MEL-SPECTROGRAM AUGMENTATION FOR SEQUENCE-TO-SEQUENCE VOICE CONVERSION

Yeongtae Hwang[1], *Hyemin Cho*[1], *Hongsun Yang*[2], *Insoo Oh*[1], *and Seong-Whan Lee*[2]

[1]Netmarble, [2]Korea University

{hwak1234, chme}@netmarble.com, y_hs@korea.ac.kr, ioh@netmarble.com, sw.lee@korea.ac.kr

ABSTRACT

When training the sequence-to-sequence voice conversion model, we need to handle an issue of insufficient data about the number of speech tuples which consist of the same utterance. This study experimentally investigated the effects of Mel-spectrogram augmentation on the sequence-to-sequence voice conversion model. For Mel-spectrogram augmentation, we adopted the policies proposed in SpecAugment [1]. In addition, we propose new policies for more data variations. To find the optimal hyperparameters of augmentation policies for voice conversion, we experimented based on the new metric, namely deformation per deteriorating ratio. We observed the effect of these through experiments based on various sizes of training set and combinations of augmentation policy. In the experimental results, the time axis warping based policies showed better performance than other policies.

1. INTRODUCTION

Recently developed speech-synthesis techniques [2, 3] can produce synthesized speech close to that of the target speaker. The biggest reason for this recent success is that encoder-decoder models with attention mechanisms have been adapted to text-to-speech (TTS) model. Speaker-adaptation has been investigated to leverage a large amount of speech data that accumulates every year and to generate a synthesized voice for a new speaker [4, 5]. These studies showed impressive results in which synthesized voices are generated by adaptation using a few samples.

Voice conversion (VC) is another speech-synthesis technique. The purpose of VC is to switch the speech of a source speaker into that of a target without changing the linguistic content. It acts in a similar manner to the speaker-adaptation technique if it is attached to a TTS system. In the frame-to-frame VC approaches based on acoustic models, i.e,. joint density Gaussian mixture models (JD-GMM) [6, 7], deep neural networks (DNN) [8, 9] and recurrent neural networks (RNN) [10, 11], frame alignment using dynamic time warping algorithms must be used during training. The application of the encoder-decoder models to VC generates natural speech without frame alignment. More recently, a variety of techniques have been proposed to improve sequence-to-sequence(Seq2Seq) VC by adding bottleneck features [12, 13] and text supervision [13, 14].

Thus far, the main problem with VC is the lack of data consisting of speech tuples containing the same utterance. To overcome this situation, data augmentation approaches have been studied based on audio processing [15, 16], text alignment [13], and synthetic data [14]. Other speech-related fields, i.e., automatic speech recognition, SpecAugment [1], vocal track length perturbation (VTLP) [17], and improved vocal track length perturbation (IVTLP) [18] have been proposed based on Mel-spectrogram processing for data augmentation.

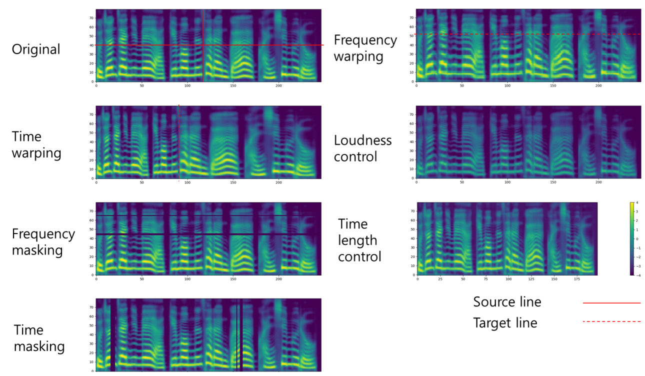

Fig. 1: *Example Mel-spectrograms for each augmentation policy. Left: original Mel-spectrogram and transformed Mel-spectrograms with the policies proposed in SpecAugment, namely time warping, frequency masking, and time masking. Right: transformed Mel-spectrograms with the policies proposed herein, namely frequency warping, loudness control, and time length control.*

Inspired by these, we set the goal of this paper as the determination of the effectiveness of Mel-spectrogram augmentation for the Seq2Seq VC model. Thus, we adopted policies proposed in SpecAugment for VC. We propose new policies for more Mel-spectrogram variants. Choosing hyperparameters for Mel-spectrogram augmentation has a large impact on the Seq2Seq VC model training. To select reasonable hyperparameters for each policy, we experimented based on our proposed metric, namely deformation per deteriorating (DPD) ratio. To evaluate the effectiveness of Mel-spectrogram augmentation, we conducted experiments that is one to one VC task with various sizes of training data and combinations of augmentation policies. In the experimental results, time warping-based policies showed character error rate better than other policies. Among them, our proposed time length control was most effective when it applied to the source and target Mel-spectrogram in the same way. The audio samples of this study are shown on our demo web[1].

2. MEL-SPECTROGRAM AUGMENTATION

We adopted policies proposed in SpecAugment, namely, time masking, frequency masking, and time warping, to deform the time axis, partial loss of time axis, and partial loss of frequency axis. For more variety of Mel-spectrogram variants, we propose the new policies of frequency warping, loudness control, and time-length control to ad-

[1]Audio samples: https://chmenet.github.io/demo/

Table 1: *Definition of the maximum ratio of deformation D_p, where p is the hyperparameter for the augmentation policy.*

p	T, N_t	F, N_f	W	H	L	Λ
D_p	$\frac{T \times N_t}{E(\tau)}$	$\frac{F \times N_f}{\nu}$	W	$\frac{H}{\nu}$	L	Λ

just the pitch, loudness, and speed of speech. The frequency warping of VTLP and IVTLP is similar to one of our frequency warping cases in which the source frequency point is fixed in the middle of the frequency. Thus, our frequency warping allows for greater frequency variation. Note that the aforementioned policies are applicable online during training. Fig. 1 shows how each policy transforms the Mel-spectrogram.

2.1. Augmentation policy

Given a Mel-spectrogram with τ lengths on time axis and ν lengths on frequency axis, the following policies can be used.

Time masking (TM): t consecutive time steps $[t_0, t_0 + t)$ is selected, where t is a discrete random variable $\in [0, T]$, T is the time masking parameter, t_0 is chosen from $[0, \tau - t]$. Selected region is replaced by the minimum value. This process is repeated N_t times.

Frequency masking (FM): f consecutive Mel-frequency channels $[f_0, f_0 + f)$ is selected, where f is a discrete random variable $\in$ [0, F], F is the frequency masking parameter, f_0 is chosen from $[0, \nu - f]$. Selected region is replaced by the minimum value. This process is repeated N_f times.

Time warping (TW): The source point in the time axis is chosen from $[\lfloor\tau/4\rfloor, \tau - \lfloor\tau/4\rfloor]$. It is to be warped by a time distance $w \in [-W\tau, W\tau]$, where W is the time warp parameter. The voice speeds of the two parts based on the target point differ.

Frequency warping (FW): The source point in the frequency axis is chosen from $[\lfloor\nu/4\rfloor, \nu - \lfloor\nu/4\rfloor]$. The source points with all time points are to be warped by a frequency distance $h \in [-H, H]$, where H is the frequency warp parameter. It increases or decreases the level of the pitch.

Time length control (TLC): The source point in the time axis is τ. A line parallel to the frequency axis with the source point warped by a time distance $l \in [-L\tau, L\tau]$, where L is the time length control parameter. It preserves or decreases the speed of the speech.

Loudness control (LC): Subtract the minimum to all Mel-spectrogram values and multiply them by $1 - \lambda$ where $\lambda \in [0, \Lambda]$, Λ is the loudness control parameter. Then add the minimum value to them. It makes the loudness of the speech either down or not.

2.2. Deformation per deteriorating ratio

A good parameter for the Mel-spectrogram augmentation gives maximum variation without losing speech quality. To fit this definition, we propose a new metric, the DPD ratio, which is described by the following equation:

$$\mathrm{DPD}_p = D_p / |E_p - E_o| \quad (1)$$

where D_p is the maximum ratio of deformation for p, E_p is the expectation value of character error rate (CER) for p, P is $\{\{T, N_t\}, \{F, N_f\}, W, H, L, \Lambda\}$ the set of hyperparameter for Mel-spectrogram augmentation, p is an element of P, E_o is the expectation value of CER without augmentation policy. $|E_p - E_o|$ represents deteriorating effects for each hyperparameter. Table 1 shows the definition of D_p for each policy.

2.3. Hyperparameter searching

To determine the best hyperparameter for Mel-spectrogram augmentation that satisfies the aforementioned definition, we conducted the following experiments. The voices for searching the best hyperparameter were 64 audios of the Korean single speaker speech

Table 2: *DPDs on validation set of KSS dataset. The variables not recorded in the table are as follows. $E_o = 0.201, E(\tau) = 217.0$ and $\nu = 80$. The maximum DPD and the selected hyperparameter for each policy are highlighted in bold.*

Time masking ($N_t = 1$)								
T	2	4	6	**8**	10	12	14	16
D_{T,N_t}	0.009	0.018	0.028	0.037	0.046	0.055	0.065	0.074
E_{T,N_t}	0.215	0.217	0.225	0.222	0.232	0.234	0.240	0.248
DDP$_{T,N_t}$	0.643	1.125	1.167	**1.762**	1.484	1.667	1.667	1.574
Frequency masking ($N_f = 1$)								
F	2	4	**6**	8	10	12	14	16
D_{F,N_f}	0.025	0.050	0.075	0.100	0.125	0.150	0.175	0.200
E_{F,N_f}	0.217	0.227	0.235	0.271	0.266	0.302	0.340	0.347
DDP$_{F,N_f}$	1.563	1.923	**2.206**	1.429	1.923	1.485	1.259	1.370
Time warping								
W	0.020	0.040	0.060	**0.080**	0.100	0.120	0.140	0.160
D_W	0.020	0.040	0.060	0.080	0.100	0.120	0.140	0.160
E_W	0.218	0.217	0.220	0.223	0.242	0.256	0.265	0.280
DDP$_W$	1.176	2.500	3.158	**3.636**	2.439	2.182	2.188	2.025
Frequency warping								
H	2	**4**	6	8	10	12	14	16
D_H	0.025	0.050	0.075	0.1	0.125	0.15	0.175	0.2
E_H	0.225	0.237	0.286	0.341	0.400	0.437	0.515	0.545
DDP$_H$	1.042	**1.389**	0.882	0.714	0.628	0.636	0.557	0.581
Time length control								
L	0.020	0.040	0.060	0.080	0.100	**0.120**	0.140	0.160
D_L	0.020	0.040	0.060	0.080	0.100	0.120	0.140	0.160
E_L	0.211	0.210	0.220	0.211	0.216	0.205	0.219	0.213
DDP$_L$	2.000	4.444	3.158	8.000	6.667	**30.00**	7.778	13.333
Loudness control								
λ	0.020	0.040	0.080	**0.160**	0.320	0.640	-	-
D_λ	0.020	0.040	0.080	0.160	0.320	0.640	-	-
E_λ	0.213	0.217	0.218	0.221	0.254	0.406	-	-
DDP$_\lambda$	1.667	2.500	4.706	**8.000**	6.038	3.122	-	-

Table 3: *DPDs on validation set of KSS dataset for masking policies. The maximum DPD and the selected hyperparameter for each policy are highlighted in bold.*

Time masking				
T, N_t	1,8	2,4	**4,2**	8,1
D_{T,N_t}	0.037	0.037	0.037	0.037
E_{T,N_t}	0.216	0.218	0.212	0.222
DPD$_{T,N_t}$	2.467	2.176	**3.364**	1.762
Frequency masking				
F, N_f	1,6	2,3	**3,2**	6,1
D_{F,N_f}	0.075	0.075	0.075	0.075
E_{F,N_f}	0.218	0.213	0.212	0.235
DPD$_{F,N_f}$	4.412	6.250	**6.818**	2.206

(KSS) datasets [19]. These selected audios were converted to Mel-spectrograms. The Mel-spectrogram augmentation for each hyperparameter was performed ten times to compute E_p. Because Korean is sensitive to spacing, CER is more reliable for E_p than word error rate. CER was calculated using the recognition result of the Google Speech API. The audios for computing CER was decoded using Griffin-Lim [20] vocoder from a Mel-spectrogram with or without doing augmentation. Through E_p in Table 2, You can see the degree of deterioration by adjusting p for each policy.

In this experiment, p was increased in the arithmetic sequence except Λ. Because the policy LC only controls audio volume and there is substantial difference in CER performance according to adjust Λ. Thus, Λ was increased to a geometric sequence in this experiment. The hyperparameters determined by choosing the maximum DPD_p are shown in Table 2. In addition, Time masking and frequency masking have two hyperparameters. To determine the optimal combination for these, we first set N_f and N_t to one to find the best D_{T,N_t} and D_{F,N_f}. With fixed D_{T,N_t} and D_{F,N_f} values, we experimented with all possible combinations for T, N_t and F, N_f. Table 3 shows the best T, N_t and F, N_f to maximize DPD_{T,N_t} and DPD_{F,N_f}. The determined hyperparameters are in bold in Tables 2 and 3. In addition, all of them are used in further experiments to evaluate the efficiency of the Mel-spectrogram augmentation.

3. VOICE CONVERSION MODEL

We used a simple model, independent of other models to extract bottleneck features and phoneme labels. Our VC model is based on Tacotron2. The input and output of this model are the Mel-spectrogram. The layers in attention and decoder are persisted. Only the encoder has been modified to FC, FC, and LSTM in a similar manner to the decoder. This is because the decoder effectively represents the Mel-spectrograms. The number of nodes in the layer was determined by referring to the SCENT [12]. The model configurations are shown in Table 4. The final waveform is generated using the Wavenet [21] neural vocoder conditioned on the Mel-spectrogram.

4. EXPERIMENTAL RESULT

4.1. Experimental condition

Two datasets were used in our experiment. For the source speaker, we used the KSS dataset, which consists of 12,853 Korean utterances from a female speaker (approximately 12+ hours). For the target speaker, we used an internal dataset by recording based on the transcript of the KSS dataset from a female speaker. After trimming the silence of both, the pair dataset is constructed with containing 12,798 utterances (approximately 8+ hours for each). We used 64 utterances as the validation set and 64 utterances as the test set; the rest were used as training sets.

All of VC networks were trained for 1×10^5 iterations using the Adam optimizer [22], with a batch size of 32 and a step size of 1×10^{-3}. Wavenet networks were trained for for 16×10^4 iterations using the Adam optimizer, with 8bit mu-law quantization for audio amplitude, a batch size of 16 and a step size of 1×10^{-3}.

4.2. Validation metric

There have been efforts to make the attention alignment diagonal through guides when learning [23]. One study proposes band diagonality [24] to analyze the importance of self-attention. Inspired by previous research, we propose a new metric called attention alignment diagonality (AAD) to avoid overfitting of Seq2Seq VC. AAD is defined as the length of the attention alignment path divided by the length of the diagonal path. The attention alignment path is the line connecting the maximum attention weight for each target vector in the attention weight matrix. This metric represents the degree of learning of the relationship between the encoder and the decoder. Using ADD, we performed early stopping (ES) for all experiments. The point of ES was defined as the point at which the value was minimized within 1×10^5 iterations. We expected our ES to have the effect of selecting the best model for the validation performance.

4.3. Evaluation metric

Researches [12, 13] have been adopted mel-cepstrum distortion (MCD) as an evaluation metric to evaluate the acoustic similarity between the synthesized audio and the target audio. To measure the linguistic expressiveness, one VC study [26] used ARS metrics, such as WER and CER. With reference to the aforementioned studies, MCD and CER were adopted as evaluation metrics to measure the performances of each experiment. In addition, we set the failure to the evaluation metric. Failure is defined as the number of failures of gate prediction on the test set. It indicates the stability of the model. MCD, CER, and failure were reported on the test set in Tables 5, 6, 7, and 8.

Table 4: *Details of model configurations.*

VC	Encoder	FC-ReLU-Dropout(0.5), 256 cells ×2 Forward-LSTM, 256 cells
	PreNet	FC-ReLU-Dropout(0.5), 256 cells ×2
	Decoder	Attention LSTM, 256 cells; Decoder LSTM, 256 cells; Linear project FC, 80 cells; Gate FC, 1 cell and sigmoid activation
	PostNet	1D convolution-BN-ReLU-Dropout(0.5), 256 channels and 5 kernels ×4; 1D convolution-BN-ReLU-Dropout(0.5), 80 channels and 5 kernels
Vocoder	Upsampling	Subpixel [25] convolution, 3×3 kernels and 1×11 strides; Subpixel convolution, 3×3 kernels and 1×25 strides
	WaveNet	20 layers dilated convolution layers, with dilation $d = 2^{k \bmod 10}$ for $k = [0, ..., 19]$, 256 softmax output

FC represents fully connected, LSTM represents long short-term memory, BN represents batch normalization, ReLU represents rectified linear unit.

4.4. Baseline performance

In order to observe the performance change of VC model according to data usage without Mel-spectrogram augmentation, we experimented by reducing the number of training data to half of it each time from the whole training set till it reaches to the 1/16 training set. The metrics obtained in this experiment were used as a criterion for determining the degree to which the augmentation policy has improved performance.

Table 5 shows the results with 100k iterations and ES. The ES results based on the minimum AAD do not guarantee better performance in all respects. The CER performance is directly proportional to the amount of training data. However, MCD and failure are not directly related to the amount of training data. To observe the change in linguistic expressive power according to data usage, we set the minimum CER values for each experiment to the baseline performance for each size of training set.

4.5. Effectiveness of augmentation policy

In this experiment, all augmentation policies were applied to the source Mel-spectrogram. One-to-many mapping data in the training set makes the model difficult to converge. In general, the augmentation is not applied to target data. However, if the speeds of the source audio and target audio are changed to the same ratio, this is one-to-one mapping and means augmenting pair data. Therefore, we experimented with two cases, namely, applying TLC only to source audio, and applying TLC to both source and target. The second case is denoted “TLC both.”

Single policy: To verify the effectiveness of each augmentation policy, we experimented with the 1/16 training set. The results for each policy are shown in Table 6. Policies showing improved CER

Table 5*: Evaluation results using various sizes of training data without Mel-spectrogram augmentation. The minimum CER values for each size of training set are highlighted in bold. We set them to the baseline performance for each size of training set.*

	Size	1	1/2	1/4	1/8	1/16
10^5 iterations	AAD	1.333	1.238	1.236	1.400	1.643
	MCD	6.873	7.123	6.759	6.850	7.367
	CER	0.143	0.159	0.225	**0.323**	**0.479**
	Failure	1	2	0	3	7
ES	AAD	1.177	1.137	1.162	1.235	1.465
	MCD	6.666	7.002	6.709	7.032	7.456
	CER	**0.130**	**0.154**	**0.182**	0.343	0.507
	Failure	1	2	0	2	2

Table 6*: Evaluation results by applying a single augmentation policy on the 1/16 training set. The lower CER values against baseline performance on the 1/16 training set are highlighted in bold.*

	Policy	**TLC**	**TLC both**	TM	**TW**	FM	FW	LC
10^5 iterations	AAD	1.266	1.538	1.865	1.53	1.726	1.691	1.573
	MCD	7.318	7.392	7.512	7.281	7.439	7.572	7.401
	CER	**0.426**	**0.397**	0.575	**0.423**	0.547	0.641	0.489
	Failure	6	7	17	4	6	3	7
ES	AAD	1.219	1.377	1.688	1.378	1.626	1.503	1.518
	MCD	7.279	7.470	7.629	7.355	7.510	7.623	7.400
	CER	**0.448**	**0.450**	0.573	**0.464**	0.549	0.559	0.507
	Failure	10	7	15	8	6	4	8

against baseline performance on the 1/16 training set were TLC, "TLC both," and TW. Policies based on the time axis warping cause differences in the speed of speech. Improved CER performances can be interpreted as policies based on the time axis can yield different distributions to source speech with less loss of the speaker's speech characteristics. Masking policies hinder learning because it gives a loss of information in the source. Frequency axis warping produces a phonetic distribution that differs from the actual speaker, which seems to adversely affect the conversion using the actual speaker's speech. LC only reduces the Mel-spectrogram value. Thus, it shows a similar performance that of the baseline.

Multiple policy: How much did using the combination of policies improve the performance of CER? To determine the answer to this question, we conducted an experiment combining TLC, "TLC both," TW, and LC. We experimented with the 1/16 training set. The results for multiple policies are shown in Table 7. No combinations showed an improved performance of CER. We interpreted a reason for this as follows. Applying multiple policies makes more changes to the original speech. Thus, it seems to have a bad influence on VC learning.

Policy effectiveness: TLC, "TLC both," TW were tested on all sizes of training data in Table 5. The results are shown in Table 8. The CER values on the 1/16 experiment outperform the baseline performance. In the 1/2 and 1/8 experiments, the CER values show some or no improvements. In experiments for all sizes of training set, the lowest CER was mostly in "TLC both." In experiments on the whole training set and 1/4 training set, there is no significant difference between the minimum CER of "TLC both" and the baseline performance. Therefore, we interpreted the experimental results of applying TLC both to Seq2Seq VC could be an option to improve the linguistic expressiveness.

Table 7*: Evaluation results by applying multiple augmentation policies. There are no lower CER values against baseline performance on the 1/16 training set.*

	Policy	TLC TW	TLC both TW	TLC LC	TLC both LC	TLC TW LC	TLC both TW LC
10^5 iterations	AAD	1.935	2.221	1.533	1.848	1.568	1.691
	MCD	7.706	7.626	7.415	7.380	7.390	7.572
	CER	0.723	0.582	0.585	0.562	0.616	0.641
	Failure	17	13	14	5	7	3
ES	AAD	1.765	1.908	1.486	1.609	1.451	1.503
	MCD	7.707	7.501	7.428	7.292	7.355	7.623
	CER	0.764	0.524	0.537	0.544	0.579	0.559
	Failure	23	9	6	8	9	4

Table 8*: Evaluation results by applying each augmentation policy on various data volumes. The lower CER values against baseline performance on each volume dataset are highlighted in bold. The minimum CER values within results on the same size of training set are highlighted in underlines.*

	Size	1	1/2	1/4	1/8	1/16
TLC both 10^5	AAD	1.259	1.251	1.377	1.445	1.538
	MCD	6.641	7.052	6.690	6.740	7.392
	CER	0.134	0.198	0.185	**0.282**	**0.397**
	Failure	5	1	0	3	7
TLC both ES	AAD	1.128	1.214	1.171	1.329	1.377
	MCD	6.722	7.094	6.608	6.950	7.470
	CER	0.144	**0.148**	0.245	**0.286**	**0.450**
	Failure	4	0	0	3	7
TLC 10^5	AAD	1.142	1.149	1.149	1.379	1.266
	MCD	6.829	7.126	6.680	6.997	7.318
	CER	0.167	0.158	0.202	**0.290**	**0.426**
	Failure	9	3	0	3	6
TLC ES	AAD	1.115	1.141	1.141	1.295	1.219
	MCD	6.835	7.005	6.837	7.084	7.279
	CER	0.171	**0.145**	0.226	0.342	**0.448**
	Failure	4	0	2	5	10
TW 10^5	AAD	1.271	1.234	1.377	1.692	1.530
	MCD	6.852	6.935	6.692	6.820	7.281
	CER	0.158	**0.143**	0.236	**0.308**	**0.423**
	Failure	2	0	0	1	4
TW ES	AAD	1.147	1.161	1.271	1.474	1.378
	MCD	6.657	7.042	6.710	6.900	7.355
	CER	0.159	**0.137**	0.218	0.338	**0.464**
	Failure	1	3	0	2	8

5. CONCLUSION

This paper describes the effect of Mel-spectrogram augmentation on the one-to-one Seq2Seq VC model. We adopted policies from SpecAugment and proposed new policies for Mel-spectrogram augmentation. We selected appropriate hyperparameters for each policy through experiments based on our proposed DPD metric. The experimental results showed that the relationship between the size of the training data and the linguistic expressiveness of the VC model is directly proportional. In addition, the policies based on the time axis warping showed lower CER than other policies. These results indicate that the use of policies based on the time axis warping is more efficiently training for developing the VC model with the insufficiency size of training set.

6. REFERENCES

[1] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[2] 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: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[3] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al., "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[4] Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al., "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Advances in neural information processing systems*, 2018, pp. 4480–4490.

[5] Yutian Chen, Yannis Assael, Brendan Shillingford, David Budden, Scott Reed, Heiga Zen, Quan Wang, Luis C Cobo, Andrew Trask, Ben Laurie, et al., "Sample efficient adaptive text-to-speech," *arXiv preprint arXiv:1809.10460*, 2018.

[6] Alexander Kain and Michael W Macon, "Spectral voice conversion for text-to-speech synthesis," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1998, vol. 1, pp. 285–288.

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

[8] Srinivas Desai, Alan W Black, B Yegnanarayana, and Kishore Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 954–964, 2010.

[9] Ling-Hui Chen, Zhen-Hua Ling, Li-Juan Liu, and Li-Rong Dai, "Voice conversion using deep neural networks with layer-wise generative training," *ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 1859–1872, 2014.

[10] Lifa Sun, Shiyin Kang, Kun Li, and Helen Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4869–4873.

[11] Jiahao Lai, Bo Chen, Tian Tan, Sibo Tong, and Kai Yu, "Phone-aware lstm-rnn for voice conversion," in *International Conference on Signal Processing (ICSP)*. IEEE, 2016, pp. 177–182.

[12] Jing-Xuan Zhang, Zhen-Hua Ling, Li-Juan Liu, Yuan Jiang, and Li-Rong Dai, "Sequence-to-sequence acoustic modeling for voice conversion," *ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 27, no. 3, pp. 631–644, 2019.

[13] Jing-Xuan Zhang, Zhen-Hua Ling, Yuan Jiang, Li-Juan Liu, Chen Liang, and Li-Rong Dai, "Improving sequence-to-sequence voice conversion by adding text-supervision," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6785–6789.

[14] Fadi Biadsy, Ron J Weiss, Pedro J Moreno, Dimitri Kanvesky, and Ye Jia, "Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation," *arXiv preprint arXiv:1904.04169*, 2019.

[15] Riku Arakawa, Shinnosuke Takamichi, and Hiroshi Saruwatari, "Implementation of dnn-based real-time voice conversion and its improvements by audio data augmentation and mask-shaped device," in *The 10th ISCA Speech Synthesis Workshop (to appear)*, 2019.

[16] Eliya Nachmani and Lior Wolf, "Unsupervised singing voice conversion," *arXiv preprint arXiv:1904.06590*, 2019.

[17] Navdeep Jaitly and Geoffrey E Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, 2013, vol. 117.

[18] Chanwoo Kim, Minkyu Shin, Abhinav Garg, and Dhananjaya Gowda, "Improved vocal tract length perturbation for a state-of-the-art end-to-end speech recognition system," *Proc. Interspeech 2019*, pp. 739–743, 2019.

[19] Kyubyong Park, "Kss dataset: Korean single speaker speech dataset," https://kaggle.com/bryanpark/korean-single-speaker-speech-dataset, 2018.

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

[21] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[22] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[23] Kou Tanaka, Hirokazu Kameoka, Takuhiro Kaneko, and Nobukatsu Hojo, "Atts2s-vc: Sequence-to-sequence voice conversion with attention and context preservation mechanisms," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6805–6809.

[24] Gaël Letarte, Frédérik Paradis, Philippe Giguère, and François Laviolette, "Importance of self-attention for sentiment analysis," in *EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2018, pp. 267–275.

[25] Wenzhe Shi, Jose Caballero, Lucas Theis, Ferenc Huszar, Andrew Aitken, Christian Ledig, and Zehan Wang, "Is the deconvolution layer the same as a convolutional layer?," *arXiv preprint arXiv:1609.07009*, 2016.

[26] Gokce Keskin, Tyler Lee, Cory Stephenson, and Oguz H Elibol, "Measuring the effectiveness of voice conversion on speaker identification and automatic speech recognition systems," *arXiv preprint arXiv:1905.12531*, 2019.