Pre-training Techniques for Improving Text-to-Speech Synthesis by Automatic Speech Recognition Based Data Enhancement

Yazhu Liu, Shaofei Xue(✉), and Jian Tang

AIspeech Ltd., Suzhou, China
{yazhu.liu,shaofei.xue,jian.tang}@aispeech.com

Abstract. As the development of deep learning, neural network (NN) based text-to-speech (TTS) that adopts deep neural networks as the model backbone for speech synthesis, has now become the mainstream technology for TTS. Compared to the previous TTS systems based on concatenative synthesis and statistical parametric synthesis, the NN based speech synthesis shows conspicuous advantages. It needs less requirement on human pre-processing and feature development, and brings high-quality voice in terms of both intelligibility and naturalness. However, robust NN based speech synthesis model typically requires a sizable set of high-quality data for training, which is expensive to collect especially in low-resource scenarios. It is worth investigating how to take advantage of low-quality material such as automatic speech recognition (ASR) data which can be easily obtained compared with high-quality TTS material. In this paper, we propose a pre-training technique framework to improve the performance of low-resource speech synthesis. The idea is to extend the training material of TTS model by using ASR based data augmentation method. Specifically, we first build a frame-wise phoneme classification network on the ASR dataset and extract the semi-supervised <linguistic features, audio> paired data from large-scale speech corpora. We then pre-train the NN based TTS acoustic model by using the semi-supervised <linguistic features, audio> pairs. Finally, we fine-tune the model with a small amount of available paired data. Experimental results show that our proposed framework enables the TTS model to generate more intelligible and natural speech with the same amount of paired training data.

Keywords: Pre-training techniques · neural network · text-to-speech · automatic speech recognition

1 Introduction

Recent advances in neural network (NN) based text-to-speech (TTS) have significantly improved the naturalness and quality of synthesized speech. We are now able to generate high-quality human-like speech from given text with less

L. Zhenhua et al. (Eds.): NCMMSC 2022, CCIS 1765, pp. 162–172, 2023.
https://doi.org/10.1007/978-981-99-2401-1_15

requirement on human pre-processing and feature development [1–5]. However, such models typically require tens of hour transcribed dataset consisting of high-quality text and audio training pairs, which are expensive and time consuming to collect. Requiring large amounts of data limits the overall naturalness and applicability especially in low-resource scenarios.

A series of extended technologies have been developed to improve the data efficiency for NN based TTS training. Most of these existing methods can be grouped into three categories: dual transformation, transfer learning and self-supervised/semi-supervised training. Firstly, dual transformation mainly focuses on the dual nature of TTS and automatic speech recognition (ASR). TTS and ASR are two dual tasks and can be leveraged together to improve each other. Speech chain technique is presented in [6] to construct a sequence-to-sequence model for both ASR and TTS tasks as well as a loop connection between these two processes. The authors in [7] develop a TTS and ASR system named LRSpeech which use the back transformation between TTS and ASR to iteratively boost the accuracy of each other under the extremely low-resource setting. In [8], it proposes an almost unsupervised learning method that only leverages few hundreds of paired data and extra unpaired data for TTS and ASR by using dual learning. Secondly, although paired text and speech data are scarce in low-resource scenarios, it is abundant in rich-resource scenarios. Transfer learning approaches try to implement adaptation methods and retain the satisfactory intelligibility and naturalness. Several works attempt to help the mapping between text and speech in low-resource languages with pre-training the TTS models on rich-resource languages [9–12]. In order to alleviate the difference of phoneme sets between rich and low-resource languages. The work in [13] proposes to map the embeddings between the phoneme sets from different languages. In [14], international phonetic alphabet (IPA) is adopted to support arbitrary texts in multiple languages. Besides that, voice conversion (VC) [15,16] is also an effective way to improve the data efficiency in low-resource TTS training. Recent work in [17] brings significant improvements to naturalness by combining multi-speaker modelling with data augmentation for the low-resource speaker. This approach uses a VC model to transform speech from one speaker to sound like speech from another, while preserving the content and prosody of the source speaker. Finally, self-supervised/semi-supervised training strategies are leveraged to enhance the language understanding or speech generation capabilities of TTS model. For example, paper [18] aims to lower TTS systems' reliance on high quality data by providing them the textual knowledge, which is extracted from BERT [19] language models during training. They enrich the textual information through feeding the linguistic features that extracted by BERT from the same input text to the decoder as well along with the original encoder representations. In [20], the researchers propose a semi-supervised training framework to allow Tacotron to utilize textual and acoustic knowledge contained in large, publicly available text and speech corpora. It first embeds each word in the input text into word vectors and condition the Tacotron encoder on them. Then an unpaired speech corpus is used to pre-train the Tacotron decoder in the acoustic domain. Finally, the model is fine-tuned using available paired data. An unsupervised pre-training mechanism that uses Vector-Quantization Variational-Autoencoder (VQ-VAE) [21] to extract the unsupervised linguistic units from the untranscribed speech is investigated in [22]. More recently, an unsupervised TTS system based on an alignment module that outputs pseudo-text and another synthesis module that uses pseudo-text for training and real text for inference, is presented in [23].

The motivation of this work is to develop novel techniques to alleviate the data demand for training NN based TTS. We propose a semi-supervised pre-training technique framework to improve the performance of speech synthesis by extending the training material of TTS model with ASR based data augmentation. Specifically, we first build a frame-wise phoneme classification network on ASR dataset and extract the semi-supervised <linguistic features, audio> paired data from large-scale speech corpora. Then, we pre-train the NN based TTS acoustic model by using the semi-supervised <linguistic features, audio> pairs. Finally, we fine-tune the model with a small amount of available paired data.

It should be noticed that similar semi-supervised pre-training work has been related in [20]. However, our work is different in several ways, constituting the main contributions of our work. Firstly, the semi-supervised <linguistic features, audio> paired data for pre-training TTS model are extracted from a frame-wise phoneme classification network, which is built from the beginning based on the ASR dataset. It makes us possible to pre-train the entire TTS acoustic model, while the encoder and decoder are separately pre-trained in [20]. Secondly, the acoustic model of TTS system implemented in our work is different. We choose to use AdaSpeech [5] which involves the adaptive custom voice technique by inserting speaker embedding as the conditional information. Finally, we investigate and analyze the effectiveness of building low-resource language TTS systems with the help of semi-supervised pre-training on the rich-resource language.

The rest of this paper is organized as follows: In Sect. 2, we briefly review the architecture of TTS model used in this work. In Sect. 3, our proposed novel techniques to improve the performance of low-resource TTS are described. Section 4 shows our experimental setups and detailed results on Mandarin and Chinese Dialects tasks. Several conclusions are further drawn in Sect. 5.

2 TTS Model

As the development of deep learning, NN based TTS that adopts deep neural networks as the model backbone for speech synthesis, has now become the mainstream technology for TTS. Compared to the previous TTS systems based on concatenative synthesis and statistical parametric synthesis, the NN based speech synthesis shows conspicuous advantages. It needs less requirement on human pre-processing and feature development, and brings high-quality voice in terms of both intelligibility and naturalness. A NN based TTS system often consists of three basic components: a text analysis module, an acoustic model (abbr. TTS-AM), and a vocoder. The text analysis module converts a text sequence into the linguistic features, and then TTS-AM transforms linguistic features to the acoustic features, finally the vocoder synthesizes the waveform based on the acoustic features.

2.1 Text Analysis Module

In the TTS system, the text analysis module has an important influence on the intelligibility and naturalness of synthesized speech. The typical text analysis module in a Chinese TTS system consists of text normalization (TN), Chinese word segmentation (CWS), part-of-speech (POS) tagging, grapheme-to-phoneme (G2P) conversion, and prosody prediction. It extracts various linguistic features from the raw text, aiming to provide enough information for training the TTS-AM.

2.2 Acoustic Model

In this work, we choose to use AdaSpeech, which is a non-autoregressive model based on the transformer architecture. The basic model backbone consists of a phoneme encoder, a spectrogram decoder, an acoustic condition modeling that captures the diverse acoustic conditions of speech in speaker level, utterance level and phoneme level. And a variance adaptor which provides variance information including duration, pitch and energy into the phoneme hidden sequence. The decoder generates spectrogram features in parallel from the predicted duration and other information.

2.3 Vocoder

The vocoder in our work is based on LPCNet [24–26]. It introduces conventional digital signal processing into neural networks, and uses linear prediction coefficients to calculate the next waveform point while leveraging a lightweight RNN to compute the residual. This makes it possible to match the quality of state-of-the art neural synthesis systems with fewer neurons, significantly reducing the complexity. The LPCNet is a good compromise between quality and inference speed for a TTS system. As the LPCNet uses bark-frequency cepstrum as input, we modify the AdaSpeech to generate bark-frequency cepstrum as output. No external speaker information, such as speaker embedding, is referred in the building of LPCNet model.

3 The Proposed Approach

In this section, the proposed semi-supervised pre-training framework on TTS modeling is detailed. The Illustration of our general framework is shown in Fig. 1. We first describe the structure of frame-wise phoneme classification model and the alignment module that greedily proposes a pairing relationship between speech utterances and phoneme transcripts. After that, the pre-training and fine-tuning procedures of our method are presented.

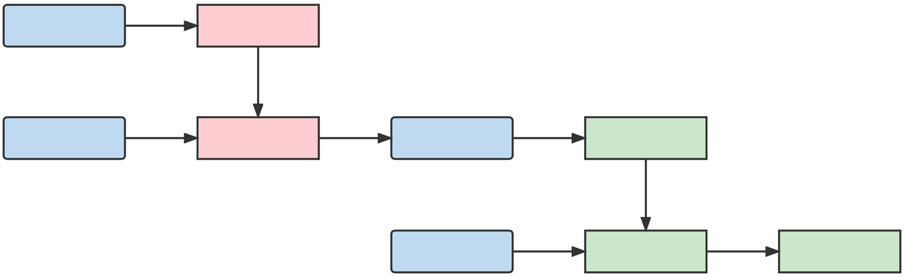

Fig. 1. Illustration of the proposed semi-supervised pre-training framework on TTS modeling.

3.1 Frame-Wise Phoneme Classification

DFSMN Model. DFSMN is an improved FSMN architecture by introducing the skip connections and the memory strides [27]. The DFSMN component consists of four parts: a ReLU layer, a linear projection layer, a memory block and a skip connection from the bottom memory block, except for the first one that without the skip connection from the bottom layer. By adding the skip connections between the memory blocks of DFSMN components, the output of the bottom layer memory block can directly flow to the upper layer. During back-propagation, the gradients of higher layer can also be assigned directly to lower layer that help to overcome the gradient vanishing problem. Since the information of adjacent frames in speech signals always have strong redundancy due to the overlap. The strides for look-back and look-ahead are used to help the DFSMN layer remove the redundancy in adjacent acoustic frames. DFSMN is able to model the long-term dependency in sequential signals while without using recurrent feedback. In practice, DFSMN models usually contain DFSMN layers around ten to twenty. We follow the model topology in [28] and implement a DFSMN with ten DFSMN layers followed by two fully-connected ReLU layers, a linear layer and a softmax output layer. To avoid the mismatch in G2P conversion, we share the same phoneme set between phoneme classification and TTS tasks. We adopt IPA as described in [14] to support arbitrary texts in Mandarin and multiple Chinese Dialects evaluations in our work.

Alignment Module. After training DFSMN based phoneme classification model, the semi-supervised paired data for pre-training TTS model has to be prepared. Pseudo phoneme transcript of the training set is first generated by greedy decoding over the output of DFSMN. Instead of extracting the phoneme duration with soft attention mechanism as described in [20, 22], the alignment between pseudo phoneme transcript and speech sequence is derived from a forced alignment procedure computed by Kaldi [29] with a phonetic decision tree. This improves the alignment accuracy and reduces the information gap between the model input and output.

3.2 Semi-supervised Pre-training

In the baseline AdaSpeech, the model should simultaneously learn the textual representations, acoustic representations, and the alignment between them. The encoder takes a source phoneme text as input and produces sequential representations of it. The decoder is then conditioned on the phoneme representations to generate corresponding acoustic representations, which are then converted to waveforms. [20] proposes two types of pre-training methods to utilize the external textual and the acoustic information. For textual representations, they pre-train encoder by the external word-vectors. For acoustic representations, they pre-train the decoder by untranscribed speech. Although [20] shows that 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 instead use the semi-supervised paired data generated by the frame-wise phoneme classification model as described in Sect. 3.1. It helps to alleviate the mismatch problem and makes pre-training the entire model possible.

3.3 AdaSpeech Fine-Tuning

The AdaSpeech in pre-trained model is applied as a multi-speaker TTS-AM, which means we do not use the adaptive custom voice technique as described in [5]. After that, the AdaSpeech is fine-tuned with some high-quality paired speech data from the target speaker. In this procedure, the inputs of the model are phoneme sequences derived from the normalized text.

4 Experiments

In this section, we evaluate the performance of the proposed approach on two type of TTS tasks including single-speaker Mandarin dataset and multi-speaker Chinese Dialects dataset. For both two experiments, we use a 3000-hour Mandarin dataset which consists of 1000-hour low-quality transcribed ASR data (1000 h-TD) and 2000-hour low-quality untranscribed data (2000 h-UTD) for pre-training. The data are collected from many domains, such as voice search, conversation, video and the sample rate of the data is 16 kHz.

In the ASR setup, waveform signal is analyzed using a 25-ms Hamming window with a 10-ms fixed frame rate. 40-dimensional filterbank features are used for training DFSMN phoneme classification models. The features are pre-processed with the global mean and variance normalization (MVN) algorithm. We use

11 frames (5-1-5) of filter-banks as the input features of neural networks. The DFSMNs model stacked with 1 Relu Layer (2048 hidden nodes), 10 DFSMN layers (2048 memory block size, 512 projection size, 10*[2048-512]), 4 ReLU layers (2*2048-1024-512) and 1 Softmax output layer.

In the TTS setup, 20-dimensional features, which consist of 18 Bark-scale cepstral coefficients and 2 pitch parameters (period, correlation), are extracted from 16k audio using a 25-ms Hamming window with a 10-ms fixed frame rate. The TTS-AM of AdaSpeech consists of 6 feed-forward Transformer blocks for the phoneme encoder and the decoder. The hidden dimension of phoneme embedding, speaker embedding and self-attention are all set to 384. The number of attention heads is 4 in the phoneme encoder and the decoder. The pre-train and fine-tune models are separately trained in a distributed manner using stochastic gradient descent (SGD) optimization on 16 GPUs and 4 GPUs.

4.1 Single-Speaker Mandarin Task

Experimental Setup. In the single-speaker Mandarin task, we evaluate our method with the Chinese Standard Mandarin Speech Corpus (CSMSC). CSMSC has 10,000 recorded sentences read by a female speaker, totaling 12 h of natural speech with phoneme-level text grid annotations and text transcriptions. The corpus is randomly partitioned into non-overlapping training, development and test sets with 9000, 800 and 200 sentences respectively. We conduct several experimental setups to investigate the influence of semi-supervised pre-training. All parameters of the TTS-AM are directly updated during the fine-tuning stage. For better comparing the efficiency of pre-training on TTS-AM, we use the same LPCNet which is trained on full 12 h CSMSC. The performance of the overall quality samples is evaluated using the mean opinion score (MOS). Listeners are asked to rate the overall naturalness and prosodic appropriateness of samples on a scale from 1 and 5. Then these synthesized samples are mixed with real speech samples and presented to listeners independently in random order. 15 raters who are native Mandarin speakers are included in the subjective test.

Performance of Phoneme Classification Models The phoneme error rate (PER) performance of using different amounts of low-quality transcribed data to build phoneme classification models is shown in Table 1. We evaluate with three test sets, including the CSMSC development set (Test-c), a 5 h dataset that randomly sampled from the 1000 h-TD (Test-i) and exists in each training data, a 4 h dataset that randomly sampled from the 2000 h-UTD which never exists in the training sets (Test-o). It can be observed that increasing the amount of training data yields a large improvement on the PER. To better evaluate the relationship between PER and pre-training efficiency, we use the recognized phoneme transcripts of all training sets for alignment to generate the semi-supervised paired data.

Table 1. PER% performance of phoneme classification models on three evaluation tasks.

Training Data Size	Test-c	Test-i	Test-o
100 h	9.3	18.8	26.9
1000 h	6.6	12.2	15.4

Results on Different Phoneme Classification Models In this section, the results of implementing different phoneme classification models to generate semi-supervised data for TTS-AM pre-training are presented. The mean MOS scores on CSMSC test set using two different DFSMN models are gradually explored. DFSMN-1000 h indicates that we use the 1000 h-TD to train the DFSMN model. DFSMN-100 h means the DFSMN model is built with 100 h subset of the 1000 h-TD. When generating the semi-supervised data for pre-training TTS-AM, we choose to use the same 100 h subset data. The results shown in Table 2 confirm that our proposed semi-supervised pre-training method brings conspicuous improvement on the MOS especially when we utilize only 15 min paired data. Besides that, the MOS on DFSMN-1000 h pre-trained model is slightly better then the DFSMN-100 h pre-trained model. It indicates that achieving higher accuracy semi-supervised paired data is also a feasible way for improving the intelligibility and naturalness of synthesized speech. In the next experiment, we choose DFSMN-1000 h model to generate all semi-supervised data.

Table 2. The mean MOS on CSMSC of using different phoneme classification models to generate TTS-AM pre-training data.

Fine-tuning Data Size	Without Pre-training	Pre-trained Model		Ground Truth
		DFSMN-100 h	DFSMN-1000 h	
15 min	2.80	3.55	**3.70**	4.71
2 h	3.99	**4.09**	4.03	
10 h	4.11	4.05	**4.16**	

Results on Different Amounts of Pre-training Data. In this experiment, we compare the results of using different amounts of data for pre-training. We utilize Size-N for labelling the data size used in TTS-AM pre-training. Thus, Size-100 h indicates that we use the same 100 h subset data as in above experiment. Size-1000 h stands that we use the whole 1000 h-TD for TTS-AM pre-training. Size-3000 h means we expand the dataset for pre-training TTS-AM by including the 2000 h-UTD. As shown in Table 3, several conclusions can be drawn from the results. Firstly, the results suggest that expanding the pre-training data size directly helps the speech synthesis performance. The MOS of fine-tuning Size-3000 h TTS-AM with 15 min paired data is similar with directly training on 2 h paired data. Secondly, it seems that the difference between using pre-training model and without pre-trained model is small when enough paired TTS data

Table 3. The mean MOS on CSMSC of using different amounts of pre-training data.

Fine-tuning Data Size	Without Pre-training	Pre-train Data Size			Ground Truth
		Size-100 h	Size-1000 h	Size-3000 h	
15 min	2.80	3.70	3.83	**3.93**	4.71
2 h	3.99	4.03	**4.09**	4.01	
10 h	4.11	**4.16**	4.10	4.08	

has been involved in the building process. For example, when fine-tuning on 10 h TTS data, we observe no conspicuous improvement on the MOS evaluation.

4.2 Multi-speaker Chinese Dialects Task

Experimental Setup. In the multi-speaker Chinese Dialects task, we evaluate our method with the two Chinese Dialect speech corpuses including Shanghainese and Cantonese. The Shanghainese corpus consists of 3 female speakers, each has 1 h recorded speech. The Cantonese corpus has 2 female speakers, each also has 1 h recorded speech. The pre-trained TTS-AM model used in this task is trained with 3000 h Mandarin dataset. For better comparison, the fine-tuning procedure is separately utilized for each speaker. And we build the speaker-dependent LPCNet to convert acoustic features into wave. We also use the MOS score to evaluate the performance of overall quality samples and share the same rating rules as in above experiments. We find 15 native Shanghainese speakers and 15 native Cantonese speakers to implement the subjective test.

Results on Low-Resource Languages Table 4 and Table 5 show the TTS performance of our proposed method on low-resource Shanghainese and Cantonese Corpuses. We investigate the MOS on fine-tuning with 15 min and 1 h datasets. It is obvious that the pre-training on rich-resource Mandarin benefits the building of low-resource Chinese Dialects TTS-AMs. For example, when conducting experiment on 15 min dataset, the mean MOS score of Shanghainese increases from 2.97 to 3.30 and the mean MOS score of Cantonese increases

Table 4. The mean MOS on Shanghainese.

Fine-tuning Data Size	Without Pre-training	Pre-trained Model	Ground Truth
15 min	2.97	**3.30**	4.30
1 h	3.39	**3.51**	

Table 5. The mean MOS on Cantonese.

Fine-tuning Data Size	Without Pre-training	Pre-trained Model	Ground Truth
15 min	2.66	**3.09**	4.53
1 h	3.54	**3.99**	

from 2.66 to 3.09. With the help of proposed technique, we can generate more intelligible and natural speech with the same amount of low-resource data.

5 Conclusions

In this paper, a novel semi-supervised pre-training technique framework that extends the training material of TTS model by using ASR based data augmentation method is proposed to improve the performance of speech synthesis. We first build a frame-wise phoneme classification network on the ASR dataset and extract the semi-supervised <linguistic features, audio> paired data from large-scale speech corpora. After that, the semi-supervised <linguistic features, audio> pairs is used to pre-train the NN based TTS acoustic model. Finally, we fine-tune the model with a small amount of available paired data. Experimental results show that our proposed framework can benefits the building of low-resource TTS system by implementing semi-supervised pre-training technique. It enables the TTS model to generate more intelligible and natural speech with the same amount of paired training data.

References

1. Wang, Y., et al.: Tacotron: towards end-to-end speech synthesis. arXiv preprint arXiv:1703.10135 (2017)
2. Shen, J., et al.: Natural TTS synthesis by conditioning WaveNet on MEL spectrogram predictions. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4779–4783 (2018)
3. Ren, Y., et al.: FastSpeech: fast, robust and controllable text to speech. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
4. Ren, Y., et al.: FastSpeech 2: fast and high-quality end-to-end text to speech. arXiv preprint arXiv:2006.04558 (2020)
5. Chen, M., et al.: AdaSpeech: adaptive text to speech for custom voice. arXiv preprint arXiv:2103.00993 (2021)
6. Tjandra, A., Sakti, S., Nakamura, S.: Listening while speaking: speech chain by deep learning. In: IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 301–308 (2017)
7. Xu, J., et al.: LRSpeech: extremely low-resource speech synthesis and recognition. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 2802–2812 (2020)
8. Ren, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., Liu, T.-Y.: Almost unsupervised text to speech and automatic speech recognition. In: International Conference on Machine Learning, pp. 5410–5419. PMLR (2019)
9. Azizah, K., Adriani, M., Jatmiko, W.: Hierarchical transfer learning for multilingual, multi-speaker, and style transfer DNN-based TTS on low-resource languages. IEEE Access **8**, 179 798–179 812 (2020)
10. de Korte, M., Kim, J., Klabbers, E.: Efficient neural speech synthesis for low-resource languages through multilingual modeling. arXiv preprint arXiv:2008.09659 (2020)

11. Zhang, W., Yang, H., Bu, X., Wang, L.: Deep learning for Mandarin-Tibetan cross-lingual speech synthesis. IEEE Access **7**, 167 884–167 894 (2019)
12. Nekvinda, T., Dušek, O.: One model, many languages: meta-learning for multilingual text-to-speech. arXiv preprint arXiv:2008.00768 (2020)
13. Tu, T., Chen, Y.-J., Yeh, C.-C., Lee, H.-Y.: End-to-end text-to-speech for low-resource languages by cross-lingual transfer learning. arXiv preprint arXiv:1904.06508 (2019)
14. Hemati, H., Borth, D.: Using IPA-based Tacotron for data efficient cross-lingual speaker adaptation and pronunciation enhancement. arXiv preprint arXiv:2011.06392 (2020)
15. Mohammadi, S.H., Kain, A.: An overview of voice conversion systems. Speech Commun. **88**, 65–82 (2017)
16. Karlapati, S., Moinet, A., Joly, A., Klimkov, V., Sáez-Trigueros, D., Drugman, T.: CopyCat: many-to-many fine-grained prosody transfer for neural text-to-speech. arXiv preprint arXiv:2004.14617 (2020)
17. Huybrechts, G., Merritt, T., Comini, G., Perz, B., Shah, R., Lorenzo-Trueba, J.: Low-resource expressive text-to-speech using data augmentation. In: ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6593–6597 (2021)
18. Fang, W., Chung, Y.-A., Glass, J.: Towards transfer learning for end-to-end speech synthesis from deep pre-trained language models. arXiv preprint arXiv:1906.07307 (2019)
19. Devlin, J., Cheng, M.-W., Kenton, L., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT, pp. 4171–4186 (2019)
20. Chung, Y.-A., Wang, Y., Hsu, W.-N., Zhang, Y., Skerry-Ryan, R.: 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), pp. 6940–6944 (2019)
21. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
22. Zhang, H., Lin, Y.: Unsupervised learning for sequence-to-sequence text-to-speech for low-resource languages. arXiv preprint arXiv:2008.04549 (2020)
23. Ni, J., et al.: Unsupervised text-to-speech synthesis by unsupervised automatic speech recognition. arXiv preprint arXiv:2203.15796 (2022)
24. Valin, J.-M., Skoglund, J.: LPCNet: improving neural speech synthesis through linear prediction. In: 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5891–5895 (2019)
25. Skoglund, J., Valin, J.-M.: Improving Opus low bit rate quality with neural speech synthesis. arXiv preprint arXiv:1905.04628 (2019)
26. Valin, J.-M., Skoglund, J.: A real-time wideband neural vocoder at 1.6 kb/s using LPCNet. arXiv preprint arXiv:1903.12087 (2019)
27. Zhang, S., Lei, M., Yan, Z., Dai, L.: Deep-FSMN for large vocabulary continuous speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5869–5873 (2018)
28. Zhang, S., Lei, M., Yan, Z.: Automatic spelling correction with transformer for CTC-based end-to-end speech recognition. arXiv preprint arXiv:1904.10045 (2019)
29. Povey, D., et al.: The Kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (2011)