INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

GenerTTS: Pronunciation Disentanglement for Timbre and Style Generalization in Cross-Lingual Text-to-Speech

Yahuan Cong, Haoyu Zhang, Haopeng Lin, Shichao Liu, Chunfeng Wang, Yi Ren, Xiang Yin, Zejun Ma

ByteDance

{congyahuan, zhanghaoyu.aries, linhaopeng, liushichao, wangchunfeng, ren.yi, yinxiang.stephen, mazejun}@bytedance.com

Abstract

Cross-lingual timbre and style generalizable text-to-speech (TTS) aims to synthesize speech with a specific reference timbre or style that is never trained in the target language. It encounters the following challenges: 1) timbre and pronunciation are correlated since multilingual speech of a specific speaker is usually hard to obtain; 2) style and pronunciation are mixed because the speech style contains language-agnostic and language-specific parts. To address these challenges, we propose GenerTTS, which mainly includes the following works: 1) we elaborately design a HuBERT-based information bottleneck to disentangle timbre and pronunciation/style; 2) we minimize the mutual information between style and language to discard the language-specific information in the style embedding. The experiments indicate that GenerTTS outperforms baseline systems in terms of style similarity and pronunciation accuracy, and enables cross-lingual timbre and style generalization[1].

Index Terms: Cross-lingual text-to-speech, self-supervised learning, style transfer

1. Introduction

Cross-lingual timbre and style generalization text-to-speech (TTS) generates speech with any specific speaker and style that are unseen for the target language in training. This technique is important and useful for several scenarios and applications: 1) training a multi-lingual expressive TTS while the timbre and speech style can not cover all languages in training dataset, especially for some low-resource languages: we can transfer style and timbre learned in rich-resource languages to the target low resource language. 2) Automatic dubbing, which aims to replace all speech contained in a video with that in a different language, matching the original timbre, and rhythm [1, 2]: we can synthesize target language speech with the same style and timbre using very limited training data via cross-lingual timbre and style generalization.

The main challenge of cross-lingual timbre and style generalizable TTS is the decoupling of pronunciation, timbre and style from each other. To be specific, 1) timbre and pronunciation are correlated since multilingual speech of a specific speaker is usually hard to obtain in the training dataset; 2) style and pronunciation are usually strongly mixed since the speech style contains language-agnostic and language-specific parts and language is strongly related to pronunciation.

One way to address the challenge is domain adversarial training. Zhang et al. [3] applied gradient reversal layer (GRL) and residual encoder to reduce the timbre on text encoder. However, training a network with GRL is known to be unstable and sensitive to the hyper-parameter setting [4]. Another way is data augmentation: Sun et al. [5] proposed a system with augmented data generated by voice conversion. However, the data construction process is relatively complicated, and using the generated speech as the ground truth data for the training process will lead to a decrease in the quality of the synthesized speech.

Recently, many useful speech representations have been proposed, such as Phonetic PosteriorGrams (PPGs) [6], ASR bottleneck features (ASR-BNFs) [7, 8] and self-supervised learning (SSL)-based features (wav2vec 2.0 [9] and HuBERT [10]). The most important feature of these representations is that they can disentangle the speech into pronunciation, timbre and other components, which can be acted as the information bottleneck. Taking HuBERT representation as an example, we conduct some analyses on HuBERT (see Section 4.3) and find that it is proficient at preserving style and pronunciation information while removing timbre information with an appropriate channel size and chosen layer.

Motivated by this observation, we propose a two-step method that effectively separates timbre, style, and pronunciation from each other: 1) to disentangle timbre from style and pronunciation, we apply HuBERT as the bottleneck feature in our TTS model, ensuring pronunciation robustness and speaker similarity in cross-lingual scenes; 2) to disentangle style from pronunciation, based on our self-supervised presentation-based structure, we propose a new style adaptor that models fine-grained style and removes language-specific characteristics by introducing mutual information (MI) minimization constraint. Since it can generalize timbre and style in cross-lingual TTS, we call it GenerTTS.

Experimental results demonstrate that GenerTTS performs well against baseline models for cross-lingual timbre and style generalization TTS in terms of style similarity and pronunciation accuracy.

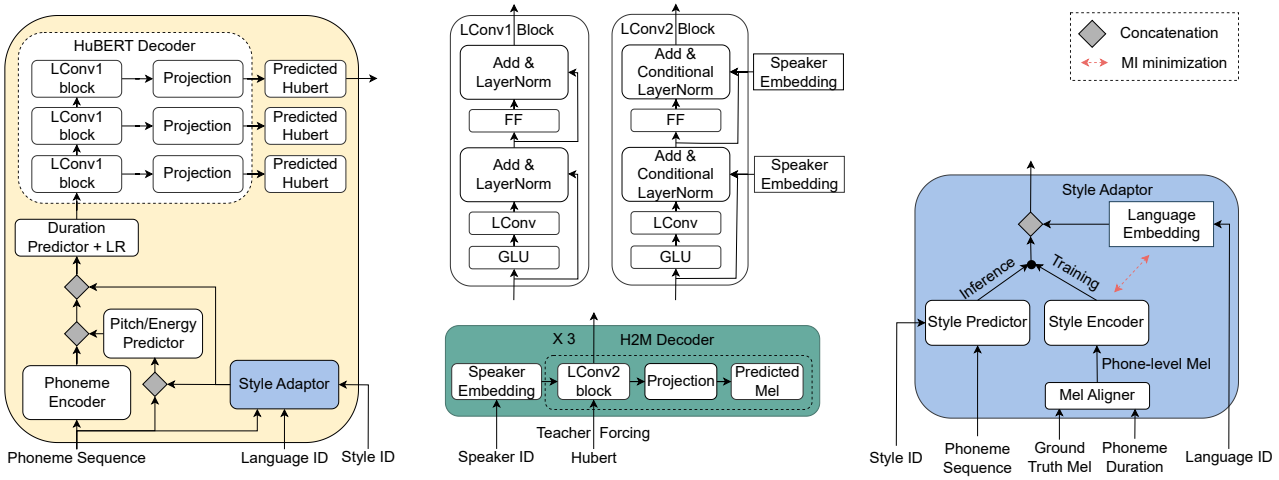
2. Background

Cross-Lingual Style Generalization: The work of Shang et al. [11] presented a method to deal with the cross-lingual transfer of timbre and style incorporating a fine-grained encoder and gradient reversal modules. However, this method does not disentangle language and style, which will lead to non-native pronunciation for cross-lingual style transfer. For example, the Chinglish phenomenon occurs when the Chinese style is transferred to the English target language.

Self-Supervised Speech Representation: Choi et al. [12] proposed a neural analysis and synthesis model by using multilingual self-supervised learning (SSL) wav2vec 2.0 features as parts of their bottleneck representations, achieving the state-of-the-art in voice conversion and successfully performing multi-

[1]The audio samples are available at https://bytecong.github.io/GenerTTS/

10.21437/Interspeech.2023-1727



(a) The diagram of P2H encoder (b) The diagram of H2M encoder and LConv block (c) The diagram of style adaptor

Figure 1: *The overall architecture for the proposed system.*

lingual voice conversion. Du et al. [13] improved the performance of TTS by utilizing a self-supervised VQ acoustic feature instead of traditional mel-spectrogram. Besides, SSL HuBERT [10, 14, 15] had better generation results than wav2vec 2.0 and CPC [16], and outperformed VQ-VAE [17] in both speech synthesis and voice conversion. Compared with other supervised bottleneck representations, SSL representation can be trained by plenty of unlabelled speech data.

3. Proposed Approach

In this section, we first describe the HuBERT representation. Then we introduce the overall design of the self-supervised representation-based TTS structure proposed in this paper. Finally, we introduce the key techniques to address challenges in cross-lingual style and timbre generalization.

3.1. Self-supervised acoustic feature: HuBERT

HuBERT is a self-supervised representation learning model, which shows the potential to disentangle timbres and other features [16, 18, 9, 10]. It uses "acoustic unit discovery system" to generate pseudo labels as the target of iterative training. Besides, a masking strategy similar to BERT [19] is applied in the pretraining to reduce prediction error and help learn representations of long-range temporal relationships. There are three iterations for HuBERT pertaining. K-means on MFCC is the training target for the first iteration, and the output of the trained model can be expected as a better representation than MFCCs. Then, for the second and third iterations, the K-means on the output of the middle layer for the previous iteration are used as the training target of the current iteration. Through such three iterations, a more refined and better continuous representation of pseudo-label can be obtained.

Previous research further discretizes this continuous embedding and applies it to generation tasks. However, discretization leads to the loss of prosodic information, which we expect more to be preserved in Hubert. Therefore, we use continuous embedding and verify that continuous embedding shows great performance in removing timbre information while retaining pronunciation and style. We will show our experiments and analysis in Section 4.3. Given the aforementioned properties of HuBERT, we are motivated to utilize it as the bottleneck feature in our GenerTTS system to disentangle timbre and pronunciation, as well as timbre and style.

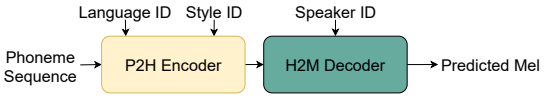


Figure 2: *The overall pipeline of our system.*

3.2. SSL-based TTS system

Our GenerTTS consists of a P2H (phoneme sequence to HuBERT) encoder and an H2M (HuBERT to Mel-spectrogram) decoder, as shown in Figure 2. P2H as text encoder is used to predict HuBERT embedding from the input phoneme sequence. H2M is designed for timbre adaptation, which refers to taking HuBERT as input and generating matching Mel-spectrograms according to different speaker embedding conditions. Waveform is synthesized from predicted Mel-spectrogram by a neural vocoder, which in our experiment is Melgan [20].

3.2.1. P2H encoder

Our P2H encoder provides style and pronunciation information. As shown in Figure 1(a), the structure of P2H consists of a phoneme encoder, a pitch/energy predictor, a duration predictor with length regulator (LR), and a HuBERT decoder. The P2H also includes a style adaptor for modeling fine-grained style, which is described in Section 3.2.3. Our P2H is based on parallel tacotron [21], which is a highly parallelizable neural TTS model with self-attention and lightweight convolutions. In the P2H encoder, a phoneme encoder learns latent representation from phoneme sequence, and then the representation is fed to a duration predictor to predict phoneme duration. And according to the duration information, the LR upsamples the phoneme encoder output to the length of the target frame sequence. Furthermore, similar to the parallel tacotron, an iterative loss is also used for the predicted output in our P2H encoder.

Innovatively, we propose a decoder to predict the target HuBERT embedding from the upsampled phoneme encoder output, instead of directly predicting Mel-spectrogram. Our HuBERT decoder contains three lightweight convolution blocks

(LConv1 blocks) with a fully-connected layer. The LConv1 block consists of a gated linear unit (GLU), a lightweight convolution (LConv), a feedforward (FF) layer, and two residual connections with layer normalization as shown in the upper part of Figure 1(b), which follows [21] [22].

In addition, to increase the stability of our proposed model and explicitly model prosody information, pitch predictor and energy predictor are added to our proposed model as in [23]. Specifically, we extract pitch and energy from speech waveform and map them to phoneme-level features depending on phoneme duration. We take them as conditional inputs in training and use predicted values in inference. Pitch and energy predictors are jointly trained with the proposed model.

3.2.2. H2M decoder

Our H2M decoder, shown in the lower part of Figure 1(b), predicts Mel-spectrogram from HuBERT and supports timbre information for synthesized speech. The network includes a 3-layer LConv2 Block with a fully-connected layer, and the target is 80-dimensional Mel-spectrogram. For high timbre adaptation quality, LConv2 Block is conditioned on speaker embedding by replacing the layer normalization of LConv1 Block with conditional layer normalization, which has been verified efficiency for timbre adaptation in [24]. Moreover, teacher forcing is used for H2M training, by making the ground truth HuBERT embedding as input to the H2M decoder during training and using the predicted HuBERT during inference. The iterative loss is also used between the predicted spectrogram and the ground truth Mel-spectrogram.

3.2.3. Cross-lingual style adaptor

As shown in Figure 1(c), our style adaptor mainly consists of four parts: Mel-spectrogram aligner (Mel aligner), style encoder, style predictor, and language embedding with mutual information constraint.

We utilize a style encoder to model fine-grained style embedding from phoneme-level spectrogram and concatenate it with the output of the phoneme encoder. Mel aligner is used to map the frame-level Mel-spectrogram into a phoneme-level spectrogram according to phoneme duration information. Since variance features such as pitch and energy are closely related to style, we add style embedding to the pitch and energy predictor.

For cross-lingual style transfer, the spoken content and even the language of the Mel-spectrogram referenced during inference are inconsistent with the input in the style encoder. Therefore, a style predictor is utilized to address these inconsistencies. The style predictor adopts text-related information to predict the fine-grained style conditioned on the style ID. In training, the output of the style encoder is used as the prediction target of the style predictor after the stop gradient. For inference, we use the style predictor to predict style depending on the target style ID and input phoneme sequence. The networks of style encoder and style predictor are the same with [11].

In addition, for cross-lingual style transfer, we want the synthesized speech to have a high style similarity while still having the native pronunciation of the target language. In order to improve the pronunciation nativeness for style transfer, we further decouple the language-specific information from the style embedding. We model the language embedding from language ID and minimize the MI between language embedding and style embedding. Due to the difficulty of estimating MI in high-dimensional space, we minimize the upper limit of MI measured by the variational contrastive log-ratio upper bound (vCLUB) [25].

4. Experiments and Results

4.1. Experimental setup

We evaluate our proposed model using 94.6 hours of Mandarin (zh-CN) from 4 speakers and 43.0 hours of English (en-US) from 5 speakers. The sampling rate of those datasets is 24k.

For the HuBERT embedding in our experiments, we train HuBERT models using 2000 hours of proprietary datasets in Chinese and English. The sampling rate of this dataset is 16k. Our HuBERT model based on fairseq [26] consists of 7 CNN layers and 12 layers transformers. Some former researches show that outputs from different layers of the SSL model may contain different information [27, 12, 10]. Empirically, we used the continuous embedding output by the 9th layer as the target of P2H of our proposed system due to retaining more style information and pronunciation information.

To validate the performance of our proposed model, two baselines are implemented:

- **Para**: Parallel tacotron. The model details follow [21]. And for comparison, we add pitch and energy to the system, which is the same as our proposed system.
- **M3**: A multi-speaker multi-style multi-language speech synthesis baseline system [11]. We reproduce the baseline system based on the basic structure of Para.
- **Ours**: Our proposed cross-lingual speech synthesis system, which is based on HuBERT and style adaptor.

4.2. Result

4.2.1. System evaluation

For cross-lingual timbre and style generalization, we evaluate the cross-lingual synthetic speech generated by transferring one timbre and two styles from Chinese to English. Both style and timbre are unseen in the target language during training. Moreover, the timbre comes from female voice data, and the two styles are female customer service and male novel narration. We also evaluate the within-lingual synthetic speech generated by this timbre and these two styles in Chinese.

For objective evaluation, we utilize cosine similarity to measure speaker similarity, and employ Character Error Rate (CER) for zh-CN and Word Error Rate (WER) for en-US to evaluate pronunciation accuracy. For subjective evaluation, we conduct a mean opinion score (MOS) experiment on a 5-point scale (5:excellent, 4:good, 3:fair, 2:poor, 1:bad) to evaluate style similarity. Moreover, pronunciation nativeness MOS is evaluated on a 5-point scale, which is only available for cross-lingual. Table 1 shows the experimental results.

Compared with the two baseline systems, our system has obvious advantages in pronunciation accuracy for cross-lingual speech synthesis, which is reflected in the pronunciation of native and ER (CER/ WER). And our system has also shown improvement in style similarity, especially compared with Para. It should also be noted that the style similarity in cross-lingual is higher than in within-lingual. This is because, in cross-lingual style transfer, the target style and the synthesized speech are in different languages, which makes the evaluators prioritize global style and pay less attention to style details. However, the style MOS criteria are the same for both within-lingual and cross-lingual evaluations.

Table 1: *Evaluation score of speaker similarity, style similarity, pronunciation nativeness MOS and ER (CER/WER). W indicates within-lingual and C indicates cross-lingual in the table.*

Model	Speaker Similarity	Style Similarity	Pronunciation Nativeness	ER (%)
Para	**0.922** (W:0.967; C:0.877)	3.43 (W:3.32; C:3.54)	3.35	13.42 (W:4.01; C:22.82)
M3	0.920 (W:0.966; C:0.875)	3.71 (W:3.35; C:4.07)	3.71	12.12 (W:3.96; C:20.28)
Ours	0.921 (W:0.964; C:0.878)	**3.99** (W:3.90; C:4.07)	**4.13**	**9.97** (W: 4.02; C:15.93)
Ours w/o MI	/	3.92 (W:3.90; C:3.95)	3.71	12.38 (W:3.58; C:21.19)
Ours w/o Adaptor	/	3.81 (W:3.57; C:4.06)	3.63	12.54 (W:3.87; C:21.20)

Table 2: *HuBERT-based voice conversion, where F indicates female speaker and M indicates male.*

Model	Style Similarity	Cosine Similarity
M2M	3.98	0.948
F2M	3.86	0.946
F2F	4.06	0.970
M2F	4.22	0.946

4.2.2. Ablation study

We further conduct an ablation study to validate the components in GenerTTS, which include SSL HuBERT and style adaptor with MI constraint on the linguistic information of style features, as shown in Table 1. The MI removed system has a sharp decline in pronunciation nativeness MOS and increases in cross-lingual Pronunciation ER. Although MI leads to an increase in pronunciation ER for within-lingual, the results are still basically the same as the baseline model. After removing the style adaptor, the performance deteriorates further, but higher than Para. These results support that the proposed style adaptor and MI are crucial for the disentanglement of language and style, and enable style transfer across languages while maintaining native pronunciation.

In comparison to the Para (ours w/o HuBERT and style adaptor), the system with HuBERT (ours w/o style adaptor) demonstrated higher style similarity and pronunciation nativeness MOS score, as well as lower pronunciation ER in within-lingual and cross-lingual synthesis. This shows that HuBERT has played a significant role in the decoupling of timbre and pronunciation/style. These results demonstrate the effectiveness of each component in our proposed model.

4.3. HuBERT analysis

In this section, we separately validate the pronunciation, timbre, and style information in HuBERT. We first evaluate the pronunciation information of different layers. Same with the original HuBERT paper, we apply k-means features of all 13 layers of the HuBERT model into 200 clusters, 500 clusters, and 1000 clusters. Cluster Purity and Phone-Normalized Mutual Information (PNMI) are also calculated as shown in Figure 3. A higher indicator means the embedding for clustering is more pronunciation-related. Hence we use the embedding from the 9th layer for its relatively higher numbers.

We then examine the timbre and style information in HuBERT by voice conversion. We train an independent H2M decoder to serve as voice conversion. HuBERT embeddings are extracted from data of two unseen Chinese speakers and used as input to the voice conversion.

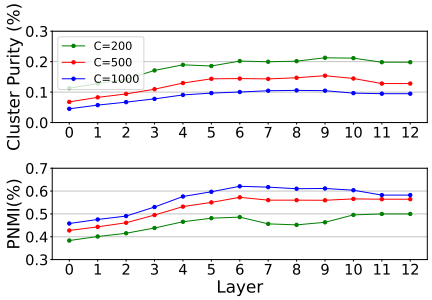


Figure 3: *Quality of cluster assignments by running k-means on HuBERT embedding from different layers.*

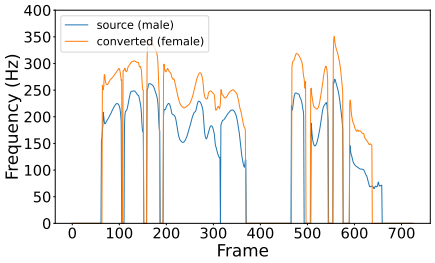


Figure 4: *The pitch of the source audio and converted audio.*

We calculate the cosine similarity of speaker verification embedding between generated audio and target timbre audio. As shown in Table 2, the H2M decoder can generate audio with high speaker cosine similarity, which indicates that little timbre-related information exists in HuBERT embedding. In addition, the higher style similarity means that style information is restored in HuBERT. We also extract F0 from the input and output voices by straight [28] and they show similar variation trends in Figure 4. Those results indicate that HuBERT embedding might be an effective bottleneck feature to remove timbre while retaining other factors, such as style, and pronunciation.

5. Conclusion

In this paper, we propose GenerTTS to address cross-lingual timbre and style generalizable speech synthesis with a specific timbre or style that is never trained in the target language. Experimental results show that GenerTTS outperforms two baseline systems in terms of pronunciation accuracy and style similarity. In addition, ablation experiments show that the HuBERT-based TTS system can improve the pronunciation accuracy of cross-lingual TTS. And our proposed MI can reduce language-specific information in style and improve style similarity and pronunciation nativeness for cross-lingual style transfer.

6. References

[1] M. Federico, R. Enyedi, R. Barra-Chicote, R. Giri, U. Isik, A. Krishnaswamy, and H. Sawaf, "From speech-to-speech translation to automatic dubbing," *arXiv preprint arXiv:2001.06785*, 2020.

[2] G. Cong, L. Li, Y. Qi, Z. Zha, Q. Wu, W. Wang, B. Jiang, M.-H. Yang, and Q. Huang, "Learning to dub movies via hierarchical prosody models," *arXiv preprint arXiv:2212.04054*, 2022.

[3] Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Z. Chen, R. Skerry-Ryan, Y. Jia, A. Rosenberg, and B. Ramabhadran, "Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning," 2019.

[4] S. H. Mun, M. H. Han, M. Kim, D. Lee, and N. S. Kim, "Disentangled speaker representation learning via mutual information minimization," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2022, pp. 89–96.

[5] Q. Sun and K. Nagamatsu, "Building multi lingual tts using cross lingual voice conversion," *arXiv preprint arXiv:2012.14039*, 2020.

[6] Y. Cao, S. Liu, X. Wu, S. Kang, P. Liu, Z. Wu, X. Liu, D. Su, D. Yu, and H. Meng, "Code-switched speech synthesis using bilingual phonetic posteriorgram with only monolingual corpora," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7619–7623.

[7] X. Zhu, Y. Lei, K. Song, Y. Zhang, T. Li, and L. Xie, "Multi-speaker expressive speech synthesis via multiple factors decoupling," *arXiv preprint arXiv:2211.10568*, 2022.

[8] D. Dai, Y. Chen, L. Chen, M. Tu, L. Liu, R. Xia, Q. Tian, Y. Wang, and Y. Wang, "Cloning one's voice using very limited data in the wild," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8322–8326.

[9] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[10] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[11] Z. Shang, Z. Huang, H. Zhang, P. Zhang, and Y. Yan, "Incorporating cross-speaker style transfer for multi-language text-to-speech." in *Interspeech*, 2021, pp. 1619–1623.

[12] H.-S. Choi, J. Lee, W. Kim, J. Lee, H. Heo, and K. Lee, "Neural analysis and synthesis: Reconstructing speech from self-supervised representations," *Advances in Neural Information Processing Systems*, vol. 34, pp. 16 251–16 265, 2021.

[13] C. Du, Y. Guo, X. Chen, and K. Yu, "Vqtts: high-fidelity text-to-speech synthesis with self-supervised vq acoustic feature," *arXiv preprint arXiv:2204.00768*, 2022.

[14] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed *et al.*, "On generative spoken language modeling from raw audio," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1336–1354, 2021.

[15] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed, and E. Dupoux, "Speech resynthesis from discrete disentangled self-supervised representations," *arXiv preprint arXiv:2104.00355*, 2021.

[16] A. Van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv e-prints*, pp. arXiv–1807, 2018.

[17] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[18] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[20] K. Kumar, R. Kumar, T. De Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," *Advances in neural information processing systems*, vol. 32, 2019.

[21] I. Elias, H. Zen, J. Shen, Y. Zhang, Y. Jia, R. J. Weiss, and Y. Wu, "Parallel tacotron: Non-autoregressive and controllable tts," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5709–5713.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[23] H. Zhan, X. Yu, H. Zhang, Y. Zhang, and Y. Lin, "Exploring timbre disentanglement in non-autoregressive cross-lingual text-to-speech," *arXiv preprint arXiv:2110.07192*, 2021.

[24] M. Chen, X. Tan, B. Li, Y. Liu, T. Qin, S. Zhao, and T.-Y. Liu, "Adaspeech: Adaptive text to speech for custom voice," *arXiv preprint arXiv:2103.00993*, 2021.

[25] P. Cheng, W. Hao, S. Dai, J. Liu, Z. Gan, and L. Carin, "Club: A contrastive log-ratio upper bound of mutual information," in *International conference on machine learning*. PMLR, 2020, pp. 1779–1788.

[26] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," *arXiv preprint arXiv:1904.01038*, 2019.

[27] J. Shah, Y. K. Singla, C. Chen, and R. R. Shah, "What all do audio transformer models hear? probing acoustic representations for language delivery and its structure," *arXiv preprint arXiv:2101.00387*, 2021.

[28] H. Kawahara, "Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds," *Acoustical science and technology*, vol. 27, no. 6, pp. 349–353, 2006.