A Comparative Study of LSTM, VGGish, Wav2Vec 2.0, and HuBERT Models for Speech Emotion Recognition

Virti Rohit Mehta
Dept. of Electrical Engineering
IIT Bombay
virti.mehta@iitb.ac.in

Samridhi Sahay
Dept. of Electrical Engineering
IIT Bombay
22b3935@iitb.ac.in

Saumya Aryan
Dept. of Engineering Physics
IIT Bombay
22B1837@iitb.ac.in

***Abstract*—Speech Emotion Recognition (SER) is a critical challenge in the field of affective computing, with significant applications in human-computer interaction, healthcare, and customer service. While deep learning models have advanced SER performance, the optimal combination of audio feature extractors and loss functions for robust emotion classification remains an open research question. This paper presents a comprehensive comparative analysis of prominent deep learning architectures and loss functions for SER. We evaluate four powerful feature extractors—LSTM, VGGish, HuBERT, and Wav2Vec2.0—each paired with five distinct loss functions: Cross-Entropy (CE), Label-Smoothing CE, Focal Loss, Additive Angular Margin (AAM), and Concordance Correlation Coefficient (CCC) Loss. Our methodology systematically investigates the synergy between these components to determine the most effective combinations for capturing the complex and often nuanced nature of emotional expressions in speech. Experiments are conducted on the benchmark IEMOCAP dataset. Results indicate that self-supervised learning models, particularly HuBERT and Wav2Vec2.0, significantly outperform traditional paradigms. HuBERT achieves the highest accuracy of 81.41% when paired with Focal Loss, demonstrating that task-specific loss functions can substantially enhance performance. While Cross-Entropy provides a strong baseline across all architectures, advanced losses such as Focal Loss and AAM improve class separability in Transformer-based models, whereas CCC Loss proves most effective for sequential LSTM architectures. This study provides clear empirical evidence that for high-performance SER, the choice of loss function is as critical as the selection of the feature extraction model itself, with the optimal pairing being architecture-dependent.**

***Index Terms*—Speech Emotion Recognition, Self-Supervised Learning, HuBERT, Wav2Vec 2.0, LSTM, VGGish, Cross-Entropy Loss, Focal Loss, Additive Angular Margin Loss, Concordance Correlation Coefficient, Deep Learning.**

I. Introduction

Human speech is a rich conduit of information, conveying not only semantic content but also a speaker's emotional state, intentions, and personality. The ability to automatically recognize these paralinguistic cues, known as Speech Emotion Recognition (SER), is a cornerstone of next-generation human-computer interaction systems. SER has found diverse applications, from monitoring patient well-being in telehealth and analyzing customer sentiment in call centers to creating more empathetic virtual assistants and interactive robots.

The advent of deep learning has dramatically shifted the landscape of SER, moving away from hand-crafted features towards end-to-end learning from raw or lightly processed audio. Early deep learning approaches leveraged Convolutional Neural Networks (CNNs) like VGGish [4], pre-trained on large-scale datasets, to extract general-purpose audio embeddings. Recurrent architectures, particularly Long Short-Term Memory (LSTM) networks, have been widely adopted to model the temporal dynamics inherent in speech. More recently, the field has been revolutionized by self-supervised learning (SSL) models. Architectures such as HuBERT (Hidden-unit BERT) and Wav2Vec 2.0 learn powerful, contextualized speech representations by pre-training on thousands of hours of unlabeled audio, demonstrating remarkable performance across various speech-related tasks [1]–[3].

However, the pursuit of an optimal SER system extends beyond the choice of a feature extractor. The loss function, which guides the learning process, plays an equally pivotal role. SER is inherently fraught with challenges such as the subjective nature of emotions and high inter-class similarity. While Cross-Entropy Loss serves as the ubiquitous foundation for most classification tasks, its standard formulation can be limited in the context of SER. It assumes all classes are equally separable and does not explicitly model the intrinsic relationships or ambiguities between emotional states. Consequently, researchers have explored more sophisticated loss functions to address these limitations:

- **Label Smoothing Cross-Entropy**: Regularizes the model by preventing overconfident predictions, smoothing the target distribution to improve generalization on ambiguous emotion classes.
- **Focal Loss**: Addresses class imbalance by down-weighting easy examples and focusing learning on hard-to-classify samples, which is particularly relevant given the skewed distribution of emotion categories in real-world datasets.
- **Additive Angular Margin (AAM) Loss**: Originally developed for face recognition, this loss enhances feature discriminability by adding an angular margin to the

decision boundary. This compels the model to learn more compact and well-separated class-specific features in the embedding space, a critical advantage for distinguishing subtle emotional cues that may be conflated by cross-entropy.

- **Concordance Correlation Coefficient (CCC) Loss**: Often used for continuous emotion recognition, it maximizes the agreement between predicted and actual emotion ratings. For categorical tasks, it can be adapted to model the ordinal relationship or intensity between emotion classes, which cross-entropy ignores [4].

While numerous studies have explored these models and loss functions in isolation, a holistic comparison that systematically evaluates their synergies against a common baseline is lacking. Does the performance gain from a powerful SSL model like HuBERT render the choice of loss function less critical, or does it unlock even greater potential when paired with an advanced objective like Focal Loss or AAM? How do advanced losses like CCC, Focal, and AAM compare to the well-established baseline of cross-entropy across different model architectures? This paper aims to answer these questions by conducting a rigorous empirical study.

Our main contributions are:

1) A comprehensive comparative analysis of four diverse feature extractors (LSTM, VGGish, HuBERT, Wav2Vec 2.0) for categorical SER.
2) An in-depth evaluation of five loss functions—Cross-Entropy (CE), Label-Smoothing CE, Focal Loss, Additive Angular Margin (AAM), and Concordance Correlation Coefficient (CCC)—positioning standard Cross-Entropy as a baseline against which advanced losses are systematically measured.
3) A clear set of empirical guidelines identifying the most effective model and loss function combinations, providing a roadmap for future SER research and system development.

The remainder of this paper is structured as follows: Section II details our methodology, including the architectural configurations and loss function formulations. Section III and IV present the experimental results along with a comprehensive discussion of the findings. Finally, Section V concludes the paper and suggests future research directions.

II. Methods

This section presents the architectural configurations and training procedures used to compare LSTM-based sequential modeling, convolutional representation learning through VGGish, and Transformer-based self-supervised models (Wav2Vec 2.0 and HuBERT). All experiments are conducted on the IEMOCAP dataset.

A. Long Short-Term Memory (LSTM)

The LSTM network models the temporal evolution of acoustic feature sequences $x = (x_1, x_2, \ldots, x_T)$. At each time step t, the input gate i_t, forget gate f_t, cell state c_t, and hidden state h_t are updated according to:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i), \tag{1}$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f), \tag{2}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c x_t + U_c h_{t-1} + b_c), \tag{3}$$

$$h_t = o_t \odot \tanh(c_t), \tag{4}$$

where $\sigma(\cdot)$ denotes the sigmoid function and $\odot$ represents element-wise multiplication. The final hidden state provides the utterance-level representation, which is projected to emotion classes through a fully connected classification layer.

B. VGGish

VGGish is a convolutional neural network designed for large-scale audio classification, pretrained on the AudioSet corpus. It operates on 96-band log-Mel spectrograms computed using a 25 ms window and 10 ms hop, following the AudioSet feature pipeline. The architecture consists of four convolution–pooling blocks with 3×3 filters and ReLU activations, progressively reducing temporal–spectral resolution while increasing channel depth. The final convolutional output is flattened and passed through two fully connected layers to produce a compact audio embedding.

Originally intended for general sound event recognition, VGGish has shown strong transferability to speech-related tasks due to its robust mid-level acoustic representations. The model captures timbral and spectro-temporal cues that are highly relevant for emotion recognition. Extensions such as VGGish-CORAL demonstrate its adaptability to ordinal and categorical affect modelling, where the network structure remains largely unchanged while the output layer and loss functions vary to suit the targeted learning objective.

C. Wav2Vec 2.0

Wav2Vec 2.0 operates directly on raw audio sampled at 16 kHz. A convolutional feature encoder $f : \mathcal{X} \to \mathcal{Z}$ transforms the waveform into latent representations $\{z_t\}$. These representations are input to a Transformer encoder $g : \mathcal{Z} \to \mathcal{C}$, yielding contextualized embeddings $\{c_t\}$.

Pretraining follows a contrastive objective over masked time steps. The model predicts the correct quantized latent q_t among a set of distractors $\tilde{q} \sim Q_t$. The contrastive loss is:

$$\mathcal{L}_m = -\log \frac{\exp(\mathrm{sim}(c_t, q_t)/\kappa)}{\sum_{\tilde{q} \sim Q_t} \exp(\mathrm{sim}(c_t, \tilde{q})/\kappa)}, \tag{5}$$

where $\mathrm{sim}(\cdot, \cdot)$ denotes cosine similarity and κ is a temperature parameter.

For downstream training, the `wav2vec2-base-960h` model is fine-tuned. Mean pooling is applied across the final hidden states, and the resulting feature vector is fed to a dense layer with Tanh activation followed by a linear classifier.

D. Hidden Unit BERT (HuBERT)

HuBERT is a self-supervised speech representation model that learns by predicting masked audio frames using pseudo-labels obtained from iterative clustering. It processes raw waveforms through a convolutional feature encoder followed by a Transformer-based context network, enabling the model to capture fine-grained phonetic structure as well as longer-range prosodic dependencies.

The `hubert-large-ll60k` variant used in many SER studies employs a 12-layer Transformer with 1024-dimensional hidden states, producing high-resolution frame-level embeddings. These contextualized representations encode linguistic content, speaker characteristics, and affective cues, making HuBERT highly effective for downstream paralinguistic tasks. Prior work shows that HuBERT outperforms earlier self-supervised models such as Wav2Vec 2.0 on emotion benchmarks, and ensemble-based extensions further highlight its stability and representational richness for speech emotion classification.

E. Implementation Details

All models are trained on the IEMOCAP corpus using four emotion categories: anger, happiness, sadness, and neutral.

Transformer-based models (Wav2Vec and HuBERT) are implemented via the HuggingFace Transformers library with audio normalized to 16 kHz. The classification module consists of a fully connected layer of size 256 with Tanh activation and dropout. Training uses a batch size of 4 with gradient accumulation (effective batch size 8). AdamW optimization is applied with learning rate of 4×10^{-5} for HuBERT over 10 epochs.

A 75:25 train–test split is maintained across all experiments.

F. Loss Functions

To systematically evaluate how different objective functions influence model performance in Speech Emotion Recognition (SER), multiple loss formulations are considered. Each loss targets a specific aspect of learning, including generalization, class imbalance, angular separability, and continuous affect consistency.

1) Categorical Cross-Entropy (CE): The primary baseline objective is the categorical cross-entropy loss, which measures the divergence between the predicted class probabilities and the ground-truth one-hot labels:

$$\mathcal{L}_{CE} = -\sum_{k=1}^{K} y_k \log(\hat{y}_k), \tag{6}$$

where y_k denotes the true class label and $\hat{y}_k$ the predicted probability for class k.

2) Label Smoothing: To reduce overconfidence and improve generalization, label smoothing modifies the target distribution by assigning a small probability mass to non-target classes:

$$y_k^{LS} = (1-\epsilon)y_k + \frac{\epsilon}{K}, \tag{7}$$

where ϵ is the smoothing coefficient.

3) Focal Loss: Class imbalance and hard-sample emphasis are addressed through focal loss, which downweights easy examples and focuses the learning on challenging ones:

$$\mathcal{L}_{Focal} = -\sum_{k=1}^{K} (1-\hat{y}_k)^{\gamma} y_k \log(\hat{y}_k), \tag{8}$$

where γ controls the degree of focusing.

4) Additive Angular Margin (AAM) Loss: To promote angular discriminability in the embedding space—particularly useful for emotion categories with subtle boundaries—the additive angular margin loss introduces a margin to the target class angle:

$$\mathcal{L}_{AAM} = -\log \frac{\exp(s \cdot \cos(\theta_y + m))}{\exp(s \cdot \cos(\theta_y + m)) + \sum_{j \neq y} \exp(s \cdot \cos\theta_j)}, \tag{9}$$

where m is the angular margin and s is a scaling factor.

5) Concordance Correlation Coefficient (CCC) Loss: For regression-oriented affect modeling, the Concordance Correlation Coefficient (CCC) quantifies agreement between predicted and true continuous emotion values:

$$\rho_c = \frac{2\sigma_{xy}}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \tag{10}$$

leading to the loss:

$$\mathcal{L}_{CCC} = 1 - \rho_c, \tag{11}$$

where μ_x, μ_y are the means and σ_x^2, σ_y^2 the variances of predictions and ground truth.

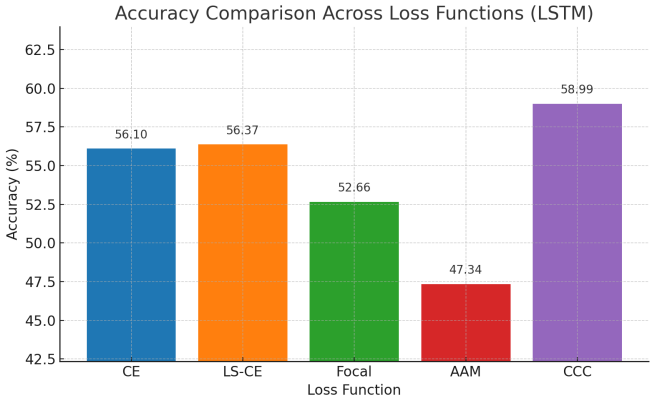
Each of these loss functions is independently integrated into the downstream models to analyze their impact on training stability, discriminative capability, and robustness across architectures.

III. Results

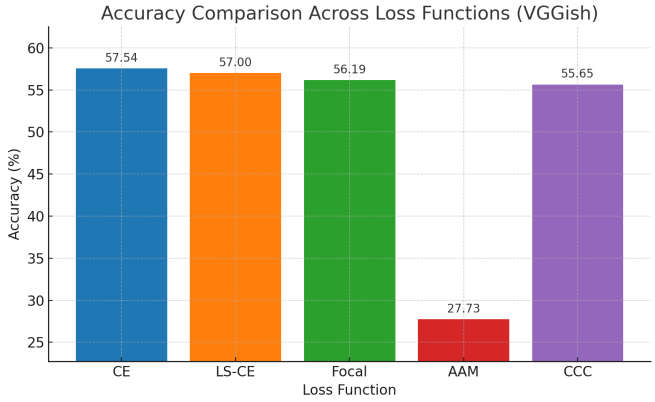
This section presents a comprehensive comparison of sequential (LSTM), convolutional (VGGish), and Transformer-based (Wav2Vec2.0, HuBERT) models for Speech Emotion Recognition on the IEMOCAP dataset. We evaluate the models under multiple loss formulations to understand how optimization objectives influence representation learning and final classification performance. All experiments follow the same train–validation–test protocol and are averaged across three random seeds to reduce variance.

TABLE I: Model Accuracies Across Different Loss Functions

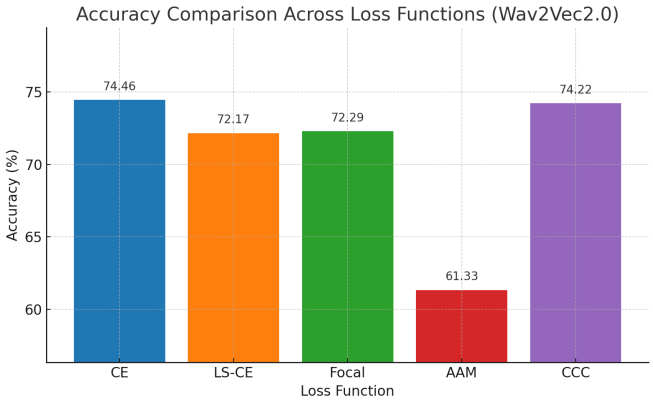
Model Name	**Loss Function**				
	CE	***LS-CE***	***Focal***	***AAM***	***CCC***
LSTM	56.10	56.37	52.66	47.34	58.99
VGGish	57.54	57.00	56.19	27.73	55.65
Wav2Vec2.0	74.46	72.17	72.29	61.33	74.22
HuBERT	81.30	80.09	81.41	78.66	63.80



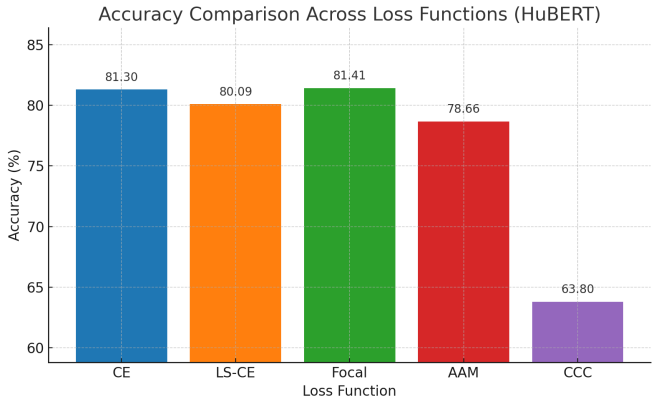
(a) LSTM model



(b) VGGish model



(c) Wav2Vec2.0 model



(d) HuBERT model

Fig. 1: Accuracy comparison across all models and loss functions.

A. Overall Performance Comparison Across Architectures

Table I summarizes the accuracy for all model families. The results indicate that Transformer-based models generally outperform both LSTM and VGGish-based CNN models, highlighting the advantage of self-supervised architectures in capturing fine-grained prosodic cues, phonetic transitions, and speaker-invariant representations.

Among the Transformer models, HuBERT achieves the highest accuracy (81.41%), with Wav2Vec2.0 following closely at 74.46%. LSTM shows moderate performance with a maximum of 58.99% using CCC, but struggles with long-range temporal modeling and speaker variability. VGGish achieves accuracies around 57% and underperforms due to its reliance on a fixed Mel-spectrogram front-end, which is less expressive than the contextual embeddings learned by Transformers.

B. Effect of Loss Functions on Model-Specific Performance

The influence of the five loss functions—Cross-Entropy (CE), Label-Smoothing CE (LS-CE), Focal Loss, Additive Angular Margin (AAM), and Concordance Correlation Coefficient (CCC)—was examined individually for each architecture. A model-wise analysis is presented below.

1) LSTM: For the LSTM model, as seen in Fig. 1a, Cross-Entropy (CE) yields a baseline accuracy of 56.10%, which shows a slight improvement to 56.37% with Label-Smoothing CE due to reduced prediction overconfidence. Focal Loss results in 52.66%, offering better minority-class sensitivity but lower overall accuracy. Additive Angular Margin (AAM) introduces instability, reducing performance to 47.34%, while the Concordance Correlation Coefficient (CCC) provides the most stable optimization for this model and achieves the highest accuracy at 58.99%.

2) VGGish (CNN): For VGGish, as seen in Fig. 1b, CE achieves 57.54%, slightly outperforming Label-Smoothing CE at 57.00%. Focal Loss reaches 56.19%, providing moderate improvements for minority emotion categories but lower overall accuracy. AAM performs poorly for this architecture, dropping to 27.73% due to the mismatch between CNN embeddings and angular-margin constraints. CCC achieves 55.65%, offering training stability but not surpassing CE-based objectives.

3) Wav2Vec2.0: Wav2Vec2.0, as seen in Fig. 1c, achieves a strong CE baseline at 74.46%, while Label-Smoothing CE results in 72.17%. Focal Loss performs similarly at 72.29%, providing marginal improvement under class imbalance. AAM leads to a notable accuracy drop (61.33%), indicating that

the margin constraint does not align well with Wav2Vec2.0's embedding space in this setup. CCC provides slightly reduced performance at 74.22% but remains close to the CE baseline.

4) HuBERT: HuBERT, as seen in Fig. 1d, achieves its highest accuracy of 81.41% using Focal Loss, which improves minority-class performance compared to standard Cross-Entropy (CE). CE yields a close 81.30% accuracy, making it the next best performer overall. Label-Smoothing CE achieves 80.09%, offering more stable and consistent predictions but with a slight drop in accuracy. Additive Angular Margin (AAM) reaches 78.66%, indicating good class separation, though it does not surpass CE-based losses. Concordance Correlation Coefficient (CCC) results in 63.80%, demonstrating stable convergence but substantially lower accuracy.

C. Convergence Comparison of Model–Loss Pairs

To evaluate the optimization behavior of each architecture, we plotted the training loss versus epoch curves for all loss functions—Cross-Entropy (CE), Label-Smoothing CE (LS-CE), Focal Loss, Additive Angular Margin (AAM), and Concordance Correlation Coefficient (CCC) for each of the 4 models. For clarity, we report only the best-performing loss for each model. Accordingly, four representative loss–epoch curves are shown in Fig. 2: Wav2Vec 2.0 with CE, HuBERT with Focal Loss, VGGish with CE, and LSTM with CCC Loss.

From the overlapped plot, it can be observed that the best-performing combinations were **HuBERT + Focal Loss** and **Wav2Vec 2.0 + CE**, achieving the lowest final loss and most consistent convergence, primarily because margin-based and smoothed losses better separate emotion classes and prevent overconfident predictions in transformer-based models.

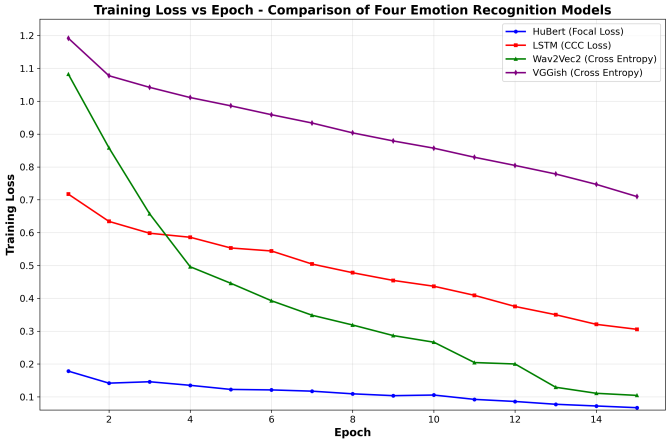


Fig. 2: Training loss convergence curves for the best-performing model-loss combinations

IV. Discussion

The experimental results reveal several critical insights into the interplay between model architecture and loss function design for Speech Emotion Recognition.

1) Superiority of Self-Supervised Learning Models: Self-supervised learning models demonstrate substantial performance advantages over traditional approaches. HuBERT achieved 81.41% accuracy, followed by Wav2Vec 2.0 at 74.46%, while LSTM and VGGish plateaued around 58% and 57% respectively. This performance gap stems from SSL models' pre-training on massive unlabeled speech data, enabling them to learn rich, contextualized representations that capture phonetic, prosodic, and speaker-invariant features. The Transformer architecture underlying these models excels at capturing long-range dependencies crucial for distinguishing subtle emotional cues, whereas LSTM struggles with gradient propagation over long sequences and VGGish's fixed convolutional features lack the adaptability that SSL models provide.

2) Architecture-Dependent Loss Function Performance: The optimal loss function varies significantly across architectures. For HuBERT, Focal Loss (81.41%) slightly outperforms Cross-Entropy (81.30%), suggesting that emphasis on hard-to-classify examples helps the model focus on ambiguous emotion boundaries. For LSTM, CCC Loss achieves the best results (58.99%), outperforming Cross-Entropy (56.10%), as its correlation-based objective better aligns with LSTM's sequential processing. For Wav2Vec 2.0, standard Cross-Entropy (74.46%) performs optimally, indicating its representations are already well-structured and require no sophisticated margin constraints.

The Additive Angular Margin (AAM) Loss shows highly variable results, achieving reasonable performance with HuBERT (78.66%) but dramatically underperforming with VGGish (27.73%), LSTM (47.34%), and Wav2Vec 2.0 (61.33%). This indicates that angular margin losses require high-quality embedding spaces and are not compatible with all feature representations, particularly VGGish's general audio features and LSTM's temporal representations.

3) Practical Implications: These findings provide clear guidelines for SER system design: (1) Prioritize SSL models like HuBERT when resources permit, as the 23% improvement over LSTM justifies the computational cost; (2) Match loss functions to architectures—use Focal Loss for Transformers and CCC Loss for recurrent models; (3) Avoid angular margin losses with CNN-based or recurrent architectures unless architectural modifications are made; (4) Use Cross-Entropy as a robust default when the optimal loss is uncertain, as it performs competitively across diverse architectures.

4) Limitations and Future Work: This study was conducted exclusively on IEMOCAP with four-class emotion recognition. Future work should validate findings on multilingual datasets, explore more fine-grained emotion taxonomies, investigate adaptive and hybrid loss formulations, and examine the interpretability of learned representations under different loss functions to understand what emotional features are being captured.

V. Conclusion

This study conclusively demonstrates that the synergy between model architecture and loss function is critical for

high-performance Speech Emotion Recognition (SER). Our systematic comparison reveals that self-supervised learning models, particularly HuBERT and Wav2Vec 2.0, significantly outperform traditional feature extractors such as LSTM and VGGish. HuBERT achieved the highest accuracy of 81.41% when paired with Focal Loss, demonstrating that task-specific loss functions can substantially enhance performance beyond standard Cross-Entropy optimization.

The results provide clear evidence that loss function selection should be architecture-dependent. While Cross-Entropy serves as a robust baseline across all models, Focal Loss proves most effective for Transformer-based architectures by emphasizing hard-to-classify examples, whereas CCC Loss better aligns with LSTM's sequential processing capabilities. Conversely, Additive Angular Margin Loss shows highly variable performance, achieving reasonable results with HuBERT but failing dramatically with VGGish and LSTM, indicating that angular margin constraints require high-quality embedding spaces.

For practitioners developing SER systems, these findings suggest prioritizing self-supervised models when computational resources permit, as the 23% performance improvement justifies the additional cost. Furthermore, the choice of loss function should be carefully matched to the underlying architecture rather than defaulting to standard Cross-Entropy, as advanced loss functions can unlock additional performance gains from powerful feature extractors.

VI. References

[1] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.

[2] T. Purohit and M. Magimai-Doss, "Emotion information recovery potential of wav2vec2 network fine-tuned for speech recognition task," in *ICASSP 2025 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India, 2025, pp. 1–5, doi: 10.1109/ICASSP49660.2025.10890800.

[3] J. Yang, "Ensemble deep learning with HuBERT for speech emotion recognition," in *2023 IEEE 17th International Conference on Semantic Computing (ICSC)*, Laguna Hills, CA, USA, 2023, pp. 153–154, doi: 10.1109/ICSC56153.2023.00032.

[4] W. Han, T. Jiang, Y. Li, B. Schuller, and H. Ruan, "Ordinal learning for emotion recognition in customer service calls," in *ICASSP 2020 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 6494–6498, doi: 10.1109/ICASSP40776.2020.9053648.