Speech De-warping: Unsupervised Pre-training for Data-Efficient Text-to-Speech on Low Resource Languages

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

Abstract

Neural text-to-speech (TTS) models can synthesize natural human speech when being trained on large amounts of transcribed speech. However, collecting such large-scale transcribed data is expensive. In this paper, we propose an unsupervised pre-training method for reducing the amount of paired data required to train a sequence-to-sequence TTS model, utilizing large untranscribed speech data. The main idea is to pre-train the model to reconstruct de-warped mel-spectrograms from warped ones. For semantically meaningful warping/de-warping, we train a self-supervised phoneme segmentation model and use the segments to warp the spectrograms in a pseudo phoneme level. In addition, as a byproduct of our pre-training process, we can optionally leverage the segment-based data augmentation in fine-tuning stage to further improve the data-efficiency. We empirically demonstrate the effectiveness of our method in a low-resource language scenario, achieving outstanding performance compared to various baselines.

1. Introduction

Parallel with the great success of pre-training neural networks on large-scale datasets in computer vision and natural language processing (Girshick et al., 2014; Devlin et al., 2019), a variety of pre-training techniques have introduced impact in both audio and speech applications. For example, Kunze et al. (2017) and Joshi et al. (2020) train an automated speech recognizer on a large-scale dataset of a transcribed language and transfer the knowledge to another language in a supervised way. Furthermore, it has been shown that models can learn meaningful representations from unlabeled examples via semi-supervised or self-supervised pre-training for various applications, such as emotion recognition (Lian et al., 2019), speech recognition (Schneider et al., 2019; Baevski et al., 2020), speaker verification (Fan et al., 2021; Chen et al., 2022), and language identification (Fan et al., 2021).

In particular, we focus on the text-to-speech (TTS) application, which requires a large amount of transcribed speech data to produce plausible human-like speech by neural TTS models (Wang et al., 2017; Shen et al., 2018). Constructing such large-scale text-annotated speech is time-consuming and costly, and even infeasible for low-resource languages. To mitigate such labeled data deficiency, some works (Ren et al., 2019; Liu et al., 2020; Tu et al., 2020) proposed a semi-supervised framework, where the duality of automatic speech recognition (ASR) and TTS is explicitly leveraged.

Recently, pre-training methods for TTS systems have been started to be investigated (Chen et al., 2018; Moss et al., 2020; Chung et al., 2019; Zhang & Lin, 2020). The capabilities needed to be prepared for TTS models are hinted in (Chung et al., 2019; Zhang & Lin, 2020); a sequence-to-sequence TTS model typically attempts to learn 1) *attention alignment* between the input and output sequences, and 2) *autoregressive prediction* of acoustic features. Thus, those pre-training methods are specifically designed to induce either of such capabilities.

Clearly, the supervised pre-training methods (Chen et al., 2018; Moss et al., 2020) directly inject necessary capabilities for TTS through supervision. It is not annotation efficient. Different from the supervised pre-training, Chung et al. (2019) present an unsupervised pre-training method that pre-trains the decoder of Tacotron (Wang et al., 2017) as an autoregressive speech generator. Improving this work, Zhang & Lin (2020) pre-train Tacotron 2 (Shen et al., 2018) to predict speech from unsupervised linguistic units extracted by an external Vector-quantization Variational-Autoencoder (VQ-VAE) (Chorowski et al., 2019).

The goal of this paper is to 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*. Our key idea is to train a TTS model to recover original spectrograms from warped ones, *i.e.*, *de-warp* them. This task encourages the model to learn both prelinimary

[*] Equal contribution [1] CNAI, Seoul, Korea [2] Dept. of EE, POSTECH, Korea. Correspondence to: T.-H. Oh <taehyun@postech.ac.kr>.

ICML Workshop on Machine Learning for Audio Synthesis, Baltimore, Maryland, USA, 2022.

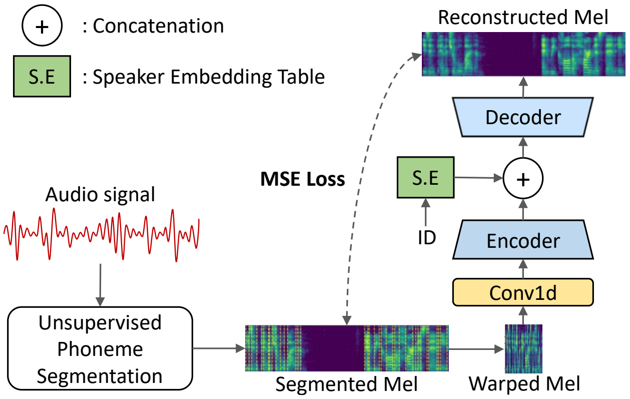

Figure 1. An overview of our self-supervised pre-training method, *Speech De-warping*. First, we train an unsupervised phoneme segmentor (Kreuk et al., 2020). Using it, we obtain the segmented mel-spectrogram and resize each segment to the same step size. Then, we pre-train Tacotron 2 (Shen et al., 2018) to predict the original spectrogram from the warped spectrogram, which is followed by fine-tuning on the TTS task with few transcribed speech.

knowledge of attention alignment and autoregressive prediction. We train a segmentor to extract phoneme-like segments in a self-supervised way, and thereby we warp spectrograms by resizing each phonetic segment to a fixed size. Finally, we fine-tune the model using few paired speech of a target speaker, possibly in a low resource language. Compared to the previous studies, e.g., Chung et al. (2019), our method does not suffer from the model mismatch problem between pre-training and fine-tuning since both the encoder and decoder of the TTS model are pre-trained. In addition, the unsupervised phoneme segmentor used for our pre-training can optionally be leveraged for a data-augmentation technique during fine-tuning to further improve performance.

Our main contributions are summarized as follows: 1) defining speech de-warping by phoneme-like segments as a self-supervised task, 2) demonstrating improved data efficiency, and 3) showing cross-language effectiveness of our method.

2. Proposed Method

We first describe the pre-training and then the fine-tuning procedure of our method. We adopt Tacotron 2 (Shen et al., 2018) as our baseline model and denote it by Tacotron for simplicity. We illustrate the overall process of the proposed method in Figure 1.

2.1. Pre-training: Unsupervised Speech De-warping

Given a series of segmentation of each audio in the pre-training dataset, we extract mel-spectrogram segments of them. Using 80 channel mel-spectrograms, a speech signal is represented as an array of shape $(80, N)$, where N is the number of timesteps. The extracted segments are represented as arrays of shape $(80, N_1), (80, N_2), ..., (80, N_s)$, where s is the number of segments of a single speech signal and $N = \sum_{i=1}^{s} N_i$.

For these segments, we construct a warped mel-spectrogram by resizing the segments to the fixed number of timesteps H. That is, each segment $(80, N_i)$ becomes an array of shape $(80, H)$. Then, the segments are concatenated by the time axis. As a result, the original mel-sepctogram of shape $(80, N)$ is warped to have shape $(80, sH)$. We use linear interpolation for resizing and set $H = 1$.

We pre-train Tacotron to reconstruct the original mel spectrograms from those corresponding warped mel-spectrogram inputs. We call this task *Speech De-warping*. In order to feed the warped spectrograms to the model, we replace the text embedding look-up-table of Tacotron to a simple 1D convolutional layer which maps the mel dimension to the embedding dimension of the Tacotron encoder. Following (Zhang & Lin, 2020), we concatenate speaker embeddings obtained from a speaker embedding look-up-table with the outputs of the encoder on each timestep.[1]

Arbitrary warp-and-dewarp does not match the target TTS task, where each phoneme of the input sequence is spread across time to form the corresponding pronunciation. Thus, it would be desired to induce such a relationship between warped inputs and dewarped outputs. To this end, we learn to extract pseudo phoneme segments exploited for the warp-and-dewarp as follows.

Unsupervised Pseudo Phoneme Segmentation. To pre-train Tacotron with semantically meaningful de-warping, we propose to use a phoneme segmentation model as the audio segmentor in pre-training. We leverage an unsupervised phoneme segmentation model (Kreuk et al., 2020) which can be trained from untranscribed data. Next, we describe the method of (Kreuk et al., 2020) in detail.

A raw audio signal can be represented as a sequence of scalars, $\mathbf{x} = (x_1, x_2, ..., x_T)$, where $x_i \in \mathbb{R}$. A convolutional encoder maps the raw audio sequence $\mathbf{x}$ to a sequence of representation vectors $\mathbf{z} = (z_1, z_2, ..., z_L)$, where $z_i \in \mathbb{R}^N$. Then, the encoder learns to minimize the following loss:

$$\mathcal{L} = \sum\nolimits_{\mathbf{x} \in S} \sum\nolimits_{z_i \in f(\mathbf{x})} \hat{\mathcal{L}}(z_i, D_K(z_i)), \tag{1}$$

where a training set $S = \{\mathbf{x}_m | m \in \{1, 2, ..., M\}\}$, and the contrastive loss $\hat{\mathcal{L}}$ for each frame z_i is given by:

$$\hat{\mathcal{L}}(z_i, D_K(z_i)) = -\log \frac{\exp[sim(z_i, z_{i+1})]}{\sum_{z_j \in \{z_{i+1}\} \cup D_K(z_i)} \exp[sim(z_i, z_j)]}. \tag{2}$$

$D_K(z_i)$ denotes a set of K vectors randomly sampled from

[1] While this step is omitted in the main paper of Zhang & Lin (2020), the authors confirm that their approach relies on the speaker embedding and we follow the exact same way.

$D(z_i) = \{z_j : |i-j| > 1\}$, i.e., non-adjacent vectors of z_i, and $sim(u,v) = \frac{u^\top v}{\|u\|\|v\|}$.

To perform pseudo phoneme segmentation with the trained model, an audio signal $\mathbf{x}$ is encoded into a latent vector sequence $\mathbf{z}$. For $\mathbf{z}$, the score of boundary at the i-th step is calculated as $score(i) = -sim(z_i, z_{i+1})$. A higher score means a higher possibility that a phoneme-like boundary is at that index. Using a peak detection algorithm over the scores, the phoneme boundaries are obtained.

2.2. Fine-tuning: Transferring Knowledge to TTS

After pre-training Tacotron by the *Speech De-warping* task, we fine-tune the model for the TTS task with a target speaker. We use few transcribed speech data of the target speaker to train the model. To use the text embedding layer as in the original Tacotron, a learnable text embedding look-up-table for the target speaker's language is randomly initialized.

Additionally, we can optionally leverage the segmentation model for data augmentation in the fine-tuning stage to further improve the data-efficiency. Different from *Speech De-warping* in the pre-training stage, where each mel-spectrogram segment is warped to the same step size, we warp the segments randomly to augment the speech data during fine-tuning. We resize each segment by a factor randomly sampled from $[r, 2-r]$, where we set $r = \frac{1}{3}$ in our experiments. After training the model with this augmentation, we further train the model for a few steps without the augmentation to adapt the model to ground truth prosody of the target speaker, i.e., cool-down. Note that this additional technique in the fine-tuning stage is optional. While our pre-training 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 'train-clean-100' subset of the LibriTTS (Zen et al., 2019) dataset as the untranscribed pre-training set, which consists of 47.6 hours of speech from 247 English speakers. We hypothetically set Korean as a low resource language and select Korean Single speaker Speech (KSS) (Park, 2018) dataset as the transcribed fine-tuning set. Following (Chung et al., 2019; Zhang & Lin, 2020), we define 24 minutes of speech as 1 shard of data. We construct fine-tuning datasets by randomly sampling 0.5, 1, 2, 3, 5, 8, 16 shards of KSS dataset.

For evaluation, we use both objective and subjective tests. For the objective evaluation, we use Mel-cepstral Distortion with Dynamic time-warping (MCD-DTW) (Kubichek, 1993), 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 10 raters to choose the more preferred one among two synthesized audios given the text, in terms of pronunciation, delivery, and naturalness. They are allowed to choose neither.

Table 1. MCD evaluation results of several models when fine-tuned on 0.5 shards (12 minutes) of paired speech of the target speaker. Note that T-Pho leverages text annotations in pre-training.

Tac	T-Dec	T-VQ	Ours	T-Pho
11.98	12.07	11.11	**10.56**	10.40

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
Ours vs. T-VQ	**84.0**	1.5	14.5
Ours vs. T-Pho	10.5	**71.5**	18.0
Ours vs. Ours + Aug	25.5	**59.0**	15.5

Implementation details. For pre-training, we adopt Adam (Kingma & Ba, 2015) optimizer with learning rate 10^{-3}. The models are trained for 100K steps with batch size 16. For fine-tuning, we gradually decrease learning rate from 10^{-3} to 10^{-4} for 50K training steps with batch size 32. The audio is down-sampled to 16000 Hz and Griffin-Lim (Griffin & Lim, 1984) algorithm is used for fast experiment cycles.

Baselines. We use Tacotron 2 (Shen et al., 2018) as a TTS model in our experiments. We denote the model trained from scratch by Tac, the model with decoder pre-training (Chung et al., 2019) by T-Dec, the model with VQ-VAE-based pre-training by T-VQ (Zhang & Lin, 2020), the model pre-trained in a supervised manner by T-Pho, and the model with our method by Ours. T-Pho is employed as an upper bound of performance for the unsupervised pre-training methods, following (Zhang & Lin, 2020).

3.2. Results on Small Amount of Fine-tuning Data

Objective Evaluation. Table 1 presents the MCD results of various methods when fine-tuned with 0.5 shards of paired data. Our method achieves the best MCD value compared to competing baselines. Unlike Ours and T-VQ (Zhang & Lin, 2020), T-Dec (Chung et al., 2019) shows similar performance to Tac. It may be because T-Dec could not learn the appropriate attention alignment in pre-training as only the decoder of Tacotron was pre-trained. Also, different from the original setting of (Chung et al., 2019), there was a language mismatch between the pre-training and fine-tuning in our experiments.

Subjective Evaluation. Table 2 shows the results of the

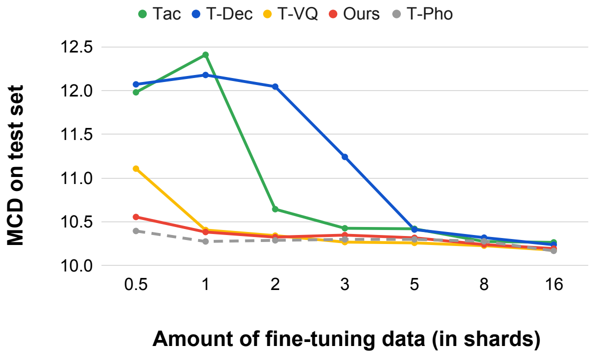

Figure 2. MCD results for several amounts of paired data. The dashed line of T-Pho indicates that it roughly plays a role of an upper bound of the performance for the unsupervised pre-training.

Table 3. MCD results of additional experiments including ablation study for 0.5 shards (12 minutes) of paired speech data. *Naive* indicates the model pre-trained with the simple up-sampling task.

NAIVE	OURS	OURS + AUG	T-PHO	T-PHO + AUG
11.37	10.56	**10.31**	10.40	**10.23**

Table 4. MCD results of several models when fine-tuned on the same language as the pre-training language, i.e., English.

MODEL	FINE-TUNING DATA (IN SHARDS)	
	0.5	1
TAC	13.79	13.79
T-VQ	11.85	**10.51**
OURS	**11.71**	10.69
T-PHO	10.61	10.21

preference test for the competitive methods using 0.5 shards of paired speech for training. Consistent with the objective results, our method outperforms T-VQ. Meanwhile, from informal listening tests, we found that Tac and T-Dec failed to produce intelligible speech.

3.3. Results on Other Amounts of Fine-tuning Data

Figure 2 presents the MCD evaluation results of the various methods on several amounts of fine-tuning data. Our method shows the best performance overall, and especially better on small amounts of data. It can be said that models tend to show better performance when using more fine-tuning data. Also, the performance of each method gets similar as the amount of fine-tuning data increases. The effect of pre-training decreases when more fine-tuning data is used.

3.4. Additional Results

Unsupervised Mel-segment Augmentation. The effectiveness of the unsupervised mel-segment augmentation technique in fine-tuning stage, *Aug*, is shown in Table 2 and Table 3. The proposed augmentation further improves the performance of both Ours and T-Pho, showing applicability to other baselines.

Ablation Study. In the proposed speech de-warping task, we resize phonetic segments of different lengths into the same timesteps. As a result, the alignment between the warped spectrogram and the original spectrogram becomes non-linear, like the alignment between text and speech. We argue that learning this monotonic yet non-linear alignment in pre-training is one of the key factors of our method. To demonstrate it, we pre-train Tacotron with a simple up-sampling task, where it learns a linear alignment between the uniformly down-sampled spectrogram and original spectrogram. Refer to the supplementary material for details. We name the simple up-sampling pre-training scheme *Naive* and report the MCD performance in Table 3. When compared to the methods from Table 1, *Naive* shows better performance than Tac and T-Dec that do not learn a monotonic alignment between encoder and decoder timesteps during pre-training. However, it shows worse performance than other methods (T-VQ, T-Pho and Ours) that learn a non-linear alignment as well as the monotonic alignment during pre-training. This results imply that learning a monotonic and non-linear alignment benefits our method.

Fine-tuning to Seen Language. Table 4 presents the MCD results from various methods when the language of the fine-tuning set is same as the language of the pre-training set, *i.e.*, English. We use the LJspeech (Ito & Johnson, 2017) dataset as the fine-tuning set. When the language is unchanged, the performance of Ours is overall similar to that of T-VQ for the small numbers of data. However, our method is more robust against overfitting to the pre-training language than T-VQ as shown in Table 1. It may be because the burden to memorize the acoustic features of the language in pre-training stage is less for our method, since some language-specific information is already given as inputs.

4. Conclusion

We proposed a self-supervised pre-training method for training a TTS model with few amounts of text-annotated speech data. Our method enables us to build a TTS system for a low resource language by leveraging a large-scale and untranscribed speech dataset which can be easily collected. We show that a data augmentation technique by virtue of the byproduct of our pre-training can be used to further improve such data efficiency. Our comprehensive experiments show the superior performance of the proposed method compared to other baselines. We empirically demonstrate that learning a non-linear alignment during pre-training of the model is beneficial compared to learning a linear alignment.

Acknowledgment. T.-H. Oh was partially supported by IITP grant funded by the Korea government (MSIT) (No.2021-0-02068, Artificial Intelligence Innovation Hub).

References

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Chen, Y., Assael, Y., Shillingford, B., Budden, D., Reed, S., Zen, H., Wang, Q., Cobo, L. C., Trask, A., Laurie, B., et al. Sample efficient adaptive text-to-speech. In *International Conference on Learning Representations (ICLR)*, 2018.

Chen, Z., Chen, S., Wu, Y., Qian, Y., Wang, C., Liu, S., Qian, Y., and Zeng, M. Large-scale self-supervised speech representation learning for automatic speaker verification. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2022.

Chorowski, J., Weiss, R. J., Bengio, S., and Van Den Oord, A. Unsupervised speech representation learning using wavenet autoencoders. *IEEE/ACM transactions on audio, speech, and language processing*, 2019.

Chung, Y.-A., Wang, Y., Hsu, W.-N., Zhang, Y., and Skerry-Ryan, R. 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.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019.

Fan, Z., Li, M., Zhou, S., and Xu, B. Exploring wav2vec 2.0 on Speaker Verification and Language Identification. In *Conference of the International Speech Communication Association (INTERSPEECH)*, 2021.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Griffin, D. and Lim, J. Signal estimation from modified short-time fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, 32(2):236–243, 1984.

Ito, K. and Johnson, L. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/, 2017.

Joshi, V., Zhao, R., Mehta, R. R., Kumar, K., and Li, J. Transfer learning approaches for streaming end-to-end speech recognition system. *Conference of the International Speech Communication Association (INTERSPEECH)*, 2020.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Kreuk, F., Keshet, J., and Adi, Y. Self-supervised contrastive learning for unsupervised phoneme segmentation. *Conference of the International Speech Communication Association (INTERSPEECH)*, 2020.

Kubichek, R. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE pacific rim conference on communications computers and signal processing*, volume 1, 1993.

Kunze, J., Kirsch, L., Kurenkov, I., Krug, A., Johannsmeier, J., and Stober, S. Transfer learning for speech recognition on a budget. *Proceedings of Workshop on Representation Learning for NLP (ACL Workshop)*, 2017.

Lian, Z., Tao, J., Liu, B., and Huang, J. Unsupervised representation learning with future observation prediction for speech emotion recognition. *Conference of the International Speech Communication Association (INTERSPEECH)*, 2019.

Liu, A. H., Tu, T., Lee, H.-y., and Lee, L.-s. Towards unsupervised speech recognition and synthesis with quantized speech representation learning. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

Moss, H. B., Aggarwal, V., Prateek, N., González, J., and Barra-Chicote, R. Boffin tts: Few-shot speaker adaptation by bayesian optimization. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

Park, K. “kss dataset: Korean single speaker speech dataset, 2018. URL https://www.kaggle.com/datasets/bryanpark/korean-single-speaker-speech-dataset.

Ren, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T.-Y. Almost unsupervised text to speech and automatic speech recognition. In *International Conference on Machine Learning (ICML)*, 2019.

Schneider, S., Baevski, A., Collobert, R., and Auli, M. wav2vec: Unsupervised pre-training for speech recognition. In *Conference of the International Speech Communication Association (INTERSPEECH)*, 2019.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018.

Tu, T., Chen, Y.-J., Liu, A. H., and Lee, H.-y. Semi-supervised learning for multi-speaker text-to-speech synthesis using discrete speech representation. *Conference of the International Speech Communication Association (INTERSPEECH)*, 2020.

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., et al. Tacotron: Towards end-to-end speech synthesis. *Conference of the International Speech Communication Association (INTERSPEECH)*, 2017.

Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J., Jia, Y., Chen, Z., and Wu, Y. Libritts: A corpus derived from librispeech for text-to-speech. *Conference of the International Speech Communication Association (INTERSPEECH)*, 2019.

Zhang, H. and Lin, Y. Unsupervised learning for sequence-to-sequence text-to-speech for low-resource languages. *Conference of the International Speech Communication Association (INTERSPEECH)*, 2020.