Emotional Speech Anonymization: Preserving Emotion Characteristics in Pseudo-speaker Speech Generation

Hua Hua[1,2], *Zengqiang Shang*[1], *Xuyuan Li*[1,2], *Peiyang Shi*[1], *Chen Yang*[1,2], *Li Wang*[1,2], *Pengyuan Zhang*[1,2]

[1]Institute of Acoustics, Chinese Academy of Sciences, China
[2]University of the Chinese Academy of Sciences, China

huahua@hccl.ioa.ac.cn

Abstract

Speech anonymization plays an important part in protecting individuals' privacy in digital communication. However, ensuring anonymity while preserving the emotional and semantic integrity of speech poses significant challenges. This paper proposes a novel approach to address these challenges by integrating speaker-independent pre-trained emotion encoding into a fully end-to-end voice conversion model. By leveraging this approach, emotional information can be preserved and consistently represented in anonymized speech to a certain degree. This paper also introduces a new method that creates pseudo-speakers through model fusion to bypass the mismatch problem between the pseudo-speaker and the target emotion. Experimental results indicate that our methodology achieves a nuanced balance, maintaining an EER exceeding 30%, while effectively enabling accurate emotion recognition of nearly 50% and achieving a WER below 5% in the VPC evaluation.

Index Terms: speech generation, voice privacy, voice conversion, speech emotion

1. Introduction

Speech serves as a rich repository of speaker-specific information, encompassing aspects such as identity, age, gender, race, and even health status [1]. However, in scenarios where such identity information is misused, potential risks of crimes or societal issues may arise. Therefore, safeguarding voice privacy has emerged as a critical concern. Current research efforts not only focus on concealing the identity of the original speaker but also emphasize maintaining the integrity of semantic information to accurately convey the intended message [2, 3, 4, 5].

The Voice Privacy Challenge (VPC) 2024 underscores the task of preserving emotional nuances during the anonymization process. Achieving a balance among privacy protection, semantic consistency, and emotion preservation serves as the benchmark for evaluating anonymization effectiveness [1]. Central to this challenge is the need to effectively model emotions within anonymization systems and address potential mismatches between the speaker embedding of the original speech and the desired emotional expression in the anonymized output.

State-of-the-art speech anonymization methods, exemplified by [2, 4], excel in privacy protection by employing sophisticated techniques within the speaker embedding space to create "pseudo-speakers" that are maximally dissimilar to the source speaker. They transform speaker vectors to hide original speaker information without damaging diversity among different speakers. These pseudo-speakers, though derived and devoid of real speech data in most cases, play a pivotal role in enhancing anonymity. However, it is likely for this reason that maintaining sufficient emotional expressiveness in speech becomes particularly complex. The creation of pseudo-speakers represents inherent sparsity: while enhancing effectiveness for privacy protection, this approach may lead to a training-inference mismatch problem in emotional speech generation, potentially impairing the accurate conveyance of emotions. On the other hand, some speech anonymization methods, especially those with digital signal processing (DSP), can well preserve emotions. These methods are based on the assumption that speaker information is time-invariant and stable. Unlike neural network-based approaches that disentangle speaker information from emotional content unsupervisedly, DSP methods focus on modifying instantaneous speech characteristics such as pitch, spectral envelope, and time scaling to alter the timbre of the original speech, creating an impression of different speakers. Throughout this process, these methods prioritize the retention of emotional expressiveness by maintaining the consistency of acoustic features that encode emotional information, such as fundamental frequency contour, vowel energy, and speech rate. However, while DSP methods often succeed in preserving emotional nuances, their anonymization efficacy may not always meet stringent requirements. The generated speech may also sacrifice content fidelity, thereby compromising the accuracy of the conveyed message.

In this paper, we have conducted some explorations to try to alleviate this trade-off problem. We base our approach on a fully end-to-end voice conversion(VC) model, with which we achieve the purpose of hiding identities by replacing the source speaker with other real or pseudo speakers. In terms of modeling emotions, we use external or adaptive emotion encoding schemes to help train the acoustic model. Beyond voice conversion, We introduce a model fusion method to create pseudo speakers to perform privacy protection. We use the evaluation scripts of VPC 2024 to measure the privacy protection capabilities and content and emotion performance of our proposed system.

2. Related Work

2.1. VPC 2024 settings and evaluation metrics

VPC 2024 specifically addresses the subgoal of voice anonymization: altering a speaker's voice to conceal their identity as much as possible while preserving the content and emotional state intact [1]. This challenge is framed as a game between a user who shares data for a desired downstream task and an attacker who tries to identify the speaker using that data. Participants are tasked with developing systems that generate speech waveforms that masks the speaker's identity at the utterance level without distorting linguistic and emotional information. The effectiveness of these systems is measured using word error rate (WER) and unweighted average recall (UAR)

for utility, obtained from automatic speech recognition (ASR) and speech emotion recognition (SER) systems, respectively. The ability to protect privacy is assessed using the equal error rate (EER) from automatic speaker verification (ASV) systems. The results of the competition are determined in this way: At the same anonymization capability level (EER), the participating systems with lower WER and higher UAR are considered to have better balanced performance. The official training and test data are all based on open-source datasets. The datasets used to measure privacy and utility are the speech emotion dataset IEMOCAP and the multi-speaker speech dataset LibriSpeech.

2.2. Baseline systems

VPC 2024 evaluation plan has released 6 different baseline systems in total. Most of these baseline systems use HifiGAN or vocoder-like modules with similar structures and functions to generate anonymized speech. In addition, some methods use audio codecs for encoding and decoding, and methods that are completely based on DSP. The details of these methods and baseline performance can be found in the evaluation plan of VPC2024 [1]. Note that the approach of using DSP alone has the highest emotion retention ability (UAR↑), with its performance far exceeding that of other baselines, but comes at the cost of a higher chance of making more mistakes in content (WER↑) and weaker privacy protection capabilities (EER↓). On the contrary, the remaining neural network speech synthesis systems show essentially the opposite situation.

3. Methodology

3.1. Pipeline

Our system processes input speech and outputs anonymized speech results through a fusion model, illustrated in Figure 1. This fusion model consists of multiple fine-tuned, single-target speaker, fully end-to-end voice conversion (VC) models, integrated at the parameter level on each node. For each speech input, we apply a random fusion weight to ensure that the speaker information carried by each output speech remains relatively distinct. Without restrictions, we may generate some similar pseudo-speakers by accident though, this does not have any impact due to the huge amount of data.

3.2. VC base model

Our system, the end-to-end VC model, is demonstrated in Figure 2, which is modified from a model that is highly capable of deceiving anti-spoofing systems [6]. On training, a piece of input speech is encoded via both a priori and a posteriori encoding. The prior part contains three different encoders. The Yaapt [7] encoder is used to extract and encode F0 information, which will effectively represent the emotional attribute information in the acoustic model. The content encoder is an automatic speech recognition (ASR) model, through which we extract the features from a certain bottleneck layer. The speaker encoder is a simple one-hot embedding module. The three parts together constitute the prior latent space variables. The posterior part accepts linear spectrum input and generates the posterior latent space variables through the transformer structure [8]. Both the prior and posterior latent space variables are reparameterized and converted to the mean and variance of a Gaussian distribution. We introduce a KL loss to constrain their consistency. We use the gradient reversal method to blur the speaker information that may be carried in the content and base frequency encoding as much as possible. The waveform generation part adopts a transposed convolution upsampling structure similar to HifiGAN [9], and the network is constrained by generative/adversarial loss and reconstruction loss.

3.3. ASR bottleneck feature

We have experienced 2 different features in this regard. One is based on a hybrid CTC/transformer structure proposed by [10]. We extract the log probability of the outputs before the ASR decoder layer to represent the content information required by the acoustic model. This feature is called phoneme posteriorgrams(PPGs). PPGs are known to contain detailed content information and are frequently utilized in voice conversion tasks [11]. However, PPGs may also inadvertently capture and leak speaker-specific timbre and style information unique to each utterance. While preserving emotional expressiveness necessitates some consideration of style information, which can be beneficial, the leakage of speaker timbre poses a challenge for anonymization tasks. Since voice conversion training is fundamentally a self-reconstruction task, it becomes challenging to ascertain during training whether the generative model utilizes the residual timbre information in PPGs for reconstructing the speaker's voice. This potential leakage complicates the effort to ensure complete speaker anonymization. We also utilize the exact content features in baseline B5, an ASR-based model built on a pre-trained wav2vec2 model with three additional TDNN-F layers. In the final TDNN-F layer, we apply vector quantization (VQ) [1]. The incorporation of VQ into this framework aims to minimize the encoding of speaker information within the bottleneck (BN) features, thereby enhancing the disentanglement properties.

3.4. Emotion disentanglement and modeling

The key to the acoustic modeling stage involves the comprehensive utilization of both speech emotion and speech content information to predict acoustic features accurately. Emotional information can take various forms: discrete category labels (such as calm, happy, sad, angry, surprised), continuous representations based on dimensional models like the arousal-valence model [12], or emotion embedding extracted from a reference speech [13]. In the context of the anonymization task, where the target speaker may not exist, we must address the challenge of emotion disentanglement and transfer. This process involves disentangling emotion from the original input sentence, obtaining a representation that is independent of the speaker and speech content, and subsequently applying appropriate methods to integrate this emotional representation into the synthesized target speech.

Inspired by previous studies in emotion voice conversion [14][15] and speech emotion transfer [16][17], we enhance the emotion modeling capability via two approaches in this paper. The first is leveraging an wav2vec based emotion encoder pretrained on MSP-Podcast according to [18], which outputs emotion embeddings and we add them to the acoustic model as a priori encoding. The second is using a global style tokens (GST) encoder composed of multi-head attention modules [13]. GST is unsupervisedly trained out of the mel spectrum extracted from the input speech, and is often regarded as an extractor of emotional features in the field of emotional TTS. During training, explicit emotion labels along with orthogonal projection discriminant loss [19] are introduced to ensure that the encoding

[1] https://github.com/deep-privacy/SA-toolkit

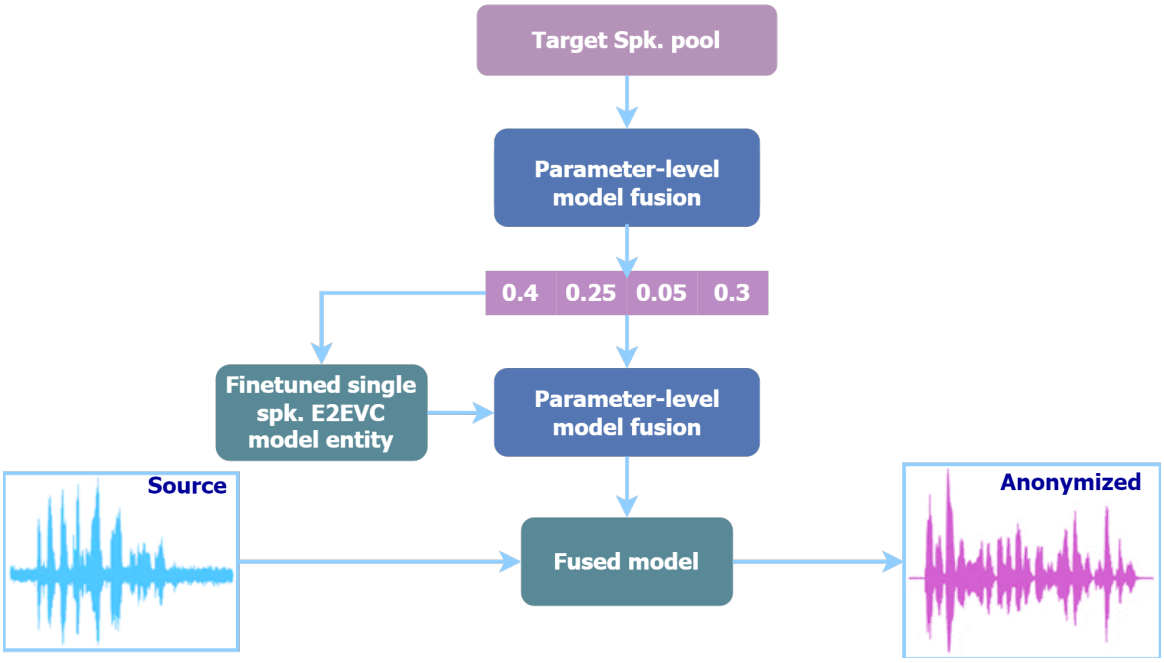

Figure 1: *System Anemone, total model pipeline*

has relatively strong emotion representation capabilities. Besides, it should be pointed out that the extracted F0 can also promote the effect of emotion generation.

3.5. Model fusion

In order to enhance the anonymization effect while minimizing the sacrifice on emotional preservation and content consistency, we aim for our pseudo-speaker construction process to avoid introducing training-inference mismatches. This approach ensures that the acoustic model and HifiGAN produce fewer "recognizable" pattern artifacts and maintain sound quality stability at the level perceptible to human hearing. Thus, we create pseudo-speakers by fusing model nodes at the parameter level. Assume that we have several end-to-end voice-conversion models that have been sufficiently fine-tuned with single-speaker data. After loading each model, we sum all the valid node weights involved in the inference process in the model network according to a certain ratio, except for those encoders used for information disentanglemnet, which do not participate in the fusing process. Then we regard the obtained weight set as a new, fine-tuned single-speaker model. When inferring with the new model, the generated speech timbre will sound like an "in-the-middle-person" among multiple fused speakers if the fusion weight is not a one-hot vector. These generated pseudo-speakers will make contributions in interfering with the ASV system. We believe that by doing so, we can lightly and effectively create a large number of pseudo-speakers even when there are only a few available high-quality speakers with sufficient data. Meanwhile, each pseudo-speaker is equipped with corresponding acoustic model inference weights, so that we can avoid the generated speech quality being poor due to sparse data. We have had similar experiences in common tasks such as TTS (text-to-speech) and SVS (singing voice synthesis). During the inference stage, the weights are randomly generated for each single utterance, and they sum to 1.

3.6. Training settings

This section shows our training details, including the pre-trained model, training data, training environment, etc. The table below shows the composition and training data of our various modules.

Table 1: *Modules and training corpora*

Modules	**Description**
Yaapt F0 encoder	baseline given
Hybrid transformer-CTC ASR	we pretrain on LibriTTS[20] train-clean-360 and train-other-500
Wav2vec emotion encoder	[18]
GST	attention-head=4, layer=6, out dim=192
HifiGAN	Resblock 1, upsampling 160x
Total training	LJ-speech[21], ESD[22], train-clean-360
Fine-tuning	LJ-speech, ESD

After the base model has been trained for 200 000 steps, it is fine-tuned for about 40 000 steps for each single speaker to make preparations for model fusion. GPU using: NVIDIA A100x4 for the base model and NVIDIA A100x1 for each fine-tuning task. The batch size is always 16 on a single GPU. The base model training takes about 12 hours, and the model fine-tuning for each speaker takes 1-2 hours. For model fusion, we fine-tune 11 different speakers and randomly select the models of 6 speakers as the source for fusion.

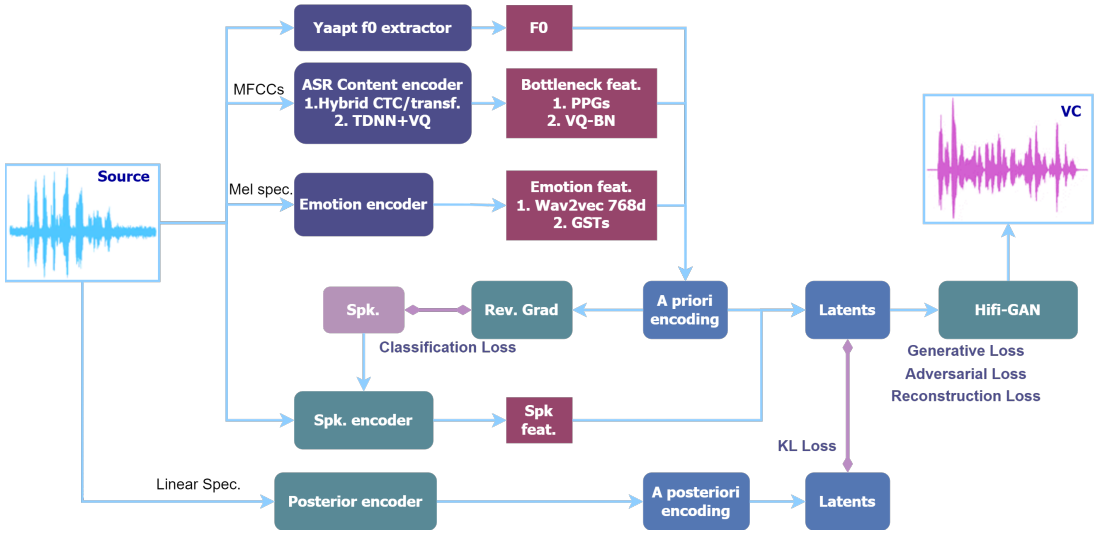

Figure 2: *System Anemone, end-to-end emotional voice conversion structure*

3.7. Training criteria

We use the following criteria to guide the model learning to convergence.

Acoustic model: *Spectral loss*

This term is used to ensure the performance of the self-reconstruction training process of the VC acoustic model. We consider L1 loss on mel-spectrograms.

Acoustic model: *Kullback-Leibler (KL) loss*

Both the prior and posterior latent space variables are reparameterized and converted to the mean and variance of a Gaussian distribution. KL loss is adopted to close their gap.

HifiGAN: *Reconstruction loss*

The constraint is that the generated speech should be as similar as possible to the input speech, which is a standard vocoder loss in self-reconstruction training process.

HifiGAN: *Generative/adversarial loss*

This part belongs to the basic loss of the GAN module, which is used to maintain generative adversarial training and jointly optimize the generator and discriminator.

Emotion encoder: *Cross Entropy (CE) or Orthogonal Projection (OP) loss*

This term is used to constrain the results of the emotion encoder to re-predict the emotion type so that the emotion encoding can better express the clustering characteristics.

Speaker rev. grad module: *Cross Entropy (CE) loss*

To further eliminate residual speaker information in the emotion encoding, we connect a classification module from the gradient inversion layer to the speaker encoder following the encoding layer. Since the training objectives on both sides of the inversion layer are opposed, higher classification accuracy indicates that the emotion encoding before the inversion layer contains less speaker information. This approach thus enhances disentanglement capability.

Weights for different losses:

$$L_{total} = L_{gen} + L_{spk} + L_{emo} + 45 * L_{spec} + 500 * L_{kl}$$

4. Results and discussion

We name our systems as in the following Table 2. Note that we only submit *ppg-w2vF0-fusion* and *ppg-GSTF0-fusion* to the VPC 2024 for ranking purposes. Others predominantly function as modules for ablation analysis.

Table 2: *Systems involved in the experiment*

System	content encoder	emotion encoder	model fusion
ppg-w2vF0-fusion	ASR-PPGs	w2v + F0	Yes
ppg-GSTF0-fusion	ASR-PPGs	GST + F0	Yes
ppg-F0-fusion	ASR-PPGs	F0 only	Yes
ppg-w2vF0	ASR-PPGs	w2v + F0	No
vqbn-w2vF0-fusion	ASR-VQ	w2v + F0	Yes

Tables 3, 4, and 5 show how our systems performed in evaluations.

It is demonstrated in the table above, that in terms of emotion preservation performance, our methods outperform other baselines except for B2. This outcome is expected, considering that the B2 baseline employs a purely DSP approach, resulting in comparatively minor modifications of the original input samples with regard to emotional features. In cases where w2v emotion encoder is not employed, the choice between PPGs or VQ-BN as content embedding features seems to appear inconsequential, as the performance of *ppg-F0-fusion* aligns closely with that of baseline B5 (B5 is roughly vqbn + F0, with a lightweighted acoustic model). Note that this comparison may be somewhat inadequate, because we have no way of knowing the specific training details of the ASR part of B5, and the data used for training also has some deviations from the data we use to train the ASR-PPG features (the overall distribution is consistent, but the specific content is not completely consistent).

In particular, we find that in the emotion test results, except for the very poor results of the sad category, the rest are comparable to the original data. We believe that this may be because the sad emotion in the 'act-out-style' training data (ESD) is more performative, while the sadness in the test data IEMO-

Table 3: *Emotion: SER (%) evaluation results*

System	ave.↑	IEMOCAP-dev↑	IEMOCAP-test↑
ppg-w2vF0-fusion	**49.96**	48.70	51.22
ppg-GSTF0-fusion	43.28	42.38	44.18
ppg-F0-fusion	38.61	37.67	39.54
ppg-w2vF0	47.36	46.10	48.62
vqbn-w2vF0-fusion	41.00	40.91	42.10
Origin	70.07	69.08	71.06
Given B1	42.74	42.71	42.78
Given B2	**54.55**	55.61	53.49
Given B3	37.83	38.09	37.57
Given B4	42.38	41.97	42.78
Given B5	38.13	38.08	38.17
Given B6	36.26	36.39	36.13

CAP [23] is mostly reflected in very low volumes, sometimes even submerged by noise, resulting in its emotion features not being significant enough. In addition, some emotional features may be reflected in supra-segmental rhythm patterns. The systems we propose attempt to model emotions at the whole utterance level, which may lack consideration in this regard.

Table 4: *Content: WER (%) evaluation results*

System	ave.↓	libri-dev↓	libri-test↓
ppg-w2vF0-fusion	4.72	4.79	4.65
ppg-GSTF0-fusion	**4.51**	4.60	4.43
ppg-F0-fusion	**4.50**	4.32	4.67
ppg-w2vF0	4.69	4.78	4.60
vqbn-w2vF0-fusion	5.19	5.25	5.13
Origin	1.82	1.80	1.85
Given B1	**2.99**	3.07	2.91
Given B2	10.20	10.44	9.95
Given B3	4.32	4.29	4.35
Given B4	6.02	6.15	5.90
Given B5	4.55	4.73	4.37
Given B6	9.39	9.69	9.09

Our proposed method is at an intermediate level between the results of various baselines considering WER, which results are shown in Table 4. Upon examining baseline 1, we observe its utilization of ASR bottleneck layer features for content encoding. Despite this, the incorporation of x-vector by the speaker mitigates pressure on the content encoder, enhancing content consistency and thereby reducing WER. Conversely, baseline 3's advantage lies in explicit phoneme decoding for control purposes. Despite this, we acknowledge that a slight increase in WER may be unavoidable and acceptable, particularly with the introduction of emotional encoding schemes like w2v or GST, which can potentially distort content due to conflicting interactions with emotional information. This parallels everyday communication scenarios where strong emotions can compromise content intelligibility and sometimes even make a little change to pronunciation, and such phenomenon is more obvious in tonal languages such as Chinese and Thai.

Table 5: *Privacy: EER (%) evaluation results*

System	ave.↑	libri-dev-f↑	libri-dev-m↑	libri-test-f↑	libri-test-m↑
ppg-w2vF0-fusion	**31.55**	32.53	29.50	33.40	30.76
ppg-GSTF0-fusion	18.58	22.98	15.37	19.68	16.29
ppg-F0-fusion	25.73	25.33	23.05	30.10	24.43
ppg-w2vF0	23.25	24.60	22.58	26.67	19.14
vqbn-w2vF0-fusion	27.86	32.41	19.08	29.90	30.03
Origin	5.16	10.51	0.93	8.76	0.42
Given B1	7.64	10.94	7.45	7.47	4.68
Given B2	6.00	12.91	2.05	7.48	1.56
Given B3	26.28	28.43	22.04	27.93	26.72
Given B4	31.48	34.37	31.06	29.37	31.16
Given B5	**34.36**	35.82	32.92	33.95	34.73
Given B6	22.10	25.14	20.96	21.15	21.14

We conclude from Table 5 that *ppg-w2vF0-fusion* can achieve a good privacy-protection ability with an EER result of more than 30%, and if without model fusion, the EER performance drops by nearly 8%. This basically indicates that using model fusion methods to increase the sparsity of the speaker space can enhance privacy protection capabilities. We are surprised to find that simply by replacing the emotion features of w2v with GST, EER significantly decreases. We speculate that the style information of GST may contain too many speaker characteristics that are difficult to filter out, and this part actually makes the clustering effect of different utterances in the speaker space better after model fusion, or in other words, the utterances of the same source speaker can find more commonalities.

5. Conclusion

In this paper, we explore the techniques of speech anonymization while preserving emotional fidelity, and propose a method based on an integrated approach of end-to-end voice conversion (VC) models and emotion disentanglement. Our method leverages speaker-independent emotion encoding and model fusion strategies to achieve a balance between privacy protection and emotional preservation in anonymized speech. Through extensive evaluations on the VPC 2024 dataset, we demonstrate that our proposed systems, particularly those integrating emotion encoders and employing model fusion, outperform those baseline methods. Notably, our approach also maintains acceptable word error rates (WER) in automatic speech recognition (ASR) tasks, indicating its effectiveness in balancing privacy preservation with semantic and emotional integrity.

However, our approach still exhibits several limitations. For instance, we encounter difficulty in elucidating the inner workings of the 'black-box' neural network, so we cannot mathemat-

ical and rigorously substantiate whether the model fusion technique really disrupts the discrimination ability of ASV systems. Moreover, the granularity of our emotional modeling remains inadequate, leaving room for enhancement in emotional transfer effectiveness.

6. References

[1] N. Tomashenko, X. Miao, P. Champion, S. Meyer, X. Wang, E. Vincent, M. Panariello, N. Evans, J. Yamagishi, and M. Todisco, "The VoicePrivacy 2024 challenge evaluation plan," 2024.

[2] X. Miao, X. Wang, E. Cooper, J. Yamagishi, and N. Tomashenko, "Speaker anonymization using orthogonal householder neural network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[3] X. Chen, G. Li, H. Huang, W. Zhou, S. Li, Y. Cao, and Y. Zhao, "System description for voice privacy challenge 2022," in *Proc. 2nd Symposium on Security and Privacy in Speech Communication*, 2022.

[4] S. Meyer, P. Tilli, P. Denisov, F. Lux, J. Koch, and N. T. Vu, "Anonymizing speech with generative adversarial networks to preserve speaker privacy," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 912–919.

[5] J. Yao, Q. Wang, L. Zhang, P. Guo, Y. Liang, and L. Xie, "Nwpu-aslp system for the voiceprivacy 2022 challenge," *arXiv preprint arXiv:2209.11969*, 2022.

[6] H. Hua, Z. Chen, Y. Zhang, M. Li, and P. Zhang, "Improving spoofing capability for end-to-end any-to-many voice conversion," in *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*, 2022, pp. 93–100.

[7] K. Kasi, "Yet another algorithm for pitch tracking:(yaapt)," Ph.D. dissertation, Citeseer, 2002.

[8] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 5530–5540.

[9] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 17 022–17 033.

[10] H. Miao, G. Cheng, C. Gao, P. Zhang, and Y. Yan, "Transformer-based online ctc/attention end-to-end speech recognition architecture," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6084–6088.

[11] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *2016 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2016, pp. 1–6.

[12] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, 2011.

[13] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International conference on machine learning*. PMLR, 2018, pp. 5180–5189.

[14] 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.

[15] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and esd," *Speech Communication*, vol. 137, pp. 1–18, 2022.

[16] T. Li, S. Yang, L. Xue, and L. Xie, "Controllable emotion transfer for end-to-end speech synthesis," in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2021, pp. 1–5.

[17] Y. Lei, S. Yang, X. Wang, and L. Xie, "Msemotts: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 853–864, 2022.

[18] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–13, 2023.

[19] K. Ranasinghe, M. Naseer, M. Hayat, S. Khan, and F. S. Khan, "Orthogonal projection loss," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12 333–12 343.

[20] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "Libritts: A corpus derived from librispeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.

[21] K. Ito and L. Johnson, "The lj speech dataset," 2017.

[22] 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.

[23] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.