INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Emotional Voice Conversion with Semi-Supervised Generative Modeling

Hai Zhu[1], *Huayi Zhan*[1,*], *Hong Cheng*[2], *Ying Wu*[3]

[1] Changhong AI Lab (CHAIR), Sichuan Changhong Electronics Holding Group Co., Ltd.
[2]Center for Robotics, University of Electronic Science and Technology of China
[3]Department of Electrical Engineering and Computer Science, Northwestern University, USA
{hai1.zhu, huayi.zhan}@changhong.com,hcheng@uestc.edu.cn,yingwu@ece.northwestern.edu

Abstract

Emotional Voice Conversion (EVC) is a task that aims to convert the emotional state of speech from one to another while preserving the linguistic information and identity of the speaker. However, many studies are limited by the requirement for parallel speech data between different emotional patterns, which is not widely available in real-life applications. Furthermore, the annotation of emotional data is highly time-consuming and labor-intensive. To address these problems, in this paper, we propose SGEVC, a novel semi-supervised generative model for emotional voice conversion. This paper demonstrates that using as little as 1% supervised data is sufficient to achieve EVC. Experimental results show that our proposed model achieves state-of-the-art (SOTA) performance and consistently outperforms EVC baseline frameworks.

Index Terms: emotional voice conversion, variational autoencoder, semi-supervised, end-to-end

1. Introduction

Emotional voice conversion (EVC) is a variety of voice conversion (VC) that aims to transform the emotional state of a spoken utterance from the source to the target, while preserving both the speaker identity and the underlying linguistic content. In recent years, EVC has gained considerable interest and attention within the field of speech technology, and offers great potential for application in human-machine interaction, encompassing expressive text-to-speech (TTS), voice assistants, and conversational robots [1, 2, 3].

Previous studies on EVC has primarily relied upon paired parallel training data, consisting of the same content spoken by the same speaker, but with varying emotions. The early-stage EVC methods learned feature mapping from these paired utterances through Gaussian mixture models (GMM) [4] and regression-based clustering [5]. Recently, deep learning techniques, including those based on Deep Neural Networks (DNN) and Recurrent Neural Networks (RNN) have shown considerable efficacy in EVC [6, 7]. However, obtaining parallel data and the aligning utterance pairs can be arduous and time-consuming. As a result, EVC techniques employing non-parallel data are better suited for real-world applications.

Recent models for non-parallel EVC can generally be categorized into GAN-based [8, 9] and disentanglement-based [10, 11]. GAN-based models, including CycleGAN [8] and StarGAN [9], utilize the cycle-consistency loss to eliminate the need for parallel training data. Disentanglement-based models generally adopt an autoencoder framework to decompose the speech into emotion and emotion-independent representations, for example, VAW-GAN [10] and SIEVC [12]. In addition, studies have shown that utilizing TTS or ASR to model linguistic information in EVC models can significantly reduce pronunciation errors and improve speech quality [13, 14]. To model the emotional information, the style vector derived from the style encoder can be employed as a one-hot vector representing the desired emotion, eliminating the requirement for explicit emotion labeling [13]. However, the style vector cannot be reliably identified during repeated training [15]. Many studies [14, 16] utilize a global emotion embedding obtained through emotion labeling to perform EVC, however, annotating emotion labels on data can be a laborious and time-intensive process. Moreover, previous EVC methods rely on spectral features as an intermediate representation and another neutral vocoder for waveform synthesis, which may affect the synthesis quality. To overcome these limitations, we have proposed a novel semi-supervised generative model, SGEVC.

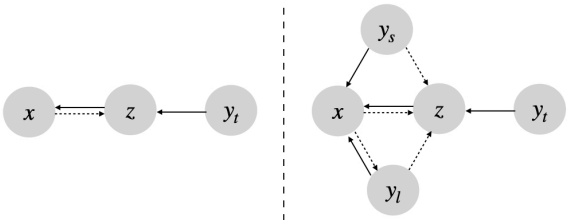


Figure 1: *The Graphical models show the generative model (solid arrow) and the variational family (dashed arrow).*

The present work presents three primary contributions. Firstly, we propose a new generative model for EVC, which enables emotional voice conversion utilizing non-parallel data. Secondly, we introduce a semi-supervised training approach that utilizes a limited number of emotional labels to effectively regulate emotions. Lastly, we demonstrate the state-of-the-art (SOTA) performance of our end-to-end EVC models. The source code and audio samples can be found on GitHub [1].

2. The Proposed Approach

2.1. Generative model

Following the framework of VAE [17], prior work [18] takes text condition $\boldsymbol{y}_t$ as an input to predict waveform $\boldsymbol{x}$, represented by $q(\boldsymbol{x}|\boldsymbol{y}_t)$, by means of a latent representation $\boldsymbol{z}$ generated by the conditional distribution $p(\boldsymbol{z}|\boldsymbol{y}_t)$. The speech waveform $\boldsymbol{x}$ is compressed into a frame-level representation $\boldsymbol{z}$ obtained from the posterior distribution $q(\boldsymbol{z}|\boldsymbol{x})$ to reconstruct the waveform

* corresponding author

[1]https://github.com/haizhu1/sgevc

10.21437/Interspeech.2023-251

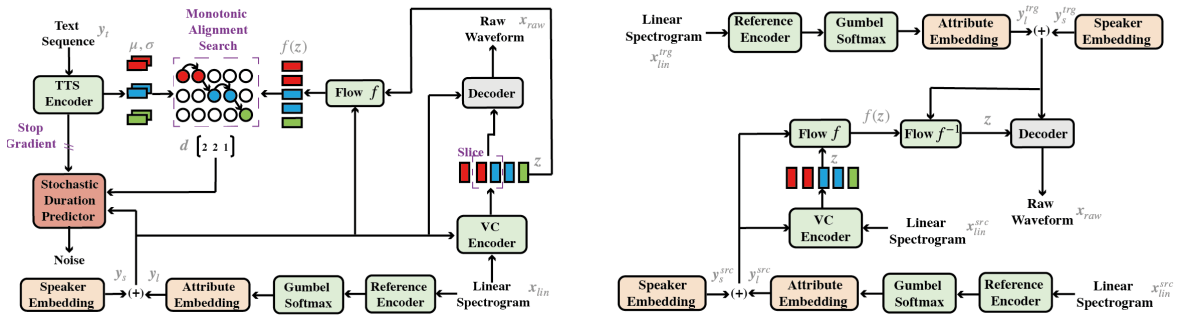


Figure 2: *Diagram of the proposed approach, showing the training procedure (left) and inference procedure(right).*

(represented as $p(\boldsymbol{x}|\boldsymbol{z})$). The above process is illustrated in the left side of Fig. 1.

In this paper, we present a novel graphical model that constructs a space of disentangled representations of linguistic, speaker identity, and emotion attribute. As depicted in the right side of Fig. 1, the conditional distribution $q(\boldsymbol{z}|\boldsymbol{y}_t)$ enables the generation of speaker-independent and emotion-independent linguistic representation through the incorporation of speaker identity $\boldsymbol{y}_s$ and emotional attribute $\boldsymbol{y}_l$ as conditions. Subsequently, a sequence of speech frames is extracted from $p(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{y}_s, \boldsymbol{y}_l)$. The generative model can be written as:

$$p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{y}_l, \boldsymbol{y}_t, \boldsymbol{y}_s) = p(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{y}_s, \boldsymbol{y}_l)p(\boldsymbol{z}|\boldsymbol{y}_t)p(\boldsymbol{y}_l). \quad (1)$$

We assume that the speaker identity $\boldsymbol{y}_s$ is observed, while the emotional attribute $\boldsymbol{y}_l$ follows a categorical distribution with weights specified by $\boldsymbol{\pi}$, i.e., $\boldsymbol{y}_l \sim Cat(\boldsymbol{\pi})$ and $p(\boldsymbol{y}_l) = k^{-1}$ is a fixed uniform prior. To determine the mixture probabilities $q(\boldsymbol{y}_l|\boldsymbol{x})$, we use the categorical re-parameterization technique with Gumbel-Softmax [19]. Specifically, the Gumbel-Softmax layer outputs the mixture probabilities based on Gumbel samples $g_i \sim Gumbel(0, 1)$ that are first sampled and then computed into y_{l_i}, where

$$y_{l_i} = \frac{exp\left(\left(log\pi_i + g_i\right)/\tau\right)}{\sum_{j=1}^{k} exp\left(\left(log\pi_i + g_i\right)/\tau\right)} \; for \; i = 1, ...k, \quad (2)$$

and τ is the temperature coefficient.

In accordance with the VAE framework, we employ a variational distribution $q(\boldsymbol{y}_l|\boldsymbol{x})q(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y}_s, \boldsymbol{y}_l)$ to approximate the posterior $q(\boldsymbol{y}_l, \boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y}_s)$. The evidence lower bound (ELBO) can be expressed as:

$$\begin{aligned}\mathcal{L} &= \mathbb{E}_{q(\boldsymbol{y}_l, \boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y}_s)}\left[log\frac{p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{y}_l, \boldsymbol{y}_t, \boldsymbol{y}_s)}{q(\boldsymbol{y}_l, \boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y}_s)}\right] \quad (3)\\ &= \mathbb{E}_{q(\boldsymbol{y}_l|\boldsymbol{x})q(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y}_s, \boldsymbol{y}_l)}\left[logp(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{y}_s, \boldsymbol{y}_l)\right]\\ &\quad - \mathbb{E}_{q(\boldsymbol{y}_l|\boldsymbol{x})}\left[KL(q(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y}_s, \boldsymbol{y}_l)||p(\boldsymbol{z}|\boldsymbol{y}_t))\right]\\ &\quad - \mathbb{KL}\left(q(\boldsymbol{y}_l|\boldsymbol{x})||p(\boldsymbol{y}_l)\right).\end{aligned}$$

Here, $q(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y}_s, \boldsymbol{y}_l)$ can be estimated via Monte Carlo sampling, and the $q(\boldsymbol{y}_l|\boldsymbol{x})$ term is differentiable due to re-parameterization with Gumbel-Softmax.

2.2. Semi-supervised training

The training objective in Eq. (3) enables emotional voice conversion by extracting the unsupervised style vector from the target speech. However, the style vector is not formally identifiable and may not consistently capture the same latent attribute across multiple training iterations [15]. This can potentially degrade the quality of emotional voice conversion. Therefore, to address these concerns, we opt to provide partial supervision to the latent variable $\boldsymbol{y}_l$. In cases where emotional labels $\boldsymbol{y}_{obs}$ are available, we introduce a supervised loss that modifies the Eq. (3) as follows:

$$\begin{aligned}\mathcal{L}_{sup} &= \mathbb{E}_{q(\boldsymbol{y}_l|\boldsymbol{x})q(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y}_s, \boldsymbol{y}_l)}[logp(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{y}_s, \boldsymbol{y}_l)] \quad (4)\\ &\quad - \mathbb{E}_{q(\boldsymbol{y}_l|\boldsymbol{x})}\left[KL(q(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y}_s, \boldsymbol{y}_l)||p(\boldsymbol{z}|\boldsymbol{y}_t))\right]\\ &\quad + \alpha log(q(\boldsymbol{y}_l = \boldsymbol{y}_{obs})|x),\end{aligned}$$

where α is a hyperparameter utilized to regulate the final term's contribution. Here, we replace the final term in the original ELBO with a supervised cross-entropy term, which encourages $q(\boldsymbol{y}_l|\boldsymbol{x})$ to match the observed label.

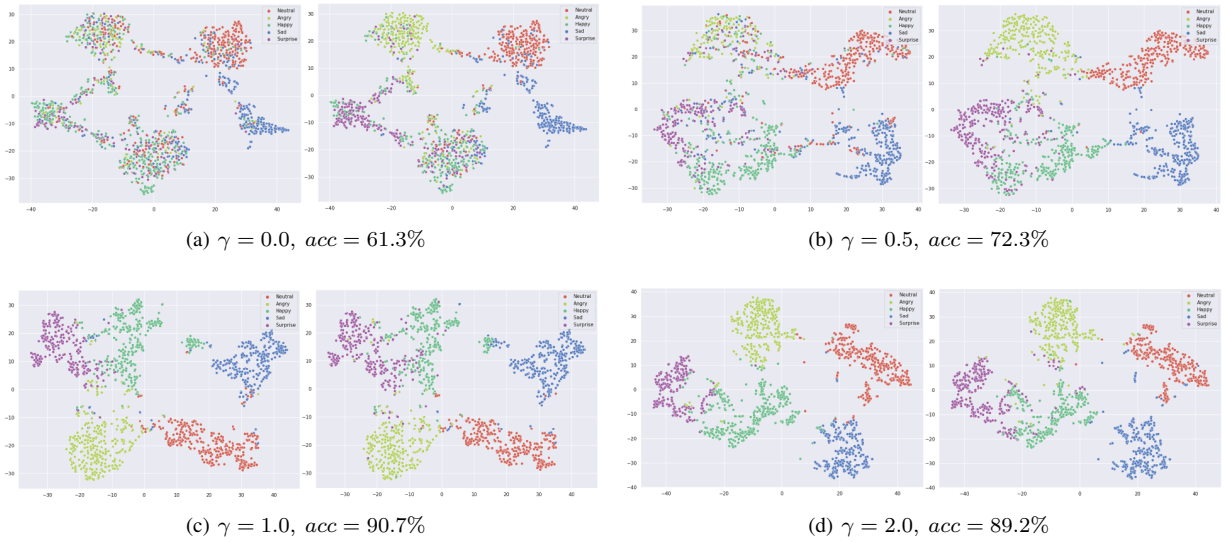
In our experiments, we observed that the model collapses into the $\boldsymbol{y}_l$ prior when using Eq. (3), which leads to low emotion classification accuracy. To address this issue, we adopt a modification to the lower bound proposed by [20], which maintains the cost from the $\boldsymbol{y}_l$ prior term at a constant value γ when it falls below a certain threshold. This allows us to use the unsupervised loss, adapted from Eq. (3), in cases where emotion labels $\boldsymbol{y}_{obs}$ are not available:

$$\begin{aligned}\mathcal{L}_{unsup} &= \mathbb{E}_{q(\boldsymbol{y}_l|\boldsymbol{x})q(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y}_s, \boldsymbol{y}_l)}[logp(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{y}_s, \boldsymbol{y}_l)] \quad (5)\\ &\quad - \mathbb{E}_{q(\boldsymbol{y}_l|\boldsymbol{x})}\left[KL(q(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y}_s, \boldsymbol{y}_l)||p(\boldsymbol{z}|\boldsymbol{y}_t))\right]\\ &\quad - max(\gamma, \mathbb{KL}(q(\boldsymbol{y}_l|\boldsymbol{x})||p(\boldsymbol{y}_l))).\end{aligned}$$

2.3. Model architecture

As illustrated in Fig. 2, we utilize neural networks to parameterize $p(\boldsymbol{z}|\boldsymbol{y}_t)$ and $p(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{y}_s, \boldsymbol{y}_l)$ as the TTS encoder and decoder. The two posteriors, $q(\boldsymbol{y}_l|\boldsymbol{x})$ and $q(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y}_s, \boldsymbol{y}_l)$, can be parameterized by latent attribute encoder and VC encoder. Our proposed architecture, based on the VITS [18], incorporates several novel modifications to implement our graphical model.

The TTS encoder comprises of a transformer-based encoder that accepts input text $\boldsymbol{y}_t$ and generates hidden representations, along with a linear projection layer that produces the mean and variance for the prior distribution $p(\boldsymbol{z}|\boldsymbol{y}_t)$ from the encoder's output. To transition $\boldsymbol{z}$ vectors from text-level to frame-level during training and prediction, we utilize methods of monotonic alignment search and stochastic duration predictor from VITS. The VC encoder takes the linear-scale spectrogram and passes it through a 1-D convolutional layer, followed by conditional WaveNet residual blocks that are conditioned by the observed speaker embedding $\boldsymbol{y}_s$ and emotion attribute embedding $\boldsymbol{y}_l$. The resulting output is then passed through two distinct linear projection layers to create the posterior distribution



(a) $\gamma = 0.0,\ acc = 61.3\%$ (b) $\gamma = 0.5,\ acc = 72.3\%$

(c) $\gamma = 1.0,\ acc = 90.7\%$ (d) $\gamma = 2.0,\ acc = 89.2\%$

Figure 3: *t-SNE plots showing the latent space of $\boldsymbol{y}_l$ under different value of γ on testing dataset (α is set to 0.1), with points colored by categories predicted by our proposed model (left), and ground truth categories (right).*

$q(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y}_s, \boldsymbol{y}_l)$ by producing the mean and variance. To improve the expressive capability of VAE, we incorporate normalizing flows between the TTS encoder and VC encoder, similar to VITS. We employ the same HiFi-GAN [21] version1 structure for the decoder as used in VITS, and also adopt the same adversarial training process introduced by VITS to enhance the speech quality of the decoder.

The speaker embedding $\boldsymbol{y}_s$ is obtained from a speaker look-up table. For the latent attribute embedding $\boldsymbol{y}_l$, we first pass the linear-scale spectrogram through a reference encoder [22], which consists of several convolutional layers and a mean pooling layer to get a single vector. The reference encoder's output, which represents the utterance-level acoustic conditions of the target speech, is then passed through a Gumbel-Softmax layer [19] to generate k-dimensional sample vectors through sampling, and is then flattened and passed through a linear projection layer to generate a fixed-dimensional latent attribute embedding. During the training stage, we add the non-linguistic information of the same speaker, which is represented by the sum of $\boldsymbol{y}_s$ and $\boldsymbol{y}_l$, to the VC encoder, stochastic duration predictor, normalizing flow f, and decoder. During the inference stage, we add the non-linguistic information of the source speaker, represented by the sum of $\boldsymbol{y}_s^{src}$ and $\boldsymbol{y}_l^{src}$, to the VC encoder and the normalizing flow f. Additionally, we add the non-linguistic information of target speaker, which is represented by the sum of $\boldsymbol{y}_s^{trg}$ and $\boldsymbol{y}_l^{trg}$, to the decoder and the inverse transformation of the normalizing flow f^{-1}.

3. Experiments and Analysis

3.1. Experiment setup

We conducted experiments on the emotional speech dataset (ESD) [10], which comprises 350 parallel utterances averaging 2.9 seconds in length and recorded by 10 native Mandarin speakers and 10 native English speakers. The corpus for each speaker includes five emotions: happy, sad, neutral, angry, and surprised. In this study, we focus only on the Mandarin-speaking subset of the dataset. For each speaker, we perform emotion conversion from neutral to happy (N2H), neutral to angry (N2A), neutral to sad (N2S1), and neutral to surprised (N2S2) within the same speaker. For each conversion pair, we partition the corpus into a training set (330 samples) and a testing set (20 samples). To ensure that our proposed model is not trained with parallel conditions, we randomly shuffle the training set and generate non-parallel utterances for each training batch. To assess the effectiveness of our proposed method, we compare its results with the state-of-the-art methods: the StarGAN-based and the PPG-based EVC models.

The StarGAN [9] can perform multi-class to multi-class emotional voice conversion and employ the WORLD vocoder for speech synthesis. As for the PPG-based EVC baseline model, we choose the BNE-Seq2seqMoL model [23, 24], which utilizes the pinch encoder to capture emotional prosody and uses HiFi-GAN version1 as the vocoder. To ensure a fair comparison, we employ available open-source implementations and train the models using the same data. All models were trained until they reached convergence. We fine-tune the well-trained BNE-Seq2seqMoL model for 10k steps and the HiFi-GAN vocoder for 1M steps with the ESD training set. For our proposed model, $\boldsymbol{y}_s$ is a 256-dimensional vector, and $\boldsymbol{y}_l$ is a 5-way categorical variable with a class dimension 32. The latent variable $\boldsymbol{z}$ is 192-dimensional vector and is assumed to have a Gaussian distribution. We obtain our linear spectrogram $\boldsymbol{x}_{lin}$ from raw waveforms $\boldsymbol{x}_{raw}$ through Short-time Fourier transform (STFT) with FFT size 1024, window size 1024, and hop size 256. We gradually decrease the temperature τ in the Gumbel-Softmax layer using a schedule $\tau = max(1.0, exp(-3\times10^{-5}t))$ of the global training step t, where τ is updated every 1000 steps. All of our proposed models were trained for 200k steps, with a batch size of 64.

3.2. Evaluation of emotion control

To evaluate the effectiveness of the semi-supervised latent variable $\boldsymbol{y}_l$ for emotion control, we trained the model detailed in

Table 1: *The accuracy of emotion classification*

Supervision level (%) (γ is set to 1.0)	Accuracy(%)		
	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 1.0$
1	66.7	84.1	82.5
5	81.4	87.7	83.9
10	82.3	91.3	87.6
20	84.4	93.5	89.3
100	86.5	93.3	91.9

Table 2: *MOS results with 95 % confidence interval to assess the speech naturalness.*

MOS	N2H	N2A	N2S1	N2S2
Ground-Truth	4.52±0.13	4.78±0.09	4.49±0.16	4.73±0.11
StarGAN	2.65±0.13	2.64±0.17	2.67±0.15	2.62±0.14
PPG	3.05±0.16	3.12±0.15	3.15±0.16	3.02±0.13
SGEVC-1	4.21±0.17	4.24±0.16	**4.37±0.14**	4.19±0.12
SGEVC-10	**4.29±0.12**	**4.28±0.12**	4.31±0.13	**4.24±0.14**

Section 2 on the ESD dataset using varying hyperparameter settings. This involved adjusting the values of γ, which resolve the KL collapse problem, and α, which regulate the supervision loss. We first visualizing the latent variable space of $\boldsymbol{y}_l$ for different values of γ in order to determine the optimal value. We then computed the accuracy of emotion classification under different levels of supervision and various values of α.

We selected 20 utterances from each emotion category in the training set as supervised data, while the rest was used as unsupervised data. To evaluate the accuracy of our proposed approach, we assigned the utterances generated by our method from the testing set to the emotion category with the highest posterior probability ($argmax_{\boldsymbol{y}_l} q(\boldsymbol{y}_l|\boldsymbol{x})$), and calculated the accuracy based on the ground truth. As illustrated in 3, it was observed that with γ being set to zero, each data point was dispersed across all clusters, resulting in a mere 61.3% accuracy in emotion classification. This phenomenon can be interpreted as the result of over-regularisation by the $\boldsymbol{y}_l$ prior. However, when γ is not zero, we observed that these embeddings became more emotion-discriminative. We found a value of $\gamma = 1.0$ achieved the highest classification accuracy of 90.7%. These results confirm the effectiveness of our modification to the lower bound.

In addition, we evaluated the degree of emotional control of our model by assessing its emotion classification accuracy on a test set with multiple levels of supervision, similar to the metric used in [15]. As shown in Table 1, We discovered that the best emotional control, as measured by classifier accuracy, is achieved when $\alpha = 0.1$. At the setting of $\alpha = 0.1$, our proposed SGEVC model is able to achieve an emotion category accuracy of 84% with only 1% supervision level (around 20 seconds per emotion for each speaker, resulting in a total of approximately 15 minutes of labeled emotional data). This suggests that the SGEVC model is highly effective in controlling emotions with a very small amount of supervised data. Annotating only 20 seconds of emotional data per speaker per emotion is feasible for most teams constructing EVC systems. We strongly encourage readers to listen to our demo page [2].

[2]https://haizhu1.github.io/sgevc/

Table 3: *MOS results with 95 % confidence interval to assess the emotion similarity.*

MOS	N2H	N2A	N2S1	N2S2
StarGAN	1.90±0.13	1.98±0.14	2.13±0.15	1.79±0.16
PPG	2.46±0.18	2.34±0.19	2.68±0.17	2.39±0.18
SGEVC-1	**3.72±0.13**	3.32±0.18	3.58±0.13	3.43±0.16
SGEVC-10	3.71±0.12	**3.64±0.12**	**3.64±0.13**	**3.63±0.15**

Table 4: *MCD results for emotional voice conversion.*

MCD	N2H	N2A	N2S1	N2S2
StarGAN	4.98	4.96	5.08	5.21
PPG	3.95	3.82	5.61	4.09
SGEVC-1	3.49	3.41	3.79	3.76
SGEVC-10	**3.45**	**3.40**	**3.77**	**3.72**

3.3. Subjective and objective evaluation

To evaluate the naturalness and similarity of the converted speech, we employed the same methodology as detailed in [25]. The naturalness of speech was measured using mean opinion score (MOS) on a 5-point scale, ranging from 1-bad to 5-excellent. For similarity evaluation, we requested the listeners to rate the similarity of speech pairs using a 4-point scale, ranging from: (1) different emotion, absolutely sure, (2) different emotion, not sure, (3) same emotion, not sure, (4) same emotion, absolutely sure. We invited 8 raters, all of whom are native Mandarin speakers. Additionally, we conducted an objective evaluation using Mel-cepstral distortion (MCD) to quantify the spectral distortion between the generated and ground truth speech. The proposed EVC models were trained and evaluated as SGEVC-1 (1% supervision level, $\gamma = 1.0$, $\alpha = 0.1$) and SGEVC-10 (10% supervision level, $\gamma = 1.0$, $\alpha = 0.1$). For comparison purposes, we also trained and evaluated two baseline models, referred to as StarGAN and PPG.

Table 2 illustrates the superior performance of the SGEVC-10 model in all emotion conversion pairs. Surprisingly, the naturalness of SGEVC-1 model is scarcely impacted, despite having only 1% emotional supervision. Both models exhibit state-of-the-art (SOTA) performance, as their naturalness scores exceed 4.0. We argue that the end-to-end architecture of our EVC framework has played a pivotal role in augmenting the quality of speech. Moreover, our proposed models outperform other baseline models in the similarity metric, indicating better emotion similarity as per human perceptual evaluation, as shown in Table 3. Table 4 further reports the MCD results of the emotion conversion pairs, demonstrating that SGEVC-10 outperforms other models in terms of transfer performance.

4. Conclusions

This paper presents the SGEVC, an end-to-end semi-supervised generative model designed for emotional voice conversion. The proposed method leverages VAE frameworks to disentangle linguistic, speaker identity, and emotion spaces. In addition, this paper introduce a semi-supervised training approach that utilizes a limited number of emotional labels to effectively control emotions. Experiments show that our proposed model can effectively conduct emotional voice conversion with only 1% supervised data, and it achieves state-of-the-art performance.

5. References

[1] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," *arXiv preprint arXiv:2106.15561*, 2021.

[2] Y. Zheng, R. Zhang, M. Huang, and X. Mao, "A pre-training based personalized dialogue generation model with persona-sparse data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 9693–9700.

[3] J. Crumpton and C. L. Bethel, "A survey of using vocal prosody to convey emotion in robot speech," *International Journal of Social Robotics*, vol. 8, pp. 271–285, 2016.

[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] C.-H. Wu, C.-C. Hsia, C.-H. Lee, and M.-C. Lin, "Hierarchical prosody conversion using regression-based clustering for emotional speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1394–1405, 2009.

[6] J. Lorenzo-Trueba, G. E. Henter, S. Takaki, J. Yamagishi, Y. Morino, and Y. Ochiai, "Investigating different representations for modeling and controlling multiple emotions in dnn-based speech synthesis," *Speech Communication*, vol. 99, pp. 135–143, 2018.

[7] H. Ming, D.-Y. Huang, L. Xie, J. Wu, M. Dong, and H. Li, "Deep bidirectional lstm modeling of timbre and prosody for emotional voice conversion." in *Interspeech*, 2016, pp. 2453–2457.

[8] K. Zhou, B. Sisman, and H. Li, "Transforming spectrum and prosody for emotional voice conversion with non-parallel training data," *arXiv preprint arXiv:2002.00198*, 2020.

[9] 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.

[10] 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.

[11] M. Elgaar, J. Park, and S. W. Lee, "Multi-speaker and multi-domain emotional voice conversion using factorized hierarchical variational autoencoder," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7769–7773.

[12] X. Chen, X. Xu, J. Chen, Z. Zhang, T. Takiguchi, and E. R. Hancock, "Speaker-independent emotional voice conversion via disentangled representations," *IEEE Transactions on Multimedia*, 2022.

[13] T.-H. Kim, S. Cho, S. Choi, S. Park, and S.-Y. Lee, "Emotional voice conversion using multitask learning with text-to-speech," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7774–7778.

[14] S. Liu, Y. Cao, and H. Meng, "Multi-target emotional voice conversion with neural vocoders," *arXiv preprint arXiv:2004.03782*, 2020.

[15] R. Habib, S. Mariooryad, M. Shannon, E. Battenberg, R. Skerry-Ryan, D. Stanton, D. Kao, and T. Bagby, "Semi-supervised generative modeling for controllable speech synthesis," in *International Conference on Learning Representations*, 2019.

[16] K. Zhou, B. Sisman, and H. Li, "Limited data emotional voice conversion leveraging text-to-speech: Two-stage sequence-to-sequence training," *arXiv preprint arXiv:2103.16809*, 2021.

[17] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[18] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.

[19] E. Jang, S. Gu, and B. Poole, "Categorical reparametrization with gumble-softmax," in *International Conference on Learning Representations (ICLR 2017)*. OpenReview. net, 2017.

[20] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling, "Improved variational inference with inverse autoregressive flow," *Advances in neural information processing systems*, vol. 29, 2016.

[21] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[22] M. Chen, X. Tan, B. Li, Y. Liu, T. Qin, T.-Y. Liu *et al.*, "Adaspeech: Adaptive text to speech for custom voice," in *International Conference on Learning Representations*.

[23] S. Liu, J. Zhong, L. Sun, X. Wu, X. Liu, and H. Meng, "Voice conversion across arbitrary speakers based on a single target-speaker utterance," in *Proc. Interspeech 2018*, 2018, pp. 496–500. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1504

[24] S. Liu, Y. Cao, D. Wang, X. Wu, X. Liu, and H. Meng, "Any-to-many voice conversion with location-relative sequence-to-sequence modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1717–1728, 2021.

[25] Y. Zhao, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z. Ling, and T. Toda, "Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion," *arXiv preprint arXiv:2008.12527*, 2020.