INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Unsupervised Learning For Sequence-to-sequence Text-to-speech For Low-resource Languages

Haitong Zhang, Yue Lin

NetEase Games AI Lab

{zhanghaitong01, gzlinyue}@corp.netease.com

Abstract

Recently, sequence-to-sequence models with attention have been successfully applied in Text-to-speech (TTS). These models can generate near-human speech with a large accurately-transcribed speech corpus. However, preparing such a large data-set is both expensive and laborious. To alleviate the problem of heavy data demand, we propose a novel unsupervised pre-training mechanism in this paper. Specifically, we first use Vector-quantization Variational-Autoencoder (VQ-VAE) to extract the unsupervised linguistic units from large-scale, publicly found, and untranscribed speech. We then pre-train the sequence-to-sequence TTS model by using the <unsupervised linguistic units, audio> pairs. Finally, we fine-tune the model with a small amount of < text, audio > paired data from the target speaker. As a result, both objective and subjective evaluations show that our proposed method can synthesize more intelligible and natural speech with the same amount of paired training data. Besides, we extend our proposed method to the hypothesized low-resource languages and verify the effectiveness of the method using objective evaluation.

Index Terms: unsupervised learning, sequence-to-sequence text-to-speech, low-resource languages

1. Introduction

Sequence-to-sequence text-to-speech (S2S TTS) models consisting of an encoder-decoder-with-attention framework can generate natural speech [1–5]. However, training these S2S TTS models requires tens of hour transcribed speech to produce audios with near-human naturalness. Although less data is required to produce intelligible speech, it limits overall naturalness and the model is prone to make undesirable mistakes.

Since collecting such a large transcribed speech corpus is both expensive and laborious, researchers have started to investigate the problem of data efficiency in TTS. Some researches focused on adapting a TTS model to new speakers using a small amount of data. Some proposed to fine-tune all or parts of the pre-trained model using a small amount of data from target speakers [6, 7]. Some investigated modeling speaker identities using speaker embeddings in TTS [8, 9]. Some also explored a combination of speaker embeddings and fine-tuning [10, 11]. Some even worked on zero-shot speaker adaptation [9, 12].

Other researches explored building TTS model with the aid of universal data. Some studied introducing distributional textual or linguistic information into TTS within the traditional TTS paradigm [13–15]. Some investigated training TTS models using Automatic Speech Recognition data or found data through data selection or analysis [16–19]. Recently, [20] proposed a simple yet effective semi-supervised approach to pre-train the decoder in end-to-end TTS by using only speech.

There has been some work on data efficiency in TTS for low-resource languages. It is shown that train a multi-lingual statistical parametric speech synthesis (SPSS) model can facilitate the adaptation to new languages with a small amount of data [21, 22] . A recent work [23] investigated transfer learning from high-resource languages to low-resource languages.

This work aims to alleviate the data demand for training S2S TTS by utilizing large-scale, publicly found, and untranscribed speech data. We propose an unsupervised framework for training Tacotron [2], a state-of-the-arts S2S TTS model. Specifically, we first use Vector-quantization Variational-Autoencoder (VQ-VAE) to extract the unsupervised linguistic units from the untranscribed speech. We then pre-train Tacotron by using the <unsupervised linguistic units, audio> pairs. Finally, we fine-tune the model with a small amount of <text, audio> paired data from target speakers.

It should be noticed that our work is related to [20]. However, our work is different from [20] in several ways, constituting the main contributions of our work. The first and most significant difference is that our approach utilizes unsupervised learning to extract phone-alike linguistic units, which made it possible to pre-train the entire TTS model, while [20] separatively pre-trains each part of the model. Secondly, we also verify our approach in the hypothesized low-resource languages. Lastly, we mainly use publicly accessible data in our experiments, which can be reproduced easily.

In Section 2, we review the semi-supervised pre-training in [20] and describe our proposed unsupervised method. Section 3 details the experiment settings and results. The paper is closed with a conclusion in Section 4.

2. Proposed Method

We use a baseline Tacotron model architecture [2], where we use location-sensitive-attention (LSA) and phoneme sequence derived from the text. To convert the predicted spectrograms into waveforms, we use Griffin-Lim algorithm [24] for fast experiment cycles, since we focus on the problem of data efficiency rather than generating high-fidelity speech. In the baseline model, the model is trained from scratch, which means all the model parameters are trained by paired data.

2.1. Semi-supervised pre-training

In the baseline Tacotron model, the model should simultaneously learn the textual representations, acoustic representations, and the alignment between them. [20] propose two types of model pre-training to utilize external textual and acoustic information. For textual representations, they pre-train Tacotron's encoder by the external word-vectors; for acoustic representations, they pre-train the decoder by untranscribed speech.

[20] then fine-tune the whole model using paired data. At this step, the model focuses on learning the alignments between textual representations and acoustic ones.

http://dx.doi.org/10.21437/Interspeech.2020-1403

Step-1 :

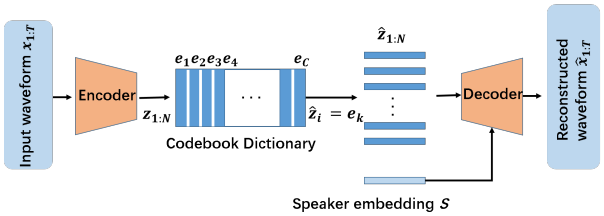

Figure 1: *VQ-VAE for extracting linguistic units.*

Step-2/3 :

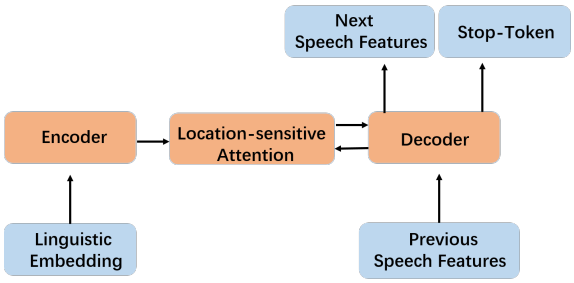

Figure 2: *Tacotron model architecture studied.*

2.2. Unsupervised Learning for pre-training

Although [13] shows the proposed semi-supervised pre-training helps the model synthesizes more intelligible speech, it finds that pre-training the encoder and decoder separately at the same time does not bring further improvement than only pre-training the decoder. However, there is a mismatch between pre-training only the decoder and fine-tuning the whole model. To avoid potential error introduced by this mismatch and further improve the data efficiency by using only speech, we propose to extract unsupervised linguistic units from untranscribed speech to pre-train the entire model.

Our proposed framework is provided in Algorithm 1. The whole framework includes two models: an unsupervised model for extracting phone-alike linguistic units (see Figure 1) and Tacotron model (see Figure 2).

2.2.1. *Unsupervised linguistic units*

Unsupervised speech representation has gained a great improvement in both representation and disentangling [25–30]. Among them, discretized representations are popular in language and speech community, because it is believed that language or speech is composed of a limited set of discretized units, such as characters in text and phonemes in speech. In this paper, we utilize VQ-VAE model [28] as the extractor of discretized linguistic units.

In this case, VQ-VAE acts as a recognition model similar to an automatic speech recognition (ASR) model. However, the main difference between VQ-VAE and ASR model is that VQ-VAE is trained in an unsupervised fashion while the ASR model trained in a supervised mode. This difference matters as far as low-resource languages are concerned. Whereas an ASR model for low-resource languages is not typically available, the proposed unsupervised method remains helpful in extracting linguistic units for low-resource languages.

Algorithm 1: Proposed Method

Step1: Training VQ-VAE using untranscribed speech
Step2: Tacotron Pre-training:
2.1 Unsupervised linguistic units extraction:
for *utt in untranscribed speech* **do**
1. feed utt into the trained VQ-VAE, and extract the nearest embeddings as the unsupervised linguistic units;
2. delete the consecutive repeated unit from the sequence;

2.2 pre-train Tacotron using <linguistic unit, audio> pair;
Step3: Tacotron Fine-tuning using <text, audio> pair.

VQ-VAE has an encoder-decoder architecture and a codebook dictionary $e = C * D$, where C is the number of latent embeddings in the dictionary and D is the dimension of each embedding. The encoder E takes raw waveform $x_{1:T} = x_1, x_2, \ldots, x_T$ as inputs, and produces the encoded representation $z_{1:N} = E(x_{1:T})$, where N depends on the length T and the number of down-sampling layers in the encoder. Then the continuous latent representations $z_{1:N}$ can be mapped into $\hat{z}_{1:N}$ by finding the nearest pre-defined discretized embedding in the dictionary as $\hat{z} = e_k$, where $k = argmin_j \, ||z - e_j||$, and e_j is the j-th embedding in the codebook dictionary, and $j \in 1, 2, \ldots, C$. Finally the latent embeddings $\hat{z}_{1:N}$ and the speaker embedding s are together passed into decoder D to reconstruct the raw-waveform $\hat{x} = D(\hat{z}, s)$.

Since the model input and output are the same, the model can be trained as an auto-encoder. However, the gradients cannot be gained from the argmin operation, thus [28] uses straight-through gradient estimation to approximate them. Then the final loss of the entire model is

$$\begin{aligned} L = & -log\left(x \mid \hat{z}(x)\,,\, s\right) \\ & + ||sg(z(x)) - e_j)||_2^2 \\ & + \beta * ||z(x) - sg(e_j)||_2^2 \end{aligned} \tag{1}$$

where the first term is the negative log-likelihood to update the whole model. The second term updates the codebook dictionary, with sg denotes stop-gradient operation. The third term, referred to the commitment loss, encourages the encoder output z to get close to the codebook embeddings, with the hyper-parameter β to weight the term.

2.2.2. *Tacotron Pre-training & Fine-tuning*

After VQ-VAE is trained, we extract the unsupervised linguistic units for each utterance. We then randomly initialize an embedding table for all the unsupervised linguistic units, and the linguistic embedding sequence by looking up the table is used as the input of Tacotron. Thus, we can pre-train Tacotron by <linguistic embedding, audio> pairs.

After the model is pre-trained as mentioned above, we fine-tune the model with some paired speech data. In the step, the inputs of the model are phoneme sequences derived from the normalized text.

Table 1: *Result of MCD objective test of four model variants, the smaller is better. All the models are trained using 24-minute speech. The best model (except for the upper-bound of the model) is marked in bold.*

Tac	T-Dec	T-VQ	T-Phone
22.24	19.57	**19.06**	18.85

Table 2: *Results of AB test of each pair of model variants. All the models are trained using 24-minute of paired data.*

Model Pair	Preference %		
	Former	Latter	N/A
Tac vs. T-Dec	1.25	80.25	18.5
Tac vs. T-VQ	0	**97.5**	2.5
T-Dec vs. T-VQ	4.25	**85**	10.75
T-VQ vs. T-Phone	6.25	20	**73.5**

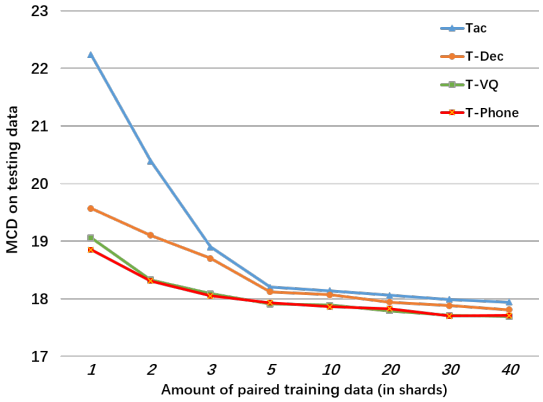

Figure 3: *MCD test results of all model variants on various amount of paired data, with 1 shard equals to 24 minutes.*

3. Experiment

3.1. Experimental setup

We conduct experiments to show the effectiveness of our proposed method. We use the LJspeech dataset [31] for model fine-tuning. The architecture of VQ-VAE investigated in this paper is similar to [30]. When training VQ-VAE, we use 39-dimension MFCCs as the model input. After our preliminary study, we set the codebook size into 256, and the dimension of each embedding 64. The jitter rate and β is 0.12 and 0.25, respectively. We encourage readers to read [30] for more details.

[20] found that 24-minute speech is the maximum amount of data that could rarely successfully build a baseline Tacotron to produce intelligible speech. Thus, in the next section, we focus on comparing all model variants trained with only 24-minute paired data.

3.2. Results on 24-minute data

The model variants investigated in this section include:

- Tac: Tacotron trained by only LJspeech;
- T-Dec: Tacotron pre-trained by external speech in the semi-supervised mode, then fine-tuned by LJspeech;
- T-VQ: Tacotron pre-trained by external speech in the proposed mode, then fine-tuned by LJspeech;
- T-Phone: Tacotron pre-trained by external speech in the supervised mode, then fine-tuned by LJspeech, referred to the upper-bound of the model.

We use VCTK [32] as the external speech dataset in this section. As mentioned above, we only use speech data in VCTK when pre-training T-Dec and T-VQ. For T-Phone, we use <text, audio> paired data in VCTK for pre-training to provide the upper-bound performance in this scenario.

We conduct both objective and subjective evaluations to measure synthesis quality. For the objective evaluation, we compute Dynamic-time-warping Mel-cepstral Distortion (DTW MCD) [33], which measures the distance between the synthesized and ground-truth speech, and the smaller is better. We use about 20-minute unseen speech as the evaluation data. For the subjective one, we conduct a series of AB preference tests using 20 unseen utterances of various lengths. [1] 20 raters (with ten males and ten females) who are native Mandarin speakers and proficient in English are included in the subjective test.

3.2.1. MCD objective test

The MCD results are provided in Table 1. As in [20], only pre-training the decoder can lower MCD. However, the proposed framework provides the best performance, whose MCD is 14.30% lower than the baseline Tacotron. We also find that T-VQ's performance is close to that of the upper-bound of the model (i.e. T-Phone).

3.2.2. AB subjective test

The results of the AB test are shown in Table 2. It is clear that all pre-training techniques help to improve model performance. There is a large performance difference between baseline Tacotron and models with pre-training (i.e. T-Dec and T-VQ). We find that training the model from scratch with LJspeech can hardly get intelligible speech, partly because the quality of LJspeech is not satisfactorily high.

In the ABtest between T-Dec with the proposed T-VQ, T-VQ gets more preference from raters. From the informal listening test, we notice that synthesized speech by T-Dec is moderate in intelligibility, while T-VQ produces more intelligible speech. This indicates that pre-training by both unsupervised linguistic units and audio can further improve model performance. The reason is that in the proposed pre-training step, the model can not only learn the acoustic representation, but also the alignment between the acoustic and textual representation. Although the unsupervised linguistic embeddings are not used in fine-tuning the model, we believe that the proposed pre-training is beneficial to textual representation learning since these unsupervised linguistic units have been proven to be phone-alike [30].

In the comparison between Tac-VQ and T-Phone, most raters show no preference, although raters prefer T-Phone over T-VQ by 20%.

[1] Speech demos are available at https://haitongzhang.github.io/DE-TTS/

3.3. Results on other amounts of data

We also conduct MCD objective evaluation on all model variants with various amounts of data. The results are provided in Figure 3. Each curve in Figure 3 represents the MCD between the ground-truth speech and synthesis by each model variant with various amounts of paired data for model training/fine-tuning. From the figure, we can see that there is a large margin between the baseline Tacotron and other model variants at 1shard (i.e. 24-minute). Another obvious trend is that with the amount of paired data increasing, the MCD differences decrease, which denotes the reducing effect of pre-training. However, regardless of the amount of paired data, T-VQ and T-Phone always achieve a lower MCD than Tac and T-Dec.

3.4. Results on low-resource languages

In this section, we verify the effectiveness of the proposed approach for two hypothesized low-resource languages. In the section, we hypothesize that English and Chinese Mandarin are two low-resource languages, in which large-scale and publicly-found speech in these two languages can not be easily collected. Thus, we resort to pre-training the model by the publicly-found speech in other languages. In this section, we mainly focus on answering the two following questions:

1. Is our proposed method beneficial to improving the data efficiency in this case ?
2. What pre-training languages are more efficient in the proposed framework? Those acoustically-closely-related to target language or those acoustically dissimilar?

In this section, the paired data for English TTS remains LJspeech, and that for Mandarin comes from an internal news-style corpus recorded by a female speaker - Xiaomin. For training VQ-VAE and pre-training Tacotron, we use open-source corpus in the following five languages: Korean [34], Japanese [2] [35], Spanish, French, German[3] [36]. As mentioned above, we only use speech for training VQ-VAE and pre-training Tacotron. Only one modification is made in training VQ-VAE: the codebook size changes from 256 to 512, since multi-lingual data is used in this scenario. In building the English and Mandarin TTS model, we investigate the following three model variants:

- Tac: Tacotron trained by paired data from LJspeech or Xiaomin;
- T-VQ-A: Tacotron pre-trained, in the proposed mode, by external speech in Asian languages (i.e Korean and Japanese), then fine-tuned by paired data from LJspeech or Xiaomin;
- T-VQ-E: Tacotron pre-trained, in the proposed mode, by external speech in European languages (i.e. Spanish, French, and German), then fine-tuned by paired data from LJspeech or Xiaomin;

To alleviate the burden of raters, we only provide MCD objective test results in this section. The MCD results of English and Mandarin TTS are provided in Table 3 and 4, respectively. It clearly shows that our proposed pre-training approach improves the quality of the synthesized speech, which is important to low-resource languages since collecting paired data would be much more difficult. Besides, T-VQ-E out-performs than T-VQ-A in English TTS, and T-VQ-A out-performs slightly than T-VQ-E in the Mandarin experiment in most cases. This result indicates that pre-training with the speech in acoustically-close languages is more efficient than with acoustically dissimilar speech. Also, we found a similar decreasing trend in MCD with the increasing amount of fine-tuning data as in the previous section. Lastly, by comparing the best model variant in this section in English TTS (i.e. T-VQ-E) with the best model variant in the previous section (T-VQ), we found that there is still a non-negligible gap between pre-training with the speech in the target language and in the acoustically-close language (see the rows in bold in Table 3), which is needed further investigation.

[2]https://sites.google.com/site/shinnosuketakamichi/publication/jsut

[3]https://voice.mozilla.org/en /datasets

Table 3: *MCD results of model variants by various amount of paired data in English TTS. Better results are marked in bold.*

Model	Training/Fine-tune Paired data (in shards)					
	0.5	1	1.5	2	2.5	3
Tac	28.72	22.24	21.10	20.39	19.10	18.90
T-VQ-A	25.25	20.77	19.64	18.92	18.62	18.50
T-VQ-E	**24.2**	**20.14**	**18.73**	**18.54**	**18.56**	**18.45**
T-VQ	-	**19.06**	-	**18.33**	-	**18.09**

Table 4: *MCD results of model variants by various amount of paired data in Mandarin TTS. Better results are marked in bold.*

Model	Training/Fine-tune Paired data (in shards)					
	0.5	1	1.5	2	2.5	3
Tac	24.18	23.55	22.59	21.67	20.13	19.73
T-VQ-A	**23.48**	**18.44**	16.91	16.31	**15.81**	**15.49**
T-VQ-E	23.69	18.63	**16.81**	**16.29**	16.02	15.93

4. Conclusion

In this paper, we propose using unsupervised learning for improving data efficiency in sequence-to-sequence TTS for low-resource languages. Our method utilizes large-scale untranscribed speech to externally provide textual and acoustic information to Tacotron. We have shown that our proposed approach works in sequence-to-sequence TTS framework. Specifically, with the proposed pre-training method, Tacotron can produce intelligible speech with less paired training data. Although we conduct our experiments using Tacotron architecture, we believe that our proposed framework should be feasible in other sequence-to-sequence TTS models. We also verify the effectiveness of the method on two hypothesized low-resource languages. This promisingly indicates that even with the non-target untranscribed speech, our proposed approach could provide a significant performance improvement. Although we use hypothesized low-resource languages, we believe that our method can generalize to real low-resource languages. This significant result also sheds light on data collection for both mono-language and multi-language TTS.

Although promising results are given, there is a lot to be investigated. For example, there are many other unsupervised models to be investigated. Besides that, since we focus on the validation of the effectiveness of the proposed framework, we use Griffin-Lim as the spectrogram-to-waveform algorithm. To fully realize small paired-data sequence-to-sequence TTS, we need to investigate the adaptation of neural vocoders using small data.

5. References

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *Proc. Interspeech 2017*, pp. 4006–4010, 2017.

[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[3] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep Voice 3: Scaling text-to-speech with convolutional sequence learning," in *International Conference on Learning Representations*, 2018.

[4] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, and Y. Bengio, "Char2wav: End-to-end speech synthesis," 2017.

[5] W. Ping, K. Peng, and J. Chen, "Clarinet: Parallel wave generation in end-to-end text-to-speech," in *International Conference on Learning Representations*, 2018.

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

[7] H. B. Moss, V. Aggarwal, N. Prateek, J. González, and R. Barra-Chicote, "BOFFIN TTS: Few-shot speaker adaptation by bayesian optimization," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7639–7643.

[8] J. Park, K. Zhao, K. Peng, and W. Ping, "Multi-speaker end-to-end speech synthesis," *arXiv preprint arXiv:1907.04462*, 2019.

[9] E. Nachmani, A. Polyak, Y. Taigman, and L. Wolf, "Fitting new speakers based on a short untranscribed sample," in *International Conference on Machine Learning*, 2018, pp. 3683–3691.

[10] Y. Deng, L. He, and F. Soong, "Modeling multi-speaker latent space to improve neural tts: Quick enrolling new speaker and enhancing premium voice," *arXiv preprint arXiv:1812.05253*, 2018.

[11] S. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," in *Advances in Neural Information Processing Systems*, 2018, pp. 10 019–10 029.

[12] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. L. Moreno, Y. 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.

[13] O. S. Watts, "Unsupervised learning for text-to-speech synthesis," Ph.D. dissertation, The University of Edinburgh, 2012.

[14] O. Watts, Z. Wu, and S. King, "Sentence-level control vectors for deep neural network speech synthesis," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[15] P. Wang, Y. Qian, F. K. Soong, L. He, and H. Zhao, "Word embedding for recurrent neural network based tts synthesis," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4879–4883.

[16] E. Cooper and X. Wang, "Utterance selection for optimizing intelligibility of TTS voices trained on asr data," *Interspeech 2017*, vol. 1, 2017.

[17] F.-Y. Kuo, S. Aryal, G. Degottex, S. Kang, P. Lanchantin, and I. Ouyang, "Data selection for improving naturalness of TTS voices trained on small found corpuses," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 319–324.

[18] F. Kuo, I. Ouyang, S. Aryal, and P. Lanchantin, "Selection and training schemes for improving TTS voice built on found data," *Proc. Interspeech 2019*, pp. 1516–1520, 2019.

[19] J. Fong, P. O. Gallegos, Z. Hodari, and S. King, "Investigating the robustness of sequence-to-sequence text-to-speech models to imperfectly-transcribed training data," in *Proc. Interspeech*, 2019.

[20] Y.-A. Chung, Y. Wang, W.-N. Hsu, Y. Zhang, and R. Skerry-Ryan, "Semi-supervised training for improving data efficiency in end-to-end speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6940–6944.

[21] Q. Yu, P. Liu, Z. Wu, S. K. Ang, H. Meng, and L. Cai, "Learning cross-lingual information with multilingual BLSTM for speech synthesis of low-resource languages," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5545–5549.

[22] A. Gutkin, "Uniform multilingual multi-speaker acoustic model for statistical parametric speech synthesis of low-resourced languages," *Proc. Interspeech 2017*, pp. 2183–2187, 2017.

[23] Y.-J. Chen, T. Tu, C.-c. Yeh, and H.-Y. Lee, "End-to-end text-to-speech for low-resource languages by cross-lingual transfer learning," *Proc. Interspeech 2019*, pp. 2075–2079, 2019.

[24] N. Perraudin, P. Balazs, and P. L. Søndergaard, "A fast Griffin-Lim algorithm," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2013, pp. 1–4.

[25] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Advances in neural information processing systems*, 2009, pp. 1096–1104.

[26] J. Glass, "Towards unsupervised speech processing," in *2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*. IEEE, 2012, pp. 1–4.

[27] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," in *Advances in neural information processing systems*, 2017, pp. 1878–1889.

[28] A. van den Oord, O. Vinyals, and k. kavukcuoglu, "Neural discrete representation learning," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 6306–6315. [Online]. Available: http://papers.nips.cc/paper/7210-neural-discrete-representation-learning.pdf

[29] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. Glass, "An unsupervised autoregressive model for speech representation learning," *Proc. Interspeech 2019*, pp. 146–150, 2019.

[30] J. Chorowski, R. J. Weiss, S. Bengio, and A. van den Oord, "Unsupervised speech representation learning using WaveNet autoencoders," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 12, pp. 2041–2053, 2019.

[31] K. Ito, "The LJ speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[32] C. Veaux, J. Yamagishi, and K. MacDonald, "Superseded - CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," 2017. [Online]. Available: http://datashare.is.ed.ac.uk/handle/10283/2651

[33] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, vol. 1. IEEE, 1993, pp. 125–128.

[34] "Zeroth-Korean, http://www.openslr.org/40/."

[35] R. Sonobe, S. Takamichi, and H. Saruwatari, "JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis," *arXiv preprint arXiv:1711.00354*, 2017.

[36] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common Voice: A massively-multilingual speech corpus," in *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 4218–4222.