



CALM: Contrastive Cross-modal Speaking Style Modeling for Expressive Text-to-Speech Synthesis

Yi Meng^{1,†}, Xiang Li¹, Zhiyong Wu^{1,2,*}, Tingtian Li³,
Zixun Sun³, Xinyu Xiao³, Chi Sun³, Hui Zhan³, Helen Meng^{1,2}

¹Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems,
Shenzhen International Graduate School, Tsinghua University, Shenzhen, China

²Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Hong Kong SAR, China

³Tencent, Shanghai, China

{my20, xiang-li20}@mails.tsinghua.edu.cn, {zywu, hmmeng}@se.cuhk.edu.hk,
tingtian.li@outlook.com, {zixunsun, lawxiao, collinsun, huizhan}@tencent.com

Abstract

To further improve the speaking styles of synthesized speeches, current text-to-speech (TTS) synthesis systems commonly employ reference speeches to stylize their outputs instead of just the input texts. These reference speeches are obtained by manual selection which is resource-consuming, or selected by semantic features. However, semantic features contain not only style-related information, but also style irrelevant information. The information irrelevant to speaking style in the text could interfere the reference audio selection and result in improper speaking styles. To improve the reference selection, we propose Contrastive Acoustic-Linguistic Module (CALM) to extract the Style-related Text Feature (STF) from the text. CALM optimizes the correlation between the speaking style embedding and the extracted STF with contrastive learning. Thus, a certain number of the most appropriate reference speeches for the input text are selected by retrieving the speeches with the top STF similarities. Then the style embeddings are weighted summarized according to their STF similarities and used to stylize the synthesized speech of TTS. Experiment results demonstrate the effectiveness of our proposed approach, with both objective evaluations and subjective evaluations on the speaking styles of the synthesized speeches outperform a baseline approach with semantic-feature-based reference selection.

Index Terms: text-to-speech synthesis, speaking style, reference selection, style-related text features

1. Introduction

Human speeches naturally contain speaking style variations as an important component to express the intention and emotional states of the speaker, such as intonation and speaker rates. With the development of deep learning and speech processing [1, 2], many approaches have been proposed in text-to-speech (TTS) synthesis to enhance the speaking style of the synthesized speech. Conventionally, such enhancement is achieved by directly predicting the speaking styles from the given input texts [3, 4]. Variational autoencoder (VAE) [5] is employed to model the correlation between the speaking styles and the word embeddings of given texts [6]. Neural network based method is further proposed and employs a CBHG [7] network to predict the speaking styles from the word embeddings [3].

But there are differences between the linguistic information in text and the acoustic information in speaking style. Since style embedding is an entangled representation of many acoustic factors, it is difficult to predict style from text alone [8]. Besides, there might be many appropriate styles for the given text, thus facing the one-to-many problem, which makes the prediction even harder [9]. Therefore, many methods are proposed to learn the patterns of speaking styles from reference speeches. Global style embeddings (GSE) [10] are firstly introduced to extract the speaking style of a reference speech. Global style tokens (GST) [9] are further proposed to embed the GSE of the reference speech into a weighted sum of style tokens. These two approaches employ the ground-truth speeches as the references for training. However, since no ground-truth reference is available during inference, the synthesized speeches are observed to suffer from low content quality. Multiple references TTS (MRTTS) [8] is proposed and employs multiple references in both training and inference to overcome this issue and improve the robustness of TTS systems.

The reference selection in MRTTS is based on the semantic information of the text. However, semantic information contains not only style-related information, but also style irrelevant information. For example, semantic similarity is utilized in natural language processing to select the answers to a given question [11, 12, 13], while the actual speaking styles of the question and selected answers could differ considerably. Take the question ‘How is the weather?’ as an example, its semantic information is close to a typical answer like ‘Such a bad weather!’, since their topics are both about the weather. But the selected answer is spoken in a hateful exclamatory tone, while the question is spoken in an ordinary interrogative tone. In this situation, MRTTS may synthesize speech with improper speaking styles.

To overcome this issue, the reference speech selection should focus on the style-related information in the text. Inspired by Contrastive Language-Image Pre-training (CLIP) [14], which learns perception from the supervision contained in text-image pairs, we propose a multiple reference TTS system with Contrastive Acoustic-Linguistic Module (CALM). CALM is a contrastive-learning-based cross-modal speaking style selecting module, which learns to extract Style-related Text Features (STF) from text via audio supervision. CALM utilizes a style encoder to extract style embeddings from speeches, and a linguistic encoder to extract STF from the text. During the training stage, the style encoder is first trained along with the

[†] Work conducted when the first author was intern at Tencent.

*Corresponding author.

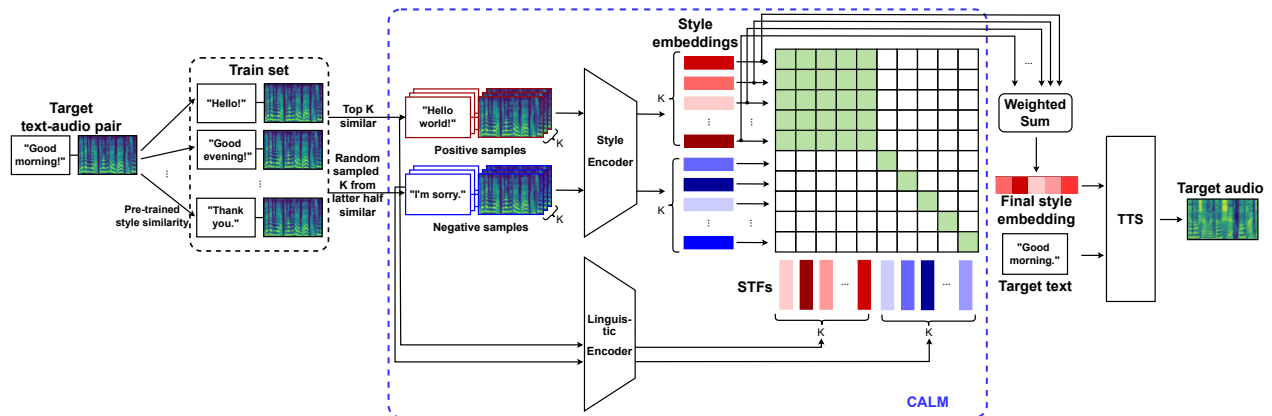


Figure 1: An overview of the training process of the proposed method

TTS system. Then we jointly optimize the style encoder and linguistic encoder by pushing the STFs with similar style embedding close to each other, while pushing away STFs with dissimilar style embeddings. Eventually, CALM learns to map the style and text of speech into the same embedding space shared by style embeddings and STFs. Thus, during the inference stage, to obtain the reference speeches whose speaking styles are most appropriate for the input text, we can directly select the speeches with the most similar STFs. Then the speaking styles are weighted summarized to stylize the synthesized speeches of TTS.

Experimental results demonstrate the effectiveness of our proposed approach. Both of the objective and subjective evaluations on Mandarin emotional dataset and English audiobook dataset confirm that CALM outperforms the semantic-feature-based baseline model on generating stylized speeches.

2. Methodology

As shown in Figure 1, our proposed model is based on the end-to-end TTS framework FastSpeech 2 [15]. Its style input comes from the Contrastive Acoustic-Linguistic Module (CALM), which consists of a linguistic encoder and a style encoder.

2.1. Contrastive Acoustic-Linguistic Module (CALM)

We use the linguistic encoder to extract Style-related Textual Feature (STF) from text, and the style encoder to extract style embedding from speech. CALM maps the style and text into the same embedding space and change the distance between their corresponding embeddings by contrastive learning. Specifically, CALM forces texts with similar speech styles (positive samples) to have closer STFs. Conversely, the STFs with different speech styles (negative samples) are dispersed as far as possible in the embedding space. During the inference stage, STF is used to guide the reference audio selection in TTS.

2.1.1. Positive and negative samples selection

To select the positive and negative samples for contrastive training, we first train the style encoder on the training set. Then we utilize the trained style encoder to extract style embeddings for each training audio, which is used to measure the style similarity among different samples. For each (speech, text) pair

in the training dataset, we sort all the other pairs in the training set by their style embedding cosine similarity to the style embedding of current speech. The top- K of the sorting results are selected as positive samples, while negative samples are obtained by randomly selecting K samples from the latter half of the sorted pairs. We do not select the last K samples as negative samples because the last K sample is too easy for the model to distinguish.

2.1.2. CALM training

As shown in the blue dash line in Figure 1, given $2K$ (speech, text) pairs as the input, CALM tries to predict a $2K \times 2K$ matrix M' , where the value at the position of the i^{th} row and j^{th} column indicates the cosine similarity between the style embedding of the i^{th} speech obtained by the style encoder and the STF embedding of the j^{th} text obtained by the linguistic encoder. The ground truth matrix M only consists of 1 and -1 values: (i) 1 for the elements whose STF embedding and style embedding are both from positive samples set or extracted from an identical pair in negative samples set; (ii) -1 for the others. In detail, the first K (speech, text) pairs are positive samples, and the last K pairs are negative samples. The loss function \mathcal{L}_{CALM} is the mean square error (MSE) between the predicted matrix M' and the ground-truth matrix M .

By these settings, CALM can learn a multi-modal embedding space by jointly training a style encoder and a linguistic encoder to maximize the cosine similarity of the embeddings between the positive samples and minimize the cosine similarity between the negative samples.

2.1.3. Model details

The setting of the style encoder of CALM is identical to the style encoder part of a GST enhanced FastSpeech 2 [9, 15]. The linguistic encoder of CALM aims to obtain sentence level feature from text. It contains a pre-trained BERT [16, 17] that takes a sentence with L words as input and outputs word level features with shape $(L \times D)$, which will be fed into a net with two layers GRU [18], dropout with rate r and a linear unit to adjust the output dimension to the same dimension with style embedding. With the help of the training process mentioned above, the output text feature tends to be more related to the speech style, thus called Style-related Textual Feature (STF).

3. Experiments

3.1. Datasets

In this paper, we use two datasets in two languages, Mandarin and English, to demonstrate the effectiveness of the proposed model. The synthesized samples are available on our demo page¹.

3.1.1. Mandarin emotional dataset

The first dataset is an internal single-speaker emotional Mandarin corpus, including six emotion categories (angry, fear, disgust, happy, sad, surprised). A total of 9,000 utterances is evenly divided by the six emotions. For each category of emotion, 50 utterances are randomly sampled to form the test set with a total of 300 utterances. The remaining 8,700 utterances serve as the training set. Though each utterance is labeled by emotion, we do not use the emotion label as supervision. The emotion label are only used to evaluate the performance of the reference audio selection in Section 3.3.1.

3.1.2. English audiobook dataset

The second dataset is an open-source English audiobook dataset. The books are read by the 2013 Blizzard Challenge speaker, Catherine Byers, in an animated and emotive storytelling style. Some books contain very expressive character voices with a high dynamic range. We choose four of these audiobooks as our dataset, with 9,741 utterances in total. We randomly select 300 utterances as the testing set while keeping the rest as the training set.

3.2. Models

3.2.1. Baseline

We implement Multiple Reference TTS (MRTTS) [8] as the baseline approach and replace its Tacotron-2 TTS backbone [1] with FastSpeech 2 for fairness concern. MRTTS selects the reference audios by semantic similarity during both the training and inference stages. MRTTS also constrains the style embeddings by pre-trained style embeddings during the training time. The semantic similarity is defined as the cosine similarity of two sentence-level embeddings, which is generated by a pre-trained BERT. For objective comparison, we used the same pre-trained BERT on the audiobook dataset with our proposed model, in which the output word embedding is with a dimension of 768.

3.2.2. Proposed model details

We utilize FastSpeech 2 [15] as the backbone of our model. The style encoder is the same setting with GST [9]. In the CALM, the linguistic encoder GRU has 256 hidden units, and the drop rate is 0.2. The loss weight λ is set to 1. The number of reference audios K in training stage is 20. If not specified, the number of reference audios N in inference stage is set to 20, which will be discussed in 3.4. As for waveform generation, a well-trained HiFi-GAN [19] is chosen as our neural vocoder.

3.3. Evaluations

3.3.1. Objective evaluation

We first demonstrate that the STF s extracted by CALM are more related to speaking style than the semantic features ex-

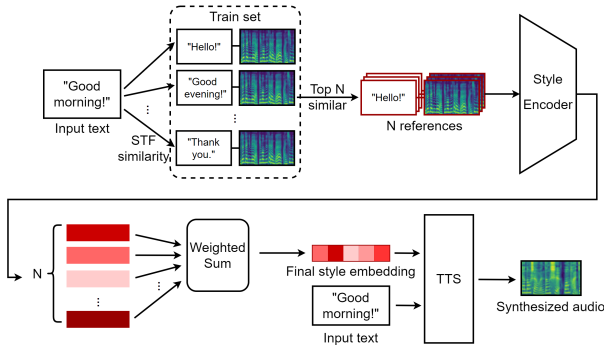


Figure 2: STF-driven multiple references selection

2.2. The TTS system with CALM

2.2.1. Training stage

During training time, as demonstrated in Section 2.1.1, we firstly train a GST enhanced FastSpeech 2, namely the TTS module in Figure 1. The pre-trained style encoder is used to extract speech styles in the training set. After that, we get positive samples and negative samples for all utterances, the style embeddings of the positive samples will not only be used in CALM, but also weighted summarized into the final style embedding used in TTS module.

We define N as the number of reference audios. Given N style embeddings extracted from reference audios $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n]$, and the STFs extract from the text of references $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_n]$, and the STF of the input text \mathbf{t}_0 , the calculation of the final style embedding can be notated as below

$$\mathbf{w} = \text{softmax}(\mathbf{T}\mathbf{t}_0^T) \quad (1)$$

$$\text{final_style_embedding} = \mathbf{w}^T \mathbf{S} \quad (2)$$

where \mathbf{w} is the weights of style embeddings. Since the positive samples in CALM training also served as reference audios to predict target audio, we set $N = K$ in training time. Taking the final style embedding and the target text as input, the TTS module produces target audio. We jointly train the entire TTS system from scratch with CALM. The total loss of the whole TTS system can be defined as

$$\mathcal{L}_{total} = \mathcal{L}_{TTS} + \lambda \mathcal{L}_{CALM} \quad (3)$$

where λ is the weight of the CALM loss. With the help of this loss function, we can jointly optimize the CALM and TTS part of the whole system.

2.2.2. Inference stage

As shown in Figure 2, during the inference time, input text is fed into the trained linguistic encoder in CALM to get the STF. All the text of (speech, text) pairs in the training set will be fed into the linguistic encoder of CALM to get STFs. Then we calculate the STF cosine similarity with all STFs in the training set and sort them in descending order. We use top- N of them as the reference audios to extract style embeddings, and weighted summarized to get the final style embedding. Taking the final style embedding and the text as input, the TTS system can synthesize audio with corresponding speaking style.

Unlike the training stage, the number of N can vary and differ from K . We will discuss the influence of N in Section 3.4.

¹<https://thuhcsi.github.io/interspeech2022-CALM-tts/>

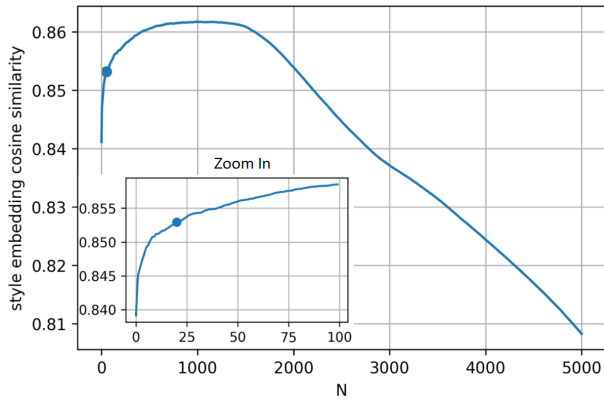


Figure 3: Style similarity with ground-truth style embedding in test set

tracted by BERT. We reveal this by calculating precision on the test set of the emotional dataset. (We can’t calculate the precision for the audiobook dataset since it do not have style labels.) A higher precision indicates that more audios with the correct emotion labels were selected. Specifically, given N selected text-audio pairs, among which N_+ samples own the same emotion label with the input, the precision is defined as:

$$Precision = \frac{N_+}{N} \quad (4)$$

We also demonstrate that our model can generate style embedding closer to ground-truth than the baseline. We calculate the cosine similarities between the generated style embedding and the ground-truth to measure that.

Table 1: Objective evaluations for different approaches

Emotional	Precision	Similarity
Baseline	0.523	0.749
Proposed	0.776	0.853
Audiobook	Precision	Similarity
Baseline	*	0.762
Proposed	*	0.877

As shown in Table 1, our model achieves higher performance than baseline. On the emotional dataset, the precision increases from 0.523 to 0.776, while cosine similarity is elevated from 0.749 to 0.853. On the audiobook dataset, the cosine similarity is elevated from 0.762 to 0.877. The objective evaluation metrics indicate that our model can obtain style embedding more similar to the ground-truth than baseline.

3.3.2. Subjective evaluation

We show that our model also achieves better performance in subjective evaluations. 30 speech files generated by the baseline and proposed approaches are rated by 23 native Mandarin speakers with good English proficiency on a scale from 1 to 5 with 1 point increments according to whether the speaking style is appropriate for the current text, from which a subjective mean opinion score (MOS) is calculated. Meanwhile, the listeners are asked to choose a preferred speech from the speeches synthesized by the baseline and proposed approaches, from which ABX preference rates are calculated.

Table 2: MOS scores with 95% confidence intervals of the baseline and the proposed approaches.

MOS	Baseline	Proposed
Audiobook	3.160 ± 0.129	3.625 ± 0.122
Emotional	2.804 ± 0.102	3.812 ± 0.089

Table 3: ABX Preference rates between the baseline and the proposed approaches. NP stands for no preference.

Preference	Baseline	NP	Proposed
Audiobook	30.5%	25.7%	43.8%
Emotional	16.5%	10.3%	73.2%

As shown in Table 2 and 3, the experimental results in both datasets demonstrate the effectiveness of our method. The MOS score increases from 3.160 to 3.625 in audiobook dataset and increases from 2.804 to 3.812 in emotional dataset. The ABX preference rate exceeds the baseline by 56.7% in emotional dataset and 13.3% in audiobook dataset. Both of the subjective evaluations indicate that our model can generate audios with more appropriate style than baseline.

3.4. Selection of N

We investigate the influence of the number of reference audios N in the test set of emotional dataset in Figure 3. The result shows that, as N grows, the cosine similarity of the weighted sum style embeddings and the ground-truth style embeddings first increases then decreases. The turning point came around $N = 1,500$. In the training set, each kind of emotion has 1,450 utterances. So we could say the linguistic encoder in CALM extracts text features more related to speaking style. The similarity increases with N , but the computational complexity also increases with N . Considering the trade-off between performance and complexity, we choose N as 20, where the improvement of similarity starts to slow down.

4. Conclusion

We propose a contrastive-learning-based multiple referenced TTS model to synthesize speech with a more appropriate speaking style for the input text. To demonstrate the effectiveness of our model, we conduct evaluations on two datasets. Our model outperforms the baseline in both objective and subjective evaluation metrics. The experimental results indicate that our model achieves a better style modeling approach and generates more natural audio. Finally, we discuss how the number of reference audios N influences the performance and choose it in consideration of both the performance and the computational complexity.

5. Acknowledgement

This work is supported by National Natural Science Foundation of China (62076144), National Social Science Foundation of China (13&ZD189), Shenzhen Key Laboratory of next generation interactive media innovative technology (ZDSYS20210623092001004) and Tsinghua University - Tencent Joint Laboratory.

6. References

- [1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-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.
- [2] H. Ze, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 7962–7966.
- [3] D. Stanton, Y. Wang, and R. Skerry-Ryan, “Predicting expressive speaking style from text in end-to-end speech synthesis,” in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 595–602.
- [4] C.-M. Chien and H.-y. Lee, “Hierarchical prosody modeling for non-autoregressive speech synthesis,” in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 446–453.
- [5] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.
- [6] Y.-J. Zhang, S. Pan, L. He, and Z.-H. Ling, “Learning latent representations for style control and transfer in end-to-end speech synthesis,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6945–6949.
- [7] 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,” *arXiv preprint arXiv:1703.10135*, 2017.
- [8] C. Gong, L. Wang, Z. Ling, J. Zhang, and J. Dang, “Using multiple reference audios and style embedding constraints for speech synthesis,” 2021.
- [9] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, “Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis,” in *International Conference on Machine Learning*. PMLR, 2018, pp. 5180–5189.
- [10] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, “Towards end-to-end prosody transfer for expressive speech synthesis with tacotron,” in *international conference on machine learning*. PMLR, 2018, pp. 4693–4702.
- [11] W. Lan and W. Xu, “Neural network models for paraphrase identification, semantic textual similarity, natural language inference, and question answering,” in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 3890–3902.
- [12] S. Minaee and Z. Liu, “Automatic question-answering using a deep similarity neural network,” in *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2017, pp. 923–927.
- [13] O. Chergui, A. Begdouri, and D. Groux-Lecllet, “Integrating a bayesian semantic similarity approach into cbr for knowledge reuse in community question answering,” *Knowledge-Based Systems*, vol. 185, p. 104919, 2019.
- [14] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.
- [15] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, “Fastspeech 2: Fast and high-quality end-to-end text to speech,” *arXiv preprint arXiv:2006.04558*, 2020.
- [16] 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.
- [17] Y. Cui, W. Che, T. Liu, B. Qin, S. Wang, and G. Hu, “Revisiting pre-trained models for chinese natural language processing,” *arXiv preprint arXiv:2004.13922*, 2020.
- [18] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using rnn encoder-decoder for statistical machine translation,” *arXiv preprint arXiv:1406.1078*, 2014.
- [19] J. Kong, J. Kim, and J. Bae, “Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.