transition agent DNN with one hidden layer and sigmoid output activation unit is adopted to produce a scalar $u_t \in (0, 1)$ for each decoder timestep. u_t can be considered as an indicator which describes the probability that the attended phone should move forward to the next one at the t-th decoder timestep. $\boldsymbol{c}_t$, $\boldsymbol{o}_{t-1}$ and $\boldsymbol{q}_t$ are concatenated as the input of this DNN. We simply integrate u_t into the calculation of $\alpha_t(n)$ as shown in Algorithm 2.

Algorithm 2 Forward Attention with Transition Agent

Initialize:

$\hat{\alpha}_0(1) \leftarrow 1$

$\hat{\alpha}_0(n) \leftarrow 0, n = 2, ..., N$

$u_0 \leftarrow 0.5$

for $t = 1$ to T **do**

$y_t(n) \leftarrow Attend(\boldsymbol{x}, \boldsymbol{q}_t)$

$\hat{\alpha}'_t(n) \leftarrow ((1 - u_{t-1})\hat{\alpha}_{t-1}(n) + u_{t-1}\hat{\alpha}_{t-1}(n-1))y_t(n)$

$\hat{\alpha}_t(n) \leftarrow \hat{\alpha}'_t(n) \Big/ \sum_{m=1}^{N} \hat{\alpha}'_t(m)$

$\boldsymbol{c}_t \leftarrow \sum_{n=1}^{N} \hat{\alpha}_t(n)\boldsymbol{x}_n$

$u_t \leftarrow DNN(\boldsymbol{c}_t, \boldsymbol{o}_{t-1}, \boldsymbol{q}_t)$

end for

The method of forward attention with transition agent can also be explained from the point of view of a product-of-experts model (PoE) [24, 25]. A PoE model combines a number of individual component models (the experts) by taking their product and normalizing the result. Each component in a product represents a soft constraint. In our proposed forward attention with transition agent, one expert $(1 - u_{t-1})\hat{\alpha}_{t-1}(n) + u_{t-1}\hat{\alpha}_{t-1}(n-1)$ describes the constraint of monotonic alignment. Another expert is the original attention probability given by $y_t(n)$. The calculation of $\hat{\alpha}_t(n)$ is based on the product of these two experts. Therefore, the alignment paths that violate the monotonic condition are expected to have low probability.

Furthermore, the transition agent provides us an opportunity to control the speed of synthetic speech conveniently, which is usually difficult for seq2seq acoustic modeling due to the lack of explicit duration models. When we add positive or negative bias to the sigmoid output units of the transition agent DNN during generation, the transition probability u_t gets increased or decreased. This can lead to a faster or slower movement of the attended phones, corresponding to a faster or slower speed of synthetic speech.

4. EXPERIMENTS

4.1. Experimental Conditions

A Mandarin speech database recorded by a female professional speaker was used in our experiments. The duration of the database was 19.8 hours, which contained 13334 utterances of 16kHz speech data. The database was divided into a training set and a test set, which had 12219 and 1115 utterances respectively. We built seq2seq acoustic models based on the framework of Tacotron [20]. The target acoustic features were log magnitude spectrogram extracted with Hamming windowing, 50 ms frame length, 12.5ms frame shift, and 2048-point Fourier transform. Griffin-Lim algorithm [26] was used to synthesize waveform from the predicted spectrogram. We extracted input features from phone sequences, which were simply composed of the phone label (61-dimension one-hot vector) and tone label (5-dimension one-hot vector) for each phone. These two vectors were first embedded into 224 and 32 dimensional descriptions respectively, and then passed to separate pre-nets. The pre-nets for phone and tone information had the same width as their embedding dimension. The outputs of both pre-nets were concatenated to form the input of the encoder. We employed the reduction trick with $r = 2$ in all experiments.

Table 1. Number of failed samples for the 9 evaluated seq2seq models, where "Window" stands for using the windowing technique, "Conv. Feats." stands for adding convolutional features and "Plain" stands for using none of these two techniques.

Model	Plain	Window	Conv. Feats.
Baseline	54	26	7
FA	5	4	**0**
FA-TA	6	3	**0**

Altogether 9 seq2seq acoustic models were built for comparison.[1] They were divided into 3 groups, which used the conventional attention method introduced in Section 2 (baseline) , the proposed forward attention method (FA), and the forward attention with transition agent (FA-TA) respectively. The 3 systems in each group adopted the windowing technique, the convolutional features, or none of them. For the windowing technique, we set $w = 2$. For using convolutional features, we used $k = 10$ and $l = 5$ in our experiments. We tried to train a system with location-based attention [21]. However, the model failed to converge in our experiments.

We also built a LSTM-based system [5] for comparison. 41-dimension mel-cepstral coefficients (MCCs), and F_0 in log-scale were extracted every 5ms using STRAIGHT [27]. The LSTM acoustic model had 2 hidden layers and 512 units per layer. The model inputs include 523 binary features for categorical linguistic contexts (e.g. phones and tones identities, stress marks) and 9 numerical linguistic contexts (e.g. the number frames and position of current frame in a phone). A separate DNN-based duration model was constructed to predict state durations at synthesis time. The DNN had 3 hidden layers and 1024 units per layer, using 523-dimension binary linguistic contexts as input.

4.2. Stability of Sequence-to-Sequence Feature Generation

We first evaluated the stability of acoustic feature generation using the 9 built seq2seq models with different attention mechanism. 120 utterances were randomly selected from the test set and synthesized using these systems. The longest utterance had about 100 phones. An experienced speech synthesis researcher was asked to listen to all these synthetic samples and label the failed samples, i.e., the synthetic utterances with repeating phones, missing phones, or any kind of perceivable mistakes. The results are summarized in Table 1.

As we can see from this table, the baseline system with plain content-based attention suffered from the mistakes made in the synthetic speech. A close examination showed that this was caused by the inappropriate alignments given by the attention probabilities. Mistakes occurred when the alignment had aliasing, became disconnected, or got stuck at the same position. By introducing the windowing technique or using convolutional features, the performance

[1] Audio samples available on `https://jxzhanggg.github.io/ForwardAttention`

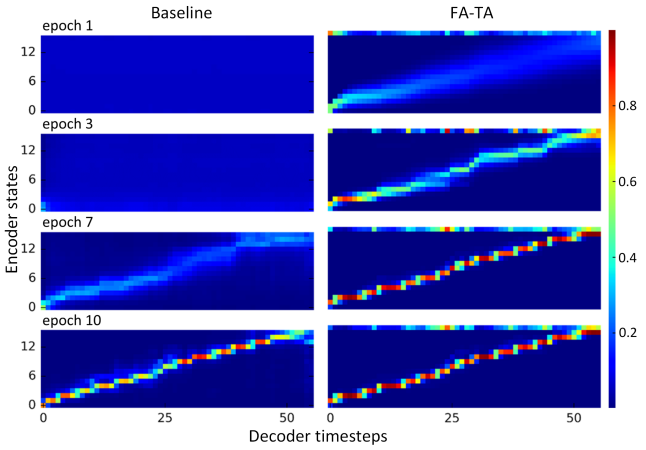

Fig. 2. Alignments of an utterance given by the baseline system and the FA-TA system after 1, 3, 7 and 10 epochs training. The top row of each subgraph in the FA-TA column shows the transition probability u predicted by the transition agent, and the rest rows show $\hat{\alpha}_t(n)$ in Algorithm 2.

Table 2. Average preference scores(%) on naturalness, where "*" stands for using convolution features. "N/P" stands for no preference. p denotes the p-value of a t-test between two systems.

FA	FA-TA	FA-TA*	Baseline	LSTM	N/P	p
22.0	**51.5**	-	-	-	26.5	$< 10^{-5}$
-	**43.0**	19.0	-	-	38.0	$< 10^{-5}$
-	**43.0**	-	13.5	-	43.5	$< 10^{-5}$
-	44.5	-	-	37.5	18.0	0.275

of stability always got improved. The two forward attention methods achieved better stability than the baseline attention method. The best systems adopted forward attention (with or without transition agent) and convolutional features. Moreover, we found that the forward attention systems converged much faster than the baseline systems. Fig. 2 shows how the alignment changed in the plain baseline system and the plain FA-TA system after 1, 3, 7 and 10 epochs of model training.

4.3. Naturalness of Synthetic Speech

Several groups of preferences were conducted to evaluate the naturalness of synthetic speech using different systems. 20 sentences which were correctly synthesized by all systems in the experiment of Section 4.2 were adopted to generate the stimuli. In each preference test, the utterances synthesized by two comparative systems were evaluated in random order by 10 native listeners using headphones. The listeners were asked to judge which utterance in each pair had better naturalness or there was no preference.

We first compared the plain FA system, the plain FA-TA system, and the FA-TA system using convolutional features. The results are shown in the first two rows of Table 2. The results show the advantage of transition agent and the negative effect of adding convolutional features on the naturalness of synthetic speech. One possible reason is that convolutional features acted as a constrain of alignment and impaired the prosodic modeling capacity of attention

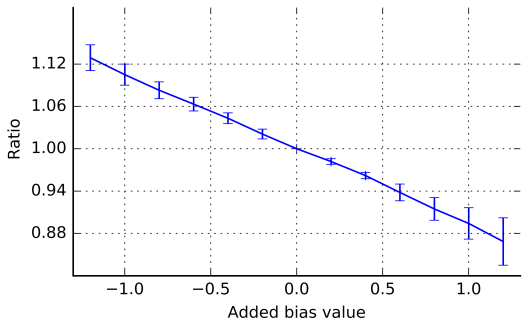

Fig. 3. Average ratios of sentence duration modification achieved by controlling the bias value in the FA-TA system. Error bars represent the standard deviations.

mechanism. Then, we conducted similar experiments to compare the FA-TA system with the plain baseline system and the conventional LSTM system. The results shown in the last two rows of Table 2 demonstrate that the FA-TA system outperformed the baseline and achieved comparable results to the LSTM system. We should notice that the LSTM system employed rich linguistic information as input while the FA-TA system only used phone and tone labels for acoustic modeling.

4.4. Speed Control Using Forward Attention with Transition Agent

In the proposed forward attention with transition agent, as we adding positive or negative bias to the sigmoid output units of the DNN transition agent during generation, the transition probability gets increased or decreased. This leads to a faster or slower of attention results. An experiment was conducted using the plain FA-TA system to evaluate the effectiveness of speed control using this property. We used the same test set of the 20 utterances in Section 4.3. We increased or decreased the bias value from 0 with a step of 0.2, and synthesized all sentences in the test set. We stopped once one of the generated samples had the problem of missing phones, repeating phones, or making any perceivable mistakes. Then we calculated the average ratios between the lengths of synthetic sentences using modified bias and the lengths of synthetic sentences without bias modification. Fig. 3 show the results in a range of bias modification where all samples were generated correctly. From this figure, we can see that more than 10% speed control can be achieved using the proposed forward attention with transition agent. Informal listening test showed that such modification did not degrade the naturalness of synthetic speech.

5. CONCLUSIONS

A forward attention method in the seq2seq acoustic modeling for speech synthesis has been proposed. Experimental results show that this method has the advantages of faster convergence during model training, higher stability of acoustic feature generation, and feasibility of controlling the speed of synthetic speech. This paper applies the proposed forward attention method to the speech synthesis task. This method can also be modified and adapted to other tasks, such as speech recognition and other seq2seq problems having the nature of monotony. Investigation on the performance of forward attention in these tasks will be apart of our future work.

6. REFERENCES

[1] Simon King, "An introduction to statistical parametric speech synthesis," *Sadhana*, vol. 36, no. 5, pp. 837–852, Oct 2011.

[2] Paul Taylor, *Text-to-speech synthesis*, Cambridge university press, 2009.

[3] Heiga Zen, Keiichi Tokuda, and Alan W Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[4] Heiga Zen, Andrew Senior, and Mike Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 7962–7966.

[5] Yuchen Fan, Yao Qian, Feng-Long Xie, and Frank K Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[6] Heiga Zen, "Acoustic modeling in statistical parametric speech synthesis - from HMM to LSTM-RNN," in *Proc. MLSLP*, 2015, Invited paper.

[7] Oliver Watts, Gustav Eje Henter, Thomas Merritt, Zhizheng Wu, and Simon King, "From HMMs to DNNs: where do the improvements come from?," in *Proc. ICASSP*, Shanghai, China, March 2016, IEEE, vol. 41, IEEE.

[8] Z. H. Ling, S. Y. Kang, H. Zen, A. Senior, M. Schuster, X. J. Qian, H. M. Meng, and L. Deng, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 35–52, May 2015.

[9] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[10] Ilya Sutskever, Oriol Vinyals, and Quoc V Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems*, 2014, pp. 3104–3112.

[11] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[12] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[13] Minh-Thang Luong, Hieu Pham, and Christopher D Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[14] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning*, 2015, pp. 2048–2057.

[15] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio, "Attention-based models for speech recognition," in *Advances in Neural Information Processing Systems*, 2015, pp. 577–585.

[16] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 4960–4964.

[17] Navdeep Jaitly, Quoc V Le, Oriol Vinyals, Ilya Sutskever, David Sussillo, and Samy Bengio, "An online sequence-to-sequence model using partial conditioning," in *Advances in Neural Information Processing Systems*, 2016, pp. 5067–5075.

[18] Wenfu Wang, Shuang Xu, and Bo Xu, "First step towards end-to-end parametric TTS synthesis: Generating spectral parameters with neural attention.," in *INTERSPEECH*, 2016, pp. 2243–2247.

[19] Jose Sotelo, Soroush Mehri, Kundan Kumar, Joao Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio, "Char2Wav: End-to-end speech synthesis," in *ICLR2017 workshop submission*, 2017.

[20] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al., "Tacotron: A fully end-to-end text-to-speech synthesis model," *arXiv preprint arXiv:1703.10135*, 2017.

[21] Alex Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[22] Yoon Kim, Carl Denton, Luong Hoang, and Alexander M Rush, "Structured attention networks," *arXiv preprint arXiv:1702.00887*, 2017.

[23] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*. ACM, 2006, pp. 369–376.

[24] G. E. Hinton, "Products of experts," in *1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470)*, 1999, vol. 1, pp. 1–6 vol.1.

[25] Max Welling, "Product of experts," *Scholarpedia*, vol. 2, no. 10, pp. 3879, 2007.

[26] Daniel Griffin and Jae Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[27] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain 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, pp. 187–207, 1999.