UNSUPERVISED PRE-TRAINING FOR DATA-EFFICIENT TEXT-TO-SPEECH ON LOW RESOURCE LANGUAGES

Seongyeon Park[1*], *Myungseo Song*[1*], *Bohyung Kim*[1] *and Tae-Hyun Oh*[2,3]

[1]CNAI, Seoul, Korea
[2]Dept. of EE and GSAI, POSTECH, Pohang, Korea
[3]Institute for Convergence Research and Education in Advanced Technology,
Yonsei University, Seoul, Korea

ABSTRACT

Neural text-to-speech (TTS) models can synthesize natural human speech when trained on large amounts of transcribed speech. However, collecting such large-scale transcribed data is expensive. This paper proposes an unsupervised pre-training method for a sequence-to-sequence TTS model by leveraging large untranscribed speech data. With our pre-training, we can remarkably reduce the amount of paired transcribed data required to train the model for the target downstream TTS task. The main idea is to pre-train the model to reconstruct de-warped mel-spectrograms from warped ones, which may allow the model to learn proper temporal assignment relation between input and output sequences. In addition, we propose a data augmentation method that further improves the data efficiency in fine-tuning. We empirically demonstrate the effectiveness of our proposed method in low-resource language scenarios, achieving outstanding performance compared to competing methods. The code and audio samples are available at: https://github.com/cnaigithub/SpeechDewarping

Index Terms— Text-to-speech, data-efficiency, pre-training, unsupervised learning, data augmentation

1. INTRODUCTION

Recent advance in deep neural networks enables us to build end-to-end text-to-speech (TTS) models [1, 2] to synthesize plausible speech. Recent research [3, 4] attributes natural and plausible speech generation of TTS models to the following capabilities to be learned: 1) *attention alignment* between the input and output sequences, and 2) *autoregressive prediction* of acoustic features. The supervised learning or pre-training methods [5, 6] directly inject the necessary capabilities for TTS through supervision using large-scale transcribed speech. However, such models require a large amount of transcribed speech data for training, which is not annotation efficient. Constructing such large-scale text-annotated speech is time-consuming, costly, and even infeasible for low-resource languages.

To mitigate the labeled data deficiency, pre-training methods for TTS systems have been investigated [3–6]. Among them, Chung et al.and Zhang et al. [3, 4] specifically designed to induce either of such capabilities in unsupervised ways by leveraging large-scale untranscribed speech data. In [3], the decoder of Tacotron [1] is pre-trained as an autoregressive speech generator. In [4], the whole model of Tacotron 2 [2] is pre-trained to predict speech from unsupervised linguistic units extracted by an external Vector-quantization

Variational-Autoencoder (VQ-VAE) [7]. It would be desirable to pre-train the full TTS model without any external model.

The goal of this paper is to further reduce the amount of transcribed speech required for TTS training. To this end, we propose an unsupervised pre-training method for Tacotron 2, *Speech De-warping*. By utilizing large-scale *untranscribed* speech, our key idea is to make the TTS model learn to reconstruct original spectrograms from warped ones, *i.e.*, learn to *de-warp*. This method does not require annotation, as we synthesize the warped spectrograms by a simple random temporal warping technique. We sample random segment boundaries and resize each segment along the temporal axis to be a fixed size. Learning to de-warp as a pre-training step encourages the model to acquire both preliminary knowledge of attention alignment and autoregressive prediction. After the pre-training, we fine-tune the model using small-scale transcribed speech data of a target speaker, possibly in a low-resource language. In addition, we extend our simple random warping technique to a data augmentation method for the fine-tuning step, which further improves performance.

Compared to the previous studies, our pre-training method does not suffer from the model mismatch problem between pre-training and fine-tuning [3] and does not require training an external model for data preparation [4]. It is also worth noting that our data augmentation does not require any external data or pre-trained models unlike other data augmentation approaches for TTS [8–13]; they typically leverage a large amount of transcribed speech to generate synthetic data with pre-trained TTS models or voice conversion models.

Our main contributions are summarized as follows: 1) proposing an unsupervised pre-training method for TTS models, *Speech De-warping*, 2) proposing a simple yet effective data augmentation method, *SegAug*, 3) demonstrating improved data efficiency, and 4) showing the cross-language effectiveness of our methods.

2. PROPOSED METHOD

2.1. Segment-based Speech Warping

Our pre-training and data augmentation methods include the procedure of warping speech. To warp the speech, we segment the mel-spectrograms of the speech along the time axis and apply a transformation per segment. To clarify the procedure, we describe the general form of the segment-based speech warping f.

Given a mel-spectrogram $\mathbf{m}$ of timesteps N, we warp $\mathbf{m}$ to generate a warped mel-spectrogram $\hat{\mathbf{m}}$, which is given by

$$\hat{\mathbf{m}} = f(\mathbf{m}; S, T), \tag{1}$$

where S is a segmentation method, and T is a transformation. The
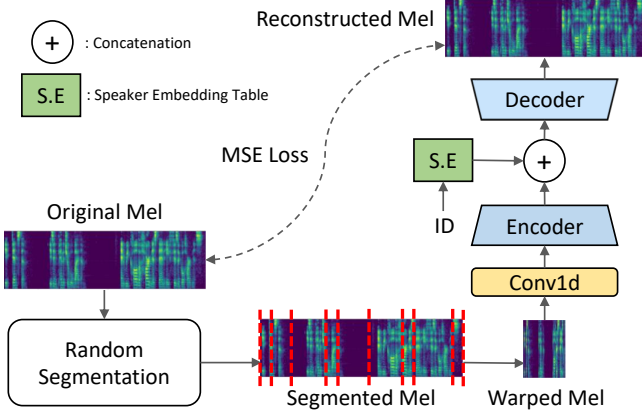

Fig. 1. Overview of our *Speech De-warping*. We randomly segment the spectrogram and warp it by resizing each segment to have equal unit timesteps. The model is learned to reconstruct the original spectrogram from the warped one.

segmentation method S segments $\mathbf{m}$ into k different spectrogram segments $\mathbf{m}_1, \mathbf{m}_2, ..., \mathbf{m}_k$ such that $N = \sum_{i=1}^{k} N_i$, where N_i is the number of timesteps of $\mathbf{m}_i$. Then, for each segment $\mathbf{m}_i$, the transformation T transforms $\mathbf{m}_i$ to a warped segment $\hat{\mathbf{m}}_i = T(\mathbf{m}_i)$. We concatenate the warped segments along the time axis to generate the warped spectrogram $\hat{\mathbf{m}} = concat(\hat{\mathbf{m}}_1, \hat{\mathbf{m}}_2, ..., \hat{\mathbf{m}}_k)$.

Theoretically, any segmentation method and transformation can be used as S and T, *e.g.*, phoneme segmentation for S. We present our specific configuration for S and T in the following subsections.

2.2. Pre-training: Unsupervised Speech De-warping

We aim to reduce the amount of transcribed speech required for TTS training. To this end, we propose an unsupervised pre-training method, *Speech De-warping* which leverages large-scale *untranscribed* speech, which is much cheaper to obtain than transcribed one. The main idea is to pre-train a TTS model to recover original spectrograms from warped ones, which is illustrated in Figure 1.

To generate pairs of input and expected output for unsupervised learning, we first generate warped spectrograms from the original spectrograms converted from the untranscribed speech by the segment-based speech warping (see Equation 1). Specifically, we use random segmentation as S, which randomly selects $k-1$ number of boundary timesteps; thus, k segments are obtained. The segment boundaries are independently sampled for each training step. We set $k = \lfloor \frac{N}{6} \rfloor$ for each spectrogram. For the transformation T, we use linear interpolation to make each segment have an equal unit timestep (*i.e.*, length 1).

We adopt Tacotron 2 [2] as our backbone TTS model and denote it as Tacotron for simplicity. Tacotron has the input format of text embedding; thus, the spectrogram inputs are not directly applicable. To feed the warped spectrograms to the model's encoder as input, we replace the text embedding look-up table of Tacotron with a simple 1D convolutional layer. It maps the mel-dimension to the embedding dimension of the Tacotron encoder during unsupervised training.

Other segmentation methods can be used as S to generate segments instead of our proposed random segmentation, *e.g.*, more semantically aligned segments like exact phonemes. For example, one can adopt the Montreal Forced Alignment (MFA) [14] tool to extract phoneme segments using text annotations. However, it is not applicable in our unsupervised pre-training setting, where no text annotation is available. Instead, one can use unsupervised pseudo phoneme segmentation [15]. We empirically show that these semantic segmentation methods can improve the performance of *Speech De-warping*, but the simple random segmentation is powerful enough to outperform other baselines.

2.3. Fine-tuning: Transferring Knowledge to TTS

After pre-training Tacotron with the pretext task, we fine-tune the model with the downstream TTS task of a target speaker. We use a few transcribed speeches, *i.e.*, text-audio pairs, of the target speaker to fine-tune the model. Before starting fine-tuning, to feed texts to the model as in the original Tacotron, the 1D convolutional layer preceding the encoder is discarded, and a learnable text embedding look-up table for the target speaker's language is randomly initialized. With this reconfiguration, we fine-tune the TTS model for the small target speaker data.

Data Augmentation. To further improve data efficiency during fine-tuning, we propose a simple data augmentation method called *SegAug*. During fine-tuning, we augment the training data by applying the segment-based speech warping (Equation 1) to the target spectrograms. Specifically, for S, we use random segmentation as in the pre-training stage. For the transformation T, we use linear interpolation to resize each segment of the input spectrogram along the time axis by a factor uniformly sampled from $[\frac{1}{3}, \frac{5}{3}]$. The resulting warped spectrograms are used as the target spectrograms for training loss. After training the model with this augmentation, we additionally train the model for a few steps without the augmentation to adapt the model to the ground truth prosody of the target speaker, *i.e.*, a cool-down step. Note that this augmentation in the fine-tuning stage is optional. While our pre-training alone empirically demonstrates favorable performance, we can further improve the performance with this augmentation during fine-tuning.

3. EXPERIMENTS

3.1. Experiment Setup

Dataset and Evaluation. We use the *train-clean-100* subset of the LibriTTS [16] dataset as the untranscribed pre-training set, which consists of 47.6 hours of speech from 247 English speakers. We set Korean as a low-resource language and use the Korean Single speaker Speech (KSS) [17] dataset as our transcribed fine-tuning set. Following [3,4], we define 24 minutes of speech as 1 shard of data. Then, we construct fine-tuning datasets by randomly sampling 0.5, 1, 2, 3, 5, 8 shards from the KSS dataset. For evaluation, we conduct both objective and subjective tests. For the objective evaluation, we use Mel-cepstral Distortion with Dynamic time-warping (MCD-DTW) [18], simply denoted as MCD. The objective results are reported as an average over the test set containing 571 utterances (about 22.7 minutes in total). For the subjective evaluation, we conduct AB preference tests on 20 utterances randomly sampled from the test set. We ask 15 native Korean raters to choose the more preferred one among two synthesized audios given the text, in terms of pronunciation, recognizability, and naturalness.

Implementation details. For pre-training, we use the Adam [19] optimizer with a learning rate of 10^{-3}. The models are pre-trained for 100K steps with batch size 16. For fine-tuning, we gradually decrease the learning rate from 10^{-3} to 10^{-4} for 50K training steps with batch size 32. Audio waveforms are down-sampled to 16 kHz, and

Table 1. MCD results of several pre-training and data augmentation methods when being fine-tuned on 0.5 or 1 shard (12 or 24 minutes) of paired speech of the target speaker. Note that T-Pho leverages text annotations in pre-training.

Augmentation in fine-tuning	Model	Supervised pre-training	Paired data (in shards)	
			0.5	1
No aug.	Tac	×	11.98	12.41
	T-Dec	×	12.07	12.18
	T-VQ	×	11.11	10.41
	T-SD (Ours)	×	**10.79**	**10.40**
	T-Pho	○	10.40	10.28
With aug.	Tac + Gaussian		12.59	12.33
	Tac + Mixup		12.06	12.04
	Tac + SpecAug		12.29	10.60
	Tac + SegAug		12.19	10.68
	T-VQ + Gaussian		10.63	10.40
	T-VQ + Mixup		11.12	10.48
	T-VQ + SpecAug		10.46	10.33
	T-VQ + SegAug		10.41	10.27
	T-SD + SegAug (Ours)		**10.28**	**10.24**

Griffin-Lim [20] algorithm is used as a vocoder for fast experiment cycles.

Compared methods. We use Tacotron 2 [2] as the TTS model in our experiments. Following the naming conventions of Zhang et al. [4] with T(acotron), we denote the model only trained with the fine-tuning data without pre-training by Tac. We denote two recent unsupervised pre-training methods, decoder pre-training [3] and VQ-VAE-based pre-training [4], by T-Dec and T-VQ, respectively. The model pre-trained with our *Speech De-warping* is denoted by T-SD, *i.e.*, ours. As an upper bound of performance for the unsupervised pre-training methods, we employ the model pre-trained in a supervised manner with text annotations and denote it by T-Pho, as suggested by Zhang et al. [4]. In addition to the pre-training methods, we compare our data augmentation with other data augmentation methods in the fine-tuning stage. We denote additive Gaussian noise-based augmentation [21] by Gaussian, mixup-based augmentation [22] by Mixup, SpecAugment [23] by SpecAug.

3.2. Results on Small Amount of Fine-tuning Data

Objective Evaluation. Table 1 presents the superior performance of the proposed methods compared to competing methods on small amounts of fine-tuning data. Without data augmentation during fine-tuning, T-SD outperforms all unsupervised pre-training methods and Tac. Both T-Dec [3] and Tac, which do not have the opportunity to pre-learn a sufficient capability of attention alignment in pre-training, show similarly lower performance than the others. In contrast, the proposed de-warping task encourages the model to learn both preliminary knowledge of attention alignment and autoregressive prediction. When data augmentation is applied during fine-tuning, T-SD with *SegAug* outperforms other combinations of pre-training and augmentation methods. *SegAug* even effectively improves the performance of other pre-training baselines and shows competitive performance compared to other augmentation methods.

Subjective Evaluation. Table 2 shows the preference test results with competitive methods using 0.5 shards of fine-tuning data. Consistent with the objective results, our methods outperform the prior art of unsupervised methods (T-VQ). Interestingly, with our proposed data augmentation applied during fine-tuning, our whole transfer learn-

Table 2. AB test results of our method over competitive baselines. All methods use 0.5 shards (12 minutes) of fine-tuning data.

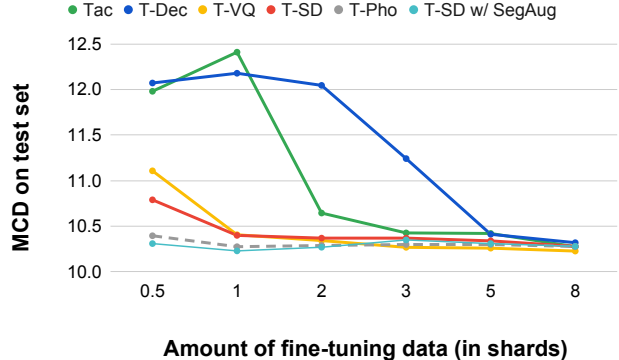| Model pair	Preference (%)		
	Former	Latter	Neutral
T-VQ vs. T-SD	15.7	**54.7**	29.6
T-VQ vs. T-SD + SegAug	3.7	**76.0**	20.3
T-Pho vs. T-SD + SegAug	23.3	**44.0**	32.7

Fig. 2. MCD results according to varying amounts of paired data. The dashed line denotes the MCD of T-Pho, which is a supervised method; thus, it can be considered a near-upper-bound performance of unsupervised pre-training methods.

ing scheme even outperforms the supervised pre-training baseline (T-Pho).

3.3. Effects of Different Amounts of Fine-tuning Data

Figure 2 presents the MCD evaluation results of the competing methods according to different amounts of fine-tuning data. Our method shows the best performance overall and is particularly better on small amounts of data. As the amount of fine-tuning data increases, the models tend to show better performance, and the performance gaps between the methods gradually decrease.

3.4. Additional Results

Comparison to an upsampling pre-text task. In our *Speech De-warping*, we resize segments of different lengths into the same timestep 1 to warp the input spectrograms. As a result, the alignment between the warped spectrogram and the original one becomes non-linear, which is analogous to the alignment characteristics between text and speech. We argue that *learning this monotonic yet non-linear alignment in pre-training is one of the critical factors* of our method. To validate this argument, we introduce a control experiment with simple upsampling pre-training as a pre-text task, called *Naive*, and compare it with our *Speech De-warping* in Table 3. Specifically, in *Naive*, instead of using the segment-wise warping to warp the spectrogram, we downsample the whole spectrogram by a single scale factor of $\frac{1}{6}$ using linear interpolation along the time axis. Thereby, the model with *Naive* learns a linear alignment between the uniformly downsampled spectrograms and the original spectrograms. Figure 3
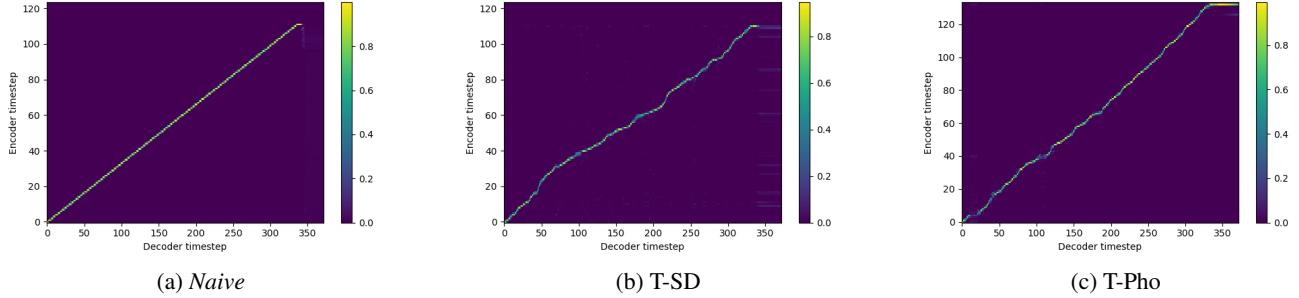
(a) *Naive*   (b) T-SD   (c) T-Pho

Fig. 3. Examples of the learned attention alignments between input and output timesteps of the decoders of the models during pre-training. While *Naive* induces the model to learn a linear alignment, our T-SD encourages the model to learn a non-linear alignment whose form is similar to the alignment between text and speech as in T-Pho.

Table 3. MCD results for the ablation study comparing our de-warping to the up-sampling pre-training. *Naive* indicates the model pre-trained with the up-sampling task.

Method	Paired data (in shards)	
	0.5	1
Naive	11.37	10.88
T-SD	**10.79**	**10.40**

Table 4. MCD results of T-VQ and our *Speech De-warping* according to different segmentation methods on two fine-tuning languages. Different and Same denote that the fine-tuning language is different or the same as the pre-training language. Note that phoneme segmentation requires text supervision.

Method	Paired data (in shards)			
	Different		Same	
	0.5	1	0.5	1
T-VQ	11.11	10.41	11.85	10.51
T-SD (Random segment)	10.79	10.40	11.57	10.63
Pseudo phoneme segment	10.56	10.38	11.71	10.69
Phoneme segment	11.35	10.48	11.19	10.46

presents examples of attention alignments learned during pre-training. The superior performance of T-SD compared to *Naive* in Table 3 verifies that learning a *monotonic and non-linear alignment* benefits our *Speech De-warping*. Note that *Naive* performs better than Tac and T-Dec in Table 1, which demonstrates the effectiveness of learning *monotonic alignment* through the upsampling task itself.

Effect of Heterogeneous Languages in Fine-tuning. We investigate the effect of using different or the same languages between pre-training and fine-tuning steps, which is the main scope of this work. As described, we use English as pre-training data. For the same language scenario, called *Same*, we use the LJspeech [24] dataset (English) for fine-tuning data. The different language scenario follows the same setup described in Sec. 3.1, called *Different*. We compare our T-SD with T-VQ to show the algorithmic behavioral differences. Table 4 shows the performance of T-SD is overall similar to T-VQ when the language between pre-training and fine-tuning is unchanged, *i.e.*, *Same*. However, as shown in the *Different* columns, T-SD is more robust against overfitting to the pre-training language than T-VQ.

We conjecture this is because the burden to memorize acoustic features of the pre-training language is less for our method since some language-specific acoustic information is already given as input for de-warping.

Effect of Segmentation Methods. We investigate the effect of different segmentation methods for the segment-based speech warping in *Speech De-warping*. In addition to the random segmentation used in our T-SD, we compare with the phoneme segmentation by using the MFA tool [14], which requires text supervision, and the pseudo phoneme segmentation by using the unsupervised phoneme segmentation model [15]. As shown in Table 4, the performance of *Speech De-warping* can be boosted by using semantically meaningful segmentation obtained from external models. The phoneme segmentation shows the best performance when the fine-tuning language is the same as the pre-training language and the worst when the fine-tuning language is unseen during pre-training. The phoneme segmentation of a specific language induces the alignment of the warped spectrograms and original spectrograms to be very similar to the alignment between text and speech of that language in pre-training. This behavior can lead to overfitting to the specific language used in pre-training.

4. CONCLUSION

We propose an unsupervised pre-training method and a data augmentation method for training TTS models with limited amounts of text-annotated speech data. Our pre-training method enables us to build a TTS system for a low-resource language by leveraging a large-scale and untranscribed speech dataset that can be easily collected. The proposed data augmentation technique can be used to further improve such data efficiency. Our comprehensive experiments show the superior performance of the proposed methods compared to various competing pre-training and data augmentation methods. We empirically demonstrate that learning a non-linear alignment during pre-training of the model is beneficial in TTS compared to learning a linear alignment. We show that our pre-training method can achieve better performance by using external models for segmentation.

5. REFERENCES

[1] 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," *Conference of the International Speech Communication Association (INTERSPEECH)*, 2017.

[2] 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 *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018.

[3] Yu-An Chung, Yuxuan Wang, Wei-Ning Hsu, Yu Zhang, and RJ Skerry-Ryan, "Semi-supervised training for improving data efficiency in end-to-end speech synthesis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019.

[4] Haitong Zhang and Yue Lin, "Unsupervised learning for sequence-to-sequence text-to-speech for low-resource languages," *Conference of the International Speech Communication Association (INTERSPEECH)*, 2020.

[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," in *International Conference on Learning Representations (ICLR)*, 2018.

[6] Henry B Moss, Vatsal Aggarwal, Nishant Prateek, Javier González, and Roberto Barra-Chicote, "Boffin TTS: Few-shot speaker adaptation by bayesian optimization," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

[7] Jan Chorowski, Ron J Weiss, Samy Bengio, and Aäron Van Den Oord, "Unsupervised speech representation learning using wavenet autoencoders," *IEEE/ACM transactions on audio, speech, and language processing*, 2019.

[8] Min-Jae Hwang, Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim, "TTS-by-TTS: TTS-driven data augmentation for fast and high-quality speech synthesis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2021.

[9] "TTS-by-TTS 2: Data-selective augmentation for neural speech synthesis using ranking support vector machine with variational autoencoder," *Conference of the International Speech Communication Association (INTERSPEECH)*, 2022.

[10] Suhyeon Oh, Ohsung Kwon, Min-Jae Hwang, Jae-Min Kim, and Eunwoo Song, "Effective data augmentation methods for neural text-to-speech systems," in *International Conference on Electronics, Information, and Communication (ICEIC)*, 2022.

[11] "Low-data? no problem: low-resource, language-agnostic conversational text-to-speech via f0-conditioned data augmentation," *Conference of the International Speech Communication Association (INTERSPEECH)*, 2022.

[12] "Cross-speaker emotion transfer for low-resource text-to-speech using non-parallel voice conversion with pitch-shift data augmentation," *Conference of the International Speech Communication Association (INTERSPEECH)*, 2022.

[13] Goeric Huybrechts, Thomas Merritt, Giulia Comini, Bartek Perz, Raahil Shah, and Jaime Lorenzo-Trueba, "Low-resource expressive text-to-speech using data augmentation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2021.

[14] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi.," in *Conference of the International Speech Communication Association (INTERSPEECH)*, 2017.

[15] Felix Kreuk, Joseph Keshet, and Yossi Adi, "Self-supervised contrastive learning for unsupervised phoneme segmentation," *Conference of the International Speech Communication Association (INTERSPEECH)*, 2020.

[16] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu, "LibriTTS: A corpus derived from librispeech for text-to-speech," *Conference of the International Speech Communication Association (INTERSPEECH)*, 2019.

[17] K Park, "Kss dataset: Korean single speaker speech dataset," https://www.kaggle.com/datasets/bryanpark/korean-single-speaker-speech-dataset, 2018.

[18] Robert Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE pacific rim conference on communications computers and signal processing*, 1993, vol. 1.

[19] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[20] 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.

[21] Birger Moëll, Jim O'Regan, Shivam Mehta, Ambika Kirkland, Harm Lameris, Joakim Gustafsson, and Jonas Beskow, "Speech data augmentation for improving phoneme transcriptions of aphasic speech using wav2vec 2.0 for the psst challenge," in *13th Language Resources and Evaluation Conference (LREC)*, 2022.

[22] Shuai Guo, Jiatong Shi, Tao Qian, Shinji Watanabe, and Qin Jin, "SingAug: Data Augmentation for Singing Voice Synthesis with Cycle-consistent Training Strategy," in *Conference of the International Speech Communication Association (INTERSPEECH)*, 2022.

[23] 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," *Conference of the International Speech Communication Association (INTERSPEECH)*, 2019.

[24] Keith Ito and Linda Johnson, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset, 2017.