Aclipse: Attention-based Cascaded Learning Enabling Privacy-preserving Speech Emotion Recognition

Jiusong Luo, Shan Chang*, Luo Zhou, Shizong Yan
Department of Computer Science and Technology, Donghua University, Shanghai, China
{1249833, zhouluo, 1222040}@mail.dhu.edu.cn, changshan@dhu.edu.cn

***Abstract*—Speech Emotion Recognition (SER), as a core technology in intelligent interaction and affective computing, faces critical privacy security challenges. Speech not only contains emotional information, but also encompasses sensitive data such as conversational content and speaker identity markers. The sensitive information could be potentially reconstructed through Automatic Speech Recognition (ASR), posing substantial privacy leakage risks. Existing privacy protection methods focus mainly on suppressing the leakage of demographic characteristics such as speaker identity, while there is a lack of systematic research on the speech content. The traditional SER and ASR joint training paradigm relies on the sequence alignment mechanism of ASR, which leads to deep coupling of emotion features with the corresponding text, causing a fundamental contradiction between "performance improvement" and "privacy protection". To address this problem, this paper proposes an attention-based cascaded learning enabling privacy-preserving (Aclipse) strategy, which is divided into two stages. In the first stage, the emotion-retaining (ER) module and semantics-obfuscating (SO) module in cascaded manner, which consist of the channel-level attention mechanism, guides the targeted injection of adversarial perturbation by identifying and strengthening high-frequency feature channels that are critical to ASR performance, thereby significantly improving the semantic obfuscation to achieve speech privacy protection while minimizing the negative impact on SER performance. In the second stage, we freeze the SO module parameters to maintain its interference effect on ASR, and fine-tune the ER to improve the performance of SER by adjusting the channel-level attention mechanism to focus on emotion-related information. The experimental results on the IEMOCAP dataset show that Aclipse achieves an effective balance between performance and privacy protection, providing a new idea to build a trustworthy SER system.**

***Index Terms*—speech emotion recognition, privacy-preserving, attention, cascaded learning, semantic obfuscation**

I. Introduction

Speech Emotion Recognition (SER), as a pivotal branch of computational phonetics, aims to analyze human emotional states from speech signals and and has broad application prospects in domains such as intelligent interaction, healthcare, and affective computing [1], [2], [3], [4], [5]. However, in practical deployments, SER faces critical privacy security challenges: raw speech signals, which serve as carriers of multimodal information, inherently encapsulate not only emotional cues but also sensitive semantic elements, including identity identifiers, conversational content, and health indicators [6]. These elements could be reverse-engineered via Automatic Speech Recognition (ASR), forming covert privacy leakage channels [7]. Therefore, while ensuring the performance of emotion recognition, applying directional distortions to the original speech signal to block the leakage of sensitive content is the core difficulty in building a trustworthy SER system, as illustrated in Figure 1.

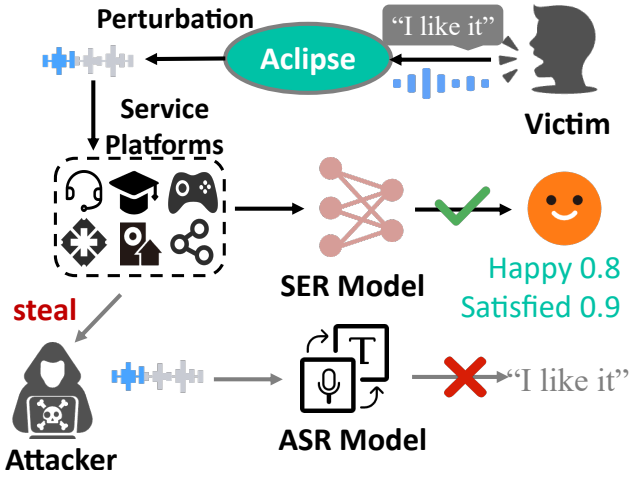

Fig. 1: Aclipse adds adversarial perturbation to the original audio for protecting the privacy of the speech content while still effectively recognize emotion.

Currently, privacy-preserving methods related in SER predominantly follow two pathways: model-level and data-level methods. Model-level methods [8], [9], [10] usually introduce differential privacy mechanism to inject gradient perturbation during SER model training to obfuscate sensitive information. On the other hand, data-level methods [11] [12], [13], [14] employ data transformations or noise injection on raw speech to eliminate privacy-revealing features. However, existing research primarily focus on suppressing the leakage of demographic characteristics such as speaker identity, while the intrinsic privacy protection of speech content remains insufficiently explored.

Inspired by multi-task learning in the protection of gender

*Corresponding Author

inference [15], [16], this paper explores the issue of SER privacy protection in joint training of SER and ASR. The traditional SER and ASR joint training paradigm usually relies on the sequence alignment capability of ASR (Connectionist Temporal Classification) to improve the performance of emotion recognition by finely aligning speech frames with text [17]. However, this deep coupling inherently causes SER models to encode speech content information while learning emotional characteristics, resulting in a fundamental conflict between performance enhancement and privacy preservation. The direct adoption of traditional multi-task learning frameworks (e.g., task negative optimization) to suppress content inference may destroy the ASR-assisted temporal alignment mechanism, which leads to a significant degradation of emotion recognition performance; conversely, if ASR alignment is fully relied upon, it is prone to retaining sensitive information in the speech content, which threatens the user's privacy.

To address this fundamental conflict, this paper proposes a multi-objective optimization-based phased training strategy, attention-based cascaded learning enabling privacy-preserving (Aclipse), aiming to improve the performance of emotion recognition while effectively suppressing speech content leakage. The framework comprises three core modules: an emotion-retaining (ER) module, a semantics-obfuscating (SO) module, and the pre-trained SER and ASR models for evaluation. The ER module and the SO module are composed of channel attention mechanism. In the first stage, the channel-level attention mechanism locates and enhances the high-frequency feature channels that are critical to the ASR task, accurately guides the injection of adversarial perturbation, thereby significantly protect speech privacy while reducing interference with SER performance. In the second stage, we freeze the SO module parameters to ensure its interference ability for ASR, and optimize the ER module to enhance the performance of SER in noisy environments by adjust the channel-level attention mechanism to capture emotional information. The main contributions are three-fold:

- We empirically reveal the privacy risks in SER and demonstrate that semantic information can be protected without compromising recognition performance.
- We propose attention-based cascaded learning enabling privacy-preserving (Aclipse), a novel multi-objective optimization phased training strategy. This method separates the optimization process of perturbation generation and emotion recognition, and achieves the focus on a single goal in each stage, thereby effectively avoiding the gradient conflict problem in multi-task learning.
- Experimental results on the IEMOCAP dataset show that the first stage improves WER by 53% while WA only decreases by 11%. Subsequently, the second stage only fine-tunes the ER module, and while ensuring that WA is basically stable, WER is further improved by 5.5%, thus achieving an effective balance between performance and privacy protection.

II. Related Work

Speech, as a composite carrier of multi-modal information, contains both emotional cues and sensitive content such as user identity and health-related data. Attackers can leverage existing technologies to infer sensitive attributes (e.g., home address, age) from speech signals, causing to severe privacy breaches [18], [19], [20]. Thus, a critical challenge for secure SER deployment lies in enhancing emotion recognition performance while effectively blocking sensitive information leakage [21], [22]. Current research primarily falls into two categories: model-level and data-level methods.

Model-level methods aim to prevent sensitive information leakage by constraining parameter updates or gradient transmission. For instance, Feng et al. [8] proposed a user-level differential privacy (UDP) method, which adds Gaussian noise to client model gradients in federated learning to achieve local differential privacy (LDP), effectively reducing gender inference accuracy to the random level. However, its protection capability degrades significantly as attackers observe more model updates. To address this, Mohammadi et al. [9] proposed a hierarchical privacy protection (GHDP) method that normalizes gradient importance and injects more noise into key layers, maintaining stable privacy preservation across multiple update rounds without significant degradation in emotion recognition performance. Additionally, Chen et al. [10] proposed an LDP method combined with a client selection strategy (LDP-FL with CSS), which reduces the negative impact on model performance by screening representative client updates, thereby retaining model performance while protecting privacy.

Data-level methods enhance privacy protection and maintain emotion recognition performance through feature transformation, noise injection, or adversarial training on speech data [23], [24]. Feng et al. [11] proposed the Cloak framework that combines adversarial training with noise injection to suppress gender attribute leakage while preserving emotion-related features by jointly optimising via joint optimization of noise functions and adversarial classifiers. Across multiple datasets, this method reduces the gender classification accuracy by more than 30%, while the emotion recognition accuracy only loses 2-5%. Similarly, Tan et al. [13] proposed a gender indistinguishability (Gender-Ind) method, which perturbs gender embedding vectors in speech features, achieving a better privacy-utility balance than traditional differential privacy methods (e.g., Voice-Ind) and improving emotion recognition accuracy by 8% under identical privacy budgets. Feng et al. [14] proposed a sensitive information neutralization conversion technology based on variational autoencoder, which decomposes speech data into sensitive attributes and emotional semantics to reduce correlations between transformed data and original sensitive information for privacy masking.

In addition, Ali et al. [15] and Zhao et al. [16] adopted multi-task learning strategies to balance privacy protection and emotion recognition performance. Ali et al. [15] proposed a dynamic adversarial training framework that simultaneously

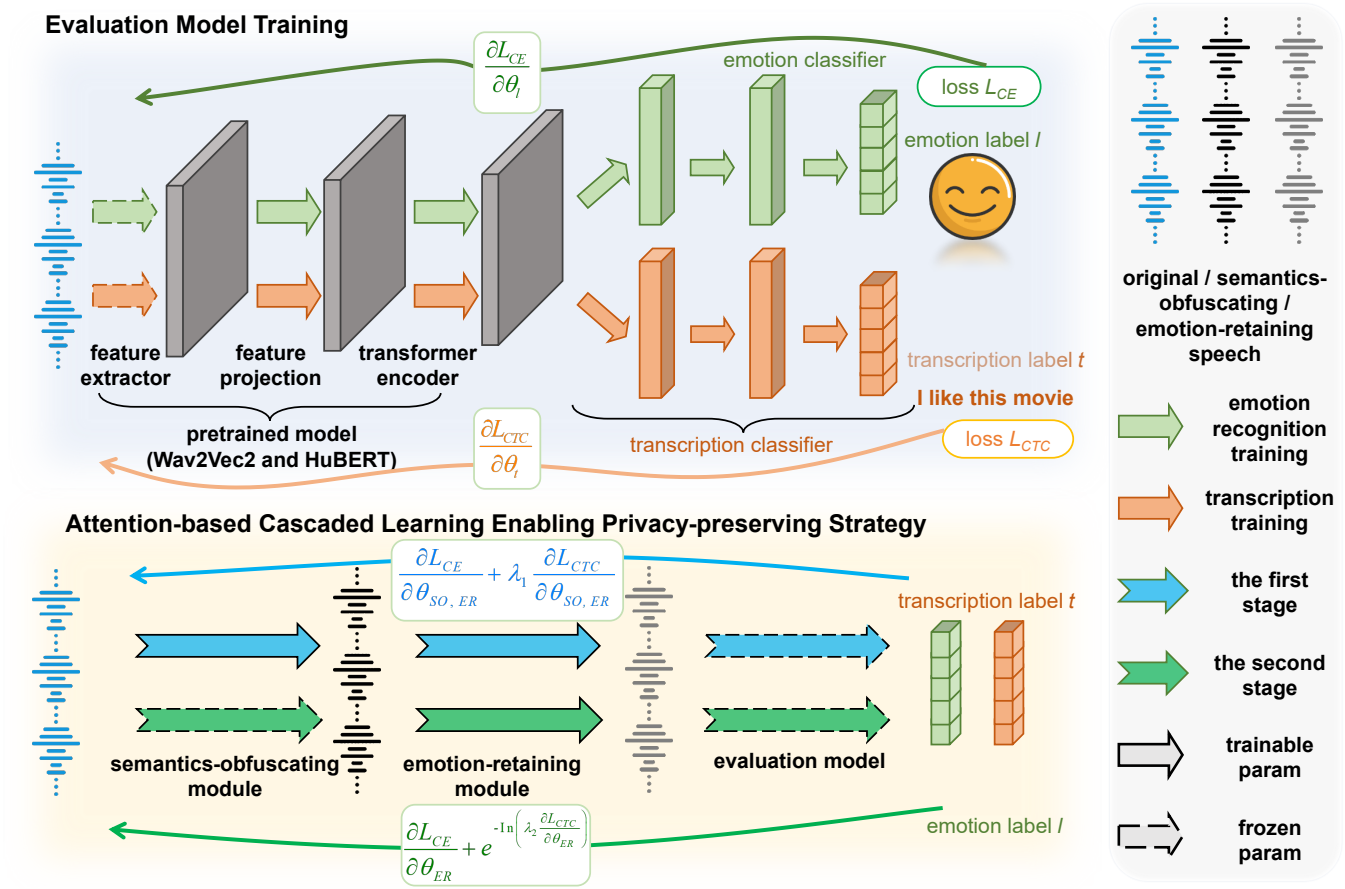

Fig. 2: Overview of Aclipse.

optimizes the emotion recognition model and adversarial classifier, using a gradient reversal layer to minimize adversarial classifier performance and remove sensitive information from features. Zhao et al. [16] introduced gender as an auxiliary task in emotion recognition via multi-task learning, reducing gender leakage through shared representations.

However, these multi-task learning approaches fail to fully account for conflicts between divergent optimization objectives. In SER tasks, deep coupling between speech frames and content may cause mutual interference between emotion recognition and privacy protection, ultimately affecting model convergence and performance.

III. Threat Model and System Goal

In the current social platform environment, voice interaction has become an important way for users to express their emotions. However, the platform's collection and storage mechanism of raw audio data brings potential privacy leakage risks. Upon access to this data by unauthorised third party, attackers can transcribe the speech content into text through ASR to extract sensitive information of users, such as identity, health status, etc. This threat not only violates user privacy, but may also trigger more serious social engineering attacks. Therefore, how to maintain the effectiveness of emotion recognition while protecting semantic privacy has become a key issue that needs to be solved urgently.

To address the above challenges, we proposed an innovative method called Aclipse, which achieves dual optimization of speech privacy protection and emotion recognition by injecting carefully designed adversarial perturbations into the original speech. Specifically, the added perturbations are designed to significantly improve the semantics obfuscating of the ASR system, making it difficult for potential threat models to recover semantic information from the audio. Meanwhile, the impact of these perturbations on SER is kept to a minimum, ensuring that the social platform can still accurately identify the user's emotional state. Consequently, Aclipse achieves an efficient balance between privacy protection and emotion recognition, providing a new solution for speech security.

IV. Design of Aclipse

A. Overview

Our core idea is to propose a training strategy to solve the problem of conflicting goals in multi-task learning, thereby achieving dual optimization of speech emotion recognition performance and privacy protection. As shown in Figure 2, it consists of two parts: evaluation model training and attention-based cascaded learning enabling privacy-preserving strategy.

First, for SER and ASR tasks, the pre-trained model is fine-tuned on the target dataset and its parameters are frozen to build an evaluation benchmark. Secondly, the emotion-

retaining (ER) module and semantics-obfuscating (SO) module are designed for adding perturbation to the original speech through attention-based cascaded learning enabling privacy-preserving strategy. The perturbad speech is fed into the evaluation model to simultaneously evaluate WA and WER.

Given a dataset D with k utterances u_i , the corresponding emotion labels are l_i. Each utterance consists of audio segment a_i and the corresponding transcription t_i, $u_i \in (a_i, t_i)$. The goal of the study is to maximize WER while maintaining WA, thereby achieving privacy protection.

B. Evaluation Model Training

Self-supervised learning (SSL)-based speech representation frameworks (such as Wav2Vec2 [25] and HuBERT [26]) efficiently capture common speech structure features from large-scale unlabeled speech data through discriminative pre-training or multi-task learning objectives. The universal representations generated by these pre-trained models show significant transfer potential in various speech recognition tasks [27], [28]. For SER and ASR tasks, this study performs end-to-end fine-tuning on the target dataset using pre-trained models.

$$e_i^a = \Phi^a(a_i) \in \mathbb{R}^{D_W}; i \in [1, k] \tag{1}$$

$$e_i^t = \Phi^t(a_i) \in \mathbb{R}^{d \times D_R}; i \in [1, k] \tag{2}$$

$$L_{CE} = CrossEntropy(e_i^a, l_i) \tag{3}$$

$$L_{CTC} = CTC\left(e_i^t, t_i\right) \tag{4}$$

where, Φ^a and Φ^t are the evaluation models of SER and ASR, respectively, which include pretrained models, fully connected layers, softmax and other operations, and are fine-tuned by CrossEntropy (CE) loss and connectionist temporal classification (CTC) loss, respectively. D_W and D_R are the number of emotion and character, respectively, and d is the number of audio frames.

After fine-tuning, the evaluation model parameters are frozen for evaluation in the subsequent two-stage tasks.

C. Attention-based Cascaded Learning Enabling Privacy-preserving Strategy

In this paper, we propose an attention-based cascaded learning enabling privacy-preserving strategy, and design ER and SO modules to generate perturbed speech in cascade manner. Both modules are composed of multiple one-dimensional convolution layers, dropout layers, channel-level attention mechanism, and ReLU activation functions. In our training strategy, the channel-level attention mechanism plays a key and differentiated role at different stages. Through flexible adjustment in two stages, the model achieve the best balance between perturbation injection and emotion feature recovery.

1) The first stage: We jointly train ER and SO to generate adversarial perturbation, which effectively improves the semantic obfuscation of the ASR system while keeping the negative impact on emotion recognition within an acceptable range. The channel-level attention mechanism is mainly used to identify and strengthen high-frequency feature channels that are critical to ASR performance. These channels often contain key information such as consonant energy distribution and vowel formant structure. By applying larger weights to these feature channels, the network can guide the injection of adversarial perturbation in a targeted manner, making the perturbation more effective in interfering with ASR while minimizing the negative impact on SER performance.

$$a_i^{perturbation} = ER(SO(a_i)) \tag{5}$$

Subsequently, the generated perturbation audios are fed into the pre-frozen evaluation model to validate the perturbation generation effect, respectively.

In this stage, we design the loss function L_1, which not only considers the contribution of perturbation to semantic protection, but also measures its destructive effect on emotion recognition performance as follows:

$$L_1 = L_{CE} + \lambda_1 L_{CTC} \tag{6}$$

where λ_1 is the weight coefficient, which is set to -1 to balance the strength of the two objectives to prioritize WRE. After the first stage of training, both ER and SO have the ability to generate effective perturbation for the ASR system, while the negative impact on the SER task is effectively suppressed.

2) The second stage : The parameters of the SO module are frozen to maintain its stable perturbation generation ability, and only the ER module is fine-tuned to focus on optimizing the emotion recognition performance. The channel-level attention mechanism is readjusted, with its focus shifted to identifying and strengthening information such as low-frequency fundamental frequency, prosodic features, and dynamic changes in intonation that are closely related to emotion recognition. This strategy is in line with the design idea of "protection first, then recognition", ensuring that privacy protection (semantic obfuscating) is achieved first, and then compensating for the adverse effects that may be caused to emotion recognition on this basis. If the opposite strategy is adopted, emotion retain is performed first, and then semantic obfuscation is performed, the latter may destroy the previously recovered emotion information in the subsequent processing process. L_2 focus on the performance of the SER as follows:

$$L_2 = L_{CE} + e^{-\ln(\lambda_2 L_{CTC})} \tag{7}$$

where λ_2 is the weight coefficient and is set to 1.

The attention-based cascaded learning enabling privacy-preserving strategy achieves effective decoupling of tasks: the first stage focus on generating adversarial perturbation to interfere with the ASR system; the second stage reduces the negative impact of perturbation on SER performance by

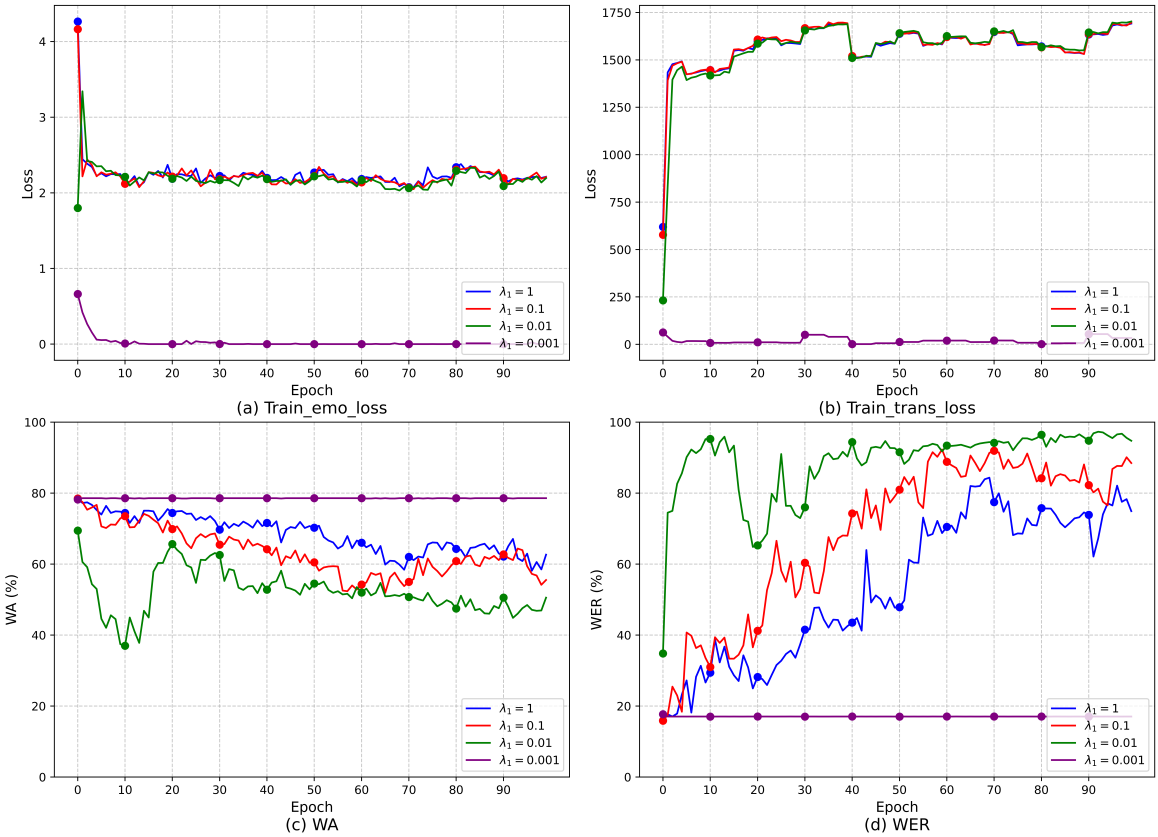

Fig. 3: The effect of the ASR task weight coefficient λ_1 on the protection semantics in the first stage.

optimizing the ER module while ensuring the perturbation generation capability. Compared with the traditional joint optimization strategy, Aclipse effectively avoids the problem of target conflict. In particular, when perturbation generation and emotion recognition are optimised simultaneously, the optimization directions may interfere with each other, thus affecting the convergence and overall performance of the model. Aclipse decouples the training process of perturbation generation and emotion recognition, allowing each stage to focus on a single goal, thereby reducing gradient conflicts and improving optimization efficiency.

V. Evaluation

A. Datasets

The Interactive Emotion Dynamic Motion Capture (IEMOCAP) dataset [29] is a benchmark dataset for the speech emotion recognition field. It comprises 12 hours of improvised and scripted audiovisual data from 10 UC theater actors (5 male and 5 female) divided into five binary sessions. Each session provides emotional information in four ways: video, audio, text transcription, and facial motion capture. The four most common emotion labels select for this study are as follows: neutral, sad, happy (combined with excited), and anger. Following the previous experimental setups [30], we employ a five-fold cross-validation, where 80% of the data are used for training and 20% for testing.

B. Experiment Setup

In this study, we select two pre-trained models of Transformer architectures that are widely used in the field of speech recognition, Wav2Vec2 and HuBERT, including two architecture variants: Base and Large.

All experiments are implemented based on the PyTorch framework and executed on a single NVIDIA GeForce RTX 3090 card, with batch size set to 2. During the fine-tuning process, the Adam optimizer is utilised, with the learning rate set to 1e-5 and the training period fixed at 50 epochs. The emotion recognition task utilises the Cross Entropy (CE) loss function, while the automatic speech recognition task employs the Connectionist Temporal Classification (CTC) loss function. In the attention-based cascaded learning enabling privacy-preserving strategy, the learning rates for the 'base' model are set to 1e-3 in the first stage and 1e-5 in the second stage, whereas for the 'large' model, the learning rates are 2e-4 and 1e-5, respectively. To mitigate the risk of overfitting, a stratified sampling strategy is employed in the attention-based cascaded learning enabling privacy-preserving strategy, selecting 400 samples as a training subset. The training spans a total of 100 epochs, with the training subset being replaced every five epochs. To limit the effect of randomness, each experimental configuration is repeated independently five times and the final results are presented as mean values.

The performance of Aclipse is evaluated with Weighted Ac-

curacy (WA) for the SER task and Word Error Rate (WER) for the ASR task to measure the protection strength of semantic privacy. We have standardized the results to quantify optimal semantic protection while maintaining the baseline emotion recognition performance.

C. Results

1) Overall Performance: We have systematically evaluated the protection effectiveness of Aclipse on different pre-trained models and compared it with the baseline evaluation model, shown in Table I. Taking the Wav2Vec2-base model as an example, after end-to-end fine-tuning on the IEMOCAP dataset, it achieves a WA of 78.57% and a WER of 15.9%, both exhibiting good performance. However, following this, the WER significantly increases by about 41%, while the WA decreases by less than 8% only. It demonstrates that Aclipse can effectively enhance semantic privacy preservation while maintaining the stability of emotion recognition as much as possible. In contrast, the large model has a stronger learning ability and is still able to extract key semantic information from the perturbed audio under the same perturbation condition. However, since the stronger learning ability may weaken the privacy protection effect, its overall performance is inferior to the base model, but still achieves a relatively ideal result in privacy protection.

TABLE I: Comparison of protect performance on different pretrained models.

Pretrained Model	Method	WA (%)	WER (%)
Wav2vec2-base	Baseline	78.57	15.90
	Our	70.25	56.57
Wav2vec2-large	Baseline	80.01	15.31
	Our	68.53	56.02
Hubert-base	Baseline	79.75	17.12
	Our	70.43	54.69
Huebrt-large	Baseline	81.56	16.26
	Our	69.98	53.21

2) Ablation Study: In the ablation experiments, we use only the experimental results of the wav2vec2-base pre-trained model. Notably, we have processed the results to ensure that the same speech emotion recognition performance is preserved and observe its optimal semantic preservation.

We investigate the effect of the ASR task weight coefficient λ_1 on the protection semantics in the first stage. The trends of SER loss (`Train_emo_loss`), ASR loss (`Train_emo_loss`), WA and WER during training when λ_1 is 0.001, 0.01, 0.1, and 1 are shown in Figure 3. Experimental results show that the trend of WER of WA is consistent when λ_1 is 0.01, 0.1 and 1. The reason is that during the training process, the loss value $Ł_{CTC}$ is always maintained at about 1500, and the loss value $Ł_{CTC}$ is only about 2, which is significantly different . As a result, if λ_1 is 0.001, the SER loss will be more heavily weighted in the L1 loss, resulting in the ASR loss (and therefore the WER) not rising steadily.

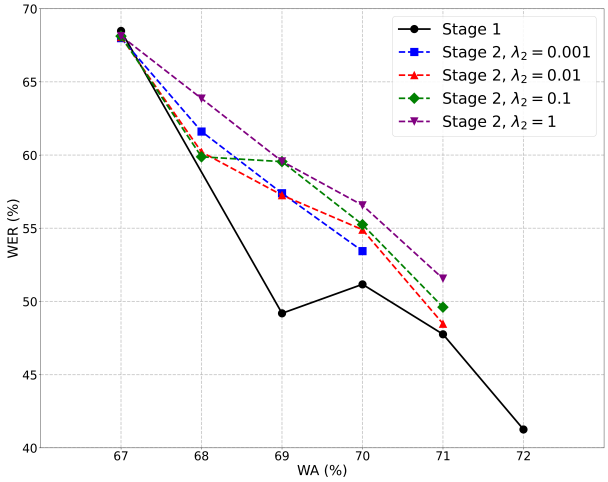

Fig. 4: The effect of different ASR weight coefficients λ_2 on emotion recognition performance recovery in the second stage.

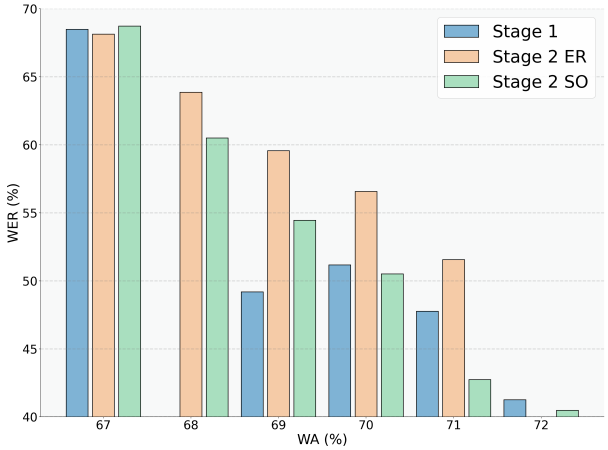

Fig. 5: The effects of fine-tuning the emotion-retaining (ER) module and semantics-obfuscating (SO) module respectively on emotion recognition recovery and speech content privacy protection in the second stage.

We systematically explore the effect of different ASR weight coefficients λ_2 on emotion recognition retain in the second stage, as shown in Figure 4. The experiment results show that the model is capable of recovering emotion recognition information at a certain level regardless of the weighting coefficients used. The model achieves the best recovery at the same WA when λ_2 is 1, and its WER improves by 5% compared to the baseline. This result validates the critical role of second-stage fine-tuning in recovering the performance of speech emotion recognition.

In this section, we also explore the effects of fine-tuning the ER module and SO module respectively in the second stage on emotion recognition recovery and speech content privacy protection, as shown in Figure 5. The parts without histograms in the figure indicate that the corresponding weighted accuracy

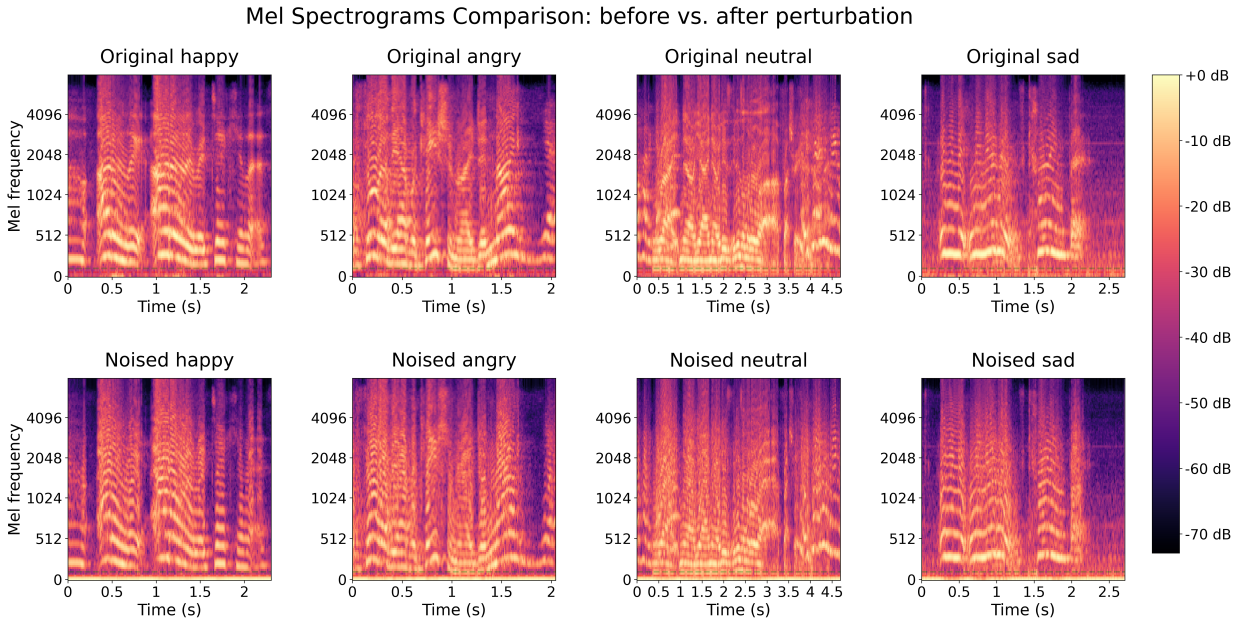

Fig. 6: The spectrogram of original and perturbed audio.

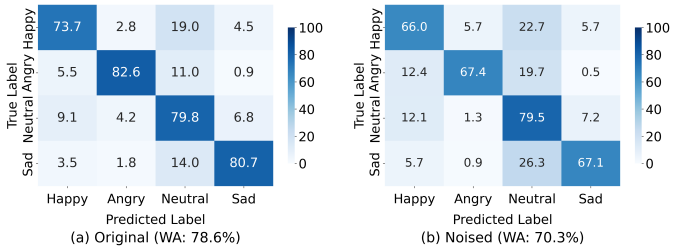

Fig. 7: The confusion matrix before and after perturbation.

(WA) value was not obtained under this condition. The experimental results show that freezing the parameters of the SO module and fine-tuning only the ER module can significantly improve the emotion classification performance and privacy preservation while ensuring stable perturbation generation. Specifically, under the same sentiment recognition accuracy (WA=70%), the WER for freezing SO to fine-tune ER is about 5% higher than that for freezing ER to fine-tune SO strategy. This suggests that during the fine-tuning process, the SO module tends to remove perturbation when in a trainable state, thus weakening its privacy-preserving capabilities. This verifies the importance of constraining the optimization of each module: freezing the SO module can effectively prevent its parameter deviation, thereby maintaining the stable injection of adversarial perturbation; and independent fine-tuning of the ER module can focus on the optimization of emotion recognition loss and avoid gradient conflict problems in multi-task training, thereby achieving a more effective balance between performance and privacy.

D. Visualization

Two visualization methods are used to verify the effectiveness of Aclipse: spectrogram comparison and confusion matrix. We first evaluate the perturbation effect by converting original and perturbed audio into spectrograms (Figure 6). In the spectrogram, transcription-related information is mainly reflected in high-frequency consonant energy (¿2000 Hz) and vowel formants (300–2000 Hz), while emotional cues depend on the fundamental frequency (50–500 Hz) and low-frequency harmonics (¡2000 Hz). The fundamental frequency trajectory captures intonation fluctuations, and the energy envelope together with speech rhythm forms rhythmic characteristics. Observations show that perturbation injection mainly reduces high-frequency energy above 2000 Hz, with minimal change below 2000 Hz. This differentiated perturbation injection effect shows that the Aclipse method has a relatively fine-grained control capability in adversarial perturbation generation, which helps to improve WER while retaining the low-frequency features related to emotion recognition as much as possible.

To further analyze the impact of Aclipse on emotion distribution, we visualize confusion matrices on the IEMOCAP dataset (Figure 7). In Figure 7(a), without perturbation, the model achieves 82.5% accuracy for ”anger”, but shows confusion between ”happy” and ”neutral”. In Figure 7(b), after applying Aclipse, overall performance remains stable, yet confusion between ”happy” and ”neutral” increases, and accuracy for ”angry” and ”sad” drops. These results suggest that while Aclipse effectively increases transcription error, it may introduce biases in recognizing certain emotions.

VI. Conclusion

In this paper, we investigate the intersection of SER and privacy preservation. We propose an attention-based cascaded learning strategy that enables effective SER while protecting speech content privacy. The core idea is to decouple multi-objective tasks by decomposing single-stage multi-objective optimization into multi-stage single-objective optimization,

achieving dual optimization of SER performance and privacy protection. Experiments on multiple pre-trained models demonstrate the effectiveness of the proposed strategy. Visualization of spectrograms and confusion matrices reveals the directional perturbation mechanism on speech features and its effect on classification decision boundaries. Future work will explore the distribution of privacy-sensitive and emotion-relevant regions in the time–frequency domain to design an adaptive spectrogram-level perturbation strategy.

ACKNOWLEDGMENT

This work was supported in part by the Natural Science Foundation of China (Grants No. 62472083, 62432008, 62501137) and the AI-Enhanced Research Program of Shanghai Municipal Education Commission (Grant No. SMEC-AI-DHUZ-01).

REFERENCES

[1] L. Tan, K. Yu, L. Lin, X. Cheng, G. Srivastava, J. C.-W. Lin, and W. Wei, "Speech emotion recognition enhanced traffic efficiency solution for autonomous vehicles in a 5g-enabled space–air–ground integrated intelligent transportation system," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 3, pp. 2830–2842, 2021.

[2] L. Pepa, L. Spalazzi, M. Capecci, and M. G. Ceravolo, "Automatic emotion recognition in clinical scenario: a systematic review of methods," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1675–1695, 2021.

[3] W. Chen, X. Xing, P. Chen, and X. Xu, "Vesper: A compact and effective pretrained model for speech emotion recognition," *IEEE Transactions on Affective Computing*, vol. 15, no. 3, pp. 1711–1724, 2024.

[4] Y. Yang, L. Yuan, J. Zhao, and W. Gong, "Content-agnostic backscatter from thin air," in *Proceedings of the 20th annual international conference on mobile systems, applications and services*, pp. 343–356, 2022.

[5] L. Yuan, C. Xiong, S. Chen, and W. Gong, "Embracing self-powered wireless wearables for smart healthcare," in *2021 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pp. 1–7, IEEE, 2021.

[6] F. Burkhardt, J. Wagner, H. Wierstorf, F. Eyben, and B. Schuller, "Speech-based age and gender prediction with transformers," in *Speech Communication; 15th ITG Conference*, pp. 46–50, VDE, 2023.

[7] C.-H. H. Yang, J. Qi, S. Y.-C. Chen, P.-Y. Chen, S. M. Siniscalchi, X. Ma, and C.-H. Lee, "Decentralizing feature extraction with quantum convolutional neural network for automatic speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6523–6527, IEEE, 2021.

[8] T. Feng, R. Peri, and S. Narayanan, "User-level differential privacy against attribute inference attack of speech emotion recognition in federated learning," *arXiv preprint arXiv:2204.02500*, 2022.

[9] S. Mohammadi, M. Mohammadi, S. Sinaei, A. Balador, E. Nowroozi, F. Flammini, and M. Conti, "Balancing privacy and accuracy in federated learning for speech emotion recognition," in *2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS)*, pp. 191–199, IEEE, 2023.

[10] H. Chen, H. Zhao, and Z. Zhang, "Gradient-level differential privacy against attribute inference attack for speech emotion recognition," *IEEE Signal Processing Letters*, 2024.

[11] T. Feng, H. Hashemi, M. Annavaram, and S. S. Narayanan, "Enhancing privacy through domain adaptive noise injection for speech emotion recognition," in *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 7702–7706, IEEE, 2022.

[12] M. Dias, A. Abad, and I. Trancoso, "Exploring hashing and cryptonet based approaches for privacy-preserving speech emotion recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2057–2061, IEEE, 2018.

[13] C. Tan, Y. Cao, S. Li, and M. Yoshikawa, "General or specific? investigating effective privacy protection in federated learning for speech emotion recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2023.

[14] T. Feng and S. Narayanan, "Privacy and utility preserving data transformation for speech emotion recognition," in *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 1–7, IEEE, 2021.

[15] H. S. Ali, F. ul Hassan, S. Latif, H. U. Manzoor, and J. Qadir, "Privacy enhanced speech emotion communication using deep learning aided edge computing," in *2021 IEEE International Conference on Communications Workshops (ICC Workshops)*, pp. 1–5, IEEE, 2021.

[16] H. Zhao, H. Chen, Y. Xiao, and Z. Zhang, "Privacy-enhanced federated learning against attribute inference attack for speech emotion recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2023.

[17] X. Cai, J. Yuan, R. Zheng, L. Huang, and K. Church, "Speech emotion recognition with multi-task learning.," in *Interspeech*, vol. 2021, pp. 4508–4512, Brno, 2021.

[18] J. Qian, F. Han, J. Hou, C. Zhang, Y. Wang, and X.-Y. Li, "Towards privacy-preserving speech data publishing," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, pp. 1079–1087, IEEE, 2018.

[19] W. Gong, L. Yuan, Q. Wang, and J. Zhao, "Multiprotocol backscatter for personal iot sensors," in *Proceedings of the 16th international conference on emerging networking experiments and technologies*, pp. 261–273, 2020.

[20] Q. Wang, S. Chen, J. Zhao, and W. Gong, "Rapidrider: Efficient wifi backscatter with uncontrolled ambient signals," in *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*, pp. 1–10, IEEE, 2021.

[21] W. Liu, S. Cao, and S. Zhang, "Multimodal consistency-specificity fusion based on information bottleneck for sentiment analysis," *Journal of King Saud University-Computer and Information Sciences*, vol. 36, no. 2, p. 101943, 2024.

[22] M. Dai, S. Chang, Y. Wang, and Z. Su, "Energy-efficient multi-access edge computing for heterogeneous satellite-maritime networks: a hybrid harvesting-and-offloading design," *IEEE Transactions on Mobile Computing*, 2025.

[23] L. Zhou, S. Chang, J. Luo, H. Wen, H. Zhu, and L. Lu, "Baroauth: Harnessing ear canal deformation for speaking user authentication on earbuds," in *2025 IEEE 45th International Conference on Distributed Computing Systems (ICDCS)*, pp. 111–121, IEEE, 2025.

[24] S. Chang, L. Zhou, W. Liu, H. Zhu, X. Hu, and L. Yang, "Combating voice spoofing attacks on wearables via speech movement sequences," *IEEE Transactions on Dependable and Secure Computing*, vol. 22, no. 1, pp. 819–832, 2024.

[25] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.

[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] E. Morais, R. Hoory, W. Zhu, I. Gat, M. Damasceno, and H. Aronowitz, "Speech emotion recognition using self-supervised features," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6922–6926, IEEE, 2022.

[28] 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*, vol. 45, no. 9, pp. 10745–10759, 2023.

[29] 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, no. 4, pp. 335–359, 2008.

[30] M. Xu, F. Zhang, X. Cui, and W. Zhang, "Speech emotion recognition with multiscale area attention and data augmentation," in *ICASSP 2021-2021 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pp. 6319–6323, Ieee, 2021.