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

Non-Parallel Voice Conversion with Autoregressive Conversion Model and Duration Adjustment

Li-Juan Liu[1], *Yan-Nian Chen*[12], *Jing-Xuan Zhang*[2], *Yuan Jiang*[12], *Ya-Jun Hu*[1], *Zhen-Hua Ling*[2], *Li-Rong Dai*[2]

[1]iFLYTEK Research, iFLYTEK Co., Ltd.
[2]National Engineering Laboratory of Speech and Language Information Processing, University of Science and Technology of China, P.R.China

{ljliu, ynchen5}@iflytek.com, {nosisi}@mail.ustc.edu.cn, {yuanjiang, yjhu}@iflytek.com
{zhling,lrdai}@ustc.edu.cn

Abstract

Although N10 system in Voice Conversion Challenge 2018 (VCC 18) has achieved excellent voice conversion results in both speech naturalness and speaker similarity, the system's performance is limited due to some modeling insufficiency. In this paper, we propose to overcome these limitations by introducing three modifications. First, we substitute an autoregressive-based model in order to improve the conversion model capability; second, we use high-fidelity WaveNet to model 24kHz/16bit waveform in order to improve conversion speech naturalness; third, a duration adjustment strategy is proposed to compensate the obvious speech rate difference between source and target speakers. Experimental results show that our proposed method can improve the conversion performance significantly. Furthermore, we validate the performance of this system for cross-lingual voice conversion by applying it directly to the cross-lingual task in Voice Conversion Challenge 2020 (VCC 2020). The released official subjective results show that our system obtains the best performance in conversion speech naturalness and comparable performance to the best system in speaker similarity, which indicate that our proposed method can achieve state-of-the-art cross-lingual voice conversion performance as well.

Index Terms: non-parallel voice conversion, autoregressive model, duration adjustment, cross-lingual voice conversion, Voice Conversion Challenge 2020

1. Introduction

Voice Conversion (VC) aims to convert the speech from a source speaker to that of a target speaker while keeping the linguistic information unchanged [1]. It has many potential applications, such as personalized text-to-speech, voice anonymization and dubbing.

The main task for VC is to learn a mapping function between the acoustic features of source and target speakers. General methods usually need a parallel corpus for training, in which source and target speakers are asked to speak the same utterances. Speeches contain entangled information of both context and speaker timbre. Using such form of training data makes learning process concentrate on timbre difference in regardless of context, thus decreases learning difficulty. These methods include models based on Gaussian mixture model (GMM) [1], frequency warping [2], non-negative matrix factorization (NMF) [3], deep neural networks, such as restricted Boltzman machine (RBM) [4], conditional restricted Boltzman machine (CRBM) [5, 6], long-short term memory (LSTM) based recurrent neural network [7] and so on. However, in most cases, using these approaches is infeasible as parallel data is hard to obtain. In order to make VC applicable in those scenarios, many researchers explore methods using non-parallel training data. These methods include phonetic posteriorgrams (PPGs) based method [8], generative adversarial network (GAN) based method [9], variational auto-encoder (VAE) based method [10] and sequence-to-sequence based method [11].

N10 system achieves the best performance in both speech naturalness and speaker similarity in the non-parallel conversion task of Voice Conversion Challenge 2018 (VCC 18) [12, 13]. It follows a similar approach to the PPGs-based method: the content-related feature is extracted from the source speech using a pre-trained speaker-independent automatic speech recognition (SI-ASR) model and is used to predict the target acoustic feature. Different from using a traditional vocoder, it adopts a WaveNet-based neural vocoder to generate converted waveforms. Although it significantly improves the non-parallel VC performance, there still remains some limitations. First, conversion results are highly correlated with the prediction precision of the acoustic features. This system adopts a simple long-short term memory with projection (LSTMP) based conversion model, so conversion model with higher capability can be further explored to improve the conversion performance. Second, speech rate is a key component of speaker features. In this system, the duration of the converted speech is not modified and is kept as the same as that of the source speech. So when source and target speakers have obviously different speech rates, the conversion similarity will be decreased. Third, although neural vocoder is used to model 10-bit waveform, quantization noise still exists.

This paper aims to further improve the N10 system performance. Our contributions in this work are as follows:

- Autoregressive models show strong acoustic modeling ability [14, 15]. In this paper, we propose to use an autoregressive conversion model in order to obtain acoustic features with higher-precision.
- A duration adjustment strategy is proposed to compensate the speech rate difference between source and target speakers.
- We use high-fidelity WaveNet-based vocoder and model waveforms with high sampling rate in order to improve converted speech quality.
- Previous study [16] proposed to apply a PPGs-based method for cross-lingual VC. In this paper, we evaluate the performance of our improved method for cross-

10.21437/VCC_BC.2020-17

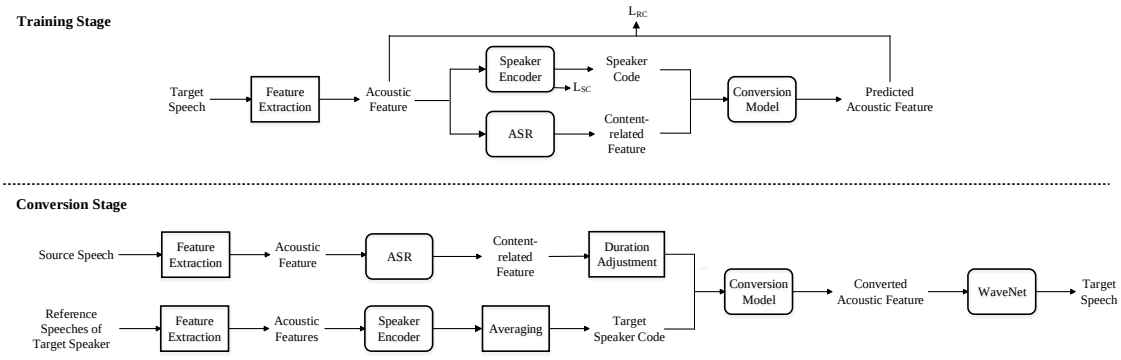

Figure 1: *The framework of the proposed system.*

lingual VC by submitting a developed system for the cross-lingual task in Voice Conversion Challenge 2020 (VCC 2020). The results indicate that this system can achieve state-of-the-art cross-lingual VC performance.

2. Baseline

In N10 system of VCC 18 [13], a pre-trained SI-ASR model was used to extract content-related feature. Then LSTMP-based conversion model was built to predict the joint acoustic feature from the content-related feature. The joint acoustic feature included mel-ceptral coefficients, static logarithmic fundamental frequency and its delta, deltadelta components, unvoiced/voiced (U/V) flag and band aperiodicities (BAPs). Finally, speaker-dependent WaveNet vocoder was built to model 16kHz/10bit waveform, in which a 1024-way categorical distribution [17] was used. Converted waveforms were reconstructed from the predicted acoustic features using the obtained WaveNet vocoder.

3. Proposed Method

3.1. Overall structure

As presented in Figure1, the proposed method mainly contains four components: ASR model, speaker encoder, conversion model and WaveNet vocoder. ASR model is pre-trained using a large multi-speaker English corpus and utilized to extract speaker-independent content-related features. The structure of speaker encoder follows our previous work [11], which is utilized to extract an utterance-level speaker code. The conversion model takes the content-related feature and speaker code as inputs and learns to transform the inputs into acoustic features. Autoregressive models show strong modeling ability for sequence signals and obtain excellent performances in many tasks such as text-to-speech (TTS). Our conversion model is constructed based on an autoregressive architecture and is expected to improve the speech naturalness. WaveNet is adopted as a neural vocoder to recover the waveform from acoustic features.

Model parameters are optimized by a two stages training strategy: pre-training and fine-tuning. This training strategy is quite essential as training data of VC task is limited. During pre-training, speaker encoder and conversion model are jointly trained using multi-speaker corpus. The speaker encoder is optimized by a speaker classification loss L_{SC}, i.e., the cross entropy loss between the predicted speaker probabilities and the target speaker labels as:

$$L_{SC} = CE(I, softmax(WC)) \tag{1}$$

where *I* is a one-hot label representing the speaker identity, *W* is a trainable weight matrix of speaker encoder, *C* is a speaker code for an input utterance, CE(·) represents the cross entropy loss function.

The conversion model is optimized by a reconstruction loss L_{RC} calculated as:

$$L_{RC} = \frac{1}{N}\sum_{i=1}^{N} \| \widehat{a}_i - a_i \|_1 \tag{2}$$

where $\widehat{a}_i$ and a_i are the predicted and target acoustic feature vectors at the i-th frame respectively, *N* is the number of acoustic frames. L_{RC} is only used to train conversion model and doesn't influence speaker encoder.

Fine-tuning process is conducted for each target speaker separately. First, a speaker code vector is generated by averaging the speaker encoder outputs over all training utterances. Then the speaker encoder network is discarded. The obtained speaker code is fed into conversion network and kept fixed during fine-tuning.

In conversion stage, the content-related feature is extracted from the source speech and duration adjustment is performed before it is fed into the conversion model. Then, WaveNet vocoder is used to generate high-fidelity waveform with converted acoustic features as conditions.

3.2. Conversion model

The conversion model maps content-related and speaker-related features to acoustic features. It follows an encoder-decoder architecture, which is similar to the model structure in our previous sequence-to-sequence voice conversion work [18]. Different from that, the input of our encoder module only includes content-related feature. Acoustic feature of source speaker is not used. Moreover, our model only need data of target speaker for training. So parallel training data are not required.

As presented in Figure 2, the encoder is built based on pyramid bidirectional LSTM (PBiLSTM) architecture [19]. Different from conventional deep bidirectional LSTM (BiLSTM) [7], the hidden cells at consecutive two steps of a layer in PBiLSTM are concatenated and fed into next layer to generate a new hidden cell, which results in shorter and higher-level linguistic representations. Since a phoneme usually corresponds to many

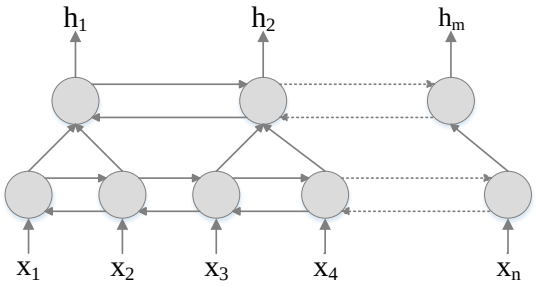

Figure 2: *The structure of the encoder.*

frames of acoustic features, it is more reasonable for the linguistic representations to have a lower sampling rate than the frame-level content-related features.

The decoder is an autoregressive model with attention, which transforms the concatenated encoder output and speaker code into the acoustic feature. At each decoder step, two frames of acoustic features are predicted. It should be noticed that the lengths of encoder and decoder outputs are equal. Although attention was used in our model, we observed no duration conversion was learned. The obtained attention alignment was nearly diagonal and only neighboring frames were focused on. Furthermore, our internal experiment showed that discarding attention module made no difference to conversion performance. We conducted an experiment by conditioning on the encoder hidden features directly. Comparable performance was obtained.

3.3. Duration adjustment

To compensate the speech rate difference between source and target speakers, we perform duration adjustment in conversion stage by interpolating the content-related feature of source speech. The interpolation coefficient for each conversion pair was separately estimated. It was obtained by a two stage estimation. We used the average duration of each phoneme over all training sentences to represent speaker's speech rate. First, a rough coefficient value was estimated by calculating the speech rate ratio between target and source speakers. Then, given the conversion model, we used the obtained interpolation coefficient to do conversion. The coefficient value would be modified if speech rate difference was perceived between converted and target speech. This process repeated until no obvious speech rate difference was perceived. The finally obtained value was used in conversion stage.

3.4. WaveNet-based neural vocoder

Using neural vocoder for waveform generation improves the quality of converted speech greatly [13]. Previous work using WaveNet with categorical distribution introduces quantization noise. To relieve this problem, we adopt the high-fidelity WaveNet model introduced in ClariNet [20] as neural vocoder, which uses a single variance-bounded Gaussian distribution for 24kHz/16bit waveform audio sample modeling. Mel-spectrogram vector is fed as conditional feature. The mean μ_t and variance σ_t of each audio sample distribution are predicted by model conditioned on the samples at all previous time-steps

Table 1: *Mean opinion score (MOS) and 95% confidence interval in t-test of proposed method and the baseline.*

	Naturalness	Similarity
baseline	3.858±0.073	3.488±0.079
proposed	4.016±0.071	3.588±0.080

Table 2: *Mean opinion score (MOS) and 95% confidence interval in t-test of proposed method with and without duration adjustment. proposed+DA denotes the proposed method with duration adjustment.*

	Naturalness	Similarity
proposed	3.719±0.077	3.676±0.080
proposed+DA	3.810±0.074	3.740±0.079

and the current acoustic feature:

$$(\mu_t, \sigma_t) = WaveNet(x_t|x_1, x_2, ..., x_t; \boldsymbol{h}) \quad (3)$$

$$p(x_t|x_1, x_2, ..., x_t; \boldsymbol{h}) = N(\mu_t, \sigma_t) \quad (4)$$

VC system generally has several minutes training data. It is difficult to train a stable WaveNet directly with such limited training data. In order to improve training stability, we follow the same training strategy proposed in [13], in which a multi-speaker model is pre-trained and used as initial model for speaker-dependent WaveNet adaptation. More details can be found in [13].

4. Experimental Conditions

We evaluated the performances of our proposed method on dataset for task1 of VCC 2020, which was intra-lingual VC task. It included 4 English source speakers (SEF1, SEF2, SEM1, SEM2) and 4 English target speakers (TEF1, TEF2, TEM1, TEM2). Each speaker contained 70 utterances. The waveform format was 24kHz/16bit. We randomly chose 64 utterances for training and the remaining 6 utterances for evaluation. For cross-lingual VC, we used the dataset of task2 in VCC 2020. It included 4 same English source speakers as task1 and 6 non-English target speakers, 2 Finish (TFF1, TFM1), 2 German (TGF1, TGM1) and 2 Mandarin speakers (TMF1, TMM1). The amount of data for training and evaluation were kept the same.

80-dimensional mel-spectrograms with 10 ms frame shift was used as acoustic features. Content-related features were extracted from a well-trained SI-ASR model using hundreds of hours of recordings with aligned phonetic transcriptions. Speaker encoder was built with two layers of BiLSTM with 512 hidden units followed by a global average pooling and a fully connected layer with 128 hidden units, which output a 128-dimensional speaker code. The number of PBiLSTM layers in the encoder was set to 2. Each PBiLSTM layer had 512 hidden units. Dropout [21] with probability of 0.2 was used at all BiLSTM layers for regularization.

We performed two stages of training to jointly optimize speaker encoder and conversion model: pre-training and fine-tuning. The pre-training was done on our internal English corpus which contained 130 hours utterances of 13 speakers. The conversion model and speaker encoder were optimized with Adam optimizer [22] with batch size 20. The learning rate was 0.001 for the first 40,000 iterations. After 40,000 iterations, it was exponentially decayed by 0.7 for every 5,000 iterations.

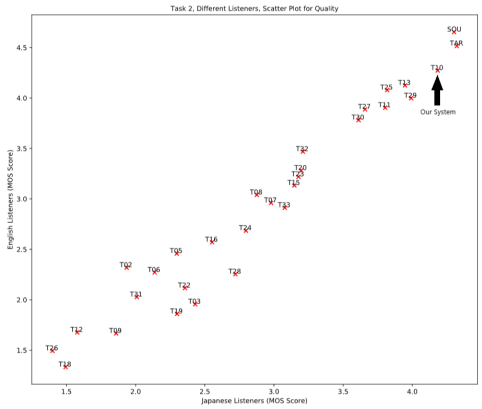

Figure 3: *Scatter plot matching naturalness of Japanese listeners and English listeners for task 2 of VCC 2020.*

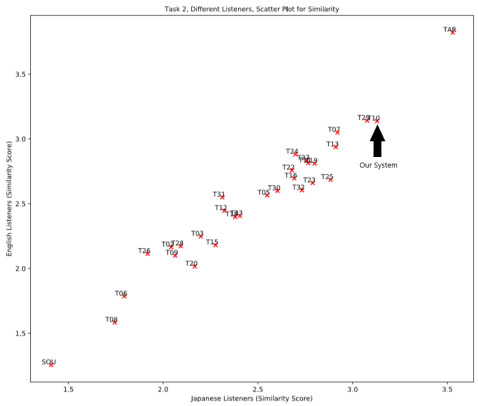

Figure 4: *Scatter plot matching similarity of Japanese listeners and English listeners for task 2 of VCC 2020.*

The fine-tuning process was done on data of each target speaker, with batch size 8. The learning rate was initialized as 0.0005 and halved every 50 iterations. The pre-training was iterated for 70,000 steps and the fine-tuning was iterated for 500 steps.

WaveNet with a structure of 32 dilated convolution layers was adopted to model 24kHz/16bit waveform. All the layers were grouped into 4 dilation cycles, i.e., the dilation rate of layer k (k = 0,1,...,31) was $2^{k \bmod 8}$. The filter width was 3. The lower bound variance was set to -10 (in log scale). For multi-speaker WaveNet training, we used an internal dataset including 45 speakers. The total amount of pre-training data was about 200 hours. All WaveNets were optimized with Adam optimizer.

To evaluate the performance of our proposed method, three systems were constructed using the task1 dataset of VCC 2020:

- baseline: The N10 system in VCC 18.
- proposed: The proposed method in this paper except that duration adjustment was not used.
- proposed+DA: The proposed method in this paper with duration adjustment.

5. Results

5.1. Comparison with baseline

We first compared our proposed method without duration adjustment to the baseline system. We conducted subjective listening tests in terms of both naturalness and similarity. 40 utterances including all conversion pairs (F-F, F-M, M-F and M-M) were randomly selected for evaluation. We used Amazon Mechanical Turk platform and 36 English listeners were involved in our experiments. They were required to use headphone during test and the test samples were presented in random order. Every listener was asked to give a 5-scale opinion score on both naturalness and similarity of each sample. And natural speeches from target speaker were offered for reference. The results of mean opinion score are demonstrated in Table 1. We can observe from the table that our proposed method obtain both higher naturalness and similarity than the baseline.

5.2. Performance of using duration adjustment

To evaluate the effectiveness of duration adjustment, a subjective listening test was conducted to compare the proposed method with and without duration adjustment. 48 English listeners participated in this listening test. The total number of utterances for evaluation was 30. The results in Table 2 demonstrate that by using duration adjustment, both naturalness and similarity are improved.

5.3. VCC 2020 official results for cross-lingual VC task[1]

To evaluate the performance of our proposed method for cross-lingual VC, a developed system with duration adjustment was submitted for the task 2 of VCC 2020. Our system is denoted as T10. The subjective listening tests were conducted on 206 Japanese listeners and 68 English listeners respectively. The naturalness results are presented in Figure 3 . It shows that our proposed method achieves the best naturalness among all the participants for both Japanese and English listeners. As for similarity results presented in Figure 4, our method achieves the best performance for Japanese listeners and comparable performance to the best system for English listeners. It proves that our method can obtain state-of-the-art naturalness and similarity results for cross-lingual VC as well.

6. Conclusion

In this paper, a non-parallel voice conversion method with autoregressive modeling and duration adjustment is proposed to enhance modeling capability for both acoustic model and neural vocoder. We utilize a well-trained ASR model to extract speaker-independent content-related features. An autoregressive conversion model transforms the content-related feature with the speaker code into the acoustic feature. An improved WaveNet vocoder generates the waveform with the acoustic feature as conditions. Besides, a duration adjustment strategy is used to compensate the speech rate difference between source and target speakers. Experimental results show that our proposed method outperforms the best system in VCC 18. Official results of VCC 2020 show that our method achieves state-of-the-art cross-lingual voice conversion performance in both naturalness and similarity.

[1]This approach was also used to construct conversion systems for SEM1-TEM1 and SEM1-TEM2 in task1. For conversion pairs of SEF1-TGM1, SEM1-TGM1 and SEM2-TGM1 in task2, logF0s of source speaker were linearly converted and fed into conversion model since we found naturalness could be further improved for these systems according to our internal experiments.

7. References

[1] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[2] D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 922–931, 2009.

[3] Z. Wu, E. S. Chng, and H. Li, "Exemplar-based voice conversion using joint nonnegative matrix factorization," *Multimedia Tools and Applications*, vol. 74, no. 22, pp. 9943–9958, 2015.

[4] L.-H. Chen, Z.-H. Ling, Y. Song, and L.-R. Dai, "Joint spectral distribution modeling using restricted boltzmann machines for voice conversion." in *Interspeech*, 2013, pp. 3052–3056.

[5] Z. Wu, E. S. Chng, and H. Li, "Conditional restricted boltzmann machine for voice conversion," in *2013 IEEE China Summit and International Conference on Signal and Information Processing*. IEEE, 2013, pp. 104–108.

[6] T. Nakashika, T. Takiguchi, and Y. Ariki, "Voice conversion using speaker-dependent conditional restricted boltzmann machine," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 1–12, 2015.

[7] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 4869–4873.

[8] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *2016 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2016, pp. 1–6.

[9] T. Kaneko and H. Kameoka, "Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 2100–2104.

[10] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2016, pp. 1–6.

[11] J.-X. Zhang, Z.-H. Ling, and L.-R. Dai, "Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 540–552, 2019.

[12] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," *arXiv preprint arXiv:1804.04262*, 2018.

[13] 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.

[14] X. Wang, J. Lorenzo-Trueba, S. Takaki, L. Juvela, and J. Yamagishi, "A comparison of recent waveform generation and acoustic modeling methods for neural-network-based speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4804–4808.

[15] O. Watts, G. E. Henter, J. Fong, and C. Valentini-Botinhao, "Where do the improvements come from in sequence-to-sequence neural tts?" in *2019 ISCA Speech Synthesis Workshop (SSW)*, vol. 10, 2019, pp. 217–222.

[16] L. Sun, H. Wang, S. Kang, K. Li, and H. M. Meng, "Personalized, cross-lingual tts using phonetic posteriorgrams." in *INTERSPEECH*, 2016, pp. 322–326.

[17] 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.

[18] J.-X. Zhang, Z.-H. Ling, L.-J. Liu, Y. Jiang, and L.-R. Dai, "Sequence-to-sequence acoustic modeling for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, pp. 631–644, 2019.

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

[20] W. Ping, K. Peng, and J. Chen, "Clarinet: Parallel wave generation in end-to-end text-to-speech," *arXiv preprint arXiv:1807.07281*, 2018.

[21] N. Srivastava, G.-E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

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