Emotion Detection in Urdu Speech Using Perception on Mel Spectrograms

Abdullah Riaz (22L-7489)
Department of Computer Science
National University of Computer and Emerging Sciences
Lahore, Pakistan
l227489@lhr.nu.edu.pk

Zainab Khan Lodhi (22L-7475)
Department of Computer Science
National University of Computer and Emerging Sciences
Lahore, Pakistan
l227475@lhr.nu.edu.pk

***Abstract*—While Speech Emotion Recognition (SER) has made significant strides in deciphering human sentiment, the landscape remains disproportionately focused on high-resource languages like English. Consequently, languages such as Urdu—spoken by millions globally—are left largely unexplored. This study seeks to bridge this divide by reimagining Urdu SER as a computer vision task, transforming raw speech into Mel spectrograms. We subsequently benchmark two distinct visual architectures on these representations: a standard Convolutional Neural Network (CNN) and a contemporary Audio Spectrogram Transformer (AST). Our primary objective is to assess how these models navigate the subtleties of a naturalistic Urdu dataset. Through this comparative analysis, we aim to establish a robust baseline for future inquiries into emotion recognition for Urdu and similar low-resource languages.**

***Index Terms*—Speech Emotion Recognition, Deep Learning, Audio Spectrogram Transformer, CNN, Mel Spectrogram, Low-Resource Languages, Urdu**

I. Introduction

A. The Need for Speech Emotion Recognition

Computers have become remarkably proficient at parsing the *content* of our speech, yet they frequently fail to grasp the *intent* behind it. This is the precise gap Speech Emotion Recognition (SER) aims to fill. The objective is to endow systems with the perceptual capability to identify states like anger, joy, or sorrow directly from vocal acoustics [1], [2]. The potential impact on human-computer interaction is substantial; consider a virtual assistant that modulates its tone upon detecting frustration, or a customer support system that prioritizes a call when it senses distress [3], [4].

B. The Mel Spectrogram Approach

Human speech is dense with paralinguistic markers. The pitch contour, cadence, and rhythm of our voices often convey as much meaning as the vocabulary we choose [5]. However, extracting these ephemeral cues from raw audio waveforms is a notoriously complex engineering challenge [1], [6].

To circumvent this, researchers have adopted a cross-modal approach: translating audio processing into an image recognition problem. We convert sound into Mel spectrograms—visual heatmaps representing frequencies over time [4], [5]. This strategy is effective for two reasons. First, the Mel scale approximates human auditory perception, amplifying the frequencies our ears naturally prioritize [5], [7]. Second, by generating an image, we can unlock the vast capabilities of modern computer vision, employing architectures like Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) [1], [4].

C. The Low-Resource Language Challenge

A significant disparity persists in the current research landscape. The overwhelming majority of SER studies utilize English or other dominant languages. Urdu, despite its extensive speaker base, is categorized as "low-resource" in this domain [8], [9]. In practice, this means researchers lack the massive, curated public datasets available for Western languages, which inevitably stifles innovation [8], [9]. This project directly tackles this scarcity by evaluating modern vision models on a naturalistic Urdu dataset.

II. Related Work and Identified Gaps

A. Evolution of SER Models

The concept of treating speech as an image is not novel, but the methodologies have matured. Early iterations relied on CNNs to detect local patterns within spectrograms [4]. While functional, standard CNNs suffer from a limited "receptive field"—they analyze small, isolated patches of the image. Since emotion is often a temporal journey evolving across a sentence, a simple CNN can miss the broader narrative. To mitigate this, researchers later introduced hybrid models, appending Long Short-Term Memory (LSTM) units to CNNs to track temporal progression [5], [10].

The paradigm has recently shifted toward Vision Transformers (ViTs). Unlike CNNs, ViTs employ a self-attention mechanism to process the image globally [6]. This architecture allows the model to capture long-range dependencies, linking, for example, a sharp intonation at the start of a phrase with a pitch drop at the end.

B. Identified Gaps

A review of existing literature reveals a consistent trend: advanced models are almost exclusively benchmarked on a handful of English-centric datasets like IEMOCAP and RAVDESS [2], [6]. This exposes two critical voids. First, a **language gap**: state-of-the-art architectures are rarely validated on low-resource languages like Urdu [11], [12]. Second,

a **data gap**: most datasets rely on actors reading scripts in controlled studios. We possess very little data on how these powerful models perform when confronted with the messy reality of natural, conversational Urdu speech.

III. Literature Review

To situate our work, we analyzed recent developments in general SER architectures alongside specific initiatives within the Urdu language domain.

Li et al. (2024) proposed MelTrans, a Vision Transformer tailored for SER [6]. While promising on IEMOCAP and EmoDB, the study focused predominantly on noise robustness rather than linguistic nuance.

Mishra et al. (2024) investigated personalized SER using ViTs for human-robot interaction [3]. Although they fine-tuned on RAVDESS, their scope was restricted to four basic emotions, limiting the emotional granularity of the model.

Mustafa et al. (2024) proved that a well-optimized CNN remains a strong contender [4]. Their model performed admirably on RAVDESS but struggled to differentiate emotions with overlapping spectral signatures.

Aljuhani et al. (2024) modernized the classic AlexNet architecture for 3D spectrograms [1]. This validated the utility of legacy models, though the study was limited by its exclusive reliance on audio, ignoring multimodal potential.

Sharan et al. (2024) designed a hybrid approach merging a pre-trained CNN with an LSTM [5]. Tested on RAVDESS, the study was somewhat constrained by the limited variety of emotion types evaluated.

Ouyang (2025) also adopted a hybrid CNN-LSTM architecture to better handle sequential data [10]. However, the model achieved a relatively low overall accuracy (61.07%), with particularly poor performance on "disgust" (38.33%), suggesting difficulty in separating spectrally similar emotions.

Dangol et al. (2023) integrated an attention mechanism into a CNN-LSTM model [2]. While this improved performance, it represents an incremental refinement of older hybrid designs rather than a pivot to pure Transformer architectures.

In the context of Urdu, efforts have been largely confined to dataset creation. **Raza et al. (2022)** released SEMOUR+, a large-scale repository of scripted speech [8]. While the audio fidelity is high, the acted nature of the speech may not fully mirror authentic emotional expression.

Asghar et al. (2022) curated a clean dataset of acted emotions [11]. A key limitation here was the use of non-professional actors repeating a narrow set of fixed sentences, which restricts lexical diversity.

Syed & Memon (2020) developed the Urdu-Sindhi Speech Emotion Corpus, gathering data via WhatsApp to diversify the speaker pool [12]. Crucially, they observed poor transferability between Urdu and Sindhi models, underscoring that emotion recognition models are often deeply language-specific.

IV. Problem Statement and Research Questions

A. Problem Statement

Detecting emotion in low-resource languages like Urdu remains a formidable challenge, compounded by data scarcity and a lack of localized benchmarks. We know that Vision Transformers (ViTs) excel at processing English spectrograms, but their efficacy on natural-sounding Urdu speech is largely unproven. Furthermore, it is unclear how they compare against established, simpler models like CNNs in this specific linguistic context.

This project aims to bridge that gap. We perform a direct comparison between a CNN and a ViT on the UrduSER dataset, setting a performance baseline and identifying which architecture is better suited for realistic Urdu speech.

B. Research Questions

Our investigation is guided by the following questions:

- How does a modern Vision Transformer (ViT) perform against a standard CNN when analyzing Urdu speech spectrograms?
- Which specific emotions are these models capable of distinguishing, and where does confusion typically arise?
- Can a model trained on a specific set of speakers successfully generalize to a new speaker it has never encountered?

V. Methodology

This section outlines our research strategy, justifying our architectural choices and defining the finetuning protocol used to adapt these models for a low-resource environment.

A. Baseline Architecture: PANNs (Cnn14)

For our baseline, we selected the **Cnn14** model from the Pre-trained Audio Neural Networks (PANNs) family [13]. PANNs are essentially a suite of deep CNNs pre-trained on AudioSet [14], providing a strong foundation in general audio pattern recognition.

Cnn14 is a 14-layer VGG-style Convolutional Neural Network. Its design includes:

- **Input Layer:** Accepts a two-dimensional Log-Mel Spectrogram.
- **Convolutional Blocks:** The backbone consists of six convolutional blocks. Following the VGG philosophy, each block contains two convolutional layers (utilizing 3×3 kernels), followed by Batch Normalization (BN) and ReLU activation.
- **Downsampling:** A 2×2 average pooling layer is applied after each block to reduce spatial dimensions.
- **Classification Head:** The final feature map is condensed via global average pooling (GAP) before passing into a linear classification layer.

B. Primary Architecture: Audio Spectrogram Transformer (AST)

To transcend the locality constraints inherent in CNNs, we employ the **Audio Spectrogram Transformer (AST)** [15] as our primary model. Unlike CNNs, which view the spectrogram as a single monolithic image, the AST processes the audio as a sequence of patches—much like a Vision Transformer handles photographs.

TABLE I
LITERATURE SURVEY OF DEEP LEARNING MODELS FOR SER

Reference	Year	Method	Dataset(s)	Limitation
Li et al. [6]	2024	Vision Transformer (MelTrans)	IEMOCAP, EmoDB	Focuses on common SER challenges like noise, not language-specific issues.
Mishra et al. [3]	2024	Vision Transformers (ViT & BEiT)	RAVDESS, TESS, MELD, Custom HRI	Focuses on only four primary emotions; limited number of participants in HRI study.
Mustafa et al. [4]	2024	Convolutional Neural Network (CNN)	RAVDESS, IEMOCAP	Acknowledges moderate misclassification for emotions with overlapping spectral features.
Aljuhani et al. [1]	2024	CNN (Modified AlexNet)	IEMOCAP	Exclusive focus on audio modality; suggests future work on multimodal approaches.
Sharan et al. [5]	2024	Pre-trained CNN (YAMNet) + LSTM	RAVDESS, DEMoS	Evaluated on a small number of emotion types and a single network architecture.
Ouyang [10]	2025	Hybrid CNN-LSTM Architecture	IEMOCAP, EmoDB, SAVEE, RAVDESS, TESS	Achieved relatively low overall accuracy (61.07%) and struggled with emotions having similar features.
Dangol et al. [2]	2023	CNN-LSTM with attention	EMODB, IEMOCAP	Focuses on improving existing hybrid models rather than exploring novel architectures like pure ViT.
Raza et al. [8]	2022	Dataset Creation (SEOUR+) & CNN/VGG-19 Benchmark	SEMOUR+ (Acted, Studio)	Acted emotions may not reflect natural emotional expression.
Asghar et al. [11]	2022	Dataset Creation & ML Benchmark	Urdu Speech Corpus (Acted, Lab)	Acted by non-professionals; limited to five fixed sentences.
Syed & Memon [12]	2020	Dataset Creation (Urdu-Sindhi) & ML Benchmark	Urdu-Sindhi Corpus	Found poor cross-lingual transferability between Urdu and Sindhi, suggesting emotion models are language-specific.

- **Patch Embedding:** The input 2D spectrogram is sliced into a sequence of 16×16 patches, which are flattened and projected into 1D embeddings.
- **Positional Encoding:** Since Transformers are permutation-invariant (they don't inherently perceive order), we add learnable positional embeddings. This is critical for retaining the temporal (time) and frequency structure of the audio.
- **Transformer Encoder:** The embeddings are processed via multi-head self-attention. This mechanism allows the model to capture global dependencies—recognizing, for instance, how a pitch rise at the beginning of a phrase relates to an intonation drop at the end—patterns often missed by the limited receptive field of a CNN.

C. Rationale for Transfer Learning in a Low-Resource Context

Our core hypothesis posits that the rich audio representations learned by PANNs and AST from massive datasets (like AudioSet) can be effectively transferred to the niche task of Urdu SER. This strategy is a deliberate response to data scarcity.

The UrduSER dataset contains approximately 3,500 samples [9]. In the deep learning landscape, this is a tiny dataset. Training complex models like Cnn14 or AST from scratch here would almost certainly result in severe overfitting. By leveraging pre-trained weights, we bypass the need for the model to learn "what sound is" from scratch. Instead, the task becomes one of adaptation rather than fundamental learning.

D. The Two-Stage Finetuning Protocol

To adapt these general-purpose models to our specific task without destroying their pre-trained knowledge, we utilized a structured two-stage finetuning protocol.

1) Stage 1: Classifier Adaptation (Head Training):

- **Objective:** To train the new classification head to map the existing, frozen backbone features to our specific emotion classes.
- **Process:** The pre-trained backbone is **frozen**. We replace the original AudioSet classifier with a new linear layer.
- **Training:** We train only the weights of this new linear layer using a relatively high learning rate (1×10^{-3}).

2) Stage 2: Full Network Finetuning (End-to-End):

- **Objective:** Once the head is stabilized, we subtly adapt the feature extraction layers themselves.
- **Process:** The entire network is **unfrozen**.
- **Training:** We use a *differential learning rate*. The backbone is updated very slowly (5×10^{-6}) to preserve learned features, while the head is refined at a slightly higher rate (1×10^{-4}).

VI. Experimental Design / Setup

Here, we detail the technical specifics of our data preparation, training configuration, and evaluation metrics.

A. Data Corpus and Preprocessing Pipeline

1) Corpus Definition: We utilize the **UrduSER** dataset (Version 3), developed by Akhtar et al. [9]. We selected this corpus specifically for the high quality and naturalism of its emotional expressions.

2) Corpus Properties: While the full dataset covers seven emotions, we focused our comparative experiments on a consolidated subset containing **2,400 samples**. This subset combines data from UrduSER and the Urdu Language Speech Dataset, focusing on the four primary emotions shared by both: **Angry, Happy, Neutral, and Sad**. This focus allows for a more rigorous comparison on high-confidence labels.

3) Data Partitioning (Speaker-Independent Split): To test the model's ability to generalize (Research Question 3), we implemented a strict **speaker-independent hold-out** strategy. We did not simply shuffle files; we separated the actors:

- **Training Partition:** 80% of speakers.
- **Test Set (Hold-out):** 20% of speakers.

The model is trained *only* on the first group and evaluated *only* on the second, ensuring it cannot memorize specific speaker characteristics.

4) Audio-to-Spectrogram Conversion: All audio files underwent an identical preprocessing pipeline:

1) **Resampling:** 32,000 Hz.
2) **STFT:** Window Length 1024, Hop Length 320.
3) **Mel-Filterbank:** 64 Mel bins, ranging from 50 Hz to 14,000 Hz.
4) **Normalization:** Per-spectrogram Z-score normalization.

B. Model Training and Hyperparameters

Experiments were conducted using PyTorch with the following settings:

- **Optimizer:** AdamW.
- **Loss Function:** Categorical Cross-Entropy Loss.
- **Regularization:** SpecAugment applied during Stage 2.

Table **??** outlines the specific configurations for each training stage.

C. Evaluation Metrics

We assess performance using:

- **Primary Metric:** Weighted F1-Score (Harmonic mean of Precision and Recall, weighted by class support).
- **Secondary Metrics:** Overall Accuracy (OA) and Unweighted Average Recall (UAR).
- **Qualitative Analysis:** Confusion Matrix for per-class error analysis.

VII. Experiments and Results

A. Experimental Setup and Dataset Consolidation

To ensure a robust analysis, we consolidated the UrduSER dataset [9] with the Urdu Language Speech Dataset, yielding a consistent corpus of **2,400 audio samples**. We focused on the four intersecting primary emotions: **Angry, Happy, Neutral, and Sad**.

Partitioning via an 80:20 split provided 1,920 samples for training and a held-out test set of 480 samples (120 per class). All training took place on a dual-GPU setup.

B. Baseline Results: PANNs (CNN14)

Our baseline PANNs (Cnn14) model achieved an overall **Accuracy of 65%** and a **Weighted F1-Score of 0.64** (Table III).

While the model was effective at identifying high-arousal negative emotions like "Angry" (F1: 0.76), it struggled significantly with the "Happy" class, managing a recall of only 39%. This suggests that the CNN's local receptive fields failed to capture the subtle prosodic nuances that distinguish happiness from other high-energy states.

TABLE III
Classification Report for Baseline PANNs (CNN14)

Emotion	Precision	Recall	F1-Score	Support
Angry	0.68	0.86	0.76	120
Happy	0.58	0.39	0.47	120
Neutral	0.64	0.62	0.63	120
Sad	0.65	0.72	0.68	120
Accuracy			**0.65**	480
Weighted Avg	0.64	0.65	0.63	480

C. Vision Transformer Results: AST

The Audio Spectrogram Transformer (AST) outperformed the baseline, validating the effectiveness of self-attention for this task. As shown in Table IV, the AST model achieved an overall **Accuracy of 75%** and a **Weighted F1-Score of 0.75**.

The most dramatic improvement appeared in the "Happy" class, where the F1-score jumped from 0.47 (CNN) to 0.64 (AST). The model also demonstrated excellent consistency on "Angry" (F1: 0.85) and maintained balanced performance across the Neutral and Sad categories.

Figure 1 illustrates the training stability of the AST model. The loss curve shows a steady decline, while the validation accuracy and F1-score rise in unison, peaking around the 25th epoch. This correlation indicates that the model is not merely optimizing for the majority class but is learning distinct features for all emotions.

D. Visual Validation and Efficiency Analysis

To further validate these metrics, we performed a visual error analysis. Figure 2 displays the confusion matrices for the AST model. The matrix shows strong diagonal consistency, particularly for "Angry" (104 correct) and "Sad" (97 correct). However, specific ambiguity persists between the "Happy" and

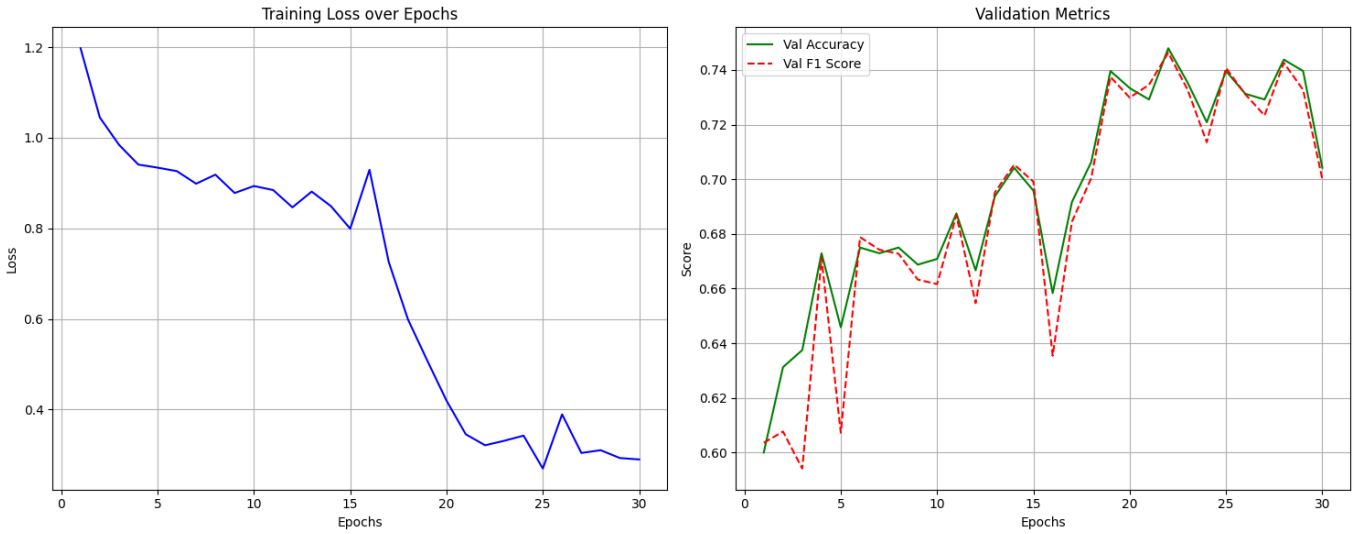

Fig. 1. Training dynamics over 30 epochs. The left panel shows the training loss convergence, while the right panel demonstrates the correlation between Validation Accuracy and F1-Score. The close tracking of F1 and Accuracy suggests a balanced learning process across emotion classes.

TABLE IV
CLASSIFICATION REPORT FOR VISION TRANSFORMER (AST)

Emotion	Precision	Recall	F1-Score	Support
Angry	0.83	0.87	**0.85**	120
Happy	0.65	0.62	**0.64**	120
Neutral	0.74	0.69	0.72	120
Sad	0.76	0.81	0.78	120
Accuracy			**0.75**	480
Weighted Avg	0.75	0.75	0.75	480

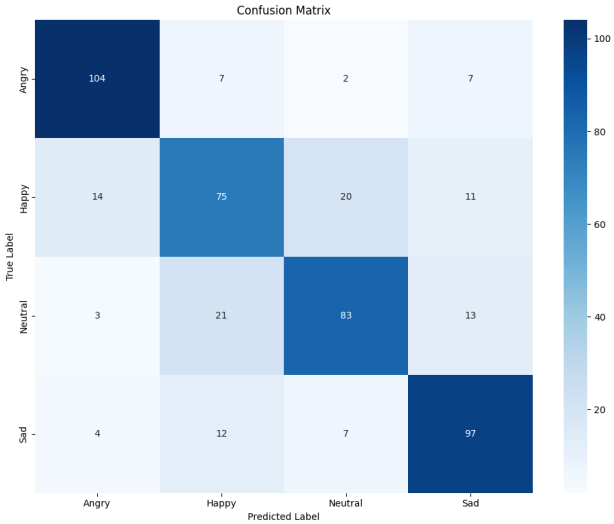

Fig. 2. Confusion Matrix for the AST model on the test set. Strong diagonal performance is observed for 'Angry' and 'Sad', while 'Happy' shows significant leakage into 'Neutral' (20 misclassified samples).

"Neutral" classes, where 20 "Happy" samples were misclassified as "Neutral", and 21 "Neutral" samples were misclassified as "Happy".

Regarding efficiency, the AST model (approx. 86M parameters) required about 2 hours and 8 minutes to converge (30 epochs), averaging 256 seconds per epoch. The PANNs model had a similar computational footprint. This indicates that the 10% accuracy gain offered by AST does not require a prohibitive increase in training time or resources compared to standard VGG-style CNNs.

VIII. DISCUSSION

A. Interpretation of Findings

These results provide a clear answer to our primary research question: Vision Transformers significantly outperform standard CNNs for Urdu Speech Emotion Recognition. The 10% gap in accuracy (75% vs. 65%) stems from the architectural differences.

CNNs rely on local convolution. They are excellent at finding specific frequency patterns—like the sharp pitch of a shout—but they struggle to integrate that information across the full duration of an utterance. The AST, utilizing self-attention, has a global receptive field. It can correlate the start of a sentence with the end, capturing the long-range prosodic contours—such as the rising intonation of happiness or the flat, slow rhythm of sadness—that define emotion in Urdu speech.

B. The Challenge of the "Happy" Class

A consistent challenge throughout our experiments was the "Happy" class. The baseline CNN recall for this emotion was remarkably low (39%). We hypothesize that this is because, spectrally, "Happy" shares high-energy characteristics with "Angry" (high amplitude, high pitch) and rhythmic traits with "Neutral" (especially in sarcastic or low-arousal contexts).

We investigated this further by visualizing specific high-confidence errors (Figure 3). As observed in the first and third

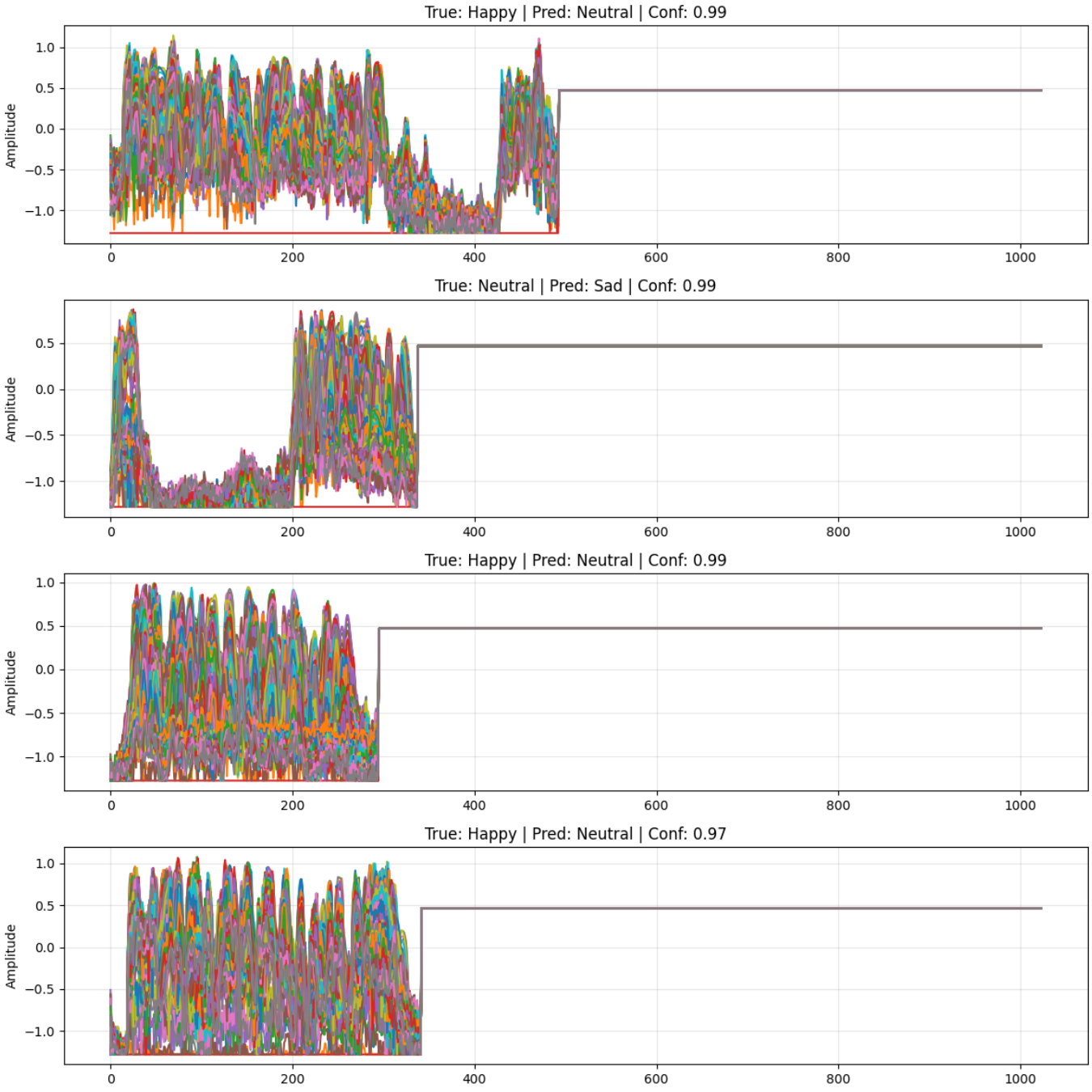

Fig. 3. Visual analysis of high-confidence errors. The samples show instances where 'Happy' audio was misclassified as 'Neutral' (and 'Neutral' as 'Sad') with near-certainty (¿0.97 confidence), highlighting the subtle boundary between these emotional states in natural speech.

panels, the model frequently predicts "Neutral" for "Happy" samples with extremely high confidence (¿0.99). This suggests that for these specific samples, the acoustic features of happiness—such as pitch variability—were likely too subtle, leading the model to interpret the speech as standard, non-emotional neutral speech.

C. Limitations and Future Work

While promising, this study has limitations. To create a balanced training set, we restricted our analysis to four primary emotions, excluding nuanced states like Disgust or Fear due to data inconsistency.

Future work will focus on two avenues:

- **Class Expansion:** Extending the AST architecture to the full 7-class spectrum by employing data augmentation to balance underrepresented classes.
- **Multimodal Integration:** Since even the best audio model capped at 75% accuracy, we propose integrating linguistic features (text/transcripts). A multimodal ViT processing both audio and text embeddings could help resolve ambiguous emotions like "Happy."

D. Conclusion

This work establishes a new baseline for Urdu SER. By demonstrating that ImageNet-style Vision Transformers can be effectively adapted to Urdu audio spectrograms, we provide a roadmap for moving beyond traditional CNNs. The clear superiority of the AST model suggests that research in low-resource languages should prioritize attention-based architectures that maximize the utility of limited data.

REFERENCES

[1] R. H. Aljuhani, A. Alshutayri, and S. Alahdal, "Alex Net-Based Speech Emotion Recognition Using 3D Mel-Spectrograms," *International Journal of Intelligent Systems and Applications in Engineering*, 2024.

[2] R. Dangol, S. P. Singh, and B. K. Balabantaray, "Speech Emotion Recognition Using CNN-LSTM and Vision Transformer," in *Innovations in Bio-Inspired Computing and Applications*, 2023.

[3] R. Mishra, A. Frye, M. M. Rayguru, and D. O. Popa, "Personalized Speech Emotion Recognition in Human-Robot Interaction using Vision Transformers," *arXiv preprint arXiv:2409.10687*, 2024.

[4] M. Mustafa, S. A. Khan, and S. A. Khan, "Speech Emotion Recognition Using Convolutional Neural Networks and Log-Mel Spectrograms," *International Journal of Scientific Research in Engineering and Technology*, 2024.

[5] R. V. Sharan, C. Mascolo, and B. W. Schuller, "Emotion Recognition from Speech Signals by Mel-Spectrogram and a CNN-RNN," in *2024 46th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2024.

[6] H. Li, J. Li, H. Liu, T. Liu, Q. Chen, and X. You, "MelTrans: Mel-Spectrogram Relationship-Learning for Speech Emotion Recognition via Transformers," *Sensors*, vol. 24, no. 17, p. 5506, 2024.

[7] R. Bautista, I. Diaz, A. Alarcon, and Y. Cardenas, "A Combined CNN Architecture for Speech Emotion Recognition," *Sensors*, vol. 24, no. 17, p. 5797, 2024.

[8] A. A. Raza, S. A. R. Rizvi, and S. A. R. Zaidi, "SEMOUR+: a Scripted EMOtional Speech Repository for Urdu," *Language Resources and Evaluation*, 2022.

[9] M. Z. Akhtar, R. Jahangir, and Q. U. Ain, "UrduSER: A Dataset for Urdu Speech Emotion Recognition," *Mendeley Data*, V3, 2024.

[10] Q. Ouyang, "Speech Emotion Detection Based on MFCC and CNN-LSTM Architecture," *arXiv preprint arXiv:2501.10666*, 2025.

[11] A. Asghar, S. Sohaib, S. Iftikhar, M. Shafi, and K. Fatima, "An Urdu speech corpus for emotion recognition," *PeerJ Computer Science*, vol. 8, p. e954, 2022.

[12] Z. S. Syed and S. A. Memon, "Introducing the Urdu-Sindhi Speech Emotion Corpus: A Novel Dataset of Speech Recordings for Emotion Recognition for Two Low-Resource Languages," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 4, 2020.

[13] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[14] J. F. Gemmeke et al., "AudioSet: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP*, 2017.

[15] Y. Gong, Y. A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proc. Interspeech*, 2021.