INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

DSE-TTS: Dual Speaker Embedding for Cross-Lingual Text-to-Speech

Sen Liu, Yiwei Guo, Chenpeng Du, Xie Chen , Kai Yu†

MoE Key Lab of Artificial Intelligence, AI Institute
X-LANCE Lab, Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China
{sen.liu, cantabile_kwok, duchenpeng, chenxie95, kai.yu}@sjtu.edu.cn

Abstract

Although high-fidelity speech can be obtained for intralingual speech synthesis, cross-lingual text-to-speech (CTTS) is still far from satisfactory as it is difficult to accurately retain the speaker timbres (i.e. speaker similarity) and eliminate the accents from their first language (i.e. nativeness). In this paper, we demonstrated that vector-quantized (VQ) acoustic feature contains less speaker information than mel-spectrogram. Based on this finding, we propose a novel dual speaker embedding TTS (DSE-TTS) framework for CTTS with authentic speaking style. Here, one embedding is fed to the acoustic model to learn the linguistic speaking style, while the other one is integrated into the vocoder to mimic the target speaker's timbre. Experiments show that by combining both embeddings, DSE-TTS significantly outperforms the state-of-the-art SANE-TTS in cross-lingual synthesis, especially in terms of nativeness.

Index Terms: cross-lingual text-to-speech, dual speaker embedding, vector-quantized acoustic feature

1. Introduction

Recent neural text-to-speech (TTS) models [1, 2, 3, 4, 5] have made great strides in synthesizing speech with high fidelity, rich prosody and remarkable speaker similarity. Nevertheless, in multilingual TTS (MTTS) scenarios, cross-lingual synthesis is still far from satisfactory as it is difficult to accurately retain the speaker's timbres and eliminate the accents from their first language. More specifically, cross-language synthesis is difficult to acquire nativeness in non-native languages while maintaining speaker similarity, while nativeness refers to the closeness of speech to the native language. Efforts have been made to mitigate the degradation in cross-lingual performance resulting from this entanglement. [6, 7] incorporated adversarial domain training, allowing them to transfer distinct voices across languages. [8] proposed to use mutual information minimization to maintain speaker consistency in cross-lingual synthesis. [9] implemented multi-task learning and joint training with a speaker classifier to enhance the overall similarity of speakers. More recently, SANE-TTS [10] presented an end-to-end multilingual TTS model based on VITS [11]. This model employed a speaker regularization loss to encourage the model to learn speaker representation independent of its language, ensuring accurate duration predictions in cross-lingual synthesis.

However, these studies typically rely on the mel-spectrogram as an acoustic feature, which is highly correlated along both time and frequency axes and contains rich speaker-dependent information, making it still challenging to disentangle correlated factors. Seeking another acoustic feature that contains less speaker identity might be crucial. Recent advancements in speech-based self-supervised learning (SBSSL) [12, 13, 14, 15, 16, 17] have enabled some TTS models to use discrete vector-quantized (VQ) speech representations as an acoustic feature, replacing the traditional mel-spectrogram for prediction. SBSSL models take raw waveform as input, which is only correlated along the time axis. As a result, the quantized output has a coarser granularity of speech features than the mel-spectrogram. This results in lower reconstruction difficulties of VQ features and potentially less speaker-dependent information. For example, [18] leverages self-supervised VQ acoustic features as an alternative to the mel-spectrogram. The VQ features are generated by an acoustic model named txt2vec and then used for waveform reconstruction by a vocoder, vec2wav. By replacing the mel-spectrogram regression task with a VQ feature classification task, [18] achieves highly competitive naturalness among publicly available TTS systems.

In this paper, by analysizing the performance of the Multi-speaker version of [18], we found that VQ acoustic feature contains little speaker-specific information. Subsequently, we conducted speaker classification experiment using different acoustic features as shown in section 3. Results show that self-supervised VQ features extracted from wav2vec 2.0 [13] contain much less speaker-specific information than mel-spectrogram and other candidates. Hence VQ features are easier to decouple timbre and linguistic information than traditional mel-spectrogram. Based on this finding, we propose DSE-TTS, a TTS system with dual speaker embedding for cross-lingual TTS that enables the system to model linguistic speaking style and speaker timbre separately. The dual speaker embedding operates by controlling different speech aspects in the acoustic model and vocoder separately in the inference stage. Experiments show that by combining both embeddings, DSE-TTS outperforms the state-of-the-art SANE-TTS in both intralingual and cross-lingual synthesis, especially in terms of nativeness. We will elaborate on the proposed methods and detailed experimental results in later sections.

2. Dual Speaker Embedding TTS

The *dual speaker embedding TTS* (DSE-TTS) is introduced in detail in this section, input representations, acoustic model architecture, and the proposed dual speaker embedding. The overall framework is shown in Figure 1.

2.1. Input representations

Following [6] and [19], the input text is initially normalized and converted to International Phonetic Alphabet (IPA) phonemes using the phonemizer [20] toolkit. To facilitate alignment between the text and speech, we preserve the tones and stresses of different languages in our input sequences. We also use

Kai Yu† is the corresponding author.

10.21437/Interspeech.2023-363

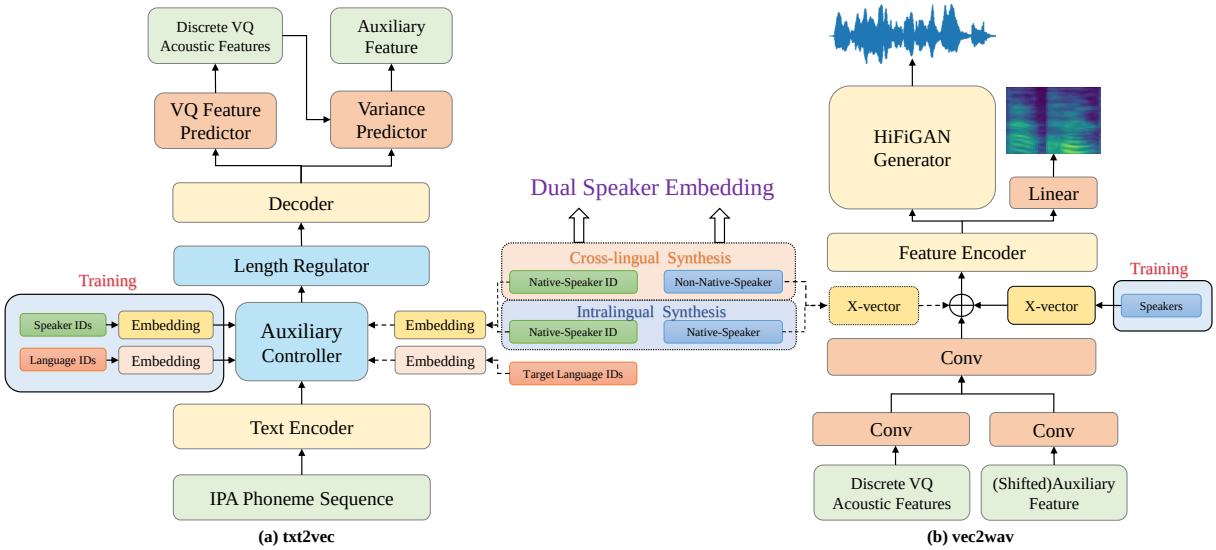

Figure 1: *The overall architecture of DSE-TTS consists of an acoustic model, txt2vec, and a vocoder, vec2wav. Dual Speaking Embedding enables DSE-TTS to model linguistic characteristics and speaker timbre separately, helping to retain the speaker's timbre in synthesized speech and eliminating the accents of their native language in cross-lingual synthesis. Solid lines represent the training stage, while dashed lines represent the inference stage.*

shared punctuation tokens across languages, categorized into four groups based on pause length, denoted as 'sp1', 'sp2', 'sp3', and 'sp4'. Additionally, we use 'sil' as the starting and ending token for each sentence. Prior to feeding the input sequence to the text encoder in txt2vec, each phoneme (token) is assigned a 384-dimensional vector using an embedding table.

2.2. Model architecture

2.2.1. Self-supervised VQ features

In this study, we extract VQ acoustic features using a wav2vec 2.0 model with two quantized codebooks, each containing 320 codewords. The wav2vec 2.0 model was pre-trained on 10,000 hours of Mandarin data. It quantizes each input speech into multiple frames with a 20ms stride, and each frame can be represented by concatenating two 256-dimensional codewords from each codebook. All possible index combinations in our mixed-language dataset are about 28.8k. The objective is to accurately predict these index pairs in order to construct high-fidelity speech. For parallel inference, we replace the VQ feature predictor with a convolutional neural network instead of LSTM [21] in the original [18]. Furthermore, we predict the index of each codebook separately instead of their combinations, resulting in two 320-class classification problems. We choose wav2vec 2.0 as our VQ feature extractor because it provides a more robust speech representation with less speaker information compared to other VQ features. The rationale for this choice will be further explained in the experiment section.

2.2.2. Phone-level(PL) auxiliary labelling

Similar to [18], we utilize log pitch, energy, and probability of voice (POV)[22] as auxiliary features. To begin, we compute and normalize phone-level representations of our mixed-language dataset. Then, we apply k-means clustering to group these representations into 128 distinct classes, with the resulting clustered index serving as the auxiliary label for PL information. We employ the ground truth PL auxiliary labels on one side for training the multilingual auxiliary controller, while on the other side, they serve as a condition for subsequent duration modeling and acoustic feature generation.

2.3. Dual speaker embedding

In previous works on cross-lingual TTS, it is difficult to accurately retain the speaker's timbres and eliminate the accent from their first languages, resulting in unnatural synthesized speech. The main reason is typically rooted in the entanglement between speakers and languages, which is often manifested in the nature of traditional acoustic features like mel-spectrogram. However, our preliminary experiments found that self-supervised VQ features contain much less speaker identity compared to traditional acoustic features. We will show these results in section 3.3. As a result, in VQ-based TTS methods, it is unnecessary to use additional techniques for the disentanglement of speaker and language within the acoustic model. This allows the model to concentrate solely on modeling textual and linguistic characteristics while the task of controlling speaker timbre is delegated to the vocoder. Thus, the VQ-based TTS model naturally learns how to speak different languages in a native way with the timbre of a non-native speaker.

From this perspective, we develop a dual speaker embedding TTS (DSE-TTS) framework to improve the nativeness and speaker similarity in cross-lingual TTS scenarios. Two speaker embeddings are used in the TTS model, where one is fed into the acoustic model txt2vec and the other for the vocoder vec2wav. In the training stage, given the text and speech pair of a native speaker, the two speaker embeddings both correspond to the same speaker. Then in the synthesis stage, no matter the intralingual or cross-lingual case, the speaker embedding of a native speaker corresponding to the input language is chosen as

the input speaker embedding to txt2vec uses a native speaker in the language of input text. In contrast, the speaker embedding in vec2wav is set as the target speaker. Hence, in the cross-lingual case, it means that we choose a native speaker's embedding in txt2vec representing linguistic speaking style and the target speaker's embedding in vec2wav that controls the timbre. In this way, language-specific speaking style and speaker timbre are naturally separated by dual embeddings.

The diagram of DSE-TTS is shown in Figure 1. For the acoustic model txt2vec, we take speaker and language IDs as input. Speaker IDs are embedded in 256-dimensional vectors, which are then projected and added to the encoder output. We handle language IDs similarly to support various languages, which are embedded in 128-dimensional vectors. These two embeddings are used to learn the linguistic characteristics of different languages. For vec2wav, we use X-vector [23] as the speaker embedding to control the timbre, which is extracted from a pre-trained speaker recognition model. Besides, to bring the timbre closer to the target speaker while doing cross-lingual synthesis, we shift the distribution of the native speaker's pitches predicted by txt2vec to match the pitch of the target speaker. It can be formulated as follows:

$$P_{\text{tgt}} = \sigma_{\text{tgt}} \frac{P_{\text{ntv}} - \mu_{\text{ntv}}}{\sigma_{\text{ntv}}} + \mu_{\text{tgt}} \quad (1)$$

where the subscripts "tgt" and "ntv" stands for target and native speaker, respectively. μ and σ are the mean and standard deviation of the target or native speaker's pitch values in the training set. We perform this pitch shift before the auxiliary features are sent to vec2wav for synthesis.

3. Experiments and results

3.1. Dataset

Our dataset comprises four languages: Mandarin (ZH), English (EN), Spanish (ES), and German (DE). We obtained the data for German and Spanish from M_AILABS [24], while the data for English and Mandarin were sourced from LibriTTS [25] and Aishell3 [26], respectively. In reality, it may be hard for some languages to collect enough data. To imitate this scenario and test our method's language adaptive ability, we randomly selected a few hours of data from German and Spanish as low-resource languages. The total duration and number of speakers involved are listed in Table 1. During training, we resampled all speech to 24 kHz and used 5% of the utterances for the validation and test set. To extract ground truth phoneme duration, we employed MFA[1], which performs forced alignment using Kaldi [27].

Table 1: *Details of the training dataset.*

Language	EN	ZH	DE	ES
Hours	74	60	6	6
#Speakers	228	142	3	3

3.2. Experimental setup

We trained our models for 200 epochs on the txt2vec and 100 epochs on the vec2wav, using batch sizes of 16 and 8, respectively. The training process was performed separately on an NVIDIA 2080Ti GPU. We utilized a publicly available pre-trained wav2vec 2.0 model [2] for VQ acoustic feature extraction. Additionally, we adopted the data balance strategy proposed by [28], with the scaling factor set to 0.2. To evaluate the performance of our model, we used the recent MTTS model, SANE-TTS, as our baseline and replicated it using the official VITS[3] implementation. We trained the SANE-TTS model for 200 epochs using a batch size of 16 while keeping all other parameters consistent with those specified in the original paper.

3.3. Speaker-independent VQ acoustic feature

To investigate the relationship between different acoustic features and speakers, we first construct a speaker classification model to evaluate the classification accuracy of various features. We compared the mel-spectrogram, a widely used acoustic feature in TTS models, with four distinct VQ features extracted from open-sourced pre-trained models, including vq-wav2vec [12], wav2vec 2.0 [13], XLSR-53 [14] and Encodec [29]. Our classification model used an X-vector architecture augmented with two linear layers to predict speaker identities. We trained the model on the LibriTTS training set, which includes more than 2000 speakers. After training the model for 80 epochs, we analyze the classification accuracy of speaker identities on the test set. As shown in Figure 2, the mel-spectrogram contains sufficient information about the speaker's identity, resulting in a high accuracy rate for speaker classification. In contrast, the VQ features have significantly less speaker information, leading to a lower accuracy rate than the mel-spectrogram. Based on our experimental results, we chose wav2vec 2.0 as our acoustic feature because it has a relatively lower speaker identification performance, indicating that it contains less speaker-dependent information.

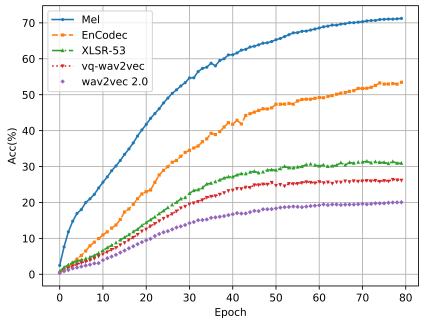

Figure 2: *Speaker classification accuracy with different acoustic features.*

3.4. Speech Synthesis Evaluation

We utilized subjective and objective measures to evaluate the quality of intralingual and cross-lingual synthesis. Our subjective measures include nativeness mean opinion score (NMOS) and similarity MOS (SMOS). NMOS is used to evaluate the nativeness of synthetic speech, while SMOS is used to assess the extent of speaker similarity. A higher NMOS score indicates that the synthesized speech is closer to the native language. MOS ratings were based on a 1-5 scale with 0.5-point increments and 95% confidence intervals. We synthesized 30 speech samples for each language using random texts from the test set and recruited multiple raters for evaluation. The raters included

[1] https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner

[2] https://github.com/TencentGameMate/chinese_speech_pretrain

[3] https://github.com/jaywalnut310/vits

Table 2: *WER, Speaker Similarity, Nativeness MOS and Similarity MOS results in cross-lingual synthesis, where SECS means Speaker Embedding Cosine Similarity and "w/o DSE" means without Dual Speaker Embedding (a.k.a., using the same target speaker's embedding in both the txt2vec and vec2wav modules during inference).*

Language	Model	EN Speaker				ZH Speaker			
		WER ↓	SECS ↑	NMOS ↑	SMOS ↑	WER ↓	SECS ↑	NMOS ↑	SMOS ↑
DE	SANE-TTS	29.6	0.44	3.77 ± 0.07	4.36 ± 0.08	29.6	0.57	3.79 ± 0.06	4.54 ± 0.08
	Ours w/o DSE	16.1	0.50	4.03 ± 0.06	4.45 ± 0.08	22.6	0.60	3.93 ± 0.05	4.60 ± 0.07
	Ours (DSE-TTS)	**14.9**	0.46	$\mathbf{4.19 \pm 0.07}$	4.40 ± 0.06	**15.5**	0.59	$\mathbf{4.11 \pm 0.06}$	4.54 ± 0.07
ES	SANE-TTS	17.9	0.46	4.03 ± 0.06	4.54 ± 0.07	21.4	0.57	3.93 ± 0.06	4.53 ± 0.08
	Ours w/o DSE	17.4	0.49	4.19 ± 0.05	4.59 ± 0.08	18.3	0.58	3.97 ± 0.06	4.57 ± 0.08
	Ours (DSE-TTS)	**13.5**	0.50	$\mathbf{4.47 \pm 0.06}$	4.50 ± 0.07	**16.0**	0.55	$\mathbf{4.26 \pm 0.07}$	4.52 ± 0.07

15 bilingual Mandarin and English speakers to assess the quality of English and Mandarin speech and 15 trilingual Mandarin-English-German and English-Mandarin-Spanish speakers to evaluate the synthesized speech in German and Spanish, respectively. For the objective metrics, we computed word error rate (WER), character error rate (CER), and speaker embedding cosine similarity (SECS) between the synthesized speech and the ground-truth speech. WER was used for Spanish, German, and English, while CER was used for Mandarin. We used pretrained ASR models, Whisper[30] for Spanish, German, and English, and a transformer[31] ASR model for Mandarin. For speaker similarity, we used an independently trained ResNet-based r-vector speaker verification model [32] and computed cosine similarity scores between 0 and 1. A larger score indicates better speaker similarity. To compare our proposed and baseline models, we synthesized 100 speech samples per language by randomly selecting sentences from the test set. Audio samples are available online[4].

Table 3: *Nativeness MOS and ASR results in intralingual synthesis, while WER is for German (DE), Spanish (ES), and English (EN), and CER is for Mandarin (ZH).*

Language	Model	NMOS ↑	WER(CER) ↓
DE	Ground truth	4.49 ± 0.07	6.4
	SANE-TTS	4.08 ± 0.08	16.4
	Ours (DSE-TTS)	$\mathbf{4.40 \pm 0.07}$	**7.8**
ES	Ground truth	4.69 ± 0.06	4.1
	SANE-TTS	4.30 ± 0.07	9.2
	Ours (DSE-TTS)	$\mathbf{4.56 \pm 0.05}$	**8.8**
EN	Ground truth	4.54 ± 0.05	4.2
	SANE-TTS	4.18 ± 0.06	5.6
	Ours (DSE-TTS)	$\mathbf{4.36 \pm 0.06}$	**5.3**
ZH	Ground truth	4.46 ± 0.06	6.8
	SANE-TTS	3.79 ± 0.07	10.6
	Ours (DSE-TTS)	$\mathbf{4.39 \pm 0.06}$	**7.9**

3.4.1. Intralingual synthesis

Table 3 shows the average NMOS and WER (CER) in intralingual evaluation. It is evident that DSE-TTS has achieved NMOS scores close to the ground truth and outperforms the baseline model on all metrics and across all languages. Specifically, DSE-TTS has attained an NMOS score above 4.3 for each language and achieved a lower WER (CER).

3.4.2. Cross-lingual synthesis

Table 2 presents the evaluation results of our cross-lingual synthesis. We observe that the results were consistent with those obtained in intralingual synthesis, as DSE-TTS outperformed SANE-TTS in terms of both NMOS and WER scores by a large margin. Specifically, in NMOS scores, raters preferred DSE-TTS against baseline by over 0.3 in all the speaker-language combinations. Moreover, the SMOS and SECS scores demonstrate that DSE-TTS maintains similar speaker characteristics to SANE-TTS. These findings suggest that DSE-TTS can synthesize high-quality German and Spanish speech in a non-native speaker's voice but with greater similarity to that of native speakers than the baseline model.

[4] https://goarsenal.github.io/DSE-TTS

3.4.3. Ablation study

We performed an ablation study to investigate the impact of dual speaker embedding (DSE) on the performance of our model. The results presented in Table 2 indicate a significant enhancement in the nativeness and decrease in WER of the synthesized speech after the integration of DSE. Our observations also suggest that the use of DSE resulted in a slight decrease in the speaker similarity scores in comparison to not using it. This may be attributed to the fact that different languages have unique linguistic speaking styles, and non-native speakers may sound slightly different when speaking a foreign language fluently. This is also evidence that DSE-TTS produces speech in a native way, though not trained with bilingual speakers.

4. Conclusions

In this paper, we propose DSE-TTS, a cross-lingual TTS model, which consists of a dual speaker embedding to model linguistic speaking style and speaker timbre separately. We first showed by a preliminary study that VQ features have fewer speaker-dependent features. Leveraging this finding, we improved our model with a novel dual speaker embedding, resulting in cross-lingual speech synthesis with high nativeness and a similar timbre to the target speaker. Our experiments demonstrated that DSE-TTS outperforms SANE-TTS in both intralingual and cross-lingual synthesis, particularly in terms of nativeness. We also verified the effectiveness of dual speaker embedding by an ablation study. In future work, we will focus on enhancing the quality of the synthesized speech in cross-lingual scenarios and expand our model into other languages.

5. Acknowledgements

This study was supported by Shanghai Municipal Science and Technology Major Project (No.2021SHZDZX0102) and the Key Research and Development Program of Jiangsu Province, China (Grant No.BE2022059-1).

6. References

[1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[2] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *Proc. International Conference on Learning Representations ICLR*. OpenReview.net, 2021. [Online]. Available: https://openreview.net/forum?id=piLPYqxtWuA

[3] C. Du and K. Yu, "Rich prosody diversity modelling with phone-level mixture density network," in *Proc. ISCA Interspeech*. ISCA, 2021, pp. 3136–3140.

[4] Y. Guo, C. Du, X. Chen, and K. Yu, "Emodiff: Intensity controllable emotional text-to-speech with soft-label guidance," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[5] B. Chen, C. Du, and K. Yu, "Neural fusion for voice cloning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1993–2001, 2022.

[6] Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Z. Chen, R. J. 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," in *Proc. ISCA Interspeech*. ISCA, 2019, pp. 2080–2084.

[7] D. Xin, Y. Saito, S. Takamichi, T. Koriyama, and H. Saruwatari, "Cross-lingual text-to-speech synthesis via domain adaptation and perceptual similarity regression in speaker space," in *Proc. ISCA Interspeech*. ISCA, 2020, pp. 2947–2951.

[8] D. Xin, T. Komatsu, S. Takamichi, and H. Saruwatari, "Disentangled speaker and language representations using mutual information minimization and domain adaptation for cross-lingual TTS," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6608–6612.

[9] J. Yang and L. He, "Cross-lingual text-to-speech using multi-task learning and speaker classifier joint training," *CoRR*, vol. abs/2201.08124, 2022.

[10] H. Cho, W. Jung, J. Lee, and S. H. Woo, "SANE-TTS: stable and natural end-to-end multilingual text-to-speech," in *Proc. ISCA Interspeech*. ISCA, 2022, pp. 1–5.

[11] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proc. International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, 2021, pp. 5530–5540.

[12] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in *Proc. International Conference on Learning Representations ICLR*. OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=rylwJxrYDS

[13] 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.

[14] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," in *Proc. ISCA Interspeech*, H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, and P. Motlícek, Eds. ISCA, 2021, pp. 2426–2430.

[15] W. Hsu, B. Bolte, Y. 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.

[16] A. Baevski, W. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "data2vec: A general framework for self-supervised learning in speech, vision and language," in *Proc. International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 162. PMLR, 2022, pp. 1298–1312.

[17] H. Zhou, A. Baevski, and M. Auli, "A comparison of discrete latent variable models for speech representation learning," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 3050–3054.

[18] C. Du, Y. Guo, X. Chen, and K. Yu, "VQTTS: high-fidelity text-to-speech synthesis with self-supervised VQ acoustic feature," in *Proc. ISCA Interspeech*. ISCA, 2022, pp. 1596–1600.

[19] A. Sánchez, A. Falai, Z. Zhang, O. Angelini, and K. Yanagisawa, "Unify and Conquer: How phonetic feature representation affects polyglot text-to-speech (TTS)," in *Proc. ISCA Interspeech*. ISCA, 2022, pp. 2963–2967.

[20] M. Bernard and H. Titeux, "Phonemizer: Text to phones transcription for multiple languages in python," *J. Open Source Softw.*, vol. 6, p. 3958, 2021.

[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735–1780, 1997.

[22] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 2494–2498.

[23] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[24] "M-ailabs speech multi-lingual dataset," https://www.caito.de/2019/01/the-m-ailabs-speech-dataset/.

[25] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A corpus derived from librispeech for text-to-speech," in *Proc. ISCA Interspeech*. ISCA, 2019, pp. 1526–1530.

[26] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "AISHELL-3: A multi-speaker mandarin TTS corpus and the baselines," *CoRR*, vol. abs/2010.11567, 2020.

[27] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[28] J. Yang and L. He, "Towards universal text-to-speech," in *Proc. ISCA Interspeech*, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 3171–3175.

[29] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *CoRR*, vol. abs/2210.13438, 2022.

[30] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *CoRR*, vol. abs/2212.04356, 2022.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Conference on Neural Information Processing Systems (NIPS)*, 2017, pp. 5998–6008.

[32] H. Zeinali, S. Wang, A. Silnova, P. Matejka, and O. Plchot, "BUT system description to voxceleb speaker recognition challenge 2019," *CoRR*, vol. abs/1910.12592, 2019.