ClapFM-EVC: High-Fidelity and Flexible Emotional Voice Conversion with Dual Control from Natural Language and Speech

Anonymous submission to Interspeech 2025

Abstract

Despite great advances, achieving high-fidelity emotional voice conversion (EVC) with flexible and interpretable control remains challenging. This paper introduces ***ClapFM-EVC***, a novel EVC framework capable of generating high-quality converted speech driven by natural language prompts or reference speech with adjustable emotion intensity. We first propose EVC-CLAP, an emotional contrastive language-audio pretraining model, guided by natural language prompts and categorical labels, to extract and align fine-grained emotional elements across speech and text modalities. Then, a FuEncoder with an adaptive intensity gate is presented to seamless fuse emotional features with Phonetic PosteriorGrams from a pre-trained ASR model. To further improve emotion expressiveness and speech naturalness, we propose a flow matching model conditioned on these captured features to reconstruct Mel-spectrogram of source speech. Subjective and objective evaluations validate the effectiveness of ClapFM-EVC.

Index Terms: emotional voice conversion, natural language prompt, contrastive language-audio pretraining, conditional flow matching

1. Introduction

Emotional voice conversion (EVC) aims to convert the emotional state of source speech to a target category while preserving original content and speaker identity [1]. Recently, EVC has garnered great attention within the speech processing realms and holds great potential for many practical applications such as voice assistant, audiobook production, and dubbing [2, 3, 4].

In general, the key challenges for EVC lie in the accurate and efficient extraction, decoupling, and use of various speech attributes, such as the emotion [5, 6], content [7, 8], and timbre [9, 10] information from speech signals. With rapid progress in deep learning, existing EVC methods are generally based on generative adversarial networks (GANs) [11, 12, 13] and autoencoder models [14, 15, 16]. StarGAN [11] used a cycle-consistent and class-conditional GAN to achieve EVC. [16] proposed AINN, an attention-based interactive disentangling network with a two-stage pipeline for fine-grained EVC. However, the converted speech of these systems lacks emotional diversity, which is crucial for realistic speech synthesis [17]. To this end, several studies [16, 17, 18] shifted towards incorporating intensity control modules into EVC framework to allow more precise manipulation of emotional expression. Emovox [18] disentangled speaker style and controlled emotional intensity by encoding emotion in a continuous space. EINet [17] predicted emotional class and intensity via an emotion evaluator and intensity mapper, incorporating controllable emotional intensity to enhance naturalness and diversity of emotion conversion.

Despite impressive advances, these approaches still face challenges. First, current EVC systems based on GANs and autoencoders, while promising, have great potential for improvements in emotional diversity, naturalness, and speech quality [17, 19]. Second, current methods typically rely on reference speech or categorical text labels as conditions to control a limited set of emotional expressions. Nevertheless, this paradigm not only imposes constraints on the user experience, but restricts the diversity of emotional expressions, while falling short in intuitiveness and interpretability of conveyed emotions.

To mitigate the aforementioned issues, this paper presents **ClapFM-EVC**, an innovative any-to-one EVC framework that enables flexible and intuitive control of emotion conversion in a user-friendly manner. To elaborate, we first propose EVC-CLAP (Contrastive Language Audio Pretraining), which is guided by both natural language prompts and emotional categorical labels, so as to extract and align the emotion features across speech-text modalities. Additionally, we introduce an end-to-end voice conversion (VC) model, termed AdaFM-VC, composed of pre-trained ASR model, FuEncoder and conditional flow matching (CFM) model. Using FuEncoder with an adaptive intensity gate (AIG), AdaFM-VC is able to integrate the captured emotional representations with Phonetic Posteriorgrams (PPGs) from a pre-trained ASR model HybridFormer [20], while allowing flexible control over the emotional intensity of the converted waveform. To further enhance emotional expressiveness, naturalness and speech quality, we incorporate a CFM-based decoder [21, 22] that samples the output of the FuEncoder from random Gaussian noise and reconstructs the Mel-spectrogram of the source speech. During inference, EVC-CLAP can generate target emotional embeddings based on the given natural language prompt, and then AdaFM-VC leverages the target emotion vectors, source PPGs, and predefined emotional intensity to reconstruct the target Mel-spectrogram, which is ultimately converted into the target speech by a pre-trained vocoder [23]. Extensive experiments and ablation studies demonstrate that our ClapFM-EVC significantly outperforms several existing EVC approaches in terms of emotional expressiveness, speech naturalness, and speech quality.

2. METHODOLOGY

2.1. System Overview

As illustrated in Fig. 1, ClapFM-EVC can be characterized as a conditional latent model, where the proposed EVC-CLAP, FuEncoder, CFM-based decoder, as well as pretrained ASR [20] and vocoder [23] models serve as its core components.

Similarly to [24, 25], we first train EVC-CLAP using a symmetric Kullback-Leibler divergence based contrastive loss (symKL-loss) along with soft labels derived from natural lan-

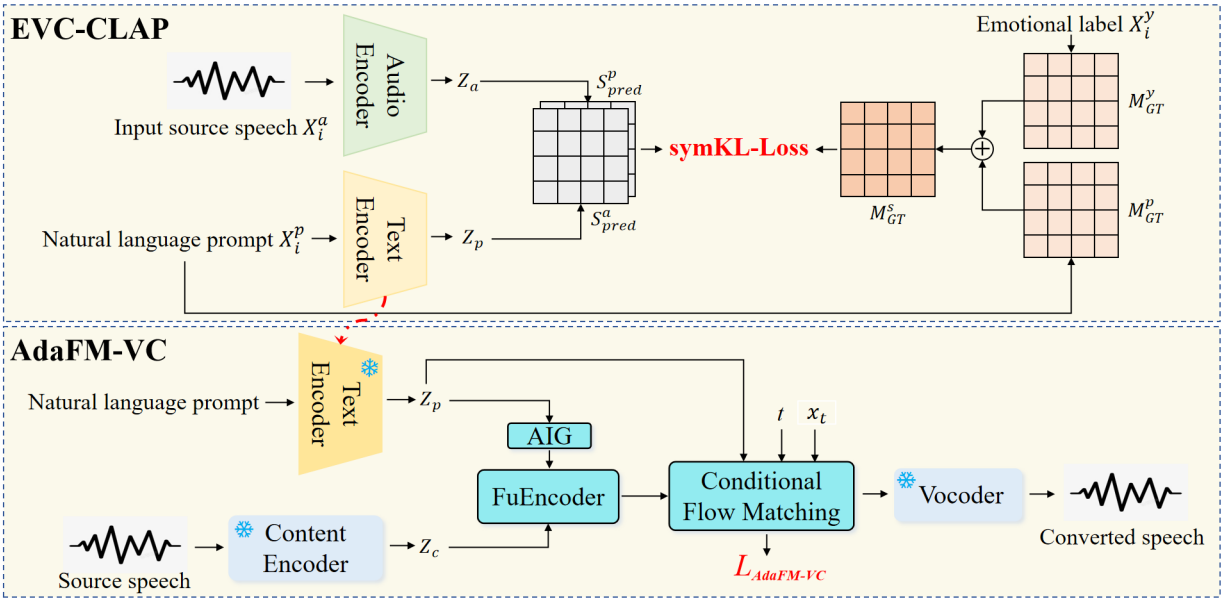

Figure 1: *Overall training architecture of the proposed ClapFM-EVC framework.*

guage prompts and their corresponding categorical emotion labels. In this way, EVC-CLAP can effectively extract and align emotional representations across audio and text modalities, while enabling ClapFM-EVC to capture fine-grained emotional information conveyed by natural language prompts. Then, we train the AdaFM-VC using the obtained emotional elements and content representations extracted by EVC-CLAP and pre-trained HybridFormer, respectively. The FuEncoder within AdaFM-VC facilitates the seamless integration of emotional and content characteristics, with its AIG module explicitly controlling the intensity of emotional conversion. Concurrently, the CFM model in AdaFM-VC samples the outputs of the FuEncoder from random Gaussian noise, and, conditioned on the target emotional vector produced by EVC-CLAP, it generates the Mel-spectrogram features of target speech. Finally, the generated Mel-spectrogram features are fed into a pre-trained vocoder to synthesize the converted speech.

During inference, it is worth noting that our ClapFM-EVC framework provides three modes for obtaining the target emotional embeddings: (1) directly based on the provided reference speech; (2) directly based on the given natural language emotional prompt; and (3) EVC-CLAP retrieves relevant data from a pre-constructed high-quality reference speech corpus using specified natural language emotional prompt, subsequently extracting the target emotion elements from the retrieved speech.

2.2. Soft-Labels-Guided EVC-CLAP

Overall, the purpose of Emo-CLAP training is to minimize the distance between data pairs within the same class, while simultaneously maximizing the distance between pairs of data from different categories.

Assume that the input data pair is $\{X_i^a, X_i^y, X_i^p\}$, where X_i^a is the source speech, X_i^y and X_i^p denote its corresponding emotional label and natural language prompt, $i \in [0, N]$ and N is the batch size. Our EVC-CLAP first adopts a pre-trained HuBERT[1] [26] based audio encoder and a pre-trained XLM-RoBERTa[2] [27] based text encoder to compress X_i^a and X_i^p into two latent variables $Z_a \in \mathbb{R}^{N \times D}$ and $Z_p \in \mathbb{R}^{N \times D}$, where D equals 512, representing the hidden state dimension. Following this, we compute their corresponding similarity matrices S_{pred}^a and S_{pred}^p as:

$$\begin{aligned} S_{pred}^a &= \varepsilon_a \times (Z_a \cdot {Z_p}^T) \\ S_{pred}^p &= \varepsilon_t \times (Z_p \cdot {Z_a}^T) \end{aligned} \tag{1}$$

where ε_a and ε_t are two learnable hyper-parameters, with their values empirically initialized to 2.3. Subsequently, we employ symKL-loss to train Emo-CLAP with the guidance of the soft labels $M_{GT}^s \in \mathbb{R}^{N \times N}$ derived from X_i^y and X_i^p.

$$M_{GT}^s = \alpha_e M_{GT}^y + (1 - \alpha_e) M_{GT}^p \tag{2}$$

where α_e is a hyper-parameter to adjust M_{GT}^y and M_{GT}^p, empirically set to 0.2 in our case. In detail, if the categorical emotional labels or natural language prompt labels of different data pairs within the same batch are identical, their corresponding ground truth is assigned a value of 1; otherwise, it is set to 0. To ensure the consistency of the label distributions across the batch, the class similarity matrices M_{GT}^E and M_{GT}^P are normalized such that the sum of each row equals 1, effectively capturing the relative similarity between data pairs. Therefore, the training loss of EVC-CLAP can be formulated as:

$$\begin{aligned} L_{symKL} = \frac{1}{4} \Bigg(& KL\left(S_{pred}^a || M_{GT}^s\right) + KL\left(\tilde{M}_{GT}^s || S_{pred}^a\right) \\ & + KL\left(S_{pred}^p || M_{GT}^s\right) + KL\left(\tilde{M}_{GT}^s || S_{pred}^p\right) \Bigg) \end{aligned} \tag{3}$$

$$\tilde{M}_{GT}^s = (1 - \alpha) \cdot M_{GT}^s + \frac{\alpha}{N} \tag{4}$$

$$KL(S||M) = \sum_{i,j} S(i,j) \log \frac{S(i,j)}{M(i,j)} \tag{5}$$

where α is a hyper-parameter, empirically set to 1×10^{-8}.

[1] https://huggingface.co/TencentGameMate/chinese-hubert-large
[2] https://huggingface.co/FacebookAI/xlm-roberta-base

2.3. AdaFM-VC

2.3.1. FuEncoder with AIG

As a pivotal intermediate component within ClapFM-EVC, FuEncoder aims to seamlessly integrate content features extracted by HybridFormer with emotional embeddings derived from EVC-CLAP, while offering flexible control over the emotion intensity through the adaptive intensity gate, namely AIG.

Detailed, FuEncoder comprises a preprocessing network (PreNet), a positional encoding module, an AIG module, an adaptive fusion module, and a linear mapping layer. First, PreNet is presented to compress the source content features Z_c to a latent space, preventing overfitting through a dropout mechanism. Next, a positional encoding module is advocated to employ sinusoidal positional encoding to extract the positional characteristics of Z_c and performs element-wise addition with Z_c to ensure that FuEncoder learns its sequential and structural information. Afterwards, we propose an AIG module to multiply a learnable hyperparameter by the EVC-CLAP's emotional features to flexibly adjust the emotional intensity. As the core of FuEncoder, the adaptive fusion module consists of multiple fusion blocks, each of which contains a multi-head self-attention layer, two emotion adaptive layer norm layers [28], and a position-wise feed-forward network layer, enabling efficient fusion of content and emotional information, thus generating rich embedding representations that contain both linguistic and emotional characteristics. The fused features are ultimately mapped to the specific dimensions $f \in R^{B\times T\times D}$ through a fully connected layer.

2.3.2. Conditional Flow Matching-based Decoder

To further enhance emotion expressiveness, speech naturalness, and quality, we incorporate an optimal transport (OT)-based CFM model to reconstruct the target Mel-spectrogram $x_1 = p_1(x)$ from a standard Gaussian noise $x_0 = p_0(x) = \mathcal{N}(x; 0, I)$. To elaborate, conditioned on the captured EVC-CLAP's emotional embeddings, an OT flow $\psi_t : [0, 1] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ is adopted to train our CFM-based decoder, which consists of 6 CFM blocks with timestep fusion. Each CFM block contains a ResNet [29] module, a multi-head self-attention [30] module and a FiLM [31] layer. By utilizing an ordinary differential equation to model a learnable and time-dependent vector field $v_t : [0, 1] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$, the flow can approximate the optimal transport path from $p_0(x)$ to the target distribution $p_1(x)$:

$$\frac{d}{dt}\psi_t(x) = v_t(\psi_t(x), t) \tag{6}$$

where $\psi_0(x) = x$ and $t \in [0, 1]$. Besides, drawing inspiration from previous works [32], which suggest adopting straighter trajectories, we simplify the OT flow formula as follows:

$$\psi_{t,z}(x) = \mu_t(z) + \sigma_t(z)x \tag{7}$$

where $\mu_t(z) = tz$, $\sigma_t(z) = (1 - (1 - \sigma_{\text{min}})t)$, and z represents the random conditioned input. σ_{min} denotes the minimum standard deviation of the white noise introduced to perturb individual samples, with its value empirically set to 0.0001.

Consequently, the training loss for AdaFM-VC is defined as:

$$\mathcal{L} = \mathbb{E}_{t,p(x_0),q(x_1)}\|(x_1-(1-\sigma)x_0)-v_t(\psi_{t,x_1}(x_0)|h)\|^2 \tag{8}$$

where $x_0 \sim p(x_0)$, $x_1 \sim q(x_1)$, $t \sim U[0, 1]$, $q(x_1)$ denotes the true yet potentially non-Gaussian distribution of the data, h refers to the conditional emotion embeddings extracted by EVC-CLAP.

3. EXPERIMENTS

3.1. Experimental Setups

3.1.1. Datasets

Since no open-source EVC corpus with comprehensive emotional natural language prompts is currently available, we leverage an internally developed expressive single-speaker Mandarin corpus for training the proposed ClapFM-EVC system. This corpus encompasses 20 hours of speech data sampled at 24 kHz. From this, we specifically selected 12,000 utterances representing 7 original categorical emotion classes (neutral, happy, sad, angry, fear, surprise, disgust). To ensure high-quality annotations, we enlisted 15 professional annotators to provide natural language prompts for the selected waveforms.

3.1.2. Implementation Details

In all experiments, the proposed EVC-CLAP and AdaFM-VC models are trained end-to-end using 8 NVIDIA RTX 3090 GPUs. For training the EVC-CLAP model, the Adam optimizer is employed with an initial learning rate of 1×10^{-5} and a batch size of 16. All models are trained for 40 epochs within the PyTorch framework. Afterwards, the AdaFM-VC approach is trained using the AdamW optimizer over 500,000 iterations on the same GPU setup, with an initial learning rate of 2×10^{-4} and a batch size of 32. During inference, the Mel-spectrogram of the target waveform is sampled using 25 Euler steps within the CFM-based decoder, with a guidance scale set to 1.0.

3.1.3. Evaluation Metric

To assess speech quality and emotion similarity of ClapFM-EVC, we perform subjective and objective evaluations. Regarding speech quality, a variety of objective metrics are used, including Mel-cepstral distortion (MCD), root mean squared error (RMSE), character error rate (CER), and predicted MOS (UTMOS). For emotion similarity, we employ a pretrained speech emotion recognition model[3] [33] to compute the cosine similarity of emotion embeddings between the converted and reference speech, referred to as the emotion embedding cosine similarity (EECS). The CER and UTMOS are calculated using pretrained CTC-based ASR[4] and MOS prediction[5] approaches. Besides, the subjective Mean Opinion Score (MOS) with a 95% confidence interval is used to measure naturalness (nMOS) and emotion similarity (eMOS). In practice, we invite 12 professional raters to participate in the evaluation. The scoring scale ranges from 1 to 5, with increments of 1, where higher scores indicate better performance. The audio samples are available online[6].

3.2. Main Results

3.2.1. EVC by Reference Speech

To evaluate the performance of ClapFM-EVC, we compare it with several existing EVC methods, i.e., StarGAN-EVC[7] [11], Seq2seq-EVC[8] [34], and MixEmo[9] [35]. Since these baselines

[3] https://github.com/ddlBoJack/emotion2vec
[4] https://huggingface.co/facebook/hubert-large-ls960-ft
[5] https://github.com/tarepan/SpeechMOS
[6] https://anonymous.4open.science/w/clapfm-evc-ACF1
[7] https://github.com/glam-imperial/EmotionalConversionStarGAN
[8] https://github.com/KunZhou9646/seq2seq-EVC
[9] https://github.com/KunZhou9646/Mixed_Emotions

Table 1: *Overall comparison results of the speech quality and emotion similarity between our proposed ClapFM-EVC system and other SOTA baseline methods using reference speech. nMOS and eMOS are presented with 95% confidence intervals.*

Model	MCD (↓)	RMSE (↓)	CER (↓)	UTMOS (↑)	nMOS (↑)	EECS (↑)	eMOS (↑)
StarGAN-EVC [11]	8.85	19.48	13.07	1.45	2.09 ± 0.12	0.49	1.97 ± 0.09
Seq2seq-EVC [34]	6.93	15.79	10.56	1.81	2.52 ± 0.11	0.54	2.23 ± 0.11
MixEmo [35]	6.28	13.84	8.93	2.09	2.98 ± 0.07	0.65	2.58 ± 0.13
ClapFM-EVC	**5.83**	**10.91**	**6.76**	**3.68**	$\mathbf{4.09 \pm 0.09}$	**0.82**	$\mathbf{3.85 \pm 0.06}$

employ the reference waveform to facilitate EVC, we first examine their performance using reference speech.

As evidenced in Table 1, our ClapFM-EVC consistently attains state-of-the-art performance in both speech quality and emotion similarity. With regard to emotion similarity, ClapFM-EVC exhibits significant advancements over baseline approaches, achieving notable relative improvements of at least 26.2% in EECS and 53.1% in eMOS, respectively. This indicates that our proposed ClapFM-EVC framework has a remarkable capability to precisely capture and effectively transfer the target emotional characteristics during emotional voice conversion. Regarding speech quality, the experimental results reveal that ClapFM-EVC exhibits superior results across multiple objective metrics. Detailed, ClapFM-EVC shows enhanced performance by achieving the lowest values in MCD, RMSE, and CER metrics, respectively. Moreover, subjective evaluation confirms that ClapFM-EVC gains the highest scores in both nMOS and UTMOS, with relative improvements of 37.2% and 49.2%, respectively, over the best-performing baseline method. These results underscore its exceptional capability in maintaining superior perceptual quality.

3.2.2. EVC by Natural Language Prompt

To compare the performance of ClapFM-EVC when using reference speech (Reference) versus natural language prompts (Prompt), we further conduct an ABX preference test.

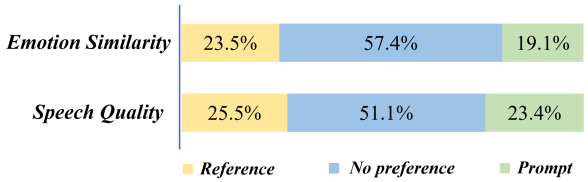

Figure 2: *The ABX preference test results compare the Reference with Prompt.*

As shown in Fig. 2, the first test aims to evaluate the emotional similarity between the converted speech driven by *Reference* and *Prompt*. 47 participants were asked to rate speech samples generated by *Prompt*, with *Reference* serving as the benchmark, on a scale from -1 to 1. Here, -1 indicates that the converted speech driven by *Reference* shows better emotion similarity, and 0 indicates no preference. The results revealed that 57.4% of participants selected "no preference," while 19.1% favored the "*Prompt*," suggesting that ClapFM-EVC can effectively control the emotional expression of converted speech via *Prompt*. In addition, we assess the quality of the converted speech relative to ground truth samples. Participants were required to choose the converted speech sample that is closer to ground truth in speech quality. As shown in the figure, the preference rates for speech driven by *Reference* and *Prompt* are 25.5% and 23.4%, showcasing that ClapFM-EVC is able to achieve high-quality EVC driven by *Prompt*.

3.3. Ablation Study

To evaluate the contributions and validity of each component of the proposed system, we conduct ablation studies. All results are summarized in Table 2.

Table 2: *Ablation results of the proposed ClapFM-EVC driven by natural language prompts. 'w/o emo label' denotes removing emotional categorical labels when training EVC-CLAP, 'w/o symKL' represents replacing symKL-loss with KL-Loss, 'w/o AIG' denotes removing the AIG module of AdaFM-VC.*

Model	**UTMOS** (↑)	**nMOS** (↑)	**EECS** (↑)	**eMOS** (↑)
ClapFM-EVC	**3.63**	$\mathbf{4.01 \pm 0.06}$	**0.79**	$\mathbf{3.72 \pm 0.08}$
w/o emo label	3.61	3.96 ± 0.11	0.66	3.01 ± 0.07
w/o symKL	3.57	3.89 ± 0.05	0.71	3.28 ± 0.08
w/o AIG	3.25	3.62 ± 0.12	0.74	3.52 ± 0.05

From the table above, we can easily reach the following conclusions: (1) In the absence of categorical emotional labels, the EECS and eMOS scores exhibit a significant decline, while the speech quality metrics remain largely unaffected. This demonstrates the effectiveness and rationality of the proposed soft-label-guided training strategy in our EVC-CLAP. (2) When training EVC-CLAP with KL-Loss, the EECS and eMOS values showed a relative performance drop of 10.1% and 11.8%, indicating that the proposed symKL-Loss can effectively enhance the emotion representation capability of EVC-CLAP. (3) The removal of the AIG module leads to a notable deterioration in speech quality and a slight performance reduction in emotional similarity. This underscores the critical role of the proposed AIG module in adaptively integrating content and emotional characteristics, thereby improving the overall performance of our proposed ClapFM-EVC system.

4. CONCLUSIONS

In this study, we propose ClapFM-EVC, an innovative and effective high-fidelity any-to-one EVC framework that features flexible and interpretable emotion control along with adjustable emotion intensity. Specifically, the proposed ClapFM-EVC initially employs EVC-CLAP to extract and align emotional elements across audio-text modalities. To enhance the emotional representational capacity, we utilize symKL-Loss to train the proposed EVC-CLAP model, guided by soft labels derived from the natural language prompts and their corresponding categorical emotion labels. Subsequently, we introduce AdaFM-VC based on the pre-trained ASR model, the proposed FuEncoder and CFM-based decoder to achieve high-fidelity and flexible emotion voice conversion. Extensive experiments indicate that our ClapFM-EVC is able to generate converted speech with precise emotion control and high speech quality driven by natural language prompts.

5. References

[1] K. Zhou, B. Sisman, C. Busso, and H. Li, "Mixed emotion modelling for emotional voice conversion," *computer*, vol. 6, p. 7, 2022.

[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] S. Chen, Y. Feng, L. He, T. He, W. He, Y. Hu, B. Lin, Y. Lin, Y. Pan, P. Tan *et al.*, "Takin: A cohort of superior quality zero-shot speech generation models," *arXiv preprint arXiv:2409.12139*, 2024.

[4] Z. Zhang, L. Li, G. Cong, H. Yin, Y. Gao, C. Yan, A. v. d. Hengel, and Y. Qi, "From speaker to dubber: movie dubbing with prosody and duration consistency learning," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 7523–7532.

[5] Y. Pan, Y. Yang, Y. Huang, J. Yao, J. Yin, Y. Hu, H. Lu, L. Ma, and J. Zhao, "Msac: Multiple speech attribute control method for reliable speech emotion recognition," *(No Title)*, 2023.

[6] Y. Pan, Y. Yang, H. Lu, L. Ma, and J. Zhao, "Gmp-atl: Gender-augmented multi-scale pseudo-label enhanced adaptive transfer learning for speech emotion recognition via hubert," *arXiv preprint arXiv:2405.02151*, 2024.

[7] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[8] Y. Yang, Y. Pan, J. Yin, and H. Lu, "Lmec: Learnable multiplicative absolute position embedding based conformer for speech recognition," *arXiv preprint arXiv:2212.02099*, 2022.

[9] H. Wang, S. Zheng, Y. Chen, L. Cheng, and Q. Chen, "Cam++: A fast and efficient network for speaker verification using context-aware masking," *arXiv preprint arXiv:2303.00332*, 2023.

[10] J. Yao, Y. Yang, Y. Lei, Z. Ning, Y. Hu, Y. Pan, J. Yin, H. Zhou, H. Lu, and L. Xie, "Promptvc: Flexible stylistic voice conversion in latent space driven by natural language prompts," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 571–10 575.

[11] 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.

[12] 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.

[13] R. Shankar, J. Sager, and A. Venkataraman, "Non-parallel emotion conversion using a deep-generative hybrid network and an adversarial pair discriminator," *arXiv preprint arXiv:2007.12932*, 2020.

[14] 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.

[15] W. Lu, X. Zhao, N. Guo, Y. Li, J. Wei, J. Tao, and J. Dang, "One-shot emotional voice conversion based on feature separation," *Speech Communication*, vol. 143, pp. 1–9, 2022.

[16] Y. Chen, L. Yang, Q. Chen, J.-H. Lai, and X. Xie, "Attention-based interactive disentangling network for instance-level emotional voice conversion," *arXiv preprint arXiv:2312.17508*, 2023.

[17] T. Qi, S. Wang, C. Lu, Y. Zhao, Y. Zong, and W. Zheng, "Towards realistic emotional voice conversion using controllable emotional intensity," *arXiv preprint arXiv:2407.14800*, 2024.

[18] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li, "Emotion intensity and its control for emotional voice conversion," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 31–48, 2022.

[19] H.-H. Chou, Y.-S. Lin, C.-C. Sung, Y. Tsao, and C.-C. Lee, "Toward any-to-any emotion voice conversion using disentangled diffusion framework," *arXiv preprint arXiv:2409.03636*, 2024.

[20] Y. Yang, Y. Pan, J. Yin, J. Han, L. Ma, and H. Lu, "Hybridformer: Improving squeezeformer with hybrid attention and nsr mechanism," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[21] J. Yao, Y. Yan, Y. Pan, Z. Ning, J. Ye, H. Zhou, and L. Xie, "Stablevc: Style controllable zero-shot voice conversion with conditional flow matching," *arXiv preprint arXiv:2412.04724*, 2024.

[22] Y. Pan, Y. Yang, J. Yao, J. Ye, H. Zhou, L. Ma, and J. Zhao, "Ctefm-vc: Zero-shot voice conversion based on content-aware timbre ensemble modeling and flow matching," *arXiv preprint arXiv:2411.02026*, 2024.

[23] S.-g. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "Bigvgan: A universal neural vocoder with large-scale training," *arXiv preprint arXiv:2206.04658*, 2022.

[24] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[25] Y. Pan, Y. Hu, Y. Yang, W. Fei, J. Yao, H. Lu, L. Ma, and J. Zhao, "Gemo-clap: Gender-attribute-enhanced contrastive language-audio pretraining for accurate speech emotion recognition," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 021–10 025.

[26] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[27] A. Conneau, "Unsupervised cross-lingual representation learning at scale," *arXiv preprint arXiv:1911.02116*, 2019.

[28] D. Min, D. B. Lee, E. Yang, and S. J. Hwang, "Meta-stylespeech: Multi-speaker adaptive text-to-speech generation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 7748–7759.

[29] 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.

[30] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[31] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[32] Y. Yang, Y. Pan, J. Yao, X. Zhang, J. Ye, H. Zhou, L. Xie, L. Ma, and J. Zhao, "Takin-vc: Zero-shot voice conversion via jointly hybrid content and memory-augmented context-aware timbre modeling," *arXiv preprint arXiv:2410.01350*, 2024.

[33] Z. Ma, Z. Zheng, J. Ye, J. Li, Z. Gao, S. Zhang, and X. Chen, "emotion2vec: Self-supervised pre-training for speech emotion representation," *arXiv preprint arXiv:2312.15185*, 2023.

[34] 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.

[35] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li, "Speech synthesis with mixed emotions," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 3120–3134, 2022.