Proceedings of the 4th International Conference on Computing Innovation and Applied Physics
DOI: 10.54254/2753-8818/86/2025.20333

Research Advanced in Speech Emotion Recognition based on Deep Learning

Zirui He[1,a,*]

[1]*College of Arts & Science, New York University, NY, USA*
a. zh2266@nyu.edu
**corresponding author*

Abstract: Speech Emotion Recognition (SER)'s burgeoning significance within intelligent systems is underscored by its transformative impact across various fields, from human-computer interaction, and virtual assistants to mental health monitoring. Over the rapid development of this technology in the past two decades, studies have continuously confronted and overcome kinds of real-world challenges, such as data scarcity, environmental noise, and cross-language differences. This survey focuses on recent innovations in SER, particularly deep learning architectures, and synthetic data augmentation, and addresses recent developments in cross-domain and multimodal SER techniques, which have expanded the applicability of SER to more diverse datasets.

Keywords: Speech emotion recognition, deep learning, data augmentation.

1. Introduction

Speech is one of humanity's most unique abilities. Compared to written text or body language, spoken language conveys emotions more directly and immediately. Even when individuals from different corners of the globe face language barriers, the tone, intonation, and rhythm of speech can evoke emotions universally. As technology grows more integrated into daily life, there is an increasing need for intelligence systems to interact with humans in a more natural, emotionlike manner. As Rosalind Picard aptly noted, "We're not going to build intelligent machines until we build, if not something we call emotion, then something that functions like our emotion systems" [1]. True intelligence, as she suggests, is that which closely resembles genuine human beings, with the expression of emotions through speech being one of the most challenging aspects to replicate.

Traditional Speech Emotion Recognition (SER) involves several steps, including speech signal pre-processing, feature extraction, dimensionality reduction, and selecting the appropriate classifier for analyzing emotion categories [2]. Since the late 20th century, researchers have attempted to leverage various machine learning and deep learning techniques for SER. As early as 2002, Chang-hyun Park et al.[3] applied recurrent neural networks (RNNs) to SER tasks, marking one of the earliest deep-learning explorations in this field. The introduction of Support Vector Machine (SVM) further advanced SER, with SVM outperforming radial basis function neural network, knearest-neighbor, and linear discriminant classifiers, achieving 85% accuracy because of its good discriminating ability [4].

In addition to these, Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs) were widely used in early SER tasks. While these traditional models performed well on small and

DOI: 10.54254/2753-8818/86/2025.20333

structured datasets, they often struggled with real-world scenarios involving noisy environments or more complex emotional nuances.

2. Recent Innovations in SER

2.1. Deep Learning Architectures

Deep learning Architectures have significantly revolutionized the field of SER. In contrast to traditional machine learning approaches, which largely depend on handed extracted features like Mel-Frequency Cepstral Coefficients (MFCC) or pitch, deep learning models can automatically derive pertinent features during training.

A new method can effectively focus on useful information in speech features, using a convolutional neural network based on an attention mechanism and a bidirectional gated Recurrent unit (BiGRU) is proposed.[5] Introducing the attention mechanism aims to achieve the primary goal of the method, which is to enhance recognition accuracy by directing the system's attention more towards the key features of speech. In the experiment phase, this method investigates the system recognition performance when the number of CNN layers changes, turns out that the number of CNN-layer is less than 5, the recognition accuracy is steadily increased. [5] And it discovers that the BiGRU architecture outperforms the more commonly used BiLSTM in terms of both recognition accuracy and computational efficiency.

Another innovative model is SERC-GCN (Speech Emotion Recognition in Conversation using Graph Convolutional Networks), which better predicts a speaker's emotional state by incorporating conversational context, speaker interactions, and temporal dependencies between utterances [6]. This two-stage graph-based SER model is designed to classify a speaker's emotional state in dyadic conversations. In the first stage, the model extracts speech features solely from utterances by converting these utterances into cyclic graphs, which are transformed through a two-layer GCN. In the second stage, a related conversation graph with vertices initialized using the utterance features from the first stage was created. And these conversation graphs capture context-sensitive and speakersensitive relationships through relational edges, which illustrate the dependencies between the speakers of the utterances.

In addition, a novel approach introduced in [7] uses contrastive pretraining to improve SER performance, especially with unlabeled data. The idea behind contrastive learning is to differentiate between positive and negative pairs of data points based on intra-speaker clusters. The learning goal aims to enhance similarity among positive pairs while reducing it for negative pairs. In this approach, positive pairs consist of utterances sampled from the same intra-speaker cluster are likely reflecting the same emotional category, while negative examples are formed using different intra-speaker clusters from the same speaker, representing varying emotional categories. The model is based on wav2vec2.0 [8], including a feature extractor and a transformer encoder, but with a more compact design featuring only 6 transformer layers. And the experiment result shows the ability of this model to perform well on unlabeled data, indicating the potential to enhance the SER where labeled data is limited.

2.2. Synthetic Data Augmentation

The ongoing research of SER, remains a persistent challenge that there is a limited quantity of large, balanced, labeled, and high-quality datasets. Popular datasets such as RAVDESS [9] contain 7,356 audio files across eight emotional categories but lacks sufficient coverage for certain subtle emotions like fear or disgust. Similarly, the IEMOCAP [10] contains 12 hours of audiovisual data featuring five sessions of two actors (one male and one female per session) engaging in conversations. However,

DOI: 10.54254/2753-8818/86/2025.20333

the imbalance in emotional categories often skews results toward more frequent emotions, highlighting the need for synthetic data to diversify and balance datasets.

Synthetic Data Augmentation (SDA) offers a solution by artificially generating or transforming speech data, thereby expanding the size and emotional variety of available datasets. A recent study [11] introduces an innovative method that employs a state-of-the-art end-to-end speech emotion conversion model to generate synthetic data for training SER models. This method relies on two different models: a generative model to synthesize speech (speech-to-speech emotion conversion) and an emotion classification model from the raw audio waveform (fine-tuned wav2vec 2.0). It uses a sequence-tosequence (ES2S) model to translate the phonetic-content unit representations, which allows the problem to be treated as a spoken language translation problem, and the objective is to learn to map discrete speech representations between different emotions. The synthesized emotional speech is generated using a modified version of the HiFi-GAN vocoder, takes synthetic data, including phonetic-content units, predicted durations, F0, speaker embeddings, and emotion, to predict speech signal waveforms, forming synthetic data for training. This method also augments the RAVDESS and IEMOCAP datasets, improves the performance of wav2vec2.0 model fine-tuned with both synthetic and real data.

Another innovative approach, presented in [12], enables emotional text-to-speech (TTS) synthesis on datasets without explicit emotion labels. The method uses a Cross-domain SER model to extract features and classify emotions. After using the SER model to assign the soft emotion labels to the TTS datasets, this approach uses a GST-based emotional TTS model to learn the emotion-related style and control the emotional expressiveness of the generated speech. Applying this method, the system can successfully generate synthetic speech with controlled emotional expressiveness while maintaining high-quality output.

2.3. Biologically Inspired Methods

Bio-inspired method is used to solve many complex problems easily by modeling the characteristics of biological species [13]. With the biologically inspired methods, offer models an alternative to conventional machine learning approaches, providing a more natural and adaptive way to process emotion recognition in speech.

A recent method that directly operates on the speech signal combines the classical source-filter model of human speech production with the liquid state machine (LSM), a biologically inspired spiking neural network (SNN) [14]. In this method, the biological element LSM based on is biological cortical neurons. Then, its original LSM design with two separate reservoirs builds upon the motor theory of human speech perception. During the preprocessing phase, the speech signal is divided into 2 parts: the source and the vocal tract, and both are fed into two separate neural reservoirs of spiking neurons within the LSM. And the output from these two reservoirs is reduced in dimensionallly in order to form more compact representations. A final classifier takes these reduced outputs, and determines which emotion the speaker is expressing based on the processed speech signal.

2.4. Multimodal Approaches

In recent years, multimodal approaches have gained increasing attention for enhancing the generalizability and resilience of speech emotion recognition. Traditional SER approaches are used to study on single language or environment, facing challenges when applied to emotional context are expressed in various data types. Multimodal Approaches can better solve these problems, and adapt to more complex and diverse needs. In the paper [15] presents a new approach integrating Automatic Speech Recognition (ASR) into SER to solve the common problem of SER's lack of data in real-world applications. The method utilizes several ASR models trained on different datasets. The model

DOI: 10.54254/2753-8818/86/2025.20333

is mainly made of 2 components: By combining ASR features, multimodal fusion of textual and auditory information improves SER performance. The first component contributes to enhance SER performance by incorporating ASR features, while the second component focuses on training a reliable joint model on labeled and unlabeled data. This study finds that ASR errors has a profound impact on SER performance, especially when the WER exceeds 25%.

3. Comparative Results of Recent Techniques

The recent advancements in Speech Emotion Recognition (SER) exhibit a diverse range of techniques, each aiming to improve recognition accuracy and adaptability in real-world scenarios. A key trend observed in the field is the integration of innovative architectures, such as attention mechanisms, biologically inspired models, and multimodal approaches, which have demonstrated substantial improvements in both performance and robustness.

As shown in Table 1, one notable contribution is the ACNN + Multi-head Selfattention model used in the study Head Fusion: Improving the Accuracy and Robustness of SER. This model, tested on the IEMOCAP and RAVDESS datasets, achieved a clean data accuracy (UA) of 72.26% The inclusion of multi-head selfattention allows the model to focus on critical parts of the input speech, improving the detection of emotional cues across the signal. More importantly, this model showcases robust performance in noisy environments, with a slower degradation of accuracy when noise is introduced. This makes it particularly valuable for real-world applications, where speech signals are often distorted by environmental noise. By leveraging the self-attention mechanism, this model effectively captures emotional nuances across various conditions.

In contrast, the Liquid State Machine (LSM) model proposed in Biologically Inspired Speech Emotion Recognition represents a shift towards bio-inspired methods. Tested on a custom emotional speech database, this model achieved a recognition rate of 82.35%. The LSM, with its spiking neurons designed to mimic biological cortical neurons, operates directly on raw speech data, bypassing traditional feature extraction. By dividing the speech signal into two components-source and vocal tract-LSM processes each through separate reservoirs, a design based on the motor theory of speech perception. This biologically inspired architecture offers a novel approach, emphasizing adaptability and cognitive realism. The high recognition rate suggests that biologically inspired models have strong potential in SER, especially in cases where traditional feature engineering may be insufficient. The GST-based TTS model combined with a Cross-domain SER model, as detailed in Emotion Controllable Speech Synthesis Using Emotion-Unlabeled Dataset, addresses a different challenge: the lack of emotion-labeled data. This method, tested on an emotion-unlabeled TTS dataset, achieved an emotion prediction accuracy of 78.75% for four emotion classes (our-4cls). It leverages a Global Style Token (GST)based approach to control emotional expressiveness in synthetic speech, reducing the reliance on labeled datasets. The cross-domain SER model further enhances performance by classifying emotions without the need for explicit emotion annotations. This approach is particularly beneficial in scenarios where collecting large amounts of labeled emotional data is impractical. Although the accuracy varies across emotion classes, this method highlights the potential of synthetic data and cross-domain learning in SER.

The wav2vec2.0+ contrastive learning strategy employed in Revealing Emotional Clusters in Speaker Embeddings shows the effectiveness of contrastive learning and multi-task learning (MTL) in improving SER performance. On the IEMOCAP dataset, the model reached an unweighted accuracy (UAR) of 73.80%, while on CREMA-D, it achieved 83.01%. The contrastive pretraining strategy differentiates between similar and dissimilar speaker embeddings, enabling the model to cluster emotions more effectively. This technique is especially powerful for datasets with limited labeled emotional data, as it enhances the model's ability to generalize by learning from unlabeled

DOI: 10.54254/2753-8818/86/2025.20333

examples. The combination of wav2vec2.0, a state-ofthe-art speech representation model, and MTL further boosts accuracy by simultaneously optimizing multiple objectives.

In Towards Improving Speech Emotion Recognition Using Synthetic Data Augmentation, the use of wav2vec 2.0 with synthetic data augmentation demonstrates the impact of augmenting training data with synthetically generated emotional speech. By augmenting the IEMOCAP and RAVDESS datasets, the model improved UAR to 76.19% and 93.05%, respectively, when trained on both real and synthetic data. This method addresses the common challenge of limited data in SER by generating synthetic emotional speech through emotion conversion techniques. The results indicate that synthetic data augmentation can significantly enhance SER models, particularly in speaker-independent setups, where generalization to new speakers is critical.

Tabel 1: Results comparison from several studies on Speech Emotion Recognition (SER) methods.

Paper	Model	Dataset	Accuracy	Key features
[18]	ACNN+Multi-head Self-attention	IEMOCAP, RAVDESS	Clean data (SNR = clean, AF = 0): UA = 72.26%	(1) Multi-head self-attention; (2) Slower accuracy degradation under noise
[14]	Liquide State Machine with dual reservoirs	Custom emotional speech database	82.35% recognition rate	(1) No feature extraction; (2) Biological elements: LSM's spiking neurons resemble biological cortical neurons; (3) Dual reservoirs
[12]	GST-based TTS model+Cross-domain SER model	Emotion-unlabeled TTS dataset	Emotion accuracy: 78.75% (our-4cls), 49.25% (full-4cls), 36.75% (base-4cls); Arousal accuracy: 91.0%; Valence accuracy: 55.5%	(1) GST-based approach for emotional speech synthesis; (2) Uses an MMD-based cross-domain SER model; (3) Reduced need for emotion-annotated data
[7]	FTwav2vec2.0 w/proposed MTL	IEMOCAP, CREMA-D	IEMOCAP UAR: 69.16% (contrastive), 73.80% (wav2vec2.0+MTL) CREMA-D UAR: 75.23% (contrastive), 83.01% (wav2vec2.0 + MTL)	(1) Contrastive pretraining strategy; (2) Multi-task learning (MTL); (3) wav2vec2.0 fine-tuning
[11]	wav2vec 2.0 + Synthetic Data Augmentation	IEMOCAP, RAVDESS	IEMOCAP UAR: 76.19% (Original + Synthetic, SD), 66.06% (SI); RAVDESS UAR: 93.05%(Original+Synthetic, SD), 81.29% (SI)	(1) Synthetic data augmentation from emotion conversion; (2) Speaker-independent setup
[19]	ELM+Utterance-level features, with SVM, RF, XGBoost, RVM, SGD	Toronto Emotional Speech Set (TESS)	Accuracy Range: 95.55% to 100%; Increment in accuracy: 3.22% to 6.28%; Peak Accuracy: 100% (ELM + RF)	(1) Utterance-level features combined with speech features ; (2) ELM + RF model

DOI: 10.54254/2753-8818/86/2025.20333

Table 1: (continued).

[5]	CNN-BiGRU with attention mechanism	CASIA, RAVDESS	CASIA: 88.92% (CNN-BiGRU + Attention); RAVDESS: 87.65% (CNN-BiGRU + Attention); 5-layer CNN Accuracy: 88.92%	(1) Attention mechanism improves recognition by focusing on key speech features; (2) BiGRU (Bidirectional Gated Recurrent Unit)
[15]	ASR-SER integration with cross-attention, W2V2	IEMOCAP, MSP Podcast	SER accuracy: 63.4% (ASR fusion)	(1) ASR-SER integration; (2) Cross-Attention Fusion Model
[6]	Two-stage GCN model with context, speaker, and temporal information	IEMOCAP	Micro-F1 (Utterance-only): 40.3%; Micro-F1 (Conversation): 51.5%; WA (Graph): 66.8% (Recency + Self-dependency);	(1) Graph Convolutional Networks; (2) Combines utterance-level and conversation-level features; (3) Recency and Self-dependency

Finally, the CNN-BiGRU model with attention mechanism, as explored in CNN-BiGRU Speech Emotion Recognition Based on Attention Mechanism, further reinforces the value of attention in SER. Tested on the CASIA and RAVDESS datasets, it achieved accuracies of 88.92% and 87.65%, respectively. The attention mechanism helps the model focus on key speech features, while the BiGRU architecture efficiently captures temporal dependencies. This combination ensures high recognition accuracy, making it a competitive approach in SER tasks.

4. Future Challenges in SER

As technology advances and innovations emerge, researchers continually devise new techniques to enhance and refine the Speech Emotion Recognition (SER) system. They aim to establish a model that yields higher accuracy and adapts effectively to real scenarios. While studies have accomplished numerous innovations and enhancements, SER is still facing with many known and unknown challenges.

4.1. Data Scarcity and Quality

One of the most persistent challenges in SER is the scarcity and quality of labeled emotional data. Accurately identifying emotions from spontaneous speech in real-world scenarios, especially in non-labeled data, remains difficult [16]. While datasets such as RAVDESS and IEMOCAP provide valuable resources, they often lack the complexity and variability seen in natural speech, such as diverse languages, accents, and subtle emotional expressions. Models trained on these datasets may struggle to generalize effectively in real-world contexts, where linguistic diversity and emotional ambiguity are prevalent. As mentioned in the Synthetic Data Augmentation section, while synthetic data can help bridge this gap, future research needs to focus on acquiring higher-quality, more representative datasets that better capture the complexity of human emotions across various contexts and cultural nuances. This will require not only expanding the amount of data but also improving its depth, diversity, and labeling accuracy.

DOI: 10.54254/2753-8818/86/2025.20333

4.2. Noise in SER

Another key challenge in SER is the impact of noise in real-world environments. As artificial intelligence increasingly mimics human capabilities, the need for SER systems to function reliably in noisy environments becomes essential. Many studies included in this survey tested models on relatively clean, noise-free datasets, which do not reflect the noisy conditions encountered in real-world scenarios. In practical applications, background noise can drastically reduce recognition accuracy by introducing distortions into the speech signal, leading to misclassifications [17]. Although some recent techniques—such as noise-robust feature extraction and classifier optimization—have shown promise in mitigating the effects of noise, they still fall short of achieving human-like performance in challenging acoustic conditions. To overcome this, future research must focus on developing more sophisticated methods for noise resistance, including the creation of more representative noisy datasets, the enhancement of noise-robust features, and improved speech enhancement techniques. These solutions will be crucial to advancing the field and making SER more viable for real-world use, particularly in applications such as call centers, healthcare, and mobile technology.

4.3. C. Cross-Domain and Multimodal SER

A major limitation of current SER models is their difficulty in adapting to different domains or environments, a challenge known as cross-domain transfer. Speech data collected in diverse settings, such as hospitals, call centers, or entertainment venues, exhibit vastly different properties. Models trained in one environment may struggle to generalize when applied to another, causing a significant drop in accuracy. This lack of cross-domain adaptability limits the widespread applicability of SER systems, which often require manual re-tuning or retraining to function effectively in new environments.

Moreover, the challenge of integrating multiple modalities—such as combining audio with visual or contextual data—presents another significant hurdle. Multimodal SER has the potential to provide richer emotional insights by leveraging diverse sources of information, yet the complexity of combining these modalities often leads to suboptimal model performance. Many existing approaches still rely heavily on unimodal data, such as speech alone, and fail to fully exploit the advantages of multimodal inputs.

5. Conclusion

The goal of this paper is to offer a comprehensive survey on recent innovative techniques in speech emotion recognition. The paper first briefly introduces the context of Speech emotion recognition and the traditional SER models, then gives detailed summaries of several recent techniques that enhance the SER in different aspects. We reviewed a wide range of recent advancements in deep learning architectures, Synthetic Data Augmentation, biologically inspired methods, and multimodal approaches. And compared the results and the key features of these innovative approaches. While significant progress has been made, several challenges remain, including the limitation of existing datasets, handling noisy real-world applications, cross-domain adaptation. With the sustained development of SER, its potential applications in affective computing, healthcare, and humancomputer interaction gives us compelling reasons and beliefs that we should continuously dedicate to the further innovation in this field.

DOI: 10.54254/2753-8818/86/2025.20333

References

[1] Higginbotham,Adam "Welcome to Rosalind Picard's touchy- feelyworld of empathictech." URL:http://www.wired.co.uklmagazine/archive/2012 /11(features/emotion-machines.)

[2] H. Zhao, N. Ye and R. Wang, "A Survey on Automatic Emotion Recognition Using Audio Big Data and Deep Learning Architectures," 2018 IEEE 4th International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing, (HPSC) and IEEE International Conference on Intelligent Data and Security (IDS), Omaha, NE, USA, 2018, pp. 139142.

[3] Chang-Hyun Park, Dong-Wook Lee and Kwee-Bo Sim, "Emotion recognition of speech based on RNN," Proceedings. International Conference on Machine Learning and Cybernetics, Beijing, China, 2002, pp. 22102213 vol.4.

[4] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi and E. Ambikairajah, "A Comprehensive Review of Speech Emotion Recognition Systems," in IEEE Access, vol. 9, pp. 47795-47814, 2021.

[5] L. Zhang, Y. Wang, J. Du and X. Wang, "CNN-BiGRU Speech Emotion Recognition Based on Attention Mechanism," 2023 2nd International Conference on Artificial Intelligence and Intelligent Information Processing (AIIIP), Hangzhou, China, 2023, pp. 85-89.

[6] D. Chandola, E. Altarawneh, M. Jenkin and M. Papagelis, "SERC-GCN: Speech Emotion Recognition In Conversation Using Graph Convolutional Networks," ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, Republic of, 2024, pp. 76-80.

[7] I. R. Ulgen, Z. Du, C. Busso and B. Sisman, "Revealing Emotional Clusters in Speaker Embeddings: A Contrastive Learning Strategy for Speech Emotion Recognition," ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, Republic of, 2024, pp. 12081-12085.

[8] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in Proceedings of the 34th International Con- ference on Neural Information Processing Systems, Red Hook, NY, USA, 2020, NIPS'20, Curran Associates Inc.

[9] Livingstone, S. R., Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE, 13(5), e0196391.

[10] Busso, C., Bulut, M., Lee, C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., Narayanan, S. S. (2008). IEMOCAP: interactive emotional dyadic motion capture database. Language Resources and Evaluation, 42(4), 335–359.

[11] K. M. Ibrahim, A. Perzo and S. Leglaive, "Towards Improving Speech Emotion Recognition Using Synthetic Data Augmentation from Emotion Conversion," ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, Republic of, 2024, pp. 10636-10640.

[12] X. Cai, D. Dai, Z. Wu, X. Li, J. Li and H. Meng, "Emotion Controllable Speech Synthesis Using Emotion-Unlabeled Dataset with the Assistance of Cross-Domain Speech Emotion Recognition," ICASSP 2021- 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 5734-5738.

[13] X.-S. Yang, S. F. Chien, and T. O. Ting, "Bio-Inspired Computing," Morgan Kaufmann, 2015. DOI: https://doi.org/10.1016/C2014-0-00501-1.

[14] R. Lotfidereshgi and P. Gournay, "Biologically inspired speech emotion recognition," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 2017, pp. 5135-5139.

[15] Y. Li, "Enhancing Speech Emotion Recognition for Real-World Applications via ASR Integration," 2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), Cambridge, MA, USA, 2023, pp. 1-5.

[16] M. S. Nair and D. P. Gopinath, "Transfer learning for Speech Based Emotion Recognition," 2022 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES), THIRUVANANTHAPURAM, India, 2022, pp. 559-564.

[17] Swapna Mol George, P. Muhamed Ilyas, A review on speech emotion recognition: A survey, recent advances, challenges, and the influence of noise, Neurocomputing, Volume 568, 2024, 127015, ISSN 0925-2312.

[18] M. Xu, F. Zhang and W. Zhang, "Head Fusion: Improving the Accuracy and Robustness of Speech Emotion Recognition on the IEMOCAP and RAVDESS Dataset," in IEEE Access, vol. 9, pp. 74539-74549, 2021.

[19] Ainurrochman and U. L. Yuhana, "Improving Performance of Speech Emotion Recognition Application using Extreme Learning Machine and Utterance-level," 2024 International Seminar on Intelligent Technology and Its Applications (ISITIA), Mataram, Indonesia, 2024, pp. 466-470.