Emotion Detection in Speech Using Lightweight and Transformer-Based Models: A Comparative and Ablation Study

1st Lucky Onyekwelu-Udoka
Electrical and Computer Engineering
Iowa State University
Ames, USA
Lucky@iastate.edu

2ndMd Shahedul Hasan
Electrical and Computer Engineering
Iowa State University
Ames, USA
shahedul@iastate.edu

3rd Md Shafiqul Islam
Electrical and Computer Engineering
Iowa State University
Ames, USA
shafiqul@iastate.edu

***Abstract*—Emotion recognition from speech plays a vital role in the development of empathetic human-computer interaction systems. This paper presents a comparative analysis of lightweight transformer-based models, DistilHuBERT and PaSST, by classifying six core emotions from the CREMA-D dataset. We benchmark their performance against a traditional CNN-LSTM baseline model using MFCC features. DistilHuBERT demonstrates superior accuracy (70.64%) and F1 score (70.36%) while maintaining an exceptionally small model size (0.02 MB), outperforming both PaSST and the baseline.**

Furthermore, we conducted an ablation study on three variants of the PaSST, Linear, MLP, and Attentive Pooling heads, to understand the effect of classification head architecture on model performance. Our results indicate that PaSST with an MLP head yields the best performance among its variants but still falls short of DistilHuBERT. Among the emotion classes, *angry* is consistently the most accurately detected, while *disgust* remains the most challenging.

These findings suggest that lightweight transformers like DistilHuBERT offer a compelling solution for real-time speech emotion recognition on edge devices. The code is available at: https://github.com/luckymaduabuchi/Emotion-detection-. The open source code and modular experiments double as a teaching toolkit, enabling instructors to demonstrate transformer fine-tuning and ablation methods in upper-level machine learning courses.

***Index Terms*—Speech Emotion Recognition, Transformer, DistilHuBERT, PaSST.**

I. Introduction

Emotion detection from speech has become an increasingly vital area of research, with applications spanning intelligent virtual assistants, affective computing, mental health monitoring, and immersive virtual environments [10], [11]. As human-computer interactions become more natural and personalized, the demand for systems capable of interpreting emotional signals in real time has increased. Emotion-aware systems enable machines to respond empathetically to users, adjust responses based on sentiment, and improve user experience through customized feedback mechanisms [16]. In domains such as telemarketing, adaptive education, and therapeutic interventions, emotion detection empowers analytical tools that optimize engagement and emotional relevance [15].

Emotion signals are typically conveyed through three modalities: facial expressions, physiological signals, and vocal audio. Among these, vocal audio presents both a rich source of emotional information and a challenging recognition problem. Compared to image-based facial cues, speech provides more dynamic, personalized, and nuanced emotional content [17]. However, the complexity of speech, driven by factors such as tone, prosody, speaker identity, and conversational context, makes emotion recognition from audio an open-ended machine learning challenge. The feature extraction process, the choice of representation (e.g. MFCC vs spectrograms), and the model architecture significantly influence the system's ability to reliably decode emotions [18].

Earlier approaches relied heavily on statistical methods such as Gaussian Mixture Models with Universal Background Models (GMM-UBM) and hybrid classifiers such as GMM-DNN [20], [21]. Although effective for constrained settings, these models struggled to scale to large, diverse datasets due to limitations in sequential modeling and robustness to noise. The introduction of deep learning models,especially CNN-LSTM architectures using hand-crafted features such as MFCCs, marked a turning point, improving both performance and temporal modeling. Ouyang et al. [22] demonstrated such improvements using a CNN-LSTM pipeline on MFCC-transformed speech data, achieving an accuracy of 61.07%.

More recently, transformer-based architectures have revolutionized speech representation learning. Self-supervised models such as DistilHuBERT leverage layer-wise knowledge distillation to offer high accuracy with minimal computational overhead [23]. In parallel, PaSST [24], designed for efficient audio classification, introduces spectrogram patching and patchout techniques to generalize effectively. Despite these advancements, a comprehensive comparison of lightweight transformers and classical baselines under consistent training and evaluation conditions remains lacking in emotion recognition tasks.

In this study, we present a comparative analysis of DistilHuBERT, PaSST, and a CNN-LSTM baseline for the classification of speech emotions using the CREMA-D dataset.

Furthermore, we conduct an ablation study on PaSST configurations to understand how architectural variations (linear and attention vs. MLP heads) and how raw audio vs. spectrogram input impact performance. Our study aims to determine the most effective and efficient model for real-time deployment, highlighting both strengths and bottlenecks in contemporary SER architectures.

This work also serves as a practical teaching tool for audio-based machine learning. The models and open-source code can be incorporated into advanced undergraduate or graduate level coursework in machine learning, speech processing, or human-centered AI. This makes the study an educational and applied contribution.

Primary learning goals include understanding and comparing different model architectures such as CNN-LSTM and Transformer-based systems for speech emotion recognition, evaluating model performance using accuracy, precision, recall, and F1 score, and analyzing the effect of architectural changes through ablation studies. Information on how input representation, whether raw waveform or spectrogram, affects classification outcomes in audio-based tasks.

To effectively engage with this study, you should have prior knowledge of deep learning fundamentals, including neural network architectures, transformers, and backpropagation. A background in audio signal processing, particularly techniques such as MFCC and spectrogram generation, is essential. Additionally, a solid understanding of model evaluation metrics and familiarity with Python and PyTorch will support hands-on exploration of the models and training routines presented.

Related Work

Early work in speech emotion recognition (SER) leveraged primarily hand-crafted acoustic features such as Mel frequency cepstral coefficients (MFCC) or spectrograms, which were input to deep learning architectures like Convolutional Neural Networks (CNN) or Recurrent Neural Networks (RNN) [20], [21]. CNNs, in particular, were effective in capturing localized spectral-temporal patterns. Fayek et al. used deep CNNs to classify emotions from spectrogram images, achieving a precision of around 60% on the SAVEE corpus [29]. Later, more advanced CNN architectures improved this, especially when combined with global pooling layers or augmented feature sets, achieving up to 70% accuracy in datasets like RAVDESS [30]. RNNs, especially long-short-term memory networks (LSTM), were also widely adopted due to their ability to model temporal dependencies in sequential speech data [32]. LSTM networks operating on MFCC sequences demonstrated performance comparable to CNNs, particularly in modeling prosodic features such as rhythm and pitch contours. Some architectures combined CNN and RNN modules, CNNs to extract spatial features and RNNs to model temporal dynamics, achieving enhanced performance [31]. For example, Trigeorgis *et al.* proposed an end-to-end convolutional recurrent network learning directly from raw waveforms [25].

The attention mechanisms further improved these models. By integrating attention layers on top of LSTMs or CNNs, models could focus on the most emotionally salient parts of an utterance. Mountzouris *et al.* achieved more than 74% accuracy on SAVEE and 77% on RAVDESS using CNN-attention hybrids [12]. However, despite these gains, CNN and RNN-based models often struggled with generalization due to limited dataset sizes and speaker variability, prompting a shift toward self-supervised and pre-trained models [14], [26].

Transformer models have more recently become prominent in SER due to their ability to model long-range dependencies and benefit from large-scale pretraining. Among them, wav2vec 2.0 is a leading self-supervised model trained in raw audio using contrastive learning [26]. It consists of a convolutional encoder followed by a Transformer that captures contextual dependencies. Pepino *et al.* demonstrated that fine-tuned wav2vec 2.0 models outperform previous CNN/LSTM models, achieving up to 73% accuracy on IEMOCAP [26].

HuBERT (Hidden Unit BERT), another Transformer-based model, differs by using masked prediction of cluster-based units derived from acoustic features [27]. Fine-tuned HuBERT models have shown even higher SER accuracy, reaching up to 79.6% on IEMOCAP and exhibiting strong performance on individual emotions such as anger and fear.

To reduce computational complexity, DistilHuBERT was proposed as a distilled version of HuBERT [23]. It compresses the model by 75% and accelerates inference while maintaining competitive performance, making it ideal for real-time applications.

Another line of work uses Transformers on spectrogram images. The Audio Spectrogram Transformer (AST) and its efficient variant PaSST (Patchout Spectrogram Transformer) apply the Vision Transformer (ViT) framework to audio spectrograms [24]. PaSST incorporates patchout regularization, which randomly drops time/frequency patches during training, reducing memory usage, and acting as augmentation. These models have achieved strong results on AudioSet and have been adapted for SER tasks.

The evolution from CNN/RNN models to Transformer-based architectures has significantly improved SER accuracy, robustness, and efficiency. Transformer models benefit from self-attention, allowing them to capture both global and fine-grained prosodic features. Pre-training on large speech corpora enables better generalization even on smaller SER datasets.

While large Transformers like wav2vec 2.0 and HuBERT deliver superior performance, they are computationally intensive. Models like DistilHuBERT and PaSST strike a balance between accuracy and efficiency, making them practical for deployment.

Problem Definition

Given a raw audio signal $\mathbf{x}(t)$, the task of Speech Emotion Recognition (SER) is to classify the signal into one of K discrete emotion classes:

$$\mathcal{Y} = \{\text{happy}, \text{sad}, \text{angry}, \text{fear}, \text{disgust}, \text{neutral}\}$$

Let $\mathbf{x} \in \mathbb{R}^T$ denote a speech waveform of duration T, and let $f_\theta : \mathbb{R}^T \to \mathbb{R}^K$ be a parameterized model (e.g., a

Transformer-based or CNN-based architecture). The goal is to learn the mapping:

$$\hat{\mathbf{y}} = f_\theta(\mathbf{x})$$

where $\hat{\mathbf{y}} \in \mathbb{R}^K$ is the predicted probability distribution over the emotion classes, and the final predicted label is:

$$\hat{y} = \arg\max_i \hat{y}_i$$

Training is performed by minimizing the categorical cross-entropy loss between the predicted distribution $\hat{\mathbf{y}}$ and the ground truth label $\mathbf{y} \in \{0, 1\}^K$:

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{i=1}^{K} y_i \log(\hat{y}_i)$$

This paper addresses the problem of identifying the most accurate and efficient model architecture for real-time SER, under consistent training and evaluation conditions. Specifically, we aim to:

- Compare the performance of a lightweight self-supervised model (DistilHuBERT), a spectrogram-based Transformer (PaSST), and a CNN-LSTM baseline.
- Evaluate the effect of different classification heads in PaSST: Linear, MLP, and Attentive Pooling.
- Identify which model offers the best trade-off between accuracy, inference time, and memory efficiency on the CREMA-D dataset.

By benchmarking these models and configurations, we seek to provide insights into optimal architectures for practical deployment of SER systems on resource-constrained devices.

Methodology

The methodology in this article is designed to be accessible and instructive to a broad audience interested in machine learning and speech processing. The model choices reflect different levels of abstraction and learning paradigms: CNN-LSTM illustrates sequential modeling from engineered features (MFCC), DistilHuBERT demonstrates self-supervised representation learning directly from waveforms, and PaSST showcases transformer-based architectures for image-like inputs such as spectrograms. These models provide a balanced overview of both classical and contemporary approaches.

This setup allows practitioners and learners alike to explore not only how different models perform but also why some models may generalize better under certain input representations or architectural designs.

Dataset and Preprocessing

The Crowd-sourced Multimodal Emotional Actors Dataset (CREMA-D) is used in this study. It contains 7,442 audio clips from 91 actors who speak 12 sentences in six basic emotional states. Anger, Disgust, Fear, Happy, Neutral, and Sad. The data set provides diverse speakers in terms of age, gender, and ethnicity, making it suitable for training robust emotion recognition models [34].

Each audio file in the CREMA-D dataset is loaded at a target sampling rate of 16 kHz and clipped or padded to a maximum duration of 10 seconds. During training, several forms of data augmentation are applied to improve model generalization. These include random gain adjustment where a gain between -6 dB and +6 dB is applied, additive Gaussian noise to simulate background interference, pitch shifting simulated through resampling to slightly higher or lower sampling rates and then converting back to 16 kHz, and random time shifting by circularly rolling the waveform forward or backward in time. These operations are applied probabilistically and independently. All audio waveforms are normalized and returned along with their categorical emotion label for supervised learning.

Models and Implementation

The baseline model used in this study is a CNN-LSTM hybrid architecture that operates on Mel frequency cepstral coefficients (MFCC) as input features. Reproduced from Ouyang et al. [22], the model consists of four convolutional layers followed by three LSTM layers and a fully connected classification head, achieving an accuracy of 61.07% on the CREMA-D dataset.

The 2D convolutional stack captures local spectral and temporal features, while the bidirectional LSTM layers model sequential dependencies in the speech signal. This combination allows the network to learn both spatial and temporal patterns, making it a strong and well-established classical baseline for speech emotion recognition tasks.

DistilHuBERT: DistilHuBERT is a lightweight, distilled version of the HuBERT speech model. It comprises a convolutional feature extractor and a 2-layer Transformer encoder, distilled from a 12-layer HuBERT model using layer-wise knowledge distillation [33]. The model takes raw waveforms as input and outputs embeddings representing phonetic and prosodic information. For classification, a linear head is attached to the CLS token representation or the mean of the hidden states.

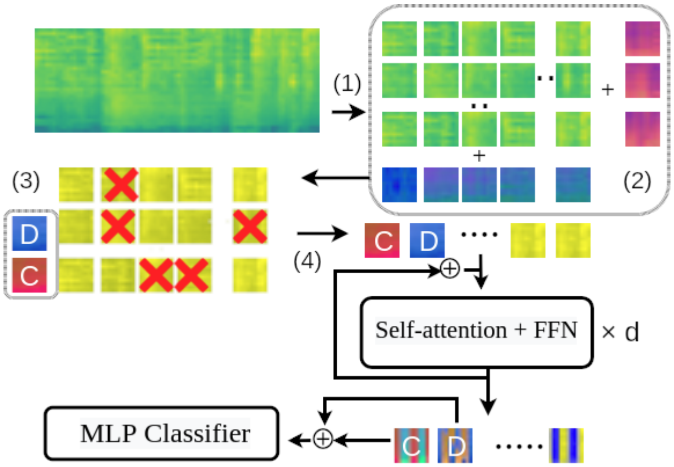


Fig. 3: PaSST

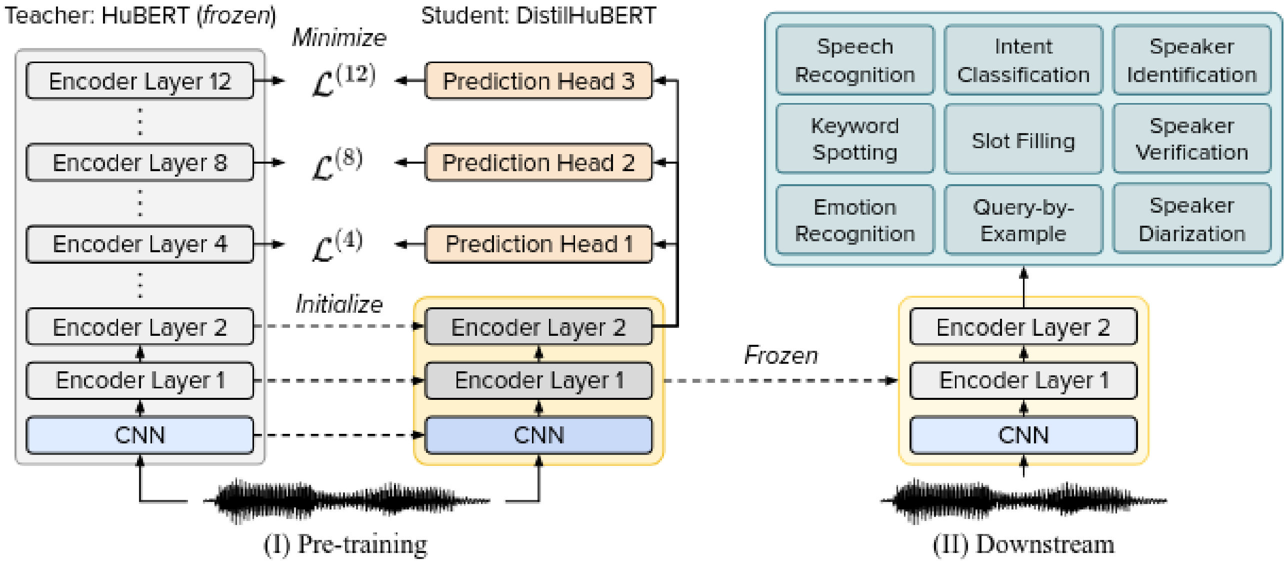


Fig. 1: DistilHuBERT architecture overview.

Patchout Spectrogram Transformer: The Patchout Spectrogram Transformer (PaSST) adapts the Vision Transformer (ViT) framework to audio spectrograms. Input spectrograms are divided into fixed-size patches that are flattened and projected into an embedding space. Two positional embeddings-time and frequency are added, and the sequence is passed through 12 transformer blocks with multihead self-attention and MLP layers.

PaSST introduces a regularization technique called *patchout*, which randomly drops time-frequency patches during training, acting as both a regularizer and an augmentation. For classification, both linear and MLP heads are tested [35].

Ablation Study Setup: To assess how architectural variations affect the performance of the PaSST model, an ablation study is conducted by experimenting with different classification heads and training configurations. All experiments use the pretrained `passt_s_swa_p16_128_ap476` backbone with patchout enabled for regularization. The specific configurations include:

Linear Head: This is the PaSST configuration, where classification is performed by applying a single linear transformation to the output of the [CLS] token. Let $\mathbf{h}_{\text{cls}} \in \mathbb{R}^d$ be the CLS embedding, then the logits are computed as:

$$\mathbf{z} = \mathbf{W}\mathbf{h}_{\text{cls}} + \mathbf{b}, \quad \mathbf{z} \in \mathbb{R}^K$$

Only the final transformer block (block 11) and the classifier layer are frozen for fine-tuning.

MLP Head: A two-layer feedforward network is applied to $\mathbf{h}_{\text{cls}}$, consisting of LayerNorm, ReLU, Dropout, and a linear output layer. The formulation is:

$$\mathbf{h}_1 = \text{ReLU}(\mathbf{W}_1 \cdot \text{LayerNorm}(\mathbf{h}_{\text{cls}}) + \mathbf{b}_1)$$

$$\mathbf{z} = \mathbf{W}_2 \cdot \text{Dropout}(\mathbf{h}_1) + \mathbf{b}_2$$

Here, $\mathbf{W}_1 \in \mathbb{R}^{256\times768}$, $\mathbf{W}_2 \in \mathbb{R}^{K\times256}$. The MLP head and the last two transformer blocks (blocks 10 and 11) are unfrozen during fine-tuning.

Attentive Pooling Head: Instead of using the CLS token, this configuration aggregates all token embeddings $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_T] \in \mathbb{R}^{T\times d}$ using attention weights:

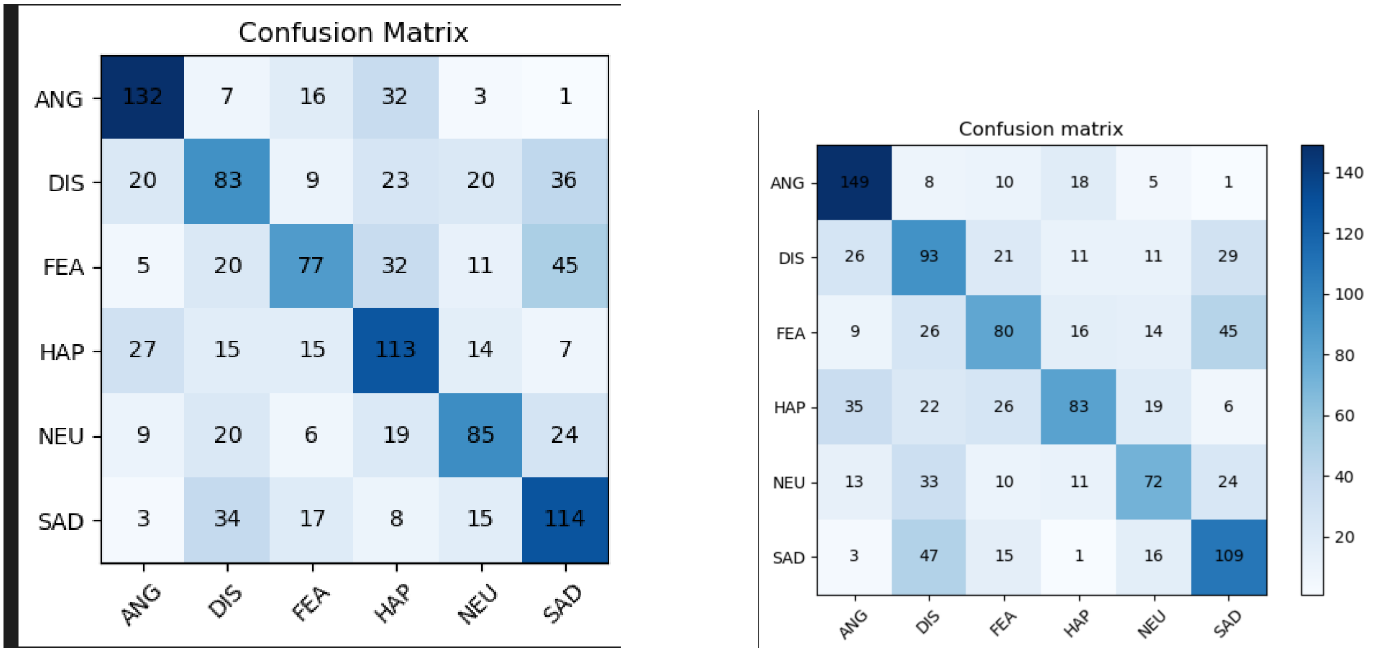
$$\alpha_t = \frac{\exp(\mathbf{w}_2^\top \tanh(\mathbf{W}_1\mathbf{h}_t))}{\sum_{j=1}^{T} \exp(\mathbf{w}_2^\top \tanh(\mathbf{W}_1\mathbf{h}_j))}$$

$$\boldsymbol{\mu} = \sum_{t=1}^{T} \alpha_t \mathbf{h}_t, \quad \boldsymbol{\sigma} = \sqrt{\sum_{t=1}^{T} \alpha_t (\mathbf{h}_t - \boldsymbol{\mu})^2}$$

$$\mathbf{z} = \mathbf{W}[\boldsymbol{\mu}; \boldsymbol{\sigma}] + \mathbf{b}$$

The attention module and the last two transformer blocks are trainable in this setup.

Figure 4 shows the training and evaluation pipeline for speech emotion recognition using DistilHuBERT and PaSST. The process begins with the CREMA-D dataset, where the audio samples undergo data augmentation. For DistilHuBERT, raw waveforms are processed by a CNN and Transformer backbone to extract contextual embeddings. In contrast, PaSST transforms audio into Mel spectrograms, applies patchout regularization, and forwards the result through a transformer. PaSST models use configurable classification heads: linear, MLP, or attentive pooling before proceeding to evaluation. Performance metrics are computed to assess the comparative effectiveness of both architectures.



(a) PaSST-MLP

(b) PaSST-Attention

Fig. 2: Confusion matrices of PaSST-MLP and PaSST-Attention showing per-emotion classification performance on the CREMA-D dataset.

Training and Evaluation

The CREMA-D dataset was randomly split into 70% training, 15% validation, and 15% test sets, ensuring speaker independence across the splits. This split strategy prevents data leakage and supports fair generalization evaluation.

All models were trained for 30 epochs using the Adam optimizer with a learning rate of 1×10^{-4}. A batch size of 16 was used and early stopping was employed based on validation accuracy with patience of 5 epochs.

Model performance was evaluated using accuracy, precision, recall, and the F1 score. In addition, the inference time per sample (in milliseconds) and the total size of the model (in megabytes) were reported to assess the feasibility of the deployment. Confusion matrices were generated to analyze class-wise recognition performance. All experiments were conducted on a single NVIDIA GPU.

Results and Comparative Analysis

Table I presents the overall performance metrics for all models evaluated in the CREMA-D test set. DistilHuBERT achieved the highest overall accuracy (70. 64%) and the F1 score (70. 36%), while requiring only 0.02 MB in size and maintaining competitive inference time, making it the most efficient and accurate among the evaluated models. Although it is a compressed version of HuBERT, its performance validates the effectiveness of knowledge distillation in preserving meaningful representations from raw audio.

Among the PaSST variants, the MLP head model performed the best (54. 07% accuracy), closely followed by the attentive pooling and linear head configurations. All variants shared the same input representation, Mel spectrograms, but differed in how the extracted Transformer features were aggregated and classified. In particular, the attentive pooling head, which summarizes temporal token features using learned statistical attention, outperformed the simpler linear projection. This challenges the notion that basic classification heads suffice when using spectrogram-based inputs, instead showing that expressive heads can extract more emotionally salient information.

In particular, not all transformer-based models outperformed traditional architectures. The CNN-LSTM baseline achieved 61. 07% precision, significantly surpassing all PaSST configurations. This result shows the strength of RNN-based temporal modeling and the value of simpler architectures, especially when dealing with moderately sized datasets such as CREMA-D. It also illustrates that the performance of a model depends not only on its architectural sophistication but also on its compatibility with the input representation and task-specific nuances.

Table III summarizes the accuracy of the classification per emotion. DistilHuBERT clearly excelled in recognizing high arousal emotions such as *Angry* (86.91%) and *Fear* (67.37%), showcasing its capacity to capture expressive variations in speech. For subtler emotions like *Neutral* and *Sad*, both CNN-LSTM and PaSST-MLP showed competitive performance, reflecting their potential to model more nuanced or flat affective tones.

The emotion that performed the worst in all models was *Disgust*, likely due to its low frequency of occurrence and ambiguous acoustic features. Surprisingly, the PaSST-MLP

Model	Accuracy	F1-score	Precision	Recall	Inf. Time (ms)	Size (MB)
DistilHuBERT	**70.64%**	**70.36%**	71.67%	70.64%	21.4	**0.02**
PaSST (MLP)	54.07%	53.82%	54.28%	54.07%	19.0	342.21
PaSST (Raw)	52.46%	52.05%	52.70%	52.46%	19.0	341.00
PaSST (Linear)	52.15%	51.29%	51.98%	51.75%	**18.5**	341.41
CNN-LSTM (Baseline)	61.07%	—	—	—	—	—

TABLE I: Model comparison on CREMA-D dataset

Configuration	Accuracy	F1-score	Notes
Linear Head	52.15%	51.29%	Minimal design using a single linear projection of the [CLS] token without additional non-linearity or pooling. Yields the lowest performance.
MLP Head	**54.07%**	**53.82%**	A two-layer feedforward network with ReLU activation and optional dropout. Applies LayerNorm. Provides the best results in classification.
Attentive Pooling Head	52.46%	52.05%	Replaces [CLS] token with attention-weighted aggregation over all frame tokens. Captures contextual relevance better than Linear but underperforms MLP.

TABLE II: Ablation study of different classification heads in PaSST for emotion recognition.

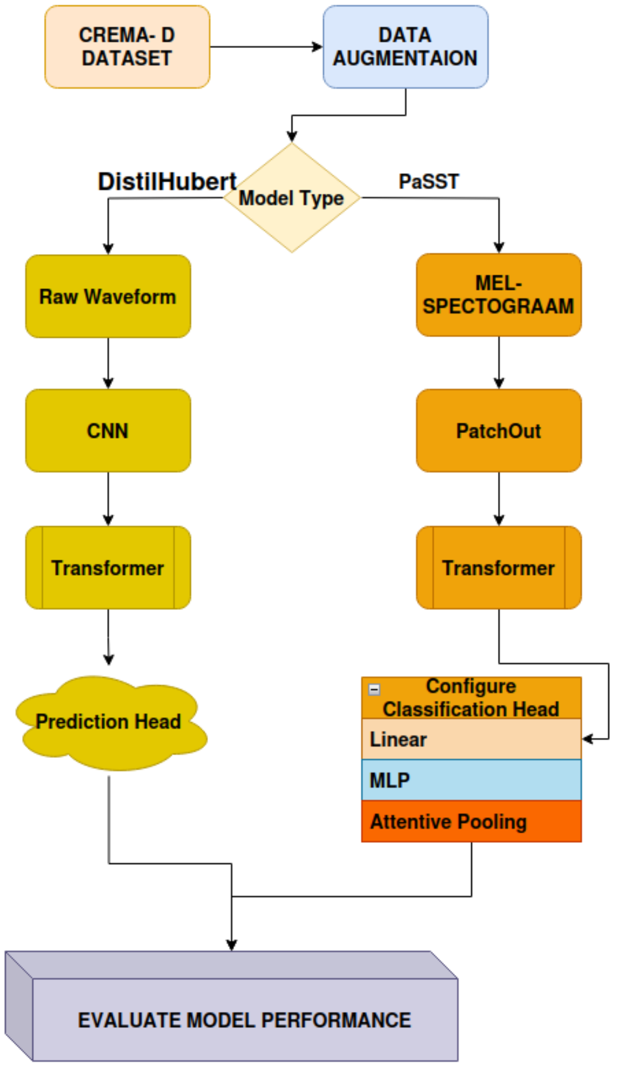


Fig. 4: Overall training pipeline for DistilHuBERT and PaSST models on the CREMA-D dataset.

Emotion	DistilHuBERT	PaSST-MLP	CNN-LSTM [22]
Angry	**86.91%**	68.22%	75.31%
Neutral	71.72%	66.49%	**71.70%**
Happy	63.35%	**59.17%**	61.18%
Sad	54.45%	54.55%	56.70%
Fear	67.37%	**60.47%**	59.04%
Disgust	40.31%	**43.46%**	38.33%

TABLE III: Per-emotion classification accuracy

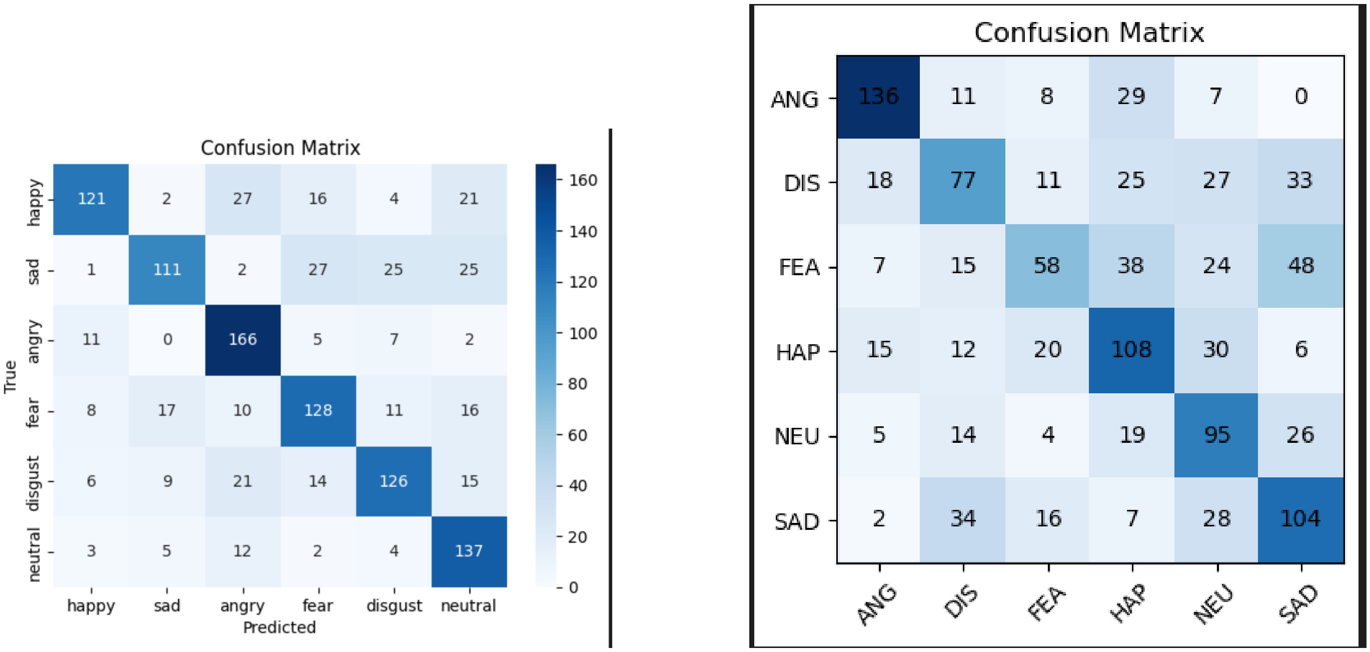
variant achieved a slight edge here (43.46%), suggesting that spectrogram-based attention may still capture isolated emotional cues better in rare categories.

These results confirm that while DistilHuBERT provides the best all-around performance, CNN-LSTM remains a strong baseline for emotions with less dynamic variation, and that raw waveform inputs can outperform spectrograms under transformer-based models, a departure from conventional audio processing wisdom.

The confusion matrices further illustrate the classification patterns and common misclassifications between models.

Visual Comparison of Model Interpretations

To further clarify how models process emotion-labeled audio, consider an example from the CREMA-D data set labeled 'Angry.' This audio sample undergoes data augmentation including gain adjustment and pitch shifting, simulating real-world recording variations. For the CNN-LSTM model, the sample is converted into MFCC features and passed through convolutional layers that capture local spectral patterns and LSTM layers that model temporal dynamics. DistilHuBERT processes the raw waveform directly, extracting contextual embeddings through its convolutional and transformer layers. PaSST, on the other hand, converts the audio into a Mel spectrogram and processes it via transformer blocks using patch-based attention.



(a) DistilHuBERT

(b) PaSST-Linear

Fig. 5: Confusion matrices of DistilHuBERT and PaSST-Linear showing per-emotion classification performance on the CREMA-D dataset.

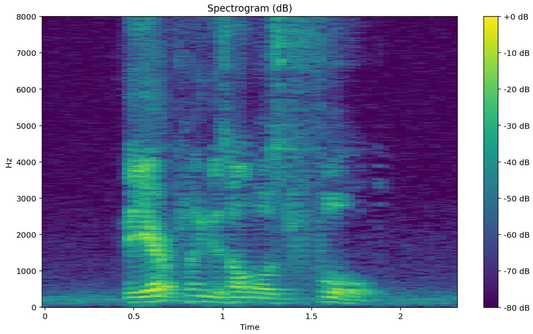


Fig. 6: Mel spectrogram of an 'Angry' utterance from the CREMA-D dataset. Energy is concentrated in the lower frequencies with noticeable bursts around 1–2 kHz and 3–4 kHz.

In this specific case, DistilHuBERT correctly classifies the sample as 'Angry' with confidence 87%, benefiting from its input of raw waveform and the ability to model prosodic cues such as pitch and tone. The model misclassifies the same sample as 'Happy', probably because of overlapping high-frequency energy in the spectrograms of both emotions. Figure 6 shows the input of the spectrogram used by PaSST. High-intensity regions (yellow-green) appear in the lower and mid frequency bands, typical of emotionally charged speech such as anger, which tends to exhibit higher pitch variation and energy bursts. PaSST processes this patchwise, potentially missing subtle temporal cues that DistilHuBERT captures from the raw waveform.

This shows how different model architectures and input representations influence classification decisions. It also demonstrates why raw waveform models such as DistilHuBERT may outperform spectrogram-based models when prosody is a key emotional signal.

Ablation Study on PaSST Variants

The PaSST architecture offers flexibility in how final classification is performed, allowing interchangeable classification heads. This ablation study evaluates three such configurations, Linear, MLP, and Attentive Pooling, under identical training conditions to isolate the impact of the classification head on model performance. In particular, all variants share the same input representation: Mel spectrograms. The difference lies solely in how the final emotion prediction is computed from the Transformer output.

As shown in Table II, the Linear Head configuration achieved the lowest performance, with an accuracy of 52.15% and an F1 score of 51.29%. This configuration directly maps the `[CLS]` token embedding to the emotion classes using a single linear layer. Although computationally efficient, its limited expressive power may constrain its ability to capture complex emotional nuances.

The MLP Head introduced a two-layer feedforward network composed of LayerNorm, ReLU, and dropout operations before the final output layer. It achieved the highest accuracy at 54.07% and an F1 score of 53.82%. The added depth and nonlinearity enable richer abstraction of features, demonstrating the effectiveness of moderately complex heads for emotion recognition. The Attention Group Head replaced the default `[CLS]` token with a statistical grouping mechanism applied to all temporal tokens. This combination computes a weighted mean and standard deviation of token characteristics, with attention weights learned during training. Although it performed slightly better than the linear head (52.46% accuracy, 52.05% F1), it still lagged behind the MLP configuration. This suggests that attention-based statistics help summarize temporal features, but may not be sufficient without additional nonlinear transformations.

These results indicate that head design plays a critical role in SER performance. Even when the backbone of the transformer and the input of the spectrogram remain constant, the classification head capacity significantly influences the model's ability to discriminate emotional states.

Conclusion and Future Work

This study presented a comparative analysis of DistilHuBERT and PaSST for speech emotion recognition using the CREMA-D data set. DistilHuBERT consistently outperformed all PaSST variants in terms of accuracy, F1 score, and model size, establishing itself as the most effective and deployable architecture for emotion recognition tasks under resource constraints. Its efficient use of raw waveform input and distilled transformer layers makes it a compelling choice for real-time applications.

However, while PaSST performed poorly relative to DistilHuBERT, it offers a modular and interpretable architecture with tunable classification heads. The ablation study revealed that the choice of classification head significantly impacts performance. Specifically, the MLP head provided the best results, indicating that shallow nonlinear transformations can help extract more discriminative features from transformer output. Although all PaSST configurations used Mel spectrogram inputs, statistical attention-based pooling (initially mislabeled as ”Raw”) did not lead to significant performance gains. This suggests that architectural decisions post-transformer (i.e., the classification head) are crucial and can rival the impact of input representation.

Overall, this study shows a key lesson, the choice of input representation and classification head can significantly affect model performance in SER. Raw waveform models such as DistilHuBERT are better suited to capture prosodic and temporal features, while spectrogram-based models like PaSST require more careful architectural tuning to compete.

In the broader research landscape, trends such as self-supervised pre-training (e.g. wav2vec 2.0, HuBERT), multimodal fusion (combining audio with facial cues), and emotion-aware contrastive objectives are being actively explored to improve SER performance. These approaches aim to improve generalization, reduce the reliance on large-labeled datasets, and improve robustness to noise and speaker variation.

Future work may explore the integration of multimodal signals, such as visual and physiological cues, to improve emotion recognition under ambiguous or low-quality audio conditions. Furthermore, extending PaSST pretraining to emotion-rich datasets and incorporating emotion-aware objectives during fine-tuning could help bridge the performance gap with DistilHuBERT. Lastly, further exploration of attention-based pooling and hierarchical heads may offer an avenue for boosting the performance of lightweight transformer models without significantly increasing computational cost.

References

[1] Qianhe Ouyang, “Speech Emotion Detection Based on MFCC and CNN-LSTM Architecture,” *Journal of Physics: Conference Series*, arXiv:2501.10666v1, 2024.

[2] Changhan Wang et al., “DistilHuBERT: Learning Speech Representation by Layer-wise Distillation,” arXiv:2110.01905.

[3] Zeineldeen et al., “PaSST: Efficient Audio Classification with Patchout Spectrogram Transformer,” arXiv:2110.05069.

[4] Neethu Mary Joy and Praveen Ravi, “Speech Emotion Recognition Using CNN and LSTM,” arXiv:1805.01055.

[5] H. K. Vydana, ”Speech emotion recognition using GMM,” in *International Journal of Engineering and Technology*, 2015.

[6] I. Shahin, ”Emotion recognition using hybrid GMM-DNN model,” in *IEEE Transactions on Affective Computing*, 2019.

[7] Q. Ouyang, ”Speech Emotion Detection Based on MFCC and CNN-LSTM Architecture,” *Journal of Physics: Conference Series*, arXiv:2501.10666v1, 2024.

[8] C. Wang *et al.*, ”DistilHuBERT: Learning Speech Representation by Layer-wise Distillation,” arXiv preprint arXiv:2110.01905, 2021.

[9] L. Pepino, P. Riera, and E. Dupoux, “Emotion recognition from speech using wav2vec 2.0 embeddings,” in *Proc. Interspeech*, 2021, pp. 3400–3404.

[10] M. El Ayadi, M. S. Kamel, and F. Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases,” *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[11] B. Schuller, S. Steidl, A. Batliner, et al., “Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends,” *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[12] G. Mountzouris, S. Parlak, M. Agrawal, and S. Narayanan, “Attention-based CNN models for speech emotion recognition,” in *Proc. Interspeech*, 2022, pp. 2003–2007.

[13] L. Pepino, P. Riera, and E. Dupoux, “Emotion recognition from speech using wav2vec 2.0 embeddings,” in *Proc. Interspeech*, 2021, pp. 3400–3404.

[14] S. Tripathi, A. W. Black, and A. Kumar, “Self-supervised learning for emotion recognition using transformers,” in *Proc. Interspeech*, 2020, pp. 3411–3415.

[15] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and J. Epps, “Deep learning for speech emotion recognition: A survey,” *IEEE Transactions on Affective Computing*, early access, 2020.

[16] M. B. Akçay and K. Oğuz, “Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers,” *Speech Communication*, vol. 116, pp. 56–76, 2020.

[17] Y. Kim, H. Lee, and E. M. Provost, “Deep learning for robust feature generation in audiovisual emotion recognition,” in *Proc. IEEE ICASSP*, 2013, pp. 3687–3691.

[18] S. Lee, D. Park, H. Kwon, and H. Ko, “A comprehensive review of the source-filter model and its applications in speech emotion recognition,” *Sensors*, vol. 21, no. 4, p. 1327, 2021.

[19] S. Zeineldeen *et al.*, ”PaSST: Efficient Audio Classification with Patchout Spectrogram Transformer,” arXiv preprint arXiv:2110.05069, 2021.

[20] H. K. Vydana, “Speech emotion recognition using Gaussian mixture models,” *International Journal of Engineering and Technology*, vol. 7, no. 2, 2015.

[21] I. Shahin, “Emotion recognition using hybrid GMM-DNN model,” *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 595–604, 2019.

[22] Q. Ouyang, “Speech Emotion Detection Based on MFCC and CNN-LSTM Architecture,” *Journal of Physics: Conference Series*, arXiv preprint arXiv:2501.10666, 2024.

[23] Z.-Q. Wang, D. Su, S. Liu, D. Yu, and Y. Yan, “DistilHuBERT: Speech Representation Learning by Layer-wise Distillation of HuBERT,” *arXiv preprint arXiv:2110.01905*, 2021.

[24] S. Zeineldeen, H. Koutini, H. Eghbal-zadeh, M. Dorfer, and G. Widmer, “Efficient Audio Classification with Patchout Spectrogram Transformer,” *arXiv preprint arXiv:2110.05069*, 2021.

[25] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. IEEE ICASSP*, 2016, pp. 5200–5204.

[26] L. Pepino, P. Riera, and J. Lorenzo-Trueba, "Emotion Recognition from Speech using wav2vec 2.0 Embeddings," in *Proc. Interspeech*, 2021, pp. 3400–3404.

[27] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[28] H. Koutini, S. Zeineldeen, H. Eghbal-zadeh, M. Dorfer, and G. Widmer, "Efficient Training of Audio Transformers with Patchout," in *Proc. IEEE ICASSP*, 2022, pp. 8562–8566.

[29] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for Speech Emotion Recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.

[30] S. Livingstone and F. Russo, "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*, vol. 13, no. 5, e0196391, 2018.

[31] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end Speech Emotion Recognition using a deep convolutional recurrent network," in *Proc. IEEE ICASSP*, 2016, pp. 5200–5204.

[32] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proc. IEEE ICASSP*, 2017, pp. 2227–2231.

[33] S. Chang, Y. Shi, and J. Glass, "DistilHuBERT: Speech representation learning by layer-wise distillation of HuBERT," in *Proc. Interspeech*, 2022, pp. 3653–3657.

[34] H. Cao, D. Livingstone, and F. D. Russo, "The CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, Oct.–Dec. 2014.

[35] Q. Koutini, H. Eghbal-Zadeh, M. Widrich, and G. Widmer, "Efficient training of audio transformers with patchout," *Proc. IEEE ICASSP*, 2022, pp. 874–878.

[36] A. Akinpelu, S. A. Akinola, A. E. Adebayo, and K. O. Oyelade, "Speech emotion recognition using Vision Transformers," in *Proc. IEEE ICAIBD*, 2022.