Emotion-Rich Cross-Speaker TTS via Contrastive Prosody Enhancement

Jen-Tzung Chien and Bryan Gautama Ngo
Institute of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
E-mail: {jtchien, bryangautama.ee11}@nycu.edu.tw

***Abstract*—Cross-speaker prosody transfer provides a meaningful approach to develop an emotional text-to-speech (TTS) system where emotionally diverse speech data are limited and only available from a single source speaker. Current approaches to prosody embedding learning for such a challenging task are suboptimal. A key issue is due to the contamination of non-prosodic features during the learning process for prosody transfer. This paper presents a new approach to disentangle non-prosodic phonetic features from emotional speech to ensure that the prosody embedding exclusively captures the prosodic characteristics from reference speech. In addition to the content disentanglement, the contrastive learning is merged to enhance the discrimination among different emotional features. Contrastive prosody modeling is performed to facilitate emotion-rich cross-speaker TTS. Experimental results demonstrate the merit of the proposed method in extracting emotional nuances and transferring them across speakers for emotional TTS in terms of mean opinion score for emotion similarity and naturalness.**

I. Introduction

The rapid advances in current text-to-speech (TTS) technology have assured significant improvement in the quality and naturalness of the synthesized speech [1], [2], [3]. However, standard TTS models only focused on producing intelligible and clear speech sounds. They often fell short of conveying and reflecting expressive emotions [4], [5] in human speech. Emotional TTS, which aims to produce the synthetic speech with appropriate emotional nuances, is increasingly important for many applications such as virtual assistants, audiobooks and conversational agents [6], [7], where the ability to express emotions can substantially enhance the user experience and engagement. One of the common approaches to develop emotional TTS is the cross-speaker prosody transfer method, which involves transferring the prosodic features from one reference speech to the synthesized speech while maintaining the specified speaker identity [8], [9]. In addition to offering a method for synthesizing emotional speech, it is also required to design a sophisticated learning approach to cope with the limitation due to data scarcity in the collection of emotionally diverse speech data from a single speaker [10], [11].

To make sure a reliable prosody transfer, the prosody embedding needs to be sufficiently extracted to reflect prosody information from a reference speech. However, achieving this goal is very challenging since speech signals contain not only prosody information like pitch, rhythm, energy and timbre but also non-prosody information such as the phonetic content [12]. Previous works have focused on disentangling the timbre which represents speaker information from the prosody representation of a speech signal. Such a disentanglement learning is useful to improve cross-speaker speech representation [8], [13]. However, the phonetic content contained in a reference speech often disturbs the prosody encoder in extracting the precise prosody information. The mixing of content and prosody in a speech signal considerably affects the cross-speaker prosody transfer for speech synthesis. Accordingly, the prosody disentanglement is crucial for transfer learning to ensure that the speaker's identity is preserved and the phonetic content remains intelligible. Alternatively, another issue in the prosody transfer for emotional TTS is to handle the learning procedure to discriminate different emotions from a small set of reference speech samples. Previous methods typically relied on the labeled emotion data and an auxiliary emotion classifier to enhance the emotion features within the prosody embeddings [14], [15]. While these approaches have shown some success, the main weakness was still caused due to the requirement of sufficient amount of emotion-labeled speech data.

To address the aforementioned issues, this paper presents a new approach to enhance emotional TTS based on a transfer learning for cross-speaker prosody features. The proposed method has twofold novelties. First, in addition to performing the speaker disentanglement, this paper introduces the content disentanglement by minimizing the mutual information [16] between prosody embedding and phoneme embedding. The ablation study shows that the proposed method assures that the prosody embedding exclusively contains the prosodic features from rhythm and pitch. Second, this paper proposes a contrastive learning method to enhance the emotional nuances within prosody features instead of relying on the additional emotion classifier loss. This study addresses different learning objectives for content disentanglement and contrastive prosody transfer. A set of experiments and analyses in both English and Mandarin settings demonstrate the merit of the proposed method in synthesizing cross-speaker emotional speech.

II. Prosody Disentanglement and Enhancement

This paper addresses the approach to enhance emotional TTS through cross-speaker prosody transfer. Before delving into the specifics of the approach, which includes feature disentanglement and contrastive learning [17], the concept of cross-speaker prosody transfer is first introduced.

979-8-3315-7206-8/25/$31.00

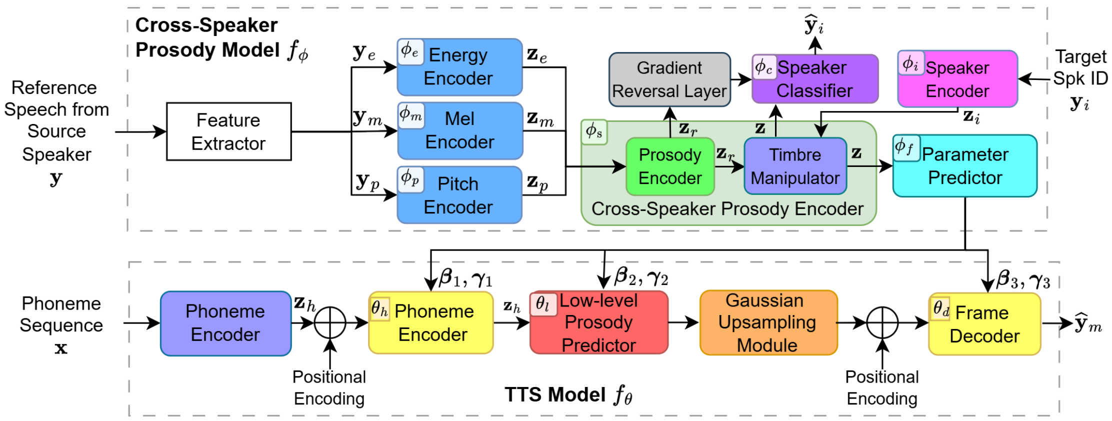

Fig. 1. Architecture for cross-speaker emotional text-to-speech driven by a cross-speaker prosody model f_ϕ and an TTS model f_θ where a source speaker in an emotional reference speech $\mathbf{y}$ is transferred to a target speaker $\mathbf{y}_i$ in a synthesized Mel spectrum $\widehat{\mathbf{y}}_m$ spoken in a phoneme sequence $\mathbf{x}$.

A. *Cross-Speaker Prosody Transfer*

This study aims to develop an emotional TTS model capable of synthesizing the Mel-spectrum of emotional speech $\widehat{\mathbf{y}}_m$ from a reference speech $\mathbf{y}$ of a source speaker through cross-speaker prosody transfer given by a phoneme sequence $\mathbf{x}$ and a target speaker identity $\mathbf{y}_i$. To achieve the goal, this paper employs a TTS architecture comprising two main components which are functioned as a cross-speaker prosody model $f_\phi = \{f_{\phi_i}, f_{\phi_c}, f_{\phi_p}, f_{\phi_m}, f_{\phi_e}, f_{\phi_s}, f_{\phi_f}\}$ and a TTS model $f_\theta = \{f_{\theta_h}, f_{\theta_l}, f_{\theta_d}\}$, as illustrated in Figure 1. The cross-speaker prosody model is designed to extract the salient prosodic features from the reference speech $\mathbf{y}$ by a source speaker and the target speaker identity $\mathbf{y}_i$, conveying the information of pitch, rhythm, and timbre. Four different inputs, consisting of speaker embedding $\mathbf{z}_i$ via f_{ϕ_i}, pitch embedding $\mathbf{z}_p$ via f_{ϕ_p}, Mel-spectrum embedding $\mathbf{z}_m$ via f_{ϕ_m}, and energy embedding $\mathbf{z}_e$ via f_{ϕ_e}, are calculated and fed into the cross-speaker prosody encoder f_{ϕ_s}. This prosody encoder calculates the reference embedding $\mathbf{z}_r$ from a source speaker, which is combined with the speaker embedding $\mathbf{z}_i$ of a target speaker $\mathbf{y}_i$ to obtain a cross-speaker mixture embedding $\mathbf{z}$ through a timbre manipulator for speech synthesis using TTS model f_θ.

To facilitate transfer learning for cross-speaker prosody information, it is crucial to disentangle original timbre features from the reference speech $\mathbf{y}$. In line with the previous work [8], this study employs a speaker classifier model f_{ϕ_c}, producing the estimated speaker identity $\widehat{\mathbf{y}}_i$ where the gradient reversal layer (GRL) [18] is applied to eliminate timbre information from the extracted prosody feature $\mathbf{z}_r$. The estimated speaker identity $\widehat{\mathbf{y}}_i$ is accordingly improved. To provide the conditional prosody information to TTS model f_θ, a parameter predictor module f_{ϕ_f} produces the control parameters $\{\beta_j, \gamma_j\}_{j=1}^{3}$, which are then fed into a feature-wise linear modulation (FiLM) conditioned on the layers of phoneme encoder f_{θ_h}, low-level prosody predictor f_{θ_l} with Gaussian upsampling module and frame decoder f_{θ_d} within the TTS model f_θ where the position encodings in both phoneme encoder and frame decoder are merged. The Mel-spectrum $\widehat{\mathbf{y}}_m$ via cross-speaker prosody transfer is estimated for synthesizing a waveform through a vocoder. This study works on cross-speaker prosody transfer, which is leveraged by learning representation based on the content disentanglement.

B. *Exclusive Prosody Disentanglement*

The concept of content disentanglement aims to minimize the dependency between the phoneme embedding $\mathbf{z}_h$ via f_{θ_h}, which holds the phonetic content, and the prosody embedding $\mathbf{z}$ via f_{ϕ_s}, which preserves the prosody information. This disentanglement ensures that $\mathbf{z}$ captures only prosody-related features $\mathbf{z}_r$, thereby enhancing the prosody representation from reference speech $\mathbf{y}$. In contrast to speaker information disentanglement, disentangling the latent content $\mathbf{z}$ from reference speech $\mathbf{y}$ is notably challenging due to the mixing connection between latent content $\mathbf{z}$ and phoneme embedding $\mathbf{z}_h$, which involves a wide range of possible candidates. Using a strategy similar to that for speaker information disentanglement may lead to suboptimal results.

To guarantee such a disentanglement, this work presents an approach to mutual information minimization between $\mathbf{z}$ and $\mathbf{z}_h$. Notably, $\mathbf{z}$ is the output of cross-speaker prosody encoder f_{ϕ_s} which contains the prosody information from the reference speech $\mathbf{y}$ and the target speaker identity $\mathbf{y}_i$. Meanwhile, the phoneme embedding $\mathbf{z}_h$ is produced exclusively from the phoneme sequence $\mathbf{x}$ through phoneme encoder f_{θ_h}. Hence, minimizing the mutual information between $\mathbf{z}$ and $\mathbf{z}_h$ paves a way to disentangle the phonetic content information $\mathbf{z}_h$ from the prosody embedding $\mathbf{z}$. Specifically, a variational form of contrastive log-ratio upper bound (vCLUB) [19] is utilized to estimate the upper bound of mutual information $I_{\text{vCLUB}}(\mathbf{z}, \mathbf{z}_h)$ between prosody embedding $\mathbf{z}$ and phoneme embedding $\mathbf{z}_h$. In the original CLUB method, the conditional probability $p(\mathbf{z}_h|\mathbf{z})$ is assumed to be known in calculation of upper bound

$$\begin{aligned} I_{\text{CLUB}}(\mathbf{z}, \mathbf{z}_h) =& \mathbb{E}_{p(\mathbf{z}, \mathbf{z}_h)} \left[\log p(\mathbf{z}_h|\mathbf{z})\right] \\ &- \mathbb{E}_{p(\mathbf{z})} \mathbb{E}_{p(\mathbf{z}_h)} \left[\log p(\mathbf{z}_h|\mathbf{z})\right]. \end{aligned} \tag{1}$$

However, since the conditional relationship between the variables is unavailable, this work employs the vCLUB method to determine the variational upper bound

$$I_{\text{vCLUB}}(\mathbf{z}, \mathbf{z}_h) = \mathbb{E}_{p(\mathbf{z}, \mathbf{z}_h)}[\log q_\psi(\mathbf{z}_h|\mathbf{z})] - \mathbb{E}_{p(\mathbf{z})}\mathbb{E}_{p(\mathbf{z}_h)}[\log q_\psi(\mathbf{z}_h|\mathbf{z})] \tag{2}$$

where $q_\psi(\mathbf{z}_h|\mathbf{z})$ denotes the variational probability in variational inference given by variational parameter ψ, which is used to approximate the intractable probability $p(\mathbf{z}_h|\mathbf{z})$ in the latent variable model. The unbiased estimator for vCLUB is then calculated from a set of latent samples $\{\mathbf{z}^i, \mathbf{z}_h^i\}_{i=1}^N$ from N speakers based on the calculation of empirical expectation

$$\mathcal{L}_{\text{mi}} \triangleq \widehat{I}_{\text{vCLUB}}(\mathbf{z}, \mathbf{z}_h) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\Big[\log q_\psi(\mathbf{z}_h^i|\mathbf{z}^i) - \log q_\psi(\mathbf{z}_h^j|\mathbf{z}^i)\Big]. \tag{3}$$

By minimizing Eq. (3), the dependency between $\mathbf{z}$ and $\mathbf{z}_h$ is reduced to pursue prosody disentanglement. The disentanglement loss based on mutual information $\mathcal{L}_{\text{mi}}$ is constructed.

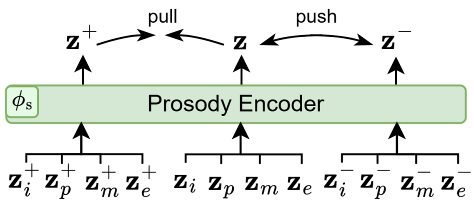

Fig. 2. Estimation of prosody encoder f_{ϕ_s} through contrastive loss which pulls together positive pairs $\{\mathbf{z}^+, \mathbf{z}\}$ and push away negative pairs $\{\mathbf{z}, \mathbf{z}^-\}$.

C. Contrastive Prosody Enhancement

In addition to guiding the prosody embedding $\mathbf{z}$ to be exclusive from phoneme embedding $\mathbf{z}_h$, this study further enhances the emotional nuances [10] in prosody embedding $\mathbf{z}$ by using the contrastive learning method [20], [21]. To do so, supervised contrastive learning [22] is explored to encourage the cross-speaker prosody model to capture the discriminative prosody embedding $\mathbf{z}$ towards different emotions by leveraging the emotion label information. The goal of this prosody guidance is to discriminate the transfer learning towards separate emotions, and make each emotion preserve its own characteristics. Four different inputs, embedded from speaker $\mathbf{z}_i$, pitch $\mathbf{z}_p$, Mel-spectrum $\mathbf{z}_m$ and energy $\mathbf{z}_e$ in the cross-speaker prosody encoder f_{ϕ_s}, denoted by $\{\mathbf{z}_i, \mathbf{z}_p, \mathbf{z}_m, \mathbf{z}_e\}$, are required in the contrastive learning process. This process is carried out within the cross-speaker prosody encoder f_{ϕ_s}. In general, the cross-speaker prosody embedding $\mathbf{z}$ can be expressed through an integrated prosody encoder

$$\mathbf{z} = f_{\phi_s}(\mathbf{z}_i, \mathbf{z}_p, \mathbf{z}_m, \mathbf{z}_e). \tag{4}$$

This process particularly utilizes the information noise-contrastive estimation (InfoNCE) [23] to calculate the contrastive loss $\mathcal{L}_c$ from the positive samples $\mathcal{D}^+ = \{\mathbf{z}_i^+, \mathbf{z}_p^+, \mathbf{z}_m^+, \mathbf{z}_e^+\}$ relative to anchor samples $\mathcal{D} = \{\mathbf{z}_i, \mathbf{z}_p, \mathbf{z}_m, \mathbf{z}_e\}$ given by the same emotion labels and the negative samples $\mathcal{D}^- = \{\mathbf{z}_i^-, \mathbf{z}_p^-, \mathbf{z}_m^-, \mathbf{z}_e^-\}$ relative to $\mathcal{D}$ given by different emotion labels as illustrated in Figure 2. Contrastive loss for emotion discrimination is calculated by

$$\mathcal{L}_c = -\log \frac{\exp((\mathbf{z}\cdot\mathbf{z}^+)/\tau)}{\exp((\mathbf{z}\cdot\mathbf{z}^+)/\tau) + \sum_{k=1}^{N^-}\exp((\mathbf{z}\cdot\mathbf{z}_k^-)/\tau)} \tag{5}$$

where $\mathbf{z}$, $\mathbf{z}^+$, and $\mathbf{z}^-$ denote the prosody embeddings associated with the anchor sample, positive sample and negative sample, produced by the cross-speaker prosody encoder f_{ϕ_s}, respectively, τ denotes a temperature parameter, and N^- denotes the number of negative samples. Notably, similar to [23], no data augmentation procedure is utilized under this supervised setting. Instead, the prosody embeddings for $\mathbf{z}^+$ and $\mathbf{z}^-$ are generated by using the reference speech samples from $\mathcal{D}^+$ and $\mathcal{D}^-$, respectively. This contrastive loss $\mathcal{L}_c$ is minimized to pull together those positive pairs $\{\mathbf{z}^+, \mathbf{z}\}$ and push away those negative pairs $\{\mathbf{z}, \mathbf{z}^-\}$ [24].

This study further explores the utilization of unsupervised contrastive learning for enhancing the emotional nuances of prosody embedding $\mathbf{z}$ since the labeled emotion data are typically limited. The key idea in unsupervised contrastive learning is based on the data augmentation for generating positive samples $\mathbf{z}^+$. This work adopts the vocal tract length perturbation (VTLP) [25] as an augmentation module f_a to obtain the positive samples from individual Mel-spectrum samples $\mathbf{y}_m$ and their augmented samples $\mathbf{y}_m^+ = f_a(\mathbf{y}_m)$. The corresponding embedding is obtained by $\mathbf{z}_m^+ = f_{\phi_m}(\mathbf{y}_m^+)$. Basically, VTLP modifies the timbre of the Mel-spectrogram of $\mathbf{y}$ by warping its frequency with a warping factor α. VTLP has been extensively used as a Mel-spectrogram augmentation method in various tasks [26], [27]. Meanwhile, the negative examples are obtained by sampling the other reference speech $\mathbf{y}$ from the dataset without implementing any augmentation. The self-supervised learning objective under this unsupervised setting is identical to Eq. (5).

D. Overall Learning Objective

With the additional losses proposed in this study, including the content disentanglement loss by minimizing $\mathcal{L}_{\text{mi}}$ and the emotion discrimination loss by minimizing $\mathcal{L}_{\text{c}}$, the prosody enhancement objective driven by the corresponding regularization parameters λ_m and λ_c, respectively, is constructed and merged in an overall learning objective $\mathcal{L}$ by

$$\mathcal{L} = \lambda_e\mathcal{L}_e + \lambda_f\mathcal{L}_f + \lambda_r\mathcal{L}_r + \lambda_p\mathcal{L}_p - \lambda_i\mathcal{L}_i + \lambda_m\mathcal{L}_{\text{mi}} + \lambda_c\mathcal{L}_c. \tag{6}$$

In this overall loss, there are some other existing losses. First, the emotion classifier loss $\mathcal{L}_e$ is calculated by a cross-entropy loss, which is minimized to pursue accurate emotion prediction. Second, the loss $\mathcal{L}_f$ for parameter predictor serves to regulate the magnitude of the FiLM parameters $\{\gamma_i, \beta_i\}_{i=1}^3$, similar to the approach in [28], while the loss $\mathcal{L}_r$ is associated with the reconstruction error between the ground-truth Mel-spectrum $\mathbf{y}_m$ and the estimated Mel-spectrum $\widehat{\mathbf{y}}_m$

$$\mathcal{L}_r = \mathcal{L}_{\text{mse}}(\mathbf{y}_m, \widehat{\mathbf{y}}_m) + \mathcal{L}_{\text{mae}}(\mathbf{y}_m, \widehat{\mathbf{y}}_m) \tag{7}$$

which consists of the mean square error (mse) and the mean absolute error (mae). Next, the prosody feature loss $\mathcal{L}_p$ is calculated by

$$\begin{aligned}\mathcal{L}_p =&\mathcal{L}_{\text{mse}}(\mathbf{y}_e, f_{\theta_l}(\mathbf{z}_h, \boldsymbol{\beta}_2, \boldsymbol{\gamma}_2)) + \mathcal{L}_{\text{mse}}(\mathbf{y}_p, f_{\theta_l}(\mathbf{z}_h, \boldsymbol{\beta}_2, \boldsymbol{\gamma}_2)) \\ &+ \mathcal{L}_{\text{mse}}(\mathbf{y}_d, f_{\theta_l}(\mathbf{z}_h, \boldsymbol{\beta}_2, \boldsymbol{\gamma}_2))\end{aligned} \tag{8}$$

where the mean square errors due to the estimation of energy $\mathbf{y}_e$, pitch $\mathbf{y}_p$ and duration $\mathbf{y}_d$ via low-level prosody prediction f_{θ_l} are involved to resemble the conventional loss for the core acoustic model as described in [2], but with an additional mean absolute error for Mel-spectrogram $\mathbf{y}_m$ as given in Eq. (7). Also, the speaker classifier loss is represented by $\mathcal{L}_i$ based on the cross-entropy loss between ground-truth speaker label $\mathbf{y}_i$ and estimated speaker posterior $\widehat{\mathbf{y}}_i = f_{\phi_i}(\mathbf{z})$ using the cross-speaker mixed prosody embedding $\mathbf{z}$ via

$$\mathcal{L}_i = -\mathbb{E}[\mathbf{y}_i \log(f_{\phi_i}(\mathbf{z}))]. \tag{9}$$

Here, the negative sign indicates the use of gradient reversal to promote the disentanglement of speaker information. The overall loss $\mathcal{L}$ is minimized with respect to cross-speaker prosody model f_ϕ and TTS model f_θ to train the emotion-rich cross-speaker TTS based on the proposed contrastive prosody transfer via latent disentanglement and discrimination.

III. Experiments

A. Experimental Settings

The experiments on cross-speaker text-to-speech were conducted in two languages including English (EN) and Mandarin (ZH). The model training was divided into two phases. First, TTS model was trained from scratch by using non-emotional speech datasets to ensure that the trained model could generate speech with acceptable quality. English setting employed the LJSpeech [29] and VCTK (https://datashare.ed.ac.uk/handle/10283/2950) datasets, while Mandarin setting employed the Aishell-3 [30] dataset for this phase. In the subsequent phase, the model was fine-tuned by using the emotional datasets with the details mentioned in Table I including emotion labels. The emotional speech dataset (ESD) [31] was utilized. This study introduced a new emotion label termed “other” in the IEMOCAP dataset [32] to account for instances labeled with multiple emotions. Additionally, the instances labeled with “fear” and “disappointed” were included under the “other” emotion category due to their infrequency.

TABLE I
Emotional datasets used in the second phase of model training. ‘Lang’ indicates language, ‘Spk’ indicates number of speakers, and ‘Dur’ indicates total duration in hours.

Dataset	Lang	Spk	Dur	Emotion
IEMO CAP	EN	10	10h	angry, excited, fear, sad, surprised, frustrated, happy, disappointed, neutral
ESD	EN	10	13h	neutral, angry, sad, happy, surprised
ESD	ZH	13	15h	neutral, angry, sad, happy, surprised

Different methods were adopted in the second phase of model training. Given the scarcity of publicly available TTS models capable of cross-speaker prosody transfer, this study emulated the previous methods by integrating an auxiliary emotion classifier, and implemented the following six distinct variants of emotion-rich cross-speaker TTS for ablation study where the contrastive loss $\mathcal{L}_c$ is further divided into the supervised contrastive loss $\mathcal{L}_{\text{sc}}$ and the unsupervised contrastive loss $\mathcal{L}_{\text{uc}}$ as shown in the following:

– FT: fine-tune the model using emotional dataset without additional loss as in the first phase [8],
– EC: incorporating the emotion classifier loss $\mathcal{L}_e$,
– MI: incorporating the mutual information loss $\mathcal{L}_{\text{mi}}$,
– SC: incorporating the supervised contrastive loss $\mathcal{L}_{\text{sc}}$,
– UC: incorporating the unsupervised contrastive loss $\mathcal{L}_{\text{uc}}$,
– MI+SC: incorporating both mutual information loss $\mathcal{L}_{\text{mi}}$ and supervised contrastive loss $\mathcal{L}_{\text{sc}}$.

All solutions shared an identical backbone architecture based on the Daft-Exprt [8] which was built for expressive TTS. FT and EC were viewed as the baselines. The model was trained with the same setting as in [8]. The emotion classifier loss in EC was implemented by using the cross-entropy loss, utilizing a classifier network. The hyperparameters of emotion classifier loss λ_e, mutual information loss λ_m and contrastive loss λ_c were set to 0.01. The remaining hyperparameters were fixed as $\lambda_f = 0.001$, $\lambda_r = 1$, $\lambda_p = 1$, and λ_i was linearly increased from 0 to 0.01 during the first 10K steps. In the SC setting, 7 samples were used as the negative samples corresponding to each emotion class. Meanwhile, for UC setting, 8 random samples were selected to serve as the negative pairs $\mathcal{D}^-$. For the UC, the uniform warping factor $\alpha \sim U(0.9, 1.1)$ was set for the VTLP module. After the TTS model training, fine-tuning was performed on a pre-trained 22KHz universal HiFi-GAN vocoder [33] to serve as the vocoder for speech synthesis [34] from the estimated Mel-spectrum $\widehat{\mathbf{y}}_m$. For data preprocessing, this paper referred to [8]. For dataset division, 3-4% of the total samples were allocated as the test set. Meanwhile, since the proposed model required sufficient training samples to obtain the desired performance, then less than 1% of the total samples were designated as the validation set.

TABLE II
Comparison of MOS for emotion similarity using different methods under different datasets.

Method	English		Mandarin
	ESD	IEMOCAP	ESD
FT	3.35 ± 0.69	3.10 ± 0.53	3.38 ± 0.39
EC	3.27 ± 0.66	3.15 ± 0.42	3.23 ± 0.49
MI	3.50 ± 0.56	3.66 ± 0.52	3.73 ± 0.58
SC	3.46 ± 0.56	3.54 ± 0.44	3.62 ± 0.78
UC	3.40 ± 0.54	3.61 ± 0.29	3.69 ± 0.29
SC+MI	$\mathbf{3.58 \pm 0.55}$	$\mathbf{3.88 \pm 0.16}$	$\mathbf{3.77 \pm 0.25}$

B. Experimental Results

This work adopted the mean opinion score (MOS) as a metric to assess the emotion similarity, evaluating the cross-speaker prosody transfer both on emotion similarity and nat-

TABLE III
COMPARISON OF MOS FOR NATURALNESS USING DIFFERENT METHODS UNDER DIFFERENT DATASETS.

Method	English		Mandarin
	ESD	IEMOCAP	ESD
FT	3.23 ± 0.72	3.17 ± 0.58	3.08 ± 0.46
EC	3.31 ± 0.37	3.15 ± 0.47	3.04 ± 0.42
MI	3.62 ± 0.54	3.68 ± 0.36	3.65 ± 0.53
SC	3.65 ± 0.30	3.61 ± 0.48	3.54 ± 0.48
UC	3.38 ± 0.47	3.66 ± 0.22	3.58 ± 0.47
SC+MI	$\mathbf{3.73 \pm 0.58}$	$\mathbf{3.76 \pm 0.28}$	$\mathbf{3.77 \pm 0.33}$

uralness to ensure the quality of the synthesized speech after embedding and processing the emotions. Speech samples were generated by conditioning on the model with the reference speech $\mathbf{y}$ for each emotion in the dataset and the random speaker identity $\mathbf{y}_i$ as the speaker. There were 15 evaluators involved in MOS evaluation. Human evaluators provided subjective ratings based on their perception of the emotional accuracy and the naturalness of the synthesized speech samples in a range between 1 (the worst) and 5 (the best). The averaged values over all test samples and evaluators were measured. Tables II and III report the MOS results for emotion similarity and naturalness, respectively. The baseline systems based on FT and EC (adding emotion classifier loss $\mathcal{L}_e$) show moderate performance, indicating that traditional cross-entropy loss for emotion classification may not be optimal. The MI model, employing the mutual information loss $\mathcal{L}_{\text{mi}}$ to disentangle the phonetic content $\mathbf{z}_h$ from the prosody information $\mathbf{z}$, shows notable improvement, demonstrating the effectiveness of content disentanglement. Both SC and UC models, utilizing the supervised and unsupervised contrastive losses $\mathcal{L}_{\text{sc}}$ and $\mathcal{L}_{\text{uc}}$, respectively, outperform the baseline models based on FT and EC. In most cases, SC shows a stronger performance than UC in both MOS metrics. However, the combined model SC+MI, where both losses $\mathcal{L}_{\text{mi}}$ and $\mathcal{L}_{\text{sc}}$ are considered, consistently achieves the highest MOS scores across all datasets, indicating that the integration of content disentanglement and emotion discrimination substantially enhances the prosody transfer for emotion information without compromising the naturalness in the cross-speaker TTS. Overall, these findings demonstrate that disentangling the content information and enhancing the prosody embeddings through information-theoretic learning and contrastive learning are effective strategies for improving both expressiveness and quality in the synthesized speech.

C. Experimental Analysis

Furthermore, Figure 3 visualizes the cross-speaker prosody embedding $\mathbf{z}$ of using standard fine-tuning (FT) and the proposed method MI+SC where the visualization using t-distributed stochastic neighbor embedding [35], [36]. The visualization shows that the proposed method results in more distinct and clustered than baseline method for individual emotions, indicating the effectiveness of contrastive learning in enhancing the prosody transfer. In contrast, the FT shows multiple mini clusters across the latent space. These mini

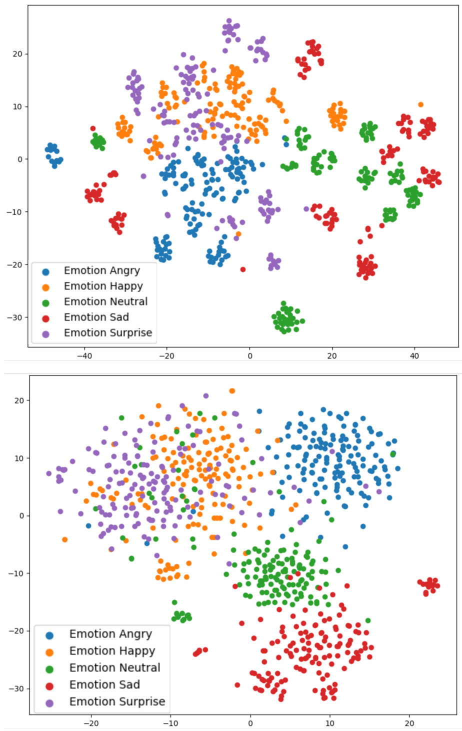

Fig. 3. Visualization of the prosody embeddings $\mathbf{z}$ by using FT (top) and the proposed MI+SC (bottom). Different emtions are shown by different colors.

clusters corresponds to different speakers since the number of speakers is almost the same as the number of mini clusters. This finding suggests that, despite the speaker disentanglement offered in [8], the prosody embedding using the method in [8] still retains some speaker-specific characteristics. Proposed method, however, eliminates these mini clusters, demonstrating that minimizing the mutual information not only disentangles the phonetic content from prosody enhancement but also effectively removes the residual speaker information, leading to more robust prosody embeddings that better capture the emotional nuances. Listening evaluation over different variants for ablation study is provided in this paper. The TTS samples using the proposed method are accessible in https://heizzzzz.github.io/.

IV. CONCLUSIONS

This study presents an enhanced emotional TTS for cross-speaker prosody transfer, which integrated different information sources from phoneme sequence, reference speech of a source speaker, and target speaker identity. The improvement was analyzed and realized through content disentanglement from the prosody embedding by minimizing the mutual information loss. Furthermore, the contrastive learning technique was employed to enhance the emotional nuances of the prosody embedding. Experimental results on cross-speaker emotion-rich TTS in both English and Mandarin demonstrate that the proposed method achieves the desirable MOS for naturalness and emotional similarity, thus confirming its efficacy in the downstream task.

References

[1] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu, "Neural speech synthesis with transformer network," in *Proc. of AAAI Conference on Artificial Intelligence*, 2019, pp. 6706–6713.

[2] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *Proc. of International Conference on Learning Representations*, 2021.

[3] Li-Jen Yang and Jen-Tzung Chien, "Continual gated adapter for bilingual codec text-to-speech," in *Proc. of Conference of the Oriental COCOSDA*, 2024, pp. 1–6.

[4] Deok-Hyeon Cho, Hyung-Seok Oh, Seung-Bin Kim, and Seong-Whan Lee, "EmoSphere++: Emotion-controllable zero-shot text-to-speech via emotion-adaptive spherical vector," *IEEE Transactions on Affective Computing*, 2025.

[5] Xiang Li, Zhi-Qi Cheng, Jun-Yan He, Junyao Chen, Xiaomao Fan, Xiaojiang Peng, and Alexander G Hauptmann, "UMETTS: A unified framework for emotional text-to-speech synthesis with multimodal prompts," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2025, pp. 1–5.

[6] Mahdin Rohmatillah, Bryan Gautama Ngo, Willianto Sulaiman, Po-Chuan Chen, and Jen-Tzung Chien, "Reliable dialogue system for facilitating student-counselor communication," in *Proc. of Annual Conference of International Speech Communication Association*, 2024, pp. 1003–1004.

[7] Jen-Tzung Chien and Yi-Chien Wu, "Empathetic response generation via regularized Q-learning," in *Proc. of Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, 2024, pp. 1–6.

[8] Julian Zaïdi, Hugo Seuté, Benjamin van Niekerk, and Marc-André Carbonneau, "Daft-Exprt: Cross-speaker prosody transfer on any text for expressive speech synthesis," in *Proc. of Annual Conference of International Speech Communication Association*, 2022, pp. 4591–4595.

[9] Tao Li, Xinsheng Wang, Qicong Xie, Zhichao Wang, and Lei 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.

[10] Bryan Gautama Ngo, Mahdin Rohmatillah, and Jen-Tzung Chien, "Learning contrastive emotional nuances in speech synthesis," in *Proc. of Conference of the Oriental COCOSDA*, 2024, pp. 1–6.

[11] Po-Chaun Chen, Mahdin Rohmatillah, You-Teng Lin, and Jen-Tzung Chien, "Convcounsel: A conversational dataset for student counseling," in *Proc. of Conference of the Oriental COCOSDA*, 2024, pp. 1–6.

[12] Kaizhi Qian, Yang Zhang, Shiyu Chang, Mark Hasegawa-Johnson, and David Cox, "Unsupervised speech decomposition via triple information bottleneck," in *Proc. of International Conference on Machine Learning*, 2020, pp. 7836–7846.

[13] Xinfa Zhu, Yi Lei, Tao Li, Yongmao Zhang, Hongbin Zhou, Heng Lu, and Lei Xie, "METTS: Multilingual emotional text-to-speech by cross-speaker and cross-lingual emotion transfer," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1506–1518, 2024.

[14] Chae-Bin Im, Sang-Hoon Lee, Seung-Bin Kim, and Seong-Whan Lee, "EMOQ-TTS: Emotion intensity quantization for fine-grained controllable emotional text-to-speech," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 6317–6321.

[15] Yuke Li, Xinfa Zhu, Yi Lei, Hai Li, Junhui Liu, Danming Xie, and Lei Xie, "Zero-shot emotion transfer for cross-lingual speech synthesis," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, 2023, pp. 1–8.

[16] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, no. 6, pp. 066138, 2004.

[17] Youzhi Tu, Man-Wai Mak, and Jen-Tzung Chien, "Contrastive self-supervised speaker embedding with sequential disentanglement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 2704–2715, 2024.

[18] Yaroslav Ganin and Victor Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. of International Conference on Machine Learning*, 2015, vol. 37, pp. 1180–1189.

[19] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin, "CLUB: A contrastive log-ratio upper bound of mutual information," in *Proc. of International Conference on Machine Learning*, 13–18 Jul 2020, vol. 119.

[20] Longxin Li, Man-Wai Mak, and Jen-Tzung Chien, "Contrastive adversarial domain adaptation networks for speaker recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 5, pp. 2236–2245, 2022.

[21] Jen-Tzung Chien and Yuan-An Chen, "Towards a unified view of adversarial training: A contrastive perspective," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 5365–5369.

[22] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan, "Supervised contrastive learning," in *Advances in Neural Information Processing Systems*, 2020, vol. 33, pp. 18661–18673.

[23] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[24] Jen-Tzung Chien, I-Ping Yeh, and Man-Wai Mak, "Collaborative contrastive learning for hypothesis domain adaptation," in *Proc. of Annual Conference of International Speech Communication Association*, 2024, pp. 3225–3229.

[25] Navdeep Jaitly and Geoffrey E Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, 2013, vol. 117, p. 21.

[26] Chak Ho Chan, Kaizhi Qian, Yang Zhang, and Mark Hasegawa-Johnson, "Speechsplit2.0: Unsupervised speech disentanglement for voice conversion without tuning autoencoder bottlenecks," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 6332–6336.

[27] Erica Cooper, Cheng-I Lai, Yusuke Yasuda, and Junichi Yamagishi, "Can speaker augmentation improve multi-speaker end-to-end TTS?," in *Proc. of Annual Conference of International Speech Communication Association*, 2020, pp. 3979–3983.

[28] Boris N. Oreshkin, Pau Rodriguez, and Alexandre Lacoste, "TADAM: task dependent adaptive metric for improved few-shot learning," in *Advances in Neural Information Processing Systems*, 2018, p. 719–729.

[29] Keith Ito and Linda Johnson, "The LJ speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[30] Yao Shi, Hui Bu, Xin Xu, Shaojing Zhang, and Ming Li, "AISHELL-3: A multi-speaker mandarin TTS corpus and the baselines," in *Proc. of Annual Conference of International Speech Communication Association*, 2020, pp. 2756–2760.

[31] Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 920–924.

[32] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[33] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems*, 2020, vol. 33, pp. 17022–17033.

[34] Li-Jen Yang, Chao-Han Huck Yang, and Jen-Tzung Chien, "Parameter-efficient learning for text-to-speech accent adaptation," in *Proc. of Annual Conference of International Speech Communication Association*, 2023, pp. 4354–4358.

[35] Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 11, pp. 2579–2605, 2008.

[36] Hsin-Yi Lin, Huan-Hsin Tseng, and Jen-Tzung Chien, "On the attractive and repulsive forces of generalized stochastic neighbor embedding with alpha-divergence," *IEEE Access*, vol. 12, pp. 90380–90394, 2024.