Speech recognition and speech emotion recognition approach for VLSP 2025

Bui Tien Dat
Viettel AI
datbt7@viettel.com

Nguyen Duy Khanh
Viettel AI
khanhnd65@gmail.com

Abstract

In this paper, we present a comprehensive multi-stage training framework that maximally exploits diverse open-source datasets with varying quality conditions for ASR and SER tasks. Following this year's competition guidelines, we strategically leverage pre-approved external datasets and pretrained models through a carefully designed multi-stage training strategy. Our framework systematically processes heterogeneous data sources across multiple training phases, utilizing a unified pre-trained model architecture for both tasks to ensure optimal knowledge transfer between stages. For SER, we implement a hybrid loss function combining cross-entropy loss with supervised contrastive learning loss to handle quality variations and improve discriminative capabilities across different data sources. During inference, we employ an interpolation strategy that integrates predictions from the multi-stage trained model with k-nearest neighbors results for robust performance. Our approach demonstrates superior performance in the ASR-SER VLSP 2025 challenge by effectively utilizing the full spectrum of available open-source resources despite their quality disparities.

1 Introduction

In conversational agents, humans convey not only explicit requests, they also implicitly express emotions. Currently, modern AI-based conversational agents often integrate Automatic Speech Recognition (ASR) and Speech Emotion Recognition (SER) models at the same time, to improve user satisfaction. This combination offers many benefits and potential applications in fields such as human-computer communication, psychological health care, advertising analysis, and many other fields. For instance, in automated customer support systems, conversational AI agents that automatically transcribe the customers' utterances using ASR also recognize their emotions using SER models to provide appropriate suggestions and responses to improve the overall user experience. Therefore, it is essential to build reliable systems that can perform ASR and SER jointly to simplify the computational requirement and increase efficiency when exchanging information with each other. However, joint training ASR and SER models at the same time is difficult due to the lack of high-quality data pairing both text and emotion and the duration imbalance between text and emotion data. Additionally, due to the scarcity of emotional data and the difficulty of recognizing emotional speech, it is challenging for the SER model to work well in practice.

Currently, ASR models using end-to-end architecture have proven effective and achieved state-of-the-art (SOTA) results (Gulati et al., 2020; Kim et al., 2022b; Gao et al., 2023; Kim et al., 2022a). Because the ASR model achieves near-optimal results in ideal environmental conditions such as conversation systems, we focus on the development SER model. We hypothesize that the language information of the encoder model of the ASR model can significantly improve SER performance by eliminating natural intonation deviations in speech.

SER is a well-studied problem in the literature, with a variety of systems proposed that achieve SOTA performance on benchmark datasets (Wagner et al., 2023; Chen and Rudnicky, 2023; Zou et al., 2022; Abdelhamid et al., 2022; Morais et al., 2022). However, most of these systems are single-task learning and development only from high-quality public datasets, with very few systems taking a multi-task learning approach. Although both tasks use speech signals as input, ASR works more at the frame level, whereas SER recognizes emotion on larger timescales. The relationship between ASR and SER is an important but understudied topic. We emphasize that better auxiliary learning tasks can help the model learn improved represen-

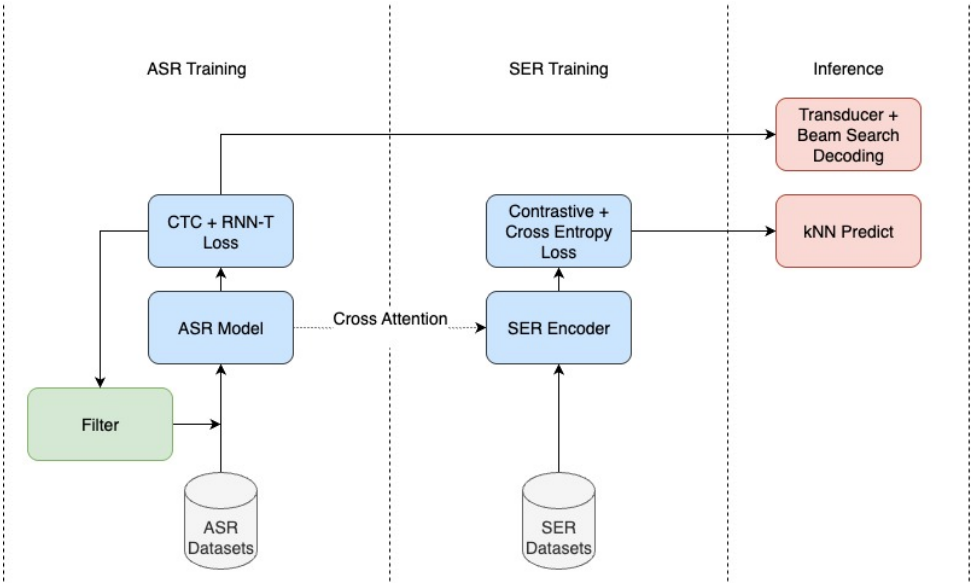

Figure 1: The overall framework of the proposed method for training SER and ASR models throughout all stages.

tations, thereby improving final SER performance. Several approaches for combined training of SER and ASR have been proposed and have achieved promising results (Li et al., 2022; Ghosh et al., 2023; Feng et al., 2020). While the combination of ASR and SER model training has gained traction in recent research, its application in real-world scenarios remains limited due to the persistent challenge of emotion labeling quality, which is inherently subjective and lacks comprehensive guidelines.

The availability of diverse open-source datasets and pre-trained models has significantly advanced the field of speech processing. Large-scale pre-trained models such as Wav2Vec2 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), and Whisper (Radford et al., 2023) have demonstrated remarkable capabilities in learning robust speech representations from massive unlabeled data. These models serve as powerful foundation architectures that can be effectively fine-tuned for downstream tasks including ASR and SER. Furthermore, the proliferation of open datasets across different languages, domains, and quality levels presents both opportunities and challenges. While datasets like LibriSpeech, Common Voice, and various emotional speech corpora provide rich training resources, their heterogeneous nature in terms of recording conditions, annotation quality, and domain specificity requires careful consideration during model development. The key challenge lies in effectively leveraging this diversity to build robust models that can generalize across different data conditions while maintaining high performance on target tasks.

In this paper, we present a comprehensive three-stage training framework for jointly developing ASR and SER models that effectively handles diverse data quality conditions. Our approach consists of: (1) Stage 1 - ASR model training using all available data types followed by systematic data quality filtering to identify high-quality samples; (2) Stage 2 - SER model fine-tuning utilizing the robust ASR encoder from Stage 1 with hybrid loss functions combining cross-entropy and supervised contrastive learning; and (3) Stage 3 - inference optimization through interpolation strategies that blend model predictions with k-nearest neighbors results. This multi-stage methodology allows us to maximally exploit heterogeneous datasets while maintaining model robustness across varying quality conditions. Experimental results demonstrate the effectiveness of our approach, achieving top-2 performance in the ASR-SER VLSP 2025 competition[1].

2 Methodology

The proposed method's diagram is shown in Figure 1 and includes three stages. The first stage focuses on comprehensive ASR model training using diverse datasets followed by systematic data quality filtering. The second stage involves SER model fine-tuning by leveraging public pretrained mod-

[1]https://vlsp.org.vn/vlsp2025/eval/asr-ser

els and ASR-SER fusion techniques. Finally, the third stage optimizes inference through advanced interpolation strategies and multi-modal fusion.

For the comprehensive training data shown in Table 1, we utilize a diverse collection of Vietnamese and international datasets spanning both ASR and SER tasks. Each dataset is strategically employed across different training stages to maximize information extraction and model robustness.

2.1 Stage 1: ASR Model Training and Data Quality Filtering

The first stage prioritizes building robust ASR models through comprehensive training and systematic data filtering, divided into two main phases.

2.1.1 Comprehensive ASR Training

We begin by fine-tuning various public pretrained models for ASR using all available Vietnamese speech datasets. Following the approach of OWSMv3.1 (Peng et al., 2024b), we employ standardized preprocessing and training procedures. The training utilizes both RNN-T (Graves, 2012) and CTC (Graves et al.) decoders with the combined loss function:

$$\mathcal{L}_{ASR} = \lambda \mathcal{L}_{RNN-T} + (1-\lambda)\mathcal{L}_{CTC} \quad (1)$$

The ASR training leverages large-scale Vietnamese datasets including VLSP2023 (300 hours), phoaudiobook (1494 hours), vivoice (1000 hours), viet_bud500 (500 hours), 28k_VigBigData (460k utterances), ViMD (100 hours), and VIVOS (15 hours). Multiple pre-trained models are evaluated including Conformer-based architectures, Whisper variants, and other state-of-the-art ASR models.

2.1.2 Training-Loop Data Filtering

Inspired by OWSM v3.1, v3.2, and v4 approaches (Peng et al., 2024b; Tian et al., 2024; Peng et al., 2025), we implement an iterative training-loop strategy to filter low-quality data samples. The process involves training initial ASR models, evaluating transcription quality, filtering samples based on dual criteria, and re-training on filtered data until convergence.

We employ WER and CTC confidence scores as complementary filtering metrics. Samples are retained only if they satisfy moderate quality thresholds:

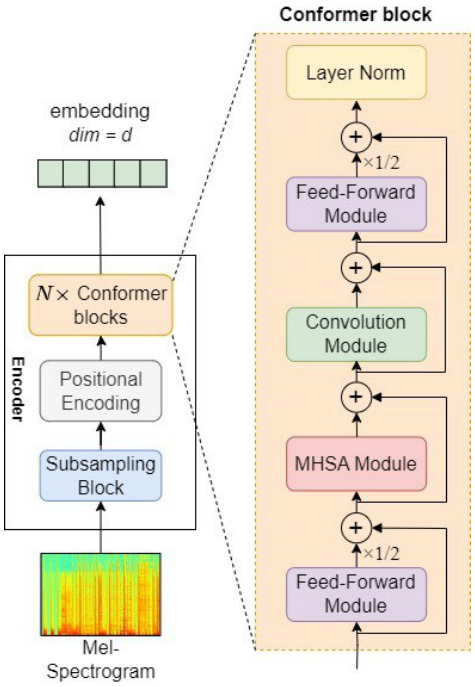

Figure 2: Architecture of the multi-stage training framework.

$$0.3 < \text{CTC\_conf} < 0.95 \text{ and } 5\% < \text{WER} < 40\% \quad (2)$$

This approach removes both overly simplistic samples (low WER + high confidence) and corrupted data (high WER + low confidence), ensuring retention of moderately challenging samples that effectively contribute to model learning.

2.2 Stage 2: SER Model Training and ASR-SER Fusion

The second stage focuses on developing robust SER models by leveraging public pretrained models and ASR-SER fusion techniques, utilizing both Vietnamese and international emotion datasets.

2.2.1 Public Pretrained Model-based SER Training

We systematically evaluate various public pretrained models specifically designed for SER tasks using a comprehensive collection of emotion datasets:

- **Vietnamese Emotion Data**: ViSEC (5400 utterances, 4 emotions) for Vietnamese-specific emotion recognition
- **International Emotion Datasets**: IEMOCAP (10k utterances), EMODB (535 utter-

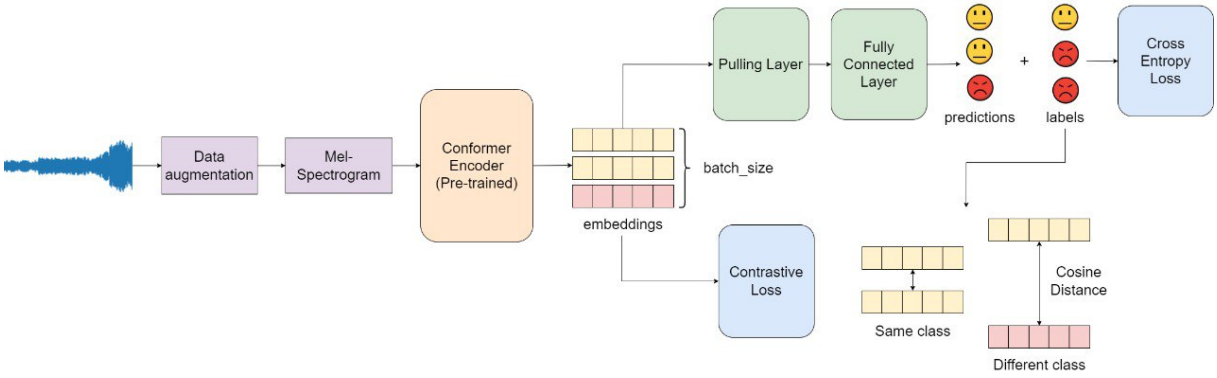

Figure 3: Fine-tuning cycle for SER task.

ances, 7 emotions), RAVDESS (1440 utterances, 7 emotions), CREMA-D (7442 utterances, 6 emotions), and EmoV-DB (7000 utterances)

- **WavLM fine-tuning**: Utilizing WavLM's enhanced speech understanding capabilities (Diatlova et al., 2024)
- **Emotion2Vec**: Leveraging specialized emotion-aware pre-trained representations (Ma et al., 2023)
- **Wav2Vec2.0**: Implementing proven SSL features, particularly from layer 9/12 which shows optimal performance for SER (Peng et al., 2024a)

2.2.2 ASR-SER Fusion Architecture

Building upon the high-quality ASR models from Stage 1, we implement advanced fusion strategies:

Cross-Attention Fusion: We employ cross-attention mechanisms to fuse semantic information extracted from the encoder layer output with acoustic information obtained from the public wav2vec-base checkpoint. The fusion architecture addresses the challenge of ASR errors degrading SER performance (Chen et al., 2024).

The SER training combines supervised contrastive learning with cross-entropy loss:

$$\mathcal{L}_{scl} = \sum_{i \in I} \frac{-1}{|P(i)|} \log \frac{\sum_{p \in P(i)} exp((x_i \cdot x_p)/\tau)}{\sum_{a \in A(i)} exp((x_i \cdot x_a)/\tau)} \quad (3)$$

$$\mathcal{L}_{SER} = (1-\alpha)\mathcal{L}_{ce} + \alpha\mathcal{L}_{scl} \quad (4)$$

where α balances the contribution of contrastive and classification losses.

2.3 Stage 3: Inference Optimization

The final stage implements advanced inference strategies combining model predictions with retrieval-based methods.

2.3.1 Multi-Modal Prediction Fusion

For SER inference, we employ a polling method, which has been proven optimal for SER tasks. The final prediction combines:

- ASR-SER fusion model outputs
- k-NN retrieval from training embeddings

2.3.2 k-NN Interpolation Strategy

Following (Wang et al., 2023), we create an embedding database from training and validation data:

$$(K, V) = \{(x_i, y_i), i \in D\} \quad (5)$$

The final prediction interpolates between model and k-NN predictions:

$$p(y|x) = \beta p_{model}(y|x) + (1-\beta)p_{knn}(y|x) \quad (6)$$

where β is optimized based on validation performance across different fusion strategies.

3 Experiment & Analysis

3.1 Datasets

We utilized a comprehensive collection of public datasets for training our emotion recognition system, as detailed in Table 1. Our dataset compilation includes both Vietnamese and international resources to ensure robust cross-lingual performance.

Dataset	Size	Label	Usage
Vietnamese ASR Datasets			
VLSP2023	300 hrs	ASR + SER	Stage 1+2
phoaudiobook	1494 hrs	audio+transcripts	Stage 1
vivoice	1000 hrs	audio+transcripts	Stage 1
viet_bud500	500 hrs	audio+transcripts	Stage 1
28k_VigBigData	460k utterances	audio+transcripts	Stage 1
ViMD	100 hrs	audio+transcripts	Stage 1
VIVOS	15 hrs	audio+transcripts	Stage 1
Emotion Recognition Datasets			
ViSEC	5400 utterances	4 emotions	Stage 2
IEMOCAP	10k utterances	Audio+emotion	Stage 2
EMODB	535 utterances	7 emotions	Stage 2
RAVDESS	1440 utterances	7 emotions	Stage 2
CREMA-D	7442 utterances	6 emotions	Stage 2
EmoV-DB	7000 utterances	emotion label	Stage 2

Table 1: Comprehensive dataset collection for multi-stage training framework.

Rank	User	WER (%)	SER Acc (%)	Final Score
1	hynguyenthien	**9.07**	82.21	**88.31**
2	ishowspeech (Ours)	11.38	79.13	85.77
3	dangnguyen-VLSP	12.66	80.84	85.39
4	SoFarSoGood	19.12	79.50	80.47
5	CodeSERSai	25.22	**85.79**	78.08
6	SoulSound	20.87	66.50	75.34
7	nhitny	23.56	71.76	75.04

Table 2: Performance comparison of teams in the ASR-SER VLSP 2025 competition leaderboard.

Vietnamese Datasets: We incorporated several Vietnamese speech datasets including VLSP2023 (300 hours), phoaudiobook (1,494 hours), vivoice (1,000 hours), viet_bud500 (500 hours), 28k_VigBigData (460k utterances), ViMD (100 hours), and VIVOS (15 hours) primarily for acoustic modeling and speech representation learning. Additionally, ViSEC (5,400 utterances) provides Vietnamese emotional speech data with 4 emotion categories.

International Emotion Datasets: For emotion-specific training, we employed established emotion recognition datasets: IEMOCAP (10k utterances), EMODB (535 utterances with 7 emotions), RAVDESS (1,440 utterances with 7 emotions), CREMA-D (7,442 utterances with 6 emotions), and EmoV-DB (7,000 utterances). These datasets provide diverse emotional expressions across different languages and recording conditions.

Data Preprocessing: We filtered out audio files with insufficient duration (< 0.5 seconds) and empty audio files from all datasets. We extracted 128-channel filterbank features using a 25ms window with 10ms stride.

Data Augmentation: We applied multi-domain augmentation strategies to enhance model robustness. In the temporal domain, we incorporated background noise injection, impulse response convolution, and pitch shifting to create acoustic variations while preserving emotional content. For spectral augmentation, we employed SpecAugment (Park et al., 2019) with frequency masking ($F = 27$) and temporal masking with maximum ratio ($pS = 0.05$), where mask duration is proportional to utterance length. The pitch shifting technique particularly benefits emotion recognition by generating diverse vocal pitch variations for identical emotional categories.

3.2 Model configuration

After careful consideration, we decided to implement an ASR model from scratch while utilizing a pre-trained Wav2Vec2.0 encoder for the SER component.

ASR Model Architecture: The ASR encoder module comprises one subsampling block that provides 4 times temporal dimension reduction for the input sequences and 12 layers of the conformer model. Each conformer layer has 512 input dims and 2048 hidden dims with 8 heads of self-attention. The pre-training uses mask config same as 3.1. Since the encoder has 4 times temporal dimension reduction, the quantization with random projections stacks every 4 frames for projections. The vocab size of the codebook is 8192 and the dimension is 16. The model has 80 million total learnable parameters.

We adopt the grapheme-based tokenizer scheme from the top-performing solution at this 2023 challenge, utilizing a vocabulary of 804 tokens for Vietnamese multilingual speech recognition.

The ASR model employs CTC (Graves et al.) and RNN-T (Graves, 2012) for the decoder, which consists of a predictor layer, a joint dense layer, and an output layer with softmax non-linearity. The

predictor network contains a layer of unidirectional LSTM, where the hidden dimension of the LSTM is 640. The joint network layer has a size of 512 and the output layer has 804 dimensions, representing 804 graphemes.

SER Model Architecture: For the Speech Emotion Recognition component, we employ a pre-trained Wav2Vec2.0 encoder to leverage its robust feature extraction capabilities for emotional speech understanding.

We set λ to 0.5 for the loss function in both loss functions 1 and 4.

All networks are trained using a transformer learning rate schedule (Vaswani et al., 2017). The training of the model uses Adam optimizer (Kingma and Ba, 2014) with a 0.004 peak learning rate and 10000 warmup steps. All training is done using two NVIDIA A100 GPUs with a batch size of 32 and the number of epochs is limited to 100.

3.3 Results

Model performance is assessed using Word Error Rate (WER_{ASR}) and Emotion Recognition Accuracy (ACC_{SER}) metrics.

$$WER_{ASR} = \frac{S + D + I}{N} \tag{7}$$

where S represents substitution errors, D denotes deletion errors, I indicates insertion errors, C is the count of correctly recognized words, and N is the total number of words in the reference transcription ($N = S + D + C$).

$$ACC_{SER} = \frac{NEU_{Correct}}{NEU \times 2} + \frac{NEG_{Correct}}{NEG \times 2} \tag{8}$$

where $NEU_{Correct}$ denotes correctly classified neutral emotion utterances, NEU represents the total neutral utterances, $NEG_{Correct}$ indicates correctly classified negative emotion utterances, and NEG represents the total negative utterances.

The final evaluation score is computed as:

$$Final\_Score = 0.7 \times (1 - WER_{ASR}) + 0.3 \times ACC_{SER} \tag{9}$$

As demonstrated in Table 2, our proposed approach achieved second place with a competitive final score, with particularly strong performance in the ASR component compared to competing teams.

4 Conclusion

This work presents an innovative training framework for jointly optimizing Automatic Speech Recognition and Speech Emotion Recognition tasks through a unified pre-trained model architecture. Our integrated approach enables the development of an end-to-end ASR-SER system with the capability to leverage diverse data modalities during the training phase. Experimental results validate the effectiveness of this methodology, especially in scenarios with limited or low-quality emotional speech data. The practical value of our approach is demonstrated through achieving second place in the ASR-SER VLSP 2025 competition, establishing its competitiveness in real-world applications.

References

Abdelaziz A. Abdelhamid, El-Sayed M. El-Kenawy, Bandar Alotaibi, Ghada M. Amer, Mahmoud Y. Abdelkader, Abdelhameed Ibrahim, and Marwa Metwally Eid. 2022. Robust Speech Emotion Recognition Using CNN+LSTM Based on Stochastic Fractal Search Optimization Algorithm. *IEEE Access*, 10:49265–49284. Conference Name: IEEE Access.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *arXiv preprint*. ArXiv:2006.11477.

Li-Wei Chen and Alexander Rudnicky. 2023. Exploring Wav2vec 2.0 Fine Tuning for Improved Speech Emotion Recognition. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. ISSN: 2379-190X.

Mingjie Chen, Hezhao Zhang, Yuanchao Li, Jiachen Luo, Wen Wu, Ziyang Ma, Peter Bell, Catherine Lai, Joshua Reiss, Lin Wang, and 1 others. 2024. 1st place solution to odyssey emotion recognition challenge task1: Tackling class imbalance problem. *arXiv preprint arXiv:2405.20064*.

Daria Diatlova, Anton Udalov, Vitalii Shutov, and Egor Spirin. 2024. Adapting wavlm for speech emotion recognition. *arXiv preprint arXiv:2405.04485*.

Han Feng, Sei Ueno, and Tatsuya Kawahara. 2020. End-to-End Speech Emotion Recognition Combined with Acoustic-to-Word ASR Model. In *Interspeech 2020*, pages 501–505. ISCA.

Zhifu Gao, Shiliang Zhang, Ian McLoughlin, and Zhijie Yan. 2023. Paraformer: Fast and Accurate Parallel Transformer for Non-autoregressive End-to-End Speech Recognition. *arXiv preprint*. ArXiv:2206.08317.

Sreyan Ghosh, Utkarsh Tyagi, S. Ramaneswaran, Harshvardhan Srivastava, and Dinesh Manocha. 2023. MMER: Multimodal Multi-task Learning

for Speech Emotion Recognition. *arXiv preprint*. ArXiv:2203.16794.

Alex Graves. 2012. Sequence Transduction with Recurrent Neural Networks. *arXiv preprint*. ArXiv:1211.3711.

Alex Graves, Santiago Fernandez, Faustino Gomez, and Jurgen Schmidhuber. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. *arXiv preprint*. ArXiv:2005.08100.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *arXiv preprint*. ArXiv:2106.07447.

Kwangyoun Kim, Felix Wu, Yifan Peng, Jing Pan, Prashant Sridhar, Kyu J. Han, and Shinji Watanabe. 2022a. E-Branchformer: Branchformer with Enhanced merging for speech recognition. *arXiv preprint*. ArXiv:2210.00077.

Sehoon Kim, Amir Gholami, Albert Shaw, Nicholas Lee, Karttikeya Mangalam, Jitendra Malik, Michael W. Mahoney, and Kurt Keutzer. 2022b. Squeezeformer: An Efficient Transformer for Automatic Speech Recognition. *Advances in Neural Information Processing Systems*, 35:9361–9373.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization.

Yuanchao Li, Peter Bell, and Catherine Lai. 2022. Fusing ASR Outputs in Joint Training for Speech Emotion Recognition. *arXiv preprint*. ArXiv:2110.15684.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2023. emotion2vec: Self-supervised pre-training for speech emotion representation. *arXiv preprint arXiv:2312.15185*.

Edmilson Morais, Ron Hoory, Weizhong Zhu, Itai Gat, Matheus Damasceno, and Hagai Aronowitz. 2022. Speech Emotion Recognition Using Self-Supervised Features. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6922–6926. ISSN: 2379-190X.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Interspeech 2019*, pages 2613–2617. ISCA.

Feichong Peng, Sanshuai Cui, and Zhun Ling. 2024a. Enhancing end-to-end speech emotion recognition using dual-stream with multi-task learning. In *2024 10th International Conference on Computer and Communications (ICCC)*, pages 917–921. IEEE.

Yifan Peng, Shakeel Muhammad, Yui Sudo, William Chen, Jinchuan Tian, Chyi-Jiunn Lin, and Shinji Watanabe. 2025. Owsm v4: Improving open whisper-style speech models via data scaling and cleaning. *arXiv preprint arXiv:2506.00338*.

Yifan Peng, Jinchuan Tian, William Chen, Siddhant Arora, Brian Yan, Yui Sudo, Muhammad Shakeel, Kwanghee Choi, Jiatong Shi, Xuankai Chang, and 1 others. 2024b. Owsm v3. 1: Better and faster open whisper-style speech models based on e-branchformer. *arXiv preprint arXiv:2401.16658*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Jinchuan Tian, Yifan Peng, William Chen, Kwanghee Choi, Karen Livescu, and Shinji Watanabe. 2024. On the effects of heterogeneous data sources on speech-to-text foundation models. *arXiv preprint arXiv:2406.09282*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need.

Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W. Schuller. 2023. Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10745–10759. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

Xuechen Wang, Shiwan Zhao, and Yong Qin. 2023. Supervised Contrastive Learning with Nearest Neighbor Search for Speech Emotion Recognition. In *INTERSPEECH 2023*, pages 1913–1917. ISCA.

Heqing Zou, Yuke Si, Chen Chen, Deepu Rajan, and Eng Siong Chng. 2022. Speech Emotion Recognition with Co-Attention Based Multi-Level Acoustic Information. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7367–7371. ISSN: 2379-190X.