LANCET: Lightweight Attention-enhanced Network for Robust Speech Emotion Recognition

Yassin Terraf, *Student Member, IEEE,* Youssef Iraqi, *Senior Member, IEEE*

***Abstract*—Speech emotion recognition (SER) involves classifying emotional states from speech signals based on acoustic characteristics. Although recent advances in transformer-based models achieve state-of-the-art performance through large-scale pretraining on diverse datasets, their computational requirements limit practicality in low-resource settings. CNN-based approaches with attention mechanisms have been proposed to balance computational efficiency and performance but often struggle to extract robust features in challenging acoustic environments. To address these limitations, we propose LANCET, a lightweight attention-enhanced network for robust emotion recognition under diverse recording conditions. LANCET integrates multiscale channel-wise attention to focus on noise-resilient spectral features, Temporal Convolutional Network (TCN) to model short-term and long-term temporal dependencies, and frame-wise attention to prioritize frames relevant for speech emotion recognition. Experiments on the IEMOCAP corpus, including its subsets: improvisation, script, and full, used clean and augmented datasets with babble, music, and ambient noise. Results demonstrate that LANCET outperforms CNN-based methods and achieves superior performance compared to HuBERT, a transformer-based model, in both clean and challenging conditions, showcasing robustness, effectiveness, and efficiency with fewer parameters. The code for LANCET is available at https://github.com/YassinTERRAF/LANCET.**

***Index Terms*—Speech Emotion Recognition, Attention Mechanism, Temporal Convolutional Network, Noise Robustness, Challenging Acoustic Environments.**

I. Introduction

Speech emotion recognition (SER) is the process of identifying emotional states from speech signals through the extraction and analysis of acoustic patterns. Accurate SER is essential for improving human-machine interaction and has wide-ranging applications in fields such as education [1], healthcare [2], and customer service, where understanding emotional states improves communication and decision-making. Recent advances in SER have focused mainly on transformer-based models, such as HuBERT [3] and WavLM [4], which leverage large-scale pretraining on diverse datasets. This pretraining enables these models to learn rich contextual representations of speech signals, which are critical for effective emotion recognition. However, their reliance on substantial computational resources makes them less practical for deployment in resource-constrained environments, such as mobile or embedded devices [5]. Alternatively, CNN-based approaches that incorporate attention mechanisms have been proposed to achieve a balance between good performance and reduced computational complexity [6]–[8]. Moreover, these approaches are typically evaluated in controlled environments using noise-free speech signals, which do not accurately reflect real-world conditions where background noise can significantly degrade emotional cues. This gap in evaluation settings limits their robustness and practical applicability in scenarios involving noisy or unpredictable acoustic environments.

To address these limitations, we propose LANCET, a novel lightweight attention-enhanced CNN-based network for emotion recognition. LANCET integrates multiscale channel-wise attention to dynamically enhance noise-resilient spectral features by weighting frequency channels according to their relevance. These refined spectral features are further processed by Temporal Convolutional Network (TCN), which models temporal patterns on varying time scales to capture the emotional dynamics of speech. Finally, frame-wise attention highlights the most emotionally salient frames, producing a robust feature representation that performs effectively in both clean and challenging acoustic environments.

The main contributions of this paper can be summarized as follows.

- We introduce LANCET, a novel lightweight attention-enhanced CNN-based emotion recognition network that integrates multiscale channel-wise attention, TCN, and frame-wise attention to improve emotion recognition under diverse acoustic conditions.
- We augment the IEMOCAP dataset with various noise types, including babble, music, and ambient noise, applied at different Signal-to-Noise Ratio (SNR) levels, to simulate real-world conditions and evaluate the robustness of the proposed method.
- Extensive experiments were conducted on the IEMOCAP dataset and its augmented versions, demonstrating the effectiveness of the proposed approach in both clean and challenging acoustic environments.

The remainder of this paper is organized as follows: Section II describes the proposed LANCET approach. Section III presents the experimental results, and Section IV concludes the paper.

II. Methodology

This section presents the proposed LANCET architecture. As illustrated in Figure 1, LANCET integrates three core components: multiscale channel-wise attention, TCN blocks, and frame-wise attention.

A. Multiscale Channel-wise Attention

In speech emotion recognition, local patterns within the log mel spectrogram vary across frequency channels, with certain channels carrying more discriminative features than

ISBN: 978-9-46-459362-4

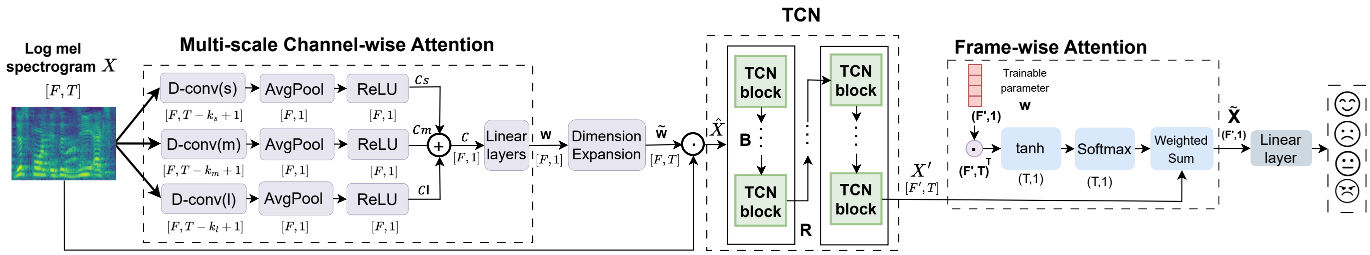

Fig. 1. The overall architecture of the Lightweight Attention-Enhanced Network for Comprehensive Emotion Recognition with TCN (LANCET). ⊙ indicates the dot product operation.

others. This variability is particularly important in noisy environments, where some channels may be affected by degradation, while others retain salient features. To address this, a multiscale channel-wise attention mechanism is introduced to dynamically assign importance to each channel, emphasizing those that contribute the most effectively to emotion recognition.

The log mel spectrogram $X \in \mathbb{R}^{F\times T}$, where F is the number of frequency channels and T is the sequence length, is used as input. Inspired by [9], temporal features at different scales are extracted for each frequency channel using parallel 1-D depthwise convolutions with kernel sizes k_s, k_m, and k_l applied along the time axis. Specifically, k_s captures fine-grained temporal details, k_m focuses on mid-range patterns, and k_l extracts long-term temporal dependencies, providing the context necessary to dynamically weight each frequency channel in each frame based on its relevance. Following the convolutional operations, average pooling is applied along the time axis to reduce the dimensionality of the features, which is followed by a ReLU activation function to introduce nonlinearity. This process produces pooled features for each kernel size, C_s, C_m, and C_l, representing the temporal characteristics of each frequency channel on different scales. These features are then combined through a fully connected layer, resulting in a unified representation $C \in \mathbb{R}^F$ that encodes the multiscale channel-wise features.

To further model dependencies between channels, two fully connected layers are used. The first layer introduces a bottleneck by reducing the dimensionality of C, and the second restores the original dimensionality while applying a sigmoid activation to compute the final attention weights $W \in \mathbb{R}^F$, which represent the relative importance of each frequency channel. The computed attention weights W are broadcast across all time frames to match the dimensions of the input X, producing expanded attention weights $\tilde{W} \in \mathbb{R}^{F\times T}$. The weighted log mel spectrogram $\hat{X} \in \mathbb{R}^{F\times T}$ is then obtained via element-wise multiplication:

$$\hat{X} = \tilde{W} \odot X \tag{1}$$

where ⊙ denotes element-wise multiplication.

B. Temporal Convolutional Network (TCN)

The TCN [10] is an architecture specifically designed to model temporal sequences, enabling effective learning of temporal dependencies [11]. It has been successfully applied in various speech processing tasks, such as speech separation and overlapping speech detection [12], [13]. A typical TCN block comprises three key components: an input 1×1 convolution, a depthwise dilated convolution, and an output 1×1 convolution. Between these layers, parametric ReLU activations and normalization layers are applied to enhance learning stability, while residual connections mitigate the vanishing gradient problem, preserving critical temporal features.

Our proposed architecture, inspired by Conv-TasNet [14], employs stacked TCN blocks to effectively capture both short-term and long-term temporal dependencies in speech signals, where short-term dependencies refer to how features change over adjacent frames, and long-term dependencies model how features evolve across distant frames over time. The weighted log mel spectrogram $\hat{X} \in \mathbb{R}^{F\times T}$ is passed through N stacked TCN blocks, repeated R times, where N and R are hyperparameters. In each block, the dilation factor increases exponentially as $2^0, 2^1, \dots, 2^{X-1}$, expanding the receptive field and allowing the model to capture temporal patterns on varying time scales. The output of the stacked TCN blocks is a refined temporal feature representation $X' \in \mathbb{R}^{F'\times T}$, where F' represents the enhanced feature dimension, and T remains the sequence length.

C. Frame-wise Attention

After the TCN extracts the refined temporal feature map $\mathbf{X}'$, which captures both short-term and long-term temporal dependencies across frames, not all frames in the sequence are equally informative for emotion recognition. To address this, we introduce a frame-wise attention mechanism that dynamically assigns importance to each frame, allowing the model to focus on the most relevant parts of the sequence.

Given the refined temporal feature map $\mathbf{X}'$, we compute an attention score for each frame using a learned vector $\mathbf{w} \in \mathbb{R}^{F'}$, which captures the contribution of each feature to the overall importance of a frame.

For each frame t, we calculate the attention score e_t by performing a dot product between the feature vector at frame t and the learned attention vector $\mathbf{w}$:

$$e_t = \mathbf{w}^T \mathbf{X'}_t, \quad t = 1, 2, \dots, T. \tag{2}$$

Attention scores e_t are then passed through a softmax function to produce attention weights:

$$\alpha_t = \frac{\exp(e_t)}{\sum_{t'=1}^{T} \exp(e_{t'})}, \quad \alpha_t \in [0, 1], \tag{3}$$

where α_t represents the attention weight assigned to the t-th frame, ensuring that the weights sum to 1 across all frames. These weights effectively measure the relative importance of each frame within the sequence.

Finally, we compute a weighted sum of frame features using the attention weights α_t, resulting in a condensed feature representation $\tilde{\mathbf{X}} \in \mathbb{R}^{F'}$:

$$\tilde{\mathbf{X}} = \sum_{t=1}^{T} \alpha_t \cdot \mathbf{X}'_t, \tag{4}$$

where $\tilde{\mathbf{X}}$ is a weighted combination of the frame features, emphasizing the most informative frames for emotion recognition. This condensed feature representation $\tilde{\mathbf{X}}$ is then passed through a linear layer to map it directly to the emotion classes.

III. Experiments

In this section, we describe the experimental setup, compare the performance of the proposed LANCET with state-of-the-art methods, and conduct an ablation study to evaluate the contribution of individual components of LANCET.

A. Dataset and Feature Extraction

The experiments are conducted on the IEMOCAP benchmark corpus [15], a widely used dataset for SER research [6], [7], [16]–[21]. IEMOCAP contains 12 hours of emotional speech from 10 actors, covering diverse speaking styles and spontaneous interactions. This dataset setup follows the standard evaluation methodology in speech emotion recognition, as used in prior studies [6], [7], [19]–[21], which have **exclusively used IEMOCAP as a benchmark dataset**, ensuring fair comparisons with existing methods. We assess performance across three subsets: (1) the *improvisation* subset, which captures natural emotional variations in spontaneous speech; (2) the *scripted* subset, which contains controlled speech with explicit emotional portrayals; and (3) the *full* subset, combining both for a comprehensive evaluation. For feature extraction, we compute log mel spectrograms with 128 mel filter banks. We process audio signals with a 25 ms frame size, a 10 ms frame shift, and a 16 kHz sampling rate. To minimize spectral leakage, we apply a Hamming window.

1) Data Augmentation: To simulate real-world acoustic conditions, the IEMOCAP dataset was augmented with noise from the MUSAN dataset [22], including babble, music, and ambient noise. Babble noise was created by mixing three to eight speech files from the "us-gov" section of MUSAN. Noise was added to each audio sample at five SNR levels: 0, 5, 10, 15, and 20 dB. To evaluate the generalizability of the proposed approach, separate subsets of noise files were used for training/validation and testing, with no overlap between the subsets to ensure that the test set remained independent of the training and validation sets. For each IEMOCAP subset (improvisation, scripted, and full), the augmentation process produced 15 noisy datasets per subset, corresponding to 3 noise types combined with 5 SNR levels. This resulted in a total of 45 augmented datasets in all subsets.

2) Baselines and Evaluation Metrics: To evaluate our proposed model, we compare it with several state-of-the-art CNN-based models, including APCNN [6], MHCNN [7], AACNN [19], GLAM [20], E-GLAM [21], TC-Net [23], and HuBERT-Base [3], a transformer-based approach.

For performance evaluation, we use **Weighted Accuracy (WA)** and **Unweighted Accuracy (UA)**, which are commonly applied in SER research [20], [21]. Since WA and UA may peak at different models, we report their average as a single accuracy metric, following previous works [6], [19].

To assess the **statistical significance** of performance differences, we employ the Approximate Randomization test [24].

3) Implementation Details: In our experiments, we follow the evaluation protocol in [25], using the neutral, sad, angry, and merged happy-excited classes. The IEMOCAP dataset was split via 10-fold speaker-independent cross-validation, with one speaker for testing, eight for training, and one for validation. All models were retrained under this protocol to ensure fair evaluation. To align with baselines, dataset clips were segmented into 2-second windows with a 1-second overlap for training and 1.6-second clips for validation/testing. Predictions were averaged across segments. LANCET's multiscale channel-wise attention uses kernel sizes $k_s = 3$, $k_m = 5$, and $k_l = 10$ for fine-grained, mid-range, and long-term temporal features. TCN consists of 5 stacked layers with 3 repetitions. We optimize with cross-entropy loss using Adam (10^{-6} weight decay), an initial learning rate of 10^{-4} (decayed by 0.95 per epoch), and train for 50 epochs with batch size 32. To enhance generalization, we applied a mixed training method [26] with a mixing rate of 0.5.

B. Results and Discussion

Experiments were conducted on the IEMOCAP dataset to evaluate the performance of the proposed LANCET approach under both clean and noisy recording conditions. Table I provides a performance comparison of the proposed method with state-of-the-art CNN-based approaches and HuBERT, under clean conditions, while Figure 2 illustrates the results under different types of noise applied at varying SNR levels.

1) Performance Comparison in Clean Recording Conditions: Table I presents the performance of the proposed LANCET approach compared to state-of-the-art CNN-based models under clean recording conditions across the IEMOCAP subsets: improvisation, script, and full. LANCET significantly outperforms all baseline models across all subsets. In the improvisation subset, LANCET achieves a WA of 74.35%, outperforming the second-best model, HuBERT, by 2.16%. In the script subset, LANCET reaches a WA of 63.12%, exceeding HuBERT by 1.68%. In the full subset, which combines the script and improvisation subsets, LANCET attains a WA of 68.04%, surpassing HuBERT by 0.5%. These results

TABLE I

PERFORMANCE COMPARISON OF LANCET WITH OTHER APPROACHES ACROSS ALL IEMOCAP SUBSETS UNDER CLEAN CONDITIONS. **BOLD** VALUES REPRESENT THE BEST PERFORMANCE, AND UNDERLINED VALUES INDICATE THE SECOND-BEST PERFORMANCE.

Method	Improvisation		Script		Full	
	WA (↑)	UA (↑)	WA (↑)	UA (↑)	WA (↑)	UA (↑)
APCNN [6]	66.80±1.92	64.12±1.27	46.75±2.81	47.04±2.24	60.36±1.30	62.25±1.15
MHCNN [7]	68.23±2.04	66.45±2.49	54.57±2.40	52.66±2.05	64.56±2.77	66.23±2.49
AACNN [19]	69.65±1.20	67.63±0.89	57.96±1.25	56.13±1.31	66.59±0.97	67.26±0.75
GLAM [20]	70.51±2.45	68.28±2.39	60.17±2.14	58.13±2.97	66.52±0.77	67.17±1.21
E-GLAM [21]	71.36±1.79	69.30±1.17	60.76±1.98	59.13±2.07	65.57±1.29	66.52±1.30
TC-Net [23]	71.79±1.84	70.50±1.68	59.73±2.19	58.27±2.22	66.95±1.73	67.05±1.65
HuBERT [3]	72.19±0.84	71.05±1.28	61.44±1.17	60.01±1.21	67.54±1.74	68.11±0.77
LANCET	**74.35±1.17**	**71.66±0.99**	**63.12±1.36**	**61.44±1.26**	**68.04±1.06**	**69.16±1.15**

demonstrate the superior performance of LANCET in speech emotion recognition in all subsets of the IEMOCAP dataset under clean conditions.

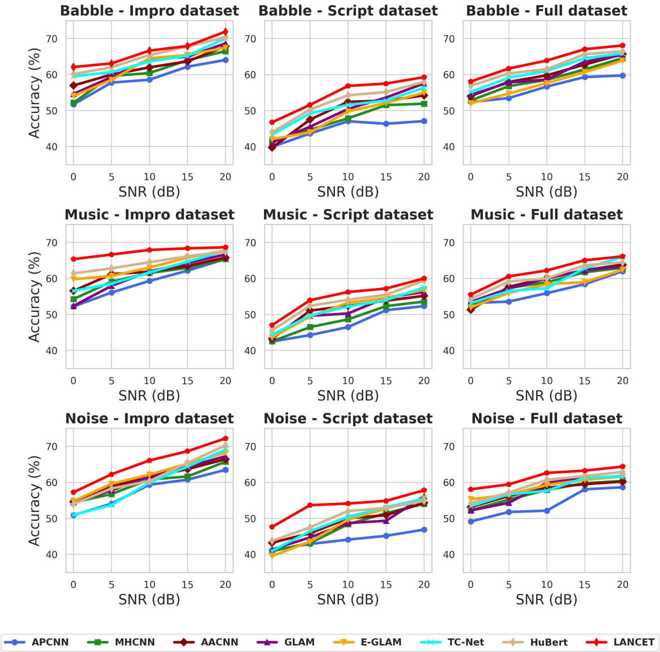

Fig. 2. Performance comparison of LANCET with other approaches across all IEMOCAP subsets under noisy conditions. Impro denotes the improvisation subset.

2) Performance Comparison in Noisy Recording Conditions: Figure 2 illustrates the performance of LANCET compared to state-of-the-art CNN-based approaches and HuBERT, a transformer-based model, under various noise conditions, including babble, music, and ambient noise, applied at five SNR levels: 0, 5, 10, 15, and 20 dB. Across all noise types and SNR levels, LANCET significantly outperforms competing methods, demonstrating its robustness in challenging acoustic environments. At lower SNR levels (0, 5, and 10 dB), LANCET shows a significant performance advantage. This is attributed to multiscale channel-wise attention, which prioritizes spectral features based on their noise resilience, and frame-wise attention, which emphasizes frames most relevant for emotion recognition. In contrast, other methods treat spectral and temporal features uniformly, limiting their effectiveness in noisy conditions. As SNR increases, the performance gap narrows, particularly beyond 10 dB, reflecting the reduced impact of noise as the recordings become cleaner. The type of noise affects LANCET's performance differently across the IEMOCAP subsets. Music noise has the greatest impact on improvisation speech due to its harmonic and melodic components overlapping with the tonal variations of spontaneous speech, making it harder to extract emotion-relevant features. Ambient noise poses the greatest challenge for the script and full subsets because its dynamic and unpredictable nature disrupts consistent emotional cues in scripted and combined speech data.

TABLE II

ABLATION STUDY OF LANCET UNDER CLEAN CONDITIONS ACROSS DIFFERENT IEMOCAP SUBSETS. FWA INDICATES FRAME-WISE ATTENTION, AND MCWA INDICATES MULTISCALE CHANNEL-WISE ATTENTION.

Method	TCN	FWA	MCWA	Accuracy (%) (↑)
		Improvisation subset		
TCN	✓	✗	✗	70.37
TCN + FWA	✓	✓	✗	70.94
TCN + MCWA	✓	✗	✓	71.04
LANCET	✓	✓	✓	**73.00**
		Script subset		
TCN	✓	✗	✗	59.14
TCN + FWA	✓	✓	✗	60.92
TCN + MCWA	✓	✗	✓	61.07
LANCET	✓	✓	✓	**62.28**
		Full subset		
TCN	✓	✗	✗	66.73
TCN + FWA	✓	✓	✗	67.39
TCN + MCWA	✓	✗	✓	68.04
LANCET	✓	✓	✓	**68.60**

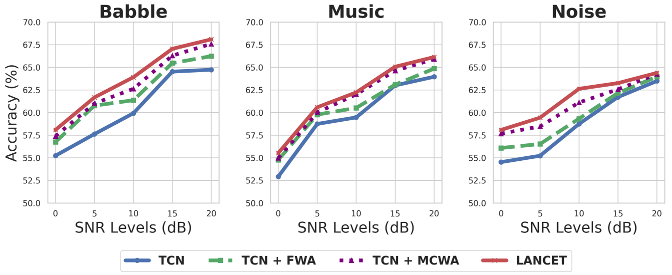

Fig. 3. Ablation study of LANCET under noisy conditions on the IEMOCAP full subset.

C. Ablation Study

We conducted ablation experiments to assess the impact of TCN, multiscale channel-wise attention, and frame-wise attention in LANCET. Table II shows results under clean conditions, while Fig. 3 presents noisy condition results for the full subset, with similar trends observed in the improvisation and script subsets.

We use TCN as the base model, evaluating the individual contributions of frame-wise and multiscale channel-wise attention. Among the two attention mechanisms, multiscale channel-wise attention has a greater impact, as it prioritizes informative frequency channels, mitigates microphone self-noise in clean conditions, and reduces environmental noise interference in noisy settings. Frame-wise attention refines the temporal representations learned by TCN, complementing its ability to model multi-scale dependencies across frames.

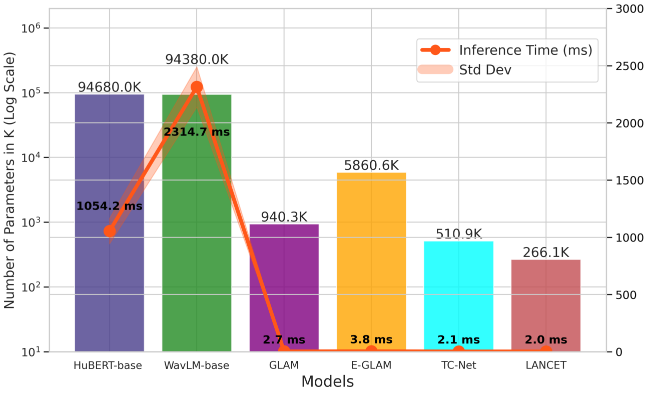

Fig. 4. Comparison of LANCET with State-of-the-Art Models, including WavLM and HuBERT, based on Parameter Count (Log Scale) and Average Inference Time (ms).

We further evaluate LANCET's computational efficiency by comparing its parameter count and inference time on 4-second speech signals (Fig. 4). Transformer-based models, such as HuBERT-base (94.68M parameters) [3] and WavLM-base (94.38M parameters) [4], have significantly higher computational costs, with inference times of 1054 ms and 2314 ms, respectively. Although HuBERT and WavLM have comparable parameter sizes, WavLM exhibits higher inference latency due to its more complex architecture.

In contrast, LANCET requires only 266.1K parameters and achieves a 2 ms inference time, making it the most efficient model among CNN- and transformer-based approaches. This highlights its suitability for real-time and resource-constrained applications.

IV. Conclusion

In this paper, we proposed a novel lightweight approach for emotion recognition under diverse recording conditions, LANCET, which integrates a TCN to effectively model short- and long-term dependencies in speech signals, a multiscale channel-wise attention mechanism to emphasize discriminative and noise-resilient frequency channels, and a frame-wise attention mechanism to focus on the most relevant frames. Comprehensive experiments on the IEMOCAP dataset, including its original and augmented versions, demonstrate that LANCET achieves state-of-the-art performance, significantly outperforming existing methods in both clean and challenging conditions.

References

[1] Dahiru Tanko, Sengul Dogan, Demir, et al. Shoelace pattern-based speech emotion recognition of the lecturers in distance education: ShoePat23. *Appl. Acoust*, 190:108637, 2022.

[2] Meishu Song, Andreas Triantafyllopoulos, Yang, et al. Daily mental health monitoring from speech: A real-world japanese dataset and multitask learning analysis. In *ICASSP 2023-2023 I*, pages 1–5. IEEE, 2023.

[3] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. 29:3451–3460, October 2021.

[4] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Xiong Liu, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.*, 16(6):1505–1518, 2022.

[5] Xiaofen Xing Zhipeng Li and al. Multi-scale temporal transformer for speech emotion recognition. In *INTERSPEECH 2023*, 2023.

[6] Pengcheng Li, Yan Song, Ian Vince McLoughlin, Wu Guo, and Li-Rong Dai. An attention pooling based representation learning method for speech emotion recognition. 2018.

[7] Mingke Xu, Fan Zhang, and Samee U Khan. Improve accuracy of speech emotion recognition with attention head fusion. In *CCWC 2020*, pages 1058–1064. IEEE, 2020.

[8] Yassin Terraf and Youssef Iraqi. Robust feature extraction using temporal context averaging for speaker identification in diverse acoustic environments. *IEEE Access*, 12:14094–14115, 2024.

[9] Chenglin Xu, Wei Rao, Eng Siong Chng, and Haizhou Li. Spex: Multi-scale time domain speaker extraction network. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 28:1370–1384, 2020.

[10] Colin Lea, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks: A unified approach to action segmentation. In *ECCV Workshops*, pages 47–54. Springer, 2016.

[11] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[12] Cunhang Fan, Jianhua Tao, Bin Liu, Jiangyan Yi, Zhengqi Wen, and Xuefei Liu. End-to-end post-filter for speech separation with deep attention fusion features. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 28:1303–1314, 2020.

[13] Samuele Cornell, Maurizio Omologo, Stefano Squartini, and Emmanuel Vincent. Detecting and counting overlapping speakers in distant speech scenarios. In *INTERSPEECH 2020*, 2020.

[14] Yi Luo and Nima Mesgarani. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 27(8):1256–1266, 2019.

[15] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Kazemzadeh, et al. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359, 2008.

[16] Yuanyuan Zhang, Jun Du, Zirui Wang, Jianshu Zhang, and Yanhui Tu. Attention based fully convolutional network for speech emotion recognition. In *APSIPA ASC 2018*, pages 1771–1775. IEEE, 2018.

[17] Lorenzo Tarantino, Philip N Garner, Alexandros Lazaridis, et al. Self-Attention for Speech Emotion Recognition. In *Interspeech*, pages 2578–2582, 2019.

[18] Jiang Li and Wang et al. CFN-ESA: A Cross-Modal Fusion Network With Emotion-Shift Awareness for Dialogue Emotion Recognition. *IEEE Trans. Affect. Comput.*, pages 1–16, 2024.

[19] Mingke Xu, Fan Zhang, Xiaodong Cui, and Wei Zhang. Speech emotion recognition with multiscale area attention and data augmentation. In *ICASSP 2021-2021 I*, pages 6319–6323. IEEE, 2021.

[20] Wenjing Zhu and Xiang Li. Speech emotion recognition with global-aware fusion on multi-scale feature representation. In *ICASSP 2022-2022 I*, pages 6437–6441. IEEE, 2022.

[21] Lingli Yu, Fengjun Xu, Yundong Qu, and Kaijun Zhou. Speech emotion recognition based on multi-dimensional feature extraction and multi-scale feature fusion. *Appl. Acoust*, 216:109752, 2024.

[22] David Snyder, Guoguo Chen, and Daniel Povey. Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*, 2015.

[23] Muhammad Ishaq, Mustaqeem Khan, and Soonil Kwon. TC-Net: A Modest & Lightweight Emotion Recognition System Using Temporal Convolution Network. *Comput. Syst. Sci. Eng.*, 46(3):3355–3369, 2023.

[24] Eric W Noreen. *Computer-intensive methods for testing hypotheses*. Wiley New York, 1989.

[25] Nikolaos Antoniou, Athanasios Katsamanis, and Giannakopoulos. Designing and Evaluating Speech Emotion Recognition Systems: A reality check case study with IEMOCAP. In *ICASSP 2023*. IEEE, 2023.

[26] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond Empirical Risk Minimization, 2018.