Speaker-Independent Emotional Voice Conversion via Disentangled Representations

Xunquan Chen, Xuexin Xu, Takashi Kamihigashi, Jinhui Chen*,
Zhizhong Zhang, Tetsuya Takiguchi, and Edwin R. Hancock, *Fellow, IEEE*

***Abstract*—Emotional Voice Conversion (EVC) technology aims to transfer emotional state in speech while keeping the linguistic information and speaker identity unchanged. Prior studies on EVC have been limited to perform the conversion for a specific speaker or a predefined set of multiple speakers seen in the training stage. When encountering arbitrary speakers that may be unseen during training (outside the set of speakers used in training), existing EVC methods have limited conversion capabilities. However, converting the emotion of arbitrary speakers, even those unseen during the training procedure, in one model is much more challenging and much more attractive in real-world scenarios. To address this problem, in this study, we propose SIEVC, a novel speaker-independent emotional voice conversion framework for arbitrary speakers via disentangled representation learning. The proposed method employs the autoencoder framework to disentangle the emotion information and emotion-independent information of each input speech into separated representation spaces. To achieve better disentanglement, we incorporate mutual information minimization into the training process. In addition, adversarial training is applied to enhance the quality of the generated audio signals. Finally, speaker-independent EVC for arbitrary speakers could be achieved by only replacing the emotion representations of source speech with the target ones. The experimental results demonstrate that the proposed EVC model outperforms the baseline models in terms of objective and subjective evaluation for both seen and unseen speakers.**

***Index Terms*—Emotional voice conversion, disentangled representation learning, adversarial learning, mutual information, speaker-independent**

I. Introduction

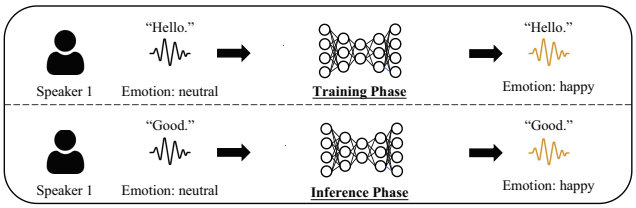
EMOTIONAL voice conversion (EVC) is a voice conversion technique where only emotional information in source speech is converted while retaining the linguistic information and speaker identity. Recently, EVC has attracted considerable attention in the field of speech processing. This technology can be applied in various domains such as voice assistants, conversational agents, and expressive text-to-speech (TTS) [1], [2], [3].

Previously developed methods for EVC can be roughly categorized into two types based on the use of training data. Early studies on EVC focused mainly on parallel training data. In other words, the mapping function is trained on paired utterances of the same linguistic content spoken in different emotional states. Among these approaches, the Gaussian mixture model (GMM) has been commonly used, and many improvements have been proposed for GMM-based EVC [4]. Other EVC methods, such as those based on non-negative matrix factorization (NMF) [5] or deep belief networks (DBNs) [6], have also been proposed. Although these methods have demonstrated their effectiveness, they require accurately aligned parallel data. Collecting parallel data and aligning the source and target utterances can be costly and time-consuming. These limitations have motivated research to explore non-parallel EVC approaches.

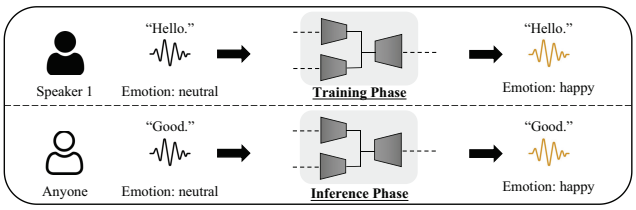
Recent works in non-parallel EVC using deep neural network models can be roughly divided into two categories: GAN-based models [7], [8], [9] and disentanglement-based models [10], [11]. To eliminate the need for parallel training data, GAN-based models such as CycleGAN-EVC [7] and StarGAN-EVC [8] have employed cycle consistency to ensure that the resultant invertible mapping is identical to the source input. Moreover, adversarial loss derived from the discriminator encourages the generator to generate speech that sounds similar to the target emotion. Because there is no guarantee that GAN-based models will learn explicit latent features, these direct transformation approaches cannot control the emotion state explicitly. Recent studies have shown that disentangled representation learning achieves remarkable performance in style transfer tasks. Disentanglement-based models usually design an autoencoder framework to decompose the speech into emotion and content representations with proper constraints. These methods can easily achieve EVC by simply replacing the emotion representation. Gao *et al.* [10] used an autoencoder framework to disentangle the emotional style and linguistic content from speech; thus, the emotional style could be modified independently without changing the linguistic content. Zhou *et al.* proposed a two-stage training strategy for EVC and used the corresponding phoneme transcription to guide the disentanglement of emotional styles and linguistic content. However, because there are no explicit constraints between different speech representations, these disentanglement-based models generally suffer from the untangle-overlapping problem. Hence, they tend to exhibit poor audio quality and transfer performance.

Although these GAN-based model and disentanglement-model methods can achieve subjectively satisfactory performance without the need for a parallel corpus, as shown in

*Corresponding author: Jinhui Chen (E-mail: chen@rieb.kobe-u.ac.jp)
X. Chen is with the Graduate School of System Informatics, Kobe University, Kobe, Japan.
X. Xu and Z. Zhang are with the Xiamen University, Xiamen, China.
T. Kamihigashi is with the Center for Computational Social Science, Kobe University, Kobe, Japan.
J. Chen is with the Prefectural University of Hiroshima, Hiroshima, Japan.
T. Takiguchi is with the Graduate School of System Informatics, Kobe University, Kobe, Japan.
E. R. Hancock is with the Department of Computer Science, The University of York, York, UK



(a) Speaker-dependent EVC



(b) Speaker-independent EVC

Fig. 1. Comparison between traditional speaker-dependent EVC methods and the proposed speaker-dependent EVC method. (a) Speaker-dependent EVC performing conversion for the specific speaker seen in training stage. (b) Speaker-independent EVC performing conversion for any speakers, including those unseen in training stage.

Fig. 1(a), they are limited to perform the conversion for a specific speaker seen in the training stage. Here, EVC frameworks, which are designed for a specific speaker, can be called speaker-dependent EVC. When encountering arbitrary speakers that may be unseen during training (outside the set of speakers used in training), existing EVC methods have limited conversion capabilities. This significant problem limits the real-world application of such models.

To overcome this limitation, as shown in Fig. 1(b), we propose a novel speaker-independent EVC (SIEVC) method for any speakers, even those unseen in training stage. This method can be easily generalized to unseen speakers without retraining or fine-tuning to improve the generalization ability, which could significantly reduce computations and the users' waiting time in practical use scenarios. To perform the conversion for arbitrary speakers, the proposed SIEVC model should only transfer speaker-independent emotion information while preserving emotion-independent elements. Although each speaker may express their emotions in a different manner, we can easily recognize their emotional states through speech, even in different languages. It has been demonstrated that there is a common code among different speakers for each pair of emotion classes in the domain of speaker-independent speech emotion recognition [12], [13], [14], [15]. Therefore, it is possible to extract a speaker-independent emotion representation from different speakers.

Inspired by disentangled representation learning in image-style transfer [16], [17], [18] and speech-style transfer [19], [20], [21], the proposed SIEVC method employs an autoencoder framework with an adversarial training strategy to disentangle the emotion information from the emotion-independent information of each input speech into separated representation spaces. There are two training stages during the entire training procedure:1) Stage I: autoencoder training and 2) Stage II: adversarial training. During the autoencoder training stage, we apply an emotion encoder to capture emotion representations and an emotion-independent encoder to encode emotion-independent representations. A decoder uses emotion and emotion-independent representations as inputs to reconstruct the input speech. Here, reconstruction loss is used to ensure that the autoencoder architecture does not lose too much information. Recent studies [22], [23], [24], [25], [26], [27] have demonstrated that mutual information minimization is an effective method for extracting disentangled representations in various style transfer tasks. To achieve better disentanglement of the emotion representations and emotion-independent representations of the input speech, we incorporate mutual information minimization into the autoencoder training process. To the best of our knowledge, this is the first study to apply mutual information to speaker-independent EVC. We further train the autoencoder based on an adversarial training strategy to enhance the quality of generated audio signals. During the adversarial training procedure, we utilize the autoencoder framework (emotion encoder, emotion-independent encoder, and decoder) as the GAN generator, which aims to fool the discriminator by generating high-quality and realistic audio signals. After training, the proposed model successfully separates emotion and emotion-independent information into independent representation spaces. Therefore, to convert the emotion for any speaker during run-time inference, we can simply replace the emotion representation of the source speech with that of the target emotion.

The main contributions of this study are summarized as follows.

- We propose SIEVC, a novel speaker-independent emotional voice conversion framework for arbitrary speakers via disentangled representation learning. The proposed method can be easily generalized to unseen speakers without any retraining or finetuning, which is much more challenging while being much more attractive in real-world scenarios.
- A novel optimization objective based on mutual information is proposed to achieve better disentanglement. Only reconstruction loss cannot guarantee that emotion representation and emotion-independent representation are disentangled without inter-dependency between them, leading to EVC performance degradation. To alleviate this issue, we incorporate mutual information theory into the autoencoder training stage.
- To achieve high-quality converted speech with the target emotion, the adversarial training strategy is also adopted in the training procedure. Through various subjective and objective evaluations, we demonstrate that the proposed method has better disentanglement and transfer performance than those of the baselines in both seen and unseen speaker scenarios on a real-world dataset.

The remainder of this paper is organized as follows. In Section II, we present related works in detail and discuss the relationship between our proposed model and alternative methods. Section III presents the proposed method. Section IV provides details of the experimental evaluations. Finally, conclusions are presented in Section V.

II. Related Works

Based on GAN or disentangled representation learning, several related works have proposed various speaker-dependent EVC models. In contrast, this study presents a novel speaker-independent EVC based on GAN combined with disentangled representation learning to achieve emotion conversion for an arbitrary speaker. In this section, we discuss the detailed differences between the proposed approach and previous studies and summarize the existing gaps in the current literature to determine our novel contributions. Furthermore, a study of mutual information is presented because it will be used for better disentanglement in this study.

A. Generative Adversarial Networks

In recent years, researchers have started to explore the feasibility of emotional voice conversion using non-parallel training data due to the difficulties in collecting accurately-aligned parallel data. An appealing solution to this problem is based on generative adversarial networks (GANs) [28], most of which have been inspired by recent advances in unpaired image style transfer [29], [30]. These GAN-based image-style transfer approaches generally utilize an adversarial training procedure to learn a mapping function that can map from the source domain to the target domain, which also applies to an EVC task.

The first studies were based on CycleGAN [29]. Zhou *et al.* [7] proposed CycleGAN-EVC to model the spectrum and prosody mapping between source speech and target speech. This has been widely acknowledged as an effective way to achieve one-to-one conversion with non-parallel data. However, using only one model to achieve many-to-many conversions is more attractive for a wide range of applications. As an improvement to CycleGAN, StarGAN [30] allows many-to-many domain mapping using a single model. Inspired by StarGAN [30], Rizos *et al.* proposed StarGAN-EVC [8] to train the spectral mapping between multiple emotional domains. Another study adopted an improved CycleGAN in the variational autoencoder (VAE)-GAN framework for EVC [9], in which a supervised strategy was used to extract more reliable emotion-related representations.

The effectiveness of GANs is because an adversarial training scheme forces the generated data to be indistinguishable from real data. Thus, the GAN has been adopted as the basic framework of our proposed method. However, the lack of explicit latent modeling in a GAN discourages the disentanglement between emotion information and emotion-independent information, thereby reducing the effectiveness of emotion representation in controlling the emotion state.

B. Disentangled Representation Learning

Recently, to address the limitations mentioned above, several studies [10], [11], [31], [32] based on speech representation disentanglement have attempted to decompose the speech into different representations. These methods can easily achieve emotional voice conversion by simply replacing the emotion-related representations.

Disentangled representation learning aims to encode input data into mutually independent latent subspaces with respect to different representations. [33], [34], [35]. Therefore, it is beneficial to obtain representations that contain certain attributes or that extract discriminative features. Drawing inspiration from recent studies on image-style transfers, adversarial and reconstruction-based autoencoder training have been widely used to obtain disentangled representations. Gao *et al.* [10] proposed a non-parallel EVC approach based on style transfer autoencoders, which consists of an encoder and decoder for each emotion domain. To use a limited amount of emotional speech data for unseen speakers, Zhou *et al.* [11] proposed a two-stage training strategy and used the corresponding phoneme transcription to guide the disentanglement of the emotional style and linguistic content. Choi *et al.* [31] used an emotion encoder and an additional speaker encoder to utilize various emotional characteristics of multiple speakers. Thus, conversion can be performed for a predefined set of multiple speakers, as seen in the training stage. However, these models are limited in their ability to convert for a specific speaker or a predefined set of multiple speakers. When encountering arbitrary speakers that may be unseen during training (outside the set of speakers used in training), existing EVC methods have limited conversion capabilities.

Thus far, there are few researchers that have explored speaker-independent emotional voice conversion techniques. To the best of our knowledge, only one model based on the VAW-GAN [36] learns to disentangle emotional elements and recompose speech by assigning a new emotional state. However, this model has no other mechanism to sufficiently decouple emotion-independent information from emotion information. This can be problematic because it can cause the model to produce low-quality audio. In this study, for better disentanglement between emotion information and emotion-independent information, mutual information is adopted to better separate the emotion information and emotion-independent information of voices into independent representation spaces.

C. Mutual Information

Mutual information (MI) is a key concept in information theory for measuring how similar two types of information are to each other. Recent studies have focused on MI estimation as a means of performing disentangled representation learning. Learning disentangled speech representation via MI minimization in speech tasks has also recently attracted increasing attention, such as text-to-speech conversion [24], [25] and speaker identity conversion [27], [26]. Hu *et al.* proposed a controllable TTS that can prevent content leakage by minimizing the mutual information between the style and content features. The mutual information between the style and content vectors was estimated using the mutual information neural estimator proposed in [37]. To achieve cross-lingual text-to-speech, Xin *et al.* [25] used a speaker encoder and a language encoder to extract speaker and language representations from acoustic features. For better disentanglement, mutual information minimization was then used to remove the entangled information within each representation.

TABLE I
ALL THE IMPORTANT NOTATIONS USED THROUGHOUT THIS PAPER.

Notation	Description
$\mathcal{X}$	an emotional dataset
X_s	source speech randomly sampled from $\mathcal{X}$
X_t	target speech randomly sampled from $\mathcal{X}$
$X_{s\to t}$	converted speech
A_s	mel-spectrogram feature of X_s
A_t	mel-spectrogram feature of X_t
$\hat{A}_{s\to s}$	reconstructed mel-spectrogram feature of X_s
$\hat{A}_{s\to t}$	converted mel-spectrogram feature of $X_{s\to t}$
Z_e^s	emotion representation of A_s
Z_c^s	emotion-independent representation of A_s
Z_e^t	emotion representation of A_t
Z_c^t	emotion-independent representation of A_t
MI	mutual information
L_{MI}	mutual information loss
L_{REC}	reconstruction loss
L_{ADV}	adversarial loss
E_e	emotion encoder
E_c	emotion-independent encoder
D_e	decoder
D_{is}	discriminator

For speaker identity conversion, Yuan *et al.* [26] proposed a disentanglement-based model that separates the style and content of voices into independent representation spaces by minimizing the mutual information between style and content representation. Wang *et al.* [27] applied a vector quantization technique for content encoding and used mutual information to achieve the proper disentanglement of content, speaker, and pitch representations. Therefore, MI estimation has been widely acknowledged as an effective way to measure dependencies between different representations for better disentanglement. Inspired by this, mutual information is applied to constrain the dependency between different representations in the proposed SIEVC framework.

Existing GAN-based or disentanglement-based EVC methods are limited in their ability to perform conversion for a specific speaker or a predefined set of multi-speakers seen in training stage. Moreover, because there are no explicit constraints between different speech representations, these disentanglement-based models generally suffer from an untangle-overlapping problem. Hence, they tend to exhibit poor audio quality and transfer performance. To alleviate these limitations, in this study, we propose a novel speaker-independent EVC based on GAN combined with disentangled representation learning to achieve conversion for an arbitrary speaker. Moreover, a novel loss function based on mutual information has been incorporated into the training process to achieve better disentanglement between different representations.

III. SPEAKER-INDEPENDENT EMOTIONAL VOICE CONVERSION

A. Notations and Preliminaries

We have defined some notations, random variables, and terminologies used throughout this paper in this section. The notations used in this study are summarized in Table I.

In EVC tasks, the input to a model consists of utterances drawn from different emotional categories. Given an emotional speech dataset $\mathcal{X}$, let us consider $X_s \in \mathcal{X}$ and $X_t \in \mathcal{X}$ as two randomly sampled speech instances from two different emotional categories. During the training procedure, A_s denotes the acoustic feature (mel-spectrogram) extracted from the source speech X_s, while A_t denotes the mel-spectrogram feature extracted from the target speech X_t.

The latent-space representation of data point A is denoted as $A = [Z_e, Z_c]$, where Z_e and Z_s are the latent emotion and emotion-dependent representation of the data point A, respectively. Our goal is to learn a generative module such that the latent representation A_s can be factorized into $[Z_e^s, Z_c^s]$. Given such a disentangled representation, we can perform EVC for any speaker by simply replacing the latent emotion representation Z_e^t with the latent emotion representation Z_e^t. For better disentanglement, in this study, mutual information has been adopted to decompose speech into independent representation spaces.

***Definition** 3.1 (Mutual Information):* Mutual information is a measure of the dependence between two random variables. For two discrete random variables X and Y, the mutual information between them is given as follows:

$$MI(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}. \tag{1}$$

Here, $p(x,y)$ is the joint probability distribution function of the random variables and $p(x)$ and $p(y)$ represent the marginal probability distribution functions of X and Y, respectively. $MI(X;Y)$ represents the amount of information shared by X and Y.

As shown in Fig. 2, our proposed SIEVC framework is based on GAN combined with disentangled representation learning. Typically, a GAN is composed of a generator and a discriminator. The generator utilized in our study is an encoder-decoder architecture consisting of three modules: an emotion encoder $E_e(\cdot)$, an emotion-independent encoder $E_c(\cdot)$, and a decoder $D_e(\cdot,\cdot)$. Inside the generator framework, the emotion encoder E_e accepts the acoustic features (mel-spectrogram) as input and extracts the emotion representation related to the emotional state of the input speech. The emotion-independent encoder E_c captures emotion-independent representations. Then, decoder D_e takes the disentangled representations as input to synthesize the converted mel-spectrogram. The discriminator distinguishes real speech from converted speech while encouraging the generator to synthesize more realistic speech. Detailed information regarding each component of the proposed framework is provided in section III-C.

If all modules are well trained, we can achieve speaker-independent emotional voice conversion via disentangled representations. Let us consider source speech X_s and target speech X_2 drawn from two different emotional categories. The emotion encoder E_e extracts the emotion representation Z_e^t from the target speech, and the emotion-independent encoder E_c extracts the emotion-independent representation Z_c^s from the source speech. The decoder takes the two extracted representations as inputs to generate the converted mel-spectrogram $\hat{A}_{s\to t}$ by changing only the source emotion to the target emotion. Finally, the converted speech waveform $\hat{X}_{s\to t}$ is

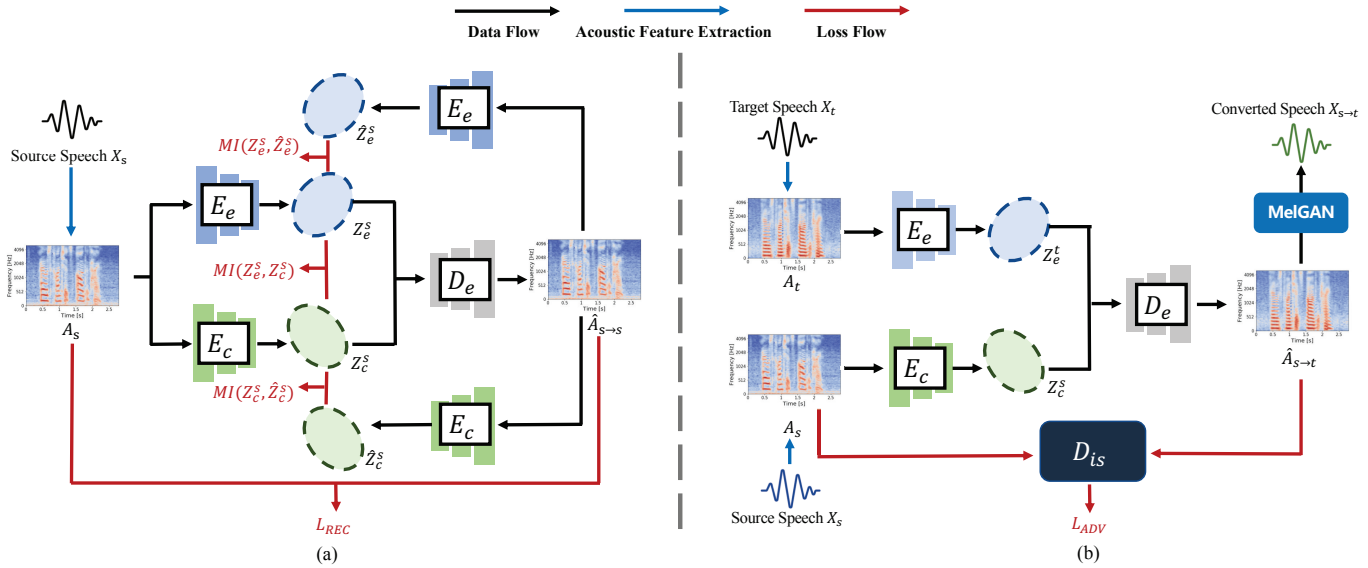


Fig. 2. The training procedure of the proposed SIEVC. The emotion encoder E_e and emotion-independent encoder E_z are employed to emotion and emotion-independent representations, respectively. The decoder D_e is used to generate the converted mel-spectrogram. The discriminator D_{is} is designed to judge whether the input mel-spectrogram comes from a real sample. Three objectives are enforced in the generative module: reconstruction loss L_{REC}, mutual information loss L_{MI}, and adversarial loss L_{ADV}. Here, mutual information loss L_{MI} is a weighted sum of $MI(Z_e^s; Z_c^s)$, $MI(Z_e^s; \hat{Z}_e^s)$, and $MI(Z_c^s; \hat{Z}_c^s)$.

reconstructed from the converted mel-spectrogram using a vocoder.

B. Training Procedure

In this section, we detail the training procedure of the proposed SIEVC framework. As shown in Fig. 2, there are two training stages during the entire training procedure: 1) Stage I: autoencoder training, and 2) Stage II: adversarial training. Three losses are introduced to train our proposed model. These are the reconstruction loss L_{REC} and mutual information loss L_{MI} for autoencoder training, and the adversarial loss L_{ADV} for adversarial training.

1) Stage I: Autoencoder traning: As depicted in Fig. 2(a), self-reconstruction from an input speech is performed to preserve the consistency during autoencoder training stage. The self-reconstruction procedure can be formulated as follows:

$$Z_e^s = E_e\left(A_s\right), Z_c^s = E_c\left(A_s\right), \hat{A}_{s\to s} = D_e\left(Z_e^s, Z_c^s\right), \quad (2)$$

where A_s is the mel-spectrogram feature extracted from input speech X_s and $\hat{A}_{s\to s}$ is the reconstructed mel-spectrogram. Z_e^s is the emotion representation captured from the emotion encoder E_e, Z_c^s denotes emotion-independent representation generated by the emotion-independent encoder E_c. The decoder D_e learns to reconstruct A_s as $\hat{A}_{s\to s}$ using concatenated latent representations.

Reconstruction loss: A reconstruction loss L_{REC} is calculated between the reconstructed mel-spectrogram and ground truth. The reconstruction loss L_{REC} is defined by measuring the differences between the input and correspond reconstructed mel-spectrogram using the $\mathcal{L}_1$ distance (norm), as follows:

$$L_{REC} = \mathbb{E}_{\mathcal{X}} \left\| \hat{A}_{s\to s} - A_s \right\|_1, \quad (3)$$

where A_s is the mel-spectrogram of the input speech signal, and $\hat{A}_{s\to s}$ is the reconstructed mel-spectrogram generated from the self-reconstruction procedure in Eq. 2. This reconstruction loss encourages well defined output mel-spectrogram and ensures that the autoencoder architecture does not lose too much information. It is also an essential part and a main objective for disentanglement-based emotional voice conversion methods [10], [31], [32].

However, using only reconstruction loss cannot guarantee that the emotion and emotion-independent representations are disentangled without residual mutual information between them. To achieve better disentanglement, we incorporate mutual information minimization into the autoencoder training process as illustrated in Fig. 2(a). Next, we explain in detail how mutual information is estimated.

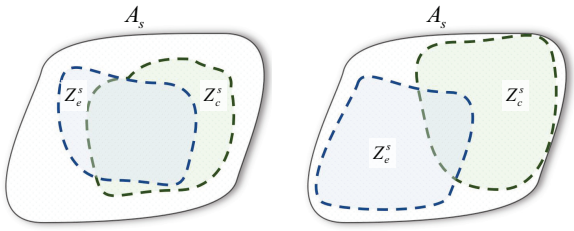
By Definition 3.1, mutual information can be equivalently expressed with the joint entropy $H(X,Y)$ and marginal entropy $H(X), H(Y)$ of two variables X and Y as follows:

$$\begin{aligned} MI(X;Y) &= H(X) - H(X \mid Y) \\ &= H(X) + H(Y) - H(X,Y), \end{aligned} \quad (4)$$

where

$$\begin{aligned} H(X) &= -\sum_x p(x) \log p(x), \\ H(Y) &= -\sum_y p(y) \log p(y), \\ H(X,Y) &= -\sum_{x,y} p(x,y) \log p(x,y). \end{aligned} \quad (5)$$

Therefore, mutual information is notoriously difficult to compute exactly because the closed form of the joint distribution is generally unknown. To solve this problem, some estimators based on deep neural networks have been proposed [37], [38], [39].



(a) Without MI loss (b) With MI loss

Fig. 3. Illustration of the disentanglement between Emotion representation Z_e^s and emotion-independent representation Z_c^s without MI loss (a) and with MI loss (b).

The proposed approach encodes the mel-spectrogram feature A_s of input speech X_s into an emotion representation $Z_e^s = E_e(A_s)$ and an emotion-independent representation $Z_c^s = E_c(A_s)$ using an emotion encoder $E_e(\cdot)$ and an emotion-independent encoder $E_c(\cdot)$, respectively. As depicted in Fig. 3, the motivation of our proposed method is to disentangle the emotion and emotion-independent latent representation spaces (ideally, there is no residual mutual information between them). Using only the reconstruction loss does not guarantee that Z_e^s and Z_c^s are separable. From the definition of mutual information, we have the following theorem:

***Theorem** 3.1:* Z_e^s and Z_c^s are fully separable or statistically independent if and only if $MI(Z_e^s; Z_c^s) = 0$.

Proof: If Z_e^s is independent of Z_c^s, then $p(Z_e^s, Z_c^s) = p(Z_e^s)p(Z_c^s)$; According to Equation 1, $MI(Z_e^s; Z_c^s)$ will be zero. ∎

Therefore, mutual information between Z_e^s and Z_c^s is expected to be minimised. By minimizing such mutual information, we learn the representations which are independent from each other.

To further faithfully represent the input data A_s, a representation of the reconstructed mel-spectrogram $\hat{A}_{s\to s}$ should be informative of A_s, that is, $MI(Z_e^s; \hat{Z}_e^s)$ and $MI(Z_c^s; \hat{Z}_c^s)$ should be large. In other words, it is desirable to maximize the mutual information between the representation pairs $(Z_e^s; \hat{Z}_e^s)$ and $(Z_c^s; \hat{Z}_c^s)$.

Consequently, the proposed objective function based on mutual information calculates the following terms:

- $MI(Z_e^s; Z_c^s)$: encourages the emotion representation Z_e^s and emotion-independent representation Z_c^s of source mel-spectrogram A_s to completely remove any redundant information present in both representations, making them independent of each other. In other words, the emotion representation Z_e^s does not contain information about the emotion-independent representation Z_c^s, and vice versa.
- $MI(Z_e^s; \hat{Z}_e^s)$ and $MI(Z_c^s; \hat{Z}_c^s)$: constrain the latent representations to be same for information preservation, before and after the self-reconstruction procedure.

Intuitively, we must minimise $MI(Z_e^s; Z_c^s)$ while maximising $MI(Z_e^s; \hat{Z}_e^s)$ and $MI(Z_c^s; \hat{Z}_c^s)$.

Mutual information loss: As a result, the overall mutual information loss L_{MI} can be minimized as follows,

$$L_{MI} = MI(Z_e^s; Z_c^s) - \beta[MI(Z_e^s; \hat{Z}_e^s) + MI(Z_c^s; \hat{Z}_c^s)], \quad (6)$$

where β is the hyperparameter controlling the relative importance of these items. In order to achieve more effective disentanglement between emotion representation Z_e^s and emotion-independent representation Z_c^s, we set β to 0.5 during training. Next, we explain how to maximise the $MI(Z_e^s; \hat{Z}_e^s)$ and $MI(Z_c^s; \hat{Z}_c^s)$, and how to minimise $MI(Z_e^s; Z_c^s)$, respectively.

To maximize $MI(Z_e^s; \hat{Z}_e^s)$ and $MI(Z_c^s; \hat{Z}_c^s)$, we estimate the lower bound of the mutual information by adopting mutual information neural estimation (MINE) [37]. Based on the Donsker-Varadhan representation [40] of the Kullback-Leibler divergence, the lower bound of mutual information can be represented by:

$$\begin{aligned} MI(X;Y) \geq \sup_{\theta\in\Theta} E_{p_{(X,Y)}(x,y)}\left[T_\theta(x,y)\right] - \\ \log\left(E_{p_X(x)p_Y(y)}\left[e^{T_\theta(x,y)}\right]\right), \end{aligned} \quad (7)$$

where $T_\theta(x, y)$ represents a function parameterised by a neural network with the parameter θ. As shown in Fig. 4(b), MINE is used to estimate the mutual information between the disentangled representations extracted from the ground-truth mel-spectrogram A_s and disentangled representation extracted from the reconstructed mel-spectrogram $\hat{A}_{s\to s}$. Each batch of training data contains B pairs of $(Z_e^s, \hat{Z}_e^s)$ and $(Z_c^s, \hat{Z}_c^s)$, where the i^{th} pair of data points is denoted by $(Z_{e,i}^s, \hat{Z}_{e,i}^s)$ and $(Z_{c,i}^s, \hat{Z}_{c,i}^s)$. The batch size is set to 32 in this paper. Consequently, the criterion for maximizing $MI(Z_e^s; \hat{Z}_e^s)$ and $MI(Z_c^s; \hat{Z}_c^s)$ can be defined as:

$$\begin{aligned} MI(Z_e^s; \hat{Z}_e^s) = \frac{1}{B}\sum_{i=1}^{B} T_\theta\left(Z_{e,i}^s, \hat{Z}_{e,i}^s\right) - \\ \log\left(\frac{1}{B}\sum_{i=1}^{B} e^{T_\theta\left(Z_{e,i}^s, \bar{\hat{Z}}_{e,i}^s\right)}\right); \end{aligned} \quad (8)$$

$$\begin{aligned} MI(Z_c^s; \hat{Z}_c^s) = \frac{1}{B}\sum_{i=1}^{B} T_\theta\left(Z_{c,i}^s, \hat{Z}_{c,i}^s\right) - \\ \log\left(\frac{1}{B}\sum_{i=1}^{B} e^{T_\theta\left(Z_{c,i}^s, \bar{\hat{Z}}_{c,i}^s\right)}\right). \end{aligned} \quad (9)$$

Minimising $MI(Z_e^s; Z_c^s)$ can be difficult if we use MINE mutual information estimator here. Because this neural estimation method estimates the lower-bound mutual information, which is inconsistent to mutual information minimization of $MI(Z_e^s; Z_c^s)$. Therefore, to effectively minimize $MI(Z_e^s; Z_c^s)$, the upper-bound mutual information should be estimated. Different from the mutual information maximization of $MI(Z_e^s; \hat{Z}_e^s)$ and $MI(Z_c^s; \hat{Z}_c^s)$, in this study, we have adopted the contrastive log-ratio upper bound (CLUB) [39] to minimize $MI(Z_e^s; Z_c^s)$ by computing its upper bound.

Given the emotion representation Z_e^s and the emotion-independent representation Z_c^s of input speech A_s, CLUB uses a probabilistic neural network to approximate the conditional distribution $q_\theta(Z_e^s \mid Z_c^s)$. In this study, a neural network parameterized by θ is used to learn the distribution. The MI minimization estimation between Z_e^s and Z_c^s through upper

bound estimator CLUB is given by

$$\begin{aligned} &MI(Z_e^s; Z_c^s) \\ &= \frac{1}{B}\sum_{i=1}^{B}\left[\log q_\theta\left(Z_{e,i}^s \mid Z_{c,i}^s\right) - \log q_\theta\left(Z_{e,i}^s \mid \bar{Z}_{c,i}^s\right)\right], \end{aligned} \tag{10}$$

where q_θ is the neural approximation and $(Z_{e,i}^s, Z_{c,i}^s)$ denotes the i^{th} pair of data points. And the data samples $\{(Z_{e,i}^s, \bar{Z}_{c,i}^s)\}_{i=1}^B$ are obtained by shuffling emotion-independent representations in a mini-batch. Here, $\bar{Z}_{c,i}^s$ denotes the ith emotion-independent representation of the mini-batch after shuffling. The batch size is set to 32 and each mini-batch consists of 32 pairs.

As shown in Fig. 4(b), using the aforementioned mutual information estimators MINE and CLUB, the proposed MI loss $L_{MI} = MI(Z_e^s; Z_c^s) - \beta[MI(Z_e^s; \hat{Z}_e^s) + MI(Z_c^s; \hat{Z}_c^s)]$ can be minimized after estimating $MI(Z_e^s; Z_c^s)$, $MI(Z_e^s; \hat{Z}_e^s)$, and $MI(Z_c^s; \hat{Z}_c^s)$,.

2) Stage II:Adversarial Training: In the second step of our method, GAN training is implemented to eliminate the training-testing mismatch and obtain a high-quality EVC. As shown in Fig. 2(b), we extract the emotion representation Z_e^t of the target emotion from the target mel-spectrogram feature A_t, then extract the emotion-independent representation Z_c^s from the source mel-spectrogram feature A_s. The decoder reconstructs $\hat{A}_{s\to t}$ using Z_e^t and Z_c^s as $\hat{A}_{s\to t} = D_e\left(Z_e^t, Z_c^s\right)$. The conversion phase is formulated as follows:

$$\boldsymbol{Z}_e^t = E_e\left(A_t\right), \boldsymbol{Z}_c^s = E_c\left(A_s\right), \hat{A}_{s\to t} = D_e\left(\boldsymbol{Z}_e^t, \boldsymbol{Z}_c^s\right). \tag{11}$$

The generator is expected to synthesize a sufficiently realistic speech mel-spectrogram $\hat{A}_{s\to t}$ to fool the discriminator. In contrast, the discriminator is trained to distinguish the generated mel-spectrogram $\hat{A}_{s\to t}$ from the ground-truth spectrogram A_t of the target speech X_t. This encourages the generator to generate realistic speech. This adversarial training stage for realistic converted mel-spectrogram can be trained with adversarial loss.

Adversarial loss: The adversarial loss can be expressed as:

$$L_{ADV} = \mathbb{E}_{\mathcal{X}}\left[\log D\left(A_t\right) + \log\left(1 - D\left(\hat{A}_{s\to t}\right)\right)\right] \tag{12}$$

The smaller the loss, the closer the converted data distribution is to a normal speech distribution. In this study, Wasserstein GAN with gradient penalty (WGAN-GP) [41] is adopted instead of the original GAN [28] to mitigate the training instability issue.

Final Objective Function: With the individual loss functions described above, shown in Fig. 4, the final objective function for training the proposed SIEVC model can be represented as:

$$L_{TOTAL} = L_{REC} + \lambda_1 L_{MI} + \lambda_2 L_{ADV}, \tag{13}$$

where λ_1 and λ_2 are hyperparameters that control the relative importance of each other. We performed hyperparameters tuning during the albation study in SectionIV-C. We note that the reconstruction loss L_{REC} is necessary for the production of speech. When the model was trained without it, the model could not generate fluent speech. The effect of each loss term is shown in Table V. Thus we set λ_1 to 0.2 and λ_1 to 0.2 during whole training procedure.

Algorithm 1 Whole training procedure of the proposed SIEVC

Required:
1. Emotional speech dataset $\mathcal{X}$;
2. Generator $G = \{E_e(\cdot), E_c(\cdot), D_e(\cdot,\cdot)\}$ and discriminator D_{is} with parameters $\theta_G = \{\theta_{E_e}, \theta_{E_c}, \theta_{D_e}\}$, and $\theta_{D_{is}}$, respectively;
3. Batch size $B = 32$.

Initialization:
Initialize the parameters θ_G and $\theta_{D_{is}}$.

Iteration:

for $(i = 1; i < n+1; i = i+1)$ **do**
  for $(j = 1; j < B+1; j = j+1)$ **do**
    Sample source speech $X_s \in \mathcal{X}$;
    Sample target speech $X_t \in \mathcal{X}$.
  end for
  Create a B-size minibatch $\{X_s, X_t\}$.
  Extract mel-spectrogram feature A_s from X_s;
  Extract mel-spectrogram feature A_t from X_t.
  latent representations:
  $Z_e^s = E_e(A_s)$, $Z_c^s = E_c(A_s)$, $Z_e^t = E_e(A_t)$
  $\hat{A}_{s\to s} = D_e\left(Z_e^s, Z_c^s\right)$, $\hat{A}_{s\to t} = D_e\left(Z_e^t, Z_c^s\right)$
  Calculate L_{REC}, L_{MI}, and L_{ADV}
  Update the weights by descending the stochastic gradient:
  $\theta_{D_{is}} \overset{+}{\leftarrow} -\nabla_{\theta_{D_{is}}} \lambda_2 L_{ADV}$
  $\theta_G \overset{+}{\leftarrow} -\nabla_{\theta_G}(L_{REC} + \lambda_1 L_{MI} + \lambda_2 L_{ADV})$
end for
return Optimized weights

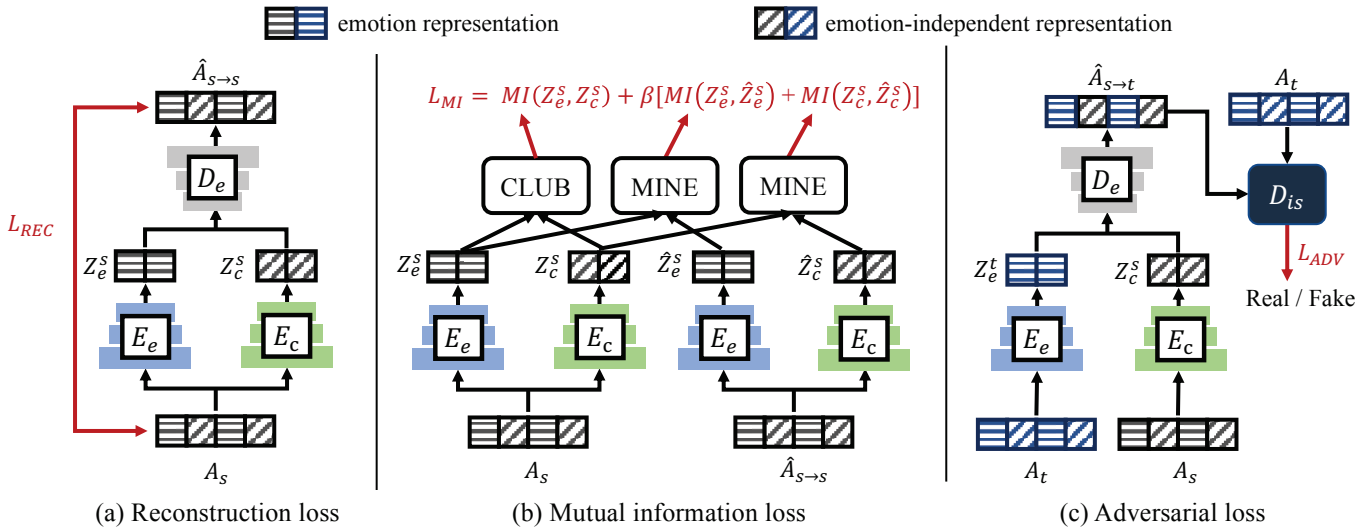
The proposed framework (*i.e.*, generator and discriminator) was trained using the ADAM optimiser [42] with a learning rate of 0.0001, $\beta_1 = 0.9$, and $\beta_1 = 0.999$. The weight decay was set to 0.0001 to prevent overfitting. The entire training procedure is presented in Algorithm 1.

C. Network Architecture

TABLE II
DETAILS OF THE MODEL CONFIGURATIONS.

Emotion Encoder
ConvBank → Conv1d-ReLU × 2 → Conv1d-ReLU-AvgPooling × 3 → Fully Connected layer → ReLU Layer
Emotion-independent Encoder
ConvBank → Conv1d × 3 → Downsample: Conv1d-IN-ReLU × 2 → Conv1d-IN-ReLU-Conv1d × 2
Decoder
Conv1d-AdaIN-ReLU × 3 → Upsample: Conv1d-PixelShuffe-AdaIN × 3 → Conv1d, with resuidual connection
Discriminator
Conv2d → Downsample: Conv2d-LReLU-IN × 5 → Fully Connected layer→Scalar output

"Conv1d" represents 1-d convolution layer and "Conv2d" represents 1-d convolution layer. "AvgPooling" represents average pooling. "IN" represents instance normalization. AdaIN represents adaptive instance normalization. "$\times N$ represents repeating the block for N times. "ReLU" and "LReLU" represent ReLU activation and leakyReLU activation, respectively.



(a) Reconstruction loss (b) Mutual information loss (c) Adversarial loss

Fig. 4. Overview of proposed training criteria. (a) Reconstruction loss L_{REC}; (b) Mutual information loss L_{MI}; (c) Adversarial loss L_{ADV}.

In this section, we provide the details of the proposed model configuration, which are summarized in Table II.

1) Generator: The proposed SIEVC framework is based on GAN combined with disentangled representation learning. Therefore, in this study the generator is an encoder-decoder architecture, which consists of three components: emotion encoder E_e, emotion-independent encoder E_c, and decoder D_e. We first obtain high-level representations from E_e and E_c, respectively, and reconstruct the speech information through D_e. The generator is composed entirely of 1d convolution neural network to capture the temporal dependencies.

Emotion Encoder: The emotion encoder is used to capture the emotion-related information from mel-spectrogram. We first employ the ConvBank layer which stacks convolution layers with different kernel sizes to enlarge the receptive field and capture long-time scale information. Subsequently, several convolution layers are applied to generate the high-level representations. The purely 1-dimensional convolution layers are implemented with a kernel size set to 5, and the stride size depends on whether downsampling of the temporal scales is required. It is important to note that we do not downsample the temporal dimension in the emotion encoder. Instead, we keep the original temporal dimension the same as the input acoustic features to preserve the overall information. We also use average pooling to enforce the emotion encoder to learn global information only and decrease the temporal resolution to match the feature shapes.

Emotion-independent Encoder: The emotion encoder is used to capture the emotion-independent information from mel-spectrogram. To enlarge the receptive field and capture long-time-scale information, we also employed the ConvBank layer [43], which stacks convolution layers of different kernel sizes. In addition, we adopt instance normalisation (IN) [44] after each convolution layer of the emotion-independent encoder to eliminate emotional style information. Then, ReLU nonlinear activation is applied after each IN layer. To mitigate the training difficulties, we also implement residual connections [45] for each pair of convolution layers with the exception of the ConvBank layer.

Decoder: The decoder is used to recover the mel-spectrogram from a combination of emotion and emotion-independent representations. A set of convolution layers with kernel size 5 and stride 1 are implemented in the decoder. To increase the temporal resolution, the PixelShuffle1d layer [46] is used for upsampling and nearest interpolation for the residual connections to match the shape. Then we use adaptive instance normalisation (AdaIN) [47] layer to provide global style information before activation layer.

2) Discriminator: The discriminator is used to distinguish a real sample of speech from a synthetic one while encouraging the generator to synthesize realistic speech of the target domain. Therefore, the discriminator is constructed with 2d convolution layers in a manner similar to [48], [49] to better capture the acoustic texture. There are 5 convolution layers with stride 2 and kernel size 5×5 to gradually, downsample the feature map. The number of filters for these convolution layers are respectively 64, 128, 256, 512 and 512. To decrease the feature channel from 512 to 32, a convolution layer with unit kernel size and stride is appended. Finally, an output layer follows and is used to obtain a measure of the degree of verisimilitude of the speech in the target domain. Instance normalisation [44] and leaky ReLu activation [50] with a slope of 0.01 are applied after each convolution layer except the last output layer.

3) Vocoder: As depicted in Fig. 2, because the output of the proposed model is the mel-sectrogram of the converted speech, we employ a vocoder to convert the mel-spectrogram to time-domain waveform. Recently, various neural vocoders [51], [52] were successfully applied to EVC for waveform reconstruction. Considering the quality and the inference speed of waveform generation, a MelGAN vocoder [51] pre-trained on the VCTK corpus [53] is applied to generate a proper speech waveform from a given mel-spectrogram. Initially, we generate the corresponding acoustic features in the required

format for the MelGAN input. Hence, in this study, we utilised mel-spectrograms for acoustic features, which can effectively imply various types of information in speech, not only linguistic but also non-linguistic, such as the speaker identity and emotion information. More precisely, we resample the audio at 22,050 HZ and perform the STFT (short-time Fourier transform) with STFT window size 1024. We then transform the magnitude of the spectrograms into an 80-bin mel-scale and then take its logarithm. Subsequently, these acoustic features will be fed into our model to optimize its parameters. Finally, we generate the converted speech through the optimized model and the vocoder.

IV. Experiments and Discussions

A. Dataset and Experimental Comparative Methods

We conducted experimental evaluations on the emotional speech dataset (ESD) [32], which consisted of 350 parallel utterances with an average duration of 2.5s recorded by 10 native English speakers and 10 native Mandarin speakers with five different emotions. In this study, we only considered five emotional categories of the 10 English speakers: angry, happy, neutral, sad, and surprised. We constructed the following four datasets: neutral-to-happy voice (Neu2Hap), neutral-to-angry voice (Neu2Ang), neutral-to-surprise voice (Neu2Sur), and neutral-to-sad voice (Neu2Sad). We trained the proposed model and baselines using speech data from the first eight speakers, which are denoted as seen speakers. The remaining two speakers who were not involved in the training phase are denoted as unseen speakers. The training and testing sets were non-overlapping utterances randomly selected from 350 utterances (330 utterances for training and 20 utterances for testing).

To evaluate the proposed method performance in converting different emotional states, we compared the results to those of several state-of-the-art methods listed below.

- **CycleGAN-EVC [7]:** The CycleGAN [29] model has been widely used in non-parallel VC tasks. Zhou *et al.*[7] extended this unsupervised learning model to develop CycleGAN-EVC, which is an effective way to achieve one-to-one emotional voice conversion. The WORLD vocoder [54] is used for speech analysis and synthesis.
- **StarGAN-EVC [8]:** StarGAN-based emotional voice conversion has been proposed for many-to-many EVC tasks. StarGAN-EVC can perform many-to-many conversion using a single model. The WORLD vocoder is used for speech analysis and synthesis.
- **VAWGAN-EVC [36]:** This is a speaker-independent method based on VAWGAN [36], which consists of two encoder-decoder structures that separately learn the spectrum and CWT-based F0 mappings. To represent different emotions, a one-hot vector was provided to the generator as an emotion identity. The WORLD vocoder is used for speech analysis and synthesis.

To make fair comparison, we reproduced their performance using the available open source implementations and with the same training data. We note that CycleGAN-EVC can only perform one-to-one conversions; thus, we trained one CycleGAN-EVC for each emotion pair separately. The remaining methods used a unified model for all the emotion pairs.

B. Evaluation Methodology

1) Objective Evaluations: To objectively measure the quality of the generated speech, we used three different metrics, *i.e.*, 1) Mel-Cepstral Distortion; 2) Root Mean Square Error; 3) Speech Emotion Recognition Accuracy.

Mel-Cepstral Distortion (MCD): In the field of EVC, MCD [55] is commonly used for assessing the quality of generated speech signals in the mel-cepstral space. To compute the MCD between the generated and ground-truth target speech, we calculated the first 24 mel-cepstral coefficients (MCCs), which can be formulated as follows:

$$MCD = (10/\ln 10)\sqrt{2\sum_{i=1}^{24}\left(mc_i^t - mc_i^c\right)^2} \tag{14}$$

where mc_i^t and mc_i^c represent the target and generated mel-cepstral, respectively. To make reasonable comparisons between the generated and ground-truth speech, Dynamic time warping (DTW) based alignment is performed before calculating MCD. A lower MCD value indicates higher similarity between the generated and ground-truth target speech.

Root Mean Suqare Error (RMSE): To evaluate the conversion error between target and converted F0 features, we calculate the RMSE as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left((F0_i^t) - (F0_i^c)\right)^2}, \tag{15}$$

where $F0_i^t$ and $F0_i^c$ denote the target and converted F0 features, respectively. We used the WORLD vocoder [54] to extract fundamental frequencies from raw audio waveforms. DTW was also applied to calculate the RMSE between the two sequences. A lower F0-RMSE value indicates a smaller distortion or prediction error.

Tables III and IV report the MCD and F0-RMSE results obtained from the neutral-to-emotional pairs. For seen speakers, as reported in Table III, our proposed model outperforms other models in spectral and F0 conversion. Table IV reports the MCD and F0-RMSE results for unseen speakers. Both methods with speaker-dependent frameworks, CycleGAN-EVC and StarGAN-EVC, achieve unsatisfactory results. Compared to speaker-independent VAWGAN-EVC model, our model exhibits better transfer performance on MCD and F0-RMSE metrics. Through the objective experiments, we empirically confirm that the proposed method effectively brings the converted acoustic feature sequence closer to the target one than that of the baseline.

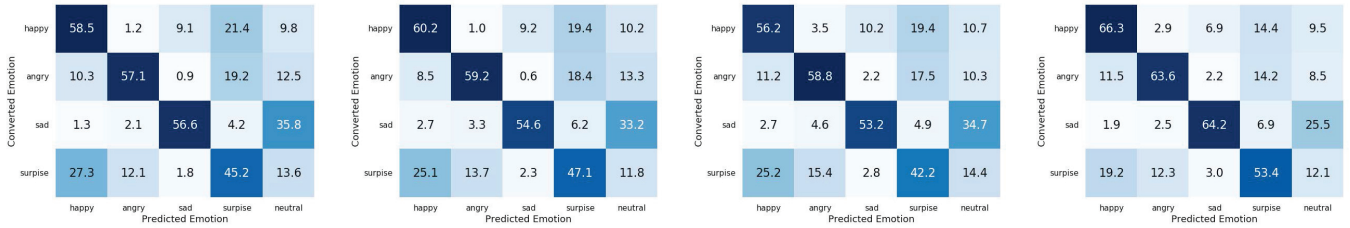
Speech Emotion Recognition Accuracy: Speech emotion recognition accuracy measures whether the converted speech belongs to the target emotion category. For a fair comparison, we used a third-party pre-trained speech emotion recognition model [56] to classify the emotion from the converted speech. The confusion matrices of the classification conducted on the

TABLE III
MCD AND F0-RMSE RESULTS FOR DIFFERENT EMOTIONS OF SEEN SPEAKERS. THE BEST PERFORMANCES ARE INDICATED IN BOLDFACE.

Seen or Unseen	Seen Speaker							
Metric	MCD [dB]				F0-RMSE [Hz]			
Conversion Pairs	Neu2Hap	Neu2Ang	Neu2Sad	Neu2Sur	Neu2Hap	Neu2Ang	Neu2Sad	Neu2Sur
Source	6.54	5.36	6.84	6.74	77.5	74.6	101.7	91.6
CycleGAN-EVC	4.54	4.81	4.57	4.60	57.2	60.8	69.5	62.7
StarGAN-EVC	4.25	**4.61**	4.67	4.58	56.5	62.1	67.9	62.2
VAWGAN-EVC	4.36	4.92	4.69	4.63	58.2	62.8	68.3	63.6
Proposed method	**3.97**	4.69	**4.23**	**4.26**	**46.3**	**49.6**	**59.4**	**56.7**

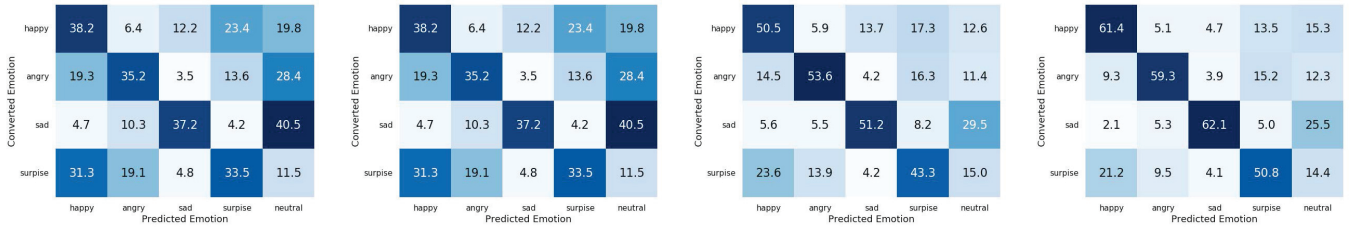
TABLE IV
MCD AND F0-RMSE RESULTS FOR DIFFERENT EMOTIONS OF UNSEEN SPEAKERS. THE BEST PERFORMANCES ARE INDICATED IN BOLDFACE.

Seen or Unseen	Unseen Speaker							
Metric	MCD [dB]				F0-RMSE [Hz]			
Conversion Pairs	Neu2Hap	Neu2Ang	Neu2Sad	Neu2Sur	Neu2Hap	Neu2Ang	Neu2Sad	Neu2Sur
Source	6.34	5.66	6.58	6.16	78.3	73.1	95.4	93.2
CycleGAN-EVC	5.89	5.79	5.77	5.83	76.1	69.8	87.5	90.6
StarGAN-EVC	5.53	5.96	5.61	5.49	75.1	70.3	79.5	82.85
VAWGAN-EVC	4.54	**4.45**	4.62	4.77	61.4	66.2	70.8	72.4
Proposed method	**4.02**	4.65	**4.52**	**4.36**	**50.4**	**51.3**	**62.7**	**61.3**



(a) CycleGAN converted voice (b) StarGAN converted voice (c) VAWGAN converted voice (d) Proposed method converted voice

Fig. 5. The confusion matrix of seen speaker.



(a) CycleGAN converted voice (b) StarGAN converted voice (c) VAWGAN converted voice (d) Proposed method converted voice

Fig. 6. The confusion matrix of unseen speaker.

seen and unseen speakers are presented in Figures 5 and 6, respectively.

In the experiment with seen speakers, we observed that our proposed model achieved the best result in all emotion states. We can see that the 66.3% recognition result of the happy emotion is better than those of the other models. Moreover, regarding the angry emotion, the proposed model achieved 63.6% recognition accuracy, which is a small improvement compared to that of the others. In the emotion state of the happy and surprise emotions, we obtained almost equal performance results compared to those of the other models. Furthermore, the happy emotion was easily confused with the surprise emotion in the classification of all models.

In the experiments with unseen speakers, our proposed model was also superiority in all emotion states compared to the speaker-dependent models. We also improved the accuracy rate in the four-class emotional recognition compared to VAWGAN-EVC, the speaker-independent EVC model.

2) Subjective Evaluations: For the subjective experiment, we also conducted evaluations on the naturalness of generated speech, and the similarity of the converted speech to the target speech. Two different subjective metrics, ”similarity” and ”naturalness”, were used as evaluation metrics. For both ”similarity” and ”naturalness”, a MOS test was conducted to evaluate the performance. The scale ranged from 1 (bad) to 5 (excellent). We carried out the MOS test for all conversion pairs, including neutral-to-angry, neutral-to-sad, neutral-to-happy, and neutral-to-surprise. To evaluate similarity, listeners were asked to listen to the speech pairs and score them from 1 to 5 according to whether they had the same emotion. For measuring naturalness, the listeners were asked to score the generated samples from 1 to 5 according to how natural the

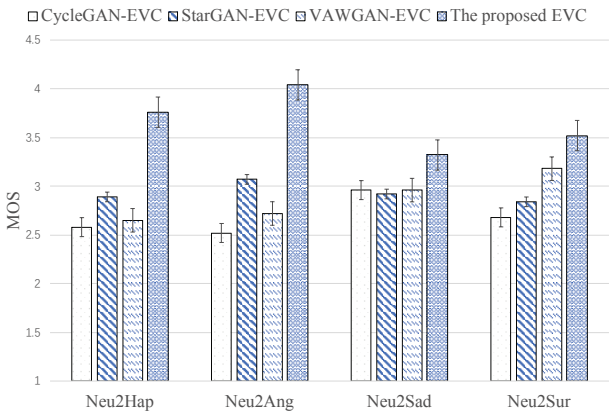


Fig. 7. MOS of the similarity evaluation.

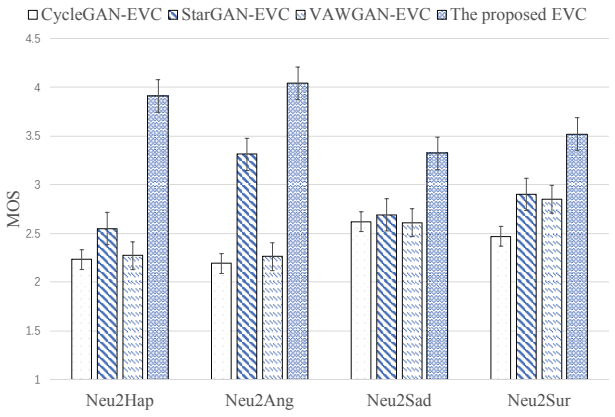


Fig. 8. MOS of the naturalness evaluation.

converted speech sounded to them.

Similarity: Fig. 7 shows MOS results of the similarity evaluation. A higher value indicates a better result. From these results, observe that all naturalness scores are above or near 2.5, indicating reasonable similarity. We can see that our proposed method achieved a better score than those of the baseline approaches, which indicates better subjective conversion quality according to human perceptual evaluation.

Naturalness: Fig. 8 shows MOS results of the naturalness evaluation. A higher value indicates a better result. Comparing the results of different models, we can see that our model also outperforms the other models in the naturalness metric. The results of the two MOS tests are determined with 95% confidence intervals. To summarize, our proposed method performed relatively well in terms of emotion similarity and naturalness for every emotion pair.

C. Ablation Study

As shown in Table. V, we present ablation study results to verify the effectiveness of each loss term used for training. We compared our model with three models trained by a part of the loss function, while keeping the other training setups unchanged, including the model structure. Once we remove adversarial loss L_{ADV}, the naturalness dramatically decreased from 3.69 to 3.4, which indicates that the adversarial learning makes the model generate high-quality and realistic audio signals. We observe a decrease in the similarity score without MI loss L_{MI}. This result indicates that the MI loss L_{MI} for better disentanglement is useful to improve the emotion transfer quality. The performance without term L_{REC} has not been reported because the model cannot generate fluent speech without reconstruction loss.

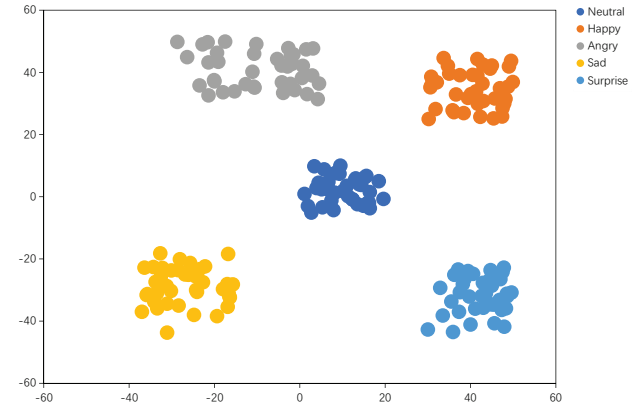


Fig. 9. t-SNE Visualization for emotion embedding of seen speakers. Each point represents an utterance and the legend indicates different emotions.

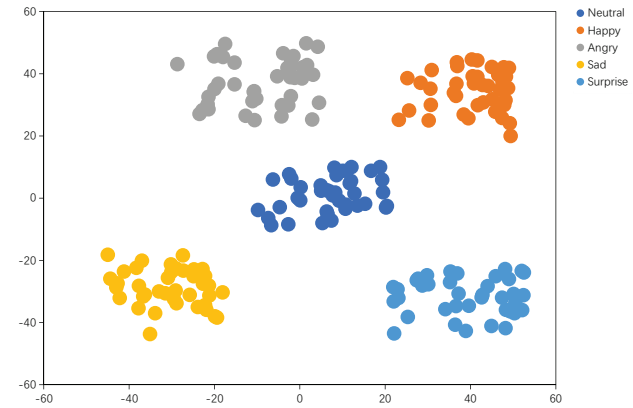


Fig. 10. t-SNE Visualization for emotion embedding of unseen speakers. Each point represents an utterance and the legend indicates different emotions.

D. Evaluation of Disentanglement

In order to further demonstrate that our proposed method can disentangle latent representations effectively, we present t-SNE visualisation [57] for emotion representations obtained from the different utterances of both seen and unseen speakers in Figures 9 and 10, respectively. Fig. 9 presents the visualization results of seen speakers. It is clear that the emotion representations of converted speech are well clustered with identical emotion, and emotion representations of converted speech are well separated among different emotions. This result indicates that our emotion encoder can extract emotion information irrelevant to other information. From Fig. 10, we observe that the emotion representations generated with unseen speakers are also located within the clusters of the same emotion, which indicates the robustness of our emotion encoder.

TABLE V
ABLATION STYDIES.

Method	MCD	F0-RMSE	SER	Naturalness	Similarity
Proposed	4.34	57.26	60.13	3.69	3.50
w/o $\mathcal{L}_{\mathcal{MI}}$	4.97	58.79	54.33	3.56	3.07
w/o $\mathcal{L}_{\mathcal{ADV}}$	4.52	59.92	54.26	3.04	3.29

TABLE VI
THE ACCURACY OF SER BASED ON LATENT REPRESENTATION.

	Without L_{MI}	With L_{MI}
Accuracy	0.36	0.19

We also perform an ablation study to verify the disentanglement of latent representations encoded by the emotion encoder and emotion-independent encoder. The disentanglement is measured as the accuracy of a speech emotion recognition based on only the emotion-independent representation. A higher value indicates that the emotion-independent representation contains the more emotion information. The results are shown in Table VI. We can see that the recognition accuracy is apparently lower when MI loss L_{MI} is applied to train the proposed model, indicating that the emotion-independent representation contains less emotion-related information. Therefore, the MI loss proposed in this paper does encourage better disentanglement for emotional VC.

V. CONCLUSION

In this paper, we have proposed a novel speaker-independent emotional voice conversion method for arbitrary speakers via disentangled representation learning, which we refer to as SIEVC. A novel optimization objective based on mutual information is proposed for better disentanglement. To achieve high-quality converted speech with target emotion, the adversarial training strategy is also adopted in the training procedure. The experimental results show that, our proposed model outperforms the baselines in both seen and unseen speaker scenarios on a real-world dataset.

ACKNOWLEDGMENT

This work was supported in part by JSPS KAKENHI (Grant No. JP21H00906).

REFERENCES

[1] R. Liu, B. Sisman, J. Li, F. Bao, G. Gao, and H. Li, "Teacher-student training for robust tacotron-based tts," in *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2020, pp. 6274–6278.

[2] J. Krivokapić, "Rhythm and convergence between speakers of american and indian english," *Laboratory Phonology*, vol. 4, no. 1, pp. 39–65, 2013.

[3] T. Raitio, L. Juvela, A. Suni, M. Vainio, and P. Alku, "Phase perception of the glottal excitation of vocoded speech," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[4] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "Gmm-based emotional voice conversion using spectrum and prosody features," *American Journal of Signal Processing*, vol. 2, no. 5, pp. 134–138, 2012.

[5] R. Aihara, R. Ueda, T. Takiguchi, and Y. Ariki, "Exemplar-based emotional voice conversion using non-negative matrix factorization," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*. IEEE, 2014, pp. 1–7.

[6] Z. Luo, T. Takiguchi, and Y. Ariki, "Emotional voice conversion using deep neural networks with mcc and f0 features," in *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*. IEEE, 2016, pp. 1–5.

[7] K. Zhou, B. Sisman, and H. Li, "Transforming spectrum and prosody for emotional voice conversion with non-parallel training data," in *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020, pp. 230–237.

[8] G. Rizos, A. Baird, M. Elliott, and B. Schuller, "Stargan for emotional speech conversion: Validated by data augmentation of end-to-end emotion recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3502–3506.

[9] Y. Cao, Z. Liu, M. Chen, J. Ma, S. Wang, and J. Xiao, "Nonparallel emotional speech conversion using vae-gan." in *INTERSPEECH*, 2020, pp. 3406–3410.

[10] J. Gao, D. Chakraborty, H. Tembine, and O. Olaleye, "Nonparallel emotional speech conversion," *arXiv preprint arXiv:1811.01174*, 2018.

[11] K. Zhou, B. Sisman, and H. Li, "Limited data emotional voicec conversion leveraging text-to-speech: two-stage sequence-to-sequence training," in *Interspeech*, 2021, pp. 811–815.

[12] S. Zhou, J. Jia, Z. Wu, Z. Yang, Y. Wang, W. Chen, F. Meng, S. Huang, J. Shen, and X. Wang, "Inferring emotion from large-scale internet voice data: A semi-supervised curriculum augmentation based deep learning approach," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 7, 2021, pp. 6039–6047.

[13] W. Nie, M. Ren, J. Nie, and S. Zhao, "C-gcn: correlation based graph convolutional network for audio-video emotion recognition," *IEEE Transactions on Multimedia*, vol. 23, pp. 3793–3804, 2020.

[14] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech emotion recognition from 3d log-mel spectrograms with deep learning network," *IEEE access*, vol. 7, pp. 125 868–125 881, 2019.

[15] T. Zhang, X. Wang, X. Xu, and C. P. Chen, "Gcb-net: Graph convolutional broad network and its application in emotion recognition," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 379–388, 2019.

[16] S. Liu and T. Zhu, "Structure-guided arbitrary style transfer for artistic image and video," *IEEE Transactions on Multimedia*, vol. 24, pp. 1299–1312, 2021.

[17] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 172–189.

[18] J. J. Virtusio, J. J. M. Ople, D. S. Tan, M. Tanveer, N. Kumar, and K.-L. Hua, "Neural style palette: a multimodal and interactive style transfer from a single style image," *IEEE Transactions on Multimedia*, 2021.

[19] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5210–5219.

[20] J.-c. Chou, C.-c. Yeh, and H.-y. Lee, "One-shot voice conversion by separating speaker and content representations with instance normalization," *arXiv preprint arXiv:1904.05742*, 2019.

[21] S.-H. Lee, J.-H. Kim, H. Chung, and S.-W. Lee, "Voicemixer: Adversarial voice style mixup," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[22] E. H. Sanchez, M. Serrurier, and M. Ortner, "Learning disentangled representations via mutual information estimation," in *European Conference on Computer Vision*. Springer, 2020, pp. 205–221.

[23] X. Peng, Z. Huang, X. Sun, and K. Saenko, "Domain agnostic learning with disentangled representations," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5102–5112.

[24] T.-Y. Hu, A. Shrivastava, O. Tuzel, and C. Dhir, "Unsupervised style and content separation by minimizing mutual information for speech synthesis," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3267–3271.

[25] D. Xin, T. Komatsu, S. Takamichi, and H. Saruwatari, "Disentangled speaker and language representations using mutual information mini-

mization and domain adaptation for cross-lingual tts," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6608–6612.
[26] S. Yuan, P. Cheng, R. Zhang, W. Hao, Z. Gan, and L. Carin, "Improving zero-shot voice style transfer via disentangled representation learning," *arXiv preprint arXiv:2103.09420*, 2021.
[27] D. Wang, L. Deng, Y. T. Yeung, X. Chen, X. Liu, and H. Meng, "VQMIVC: Vector Quantization and Mutual Information-Based Unsupervised Speech Representation Disentanglement for One-Shot Voice Conversion," in *Proc. Interspeech 2021*, 2021, pp. 1344–1348.
[28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.
[29] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.
[30] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8789–8797.
[31] H. Choi and M. Hahn, "Sequence-to-sequence emotional voice conversion with strength control," *IEEE Access*, vol. 9, pp. 42 674–42 687, 2021.
[32] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 920–924.
[33] H. Ning, X. Zheng, X. Lu, and Y. Yuan, "Disentangled representation learning for cross-modal biometric matching," *IEEE Transactions on Multimedia*, vol. 24, pp. 1763–1774, 2021.
[34] H. Xu, X. Wang, and J. Ma, "Drf: Disentangled representation for visible and infrared image fusion," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–13, 2021.
[35] M. Jia, X. Cheng, S. Lu, and J. Zhang, "Learning disentangled representation implicitly via transformer for occluded person re-identification," *IEEE Transactions on Multimedia*, 2022.
[36] K. Zhou, B. Sisman, M. Zhang, and H. Li, "Converting anyone's emotion: Towards speaker-independent emotional voice conversion," *arXiv preprint arXiv:2005.07025*, 2020.
[37] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, "Mutual information neural estimation," in *International Conference on Machine Learning*. PMLR, 2018, pp. 531–540.
[38] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.
[39] P. Cheng, W. Hao, S. Dai, J. Liu, Z. Gan, and L. Carin, "Club: A contrastive log-ratio upper bound of mutual information," in *International Conference on Machine Learning*. PMLR, 2020, pp. 1779–1788.
[40] M. D. Donsker and S. S. Varadhan, "Asymptotic evaluation of certain markov process expectations for large time, i," *Communications on Pure and Applied Mathematics*, vol. 28, no. 1, pp. 1–47, 1975.
[41] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein gans," *arXiv preprint arXiv:1704.00028*, 2017.
[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[43] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.
[44] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.
[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[46] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1874–1883.
[47] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1501–1510.
[48] J.-c. Chou, C.-c. Yeh, H.-y. Lee, and L.-s. Lee, "Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations," *arXiv preprint arXiv:1804.02812*, 2018.
[49] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 266–273.
[50] A. L. Maas, A. Y. Hannun, A. Y. Ng, *et al.*, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. icml*, vol. 30, no. 1. Citeseer, 2013, p. 3.
[51] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," *arXiv preprint arXiv:1910.06711*, 2019.
[52] L.-J. Liu, Z.-H. Ling, Y. Jiang, M. Zhou, and L.-R. Dai, "Wavenet vocoder with limited training data for voice conversion." in *Interspeech*, 2018, pp. 1983–1987.
[53] C. Veaux, J. Yamagishi, K. MacDonald, *et al.*, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2017.
[54] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.
[55] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, vol. 1. IEEE, 1993, pp. 125–128.
[56] M. Chen, X. He, J. Yang, and H. Zhang, "3-d convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018.
[57] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

Xunquan Chen received the B.S. degree with the major in Statistics from Chongqing University, China, in 2018. He received the M.E. degree from Kobe University, Japan, in 2021, where he is currently working toward the Ph.D. degree in computer science. His research interests include machine learning and speech synthesis. He is a Student Member of ASJ.

Xuexin Xu received his BSc degree in computer science from Fujian Normal University, China, in 2020. He is currently a postgraduate student in the School of Informatics, Xiamen University. His research interests are machine learning and voice conversion.

Takashi Kamihigashi received his Ph.D. degree in economics (1994) from University of Wisconsin at Madison (USA). He is currently a professor at the Research Institute for Economics and Business Administration, Kobe University, and director of the Center for Computational Social Sciences (Japan). He specializes in theoretical analysis on Macroeconomic Dynamics, and in recent years has also been developing research and dissemination activities in computational social science. He is the Editor-in-Chief of Computational Social Science. He has served as Editor-in-Chief of the International Journal of Economic Theory, President of IEFS-Japan, and a member of the Science Council of Japan.

Jinhui Chen received his Ph.D. degree (2016) in information science from Kobe University (Japan). From 2016 to 2020, he was an assistant professor at Kobe University. He is currently an associate professor at (Prefectural) University of Hiroshima (Japan). His research interests include pattern recognition (voice and image) and machine learning. He is a member of IEEE, ACM, and IEICE. He has published more than 20 publications in major journals and international conferences, such as IEEE Trans. Multimedia, IEEE/ACM Trans. Audio Speech Lang. Process., ACM MM, Interspeech etc.

Zhihong Zhang received his BSc degree (1st class Hons.) in computer science from the University of Ulster, UK, in 2009 and the Ph.D. degree in computer science from the University of York, UK, in 2013. He won the K. M. Stott prize for best thesis from the University of York in 2013. He is now an associate professor at the school of Informatics, Xiamen University, China. His research interests are wide-reaching but mainly involve the areas of pattern recognition and machine learning, particularly problems involving graphs and networks.

Tetsuya Takiguchi received the M.Eng. and Dr. Eng. degrees in information science from the Nara Institute of Science and Technology, Ikoma, Japan, in 1996 and 1999, respectively. From 1999 to 2004, he was a Researcher with IBM Research, Tokyo Research Laboratory. From 2004 to 2016, he was an Associate Professor with Kobe University, where he has been a Professor since 2016. From May 2008 to September 2008, he was a visiting scholar with the Department of Electrical Engineering, University of Washington, where he was a visiting scholar with the Institute for Learning Brain Sciences from March 2010 to September 2010. From April 2013 to October 2013, he was a visiting scholar with the Laboratoire d'InfoRmatique en Image et Systèmes d'information, INSA Lyon. His research interests include speech, image, and brain processing, and multimodal assistive technologies for people with articulation disorders. He is a member of IEICE, IPSJ, and ASJ.

Edwin R. Hancock holds a B.Sc. degree in physics (1977), a Ph.D. degree in high-energy physics (1981) and a D.Sc. degree (2008) from the University of Durham, and a doctorate Honoris Causa from the University of Alicante in 2015. He is an Emeritus Professor in the Department of Computer Science at the University of York, Adjunct Professor at Beihang University and Distinguished Visiting Professor at Xiamen University. His main research interests are in pattern recognition, machine learning and computer vision, where he has made sustained contributions to the use of graph-based methods and physics-based vision over the past 30 years. He was elected a Fellow of the Royal Academy of Engineering (the UK's national academy of engineering), in 2021. He is also a Fellow of both the International Association for Pattern Recognition and the IEEE. He was the 2016 Distinguished Fellow of the BMVA. He is currently Editor-in-Chief of the journal Pattern Recognition and was founding Editor-in-Chief of IET Computer Vision from 2006 until 2012. He has also been a member of the editorial boards of the journals IEEE Transactions on Pattern Analysis and Machine Intelligence, Pattern Recognition, Computer Vision and Image Understanding, Image and Vision Computing, and the International Journal of Complex Networks. He was Vice President of the IAPR from 2016 to 2018. He has been the recipient of the Pattern Recognition Medal (1992), the IAPR Piero Zamperoni Award (2006), a Royal Society Wolfson Research Merit Award (2008), and the IAPR Pierre Devijver Award (2018), He is an IEEE Computer Society Distinguished Visitor for the period 2021-2023