

# DiCLET-TTS: Diffusion Model based Cross-lingual Emotion Transfer for Text-to-Speech — A Study between English and Mandarin

Tao Li, Chenxu Hu, Jian Cong, Xinfu Zhu, Jingbei Li, Qiao Tian, Yuping Wang, Lei Xie, *Senior Member, IEEE*

**Abstract**—While the performance of cross-lingual TTS based on monolingual corpora has been significantly improved recently, generating cross-lingual speech still suffers from the foreign accent problem, leading to limited naturalness. Besides, current cross-lingual methods ignore modeling emotion, which is indispensable paralinguistic information in speech delivery. In this paper, we propose DiCLET-TTS, a Diffusion model based Cross-Lingual Emotion Transfer method that can transfer emotion from a source speaker to the intra- and cross-lingual target speakers. Specifically, to relieve the foreign accent problem while improving the emotion expressiveness, the terminal distribution of the forward diffusion process is parameterized into a speaker-irrelevant but emotion-related linguistic prior by a prior text encoder with the emotion embedding as a condition. To address the weaker emotional expressiveness problem caused by speaker disentanglement in emotion embedding, a novel orthogonal projection based emotion disentangling module (OP-EDM) is proposed to learn the speaker-irrelevant but emotion-discriminative embedding. Moreover, a condition-enhanced DPM decoder is introduced to strengthen the modeling ability of the speaker and the emotion in the reverse diffusion process to further improve emotion expressiveness in speech delivery. Cross-lingual emotion transfer experiments show the superiority of DiCLET-TTS over various competitive models and the good design of OP-EDM in learning speaker-irrelevant but emotion-discriminative embedding.

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

## I. INTRODUCTION

CROSS-lingual text-to-speech (TTS) [1], [2], [3] refers to the task that requires the system to generate speech in a language foreign to a target speaker. This task has many applications, such as code-mixed speech synthesis for a voice agent, foreign movie dubbing [4], and computer-assisted pronunciation teaching [5]. Due to the difficulty of obtaining a bilingual corpus produced by a highly proficient speaker

in both languages, more practically, current studies mainly build a cross-lingual TTS system based on corpora from monolingual speakers in different languages [6], [7], [8], [9]. However, these approaches mostly ignore modeling emotion aspects during speech generation, while emotion is a kind of indispensable paralinguistic information that reveals the speaker's intentions and moods. Without properly delivering such paralinguistic information, the gap between synthetic and real speech cannot be mitigated. This paper aims to address this emotional speech synthesis problem in cross-lingual TTS with only monolingual corpora available. Specifically, a *cross-lingual emotion transfer* method in the same-gender scenario is introduced. With cross-lingual emotion transfer, a cross-lingual TTS model can directly synthesize emotionally diverse speech in a language foreign to the target speaker, i.e., synthesizing emotional speech in authentic Mandarin for an English speaker by transferring the emotion from a Mandarin speaker, without employing any Mandarin emotional speech from the English speaker during system building.

Although the current studies have made many efforts to cross-lingual TTS, there is still a gap between generated speech and those of native speakers in terms of naturalness, as synthetic speech often comes with a strong *foreign accent* [1]. The reason for this phenomenon is that each speaker in the training set speaks only one language, and the entanglement between different speech factors, such as linguistic content, speaker identity, and emotion, makes it hard to only transfer the speaker's timbre across different languages. Therefore, the key to alleviating this foreign accent issue is how to properly *disentangle* the speaker and language or linguistic content [8], [9], [10].

Based on the disentanglement strategy, the existing cross-lingual approaches can be roughly divided into *implicit-based* and *explicit-based* methods [9]. Implicit-based methods mainly study the unified linguistic/phonetic representations across languages to disentangle language and speaker timbre implicitly [11], [12], [13], [14], [15], [16]. On the other hand, to further solve the foreign accent problem, the explicit-based methods prefer to adopt adversarial learning [1], [7], [9], [17] or mutual information [6] to minimize the correlation between different speech factors, thus encouraging the model to automatically learn disentangled linguistic representations. However, the disturbance caused by adversarial learning could degrade the naturalness of the generated cross-lingual speech. Furthermore, the above cross-lingual studies have not considered the emotion factor yet, while proper emotion is essential to speech expressiveness, as just mentioned.

This work was supported by the National Key Research and Development Program of China under Grant 2020AAA0108600. (Corresponding author: Lei Xie)

Tao Li, Xinfu Zhu, and Lei Xie is with the School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China. Email: taoli@npu-aslp.org (Tao Li), xfzhu@mail.nwpu.edu.cn (Xinfu Zhu), lxie@nwpu.edu.cn (Lei Xie)

Chenxu Hu is with the Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing 100084, China, Email: chenxuhu65@gmail.com

Jingbei Li is with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China, Email: lijib19@mails.tsinghua.edu.cn

Jian Cong, Qiao Tian, and Yuping Wang are with the Speech, Audio, and Music Intelligence (SAMI) Group, ByteDance, Shanghai 200233, China. Email: congjian.tts@bytedance.com (Jian Cong), tian-qiao.wave@bytedance.com (Qiao Tian), wangyuping@bytedance.com (Yuping Wang)

To improve the emotion diversity of synthetic cross-lingual speech, we need to implement a cross-lingual TTS model with the ability of cross-speaker emotion transfer as well, which can produce emotional speech for target speakers by transferring the emotion from another source speaker [18]. Reference-based style transfer is the most popular strategy for cross-speaker emotion transfer, where Reference Encoder [19], Global Style Tokens (GST) [20], and Variational Auto-Encoder (VAE) [21] are typically used to extract an emotion embedding from the reference Mel-spectrum with desired emotion. Usually, the speaker identity can be obtained from either a trainable look-up table [22] or a pre-trained speaker verification model [23].

The key to the reference-based methods is to learn speaker-irrelevant emotion embedding from the reference spectrum by disentangling the emotion and the speaker's timbre [24], [25], [18]. Otherwise, the speaker information retained in the emotion embedding could contaminate the target speaker's timbre, making synthesized speech sound somehow like uttered by the source speaker rather than the target speaker, i.e., the *speaker leakage* problem [26]. However, due to the emotion and the timbre being deeply entangled in speech, it is hard to remove the speaker-related information while avoiding the emotion information from being weakened in the emotion embedding, which could lead to *weaker emotional expressiveness* problem in the synthesized speech [27]. Furthermore, for cross-lingual emotion transfer, a unique challenge is that emotion will make the intonation change more violently [28] and then aggravate the influence of foreign accents, resulting in a serious decline in the naturalness of the emotional speech synthesized for foreign speakers. In this paper, we attempt to enable English speakers to express various emotions in Mandarin naturally and expressively, which is a more challenging scenario since Mandarin is a typical tonal language and English is a non-tonal language [29], [30].

Recently, diffusion probabilistic models (DPMs) [31], [32] have shown their superiority in various content generation tasks [33], [34], [35], including the recent attempts in speech generation tasks [36], [37], [38], [39]. A DPM aims to gradually transform the raw data into a terminal distribution (usually standard Gaussian) by a forward diffusion process and then learns a reverse diffusion process parameterized with a neural network to rebuild the raw data from the terminal distribution [31]. Importantly, DPMs show superiority in expressive data generation, which means they can generate more diverse data due to their ability to essentially preserve the semantic structure of the data. To leverage the advances of DPMs, this paper proposes **DiCLET-TTS**, a novel DPM-based TTS model for cross-lingual emotion transfer. DiCLET-TTS consists of a prior text encoder, an orthogonal projection based emotion disentangling module (OP-EDM), and a condition-enhanced DPM decoder.

Specifically, to relieve the *foreign accent* problem and improve emotional expressiveness, the prior text encoder aims to parameterize the terminal distribution of the forward diffusion process into a speaker-irrelevant but emotion-related linguistic prior, achieved by two steps. First, the linguistic encoding is constrained by speaker adversarial training to obtain a

speaker-irrelevant linguistic representation. A content loss is particularly adopted to mitigate the interference of adversarial training on linguistic encoding. An emotional adaptor is subsequently adopted to convert the speaker-irrelevant linguistic representation into a speaker-irrelevant but emotion-related linguistic prior with the condition of emotion embedding extracted from OP-EDM.

To address the *weaker emotional expressiveness* problem, the emotion embedding space learned in OP-EDM is explicitly constrained by an Orthogonal Projection Loss [40] to force the emotion embeddings to be aggregated within the same emotion category and orthogonal between different emotion categories, leading to a discriminative emotion embedding space and improved transferred emotion expressiveness in synthetic speech.

The reverse diffusion process is further parameterized with the DPM decoder to restore the target Mel-spectrum from the speaker-irrelevant but emotion-related terminal distribution. We particularly introduce a *condition-enhanced* decoder to further improve emotion expressiveness in speech delivery. Specifically, the decoder follows the Unet [41] structure in Grad-TTS [36], but differently, the speaker and emotion embeddings are fed to each ResBlock as *enhanced conditions*.

During the experimental evaluation, different emotions are transferred from the source speaker to the intra- and cross-lingual target speakers, respectively, to verify the effectiveness of DiCLET-TTS while comparing the performance difference between intra- and cross-lingual emotion transfer. Results show that although the performance of intra-lingual transfer is better than that of more challenging cross-lingual transfer, DiCLET-TTS can clearly improve speech naturalness, emotion similarity, and speaker similarity compared to three competitive methods in both intra- and cross-lingual emotion transfer scenarios. Furthermore, the embedding visualization and preference test demonstrates the advantages of OP-EDM in learning speaker-irrelevant but emotion-discriminative embedding.

The rest of this paper is organized as follows. Section II reviews the related work. Section III introduces the proposed method in detail. Section IV and Section V describe the experimental setups and results, respectively. The component analysis is introduced in Section VI. Finally, the paper concludes in Section VII. Examples of synthesized speech can be found on the project page<sup>1</sup>.

## II. RELATED WORK

This section describes related studies on cross-lingual, cross-speaker emotion transfer, and recent DPM-based TTS.

### A. Cross-lingual TTS

Most current studies realize cross-lingual TTS by mixing monolingual corpora of different languages while disentangling the speaker and language or linguistic representations in implicit or explicit ways to alleviate the foreign accent problem. Implicit methods mainly focus on exploring language-irrelevant input representations [11], [12], [14], [15]. Liu et

<sup>1</sup>The demo can be found on <https://silyfox.github.io/DiCLETdemo/>

al. [42] introduce a shared phoneme set for different languages. Language embedding is extended by tone/stress embeddings to control the accent of synthetic speech. In [11], [12], the Automatic Speech Recognition (ASR) models are employed to extract language-irrelevant Phonetic Posterior Gram (PPG) features as the input representations. Unicode bytes [13], mixed-lingual Grapheme-to-Phoneme (G2P) [14] frontend, and International Phonetic Alphabet (IPA) [43], [15], [16] are also taken as the unified phonetic representations that share pronunciation across languages [9]. These studies indicate that language-irrelevant representations can help disentangle speaker and language, but the complexity of the cross-lingual TTS pipeline is increased.

The explicit methods encourage the cross-lingual model to automatically learn disentangled representation, i.e., speaker-irrelevant linguistic representations or language-irrelevant speaker representations. Zhang et al. [1] and Nekvinda et al. [17] employ domain adversarial training to remove speaker identity entangled in linguistic representations. Xin et al. [7] construct a language-irrelevant speaker space via domain adaptation and perceptual similarity regression. In [6], mutual information minimization and domain adversarial training are adopted to disentangle the obtained language and speaker embedding, which guides cross-lingual speech synthesis. Ye et al. [9] introduce a triplet training scheme to enhance cross-lingual pronunciation by allowing previously unseen content and speaker combinations to be seen during training. Shang et al. [8] alleviate the foreign accent problem by using existing authentic style during inference and accordingly propose a style encoder through adversarial training. The above studies mainly address the foreign accent problem, while emotional speech is not considered.

### B. Cross-speaker emotion transfer

Cross-speaker emotion transfer in TTS shares similar methods with other kinds of style transfer, as emotions are expressed in a special style. For clarity, all of them are referred to as emotion transfer. Currently, there are mainly two major approaches for cross-speaker emotion transfer, i.e., label-assisted and reference-based methods. Label-assisted [44], [45] methods are proposed to predict emotion-related prosodic information, i.e., pitch and energy, from input text with speaker and emotion ID. However, since prosodic information contained in text lack residual acoustic information other than pitch and energy, these methods are prone to produce synthesized speech with average expressiveness.

The reference-based methods [19], [46], [21], [47], [48], [49] is the mainstream strategy, which learns an emotion representation [50] from reference as a condition to guide emotion transfer. Skerry-Ryan et al. [19] integrate the Tacotron [51], [52] model with an extra prosody encoder, denoted as Reference Encoder, in which the reference is encapsulated into a fixed-length embedding that is directly concatenated with the linguistic representations. Global Style Tokens (GST) [46] further extends Reference Encoder by an embedded library to learn a latent high-dimensional representation. Variational Auto-Encoder [21] is also introduced to learn the potential emotion representation from the reference to complete

emotion transfer. However, these methods ignore speaker disentanglement and aggregate all emotion-related aspects, e.g., pitch, energy, and speaker's timbre, into one hidden emotion embedding, resulting in speaker leakage. To achieve speaker disentanglement, Bian et al. [24] propose a multi-reference encoder and an intercross training scheme in which emotion and speaker are disentangled and transferred independently. Whitehill et al. [25] improve the performance of the multi-reference model on disjoint datasets by unpaired training strategy and adversarial cycle consistency scheme. Li et al. [18] introduce an emotion disentangling module, which constrains the emotion embedding to be speaker-irrelevant via an orthogonal loss with the learned speaker embedding. To summarize, the aforementioned methods mainly aim to obtain a speaker-irrelevant emotion embedding space in different ways, while the trade-off between speaker timbre and emotional expressiveness is inevitable [26], [27].

### C. DPM-based TTS

The Diffusion Probabilistic Models (DPMs) aim to convert the raw data distribution into random noise before reversing the transformations step by step to rebuild a new sample with the same distribution as the raw data [31], [53] and have achieved the SOTA results in various tasks, e.g., image generation [54], [33], super-resolution [35], [34], and TTS [36], [37], [38], [55]. One major drawback of DPM-based models is the slow sampling speed due to many iterative steps. Therefore, many previous DPM-based TTS methods focus on accelerating the sampling method to boost the inference speed [38], [39], [56], [57], [58]. Some research considers changing the training process to generate high-quality speech. Grad-TTS [36] and PriorGrad [59] transform the raw data distribution into a data-dependent prior distribution obtained from the conditional information. The studies [60], [61] also drive unconditional DPM-based models trained on untranscribed speech to generate high-quality samples by phoneme classifier guidance, where the phoneme classifier is trained separately. Liu et al. [62] and Xue et al. [63] introduce the DPM-based model into singing voice synthesis (SVS), demonstrating its superiority in the expressiveness synthesis tasks. In this study, we introduce the DPM-based model to cross-lingual emotion transfer TTS, a more challenging and unexplored task.

## III. METHODOLOGY

This section first gives a system overview of the proposed DiCLET-TTS and then introduces the design of each module in detail.

Figure 1 illustrates the proposed DiCLET-TTS architecture for cross-lingual emotion transfer, a DPM-based TTS model consisting of three major components: a prior text encoder, an orthogonal projection based emotion disentangling module (OP-EDM), and a condition-enhanced DPM decoder. As discussed, the entangled speaker and linguistic representation can lead to a *foreign accent* problem. Thus, the prior text encoder is adopted to parameterize the forward diffusion's terminal distribution as a speaker-irrelevant but emotion-related linguistic prior, to mitigating the foreign accent while improving

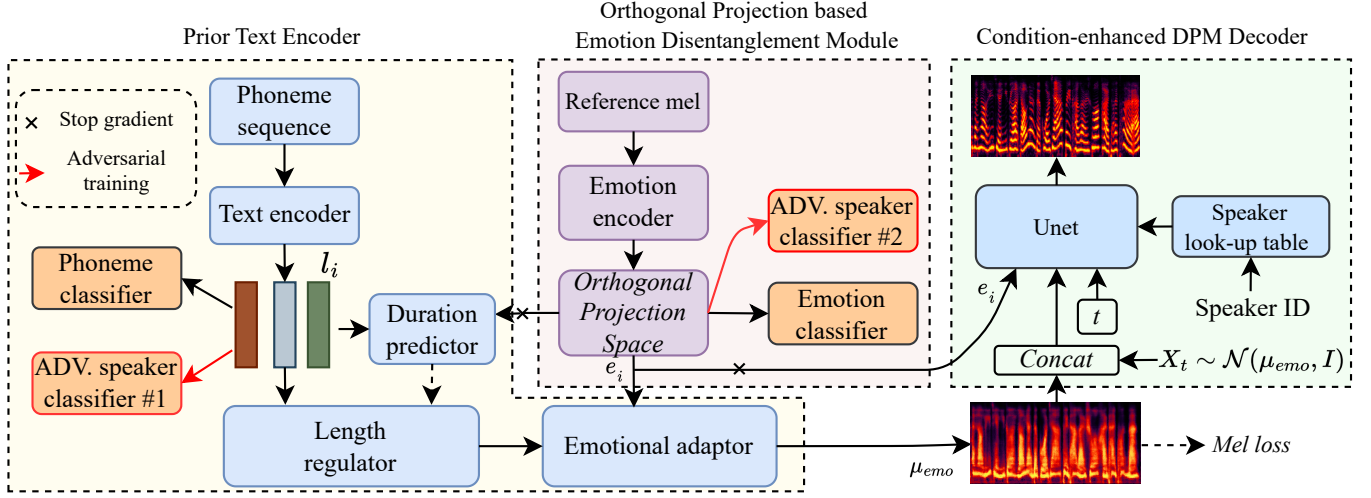


Fig. 1: The architecture of the proposed DiCLET-TTS. The input text is represented as the phoneme sequence, and speech is represented by Mel-spectrum, which can be converted to the waveform by a Hifi-Gan vocoder.

emotional expression. The speaker identity is only modeled by a speaker look-up table with speaker ID in the DPM decoder to further disentangle the speaker from other factors. Considering that the speaker disentanglement in emotion embedding could lead to *weaker emotional expressiveness* in synthesized speech, our disentangled emotion embedding space is further constrained by an introduced orthogonal projection loss to ensure that the embedding maintains intense emotion discrimination after removing speaker-related information. Finally, a condition-enhanced DPM decoder is adopted to restore the target Mel-spectrum from the speaker-irrelevant but emotion-related terminal distribution, guided by the speaker and emotion embeddings.

#### A. Prior text encoder

The prior text encoder consists of a text encoder, a length regulator, and an emotional adaptor, aiming to parameterize the terminal distribution of the forward diffusion process into a speaker-irrelevant but emotion-related linguistic prior. Specifically, to remove the speaker information from the linguistic representation  $l_i$ , the text encoder is encouraged to encode input phonemes in a speaker-irrelevant manner by introducing a speaker adversarial classifier. Then, we encode the length-regulated  $l_i$  through an emotional adaptor under the condition of emotion embedding (will be introduced in Section III-B) to obtain the speaker-irrelevant but emotion-related linguistic representation  $\mu_{emo}$ . However, adversarial training could disturb linguistic encoding to some extent. Therefore, a content loss is introduced to mitigate this disturbance. Details of the prior text encoder and the loss functions will be introduced.

1) *Text encoder*: The text encoder converts the phoneme sequence into the hidden linguistic representation  $l_i \in \mathbb{R}^{C_i \times d}$ , where  $C_i$  denotes the length of the phoneme sequence, and  $d$  denotes the dimension of representation. The speaker adversarial classifier is to make the linguistic representation  $l_i$  speaker-irrelevant through the softmax layer with gradient reversal (GR), and the loss function is defined as:

$$\mathcal{L}_{adv} = - \sum_{i=1}^n \log P(s_i | l_i), \quad (1)$$

where  $n$  is the batch size,  $P(s_i | l_i)$  is possibility of  $l_i$  belonging to the speaker  $s_i$ . So that we can minimize the speaker classification loss to reversely optimize the text encoder on the speaker classification task.

The content loss guarantees the text encoder's stability to encode the input phoneme sequence when using the speaker adversarial classifier. The corresponding loss function is defined as:

$$\mathcal{L}_c = - \sum_{i=1}^n \sum_{j=1}^{C_i} \log P(p_i^j | l_i^j). \quad (2)$$

where  $p_i^j$  denotes the ground-truth label of the  $j$ -th phoneme in the  $i$ -th input sequence, and  $l_i^j$  denotes the  $j$ -th hidden linguistic representation of the  $i$ -th input sequence.

2) *Length regulator*: The length regulator has the same architecture as that in FastSpeech [64]. It takes emotion embedding as an extra input since the duration of the same sentence in different emotions should be different. The  $l_i$  is length-regulated according to its real duration by the length regulator during training. The duration predictor is trained by the mean square error (MSE) loss with the ground-truth duration. The duration loss is denoted as  $\mathcal{L}_{dur}$ .

3) *Emotional adaptor*: The emotional adaptor aims to transform the length-regulated  $l_i$  into a speaker-irrelevant but emotion-related linguistic representation  $\mu_{emo}$  through multiple FFT blocks with Conditional LayerNorm [65], which takes the emotion embedding as the condition. The  $\mu_{emo}$  has the same dimension as the Mel-spectrum and is adopted to define the forward diffusion's terminal distribution ( $\mathcal{N}(\mu_{emo}, I)$ ). An MSE loss  $\mathcal{L}_{mel}$  constrains the  $\mu_{emo}$  from the target Mel-spectrum. Regarding the prior text encoder, the total objective function is defined as:

$$\mathcal{L}_{prior} = 0.01 * \mathcal{L}_{adv} + \mathcal{L}_c + \mathcal{L}_{dur} + \mathcal{L}_{mel}. \quad (3)$$

### B. Orthogonal projection based emotion disentanglement module

The orthogonal projection based emotion disentanglement module (OP-EDM) is to learn an emotion encoder that extracts the speaker-irrelevant emotion embedding  $e_i$  from the reference. Ideally, the embedding  $e_i$  should be free of speaker-related information and discriminative in distinguishing different emotion categories. To this end, the emotion encoder in OP-EDM is trained with two loss functions: 1) an adversarial loss to make the obtained embedding  $e_i$  speaker-irrelevant; 2) a classification loss to make the obtained embedding  $e_i$  emotion-dependent. Specifically, the emotion encoder in OP-EDM has a similar architecture as the Reference Encoder [19] to generate a 256-dimensional vector as the emotion embedding  $e_i$ .

The adversarial loss aims to make the emotion embedding  $e_i$  speaker indistinguishable. A GRL is adopted between the emotion encoder and a speaker classifier. Then, the emotion encoder is reversely optimized on the speaker classification by minimizing the following loss function:

$$\mathcal{L}_{sadv} = - \sum_{i=1}^n \log P(s_i | e_i), \quad (4)$$

where  $P(s_i | e_i)$  is the possibility of the emotion embedding  $e_i$  extracted from speech with the speaker label  $s_i$ .

The classification loss is implemented by an emotion classifier with the same structure as the above speaker classifier, to make the obtained  $e_i$  emotion-dependent. Note that the sentences of all non-emotional speakers are treated as a separate emotion category, denoted as *neutral\_N*. Thus, the softmax layer produces the probability of 8 emotion types, i.e., *neutral*, *happy*, *surprise*, *angry*, *disgust*, *fear*, *sad*, and *neutral\_N*. The corresponding objective function is:

$$\mathcal{L}_{emo} = - \sum_{i=1}^n \log P(t_i | e_i), \quad (5)$$

where  $P(t_i | e_i)$  is possibility of emotion embedding  $e_i$  belonging to the emotion label  $t_i$ .

1) *Explicit constraint for emotion embedding space*: As mentioned, the emotional information conveyed by the  $e_i$  would be weakened after removing the speaker-related information, leading to weaker emotional expressiveness in synthesized speech. To address this issue, we resort to the Orthogonal Projection Loss [40] (OPL), a potent technique to construct discriminative embedding space without learnable parameters. The objective of OPL is to enforce constraints to embedding space such that the embedding  $e_i$  for different emotion classes  $t_i$  is orthogonal to each other and the  $e_i$  for the same class is similar, which can effectively disentangle the class-specific characteristics of different emotions, further improving the emotion discrimination of  $e_i$ . The objective function is defined as:

$$\mathcal{L}_{opl} = (1 - E_{same}) + 0.5 * |E_{different}|, \quad (6)$$

where  $|\cdot|$  is the absolute value operator. When minimizing this loss  $\mathcal{L}_{opl}$ , the first term  $(1 - E_{same})$  can ensure clustering of same class samples, while the second term  $|E_{different}|$  can

ensure the orthogonality of different class samples. The  $E_{same}$  and  $E_{different}$  are defined as:

$$E_{same} = \sum_{t_i=t_j}^n \langle \mathbf{e}_i, \mathbf{e}_j \rangle, \quad E_{different} = \sum_{t_i \neq t_k}^n \langle \mathbf{e}_i, \mathbf{e}_k \rangle, \quad (7)$$

where  $\langle \cdot, \cdot \rangle$  is the cosine similarity operator applied on two emotion embeddings. The total objective function of OP-EDM is defined as:

$$\mathcal{L}_{op-edm} = 0.2 * \mathcal{L}_{sadv} + 0.8 * \mathcal{L}_{emo} + \mathcal{L}_{opl}. \quad (8)$$

### C. Condition-enhanced DPM decoder

A DPM with data-dependent prior can be seen as such: a forward diffusion converts the raw data into simple terminal distribution (usually standard Gaussian) by gradually adding Gaussian noise, then based on this terminal distribution, a reverse diffusion parameterized with a neural network learns to follow the trajectories of the reverse-time forward diffusion [32], [55]. If the forward and reverse diffusion processes have close trajectories, then the distribution of generated samples will be very close to that of the raw data.

In DiCLET-TTS, the terminal distribution of forward diffusion has been parameterized by the prior text encoder as a simple linguistic-based prior distribution  $\mathcal{N}(\mu_{emo}, I)$ , which is emotion-related but speaker-irrelevant. We parameterize the reverse diffusion with a condition-enhanced DPM decoder to further improve the emotion expressiveness in speech delivery. Specifically, the condition-enhanced DPM decoder's architecture is based on Unet and is the same as that in Grad-TTS [36], but the speaker and emotion embeddings are added to each ResBlock rather than just concatenated with the decoder's input. The speaker and emotion embeddings are produced by a speaker look-up table and the OP-EDM, respectively.

We mostly follow the formulation introduced in Grad-TTS, the forward and reverse diffusion processes of DiCLET-TTS as satisfies the following Itô stochastic differential equations (SDEs):

$$dX_t = \frac{1}{2} \Sigma^{-1} (\mu_{emo} - X_t) \beta_t dt + \sqrt{\beta_t} d\vec{W}_t, \quad (9)$$

$$dX_t = \left( \frac{1}{2} \Sigma^{-1} (\mu_{emo} - X_t) - \nabla \log p_t(X_t | X_0) \right) \beta_t dt + \sqrt{\beta_t} d\overleftarrow{W}_t, \quad (10)$$

where the  $\vec{W}_t$  and  $\overleftarrow{W}_t$  are forward and reverse-time Brownian motion. The  $X_0$  and  $X_t$  are raw and noise data, where  $X_t \sim \mathcal{N}(\mu_{emo}, I)$ ,  $t \in [0, 1]$ . The  $\beta_t$  is a noise schedule with the same definition in Grad-TTS. The  $\log p_t(X_t | X_0)$  is the log probability density function which is predicted by a learnable score function  $s_\theta(X_t, \mu_{emo}, t, E_{spk}, e_i)$  parameterized with the condition-enhanced DPM decoder  $\theta$ . The  $E_{spk}$  and  $e_i$  are speaker and emotion embeddings, respectively. The reverse diffusion (10) is solved by a defined ordinary differential equation (ODE):

TABLE I: Dataset for the cross-lingual emotion transfer TTS.

Corpus	Gender	Language	Emotion (sentences)							Usage
			Neutral	Happy	Surprise	Sadness	Angry	Disgust	Fear	
CN1	Female	Mandarin	5k	-	-	-	-	-	-	Training&Evaluation Training
CN2	Female	Mandarin	5k	-	-	-	-	-	-	
CN-emo	Female	Mandarin	5k	2k	2k	2k	2k	2k	2k	Training&Evaluation
EN1	Female	English	9k	-	-	-	-	-	-	Training&Evaluation
EN2	Female	English	9k	-	-	-	-	-	-	Training

$$dX_t = \frac{1}{2} (\mu_{emo} - X_t - s_\theta(X_t, \mu_{emo}, t, E_{spk}, e_i)) \beta_t dt. \quad (11)$$

This reverse diffusion process is trained by minimizing weighted L2 loss as follows:

$$\begin{aligned} \mathcal{L}(\theta)_{diff} = & \arg \min_{\theta} \int_0^1 \lambda_t \mathbb{E}_{X_0, X_t} \|s_\theta(X_t, \mu_{emo}, t, E_{spk}, e_i) \\ & - \nabla \log p_t(X_t | X_0)\|_2^2 dt, \end{aligned} \quad (12)$$

where the  $\lambda_t = 1 - e^{-\int_0^t \beta_s ds}$ ,  $0 < s < t$ . In brief, the reverse diffusion parameterized with the  $s_\theta(X_t, \mu_{emo}, t, E_{spk}, e_i)$  is trained to approximate gradient of log-density of  $X_t$  given  $X_0$ ,  $E_{spk}$ ,  $e_i$  and  $\mu_{emo}$ . During the inference, we first predict a speaker-irrelevant but emotion-related  $\mu_{emo}$  from input text with the emotion embedding  $e_i$ . The  $e_i$  is extracted by OP-EDM from the reference with desired emotion. Then the condition-enhanced DPM decoder gradually reconstructs the target Mel-spectrum using the score predicted from  $s_\theta$  in adjustable iterations, with the conditions of  $e_i$  and  $E_{spk}$ .

#### D. Final objective function

All modules introduced in the previous sections are trained together. The final objective function of the proposed DiCLET-TTS is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{prior} + \mathcal{L}_{op-edm} + \mathcal{L}_{diff}, \quad (13)$$

where  $\mathcal{L}_{prior}$ ,  $\mathcal{L}_{op-edm}$ , and  $\mathcal{L}_{diff}$  are loss functions of the prior text encoder, OP-EDM, and condition-enhanced DPM decoder.

### IV. EXPERIMENTAL SETUPS

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

#### A. Dataset

As shown in Table I, the dataset used in this paper comprises five female monolingual speakers, denoted as CN1, CN2, CN-emo, EN1, and EN2. Note that CN1 and CN2 are publicly available Mandarin corpus<sup>2</sup> and EN1 and EN2 are internal English corpora. Only **CN-emo** is the emotional corpus employed as the **source speaker** during emotion transfer. All data are studio-quality recorded at 48KHz.

The test set consists of 1100 sentences in total. Specifically, we randomly select 700 sentences from the *CN-emo* corpus,

and each emotion category (including *neutral*) contains 100 sentences. In addition, 400 sentences are randomly selected from the four neutral speaker corpora, and every speaker contains 100 sentences.

#### B. Model Configurations

The text encoder has the same architecture in DelightfulTTS [66], which is composed of a pre-net (3 layers of convolutions followed by a fully-connected layer), 6 Conformer blocks [67] with multi-head self-attention, and the final linear projection layer to generate 448-dimensional linguistic representation. The speaker adversarial classifier in the text encoder consists of a GRU layer, a fully connected (FC) layer, and a softmax layer. Especially, a gradient reversal layer (GRL) is adopted between the GRU and the FC layer. The content loss is implemented by a phoneme classifier consisting of two FC layers and a softmax layer. The emotional adaptor consists of 1 layer of 1D convolution, 2 FFT blocks, and a 1D convolution output layer, where each FFT block is followed by a Conditional LayerNorm [65]. We employ the same structure for speaker and emotion classifiers in OP-EDM: an FC layer and a softmax layer. The difference is that a GRL layer is inserted before the speaker classifier.

#### C. Evaluation Methods

Three types of human perceptual rating experiments are performed: 1) Mean Opinion Score (MOS) [8] is used for subjective evaluation of the naturalness, which can reflect the influence of foreign accents and emotion on synthesized naturalness. 2) Differential Mean Opinion Scores (DMOS) [18] is adopted to subjectively evaluate the synthesized speech from two aspects, emotion similarity (between the synthesized speech and compared emotional reference) and speaker similarity (between the synthesized speech and compared target speaker's reference). 3) AB preference test [68] (AB test) is adopted to compare samples synthesized by two models, where participants are asked to choose which speech sample sounds closer to the compared reference in terms of speaker or emotion. In both MOS and DMOS tests, the participants are asked to rate given speech a score ranging from 1 to 5 based on the specific purpose. The rating criteria is: *bad* = 1; *poor* = 2; *fair* = 3; *good* = 4; *great* = 5, in 0.5 point increments.

During our experiments, we found that the results of different speakers in the same language were similar in human evaluation. Consequently, to reduce the cost of human evaluation, we randomly selected one speaker from each of the two languages, i.e., CN1 and EN1, as our target speakers without loss of generality. For MOS evaluation, 20 Mandarin and 20 English sentences are randomly selected from the test set to

<sup>2</sup>The dataset is available at [http://www.data-baker.com/hc\\_znv\\_1.html](http://www.data-baker.com/hc_znv_1.html)

TABLE II: Naturalness MOS results of DiCLET-TTS with M3, CSET, and Grad-TTS in transferring emotion to the intra- and cross-lingual target speakers, with confidence intervals of 95%. The bold indicates the best performance of the four models in each emotion.

Emotion	Language	Intra-lingual scenario (target Mandarin speaker)				Cross-lingual scenario (target English speaker)			
		M3	CSET	Grad-TTS	DiCLET-TTS	M3	CSET	Grad-TTS	DiCLET-TTS
Neutral	Mandarin	4.17±0.03	4.15±0.05	4.19±0.04	<b>4.23±0.06</b>	3.92±0.06	3.81±0.04	3.87±0.06	<b>3.98±0.05</b>
	English	3.95±0.04	3.84±0.02	3.88±0.07	<b>3.99±0.05</b>	4.18±0.04	4.14±0.06	<b>4.24±0.03</b>	4.21±0.05
Fear	Mandarin	4.05±0.05	3.92±0.08	4.03±0.04	<b>4.07±0.08</b>	3.82±0.05	3.51±0.07	3.68±0.04	<b>3.90±0.06</b>
Disgust		4.08±0.09	4.05±0.10	4.10±0.09	<b>4.12±0.07</b>	3.87±0.08	3.69±0.09	3.72±0.09	<b>3.93±0.04</b>
Angry		4.00±0.10	3.93±0.07	4.02±0.05	<b>4.03±0.05</b>	3.76±0.03	3.42±0.10	3.59±0.10	<b>3.82±0.07</b>
Sadness		4.03±0.09	3.99±0.04	4.04±0.09	<b>4.06±0.08</b>	3.81±0.06	3.57±0.08	3.69±0.07	<b>3.88±0.07</b>
Happy		4.01±0.04	3.98±0.05	4.00±0.07	<b>4.04±0.03</b>	3.75±0.09	3.46±0.06	3.61±0.08	<b>3.84±0.08</b>
Surprise		4.02±0.07	3.96±0.06	4.05±0.04	<b>4.09±0.02</b>	3.73±0.07	3.44±0.11	3.64±0.08	<b>3.83±0.05</b>

synthesize speech foreign to the target speaker. For DMOS and AB tests, we randomly select 10 Mandarin sentences from the test set to synthesize speech with 6 types of emotions for two target speakers, respectively, resulting in 120 testing sentences. These synthesized emotional speech sentences also are evaluated for naturalness by MOS. Twenty Chinese native speakers with basic English skills participated in these experiments. The gender distribution was balanced, and their ages ranged from 20 to 30. The final score for each utterance was the average rating by all participants for this sample. The results are associated with 95% confidence intervals in all tests. Besides, speaker cosine similarity and embedding visualization are adopted to evaluate speaker similarity and emotion discrimination objectively.

#### D. Training setups

All the speech sentences are down-sampled to 16 KHz and represented by 80-band Mel-spectrum with a frame length of 50ms, frameshift of 12.5ms, and hop size of 200. A grapheme-to-phoneme (G2P) module converts text sentences into phoneme sequences. The phoneme duration is obtained by a pre-trained Montreal Forced Alignment (MFA) tool [64]. We train all the models for 300K iterations with a batch size of 38 on 4 NVIDIA Tesla V100 GPUs. During the inference, a well-trained Hifi-Gan [69] is adopted as the neural vocoder to reconstruct waveform from the predicted Mel-spectrum.

#### E. Compared methods

As this work, to our knowledge, is the first time that attempts to synthesize foreign emotional speech based on emotion transfer by a DPM-based model, there is no existing method that can be compared directly. Therefore, we selected the most recent relevant methods to compare with our proposed DiCLET-TTS. For fairness, some modifications are made to make the compared methods suitable for cross-lingual emotion transfer. The comparative model and the corresponding improvements are as follows: 1) **M3** [8] is a FastSpeech-based [64] multi-speaker multi-style multi-lingual speech synthesis method that introduced a fine-grained style encoder to relieve the foreign accent problem. To make M3 suitable for emotion transfer, the emotion ID and emotion classifier is introduced in the style predictor and style encoder, respectively. 2) **CSET** [18] is a reference-based cross-speaker emotion transfer method, which introduced an emotion disentangling module to Tacotron2. The text encoder and

decoder are extended by the speaker adversarial training and language embedding [1], respectively. 3) **Grad-TTS** [36] is also improved for cross-lingual emotion transfer. We follow the original setting of Grad-TTS, where the decoder's input is concatenated with the speaker embedding and emotion embedding obtained from two look-up tables with the speaker ID and emotion ID as input, respectively. The text encoder structure is the same as that in DiCLET-TTS and is trained by the speaker adversarial loss.

### V. EXPERIMENTAL RESULTS

In this section, the results of emotions transferred to the intra- and cross-lingual target speakers are presented, i.e., the comparison of DiCLET-TTS with other methods in naturalness, speaker similarity, and emotion similarity. The corresponding demos can be found on the project page<sup>1</sup>, and we recommend readers listen to those demos.

#### A. Performance on naturalness

Two MOS tests are conducted to evaluate the naturalness of Mandarin emotional speech and cross-lingual neutral speech generated by DiCLET-TTS, M3, CSET, and Grad-TTS for intra- and cross-lingual target speakers. The results are shown in Table II, and unsurprisingly, the highest MOS scores are obtained when synthesizing intra-lingual neutral speech for the target speaker in each method since the synthesized speech is unaffected by the foreign accent and emotion. DiCLET-TTS achieves higher scores in cross-lingual neutral speeches, while there is no significant difference in scores between the compared methods. This may be due to the fact that the text encoder in these three compared methods is only constrained by speaker adversarial training, which could somewhat disturb the linguistic coding. In DiCLET-TTS, this disturbance is mitigated by the content loss to stabilize the training and effectively improve the naturalness.

For synthesized Mandarin emotional speech, generally speaking, the naturalness of transferring emotion to the intra-lingual target speaker is better than transferring emotion to the cross-lingual target speaker among all methods. This phenomenon is caused by the fact that emotion could make the tone change more violently and the foreign accent more obvious. The score gap between the synthesized Mandarin emotional speech for intra- and cross-lingual target speakers in DiCLET-TTS and M3 is smaller than that in Grad-TTS and

TABLE III: Speaker and emotion similarity DMOS comparison of DiCLET-TTS, M3, CSET, and Grad-TTS in transferring the emotions to intra- and cross-lingual target speakers, with a confidence interval of 95%. The bold indicates the best performance of the four models in each emotion.

Emotion	Intra-lingual scenario (target Mandarin speaker)							
	Speaker similarity DMOS				Emotion similarity DMOS			
	M3	CSET	Grad-TTS	DiCLET-TTS	M3	CSET	Grad-TTS	DiCLET-TTS
Fear	<b>4.04±0.04</b>	3.91±0.02	4.02±0.01	4.01±0.05	3.85±0.03	3.71±0.06	3.17±0.03	<b>4.04±0.04</b>
Disgust	4.05±0.02	3.96±0.04	4.06±0.07	<b>4.08±0.04</b>	3.79±0.05	3.60±0.03	3.13±0.05	<b>3.90±0.06</b>
Angry	<b>4.01±0.06</b>	3.87±0.03	3.97±0.05	3.98±0.08	3.81±0.06	3.68±0.08	3.19±0.11	<b>3.96±0.09</b>
Sadness	4.02±0.02	3.77±0.05	<b>4.03±0.02</b>	4.00±0.06	3.90±0.04	3.89±0.04	3.28±0.08	<b>4.02±0.03</b>
Happy	<b>3.97±0.05</b>	3.79±0.04	3.94±0.06	<b>3.96±0.04</b>	3.92±0.07	3.87±0.06	3.30±0.04	<b>4.04±0.08</b>
Surprise	3.94±0.06	3.84±0.07	<b>4.01±0.04</b>	3.97±0.07	3.87±0.03	3.82±0.04	3.25±0.07	<b>3.97±0.06</b>
Cross-lingual scenario (target English speaker)								
Fear	<b>3.91±0.05</b>	3.70±0.04	3.86±0.06	3.89±0.02	3.64±0.07	3.55±0.05	3.07±0.09	<b>3.86±0.06</b>
Disgust	<b>3.94±0.04</b>	3.74±0.07	3.92±0.08	3.90±0.09	3.51±0.08	3.39±0.04	3.05±0.08	<b>3.81±0.07</b>
Angry	<b>3.81±0.05</b>	3.66±0.10	3.78±0.04	3.79±0.07	3.62±0.11	3.56±0.03	3.14±0.04	<b>3.84±0.09</b>
Sadness	<b>3.87±0.09</b>	3.64±0.05	3.76±0.03	3.85±0.06	3.57±0.09	3.41±0.08	3.19±0.07	<b>3.91±0.03</b>
Happy	3.72±0.04	3.65±0.02	3.75±0.07	<b>3.80±0.08</b>	3.68±0.04	3.64±0.06	3.21±0.10	<b>3.93±0.05</b>
Surprise	3.68±0.06	3.68±0.09	3.71±0.03	<b>3.74±0.07</b>	3.60±0.05	3.55±0.02	3.20±0.06	<b>3.79±0.04</b>

CSET. This advantage mainly comes from DiCLET-TTS and M3 adopting prosodic-related linguistic representation, which can alleviate the foreign accent problem and improve the naturalness of cross-lingual emotion transfer. Besides, DiCLET-TTS achieves the highest naturalness score in synthesized neutral and emotional speeches, indicating that the proposed method can disentangle speakers and languages while stabilizing the training, making the speakers speak foreign languages fluently and express various emotions in authentic Mandarin.

#### B. Performance on emotion transfer

Besides measuring the naturalness, the target speaker similarity and transferred emotion similarity are also evaluated. Four DMOS tests are conducted to evaluate the speaker similarity and emotion similarity of generated Mandarin emotional speech by DiCLET-TTS, M3, CSET, and Grad-TTS for intra- and cross-lingual target speakers. The results are shown in Table III, where the upper part is the speaker and emotion similarity results of transferring emotion from the source speaker to the intra-lingual target speaker, the lower part is the results of transferring emotion from the source speaker to the cross-lingual target speaker.

As seen in Table III, regarding the speaker similarity of all emotion categories in each method, the scores of synthesized emotional speech of the intra-lingual target speaker are higher than that of the cross-lingual target speaker. This phenomenon could be partially caused by emotion and language affecting participants' perception since the compared reference during the DMOS test of the cross-lingual target speaker is neutral English audio rather than Mandarin emotional audio. A similar situation also occurs in emotion similarity DMOS. These results indicate that compared with the cross-speaker emotion transfer task, which only recombines the two factors (speaker, emotion), it is more challenging to simultaneously recombine the three factors (speaker, language, and emotion), which are deeply entangled.

Specifically, regarding speaker similarity, the difference between DiCLET-TTS, M3, and Grad-TTS are not noticeable, while CSET performs the worst. Although the emotion sim-

TABLE IV: Speaker cosine similarity of synthesized speech with the cross-lingual target speaker and emotional source speaker, respectively.

Speaker	Target speaker	M3	CSET	Grad-TTS	DiCLET-TTS
Source speaker	0.18	0.23	0.29	<b>0.21</b>	0.25
Target speaker	0.80	<b>0.75</b>	0.65	0.72	0.73

ilarity of CSET is better than Grad-TTS, the poor scores in speaker similarity and cross-lingual naturalness (see Table II) indicate the weakness of CSET for the cross-lingual emotion transfer task. Grad-TTS achieves reasonable speaker similarity in transferring the emotion to intra- and cross-lingual speakers but performs poorly in emotion similarity. It is mainly caused by Grad-TTS adopting a look-up table in emotion modeling, which produces average emotion expressiveness. DiCLET-TTS outperforms Grad-TTS in terms of emotion and speaker similarity, showing that such emotion transfer performance is derived not only from the diffusion model but also from the introduced OP-EDM and emotional adaptor.

DiCLET-TTS significantly outperforms all comparison methods in emotion similarity and obtains a comparable speaker similarity score with M3. The slight speaker similarity gap between M3 and DiCLET-TTS could be caused by the stronger emotional expressiveness of DiCLET-TTS, which could affect participants on the rating of the timbre similarity. Besides, M3 and DiCLET-TTS adopt speaker adversarial training to remove speaker-related information in emotion embedding. The emotional information conveyed by such disentangled emotion embedding tends to be weakened since the speaker and emotion are deeply entangled and both related to the prosody. While in DiCLET-TTS, the emotion embedding space obtained by OP-EDM is further constrained to ensure that the emotion embedding retains high emotion discrimination after removing the speaker-related information, thus promoting the expressiveness of transferred emotions. These results show that DiCLET-TTS can well balance maintaining the target speaker's identity and enriching the transferred emotion expressiveness in intra- and cross-lingual scenarios.



TABLE V: Speaker and emotion similarity DMOS comparison of DiCLET-TTS, “w/o EA” and “w/o CE-D” in transferring the emotion to the cross-lingual target speaker, with a confidence interval of 95%, and the higher value means better performance and the bold indicates the best performance out of four models in terms of each emotion.  $\mu$  and  $\mu_{emo}$  represent emotion-irrelevant and emotion-related linguistic representation, respectively.

Emotion	Speaker similarity DMOS			Emotion similarity DMOS		
	“w/o EA” ( $\mu$ )	“w/o CE-D” ( $\mu_{emo}$ )	DiCLET-TTS ( $\mu_{emo}$ )	“w/o EA” ( $\mu$ )	“w/o CE-D” ( $\mu_{emo}$ )	DiCLET-TTS ( $\mu_{emo}$ )
Fear	<b>3.91±0.05</b>	3.83±0.12	<b>3.89±0.02</b>	3.40±0.09	3.53±0.05	<b>3.86±0.06</b>
Disgust	<b>4.00±0.04</b>	3.88±0.07	<b>3.96±0.09</b>	3.41±0.08	3.46±0.04	<b>3.81±0.07</b>
Angry	<b>3.82±0.03</b>	3.73±0.05	3.79±0.07	3.60±0.04	3.66±0.03	<b>3.84±0.09</b>
Sadness	<b>3.90±0.06</b>	3.77±0.09	<b>3.83±0.06</b>	3.43±0.07	3.61±0.08	<b>3.91±0.03</b>
Happy	<b>3.79±0.06</b>	3.66±0.04	3.72±0.08	3.66±0.10	3.72±0.06	<b>3.93±0.05</b>
Surprise	<b>3.75±0.08</b>	3.63±0.04	3.68±0.07	3.58±0.06	3.69±0.02	<b>3.79±0.04</b>

### C. Speaker similarity with target speaker and source speaker in cross-lingual emotion transfer

To objectively show the speaker leakage degree of each method, we calculate the speaker cosine similarity between synthesized speech and ground-truth neutral speech from the cross-lingual target speaker and emotional source speaker, respectively. Specifically, we adopt a pre-trained speaker verification model ECAPA-TDNN [70] to extract the x-vectors of synthesized and ground truth speech. The speaker cosine similarity with the target speaker and the source speaker has measured on 80 synthesized speech.

We first calculated the upper bound of cosine similarity within the target speaker’s ground truth speech, and the lower bound between the target speaker and the source speaker. As shown in Table IV, the upper bound is **0.80**, and the lower bound is **0.18**. Note that the synthesized speech from CSET has the highest similarity with the source speaker and the lowest similarity with the target speaker, consistent with the results shown in Section V-B. The speech synthesized by DiCLET-TTS achieves a comparable cosine similarity score with M3, and as explained above, this gap may also be caused by the stronger emotion expressiveness of DiCLET-TTS.

## VI. COMPONENT ANALYSIS

In Section V, DiCLET-TTS has shown good performance on emotion transfer in intra- and cross-lingual scenarios. In this section, the effectiveness of each proposed component, i.e., content loss, emotional adaptor, and condition-enhanced DPM decoder, is evaluated by transferring emotion to the cross-lingual target speaker. The advantages of the proposed orthogonal projection based emotion disentanglement module (OP-EDM) are also analyzed.

### A. The effectiveness of content loss and emotional adaptor on naturalness

In DiCLET-TTS, the content loss and emotional adaptor are the keys to improving the naturalness of synthesized cross-lingual speech. Besides, with the guidance of emotion embedding extracted by OP-EDM, the emotional adaptor and condition-enhanced DPM decoder are further committed to enhancing emotion expressiveness. Therefore, we first conduct an ablation study via the MOS test to verify the benefits of content loss and emotional adaptor in improving naturalness. We do not verify the benefits of the condition-enhanced DPM

decoder since it contributes little to improving naturalness. Specifically, two variants are evaluated: 1) no content loss is adopted for the text encoder’s output, which is constrained only by speaker adversarial training. We denote this variant as “w/o CTL”. 2) No emotional adaptor is adopted for the length regulator’s output. We denote this variant as “w/o EA”.

Table VI shows the naturalness MOS results of DiCLET-TTS and its two variants. Comparing DiCLET-TTS and “w/o CTL”, we can find the drop of naturalness when discarding content loss in “w/o CTL”, indicating that introducing content loss in adversarial training can effectively improve the naturalness in synthesized speech. We also find that the degradation is more prominent in some emotion categories, i.e., *happy*, *surprise*, and *angry*, since the intonation changes in these categories are more dramatic. Besides, the naturalness significantly drops in “w/o EA”, where the linguistic representation is emotion-irrelevant. This result suggests that parameterizing the terminal distribution of the diffusion process into emotion-related linguistic prior by the emotional adaptor plays an essential role in promoting naturalness.

TABLE VI: Naturalness MOS results of DiCLET-TTS, “w/o CTL”, and “w/o EA” in transferring emotion to the cross-lingual target speaker, with confidence intervals of 95%. Neutral (Mandarin) and neutral (English) represent synthesized neutral Mandarin and English speech, respectively.

Method	“w/o CTL”	“w/o EA”	DiCLET-TTS
Neutral (Mandarin)	3.93±0.04	3.86±0.05	<b>3.98±0.05</b>
Neutral (English)	4.15±0.07	4.11±0.04	<b>4.21±0.05</b>
Fear	3.84±0.09	3.71±0.07	<b>3.90±0.06</b>
Disgust	3.85±0.08	3.76±0.12	<b>3.93±0.04</b>
Angry	3.71±0.05	3.66±0.11	<b>3.82±0.07</b>
Sadness	3.83±0.10	3.74±0.07	<b>3.88±0.07</b>
Happy	3.74±0.09	3.70±0.05	<b>3.84±0.08</b>
Surprise	3.72±0.07	3.69±0.08	<b>3.83±0.05</b>

### B. The effectiveness of emotional adaptor and condition-enhanced DPM decoder on speaker and emotion similarity

The effectiveness of the emotional adaptor in improving naturalness has been verified in Section VI-A. In this section, we further present the benefits of the emotional adaptor and condition-enhanced DPM decoder in terms of the speaker and emotion similarity by two DMOS tests. Therefore, besides the variant “w/o EA”, the variant “w/o CE-D” is also taken into the test, where the emotion embedding and speaker embedding are

concatenated with the input of the decoder rather than being added to each ResBlock.

As shown in Table V, regarding the emotion similarity, the two variants in all categories have dropped compared with DiCLET-TTS, and the degradation of “w/o EA” is the most significant. The lower emotion similarity of “w/o EA” brings a weaker impact on the speaker identity of synthesized speech, resulting in a slightly better performance than DiCLET-TTS in speaker similarity. Specifically, the emotion modeling of “w/o EA” is only completed in the condition-enhanced DPM decoder under the condition of the emotion embedding learned by OP-EDM. And the linguistic prior of “w/o EA” is emotion-irrelevant. This result indicates that parameterizing the terminal distribution of the diffusion process as an emotion-related linguistic prior by the emotional adaptor can also effectively improve the expressiveness of transferred emotion. Besides, referring to the results in Table III, the emotion similarity of “w/o EA” is superior to that of Grad-TTS in terms of all emotion categories, and “w/o EA” also has an improved performance than CSET in most cases (except *disgust*). These results also reflect the effectiveness of the introduced OP-EDM in learning speaker-irrelevant emotion embedding, which can result in a good performance in terms of speaker similarity and emotional expressiveness.

For “w/o CE-D”, although it achieves better performance than “w/o EA” on emotion expressiveness, this improvement is not always significant. Emotion expressiveness is still unsatisfactory for emotions (e.g., *disgust* and *fear*) that rely on speaking speed and stress. Meanwhile, for emotions partially reflected in the changes of the source speaker’s timbre (e.g., *happy* and *surprise*), the target speaker similarity of “w/o CE-D” is dropped. All these results show that with the guidance of speaker-irrelevant emotion embedding extracted from OP-EDM, the emotional adaptor and condition-enhanced DPM decoder can effectively improve the performance of cross-lingual emotion transfer while maintaining reasonable speaker similarity and speech naturalness.

### C. Advantages of emotion embedding space with orthogonal projection

This section analyzes the benefits of the proposed orthogonal projection based emotion disentanglement module (OP-EDM) by comparing it with the variant “w/o OPL”, in which “w/o OPL” means the orthogonal projection loss in OP-EDM is removed. Ideally, the emotion embedding learned by the emotion disentanglement module is expected to be irrelevant to the speaker identities but holds high emotion discrimination. Therefore, the t-distributed stochastic neighbor embedding (t-SNE) [71] is adopted to demonstrate the capacity of emotion embedding learned from these two modules on distinguishing emotion categories or speaker identities.

1) *Emotion discrimination ability*: To display the distribution of emotion embeddings extracted by “w/o OPL” and OP-EDM, 80 speeches of each emotion category from the *CN-emo*’s test set are randomly selected, resulting in 560 reference speeches in total and then embedded as emotion embeddings by these two modules, respectively. The distributions of these embeddings are presented in Fig. 2, where each point indicates

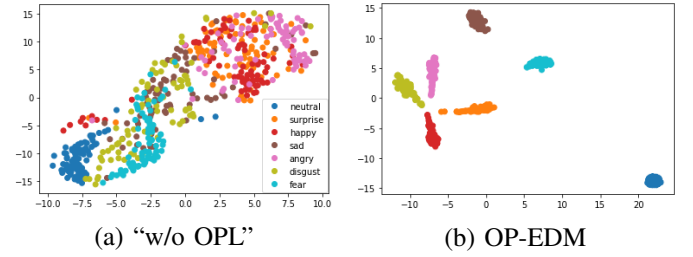


Fig. 2: Emotion distribution of the emotion embedding created by different models (a) “w/o OPL” and (b) OP-EDM. The presented data are 80 sentences randomly selected from each emotion category of the *CN-emo*’s test set.

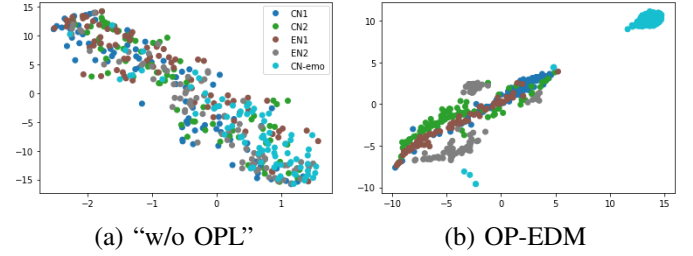


Fig. 3: Speaker distribution of the emotion embedding created by different models (a) “w/o OPL” and (b) OP-EDM module. The presented data are 80 neutral sentences randomly selected from each speaker’s test set.

an emotion embedding, and points with the same color are from the same emotion category. Smaller distances between the two points indicate that the embeddings are more similar. As shown in Fig. 2 (a), the emotional embedding generated by “w/o OPL” only retains weak emotion discrimination, where emotions with similar characteristics tend to be confused: (1) *happy*, *surprise*, and *angry* with a higher pitch and fast speech speed; (2) *sad*, *fear*, and *disgust* with a deep voice and slower speech speed. In contrast, in Fig. 2 (b), the emotion embeddings from the same emotion category are clustered together while different clusters are separated, demonstrating that benefits from the orthogonal projection loss, OP-EDM can obtain an embedding space with high emotion discrimination. We notice that although these embeddings are all from the same speaker, the neutral embeddings are far away from the others. This phenomenon could be due to the fact that emotions are mainly reflected in pitch, energy, and speech speed, and these attributes are relatively flat in neutral emotions.

2) *Speaker identity removal capability*: For speaker identity visualization, 80 neutral speeches are randomly selected from each speaker’s test set, resulting in 400 speeches. The visualization results are shown in Fig. 3, in which the same color colors the embeddings from the identical speaker. As mentioned, the emotion embedding should contain no speaker-related information but only emotional information, which implies embeddings extracted from different speakers’ speech are expected to be inseparable. As shown in Fig. 3(a), the embeddings from different speakers extracted by “w/o OPL” are indeed inseparable, while the cost is that these embeddings maintain little emotion information from the reference audio (see Fig. 2(a)). For the OP-EDM module (see Fig. 3(b)), the embeddings from the four neutral speakers’ corpus are

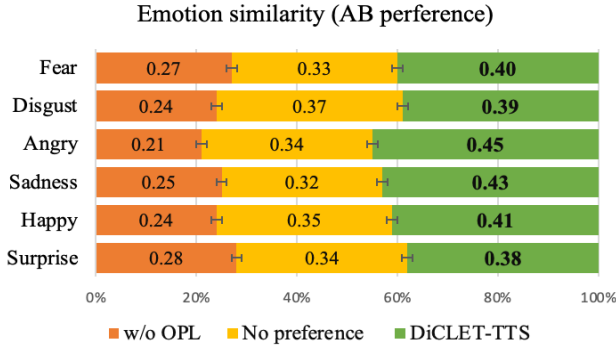


Fig. 4: Emotion similarity AB preference test for “w/o OPL” and DiCLET-TTS with confidence intervals of 95%.

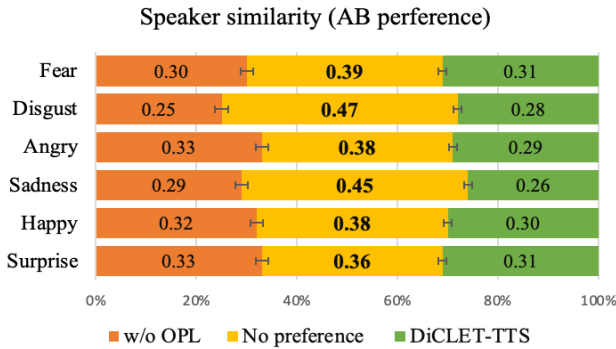


Fig. 5: Speaker similarity AB preference test for “w/o OPL” and DiCLET-TTS with confidence intervals of 95%.

clustered into one cluster. It is worth noting that the neutral speech from *CN-emo* is treated as an independent emotion category, so the embeddings from *CN-emo* are clustered into a separate cluster in Fig. 3(b). This distribution indicates that the proposed OPL-EDM can effectively remove the speaker-related information while greatly retaining the emotion-related information, resulting in speaker-irrelevant but emotion-discriminative embedding.

3) *Preference test*: To further investigate the effectiveness of using OPL in learning emotion embedding for emotion transfer. We conducted two AB tests between DiCLET-TTS and the variant “w/o OPL” regarding emotion and speaker similarity. The results are shown in Fig. 4 and Fig. 5, respectively. As shown in Fig. 4, we can find that “w/o OPL” obtains lower preference in all emotion categories, showing lower emotion similarity is perceived. In contrast, the listeners preferred DiCLET-TTS more when we inserted OPL into OP-EDM. As analyzed, the performance gain is essentially contributed by the OPL strategy in learning discriminative emotion embeddings. Regarding speaker similarity, as shown in Fig. 5, there is no significant difference between “w/o OPL” and DiCLET-TTS, i.e., most listeners give *No preference*. All the above evidence shows that OP-EDM introduced in this paper contributes to better emotion similarity without reducing speaker similarity.

## VII. CONCLUSION

This paper proposes a DPM-based cross-lingual emotion transfer model – DiCLET-TTS. We adopt prosodic information to alleviate the foreign accent problem, where a prior text encoder takes emotion embedding as a condition to parameterize the terminal distribution of the forward diffusion processes into a speaker-irrelevant but emotion-related linguistic prior. To address the weaker emotional expressiveness problem caused by removing speaker information from emotion embedding, an orthogonal projection based emotion disentangling module (OP-EDM) is proposed to learn the speaker-irrelevant but high emotion-discriminative embedding. The reverse diffusion process is parameterized by a condition-enhanced DPM decoder, where the modeling ability of the speaker and emotion is enhanced to further improve emotion expressiveness in synthetic speech. Experimental results demonstrate that DiCLET-TTS performs well in intra- and cross-lingual emotion transfer while preserving the timbre of the target speaker and synthesized naturalness. The results also prove the advantages of OP-EDM in learning speaker-irrelevant but emotion-discriminative embedding.

In this study, only the same-gender speakers are involved in our experiments while cross-gender emotion transfer is considered a difficult task itself and it can be more challenging in the cross-lingual scenario. We will further study this cross-gender task as a follow-up work.

## REFERENCES

- [1] Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Z. Chen, R. Skerry-Ryan, Y. Jia, A. Rosenberg, and B. Ramabhadran, “Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning,” *Proc. Interspeech 2019*, pp. 2080–2084, 2019.
- [2] S. K. Rallabandi and A. W. Black, “On building mixed lingual speech synthesis systems,” in *Interspeech 2017*, pp. 52–56.
- [3] Z. Cai, Y. Yang, and M. Li, “Cross-lingual multi-speaker speech synthesis with limited bilingual training data,” *Computer Speech & Language*, vol. 77, p. 101427, 2023.
- [4] C. Hu, Q. Tian, T. Li, W. Yuping, Y. Wang, and H. Zhao, “Neural dubber: Dubbing for videos according to scripts,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 16 582–16 595, 2021.
- [5] A. Pourhosein Gilakjani and R. Rahimy, “Using computer-assisted pronunciation teaching (capt) in english pronunciation instruction: A study on the impact and the teacher’s role,” *Education and information technologies*, vol. 25, no. 2, pp. 1129–1159, 2020.
- [6] 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 *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6608–6612.
- [7] D. Xin, Y. Saito, S. Takamichi, T. Koriyama, and H. Saruwatari, “Cross-lingual text-to-speech synthesis via domain adaptation and perceptual similarity regression in speaker space,” in *Interspeech*, 2020, pp. 2947–2951.
- [8] Z. Shang, Z. Huang, H. Zhang, P. Zhang, and Y. Yan, “Incorporating cross-speaker style transfer for multi-language text-to-speech,” in *Interspeech*, 2021, pp. 1619–1623.
- [9] 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 *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6072–6076.
- [10] H. Zhan, X. YU, H. Zhang, Y. Zhang, and Y. Lin, “Exploring Timbre Disentanglement in Non-Autoregressive Cross-Lingual Text-to-Speech,” in *Proc. Interspeech 2022*, 2022, pp. 4247–4251.

- [11] Y. Cao, S. Liu, X. Wu, S. Kang, P. Liu, Z. Wu, X. Liu, D. Su, D. Yu, and H. Meng, "Code-switched speech synthesis using bilingual phonetic posteriorgram with only monolingual corpora," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7619–7623.
- [12] S. Zhao, T. H. Nguyen, H. Wang, and B. Ma, "Towards natural bilingual and code-switched speech synthesis based on mix of monolingual recordings and cross-lingual voice conversion," *Proc. Interspeech 2020*, pp. 2927–2931, 2020.
- [13] B. Li, Y. Zhang, T. Sainath, Y. Wu, and W. Chan, "Bytes are all you need: End-to-end multilingual speech recognition and synthesis with bytes," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5621–5625.
- [14] S. Bansal, A. Mukherjee, S. Satpal, and R. Mehta, "On improving code mixed speech synthesis with mixlingual grapheme-to-phoneme model," *Proc. Interspeech 2020*, pp. 2957–2961, 2020.
- [15] G. Maniati, N. Ellinas, K. Markopoulos, G. Vamvoukakis, J. S. Sung, H. Park, A. Chalamandaris, and P. Tsiakoulis, "Cross-lingual low resource speaker adaptation using phonological features," *Proc. Interspeech 2021*, pp. 1594–1598.
- [16] H. Zhan, H. Zhang, W. Ou, and Y. Lin, "Improve cross-lingual text-to-speech synthesis on monolingual corpora with pitch contour information," in *Interspeech*, 2021, pp. 1599–1603.
- [17] T. Nekvinda and O. Dušek, "One model, many languages: Meta-learning for multilingual text-to-speech," *Proc. Interspeech 2020*, pp. 2972–2976, 2020.
- [18] 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.
- [19] 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.
- [20] O. Kwon, I. Jang, C. H. Ahn, and H.-G. Kang, "Emotional speech synthesis based on style embedded tacotron2 framework," *2019 34th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*, pp. 1–4, 2019.
- [21] 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," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6945–6949, 2019.
- [22] S. Ö. Arık, G. Diamos, A. Gibiansky, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 2966–2974.
- [23] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 4485–4495.
- [24] Y. Bian, C. Chen, Y. Kang, and Z. Pan, "Multi-reference tacotron by intercross training for style disentangling, transfer and control in speech synthesis," *CoRR*, vol. abs/1904.02373, 2019.
- [25] M. Whitehill, S. Ma, D. McDuff, and Y. Song, "Multi-reference neural tts stylization with adversarial cycle consistency," *Proc. Interspeech 2020*, pp. 4442–4446, 2020.
- [26] S. Karlapati, A. Moinet, A. Joly, V. Klimkov, D. Sáez-Trigueros, and T. Drugman, "Copycat: Many-to-many fine-grained prosody transfer for neural text-to-speech," 2020, pp. 4387–4391.
- [27] T. Li, X. Wang, Q. Xie, Z. Wang, M. Jiang, and L. Xie, "Cross-speaker Emotion Transfer Based On Prosody Compensation for End-to-End Speech Synthesis," in *Proc. Interspeech 2022*, 2022, pp. 5498–5502.
- [28] J. M. Levis, "Intonation in theory and practice, revisited," *TESOL quarterly*, vol. 33, no. 1, pp. 37–63, 1999.
- [29] S. Duanmu, "Tone and non-tone languages: An alternative to language typology and parameters," *Language and Linguistics*, vol. 5, no. 4, pp. 891–923, 2004.
- [30] Y. Li, C. Tang, J. Lu, J. Wu, and E. F. Chang, "Human cortical encoding of pitch in tonal and non-tonal languages," *Nature communications*, vol. 12, no. 1, p. 1161, 2021.
- [31] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International Conference on Machine Learning*. PMLR, 2015, pp. 2256–2265.
- [32] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations*, 2020.
- [33] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*, 2020.
- [34] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.
- [35] B. Kavar, M. Elad, S. Ermon, and J. Song, "Denoising diffusion restoration models," in *ICLR Workshop on Deep Generative Models for Highly Structured Data*, 2022.
- [36] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-tts: A diffusion probabilistic model for text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8599–8608.
- [37] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, N. Dehak, and W. Chan, "Wavegrad 2: Iterative refinement for text-to-speech synthesis," *arXiv preprint arXiv:2106.09660*, 2021.
- [38] M. Jeong, H. Kim, S. J. Cheon, B. J. Choi, and N. S. Kim, "Diff-tts: A denoising diffusion model for text-to-speech," *arXiv preprint arXiv:2104.01409*, 2021.
- [39] R. Huang, M. W. Lam, J. Wang, D. Su, D. Yu, Y. Ren, and Z. Zhao, "Fastdiff: A fast conditional diffusion model for high-quality speech synthesis," *arXiv preprint arXiv:2204.09934*, 2022.
- [40] K. Ranasinghe, M. Naseer, M. Hayat, S. Khan, and F. S. Khan, "Orthogonal projection loss," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 333–12 343.
- [41] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
- [42] Z. Liu and B. Mak, "Multi-lingual multi-speaker text-to-speech synthesis for voice cloning with online speaker enrollment," in *INTERSPEECH*, 2020, pp. 2932–2936.
- [43] M. Chen, M. Chen, S. Liang, J. Ma, L. Chen, S. Wang, and J. Xiao, "Cross-lingual, multi-speaker text-to-speech synthesis using neural speaker embedding," *Proc. Interspeech 2019*, pp. 2105–2109, 2019.
- [44] S. Pan and L. He, "Cross-Speaker Style Transfer with Prosody Bottleneck in Neural Speech Synthesis," in *Proc. Interspeech 2021*, 2021, pp. 4678–4682.
- [45] Q. Xie, T. Li, X. Wang, Z. Wang, L. Xie, G. Yu, and G. Wan, "Multi-speaker multi-style text-to-speech synthesis with single-speaker single-style training data scenarios," *arXiv preprint arXiv:2112.12743*, 2021.
- [46] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proc. ICML*, 2018, pp. 5180–5189.
- [47] 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," *arXiv preprint arXiv:2206.14866*, 2022.
- [48] T. Li, S. Yang, L. Xue, and L. Xie, "Controllable emotion transfer for end-to-end speech synthesis," *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 1–5, 2021.
- [49] R. Liu, B. Sisman, G. Gao, and H. Li, "Expressive tts training with frame and style reconstruction loss," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1806–1818, 2021.
- [50] R. Liu, B. Sisman, B. Schuller, G. Gao, and H. Li, "Accurate Emotion Strength Assessment for Seen and Unseen Speech Based on Data-Driven Deep Learning," in *Proc. Interspeech 2022*, 2022, pp. 5493–5497.
- [51] 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," *Proc. Interspeech 2017*, pp. 4006–4010, 2017.
- [52] 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 *Proc. ICASSP*. IEEE, 2018, pp. 4779–4783.
- [53] H. Cao, C. Tan, Z. Gao, G. Chen, P.-A. Heng, and S. Z. Li, "A survey on generative diffusion model," *arXiv preprint arXiv:2209.02646*, 2022.
- [54] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, 2021.
- [55] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. S. Kudinov, and J. Wei, "Diffusion-based voice conversion with fast maximum likelihood

- sampling scheme,” in *International Conference on Learning Representations*, 2021.
- [56] Y. Leng, Z. Chen, J. Guo, H. Liu, J. Chen, X. Tan, D. Mandic, L. He, X.-Y. Li, T. Qin *et al.*, “Binauralgrad: A two-stage conditional diffusion probabilistic model for binaural audio synthesis,” *arXiv preprint arXiv:2205.14807*, 2022.
  - [57] R. Huang, Z. Zhao, H. Liu, J. Liu, C. Cui, and Y. Ren, “Prodiff: Progressive fast diffusion model for high-quality text-to-speech,” in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 2595–2605.
  - [58] D. Yang, J. Yu, H. Wang, W. Wang, C. Weng, Y. Zou, and D. Yu, “Diffsound: Discrete diffusion model for text-to-sound generation,” *arXiv preprint arXiv:2207.09983*, 2022.
  - [59] S.-g. Lee, H. Kim, C. Shin, X. Tan, C. Liu, Q. Meng, T. Qin, W. Chen, S. Yoon, and T.-Y. Liu, “Priorgrad: Improving conditional denoising diffusion models with data-dependent adaptive prior,” in *International Conference on Learning Representations*, 2021.
  - [60] H. Kim, S. Kim, and S. Yoon, “Guided-tts: A diffusion model for text-to-speech via classifier guidance,” in *International Conference on Machine Learning*. PMLR, 2022, pp. 11 119–11 133.
  - [61] S. Kim, H. Kim, and S. Yoon, “Guided-tts 2: A diffusion model for high-quality adaptive text-to-speech with untranscribed data,” *arXiv preprint arXiv:2205.15370*, 2022.
  - [62] J. Liu, C. Li, Y. Ren, F. Chen, and Z. Zhao, “Diffsinger: Singing voice synthesis via shallow diffusion mechanism,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 11 020–11 028.
  - [63] H. Xue, X. Wang, Y. Zhang, L. Xie, P. Zhu, and M. Bi, “Learn2Sing 2.0: Diffusion and Mutual Information-Based Target Speaker SVS by Learning from Singing Teacher,” in *Proc. Interspeech 2022*, 2022, pp. 4267–4271.
  - [64] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, “Fastspeech: Fast, robust and controllable text to speech,” in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019, pp. 3171–3180.
  - [65] M. Chen, X. Tan, B. Li, Y. Liu, T. Qin, T.-Y. Liu *et al.*, “Adaspeech: Adaptive text to speech for custom voice,” in *International Conference on Learning Representations*, 2020.
  - [66] 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,” *arXiv preprint arXiv:2110.12612*, 2021.
  - [67] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, “Conformer: Convolution-augmented transformer for speech recognition,” *Proc. Interspeech 2020*, pp. 5036–5040, 2020.
  - [68] X. An, F. K. Soong, and L. Xie, “Disentangling style and speaker attributes for tts style transfer,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 646–658, 2022.
  - [69] 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.
  - [70] B. Desplanques, J. Thienpondt, and K. Demuyne, “Ecapa-TDNN: Emphasized channel attention, propagation and aggregation in tdnns based speaker verification,” *Proc. Interspeech 2020*, pp. 3830–3834, 2020.
  - [71] V. D. M. Laurens and G. Hinton, “Visualizing data using t-SNE,” *Journal of Machine Learning Research*, vol. 9, no. 2605, pp. 2579–2605, 2008.