HuBERT-CLAP: Contrastive Learning-Based Multimodal Emotion Recognition using Self-Alignment Approach

Long H. Nguyen[*]
Ton Duc Thang University
Vietnam
hoanglong.fruitai@gmail.com

Nhat Truong Pham[*]
Sungkyunkwan University
Republic of Korea
truongpham96@skku.edu

Mustaqeem Khan[†]
Mohamed Bin Zayed University of Artificial Intelligence
UAE
mustaqeemicp@gmail.com

Alice Othmani
Université Paris-Est Créteil (UPEC)
France
alice.othmani@u-pec.fr

Abdulmotaleb EI Saddik[‡]
Mohamed Bin Zayed University of Artificial Intelligence
UAE
elsaddik@uottawa.ca

Abstract

A breakthrough in deep learning has led to improvements in speech emotion recognition (SER), but these studies tend to process fixed-length segments, resulting in degraded performance. Therefore, multimodal approaches that combine audio and text features improve SER but lack modality alignment. In this study, we introduce HuBERT-CLAP, a contrastive language-audio self-alignment pre-training framework for SER to address the aforementioned issue. Initially, we employ CLIP to train a contrastive self-alignment model using HuBERT for audio and BERT/DistilBERT for text to extract discriminative cues from the input sequences and map informative features from text to audio features. Additionally, HuBERT in the pre-trained HuBERT-CLAP undergoes partial fine-tuning to enhance the effectiveness in predicting emotional states. Furthermore, we evaluated our model on the IEMOCAP dataset, where it outperformed the non-pre-training model, achieving a weighted accuracy of 77.22%. Our source code is publicly available at https://github.com/oggyfaker/HuBERT-CLAP/ for reproducible purposes.

CCS Concepts

• **Computing methodologies → Speech recognition**; **Neural networks**; **Supervised learning by classification**.

Keywords

Affective Computing, Contrastive Learning, Human-Computer Interaction, Partial Fine-Tuning, Speech Emotion Recognition

[*] Both authors contributed equally to this research.
[†] Co-corresponding author.
[‡] Corresponding author.

MMASIA '24, December 03–06, 2024, Auckland, New Zealand
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1273-9/24/12
https://doi.org/10.1145/3696409.3700183

ACM Reference Format:
Long H. Nguyen, Nhat Truong Pham, Mustaqeem Khan, Alice Othmani, and Abdulmotaleb EI Saddik. 2024. HuBERT-CLAP: Contrastive Learning-Based Multimodal Emotion Recognition using Self-Alignment Approach. In *ACM Multimedia Asia (MMASIA '24), December 03–06, 2024, Auckland, New Zealand*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3696409.3700183

1 Introduction

In audio recordings or speech signals, speech emotion recognition (SER) plays a pivotal role in identifying and understanding emotional states. As technology advances, SER has become increasingly vital in various fields, including signal processing, audio processing, and affective computing. Its widespread adoption is evident across diverse industries, such as e-learning, computer games, healthcare, human-computer interfaces, and human-robot interaction, where it facilitates enhanced communication and user experience. The expanding scope of SER applications is exciting, especially in evaluating virtual interviews, online learning environments, and instructional effectiveness.

SER typically follows two primary schemes: feature extraction and emotion classification. Traditionally, researchers relied on handcrafted features and machine learning-based classifiers to discern emotional states from speech signals, utilizing techniques such as linear discriminant analysis and support vector machine classifiers [7] or employing acoustic features with hierarchical decision tree classifiers [11].

With the advent of deep learning, SER has witnessed significant advancements. Researchers have developed end-to-end approaches that outperform traditional methods. These include deep echo state networks [6, 8], convolutional neural networks [14, 29], recurrent neural networks [15, 18], and attention mechanisms or transformers [9, 19]. Additionally, some studies have integrated the strengths of both traditional and deep learning approaches to enhance SER performance [12, 21].

Leveraging deep learning models in SER requires uniform preparation of inputs. Typically, this involves segmenting speech signals into fixed-length chunks. However, this approach may result in information loss, as crucial features might not be fully captured within the segment. To address this limitation, researchers have turned to multimodal approaches, combining information from both audio and text sources to enhance SER performance [10, 26, 28]. Nonetheless, many studies focus on concatenating multimodal features without aligning information between audio and text features.

Recently, Yang [24] employed HuBERT [5] in an ensemble learning approach for SER and achieved an accuracy of 70.24%. This demonstrates that HuBERT is suitable for downstream SER tasks despite its original design for speech representation learning. Adoma *et al.* [1] conducted a comparative study to assess the performance of pre-trained BERT [4], DistilBERT [17], and other models for text-based emotion recognition. The study revealed that the BERT-based model outperformed the DistilBERT-based model. However, the DistilBERT-based model, with fewer parameters than BERT, holds the potential for deploying SER applications on mobile and embedded devices. More recently, Wu *et al.* [23] designed a pre-training framework for audio representation learning utilizing a contrastive learning strategy. However, the mentioned framework was developed based on CLIP (contrastive language-image pre-training) [16], incorporating feature fusion and keyword-to-caption augmentation to accommodate various audio lengths, resulting in improved performance.

In this study, we developed a contrastive language-audio self-alignment pre-training approach for SER, utilizing well-known models for speech and text: HuBERT [5] and BERT [4]/DistilBERT [17]. HuBERT processed the audio input to extract audio embeddings, while BERT/DistilBERT processed the text input to derive text embeddings. These embeddings were then inputted into CLIP [16], a contrastive language-image pre-training framework, to align high-level feature representations from audio and text using the scaled pairwise cosine similarities and symmetric loss function based on the cross-entropy loss function.

Following pre-training, the pre-trained HuBERT-CLAP model was employed for downstream SER tasks. We trained and tested the proposed method on the IEMOCAP dataset. Although HuBERT-CLAP utilizes audio and text inputs for pre-training, only audio input is required during downstream SER tasks and inference, making it an unimodal SER model. Ablation and comparative analyses demonstrated that our method outperformed the latest state-of-the-art approach for SER on the same dataset. Furthermore, we conducted a case study to assess the transferability of models trained on the IEMOCAP dataset to the EmoDB dataset. In future research, we aim to employ the knowledge distillation framework to train a lighter model suitable for embedded devices and smartphones, facilitating widespread application in human-computer interactions and robotics.

The remaining sections of this paper are organized as follows: Section 2 summarizes the related research to this study. In Section 3, we introduce the proposed framework. Section 4 presents a detailed account of the experimental results and comparisons. Finally, Section 5 provides the conclusion of this study.

2 Related Work

2.1 Foundation Models

Several well-established models have recently emerged in the natural language processing (NLP) and speech processing domains, including BERT, DistillBERT, and HuBERT. These models serve as the backbone for many language-based and speech-based tasks, and they are briefly summarized below.

HuBERT [5] or Hidden-Unit BERT is a self-supervised method for speech representation learning that uses offline clustering to generate target labels, which are then applied in a BERT-like prediction task. HuBERT addresses challenges such as multiple sound units per utterance, the absence of a sound unit lexicon during pre-training, and the variable lengths of sound units without explicit segmentation. By focusing prediction loss only on masked regions, HuBERT encourages the model to learn both acoustic and linguistic features from the continuous speech input. Additionally, it prioritizes consistency in the clustering process, helping the model learn robust representations from speech data.

BERT [4], or Bidirectional Encoder Representations from Transformers, is a widely used language model known for its strong language understanding. Built on a bidirectional Transformer encoder, BERT is pre-trained on unlabeled text by considering both left and right context. After pretraining, it can be fine-tuned with a superficial output layer to excel in downstream tasks, such as question answering, sentiment analysis, and multimodal emotion recognition.

DistilBERT [17] is a compact, efficient, and lightweight Transformer model created through knowledge distillation during BERT's pretraining phase. With 40% fewer parameters than the BERT base model, it offers a 60% increase in speed while retaining over 97% of BERT's performance, as demonstrated on the GLUE language understanding benchmark.

2.2 Multimodal Emotion Recognition

Multimodal emotion recognition has advanced significantly in recent years, with researchers exploring various audio, text, and visual input combinations. Studies such as [10, 26–28] demonstrate that integrating speech and text improves emotion recognition accuracy by utilizing the complementary information from both modalities. For example, Yoon *et al.*[26] enhanced emotion recognition systems' performance by using audio and text as input modalities. Zhang*et al.*[27] further improved the fusion of audio and text features through a hybrid attention network. Similarly, Zhao*et al.* [28] extracted audio and text embeddings and fed them into a long short-term memory (LSTM) module, followed by a sliding window attention mechanism to capture inter-modal interactions. The resulting feature representations were then processed through a multi-level high-response feature reuse module before being passed to the classification head for final emotion recognition. Despite these advancements, most studies rely on simple concatenation of multimodal features without effectively aligning the information between audio and text modalities.

3 Methodology

Our proposed HuBERT-CLAP framework operates through two distinct stages, as illustrated in **Figure 1**. In the initial stage (Stage 1), we partially fine-tune HuBERT to process audio inputs, following the strategy outlined in [22]. Concurrently, we employ either DistilBERT to handle text inputs. Subsequently, CLIP is utilized to

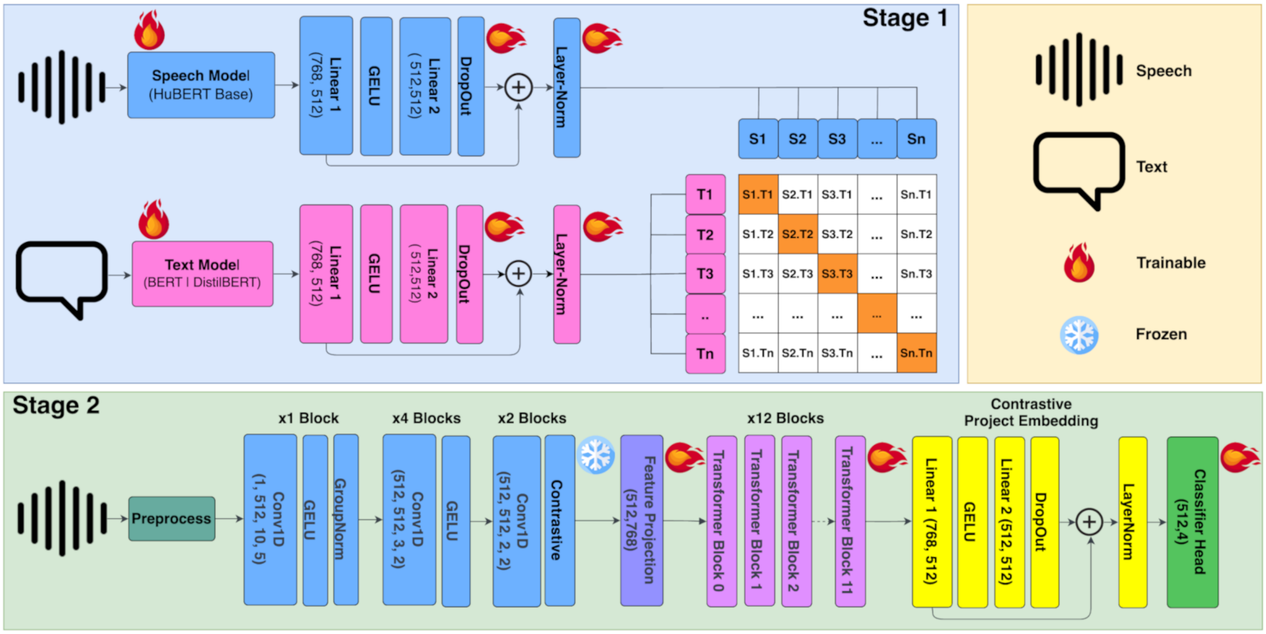


Figure 1: Architectural overview of the proposed HuBERT-CLAP, illustrating the integration of self-supervised speech representation learning and audio-text alignment for emotion recognition.

align the extracted audio and text features, with the primary objective of ensuring that features learned from audio inputs closely correspond with those derived from text inputs.

In stage 1, speech and text inputs are processed through the speech model (HuBERT) and text model (BERT/DistilBERT) to extract both modality cues. Furthermore, we processed these cues through a multilayer perceptron block with residual skip connections, followed by Layer Normalization to obtain embedding. Notably, a GELU (Gaussian Error Linear Unit) activation function is applied after the first linear layer, and dropout is applied after the second linear layer. Following the approach of CLIP [16] and the framework in [23], these text and audio embeddings are utilized to compute scaled pairwise cosine similarities and a symmetric loss function based on the cross-entropy loss function.

In Stage 2, the pre-trained HuBERT-CLAP model from Stage 1 is harnessed for the downstream emotion recognition task. It is important to note that only the HuBERT architecture within the pre-trained HuBERT-CLAP undergoes partial fine-tuning in this stage. We partially fine-tune the feature projection layer, transformer blocks, contrastive project embedding block, and classifier head for final emotional classification. Meanwhile, the feature extractor blocks remain frozen without updating the parameters. Finally, the features obtained by the classifier head are utilized to generate the final prediction using cross-entropy with the softmax function.

In our two-stage approach, audio and text features were aligned in the pre-trained HuBERT-CLAP model, and selected components were fine-tuned to specific downstream tasks within the pre-trained model. The initial stage involves aligning the extracted features from speech and text inputs, ensuring they complement each other effectively. This alignment process facilitates the creation of a unified representation that captures both modalities' relevant information. Subsequently, in the second stage, we leverage the pre-trained model to refine this representation for the emotion recognition task. To preserve the overall capabilities of the model, we only fine-tune specific components, such as the feature projection layer and the classifier head. As a result of this fine-tuning, the model is more accurate at classifying emotions from speech inputs. In SER, our approach maximizes the utility of pre-trained models.

4 Results and Discussion

4.1 Experimental Setup

4.1.1 Dataset. IEMOCAP is a well-known English emotional dataset published by Busso et al. [3] in 2008. It consists of recordings from ten performers, five male and five female, sampled at 48 kHz. Initially, there were nine different emotional categories in the dataset. However, they were merged due to insufficient utterances in specific categories; they were merged, resulting in four main categories: 1,103 utterances for anger, 1,635 utterances for pleasure, 1,708 utterances for neutrality, and 1,084 utterances for sorrow. Consequently, these four emotional categories have become the standard for comparison in a wide range of research. The distribution percentage of each emotional state in the IEMOCAP dataset is shown in **Figure 2**.

4.1.2 Implementation details. In the experiments, all audio samples from the IEMOCAP dataset were converted to mono channel and truncated to 8 seconds in length for both contrastive pre-training and downstream SER task training. No additional audio augmentation was applied. The PyTorch framework [13] was utilized to implement the proposed method.

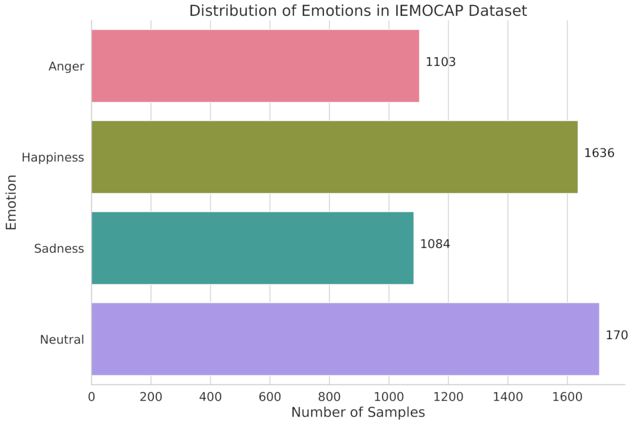


Figure 2: The distribution of emotion labels, displaying the number of occurrences for each emotion category in the IEMOCAP dataset.

In Stage 1, we utilized the pre-trained HuBERT[1], BERT[2], and DistilBERT[3] for pre-training. We utilize the CosAngularGrad optimizer with the CosineAnnealingLR scheduler. The learning rate (LR) is set to 1e-4 for HuBERT and 1e-5 for BERT and DistilBERT, with a minimum LR of 1e-6. We conduct training for 50 epochs with a batch size of 16, a temperature of 1, a precision set to BF16, and a dropout rate of 0.2.

In Stage 2, we employ the same optimizer, scheduler, and LR as in Stage 1. However, the number of epochs is set to 30, and the precision is 32-bit. Additionally, to prevent overfitting and train more generalized models, we implement 5-fold cross-validation and incorporate a dropout rate of 0.2.

To evaluate the effectiveness of the suggested approach, we employ weighted accuracy (WA) as the primary metric due to the imbalanced nature of the IEMOCAP dataset. Additionally, we utilize t-SNE [20] to visualize the learned feature representations of the proposed method.

4.2 Ablation Study

To validate the impacts and effectiveness of the contrastive pre-training approach, we conducted an ablation study in three cases, as illustrated in **Table 1**. Corresponding confusion matrices are depicted in **Figures 3a, 3b, and 3c**. Employing the contrastive pre-training approach yields superior results compared to not using it. Furthermore, utilizing BERT for text embedding features outperforms DistilBERT. The model incorporating HuBERT and BERT pre-training achieves the highest performance with a WA of 77.22%. Consequently, we designate this model as HuBERT-CLAP for subsequent comparisons. Specifically, HuBERT-CLAP outperforms the other two models by 0.31 to 7.18%.

Table 1: Ablation study comparing different architectures on the IEMOCAP dataset for emotion recognition.

Model	WA (%)
Only HuBERT without pre-training	70.04
HuBERT + DistilBERT pre-training	76.91
HuBERT + BERT pre-training	**77.22**

Table 2: Comparison of HuBERT-CLAP with recent state-of-the-art methods on the IEMOCAP dataset.

Model	Year	WA (%)
CNN+Bi-GRU [29]	2020	70.39
SPU+MSCNN [14]	2021	66.60
LightSER [2]	2022	70.23
TIM-Net [25]	2023	71.65
HuBERT-CLAP (Our)	**2024**	**77.22**

4.3 Feature Analysis

To assess the effectiveness of the contrastive pre-training approach compared to not using it, we utilized t-SNE [20] to visualize the distribution of emotions in the learned feature representations of the three models employed in the ablation study. As shown in **Figure 4**, the feature representations learned by the pre-training models exhibit greater separation than those without pre-training. Specifically, the DistilBERT model demonstrates more overlapping points between the four emotional classes than the one using BERT. This observation aligns with the findings from Adoma's study [1]. Based on the results of the pre-training stage, HuBERT-CLAP features are well aligned with text features extracted from BERT, indicating that the model can learn robust feature representations during pre-training, which can then be generalized during inference on new data, as shown in the previous section with the case study.

4.4 Performance Comparison

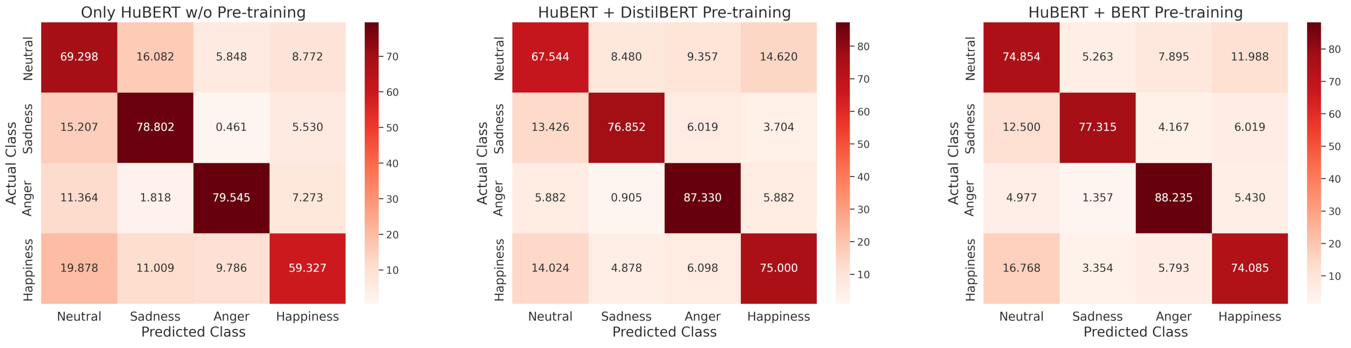
Furthermore, to validate the enhanced performance of HuBERT-CLAP, we compare it with the most recent state-of-the-art methods on the IEMOCAP dataset, as detailed in **Table 2**. Results depicted in **Table 2** indicate that HuBERT-CLAP surpasses CNN+Bi-GRU [29], SPU+MSCNN [14], LightSER [2], and TIM-Net [25] in terms of WA, with gains of 6.83%, 10.62%, 6.99%, and 5.57%, respectively.

4.5 Case Study

To further assess the effectiveness and transferability of the proposed method, we evaluated the three models trained on the IEMOCAP dataset using the EmoDB dataset. The original EmoDB dataset consists of 535 samples covering seven emotions. However, to align with the models trained on the IEMOCAP dataset, we selected four similar emotions from the EmoDB for this study: anger (127 utterances), happiness (71 utterances), neutral (79 utterances), and sadness (62 utterances). It is important to note that the three models employed in this case study were trained exclusively on the IEMOCAP dataset. The results are presented in **Table 3**.

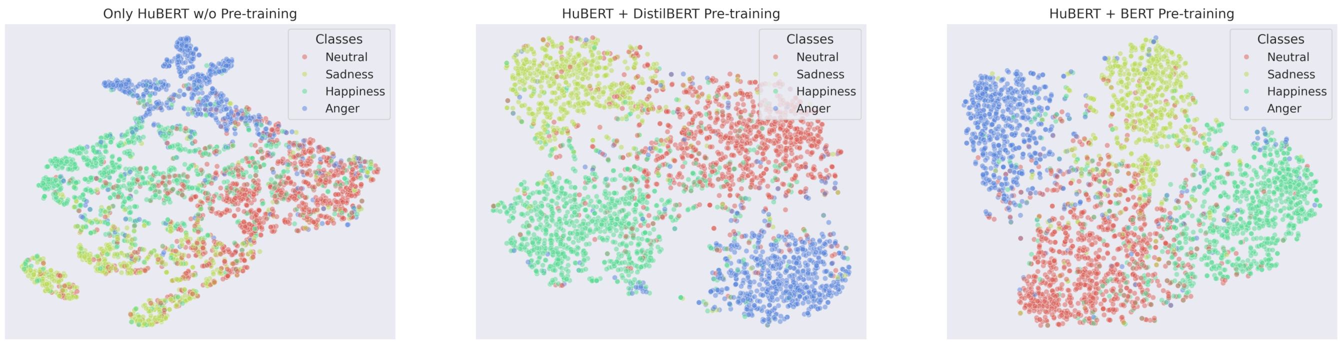
As shown in **Table 3**, the performance trends are similar to those observed in the IEMOCAP dataset. Additionally, the pre-training

[1] https://huggingface.co/facebook/hubert-base-ls960
[2] https://huggingface.co/google-bert/bert-base-uncased
[3] https://huggingface.co/distilbert/distilbert-base-uncased



(a) **(b)** **(c)**

Figure 3: Confusion matrices illustrating the performance of the proposed models on the IEMOCAP dataset: (a) HuBERT without pre-training, (b) HuBERT with DistilBERT pre-training, and (c) HuBERT with BERT pre-training for the downstream SER task. The diagonal values represent the recall for each class.



(a) **(b)** **(c)**

Figure 4: Class-level t-SNE visualization of the proposed models on the IEMOCAP dataset: (a) HuBERT without pre-training, (b) HuBERT with DistilBERT pre-training, and (c) the proposed HuBERT with BERT pre-training for the downstream SER task.

Table 3: Case study on the EmoDB dataset, evaluated using models trained on the IEMOCAP dataset.

Model	WA (%)
Only HuBERT without pre-training	60.02
HuBERT + DistilBERT pre-training	74.15
HuBERT + BERT pre-training	**79.00**

approach outperforms the non-pre-training method, with the model utilizing both HuBERT and BERT pre-training achieving the highest performance, with a WA of 79.00%.

5 Conclusion

This study developed HuBERT-CLAP, a framework that enhances SER performance through contrastive language-audio self-alignment pre-training. We use HuBERT and BERT/DistilBERT to extract audio and text embeddings within a contrastive pre-training framework. The primary objective is to align the features learned by HuBERT from audio inputs with those obtained by BERT/DistilBERT from text inputs. Subsequently, the pre-trained HuBERT from the initial stage is utilized for the downstream emotion recognition task.

Experimental results on the IEMOCAP dataset reveal that HuBERT-CLAP achieves superior performance compared to approaches that do not incorporate the contrastive pre-training technique. Moreover, it surpasses the latest state-of-the-art methods on the same dataset. In future research endeavors, we aim to explore the knowledge distillation framework to develop a more lightweight model suitable for deployment on embedded devices and smartphones. This advancement would facilitate broader human-computer interactions and robotics applications, enhancing user experiences across various domains.

References

[1] Acheampong Francisca Adoma, Nunoo-Mensah Henry, et al. 2020. Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. In *In proceeding of ICCWAMTIP-2020*. IEEE, 117–121.

[2] Arya Aftab, Alireza Morsali, et al. 2022. LIGHT-SERNET: A Lightweight Fully Convolutional Neural Network for Speech Emotion Recognition. In *In Proceeding of ICASSP 2022, Virtual and Singapore, May 23-27*. IEEE, 6912–6916. https://doi.org/10.1109/ICASSP43922.2022.9746679

[3] Carlos Busso, Murtaza Bulut, et al. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Evaluation* 42, 4 (2008), 335–359. https://doi.org/10.1007/S10579-008-9076-6

[4] Jacob Devlin, Ming-Wei Chang, et al. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1*. 4171–4186. https://doi.org/10.18653/V1/N19-1423

[5] Wei-Ning Hsu, Benjamin Bolte, et al. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE ACM Trans. Audio Speech Lang. Process.* 29 (2021), 3451–3460. https://doi.org/10.1109/TASLP.2021.3122291

[6] Hemin Ibrahim, Chu Kiong Loo, et al. 2022. Bidirectional parallel echo state network for speech emotion recognition. *Neural Comput. Appl.* 34, 20 (2022), 17581–17599. https://doi.org/10.1007/S00521-022-07410-2

[7] J Indra, R Kiruba Shankar, et al. 2022. Speech Emotion Recognition Using Support Vector Machine and Linear Discriminant Analysis. In *International Conference on Intelligent Systems Design and Applications*. Springer, 482–492.

[8] Mustaqeem Khan, Abdulmotaleb, et al. 2023. AAD-Net: Advanced end-to-end signal processing system for human emotion detection & recognition using attention-based deep echo state network. *Knowl. Based Syst.* 270 (2023), 110525. https://doi.org/10.1016/J.KNOSYS.2023.110525

[9] Mustaqeem Khan, Wail Gueaieb, et al. 2024. MSER: Multimodal speech emotion recognition using cross-attention with deep fusion. *Expert Systems with Applications* 245 (2024), 122946.

[10] Puneet Kumar, Vishesh Kaushik, et al. 2021. Towards the Explainability of Multimodal Speech Emotion Recognition.. In *In Proceeding of Interspeech-2021*. 1748–1752.

[11] Chi-Chun Lee, Emily Mower, et al. 2011. Emotion recognition using a hierarchical binary decision tree approach. *Speech Commun.* 53, 9-10 (2011), 1162–1171. https://doi.org/10.1016/J.SPECOM.2011.06.004

[12] Seong-Gyun Leem, Daniel Fulford, et al. 2024. Selective Acoustic Feature Enhancement for Speech Emotion Recognition With Noisy Speech. *IEEE ACM Trans. Audio Speech Lang. Process.* 32 (2024), 917–929. https://doi.org/10.1109/TASLP.2023.3340603

[13] Adam Paszke, Sam Gross, Francisco Massa, et al. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *In Proceeding of NeurIPS 2019, December 8-14, Vancouver, BC, Canada*. 8024–8035.

[14] Zixuan Peng, Yu Lu, et al. 2021. Efficient Speech Emotion Recognition Using Multi-Scale CNN and Attention. In *In proceeding of ICASSP 2021, Toronto, ON, Canada, June 6-11*. IEEE, 3020–3024. https://doi.org/10.1109/ICASSP39728.2021.9414286

[15] Nhat Truong Pham, Duc Ngoc Minh Dang, et al. 2023. Hybrid data augmentation and deep attention-based dilated convolutional-recurrent neural networks for speech emotion recognition. *Expert Syst. Appl.* 230 (2023), 120608. https://doi.org/10.1016/J.ESWA.2023.120608

[16] Alec Radford, Jong Wook Kim, et al. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings ICML 2021, 18-24 July 2021, Virtual Event*, Vol. 139. PMLR, 8748–8763.

[17] Victor Sanh, Lysandre Debut, et al. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR* abs/1910.01108 (2019). arXiv:1910.01108

[18] Fei Tao and Gang Liu. 2018. Advanced LSTM: A Study About Better Time Dependency Modeling in Emotion Recognition. In *In Proceeding of ICASSP 2018, Calgary, AB, Canada, April 15-20*. IEEE, 2906–2910. https://doi.org/10.1109/ICASSP.2018.8461750

[19] Lorenzo Tarantino, Philip N. Garner, et al. 2019. Self-Attention for Speech Emotion Recognition. In *Interspeech 2019, Graz, Austria, 15-19 September*. ISCA, 2578–2582. https://doi.org/10.21437/INTERSPEECH.2019-2822

[20] Laurens Van der Maaten et al. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[21] Bogdan Vlasenko, Ravi Shankar Prasad, et al. 2021. Fusion of Acoustic and Linguistic Information using Supervised Autoencoder for Improved Emotion Recognition. In *In Proceedings of ACM, Virtual Event, China, 24 October*. ACM, 51–59. https://doi.org/10.1145/3475957.3484448

[22] Yingzhi Wang, Abdelmoumene Boumadane, et al. 2021. A Fine-tuned Wav2vec 2.0/HuBERT Benchmark For Speech Emotion Recognition, Speaker Verification and Spoken Language Understanding. *CoRR* abs/2111.02735 (2021). arXiv:2111.02735

[23] Yusong Wu, Ke Chen, et al. 2023. Large-Scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation. In *In proceeding of ICASSP 2023, Rhodes Island, Greece, June 4-10*. IEEE, 1–5. https://doi.org/10.1109/ICASSP49357.2023.10095969

[24] Janghoon Yang. 2023. Ensemble deep learning with HuBERT for speech emotion recognition. In *In proceeding of ICSC-2023, Laguna Hills, CA, USA, February 1-3*. IEEE, 153–154. https://doi.org/10.1109/ICSC56153.2023.00032

[25] Jiaxin Ye, Xin-Cheng Wen, et al. 2023. Temporal Modeling Matters: A Novel Temporal Emotional Modeling Approach for Speech Emotion Recognition. In *In proceeding of ICASSP 2023, Rhodes Island, Greece, June 4-10*. IEEE, 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096370

[26] Seunghyun Yoon, Seokhyun Byun, et al. 2018. Multimodal speech emotion recognition using audio and text. In *In Proceeding of SLT-2018*. IEEE, 112–118.

[27] Shiqing Zhang, Yijiao Yang, Chen Chen, Ruixin Liu, Xin Tao, Wenping Guo, Yicheng Xu, and Xiaoming Zhao. 2023. Multimodal emotion recognition based on audio and text by using hybrid attention networks. *Biomedical Signal Processing and Control* 85 (2023), 105052.

[28] Ziping Zhao, Tian Gao, et al. 2023. SWRR: Feature Map Classifier Based on Sliding Window Attention and High-Response Feature Reuse for Multimodal Emotion Recognition. In *Proc. INTERSPEECH-2023*, Vol. 2023. 2433–2437.

[29] Ying Zhong, Ying Hu, et al. 2020. A Lightweight Model Based on Separable Convolution for Speech Emotion Recognition. In *Interspeech 2020, Virtual Event, Shanghai, China, 25-29 October*. ISCA, 3331–3335. https://doi.org/10.21437/INTERSPEECH.2020-2408