METTS: Multilingual Emotional Text-to-Speech by Cross-speaker and Cross-lingual Emotion Transfer

Xinfa Zhu, Yi Lei, Tao Li, Yongmao Zhang, Hongbin Zhou, Heng Lu, Lei Xie, Senior Member, IEEE

Abstract-Previous multilingual text-to-speech (TTS) approaches have considered leveraging monolingual speaker data to enable cross-lingual speech synthesis. However, such dataefficient approaches have ignored synthesizing emotional aspects of speech due to the challenges of cross-speaker cross-lingual emotion transfer - the heavy entanglement of speaker timbre, emotion and language factors in the speech signal will make a system to produce cross-lingual synthetic speech with an undesired foreign accent and weak emotion expressiveness. This paper proposes Multilingual Emotional TTS (METTS) model to mitigate these problems, realizing both cross-speaker and cross-lingual emotion transfer. Specifically, METTS takes DelightfulTTS as the backbone model and proposes the following designs. First, to alleviate the foreign accent problem, METTS introduces multi-scale emotion modeling to disentangle speech prosody into coarse-grained and fine-grained scales, producing language-agnostic and language-specific emotion representations, respectively. Second, as a pre-processing step, formant shift based information perturbation is applied to the reference signal for better disentanglement of speaker timbre in the speech. Third, a vector quantization based emotion matcher is designed for reference selection, leading to decent naturalness and emotion diversity in cross-lingual synthetic speech. Experiments demonstrate the good design of METTS.

Index Terms—Speech synthesis, cross-lingual, emotion transfer, disentanglement, diffusion model

I. INTRODUCTION

R ECENT years have witnessed significant progress in the quality and naturalness of synthetic speech thanks to the advances of neural text-to-speech (TTS) systems [1], [2], [3], [4], [5], [6], [7], [8]. As near human parity performance has been reported in some closed TTS domains [9], *diversity* and *controllability* have become the new chasing target, including multi-speaker [10], multi-lingual [11], multi-emotion [8] as well as multi-style [12] scenarios. At the same time, *data efficiency* is also highly desired as modern corpus-based TTS heavily relies on high-quality annotated data. This induces a variety of approaches better leveraging limited, low-resource as well as low-quality data [13], [14], [15].

This paper addresses an extremely diverse and controllable speech generation scenario – multilingual emotional text-tospeech (METTS), particularly considering data efficiency by cross-speaker and cross-lingual emotion transfer. Specifically, METTS can produce bilingual emotional speech for each speaker after system building, while in the training data, each speaker speakes only one language (monolingual speaker), and some speakers have only neutral speech. Importantly, with only the neutral native speech (L1) data for a target speaker during training, helped with another emotional speaker in the target language (L2), our METTS system is able to produce the target speaker's emotional speech in the target language (L2) with reasonable proficiency and naturalness. METTS has significant applications such as foreign movie dubbing and computer-assisted language learning (CALL). However, building METTS is not a trivial task with three challenges.

- Foreign accent problem. Different languages have quite different prosody patterns in pronunciation. In a typical cross-lingual TTS system, the accent of the source language may be inevitably delivered to the speech in the target language [11], leading to non-native synthetic speech with a strong foreign accent. This problem is prominent for cross-speaker and cross-lingual emotion transfer because emotional expressions are heavily reflected in prosody patterns through changes in pitch, loudness, speech rate and pauses.
- Speech entanglement problem. Only partial speakers in the training set have emotional speech data, while we need to enable each speaker in the training set generates emotional speech. The speaker timbre and emotion entanglement in speech may lead to speaker timbre leakage when performing cross-speaker emotion transfer [16]. In other words, the source speaker's timbre may be inevitably transferred to the speech of the target speaker, making the synthetic speech of the target speaker sounds like the source speaker. Particularly, this problem becomes more severe for cross-speaker cross-lingual emotion transfer as the entanglement of three factors in speech – speaker timbre, language and emotion complicates the disentanglement process.
- *Emotional diversity problem*. In order to synthesize diverse emotions, the TTS model usually needs additional emotional information as prior, such as an emotion ID or a reference speech sample. Compared to emotion ID, providing emotional representation through a reference encoder is apparently more diverse [17] as different references lead to different fine-grained emotion deliveries. However, such diversity raises a problem how to select an appropriate reference signal to exactly match the textual content [18]. In this paper, cross-lingual reference

Corresponding author: Lei Xie

Xinfa Zhu, Yi Lei, Tao Li, Yongmao Zhang, and Lei Xie are with Audio, Speech and Language Processing Group (ASLP@NPU), the School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China. Email: xfzhu@mail.nwpu.edu.cn (Xinfa Zhu), leiyi@npu-aslp.org (Yi Lei), taoli@npu-aslp.org (Tao Li), zym@mail.nwpu.edu.cn (Yongmao Zhang), lxie@nwpu.edu.cn (Lei Xie)

Hongbin Zhou and Heng Lu are with Ximalaya Inc., Shanghai, China. Email: hongbin.zhou@ximalaya.com (Hongbin Zhou), bear.lu@ximalaya.com (Heng Lu)

selection, i.e., selecting a reference in L1 to match the text in L2, is a brand new problem for cross-speaker emotion transfer in the multilingual TTS system.

To address the above challenges, we propose METTS, a novel approach to synthesizing bilingual emotional speech for each monolingual speaker, even though some speakers do not have emotional speech data during model training. METTS is based on DeligtfulTTS [19], a state-of-the-art nonautoregressive text-to-speech approach with improved Conformer [20] blocks to model the variation of speech prosody. Based on the skeleton of DelightfulTTS, our METTS leverages the following designs to achieve: 1) emotion transfer from a reference mel-spectrogram to synthesize speech (METTS-REF) and 2) automatically matching the most suitable reference embedding according to the input text and emotion ID to synthesize speech (METTS-ID).

First, we introduce multi-scale emotion modeling to address the foreign accent problem. We believe emotion expressions in multilingual speech can be generally factorized into the shared similar prosody pattern [21], [22] (e.g., high pitch for angry and low pitch for sad) conveyed in both languages and distinct fine details of prosody [23], [24] due to the different manners of pronunciation. This inspires us to model emotion with coarse and fine scales to respectively represent languageagnostic and language-specific emotion aspects in speech. To be specific, the coarse-grained emotion is modeled by a Global Style Token (GST) [25] layer with semi-supervised constraints to make the coarse-grained emotional representation languageagnostic; a Conditional Variational Autoencoder (CVAE) layer establishes a language-specific emotion representation by investigating the fine-grained prosody variation of speech, with the condition of language-dependent text input and the above coarse-grained emotional representation. In this way, METTS manages to successfully transfer emotion across languages via the coarse-grained language-agnostic emotional representation and produces native pronunciation without a foreign accent in emotion delivery through the fine-grained language-specific emotion representation.

Second, we employ *information perturbation* [16] to further address the issue of speech entanglement. Specifically, the reference speech is perturbed by randomly shifting its formant frequency, which allows the multi-scale emotion modeling module to generate a speaker-independent emotional representation. This speaker-independent signal can decouple the speaker timbre from reference speech in nature.

Third, to address the emotional diversity problem and select an appropriate reference signal, we propose a vector quantization (VQ) based *emotion matching* module. Instead of directly modeling the complicated regression between bilingual textual representation and emotional representation, we first use VQ to quantify coarse-grained language-agnostic emotion representation to form a reference pool and subsequently adopt an emotion matcher to match the bilingual textual representation with the reference pool. This allows METTS to realize *reference-free* inference and produce more diverse emotional speech with reasonable expressiveness for both L1 and L2 text inputs.

The proposed METTS system is extensively evaluated on

a demanding Mandarin-English bilingual TTS task. This task poses significant challenges due to substantial differences in pronunciation between Chinese and English, such as variations in syllable structure, tones, vowels, and consonant inventory, the absence of retroflex sounds in Chinese, and the presence of long vowels in English [26], [27]. Experimental results show that although the performance of intra-lingual emotional speech synthesis is better than that of more challenging cross-lingual emotional speech synthesis, METTS can produce bilingual emotional speech for each target speaker and effectively improve speech naturalness, speaker similarity, and emotion similarity compared to other competitive methods. Furthermore, the component analysis validates the good design of our proposed model. Audio samples can be found on our demo page ¹.

The remainder of this paper is organized as follows. Section II provides a comprehensive review of related work in the field. Section III presents a detailed description of the proposed approach. The experimental setups and results are described in Section IV and Section V, respectively, where we analyze the performance of METTS and present the evaluation outcomes. In Section VI, we delve into the component analysis, examining the contributions of each individual module in our system. Finally, Section VII concludes the paper, summarizing the findings and highlighting future research directions.

II. RELATED WORK

Multi-lingual speech synthesis and emotional speech synthesis are two popular topics in the literature where transfer learning approaches – transferring speaker voices across languages or transferring style/emotion across speakers – are mainstream approaches better leveraging limited data. However, to the best of our knowledge, the current studies have not yet addressed both cross-speaker and cross-lingual emotion transfer. This is a more challenging task, as discussed in Section 1. Moreover, to improve the diversity of synthetic speech, cross-lingual reference selection is another new problem in reference-based multi-lingual emotional TTS. Therefore, here we review the prior arts in multilingual speech synthesis, emotional speech synthesis, and reference speech selection, respectively.

A. Multilingual speech synthesis

Developing a robust multilingual text-to-speech (TTS) system requires a unified representation of textual input. This can be achieved by merging phoneme sets from different languages or utilizing the International Phonetic Alphabet (IPA) to represent speech sounds. Based on the unified representation of textual input, several studies have explored a more efficient method that utilizes a single acoustic model with shared parameters across languages as an alternative to training separate models for each language [28], [29], [30]. Additionally, incorporating explicit language identification (ID) enables better control over language-specific prosody and improves the naturalness of synthetic speech [31], [32].

¹https://anonymous-rep0.github.io/METTS/

However, due to distinct prosody patterns in different languages, cross-lingual synthetic audio often suffers from an undesired foreign accent problem in multilingual TTS systems. To address this challenge, some researchers investigate obtaining language-specific and speaker-independent prosodic representations through implicit disentanglement [33], [34]. Typically, domain adversarial training strategies have been employed to encourage the model to learn disentangled representations of text and speaker identity [11], [35], [32], where a gradient reversal layer is inserted before a speaker classifier. Furthermore, some studies [36], [37] propose to use a style VAE encoder to alleviate the foreign accent problem by leveraging existing authentic styles during inference. Moreover, the triplet training scheme is utilized to overcome the accent problem by combining unseen speakers and language through finetuning [38]. CrossSpeech [39] obtains disentangled speaker and language representations through the speaker-independent generator and speaker-dependent generator.

It is important to note that while these approaches have shown promise in improving the performance of multilingual speech synthesis, challenges still remain in achieving emotionally expressive and foreign accent-free speech across different languages.

B. Emotional speech synthesis

Significant progress has been made in the field of emotional speech synthesis in recent years [17], [8], [40], [41]. When categorized emotion data is available for the target speaker, a straightforward approach is to model emotional expressions based on discrete emotion IDs [42], [43]. However, the resulting synthetic speech often exhibits over-averaged emotional expressions due to the discrete nature of emotion ID control. On the other hand, transfer learning approaches have been proposed, where a reference encoder is used to extract emotional representations from a reference signal, providing guidance for emotion synthesis. These emotion transfer approaches offer more flexibility and diversify the generated emotional speech [44], [45], [46].

Emotion transfer is effective for generating emotional speech for speakers who only have neutral data, while it often faces a trade-off between speaker similarity and emotional expressiveness in synthetic speech. This trade-off results in either low speaker similarity or poor emotion similarity. To address this issue, disentangling these speech attributes and modeling them separately becomes necessary. Various schemes have been proposed for disentanglement. Some work implicitly decouples speech attributes in latent representations [47], [48]. Whitehill et al. [49] use an unpaired training strategy and adversarial cycle consistency scheme to disentangle emotion and speaker. Li et al. [40] propose an emotion-disentangling module, which learns speaker-independence emotion embedding via an orthogonal loss with the speaker embedding. On the other hand, some studies [50], [51] explicitly decouple speech attributes through bottleneck features, information perturbation, and other methods. Li et al. [16] learn emotionrelated mel-spectrogram and speaker-related mel-spectrogram through information perturbation and generate emotionally expressive speech.

Considering human emotional expression's diverse and complex patterns, some research proposes to model emotional speech at multiple scales to capture rich emotional variations. [8], [52], [53]. For instance, the approach in [8] combines global emotion representation with local emotion representation at a fine-grained level, such as phoneme- or syllable-level, resulting in more natural and expressive speech. However, in the context of multilingual emotional speech synthesis, the entanglement between emotion and language becomes more complex and requires adequate consideration.

C. Reference speech selection

The selection of appropriate reference speech plays a crucial role in ensuring the naturalness and diversity of emotional expression in emotion transfer-based TTS approaches. In practical non-parallel transfer applications, the textual contents of reference audio are different from that of generated speech during inference, while they remain the same during training. This mismatch can result in degraded speech naturalness [54], [55] and inappropriate emotional expressions. Since there is no text-matched reference available during inference, the problem of selecting suitable reference audio becomes challenging.

One straightforward approach to address the mismatch problem is leveraging multiple references during training and inference, such as through simple averaging embeddings of multiple references [56]. Recently, more sophisticated approaches based on context have been proposed for reference selection. For instance, ProsodySpeech [57] utilizes a Prosody Distributor that employs an attention mechanism to select references at the phone level. Inspired by Contrastive Language-Image Pre-training (CLIP) [58], CALM [18] incorporates a Contrastive Acoustic-Linguistic Module to select reference speeches based on the input text.

However, as human emotions and the contextual content of multilingual scripts are highly complicated, addressing emotional reference selection and accounting for the diverse contextual content in multilingual scenarios remain essential research topics.

III. METHODOLOGY

This section first gives an overview of METTS, followed by the motivation and design of each module. The training pipeline will also be introduced in this section.

A. Overview

The proposed framework is built on a multilingual text processing front-end, which supports both Chinese and English. The front end encodes Chinese text input into labels of phonemes, tones, word boundaries, and prosodic boundaries while encoding English text input into labels of phonemes. Note that the Chinese phoneme set is based on Pinyin, while English is based on CMU-Dict. Therefore, we merge the Chinese and English phoneme sets for unified textual inputs of the bilingual TTS system.

As shown in Figure 1, the backbone of METTS is based on DelightfulTTS [19], which consists of a text encoder, a



Fig. 1: The architecture of METTS.

variance adaptor, and a mel-spectrogram decoder. Notably, the improved Conformer [20] structure better model local and global dependency of mel-spectrogram, which has led DelightfulTTS to win the Blizzard Challenge 2021. In general, METTS updates DelightfulTTS with multi-scale emotion modeling to achieve natural multilingual emotional TTS by both emotion reference (METTS-REF) and ID (METTS-ID) as the control signal. Specifically, the coarse-grained emotion embedding is provided by a GST layer or an emotion matcher, while the fine-grained emotion embedding is obtained from the CVAE module. Moreover, a perturb module is introduced to distort the speaker timbre of the reference signal for decoupling timbre from speech. The speaker embedding is obtained through a lookup table and is added to the output of the variance adaptor, which is then fed into the mel-spectrogram decoder to synthesize the final mel-spectrogram.

B. Multi-scale emotion modeling

In general, the emotional expressions of multi-lingual speech could be factorized to the shared prosody pattern and distinct fine details of prosody due to the different manners of pronunciation. METTS utilizes the Global Style Tokens (GST) and Conditional Variational Autoencoder (CVAE) modules to establish coarse-grained language-agnostic and fine-grained language-specific emotional representations.

The GST module employs a reference encoder to encode mel-spectrogram into a hidden representation and utilizes multi-head attention to calculate the global emotional style tokens. Notably, L2 normalization is applied to the global emotional representation, eliminating magnitude-related information that may vary across languages and speakers. This normalization improves the generalization ability of the model, allowing for emotion embedding control based solely on angular information geometrically [59]. By mapping the emotional representation of different languages to the same global tokens, the GST module forms the language-agnostic representation.

The CVAE module focuses on learning fine-grained emotion expression from the mel-spectrogram, with text and coarsegrained emotion conditions. It utilizes a conformer block and a GRU layer to extract frame-level emotion embeddings, which are then downsampled to the phoneme level based on duration. These phoneme-level embeddings are used to derive the mean and variance of the distribution of phoneme-level prosody. Additionally, taking inspiration from VITS [7], a fine-grained predictor is designed to predict the distribution of fine-grained emotion from text. To improve the expressiveness of the predicted distribution, a normalizing flow technique is employed, enabling an invertible transformation from a simple distribution to a more complex one. By learning fine-grained emotional representations that are consistent with the text, the CVAE module forms the language-specific representation. To ensure that the extracted multi-scale representations are relevant to emotions, even for training data without emotion annotations, a semi-supervised strategy is employed, which includes an emotion classifier. Specifically, only the embeddings of annotated audio are used to supervise the emotion classifier for both coarse- and fine-grained representations. The audio without emotion annotations is not involved to optimize the emotion classifiers and is utilized to train the acoustic model by the extracted embeddings. Furthermore, the frame and phoneme emotion embeddings are processed through a GRU layer to extract a single vector, which is then used as input to the emotion classifier.

C. Speaker disentanglement based on information perturbation

In our multilingual emotional speech synthesis setup, where each speaker is mono-lingual and only some speakers in the training set have emotional speech data, the entanglement of speaker timbre with emotion and language poses a challenge, resulting in synthetic speech with low speaker similarity and unusual emotional expression and pronunciation.

To address this issue more comprehensively, we adopt a pre-processing step that utilizes a signal perturbation module to remove the speaker timbre information. Specifically, we apply a dynamic *formant* shift to the mel-spectrogram of the reference speech. Speech formants are primarily determined by the size, shape, and position of the vocal tract, which are highly specific to each speaker and represent their vocal identity [60]. By performing a formant shift function, denoted as fs, on the original waveform Wave, we obtain a speaker-independent signal denoted as Wave = fs(Wave).

Subsequently, we extract the mel-spectrogram of the perturbed wave, denoted as Mel, which serves as the input for emotion representations extraction. The perturbation module perturbs the timbre of the recordings at a random scale by each step during training, allowing the GST and CVAE modules to learn a speaker-independent representation and effectively disentangle the speaker's timbre from speech.

D. VQ based emotion matcher

During inference, different reference usually lead to different emotional expression of the synthetic speech. The selection of an appropriate reference is crucial for achieving natural speech and accurate emotional expression. Therefore, we propose an emotion matcher that automatically matches the most suitable reference embedding based on the textual content. Figure 2 illustrates the architecture of the emotion matcher, which takes the text encoder output and the language ID as input and generates the optimal reference embedding, facilitating reference-free inference in METTS.

Directly modeling the complex relationship between bilingual textual representation and emotional representation is quite challenging, so we employ VQ to quantize the coarsegrained emotion representation, forming a reference pool. Subsequently, we predict the correct codebook from the reference pool for the current utterance. This transforms the intricate regression task into a simpler classifier task, simplifying the modeling process.

Specifically, in our approach, we begin by extracting the coarse-grained emotional representation and predicted emotion ID for all training audio samples using the GST layer and the emotion classifier. Subsequently, we apply VQ to quantize the coarse-grained emotional representation. To achieve this, we employ the k-means algorithm to obtain N clusters for each emotion category, resulting in a total of $N \times M$ reference embeddings (M is the number of emotion categories), which form the reference pool.

To select the appropriate reference from the pool, the emotion matcher utilizes a multilayer perceptron (MLP) to generate a text-emotion vector. This vector captures the contextual information related to emotion by taking the text encoder output and emotion ID as inputs. Using the text-emotion vector and the embedding-candidate pool, we calculate the correlation coefficient (CC) matrix between them. The calculation of CC is similar to the process of Scaled Dot-Product Attention [61] and is defined as:

$$CC(V_t, E_c) = \operatorname{softmax}\left(\frac{V_t E_c^T}{\sqrt{d_{E_c}}}\right),$$
 (1)

where V_t and E_c represent the text-emotion vector and embedding candidates, respectively, and d_{E_c} denotes the dimension of the embedding candidates. The softmax function is applied to normalize the correlation coefficients. The embedding with the highest correlation coefficient is selected as the coarsegrained emotion representation for the TTS system.

To ensure that the selected embedding corresponds to the input text, a matcher classifier is introduced to supervise the CC matrix. It ensures that the embedding with the highest correlation coefficient is the cluster center corresponding to the input text.

E. Training and fine-tuning

For flexible control of the generated emotional expressions, we use pre-training and fine-tuning procedures to conduct multilingual emotional speech synthesis from a reference signal and manual emotion ID, respectively.

The training objective of METTS-REF is

$$\mathcal{L}_{\text{pretrain}} = 0.05 * \mathcal{L}_{\text{kl}} + \mathcal{L}_{\text{prosody}} + 0.1 * \mathcal{L}_{\text{emo}} + \mathcal{L}_{\text{ssim}} + \mathcal{L}_{\text{iter}},$$
(2)

 $\mathcal{L}_{\text{prosody}}$ the loss between the where is L1 predicted pitch/energy/duration and the ground-truth pitch/energy/duration, \mathcal{L}_{emo} the semi-supervised is crossentropy loss of emotion classifier. \mathcal{L}_{kl} is the KL divergence to predict the phoneme-level emotion distribution from text encoder output, and \mathcal{L}_{iter} is the sum of melspectrogram L1 loss between the predicted and ground-truth mel-spectrogram in each Conformer block. Moreover, we use structural similarity \mathcal{L}_{ssim} [62] to measure the similarity between predicted and ground-truth mel-spectrogram in the final Conformer block.

The purpose of fine-tuning is to support METTS-ID. During fine-tuning, we use the ground-truth clustering center as the



Fig. 2: The architecture of emotion matcher.

coarse-grained emotion representation for the TTS model and jointly optimize the emotion matcher. To stabilize the joint training, we freeze the GST-layer and emotion classifier as a discriminator to distinguish the emotion category of the generated mel-spectrogram.

The fine-tuning objective is

$$\mathcal{L}_{\text{finetune}} = \mathcal{L}_{\text{match}} + \mathcal{L}_{\text{disc}} + \mathcal{L}_{\text{base}'}, \qquad (3)$$

where \mathcal{L}_{match} is the cross entropy loss between the selected clustering centre and the actual clustering centre in the emotion matcher, \mathcal{L}_{disc} is the emotion classification loss of the predicted mel-spectrogram taking GST layer as an discriminator, and $\mathcal{L}_{base'}$ means removing \mathcal{L}_{emo} from the pre-trained model objectives.

IV. EXPERIMENTAL SETUPS

This section introduces the database configuration, training setups, compared methods, and evaluation methods.

A. Dataset

To assess the performance of METTS, we conduct a series of experiments on Chinese and English datasets, as shown in Table I. These two languages have tremendous pronunciation differences, which poses a challenge for multilingual speech synthesis. The Chinese dataset includes audio clips from two female speakers, denoted as CN1 and CN2, expressing six types of emotions (anger, fear, happiness, sadness, surprise, and neutral). The total number of audio clips was 22,205, approximately 21 hours of audio in sum. The English dataset includes audio clips from two female speakers, EN1 and EN2, for 19,676 audio clips, approximately 20 hours. There is no apparent emotional expression in English datasets. All data are studio-quality recorded at 48KHz.

B. Training setups

For all texts, the TTS front end encodes Chinese text input into phonemes, tones, word boundaries, and prosodic boundaries while decoding English text input into labels of phonemes. We down-sample all the audios into 24k Hz and set the frame and hop sizes to 1200 and 300, respectively, when extracting optional auxiliary acoustic features like pitch and mel-spectrogram. The auxiliary pitch and energy contour is extracted through WORLD [63], and the implementation of formant shifting is achieved by Praat, following the NANSY [50] model. The phoneme duration is obtained through an HMMbased force alignment model [64].

METTS takes DelightfulTTS [19] as the backbone, which consists of an encoder and decoder, both containing six conformer blocks. The dimensions of the emotion embedding and speaker embedding are set to 384. The CVAE module uses the Flow setting from VITS [7], and the dimension of the fine-grained emotion embedding is set to 16. The multi-layer perceptron (MLP) consists of one conformer block layer, six two-dimensional convolution layers, and one gated recurrent unit (GRU) layer, which outputs a 384-dimensional vector. The number of clusters in the k-means algorithm N is set to 64. All classifiers have the same structure that consists of 3 fully connected layers with the Relu activation function.

All models are trained up to 400k steps on two 2080Ti GPUs with a batch size of 12 and use a MelGAN [65] vocoder to convert the generated mel-spectrogram into waveforms.

C. Comparison methods

As this work is the first attempt, to the best of our knowledge, to synthesize foreign emotional speech through emotion transfer from reference speech or directly based on emotion ID, there are no existing methods directly comparable to our proposed approach. However, we compare our proposed METTS with the most relevant and recent methods in the field to provide a fair evaluation. To ensure fairness in the comparison, we implement the following comparison models on the delightful TTS model backbone and maintain identical training setups.

• **CET** [41] is a powerful Cross-speaker Emotion Transfer speech synthesis system, which defines several emotion tokens that are trained to be highly correlated with corresponding emotions by a semi-supervised training

C	T	Emotion (sentences)						I.L.	
Corpus	Language	Neutral	Нарру	Surprise	Sadness	Angry	Disgust	Fear	Usage
CN1	Chinese	5k	0.5k	0.5k	0.5k	0.5k	0.5k	0.5k	Training&Evaluation
CN1	Chinese	5k	2k	2k	2k	2k	2k	2k	Training&Evaluation
EN1	English	10k	-	-	-	-	-	-	Training&Evaluation
EN2	English	10k	-	-	-	-	-	-	Training&Evaluation

strategy. Speaker condition layer normalization is implemented to eliminate the down-gradation to the timbre similarity for cross-speaker emotion transfer. During inference, the model transfers emotion from a reference mel-spectrogram to the synthetic speech.

- M3 [36] is a Multi-speaker, Multi-style, and Multilingual text-to-speech system, which utilizes a speaker conditional variational encoder and conducts adversarial speaker training by the gradient reversal layer. Moreover, the model uses a Mixture Density Network (MDN) for mapping text and the extracted style vectors for each speaker. In inference time, the model predicts emotion representation according to emotion ID and text to synthesize speech.
- **METTS-REF** is the proposed model that transfers emotion from a reference mel-spectrogram to synthesize speech.
- **METTS-ID** is the proposed model that automatically matches the most suitable reference embedding according to the input text and emotion ID to synthesize speech.

D. Evaluation metrics

To evaluate the performance of the benchmark systems, we conduct a comprehensive set of evaluation methods. We prepare two test sets consisting of forty English texts and forty Chinese texts. For each speaker and emotion category, we generate samples, resulting in a total of 1,920 samples (2 languages \times 40 texts \times 4 speakers \times 6 emotions) for evaluation. For models that transfer emotion from a reference melspectrogram, we provide randomly selected mel-spectrograms from CN_spk1 as the reference. For models that synthesize speech based on emotion ID, we provide the corresponding emotion ID as input.

For subjective evaluation, we conducted two types of human perceptual rating experiments. A total of twenty-two volunteers with basic English skills participate in these experiments. Mean Opinion Score (MOS) [66] is used to evaluate the naturalness of the synthetic speech. Participants are asked to rate the speech on a scale ranging from 1 to 5, reflecting the influence of foreign accents and emotion on naturalness. The rating criteria are as follows: bad = 1, poor = 2, fair = 3, good = 4, great = 5, with 0.5-point increments. Similarity Mean Opinion Scores (SMOS) [67] is adopted to subjectively evaluate the synthetic speech from two aspects: emotion similarity and speaker similarity. Participants are asked to rate the speech's similarity to a given emotional reference and the similarity to the reference of the target speaker. The rating scale and criteria are the same as those used in the MOS evaluation.

For objective evaluation, we measure speaker cosine similarity, character error rate (CER), and word error rate (WER) for the synthetic audio. To measure speaker cosine similarity, we train an ECAPA-TDNN [68] model trained on 3,300 hours of Mandarin speech and 2,700 hours of English speech from 18,083 speakers to extract x-vectors. We extract the averaged x-vector of all utterances for each English speaker and six averaged x-vectors for each emotion category of the Chinese speaker. We then extract the x-vector of the synthetic audio and calculate the cosine distance. A higher cosine similarity indicates a more similar speaker timbre. To evaluate CER and WER, we use an open-source model provided by the WeNet community [69], which uses the U2++ conformer architecture and is trained on 10,000 hours of open-source Gigaspeech English data [70] and 10,000 hours of open-source WeNet Mandarin data [71], respectively. A higher CER or WER indicates less accurate pronunciation.

V. EXPERIMENTAL RESULTS

This section evaluates the performance of each system to produce bilingual emotional speech for Chinese and English speakers. The comparison between METTS and other methods is presented and discussed.

A. Subjective evaluation

We initially conducts a subjective evaluation to assess the performance of the generated multilingual emotional speech in terms of speech naturalness, speaker similarity, and emotion similarity for both Chinese and English speakers. The evaluation results, as presented in Table II and Table III, demonstrate that the proposed METTS family consistently outperforms the baseline models across all evaluation metrics for both Chinese and English speakers. Notably, all models exhibit a performance degradation during cross-lingual emotional speech synthesis, indicating that the synthetic Chinese speech for English speakers generally has lower quality compared to that for Chinese speakers. Nevertheless, the proposed METTS family demonstrates relatively minor degradation in performance during cross-lingual emotional speech synthesis, suggesting its capability to generate natural and fluent foreign speech for a given target speaker.

Comparing the different models in the METTS family, METTS-REF achieves the highest speaker and emotion similarity scores, indicating its effectiveness in transferring emotions from reference to synthetic speech. On the other hand, METTS-ID achieves almost the highest naturalness score and comparable emotion similarity to METTS-REF. This result validates the efficacy of the emotion matcher module in accurately matching a suitable reference embedding to synthesize

TABLE II: Results of subjective evaluation with 95% confidence interval for Chinese speakers.

		Chinese Text			English Text	
Model	Naturalness	Speaker Similarity	Emotion Similarity	Naturalness	Speaker Similarity	Emotion Similarity
METTS-REF METTS-ID CET [41] M3 [36]	4.11±0.12 4.07±0.13 3.65±0.14 2.69±0.19	3.94±0.16 3.88±0.16 3.69±0.12 3.35±0.15	4.12±0.14 3.95±0.13 4.00±0.11 3.21±0.18	4.00±0.11 4.06±0.18 3.01±0.19 2.49±0.23	3.94±0.12 3.77±0.20 3.35±0.14 3.46±0.16	3.44±0.22 3.24±0.22 3.39±0.16 2.97±0.16

TABLE III: Results of subjective evaluation with 95% confidence interval for English speakers.

		Chinese Text			English Text		
Model	Naturalness	Speaker Similarity	Emotion Similarity	Naturalness	Speaker Similarity	Emotion Similarity	
METTS-REF METTS-ID CET [41] M3 [36]	3.91±0.14 4.02±0.15 3.08±0.16 2.72±0.15	3.68±0.18 3.57±21 2.88±0.17 3.17±0.15	3.71±0.15 3.73±0.17 3.41±0.16 3.01±0.18	3.95±0.14 4.05±0.18 2.89±0.12 2.41±0.19	3.82±0.16 3.74±0.18 3.33±0.15 3.24±01.5	3.44±0.19 3.26±0.14 3.21±0.20 2.81±0.18	

more natural speech. Furthermore, there are two exceptional cases worth mentioning. In Table II, METTS-REF achieves the highest naturalness score in synthesizing Chinese emotional speech for Chinese speakers, indicating that intra-lingual emotion expressions of different speakers are similar. In Table III, METTS-ID obtains the highest emotion similarity score in synthesizing Chinese emotional speech for English speakers, which suggests that the coarse-grained emotion embedding provided by the emotion matcher module is close to that of the emotion encoder for English speakers in this particular condition.

CET demonstrates similar emotion similarity to METTS-REF under specific test conditions, indicating its powerful ability in emotion transfer. However, CET is primarily designed for inter-language emotion transfer and relies on a single-scale emotion representation, which is hard to capture the diverse emotional expressions across different languages. As a result, the synthetic speech may exhibit a heavy accent. Therefore, CET receives lower scores in naturalness and speaker similarity evaluations. In contrast, our proposed METTS model incorporates multi-scale emotion modeling to capture both language-specific and language-agnostic emotional expressions, effectively avoiding the entanglement of accents with emotions. Furthermore, M3 performs poorly across all evaluation metrics. M3 assumes a strong correlation between style coding and the speaker's attributes and content [36], which leads to an entanglement between the speaker's timbre and emotion. Additionally, the domain adversarial training used in M3 for speaker timbre disentanglement is not stable [72]. In contrast, our proposed model employs information perturbation to effectively remove the speaker's timbre, resulting in a more stable and practical approach.

B. Objective evaluation

To comprehensively evaluate the performance of our multilingual emotional TTS system, we conduct objective tests to measure speaker cosine similarity, character error rate (CER) for synthetic Chinese speech, and word error rate (WER) for synthetic English speech.

The objective test results presented in Tables IV and Table V confirm the observations from the subjective evaluation, high-

TABLE IV: Results of objective evaluation for Chinese speakers.

	Chinese Tex	t	English Text	
Model	Cosine Similarity	CER	Cosine Similarity	WER
METTS-REF	0.813	0.48	0.753	5.60
METTS-ID	0.805	0.48	0.711	5.46
CET [41]	0.726	0.35	0.638	12.65
M3 [36]	0.754	11.02	0.673	55.32

TABLE V: Results of objective evaluation for English speakers.

	Chinese Tex	t	English Text	
Model	Cosine Similarity	CER	Cosine Similarity	WER
METTS-REF	0.735	1.38	0.769	5.51
METTS-ID	0.709	1.36	0.786	2.15
CET [41]	0.659	1.24	0.663	8.05
M3 [36]	0.671	16.74	0.704	38.22

lighting the distinction between inter-lingual and cross-lingual speech synthesis. The METTS family achieves the highest speaker cosine similarity, demonstrating the effectiveness of our approach in disentangling speaker timbre from both emotion and language. Furthermore, the METTS family achieves lower CER and WER, indicating its stability in generating intelligent, natural-sounding multilingual emotional speech.

It is worth noting that CET achieves the lowest Chinese CER, showcasing its ability in intra-lingual emotion transfer. However, the higher English WER of CET reflects the significant challenges of cross-lingual emotion transfer, which aligns with the subjective evaluation results concerning naturalness. Additionally, M3 fails to effectively address accentrelated challenges in multilingual emotional speech synthesis, resulting in incorrect pronunciation and yielding the highest CER and WER. Furthermore, the results of speaker cosine similarity suggest that information perturbation for speaker timbre removal employed in our approach is more effective than the SALN method used in CET and the speaker adversarial training method in M3 in terms of speaker disentanglement in multilingual emotional speech synthesis.

TABLE VI: Results of Ablation study with 95% confidence interval for Chinese speakers.

		Chinese Text			English Text		
Model	Naturalness	Speaker Similarity	Emotion Similarity	Naturalness	Speaker Similarity	Emotion Similarity	
METTS-REF - GST - CVAE - Perturb	4.11±0.12 3.50±0.15 3.95±0.12 4.02±0.12	3.94±0.16 3.82±0.17 3.81±0.16 3.87±0.15	4.12±0.14 3.19±0.18 3.88±0.12 4.19±0.14	4.00±0.11 3.69±0.13 3.43±0.14 3.45±0.17	3.94±0.12 3.80±0.18 3.82±0.13 3.55±0.16	3.44±0.22 3.07±0.17 3.39±0.17 3.58±0.21	

TABLE VII: Results of Ablation study with 95% confidence interval for English speakers.

		Chinese Text	Chinese Text		English Text	
Model	Naturalness	Speaker Similarity	Emotion Similarity	Naturalness	Speaker Similarity	Emotion Similarity
METTS-REF	3.91±0.14	3.68±0.18	3.71±0.15	3.95±0.14	3.82±0.16	3.44±0.19
- GST	3.75±0.19	3.52±0.21	3.24±0.21	3.73±0.19	3.64±0.21	3.17±0.18
- CVAE	3.71±0.18	3.56±0.19	3.58±0.19	3.17±0.18	3.76±0.21	3.30±0.20
- Perturb	3.85±0.21	3.19±0.27	3.82±0.19	3.50±0.22	3.26±0.29	3.52±0.20



(b) Colored by speaker.

Fig. 3: T-SNE visualization of emotion embedding. The difference between (a) and (b) is that they are colored by different attributes.

C. Visual analysis of emotional representation

We further visualize We further visualize the coarse-grained emotional representation via T-SNE [73]. Specifically, we preserve 100 utterances per emotion in the Chinese training speech data and 600 in English.

Figure 3(a) presents the T-SNE visualization of the emotion embeddings for Chinese utterances, demonstrating clear clusters. This observation validates the effectiveness of our semi-supervised emotion classifier. However, in Figure 3(a), we notice that certain emotion embeddings of English utterances are intermixed with those of Chinese utterances. We hypothesize that the language-agnostic nature of the coarsegrained emotion representation enables it to capture subtle emotional expressions in English utterances, resulting in their clustering alongside the Chinese utterances. This intermixed phenomenon of coarse-grained emotion representation signifies the METTS family's ability to transfer emotions across languages.

To further explore the extent to which the emotional representation encompasses the speaker's timbre attribute, we color the T-SNE visualization based on speaker ID. Figure 3(b) illustrates that the emotion embeddings are not well clustered according to the speaker, providing evidence of the speaker's independence in the coarse-grained emotion representation. This finding reinforces the effectiveness of our approach in disentangling speaker characteristics and isolating them from the emotional representation.

VI. COMPONENT ANALYSIS

In Section V, we demonstrate the excellent performance of METTS in both intra- and cross-lingual scenarios of emotional speech synthesis. In this section, we aim to evaluate the effectiveness of each component by examining their impact on naturalness, speaker similarity, and emotion similarity. Additionally, we analyze the influence of different values of clusters on the performance of METTS-ID.

A. Ablation study of METTS-REF

We conduct ablation studies where the GST module, CVAE module, and perturb module are removed individually. The corresponding results are presented in Table VI and Table VII, respectively.

The removal of the GST module significantly affects the control of global emotional expression in bilingual speech. Without GST, METTS fails to map emotional expressions of different languages to the same global token and provide global emotion conditions. As a result, there is a significant decrease in emotion similarity and noticeable declines in naturalness and speaker similarity. This highlights the crucial role of the coarse-grained language-agnostic emotion representation in our approach.

TABLE VIII: Results of different values of N on model's performance with 95% confidence interval for Chinese speakers.

			Chinese Text		English Text			
N	Accuracy	Naturalness	Speaker Similarity	Emotion Similarity	Accuracy	Naturalness	Speaker Similarity	Emotion Similarity
32 64 96	0.916 0.854 0.656	3.91±0.13 4.07±0.13 4.00±0.12	3.60±0.15 3.88±0.16 3.76±0.14	3.51±0.19 3.95±0.13 3.68±0.16	0.920 0.875 0.664	3.81±0.13 4.06±0.18 3.86±0.14	3.60±0.15 3.77±0.20 3.63±0.17	2.95±0.21 3.24±0.22 2.98±0.20

TABLE IX: Results of different values of N on model's performance with 95% confidence interval for English speakers.

Chinese Text			English Text					
Ν	Accuracy	Naturalness	Speaker Similarity	Emotion Similarity	Accuracy	Naturalness	Speaker Similarity	Emotion Similarity
32	0.916	3.80±0.17	3.62±0.17	3.57±0.19	0.920	3.76±0.18	3.65±0.18	3.21±0.17
64	0.854	4.02±0.15	3.57±0.21	3.73±0.17	0.875	4.05±0.18	3.74±0.18	3.26±0.14
96	0.656	3.90 ± 0.20	3.72±0.15	3.64±0.17	0.664	3.74±0.16	3.52±0.18	3.10 ± 0.22

Furthermore, when the CVAE module is removed, there is a sharp decline in naturalness and a decrease in emotion similarity. This indicates that the fine-grained emotional representation learned by the CVAE module, which is consistent with the input text, not only enhances the emotional expression but also plays a vital role in addressing the foreign accent problem and improving the overall naturalness of the synthetic speech.

Regarding the perturbation module, its omission slightly increased emotion similarity in most test conditions. However, it significantly compromised naturalness and speaker similarity. This trade-off suggests a substantial entanglement between speaker timbre, emotion, and language in multilingual emotional speech synthesis. Speaker timbre entangled with language may lead to abnormal pronunciation, while speaker timbre entangled with emotion may result in slightly high emotional expressiveness but low speaker similarity. Therefore, the necessity of speaker disentanglement becomes apparent to achieve idiomatic pronunciation and natural emotional expression for each speaker.

B. Ablation study of METTS-ID

Given the significance of the codebook size in Vector Quantization (VQ), we investigate the impact of different values of N in the emotion matcher module on the performance of METTS-ID. Alongside evaluating naturalness, speaker similarity, and emotion similarity, we also examine the accuracy of the emotion matcher. For analysis, we retain 100 utterances per emotion in Chinese and 600 in English.

We first evaluate the accuracy of the emotion matcher by extracting the ground-truth cluster labels for each utterance and calculating the predicted accuracy. The results, presented in Table VIII and Table IX, demonstrate that as the value of N increases, there is an increase in the diversity of emotion embeddings. In contrast, the predicted accuracy of the emotion matcher gradually decreases. This indicates a complex trade-off between the diversity of emotion embeddings and the predicted accuracy of the emotion matcher. Notably, the predicted accuracy remains consistent across target speakers and languages, ensuring that METTS-ID can generate natural and emotionally expressive bilingual speech for each target speaker.

Furthermore, as shown in Table VIII and Table IX, the effect of N on speaker similarity is negligible, while naturalness and emotion similarity achieve their highest scores when N is set to 64. Therefore, considering the overall performance, we designate N as 64 to strike a balance between predicted accuracy, naturalness, speaker similarity, and emotion similarity.

VII. CONCLUSION

This paper proposes METTS for multilingual emotional speech synthesis, aiming at achieving natural and diverse bilingual emotional speech across speakers. First, we introduce multi-scale emotion modeling to learn emotional expressions from a language-agnostic emotion representation (coarsegrained) and a language-specific emotion representation (finegrained), effectively addressing the foreign accent problem. Meanwhile, we leverage information perturbation to address the problem of speaker timbre coupling and obtain speakerindependent multi-scale emotion representation. Moreover, we design a VQ-based emotion matcher to construct an embedding-candidate pool and select appropriate references according to the input text and emotion category for better emotional diversity and the naturalness of synthetic speech. English-Chinese bilingual experiments show that METTS can synthesize expressive bilingual speech with natural emotion and native pronunciation for each mono-lingual speaker.

During our investigation of the multilingual emotional TTS system through cross-speaker cross-lingual emotion transfer, we have identified the need for further improvements in synthesizing English emotional speech for both Chinese and English speakers. This is mainly because the English training corpus is mainly neutral in our study. We believe that leveraging emotional English corpus to train METTS will effectively improve the expressiveness of English synthetic speech in multilingual emotional text-to-speech.

REFERENCES

[1] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August* 20-24, 2017, F. Lacerda, Ed. ISCA, 2017, pp. 4006–4010.

- [2] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "Fastspeech: Fast, robust and controllable text to speech," in Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 3165–3174.
- [3] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019.* AAAI Press, 2019, pp. 6706–6713.
- [4] M. Chen, X. Tan, B. Li, Y. Liu, T. Qin, S. Zhao, and T. Liu, "Adaspeech: Adaptive text to speech for custom voice," in 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021.
- [5] M. Jeong, H. Kim, S. J. Cheon, B. J. Choi, and N. S. Kim, "Diff-tts: A denoising diffusion model for text-to-speech," in *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, and P. Motlícek, Eds. ISCA, 2021, pp. 3605–3609.
- [6] R. J. Weiss, R. J. Skerry-Ryan, E. Battenberg, S. Mariooryad, and D. P. Kingma, "Wave-tacotron: Spectrogram-free end-to-end text-to-speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021.* IEEE, 2021, pp. 5679–5683.
- [7] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event,* ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 5530–5540.
- [8] Y. Lei, S. Yang, X. Wang, and L. Xie, "Msemotts: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis," *IEEE* ACM Trans. Audio Speech Lang. Process., vol. 30, pp. 853–864, 2022.
- [9] Y. Ren, X. Tan, T. Qin, Z. Zhao, and T. Liu, "Revisiting over-smoothness in text to speech," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, *ACL 2022, Dublin, Ireland, May 22-27, 2022*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Association for Computational Linguistics, 2022, pp. 8197–8213.
- [10] Q. Wu, Q. Shen, J. Luan, and Y. Wang, "MSDTRON: A high-capability multi-speaker speech synthesis system for diverse data using characteristic information," in *IEEE International Conference on Acoustics, Speech* and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022. IEEE, 2022, pp. 6327–6331.
- [11] 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 *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, G. Kubin and Z. Kacic, Eds. ISCA, 2019, pp. 2080– 2084.
- [12] D. Min, D. B. Lee, E. Yang, and S. J. Hwang, "Meta-stylespeech : Multi-speaker adaptive text-to-speech generation," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 7748–7759.
- [13] S. Liu, S. Yang, D. Su, and D. Yu, "Referee: Towards reference-free cross-speaker style transfer with low-quality data for expressive speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022.* IEEE, 2022, pp. 6307–6311.
- [14] Y. Yan, X. Tan, B. Li, T. Qin, S. Zhao, Y. Shen, and T. Liu, "Adaspeech 2: Adaptive text to speech with untranscribed data," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP* 2021, Toronto, ON, Canada, June 6-11, 2021. IEEE, 2021, pp. 6613– 6617.
- [15] T. Saeki, K. Tachibana, and R. Yamamoto, "Drspeech: Degradationrobust text-to-speech synthesis with frame-level and utterance-level acoustic representation learning," in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022, H. Ko and J. H. L. Hansen,* Eds. ISCA, 2022, pp. 793–797.

- [16] Y. Lei, S. Yang, X. Zhu, L. Xie, and D. Su, "Cross-speaker emotion transfer through information perturbation in emotional speech synthesis," *IEEE Signal Process. Lett.*, vol. 29, pp. 1948–1952, 2022.
- [17] G. Zhang, Y. Qin, W. Zhang, J. Wu, M. Li, Y. Gai, F. Jiang, and T. Lee, "iemotts: Toward robust cross-speaker emotion transfer and control for speech synthesis based on disentanglement between prosody and timbre," *CoRR*, vol. abs/2206.14866, 2022.
- [18] Y. Meng, X. Li, Z. Wu, T. Li, Z. Sun, X. Xiao, C. Sun, H. Zhan, and H. Meng, "CALM: constrastive cross-modal speaking style modeling for expressive text-to-speech synthesis," in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022,* H. Ko and J. H. L. Hansen, Eds. ISCA, 2022, pp. 5533–5537.
- [19] Y. Liu, Z. Xu, G. Wang, K. Chen, B. Li, X. Tan, J. Li, L. He, and S. Zhao, "Delightfultts: The microsoft speech synthesis system for blizzard challenge 2021," *CoRR*, vol. abs/2110.12612, 2021.
- [20] A. Gulati, J. Qin, C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020,* H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 5036–5040.
- [21] K. Dai, H. J. Fell, and J. MacAuslan, "Comparing emotions using acoustics and human perceptual dimensions," in *Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI* 2009, Extended Abstracts Volume, Boston, MA, USA, April 4-9, 2009, D. R. O. Jr., R. B. Arthur, K. Hinckley, M. R. Morris, S. E. Hudson, and S. Greenberg, Eds. ACM, 2009, pp. 3341–3346.
- [22] B. W. Schuller, R. Müller, M. K. Lang, and G. Rigoll, "Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles," in *INTERSPEECH 2005 - Eurospeech*, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005. ISCA, 2005, pp. 805–808.
- [23] M. Kotti and F. Paternò, "Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema," *Int. J. Speech Technol.*, vol. 15, no. 2, pp. 131–150, 2012.
- [24] E. Fersini, E. Messina, G. Arosio, and F. Archetti, "Audio-based emotion recognition in judicial domain: A multilayer support vector machines approach," in *Machine Learning and Data Mining in Pattern Recognition, 6th International Conference, MLDM 2009, Leipzig, Germany, July 23-25, 2009. Proceedings*, ser. Lecture Notes in Computer Science, P. Perner, Ed., vol. 5632. Springer, 2009, pp. 594–602.
- [25] Y. Wang, D. Stanton, Y. Zhang, R. J. Skerry-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 *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, ser. Proceedings of Machine Learning Research, J. G. Dy and A. Krause, Eds., vol. 80. PMLR, 2018, pp. 5167–5176.
- [26] F. Han, "Pronunciation problems of chinese learners of english." OR-TESOL Journal, vol. 30, pp. 26–30, 2013.
- [27] F. Li, "Contrastive study between pronunciation chinese 11 and english 12 from the perspective of interference based on observations in genuine teaching contexts." *English Language Teaching*, vol. 9, no. 10, pp. 90– 100, 2016.
- [28] S. Sitaram, S. K. Rallabandi, S. Rijhwani, and A. W. Black, "Experiments with cross-lingual systems for synthesis of code-mixed text," in *The 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA, USA, 13-15 September 2016.* ISCA, 2016, pp. 76–81.
- [29] L. Xue, W. Song, G. Xu, L. Xie, and Z. Wu, "Building a mixed-lingual neural TTS system with only monolingual data," in *Interspeech 2019*, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019, G. Kubin and Z. Kacic, Eds. ISCA, 2019, pp. 2060–2064.
- [30] E. Nachmani and L. Wolf, "Unsupervised polyglot text-to-speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019.* IEEE, 2019, pp. 7055–7059.
- [31] B. Li and H. Zen, "Multi-language multi-speaker acoustic modeling for LSTM-RNN based statistical parametric speech synthesis," in *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12,* 2016, N. Morgan, Ed. ISCA, 2016, pp. 2468–2472.
- [32] Y. Peng and Z. Ling, "Decoupled pronunciation and prosody modeling in meta-learning-based multilingual speech synthesis," in *Interspeech 2022*, 23rd Annual Conference of the International Speech Communication

Association, Incheon, Korea, 18-22 September 2022, H. Ko and J. H. L. Hansen, Eds. ISCA, 2022, pp. 4257–4261.

- [33] D. Xin, Y. Saito, S. Takamichi, T. Koriyama, and H. Saruwatari, "Crosslingual text-to-speech synthesis via domain adaptation and perceptual similarity regression in speaker space," in *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020, H. Meng, B. Xu,* and T. F. Zheng, Eds. ISCA, 2020, pp. 2947–2951.
- [34] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021.* IEEE, 2021, pp. 6608–6612.
- [35] T. Nekvinda and O. Dusek, "One model, many languages: Meta-learning for multilingual text-to-speech," in *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020,* H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 2972–2976.
- [36] 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, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September* 2021, H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, and P. Motlícek, Eds. ISCA, 2021, pp. 1619–1623.
- [37] D. Rattcliffe, Y. Wang, A. Mansbridge, P. Karanasou, A. Moinet, and M. Cotescu, "Cross-lingual style transfer with conditional prior VAE and style loss," in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, H. Ko and J. H. L. Hansen, Eds. ISCA, 2022, pp. 4586–4590.
- [38] J. Ye, H. Zhou, Z. Su, W. He, K. Ren, L. Li, and H. Lu, "Improving cross-lingual speech synthesis with triplet training scheme," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022.* IEEE, 2022, pp. 6072–6076.
- [39] J. Kim, H. Yang, Y. Ju, I. Kim, and B. Kim, "Crossspeech: Speakerindependent acoustic representation for cross-lingual speech synthesis," *CoRR*, vol. abs/2302.14370, 2023.
- [40] T. Li, X. Wang, Q. Xie, Z. Wang, and L. Xie, "Cross-speaker emotion disentangling and transfer for end-to-end speech synthesis," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 1448–1460, 2022.
- [41] P. Wu, J. Pan, C. Xu, J. Zhang, L. Wu, X. Yin, and Z. Ma, "Cross-speaker emotion transfer based on speaker condition layer normalization and semi-supervised training in text-to-speech," *CoRR*, vol. abs/2110.04153, 2021.
- [42] S. An, Z. Ling, and L. Dai, "Emotional statistical parametric speech synthesis using lstm-rnns," in 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2017, Kuala Lumpur, Malaysia, December 12-15, 2017. IEEE, 2017, pp. 1613–1616.
- [43] Y. Lee, A. Rabiee, and S. Lee, "Emotional end-to-end neural speech synthesizer," CoRR, vol. abs/1711.05447, 2017.
- [44] R. J. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *Proceedings* of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, ser. Proceedings of Machine Learning Research, J. G. Dy and A. Krause, Eds., vol. 80. PMLR, 2018, pp. 4700–4709.
- [45] K. Akuzawa, Y. Iwasawa, and Y. Matsuo, "Expressive speech synthesis via modeling expressions with variational autoencoder," in *Interspeech* 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018, B. Yegnanarayana, Ed. ISCA, 2018, pp. 3067–3071.
- [46] T. Li, S. Yang, L. Xue, and L. Xie, "Controllable emotion transfer for end-to-end speech synthesis," in 12th International Symposium on Chinese Spoken Language Processing, ISCSLP 2021, Hong Kong, January 24-27, 2021. IEEE, 2021, pp. 1–5.
- [47] D. Wang, L. Deng, Y. T. Yeung, X. Chen, X. Liu, and H. Meng, "VQMIVC: vector quantization and mutual information-based unsupervised speech representation disentanglement for one-shot voice conversion," in *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021, H. Hermansky, H. Cernocký, L. Burget, L. Lamel,* O. Scharenborg, and P. Motlícek, Eds. ISCA, 2021, pp. 1344–1348.

- 846–850.
 [49] M. Whitehill, S. Ma, D. McDuff, and Y. Song, "Multi-reference neural TTS stylization with adversarial cycle consistency," in *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020,* H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 4442–4446.
- [50] H. Choi, J. Lee, W. Kim, J. Lee, H. Heo, and K. Lee, "Neural analysis and synthesis: Reconstructing speech from self-supervised representations," in Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual, M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021, pp. 16 251–16 265.
- [51] C. H. Chan, K. Qian, Y. Zhang, and M. Hasegawa-Johnson, "Speech-split2.0: Unsupervised speech disentanglement for voice conversion without tuning autoencoder bottlenecks," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022.* IEEE, 2022, pp. 6332–6336.
- [52] R. Huang, Y. Ren, J. Liu, C. Cui, and Z. Zhao, "Generspeech: Towards style transfer for generalizable out-of-domain text-to-speech synthesis," *CoRR*, vol. abs/2205.07211, 2022.
- [53] J. Ye, X. Wen, X. Wang, Y. Xu, Y. Luo, C. Wu, L. Chen, and K. Liu, "Gm-tcnet: Gated multi-scale temporal convolutional network using emotion causality for speech emotion recognition," *CoRR*, vol. abs/2210.15834, 2022.
- [54] J. R. Chang, A. Shrivastava, H. Koppula, X. Zhang, and O. Tuzel, "Style equalization: Unsupervised learning of controllable generative sequence models," in *International Conference on Machine Learning, ICML* 2022, 17-23 July 2022, Baltimore, Maryland, USA, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 2022, pp. 2917–2937.
- [55] A. T. Sigurgeirsson and S. King, "Do prosody transfer models transfer prosody?" CoRR, vol. abs/2303.04289, 2023.
- [56] C. Gong, L. Wang, Z. Ling, J. Zhang, and J. Dang, "Using multiple reference audios and style embedding constraints for speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022.* IEEE, 2022, pp. 7912–7916.
- [57] Y. Yi, L. He, S. Pan, X. Wang, and Y. Xiao, "Prosodyspeech: Towards advanced prosody model for neural text-to-speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP* 2022, Virtual and Singapore, 23-27 May 2022. IEEE, 2022, pp. 7582– 7586.
- [58] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 8748–8763.
- [59] J. Kim, S. Lee, J. Lee, H. Jung, and S. Lee, "GC-TTS: few-shot speaker adaptation with geometric constraints," in 2021 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2021, Melbourne, Australia, October 17-20, 2021. IEEE, 2021, pp. 1172–1177.
- [60] I. Yoo, H. Lim, and D. Yook, "Formant-based robust voice activity detection," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 12, pp. 2238–2245, 2015.
- [61] 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. NIPS*, 2017, pp. 5998–6008.
- [62] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.
- [63] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, vol. 99-D, no. 7, pp. 1877–1884, 2016.
- [64] K. Sjölander, "An hmm-based system for automatic segmentation and alignment of speech," 2003.
- [65] 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," in Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 14 881–14 892.

- [66] Z. Shang, Z. Huang, H. Zhang, P. Zhang, and Y. Yan, "Incorporating cross-speaker style transfer for multi-language text-to-speech." in *Inter-speech*, 2021, pp. 1619–1623.
- [67] T. Li, X. Wang, Q. Xie, Z. Wang, and L. Xie, "Cross-speaker emotion disentangling and transfer for end-to-end speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1448–1460, 2022.
- [68] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Interspeech 2020, 21st Annual Conference* of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 3830–3834.
- [69] Z. Yao, D. Wu, X. Wang, B. Zhang, F. Yu, C. Yang, Z. Peng, X. Chen, L. Xie, and X. Lei, "Wenet: Production oriented streaming and nonstreaming end-to-end speech recognition toolkit," in *Interspeech 2021*, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021, H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, and P. Motlícek, Eds. ISCA, 2021, pp. 4054–4058.
- [70] G. Chen, S. Chai, G. Wang, J. Du, W. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang, M. Jin, S. Khudanpur, S. Watanabe, S. Zhao, W. Zou, X. Li, X. Yao, Y. Wang, Z. You, and Z. Yan, "Gigaspeech: An evolving, multi-domain ASR corpus with 10, 000 hours of transcribed audio," in *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August 3 September 2021*, H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, and P. Motlícek, Eds.
- [71] B. Zhang, H. Lv, P. Guo, Q. Shao, C. Yang, L. Xie, X. Xu, H. Bu, X. Chen, C. Zeng, D. Wu, and Z. Peng, "WENETSPEECH: A 10000+ hours multi-domain mandarin corpus for speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022.* IEEE, 2022, pp. 6182–6186.
- [72] D. Acuna, M. T. Law, G. Zhang, and S. Fidler, "Domain adversarial training: A game perspective," in *The Tenth International Conference* on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net, 2022.
- [73] L. van der Maaten and G. E. Hinton, "Visualizing data using t-sne," vol. 9, 2008, pp. 2579–2605.