iEmoTTS: Toward Robust Cross-Speaker Emotion Transfer and Control for Speech Synthesis Based on Disentanglement Between Prosody and Timbre

Guangyan Zhang, Ying Qin, Wenjie Zhang, Jialun Wu, Mei Li, Yutao Gai, Feijun Jiang, and Tan Lee, Member, IEEE

Abstract—Cross-speaker emotion transfer is a common approach to generating emotional speech when speech data with emotion labels from target speakers is not available. This paper presents a novel cross-speaker emotion transfer system named iEmoTTS. The system is composed of an emotion encoder, a prosody predictor, and a timbre encoder. The emotion encoder extracts the identity of emotion type and the respective emotion intensity from the mel-spectrogram of input speech. The emotion intensity is measured by the posterior probability that the input utterance carries that emotion. Unlike many other studies which focus on disentangling speaker and style factors of speech, the iEmoTTS is designed to achieve cross-speaker emotion transfer via disentanglement between prosody and timbre. Prosody is considered the primary carrier of emotion-related speech characteristics, and timbre accounts for the essential characteristics for speaker identification. Zero-shot emotion transfer, meaning that the speech of target speakers is not seen in model training, is also realized with iEmoTTS. Extensive experiments of subjective evaluation have been carried out. The results demonstrate the effectiveness of iEmoTTS compared with other recently proposed systems of cross-speaker emotion transfer. It is shown that iEmoTTS can produce speech with designated emotion types and controllable emotion intensity. With appropriate information bottleneck capacity, iEmoTTS is able to transfer emotional information to a new speaker effectively. Audio samples are publicly available.

Index Terms—Emotion transfer, emotion intensity, cross-speaker, zero-shot, timbre, prosody.

I. INTRODUCTION

T

HE basic function of a text-to-speech (TTS) system is to produce fluent speech from text input. Early approaches to TTS included waveform concatenation [1], [2] and statistical parametric synthesis [3], [4], [5]. In recent years, TTS systems have been predominantly based on deep neural network models [6], [7], [8], [9]. State-of-the-art neural TTS systems are able to generate highly intelligible speech with human-level perceived naturalness. Nevertheless, the ability to realize expressiveness pertinent to natural speech communication is generally considered inadequate or absent. Expressiveness in natural speech is related to the speaker’s affect status, namely emotion, mood and attitude, depending on the communication scenario. The expressiveness is mainly manifested by prosody of speech [2], [10]. For applications like spoken dialogue systems, chat avatars, and video game dubbing, it is desirable to have computer-generated speech that can express specific emotions, e.g., anger and fear, to create an engaging user experience. The aim of this research is to develop a neural TTS system that can generate speech with various emotion types and controllable emotion intensity.

One of the challenges in emotional speech generation with neural TTS models is the scarcity of labelled data. Cross-speaker emotion transfer is a potential solution to the data scarcity problem in emotional speech generation. This method involves transferring emotion-related speech characteristics from source speakers to target speakers, without requiring labelled data from the target speakers. Previous studies have explored various methods to incorporate emotion information into TTS models, such as using a categorical identity of emotion type as an auxiliary input [11], [12], [13], [14]. However, this approach requires training utterances with manually labelled emotion types, which are costly and time-consuming to prepare. A common alternative is fine-tuning a pre-trained model with labelled data, but this still depends on the availability of manually labelled data for the target speakers. In contrast, cross-speaker emotion transfer only requires utterances with emotion labels from the source speakers and a batch of utterances from the target speakers.

To date, a quantitative and detailed analysis of the emotional characteristics of the speech of the target speakers is lacking in the literature. Therefore, this study aims to address the following research question: How does the speech of the target speakers...
convey emotion? Following Ortony et al. [15], this study adopts the assumption that all emotions have affective valence and none of them are affectively neutral. A potential method for generating emotional speech is to learn a transformation from the source speakers’ neutral speech to their emotional speech, and then apply it to the target speakers’ neutral speech. However, this method faces several challenges. First, it may not be feasible to obtain or assume neutral speech from the target speakers. Second, if the source speakers’ neutral speech is also unavailable, the transformation cannot be learned.

One of the key challenges in cross-speaker emotion transfer is to disentangle emotion and speaker factors in speech. Speaker factor is usually represented by an embedding, which can be obtained from either a trainable look-up table [16] or a pre-trained speaker verification system [17], [18]. Emotion factor is learned from reference speech using a reference encoder module [19]. However, the learned emotion embedding may be confounded with other speech factors, such as speaker, text content, and environment [10], [19], [20], [21]. Disentangling emotion from speaker identity is particularly difficult when some emotions are specific to certain speakers in the dataset. In this study, we focus on the disentanglement between timbre and prosody, which are two important aspects of speech. Emotion is mainly conveyed and perceived through prosody [2], [10], while timbre is the perceptual attribute that distinguishes two sounds with the same pitch, intensity, and duration [22]. Timbre is closely related to the speaker’s voice [23], and different speakers can be recognized by their timbre cues. Based on the definition of timbre, we hypothesize that there is a complementary relationship between prosody and timbre, and that disentangling timbre information from prosody would lead to more robust emotion transfer than disentangling speaker and emotion.

Speech datasets often provide emotion labels in the form of discrete affect states, such as anger, happiness, and sadness. However, emotional expressions in human speech are complex and vary in intensity. Labelling the emotion intensity of speech utterances is more difficult and costly than labelling the emotion categories. Moreover, there is no universal or standard model to describe emotion intensity in speech. Previous studies have explored unsupervised methods for labelling and controlling emotion intensity in speech [14], [24], [25], [26], [27], [28]. In this paper, we propose a novel approach to probability-based emotion intensity learning and control, which can generate emotional speech with different intensity levels, such as weak, moderate and high. Our hypothesis is that an utterance with a specific emotion type has a higher perceived emotion intensity if it can be easily distinguished from utterances with other emotion types by a speech emotion recognition (SER) model.

In zero-shot emotion transfer, emotion is transferred to the speech of target speakers. The utterances of target speakers are unseen to the model in the training process. The goal of zero-shot emotion transfer is to generate the emotional speech for an unseen target speakers and do not need to train the system again. Zero-shot voice conversion was studied extensively, while little work has been done on cross-speaker emotion transfer.

This paper describes a novel system for cross-speaker emotion transfer. The system comprises three core modules: emotion encoder, prosody predictor, and timbre encoder. The emotion encoder is used to extract discrete emotion types and emotion intensity values from input utterances. Given the emotion type and intensity, the prosody predictor generates prosodic features in accordance with the input phoneme sequence and emotion type. The timbre encoder provides the timbre information of output speech. If speech data of target speakers is available, the timbre encoder is realized as a trainable layer for embedding lookup. In the case of zero-shot emotion transfer, the timbre encoder consists of a speaker encoder and a bottleneck layer. The ground-truth prosodic features are used for system training. The system can remove the prosodic information but retain timbre-related information for the timbre encoder. The main contributions of this study are summarized as follows:

- A novel design of the TTS system named iEmoTTS for cross-speaker emotion transfer is proposed. To our knowledge, previous methods of cross-speaker emotion transfer were commonly based on speaker-emotion disentanglement. In iEmoTTS, we develop and apply a timbre-prosody disentanglement. The proposed iEmoTTS demonstrates superior performance to other emotion transfer systems reported recently. In the ablation study, iEmoTTS is compared with its variants to demonstrate the effectiveness of timbre-prosody disentanglement;
- An emotion encoder is specifically designed to produce discrete emotion type IDs for speech data of target speakers in training and is trained with other components in an end-to-end manner;
- A new method of probability-based emotion intensity control is developed and realized in the emotion encoder of iEmoTTS. With this method, iEmoTTS is able to produce speech with varying levels of emotion intensity;
- Zero-shot emotion transfer is achieved with iEmoTTS by a specifically designed timbre encoder with a bottleneck layer. Experimental results show that emotional information can be transferred to unseen target speakers while maintaining a certain degree of speaker similarity.

The remaining part of this paper proceeds as follows. Section II introduces the related work. Section III analyzes the emotional characteristics in the speech. The overview and each component of the proposed iEmoTTS are described in Section IV. Section V introduces the datasets, configurations and evaluation metrics for the experiments. Section VI presents the experiment results of the research.

II. RELATED WORKS

A. Cross-Speaker Emotion Transfer

Cross-speaker emotion or style transfer has been studied in the era of TTS [27], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41]. Existing approaches can be divided into two categories. Methods of the first category, e.g., [29], [30], [31], [32], [34], [37], [38], [39], [40], [42], require or assume that the given speech of target speakers is neutral.
Emotion transfer is carried out using a ‘neutral to emotional’ transformation learned from the speech of source speakers. Lorenzo-Trueba et al. [29] proposed an emotion transplantation method in the HMM-based speech synthesis framework. The acoustic features of neutral target speakers are transferred to the target speakers and styles by two different linear transformations. Whitehill et al. [34] applied the principle of adversarial cycle consistency [43] to encourage the synthesized speech to preserve the appropriate styles. Recently, Pan et al. [40] used a prosody bottleneck layer to predict prosodic features given source speakers’ speech and emotion type IDs. The predicted features were combined with speaker identity information of target speakers to generate the speech. Lu et al. [39] attempted to address the problem of cross-speaker emotion transfer by introducing phoneme-level latent features. Shechtman et al. [38] proposed a controllable style transfer with prosodic descriptors as input, allowing users to control the expression strength of each style.

In the second category of approaches, the speech utterances of target speakers are not presumed to be neutral. Instead, they are labelled using a SER classifier. The classifier can be either trained independently [33], [35], [36] or jointly with emotion transfer TTS systems [27]. Gao et al. [35] used a pre-trained classifier to annotate the speech of target speakers. A joint-training approach was proposed in [27]. A classifier mapped each pre-defined emotion token to an emotion type. The emotion embeddings were derived via attention between the prosodic representations [19] and emotion tokens. During the emotion transfer process, the emotion embedding was determined from emotion tokens given emotion type ID. A potential problem of this approach is that emotion embeddings in the training and transfer process might need to be more consistent.

B. Speech Factors Disentanglement

Speech factor disentanglement has been widely studied in TTS [19], [44], [45] and voice conversion [23], [46]. Skerry-Ryan et al. proposed to disentangle a prosodic representation from speech content and speaker characteristics [19]. The prosodic representation was extracted from mel-spectrograms of the reference speech. The representation is assumed to retain the pertinent prosodic information that can be used with the text transcript and speaker information to reconstruct reference speech. For cross-speaker emotion (style) transfer, a multi-reference TTS stylization system was presented in [37]. Each reference encoder was used to decompose and control a specific style class. Whitehill et al. [34] proposed a method based on adversarial cycle consistency to ensure all possible style combinations were seen. Qian et al. [46] developed a voice conversion system that disentangles speech content from timbre information. Furthermore, a voice conversion model [23] was introduced to decompose speech into four components: pitch, rhythm, content, and timbre.

C. Emotion Intensity Control

Variation of emotion intensity (strength) in synthesized speech can be achieved via a learned emotion representation (embedding) in an unsupervised manner. Straightforwardly, this can be done by applying a scalar weight to the emotion embedding [26], [27] in the inference process. This approach was shown effective in image style transfer [47]. Zhu et al. [24], [48] proposed to quantize emotion intensity by learning a rank function using the concept of relative attribute [49]. In order to learn the ranking function, prosodic features were extracted from <neutral, emotional> utterance pairs. The emotion intensity values were measured by weighting the prosodic features with learnable ranking weights and subsequently applied as an auxiliary input to train the TTS model. In [14], an emotion intensity control method was proposed to control the distinct characteristic of a target emotion category. An interpolation technique was introduced to control emotion intensity by gradually changing the emotional to neutral speech.

III. ANALYSIS OF EMOTION-RELATED CHARACTERISTICS IN SPEECH

A. Emotion Speech Corpora for Analysis

We use two corpora of Chinese speech\(^1\) to analyse the speech characteristics related to emotion. One corpus (Multi-S60-E3) has human-annotated emotion labels, and the other corpus (VA-S2) is unlabeled. The Multi-S60-E3 corpus consists of 30,000 utterances from 30 male and 30 female speakers, expressing four different emotions: anger, happiness, sadness, and neutrality. Each speaker uttered 125 sentences with different emotions. The VA-S2 corpus was designed for building TTS systems for voice assistant devices. It includes one female (F1) and one male (M1) speaker, each with about 2,000 utterances.

B. Analysis of Characteristics of Emotional Speech

We trained an SER model on the Multi-S60-E3 dataset, consisting of a feature extractor and a softmax layer. The feature extractor follows the CNN+RNN structure as in [19], [50]. It takes the mel-spectrogram of an utterance as input and outputs four-dimensional logits \(z\). The softmax layer converts the logits into a four-class posterior probability distribution. The SER model achieved an accuracy of 97.11\%. We applied the SER model to label the utterances in VA-S2, as shown in Table I. The results indicated that most utterances in VA-S2 were classified as non-neutral. Therefore, assuming that the speakers in VA-S2 were neutral would not be valid for emotion transfer.

We randomly selected 400 utterances from Multi-S60-E3 and 200 utterances from VA-S2. We visualized the hidden features of these utterances using UMAP [51] in Fig. 1. The utterances from Multi-S60-E3 formed four clusters, each corresponding to one of the four emotions. The utterances from VA-S2 were more distant from the cluster centers. Based on our informal listening

\(^1\) All corpora used in this work are from https://www.data-baker.com/.

|       | Sadness | Neutrality | Happiness | Anger |
|-------|---------|------------|-----------|-------|
| M1    | 4.15%   | 19.83%     | 73.07%    | 2.95% |
| F1    | 23.59%  | 16.38%     | 56.86%    | 3.17% |

TABLE I

The prediction results of utterances in VA-S2
test, the utterances of VA-S2 sounded perceptually consistent with the predicted emotion types but with lower intensity of emotion expression. In summary, it would be more appropriate to represent each utterance in VA-S2 with an emotion type and an intensity value.

We propose to measure the emotion intensity of an utterance using the posterior probability. During the training of the SER model, the optimizer increases $z_i$ and decreases $z_j \neq i$. The ratio of $z_i$ over $z_j \neq i$ can be regarded as the classifier’s confidence that the utterance belongs to emotion type $i$. Based on informal listening tests, we assume that the confidence level is related to the perceived emotion intensity. A higher confidence level implies stronger emotion intensity and vice versa. To normalize $z$ across different utterances, we use a modified softmax function that replaces the base $e$ with a hyper-parameter $\alpha$.

We tested different values of $\alpha$, such as 1.01, 1.2, and 2, on the VA-S2 corpus. Fig. 2 shows the histograms of $int_i$ for different emotion types. When $\alpha$ is close to 1, $int_i$ is close to $1/N$, where $N = 4$ in our experiment. When $\alpha$ is larger, such as 2, $int_i$ reaches 1. By choosing a suitable value of $\alpha$, such as 1.2, $int_i$ can vary from 0 to 1.

### IV. The Proposed iEmoTTS System

Fig. 3 illustrates the proposed iEmoTTS system. It consists of six major components: phoneme encoder, emotion encoder, prosody predictor, timbre encoder, speech decoder and vocoder. In the training process, iEmoTTS reconstructs the mel-spectrogram of an input speech utterance, given the phoneme sequence and speaker identity.

#### A. Model Overview

The phoneme sequence is mapped to phoneme embeddings by a lookup embedding layer. The embeddings are converted into phoneme encoding sequence by the phoneme encoder. The emotion type ID is passed through another lookup embedding layer and multiplied by the emotion intensity value to produce the emotion encoding. The emotion encoding is broadcast-added to the phoneme encoding sequence, producing a hidden sequence with dimension size 256 as the input of the prosody predictor. The 3-dimensional phoneme-level prosodic features are transformed into a sequence of 256-dimensional vectors by a 1-D convolutional layer. The vectors are added to the hidden sequence and up-sampled to frame-level representations. The timbre encoding is broadcast-added with the above representations and presented to the speech decoder to generate mel-spectrograms. A pre-trained vocoder converts the generated mel-spectrograms to the speech waveform.

The phoneme encoder and speech decoder remain the same as the standard FastSpeech, using the feed-forward Transformer (FFT) block as the basic unit. During modelling training process, the emotion type ID and intensity are derived from the emotion encoder given input mel-spectrograms. In the emotion transfer process, the emotion type ID and intensity are provided to the model by external control.

The iEmoTTS is designed to perform disentanglement between timbre and prosody. In this design, prosody is acoustically represented by pitch, intensity and duration. Disentanglement of prosody and timbre is achieved by (1) incorporating ground-truth prosodic features in model training; and (2) presenting the timbre encoding to the model after the prosodic features are incorporated. The prosodic information is provided first to the model. From an information-theoretic perspective, the iEmoTTS should be able to get rid of prosodic information from the timbre encoding but retain the timbre information.

#### B. Emotion Encoder

The emotion encoder extracts the emotion type ID and emotion intensity value from the mel-spectrogram of an utterance. The emotion encoder can produce discrete emotion type ID that matches the external control during inference, while being trained end-to-end. As shown in Fig. 4, the emotion encoder consists of a feature extractor, an adversarial speaker classifier, an emotion classifier, a modified softmax layer and a Straight-Through (ST) Gumbel-Softmax estimator.

Only utterances from source speakers have emotion labels. There are $N$ emotion types, including neutrality, in the source speakers’ data. The feature extractor described in Section III-B
transforms the mel-spectrograms from both source and target speakers into hidden features, which are partially de-speakerized by a gradient reversal layer [54] and a speaker classifier. The speaker classifier consists of a dense and softmax layer. The hidden features are then projected onto $M$-dimensional logits $z$ by a fully connected layer. Only the logits from source speakers are fed to the emotion classifier. The first $N$ dimensions of $z$ correspond to the $N$ emotion types. $M \geq N$ is required because some target speakers’ utterances may not have any known emotion types from source speakers. The emotion intensity scalar is obtained as the $i$th value from applying the modified softmax function to $z$.

The emotion type $i$ is obtained by using the Gumbel-Softmax distribution [55], [56] to approximate sampling from a categorical distribution $\pi$: 

$$y_i = \frac{\exp((\log(\pi_i) + g_i)/\tau)}{\sum_{j=1}^{M} \exp((\log(\pi_j) + g_j)/\tau)}, \quad i = 1, 2, \ldots, M$$  \hspace{1cm} (1)

where $g_i$ are i.i.d samples drawn from Gumbel(0,1) and $y$ denotes a sample from the Gumbel-Softmax distribution. This distribution allows us to obtain the emotion type $i$ as a discrete quantity that can be back-propagated. To do so, we use a small but non-zero temperature $\tau$ that controls the smoothness of the approximation. We also approximate $\log \pi_i$ by $z_i$. We then convert $y$ into one-hot discrete data $y'$ using $\arg\max$ operation to represent the emotion type. Finally, we use the gradient of $y'$ in the backward pass to approximate the gradient of $y$.

C. Prosody Predictor

The prosody predictor aims to predict F0, intensity and duration features. It consists of a six-layer 1D-convolutional network with ReLU activation, followed by layer normalization, a dropout layer and an extra linear layer. In addition to model size, the proposed prosody predictor differs from the variance adaptor [8] in several aspects. First, the variance adaptor consists of a sequence of independent predictors, each for a specific prosodic feature. The proposed predictor outputs a multiple-dimension vector in which F0, duration and intensity are represented by different dimensions. This design may help the model to capture the potential dependence among different prosodic features. Second, the proposed predictor generates prosodic features at the phoneme level instead of the frame level. Phoneme-level F0 and intensity values are obtained by averaging the frame-level values over the time intervals of the respective phonemes. On the contrary, the variance predictor outputs frame-level pitch and energy. Frame-level features are challenging to predict and would affect the model generalization capacity [57]. Lastly, prosodic features in iEmoTTS are mean-variance-normalized on individual speakers.

D. Timbre Encoder

Timbre encoder provides timbre-related information to the TTS model. For the cross-speaker emotion transfer task where the target speakers are seen in model training, a trainable lookup embedding layer with speaker ID as input is used to produce the timbre encoding. For zero-shot emotion transfer, a pre-trained speech encoder is employed to extract a speaker embedding [18] from the mel-spectrograms. The speech embedding is passed through a bottleneck layer to generate timbre encoding. The pre-trained speaker encoder follows the design in [17], which consists of a stack of two LSTM layers. The last time output of the LSTM layers is selected and projected down to a speaker embedding with a fully connected layer. The speaker embedding might contain redundant information regarding prosody and content. A bottleneck layer is leveraged to remove the
redundant information and keep most of the timbre information. The bottleneck layer takes the speaker embedding as the input and outputs timbre encoding. This paper implements the bottleneck layer by a modified Vector Quantized VAE (VQ-VAE) quantized layer [20], [58]. It learns a dictionary (codebook) \( \mathbf{E} \) with dictionary size as \( K \) and group number \( G \). The speaker embedding is divided equally into \( G \) groups and arranged as a matrix, where each column can be encoded by an integer index independently. Each index queries a corresponding embedding from the dictionary \( \mathbf{E} \), and all corresponding embeddings concatenate together as the timbre encoding. With a fixed dictionary size, the information bottleneck (IB) capacity of the bottleneck layer (i.e., \( G \log K \)) can be controlled by changing the group number \( G \). The larger the number \( G \), the higher the IB capacity, and the more information can pass through the bottleneck layer. This study sets the dictionary size \( K \) and group number \( G \) as 32 and 4, respectively. In training process, timbre encoding is different for each sentence. At the inference stage, the timbre encoding for each speaker is generated by feeding one or several utterance(s) of the same speaker to the timbre encoder and averaging the resulting sentence-level encoding.

\[ L = L_{\text{mel}} + \lambda_1 L_{\text{pros}} + \lambda_2 L_{\text{adv_spk}} + \lambda_3 L_{\text{emo_source}} \]  

where \( L_{\text{mel}} \) is the Mean Absolute Error (MAE) between the synthesized mel-spectrograms and the ground truth, \( L_{\text{pros}} \) is the L2 loss between the predicted and ground-truth prosodic features, and \( L_{\text{adv_spk}} \) and \( L_{\text{emo_source}} \) denote the cross-entropy losses for the adversarial speaker classifier and the emotion classifier, respectively. \( \lambda_1, \lambda_2, \text{ and } \lambda_3 \) are the hyper-parameters to balance the contributions of different losses. Since emotion and speaker are highly entangled in training data, a very small weight \( \lambda_2 \) is assigned to the loss function of the adversarial classifier \( L_{\text{adv_spk}} \) to prevent emotion information from being reduced. The \( \lambda_1, \lambda_2, \text{ and } \lambda_3 \) are set as 0.8, 0.01 and 0.5, respectively.

V. EXPERIMENTAL PROTOCOL

A. Data Preparation

Four speech corpora of Mandarin speech were used in the experiments. They are \textbf{Multi-S60-E3}, \textbf{Child-S1-E6}, \textbf{VA-S2} and \textbf{Read-S40}. The emotion labels of \textbf{Multi-S60-E3} and \textbf{Child-S1-E6} are available. \textbf{Multi-S60-E3} is a multi-speaker emotional corpus as described in Section III. The \textbf{Child-S1-E6} contains 12,000 speech utterances by a female voice actor imitating a child-like voice. There are six emotion types: happiness, amazement, anger, disgust, poorness, and fear. Each type has 2000 spoken utterances. The utterances in \textbf{VA-S2} and \textbf{Read-S40} are without emotion labels. \textbf{VA-S2} is described in Section III. The other one, named \textbf{Read-S40}, contains 20 male and 20 female speakers, each having about 500 utterances in reading style.

1) Speech Data of Source Speakers: The dataset \textbf{Child-S1-E6} covers six different emotion types. This corpus was included in all experiments. For cross-speaker emotion transfer, we involved additionally three female and three male speakers from \textbf{Multi-S60-E3} as source speakers. Hence there were seven source speakers and seven emotion types covered. While for zero-shot emotion transfer, in addition to \textbf{Child-S1-E6}, 15 female and 15 male speakers from \textbf{Multi-S60-E3} were used as source speakers to improve model generalizability [46]. As a result, there were 31 source speakers with seven emotion types. It should be noted that some emotion types, namely, amazement, disgust, fear, and poorness, were associated with source speakers only from \textbf{Child-S1-E6}.

2) Speech Data of Target Speakers: Two speakers from corpus \textbf{VA-S2} were used as target speakers for the cross-speaker emotion transfer task. The model did not see the target speakers in the training process of zero-shot emotion transfer. In this case, one female and one male speaker were selected randomly as target speakers from \textbf{Read-S40}. Only five utterances (around the 20 s) were used for each speaker.

B. Implementation Details

The raw Chinese text was converted into phoneme sequence by an open-sourced grapheme-to-phoneme tool.\(^2\) The Montreal forced alignment (MFA) tool\(^3\) was employed to obtain the phoneme boundaries and duration of the speech. 80-band mel-spectrograms were computed from raw speech waveforms with a frame length of 50 ms and frameshift of 12.5 ms. A pre-trained vocoder based on full-band HiFi-Gan [59] was adopted to transform the predicted mel-spectrograms to a speech waveform. The \( N \) and \( M \) in Section IV-B were set as 8 and 10. The iEmoTTS was trained with a batch size of 32 sentences on 4 NVIDIA V100 GPUs, using the Adam optimizer [60] and the learning rate schedule in [61]. It took 200 k steps for training until convergence.

C. Subjective Evaluation Methods

Crowdsourced Mean Opinion Score (MOS) was used for subjective evaluation of the aspects of emotion similarity, speaker similarity and voice quality of synthesized speech. The scores range from 1 to 5 in 0.5 point increments [62].

\textbf{Emotion Similarity:} Emotion similarity measures how well the synthesized speech sounds like the designated emotion types. Subjective evaluation started by arranging for each listener to listen to 5-10 sample utterances of each emotion type to become familiar with this emotion type. Subsequently, the listeners were asked to evaluate each audio sample based on how much it sounded like a given emotion type. The listeners were advised to ignore the contents of the utterance when scoring the emotion similarity. Listeners can move back to listen to the audio samples during the evaluation process.

\textbf{Speaker Similarity:} The speaker similarity test evaluates whether the synthesized utterances carry the speaker characteristics of target speakers. Before evaluating the test samples of each target speaker, listeners were asked to listen to 10 utterances

\(^2\)[Online]. Available: https://github.com/mozillazh/python-pinyin
\(^3\)[Online]. Available: https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner
to create an overall impression of this speaker. Then, listeners were asked to evaluate each audio sample based on how much it sounded like this target speaker.

Voice Quality: For voice quality assessment, listeners were presented with one utterance each time and asked to give a score regarding speech naturalness and pronunciation correctness.

The evaluation of emotion intensity control was carried out via an independent subjective test for each evaluation metric. We recruited 30 native Chinese speakers to listen to synthesized audio samples from different systems with the same text content and high, moderate and low emotion intensity values. The three utterances were presented in random order. The listener was asked to classify them as high, moderate and low emotion intensity. A good classification accuracy implies that control of emotion intensity is effective.

Each table in the following sections reports the results of an independent subjective test for each evaluation metric. We compared iEmoTTS with two recent cross-speaker emotion transfer models: Trans-CLN [27] and Trans-Pros [40]. Both models shared similar model parameters with iEmoTTS. Trans-CLN used a set of emotion tokens to represent different types of emotions. It adopted a semi-supervised training strategy to map the utterances of target speakers to the emotion tokens.

### VI. Experimental Results

#### A. Performance Evaluation

We compared iEmoTTS with two recent cross-speaker emotion transfer models: Trans-CLN [27] and Trans-Pros [40]. Both models shared similar model parameters with iEmoTTS. Trans-CLN used a set of emotion tokens to represent different types of emotions. It adopted a semi-supervised training strategy to map the utterances of target speakers to the emotion tokens. Trans-Pros predicted the bottlenecked prosodic features given the information of source speakers. The bottlenecked features were combined with the identity of target speaker and text to generate the mel-spectrograms. We evaluated the performance of iEmoTTS and the reference systems on three criteria: emotion similarity, speaker similarity, and voice quality. The results are shown in Table II. iEmoTTS outperforms both reference systems on most criteria. On emotion similarity, iEmoTTS and Trans-Pros show comparable performance, and both are significantly better than Trans-CLN. Trans-CLN does not achieve good emotion similarity, especially on low-arousal emotions such as sadness and poorness. On speaker similarity, iEmoTTS and Trans-CLN show similar performance, while Trans-Pros performs poorly. On voice quality, iEmoTTS performs significantly better than Trans-Pros and Trans-CLN. The poor performance of Trans-Pro might be due to two reasons: (1) The prosodic features were predicted from one source speaker and might retain speaker-specific information from that speaker. (2) The speaker embedding of the target speaker might retain the prosodic information of the target speaker.

| Emotion | Emotion Similarity MOS | Speaker Similarity MOS | Voice Quality MOS |
|---------|------------------------|------------------------|------------------|
|         | Trans-CLN | Trans-Pros | iEmoTTS | Trans-CLN | Trans-Pros | iEmoTTS |
| Happiness | 4.32 ± 0.07 | 4.53 ± 0.08 | 4.38 ± 0.06 | 4.00 ± 0.10 | 3.50 ± 0.14 | 4.03 ± 0.10 |
| Sadness  | 2.59 ± 0.08 | 4.32 ± 0.07 | 3.98 ± 0.08 | 3.60 ± 0.14 | 3.12 ± 0.12 | 3.79 ± 0.14 |
| Poorness  | 3.05 ± 0.10 | 4.12 ± 0.09 | 4.30 ± 0.07 | 3.74 ± 0.09 | 2.00 ± 0.16 | 3.87 ± 0.17 |
| Fear     | 3.22 ± 0.09 | 4.39 ± 0.08 | 4.61 ± 0.07 | 3.34 ± 0.34 | 1.50 ± 0.17 | 2.84 ± 0.12 |
| Anger    | 3.86 ± 0.09 | 3.80 ± 0.10 | 3.93 ± 0.08 | 4.31 ± 0.09 | 3.88 ± 0.06 | 4.22 ± 0.18 |
| Amusement| 3.58 ± 0.10 | 3.94 ± 0.10 | 3.81 ± 0.08 | 3.72 ± 0.10 | 3.38 ± 0.14 | 3.50 ± 0.15 |
| Disgust  | 3.42 ± 0.08 | 3.42 ± 0.09 | 3.64 ± 0.08 | 3.61 ± 0.12 | 3.88 ± 0.06 | 4.07 ± 0.07 |
| Average  | 3.43 ± 0.04 | 4.07 ± 0.03 | 4.09 ± 0.03 | 3.76 ± 0.04 | 3.04 ± 0.08 | 3.76 ± 0.05 |

The emotion similarity, speaker similarity and voice quality MOS with a 95% confidence interval are reported. A higher MOS value indicates better performance. The MOS values significantly higher than other models are in bold.

### B. Effectiveness of Disentanglement

To investigate the effect of timbre-prosody disentanglement on cross-speaker emotion transfer, we compared iEmoTTS with three variants that violated this condition: iEmoTTS-NP, iEmoTTS-SE and iEmoTTS-SENP. These variants had similar structures to iEmoTTS except for the following differences: iEmoTTS-NP did not use prosodic features as model input, and only combined emotion and phoneme encoding before up-sampling. The prosody predictor was only used to predict duration information for up-sampling. iEmoTTS-SE placed timbre encoding before the prosody predictor and combined it with emotion and phoneme encoding. The combined features were fed to the prosody predictor and then added with prosodic information as input to the speech decoder. iEmoTTS-SENP combined both differences of iEmoTTS-NP and iEmoTTS-SE. We evaluated the performance of iEmoTTS and its variants on three criteria: emotion similarity, speaker similarity, and voice quality. The results are shown in Table III.

The iEmoTTS model outperforms the three variant models in terms of emotion similarity for all emotion types. The differences are especially noticeable for sadness and poorness, which have low arousal levels and are rarely expressed by the target speakers who tend to have high arousal levels. Unlike the iEmoTTS model, which forces the timbre encoding to remove the prosodic information and retain only the timbre information, the variant models may preserve some speaker-related prosodic information in the timbre encoding. This makes it more difficult for them to synthesize emotions that are not present in the target speakers’ speech. As for speaker similarity, iEmoTTS-SE
and iEmoTTS-SENP achieve significantly higher scores than iEmoTTS and iEmoTTS-NP, respectively. This suggests that prosody cues also contribute to speaker perception, although timbre is the main factor. However, if the prosody of a synthesized utterance deviates too much from the target speakers’ recordings, the listeners may perceive lower speaker similarity. For instance, when we synthesize the emotion “fear” for a target speaker using the iEmoTTS model, the pitch level of the utterance is much lower than the average pitch of that speaker. Consequently, the speaker similarity score for “fear” is much lower than for other emotions. Previous studies have also reported that pitch level affects listeners’ perception of male and female voices [63]. The voice quality of iEmoTTS-SE and iEmoTTS-SENP is significantly higher than that of iEmoTTS and iEmoTTS-NP, respectively. This may be due to the fact that the prosodic features generated by iEmoTTS and iEmoTTS-NP differ from the prosodic features in the target speakers’ recordings. The mismatched or novel prosodic features may impair the model performance on voice quality. The lower performance of iEmoTTS-NP compared to iEmoTTS can be explained by the lack of F0 and intensity features. This is in line with the findings of [8], [64], which show that incorporating the F0 feature during training enhances the quality of synthesized speech.

C. Emotion Intensity Control

We set the intensity value to the posterior probability value of 0.1 for ‘low’ and 1 for ‘high’ emotion intensity levels, and used the median value of the training data statistics for the ‘moderate’ level. We compared this method with the method proposed by [26], [28], which we implemented by removing the modified softmax layer in iEmoTTS. We call the resulting system iEmoTTS-S. The iEmoTTS-S controls the emotion intensity by multiplying the emotion encoding by a scalar factor at the inference stage. The scalar is 0.5, 1.5, and 2.5 for the ‘low’, ‘moderate’, and ‘high’ intensity levels, respectively. Table IV shows the emotion intensity ranking test results for iEmoTTS and iEmoTTS-S. Fig. 5 illustrates the prosodic variation achieved by iEmoTTS and iEmoTTS-S by plotting the F0 curves of synthesized speech with different emotion intensity levels.

The proposed method can successfully and consistently realize different levels of emotion intensity. However, the performance is better for the source speakers than for the target speakers, as shown by Table IV. This may be due to the fact that prosody is not the only factor that affects the perception of emotion intensity. The proposed method outperforms the compared method in three aspects. First, it realizes emotion intensity with more nuanced and diverse variations. For example, higher emotion intensity results in a slower speaking rate for ‘sadness’, ‘fear’, ‘poorness’, and ‘disgust’. For ‘amazement’, higher intensity produces a low F0 at the beginning and a high F0 and slow speaking rate at the end to convey surprise. For ‘fear’, higher intensity generates speech with a sudden rise of F0 at the end of the first phrase. In contrast, iEmoTTS-S synthesizes speech with low pitch and slow speaking rate for low intensity and high pitch and fast speaking rate for high intensity for all emotion types. Second, in most cases, the proposed method achieves better classification performance in ranking the utterances by their intensity levels. The advantage is especially evident for ‘sadness’ and ‘poorness’. One unexpected finding is that the scalar-based method performs similarly to our method for some emotion types with high arousal levels. The scalar-based method creates speech with similar prosodic variations for the same emotion intensity level. The feedback from the test participants suggests that speech with relatively low pitch and slow tempo is often perceived as low intensity.
and speech with high pitch and fast tempo as high intensity for emotion types with high arousal, regardless of the actual emotion expression. Third, the voice quality of speech generated by iEmoTTS is higher than iEmoTTS-S. In the training of iEmoTTS-S, the emotion intensity value is always 1. However, in the inference, iEmoTTS-S uses different values of intensity. The mismatch between training and inference processes may cause the quality degradation.

D. Zero-Shot Emotion Transfer

The performance of iEmoTTS on zero-shot emotion transfer was evaluated in this section. In addition to the default setting, we evaluated different information bottleneck (IB) capacities for the bottleneck layer of the timbre encoder. More specifically, IB-full refers to the bottleneck layer removed after the speaker encoder. All information in the speaker vector is fed directly to the model with no constraint. IB-large is the system obtained by setting the group number in the modified VQ-VAE to 8, which is larger than the default setting. Therefore, the IB capacity of the bottleneck layer is larger, allowing more information to be passed through the bottleneck layer. IB-small works with a smaller IB capacity with a group number of 2. Subjective evaluation on emotion similarity and speaker similarity are detailed in Table V.

The results suggest that synthesized speech with the default IB setting achieves significantly better performance on emotion similarity than other settings and maintains good speaker similarity on the zero-shot emotion transfer task. It is interesting to note that the MOS on emotion similarity increases as the IB capacity decreases. The growth of emotion similarity is attributed to the narrowing bottleneck, which allows only the information pertinent to the timbre to remain in the timbre encoding. Compared with the case of seen target speakers (Table II), IB-default and IB-small settings can achieve comparable performance on unseen target speakers in the zero-shot scenario. As the IB capacity decreases, the MOS on speaker similarity decreases.
This may be explained by the model removing some of the timbre information when the information bottleneck is made narrow. It is also noted that the MOS on speaker similarity decreases slightly from IB-full to IB-default and drops abruptly from IB-default to IB-small setting. As the IB capacity shrinks, the model would be biased to discard most other information and a small portion of timbre information. If the information bottleneck is too narrow, the model may discard too much timbre information. This suggests a trade-off between emotion and speaker similarity as the IB capacity varies.

### E. Effectiveness of Emotion Encoder

This section evaluates the benefits of the proposed emotion encoder by comparing two variants of the iEmoTTS system: iEmoTTS-SER and iEmoTTS-WoEI. In iEmoTTS-SER, a SER model was trained with utterances of the source speakers following the structure described in Section III. Then, each utterance from target speakers was assigned an emotion label by inputting its mel-spectrograms into the SER model. Since all utterances had either ground-truth or predicted emotion labels, the emotion encoder and emotion intensity module were removed from iEmoTTS. In iEmoTTS-WoEI, the emotion intensity module of the iEmoTTS was removed such that the output of the emotion encoder contained only the discrete emotion type ID. The other components and parameters of iEmoTTS-SER and iEmoTTS-WoEI were identical to iEmoTTS.

Table VI presents the results on emotion similarity and voice quality for the three systems. The systems have comparable performance on voice quality. However, iEmoTTS-WoEI outperforms iEmoTTS-SER in terms of emotion similarity. A possible explanation is that the emotion encoder in iEmoTTS-WoEI was jointly trained with other components, which may enable it to better model emotion-related information in mel-spectrograms. The results also indicate that iEmoTTS achieves significantly higher emotion similarity than iEmoTTS-WoEI. This improvement could be attributed to the better emotion representation in iEmoTTS, where both emotion type ID and emotion intensity were utilized in model training.

### F. Effectiveness of Semi-Supervised Strategy

In this section, we evaluated the benefits of the semi-supervised strategy. In iEmoTTS, the emotion types for the utterances of target speakers were derived from semi-supervised learning and needed not to be neutral. Two variants of iEmoTTS were created for comparison, namely iEmoTTS-NEU and iEmoTTS-NEU2. In two variant models, utterances of target speakers were provided with neutral labels. The iEmoTTS-NEU and iEmoTTS-NEU2 had the same model structure as iEmoTTS. The only difference between the two variants was the neutral labels for target speakers. In iEmoTTS-NEU, the neutral label for target speakers was considered identical to that for source speakers. In iEmoTTS-NEU2, the target speakers were given a new neutral label, which was to be differentiated from the neutral label for source speakers. This variant was motivated by [28], where the model treated the labels of target speakers as a new type of neutral style. The evaluation results on emotion similarity and voice quality are shown in Table VII.

There is no significant difference between the three systems in the aspect of emotion similarity. However, iEmoTTS achieve better voice quality than both iEmoTTS-NEU and
iEmoTTS-NEU2. This reveals the effectiveness of the semi-supervised training strategy. In the training of iEmoTTS-NEU and iEmoTTS-NEU2, emotion labels and target speaker ID combinations were not seen. The generated mel-spectrograms tended to be degraded in the synthesis stage.

VII. CONCLUSION

An end-to-end TTS model for a cross-speaker emotion transfer system has been developed based on timbre-prosody disentanglement. The system, iEmoTTS, encodes emotion-related information in speech in terms of a discrete emotion type and a probability-based emotion intensity value. It is able to generate emotional speech for seen or unseen target speakers via a process of cross-speaker emotion transfer. Notably, the cross-speaker emotion transfer can be done even if the speaker and emotion are highly entangled in the given speech data of source speakers. Extensive experiments on subjective evaluation have been performed to assess iEmoTTS in various aspects. The evaluation results show that iEmoTTS performs better than a variety of reference systems and variants of itself. iEmoTTS shows positive results on zero-shot transfer to unseen target speakers. With a properly chosen information bottleneck capacity, the proposed model is able to achieve a good balance between speaker similarity and emotion similarity. Lastly, the proposed method of emotion intensity control has been shown effective such that the desired intensity level of speech emotion can be well recognized by human listeners.

REFERENCES

[1] A. J. Hunt and A. W. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 1996, vol. 1, pp. 373–376.
[2] P. A. Taylor, Text-to-Speech Synthesis. Cambridge, U.K.: Cambridge Univ. Press, 2009.
[3] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2000, vol. 3, pp. 1315–1318.
[4] H. Zen, K. Tokuda, and A. W. Black, “Statistical parametric speech synthesis,” Speech Commun., vol. 51, no. 11, pp. 1039–1064, 2009.
[5] Y. Qian, Y. Fan, W. Hu, and F. K. Soong, “On the training aspects of deep neural network (DNN) for parametric TTS synthesis,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2013, pp. 3829–3833.
[6] W. Wang et al., “Tacotron: Towards end-to-end speech synthesis,” in Proc. Interspeech, 2017, pp. 4006–4010.
[7] J. Shen et al., “Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2018, pp. 4779–4783.
[8] Y. Ren et al., “FastSpeech 2: Fast and high-quality end-to-end text to speech,” in Proc. Int. Conf. Learn. Representations, 2021.
[9] J. Kim, J. Kong, and J. Son, “Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech,” in Proc. Int. Conf. Mach. Learn., 2021, pp. 5530–5540.
[10] G. Zhang, S. Qiu, Y. Qin, and T. Lee, “Estimating mutual information in prosody representation for emotional prosody transfer in speech synthesis,” in Proc. IEEE 11th Int. Symp. Chin. Spoken Lang. Process., 2021, pp. 1–5.
[11] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, “Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis,” IEICE Trans. Inf. Syst., vol. 88, no. 3, pp. 502–509, 2007.
[12] Y. Lee, S.-Y. Lee, and A. Rabii, “Emotional end-to-end neural speech synthesizer,” in Proc. Adv. Neural Inf. Process. Syst. Neural Inf. Process. Syst. Found., 2017.
[13] J. Lorenzo-Trueba, G. E. Henter, S. Takaki, J. Yamagishi, Y. Morino, and Y. Ochiai, “Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis,” Speech Commun., vol. 99, pp. 135–143, 2018.
[14] S.-Y. Um, S. Oh, K. Byun, I. Jung, C. Ahn, and H.-G. Kang, “Emotional speech synthesis with rich and granularized control,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2020, pp. 7254–7258.
[15] A. Ortony and T. J. Turner, “What’s basic about basic emotions?,” Psychol. Rev., vol. 97, no. 3, 1990, Art. no. 315.
[16] A. Gibiansky et al., “Deep voice 2: Multi-speaker neural text-to-speech,” in Proc. Adv. Neural Inf. Process. Syst., 2017, vol. 30, pp. 2966–2974.
[17] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, “Generalized end-to-end loss for feature verification,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2018, pp. 4879–4883.
[18] Y. Jia et al., “Transfer learning from speaker verification to multispeaker text-to-speech synthesis,” in Proc. Adv. Neural Inf. Process. Syst., 2018, pp. 4485–4495.
[19] R. Skerry-Ryan et al., “Towards end-to-end prosody transfer for expressive speech synthesis with tacotron,” in Proc. Int. Conf. Mach. Learn., 2018, pp. 4693–4702.
[20] G. Zhang, Y. Qin, D. Tan, and T. Lee, “Applying the information bottleneck principle to prosodic representation learning,” in Proc. Interspeech, 2021, pp. 3156–3160.
[21] W.-N. Hsu et al., “Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2019, pp. 5901–5905.
[22] S. American National Standards Institute. Committee on Bioacoustics and M. Sonn, American National Standard Psychoacoustical Terminology. Washington, D.C., USA: American National Standards Institute, 1973.
[23] K. Qian, Y. Zhang, S. Chang, M. Hasegawa-Johnson, and D. Cox, “Unsupervised speech decomposition via triple information bottleneck,” in Proc. Int. Conf. Mach. Learn., 2020, pp. 7836–7846.
[24] X. Zhu, S. Yang, G. Yang, and L. Xie, “Controlling emotion strength with relative attribute for end-to-end speech synthesis,” in Proc. Autom. Speech Recognit. Understanding, 2019, pp. 192–199.
[25] A. Sorin, S. Shechtman, and R. Hoory, “Principal style components: Expressive style control and cross-speaker transfer in neural TTS,” in Proc. Interspeech, 2020, pp. 3411–3415.
[26] T. Li, S. Yang, L. Xue, and L. Xie, “Controllable emotion transfer for end-to-end speech synthesis,” in Proc. IEEE 12th Int. Symp. Chin. Spoken Lang. Process., 2021, pp. 1–5.
[27] P. Wu et al., “Cross-speaker emotion transfer based on speaker condition layer normalization and semi-supervised training in text-to-speech,” 2021, arXiv:2110.04153.

[28] 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.

[29] J. Lorenzo-Trueba, R. Barra-Chicote, O. Watts, and J. M. Montero, “Towards speaking style transplantation in speech synthesis,” in Proc. 5th ISCA Workshop Speech Synth., 2013, pp. 159–163.

[30] Y. Ohtani, Y. Nasu, M. Morita, and M. Akamine, “Emotional transplant emotional speech synthesis based on emotion additive model,” in Proc. Interspeech, 2015, pp. 274–278.

[31] K. Inoue, S. Hara, M. Abe, N. Hojo, and Y. Iijima, “An investigation to transplant speaker emotions in DNN-based TTS synthesis,” in Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Ann. Summit Conf., 2017, pp. 1253–1258.

[32] L. Chen, N. Braunischweiler, and M. J. Gales, “Speaker and expression factorization for audiobook data: Expressiveness and prosody,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 23, no. 4, pp. 605–618, Apr. 2015.

[33] Z. Hodari, O. Watts, S. Ronanki, and S. King, “Learning interpretable control dimensions for speech synthesis by using external data,” in Proc. Interspeech, 2018, pp. 32–36.

[34] M. Whitsill, S. Ma, D. McDuff, and Y. Song, “Multi-reference neutral TTS stylization with adversarial cycle consistency,” in Proc. Interspeech, 2020, pp. 4442–4446.

[35] Y. Gao, W. Zheng, Z. Yang, T. Köhler, C. Fuegen, and Q. He, “Interactive text-to-speech system via joint style analysis,” in Proc. Interspeech, 2020, pp. 4447–4451.

[36] X. Cai, D. Dai, Z. Wu, X. Li, J. Li, and H. Meng, “Emotion controllable speech synthesis using emotion-unlabeled dataset with the assistance of cross-domain speech emotion recognition,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2021, pp. 5734–5738.

[37] Y. Bian, C. Chen, Y. Kang, and Z. Pan, “Multi-reference tacotron by intercross training for style disentangling, transfer and control in speech synthesis,” in Proc. Interspeech, 2019, pp. 861–865.

[38] S. Shechtman, R. Fernandez, A. Sorin, and D. Haws, “Synthesis of expressive speaking styles with limited training data in a multi-speaker, prosody-controllable sequence-to-sequence architecture,” in Proc. Interspeech, 2021, pp. 4693–4697.

[39] C. Lu, X. Wen, R. Liu, and X. Chen, “Multi-speaker emotional speech synthesis with fine-grained prosody modeling,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2021, pp. 5729–5733.

[40] S. Pan and L. He, “Cross-speaker style transfer with prosody bottleneck in neural speech synthesis,” in Proc. Interspeech, 2021, pp. 979–983.

[41] Y. Shin, Y. Lee, S. Jo, Y. Hwang, and T. Kim, “Text-driven emotional style control and cross-speaker style transfer in neural TTS,” in Proc. Interspeech, 2022, pp. 2313–2317.

[42] L. Xue, S. Pan, L. He, L. Xie, and F. K. Soong, “Cycle consistent network for end-to-end style transfer TTS training,” Neural Netw., vol. 140, pp. 223–236, 2021.

[43] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 2223–2232.

[44] Y. Wang et al., “Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis,” in Proc. Int. Conf. Mach. Learn., 2018, pp. 5180–5189.

[45] X. An, F. K. Soong, and L. Xie, “Disentangling style and speaker attributes for TTS style transfer,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 30, pp. 846–858, 2022.

[46] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, “Autovc: Zero-shot voice style transfer with only autoencoder loss,” in Proc. Int. Conf. Mach. Learn., 2019, pp. 5210–5219.

[47] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 694–711.

[48] Y. Lei, S. Yang, and L. Xie, “Fine-grained emotion strength transfer, control and prediction for emotional speech synthesis,” in Proc. IEEE Spoken Lang. Technol. Workshop, 2021, pp. 423–430.

[49] D. Parikh and K. Grauman, “Relative attributes,” in Proc. IEEE Int. Conf. Comput. Vis., 2011, pp. 503–510.

[50] A. Satt, S. Rozenberg, and R. Hoory, “Efficient emotion recognition from speech using deep learning on spectrograms,” in Proc. Interspeech, 2017, pp. 1089–1093.
Jialun Wu received the M.S. degree from the School of Informatics, University of Edinburgh, Edinburgh, U.K., in 2017. He is currently a Senior Algorithm Engineer with Alibaba, Hangzhou, China. His research interests include speech processing, speech synthesis, natural language processing, and deep learning.

Mei Li received the M.S. degree in computer science from Northwestern Polytechnical University, Xi’an, China, in 2017. She is currently a Senior Algorithm Engineer with Alibaba, Hangzhou, China. Her research interests include deep learning models for speech processing, speech, and video synthesis.

Yutao Gai received the B.Eng. and M.Eng. degrees from the School of Software and Microelectronics, Northwestern Polytechnical University, Xi’an, China, in 2012 and 2015, respectively. He is currently a Staff Algorithm Engineer, the leader of Intelligent Synthesis Team of Artificial Intelligence Department in Intelligent Connectivity Business Unit, Alibaba Group. His research interests include speech synthesis, speech processing, speech recognition, speech enhancement, machine learning, and deep learning.

Feijun Jiang received the B.Eng. degree from Zhejiang University, Hangzhou, China, and the Ph.D. degree from the Hong Kong University of Science and Technology, Hong Kong. After Ph.D. degree, he joined the Alibaba Group as a Software Engineer on algorithm development. He is currently a Senior Staff Engineer, and the Director of the Department of Artificial Intelligence in Intelligent Connectivity Business Unit. He is currently directing the software system of Tmall Genie. He has authored or coauthored several papers in the conferences such as NeurIPS, SIGKDD, SIGIR.

Tan Lee (Member, IEEE) is currently a Professor with the Department of Electronic Engineering and an Associate Dean (Education) of the Faculty of Engineering, The Chinese University of Hong Kong, Hong Kong. His research interests include speech and audio signal processing, spoken language technology, deep learning models for speech and language, paralinguistics in speech, and neurological basis of speech and language. He led the effort on developing Cantonese-focused spoken language technologies that have been widely licensed for industrial applications. His recent work is focused on applying signal processing and machine learning methods to the classification and generation of atypical speech in challenging communication scenarios across different age groups. Tan Lee was an Associate Editor for the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING, an Area Chair of the Technical Programme Committees of INTERSPEECH 2014, 2016, and 2018, the General Co-Chair of ISCSLP 2018 and 2021. He was the Vice Chair of ISCA Special Interest Group on Chinese Spoken Language Processing. Tan Lee’s students were the recipient of the Best Student Paper awards at ICASSP 2005 and ICASSP 2019.