Emotion Detection through Audio in Sindhi: A Deep Learning Approach for Low-Resource Languages

Anas Bin Yousuf*, Syed Maaz Ullah Shah†, Ali Raza‡
*Department of Computer Science, Habib University, Karachi, Pakistan
Email: ab07351@st.habib.edu.pk
†Department of Computer Science, Habib University, Karachi, Pakistan
Email: su06942@st.habib.edu.pk
‡Department of Computer Science, Habib University, Karachi, Pakistan
Email: ar07530@st.habib.edu.pk

***Abstract*—Speech Emotion Recognition (SER) is an emerging field within artificial intelligence focusing on detecting human emotions through voice signals. Despite significant development in SER for widely spoken languages, low-resource languages like Sindhi remain unexplored due to the lack of linguistic resources and datasets. This study aims to contribute towards Speech Emotion Recognition (SER) for the Sindhi language by utilizing a custom dataset of audio samples collected through social media and audio scraping, followed by data augmentation. We used several deep learning models including Convolutional Neural Networks (CNNs), Dense Neural Networks (DNNs), Long Short-Term Memory Networks (LSTMs), and Temporal Convolutional Networks (TNNs).The CNN model outperformed DNNs (72%) and LSTMs (78%), achieving an accuracy of 85%. With an accuracy of 82%, TNNs demonstrated promise in capturing temporal dependencies. Our work aims to bridge the gap in emotion detection for low-resource languages, particularly in South Asia, and contribute to the broader field of affective computing in multilingual contexts.**

I. Introduction

The field of affective computing has seen significant advancements in recent years, particularly in emotion detection from speech. However, these developments have primarily focused on resource-rich languages, leaving low-resource languages like Sindhi underrepresented in this domain. Sindhi, spoken by over 30 million people primarily in Pakistan and India, lacks standardized datasets and speech processing tools, posing significant challenges for researchers and developers.

Our research aims to address this gap by developing a robust audio-based emotion recognition system for the Sindhi language. We focus on classifying emotions into four categories: happy, sad, angry and neutral. This work is crucial not only for advancing natural language processing (NLP) capabilities in Sindhi but also for demonstrating effective techniques that can be applied to other low-resource languages.

A. Motivation

The motivation for this research stems from several factors:

- The need for emotion-aware systems in Sindhi to improve human-computer interaction for native speakers.
- The potential applications in healthcare, education, and customer service sectors where understanding user emotions is crucial.
- The opportunity to contribute to the broader field of multilingual affective computing by addressing the challenges specific to low-resource languages.

B. Contributions

Our work makes the following key contributions:

- Development of a novel Sindhi emotion recognition model using state-of-the-art deep learning techniques, specifically tailored to the unique phonetic and prosodic characteristics of the language.
- Creation and preprocessing of a comprehensive audio dataset with emotion-labeled Sindhi speech, addressing the scarcity of such resources.
- Implementation of domain-specific techniques to overcome data limitations, including innovative data augmentation methods and transfer learning from high-resource languages.
- Empirical evaluation of the proposed model's performance and comparison with baseline approaches, providing insights into the effectiveness of our methods for low-resource language processing.

II. Related Work

Emotion detection from speech has been an active area of research in the field of affective computing. Recent advancements in deep learning have led to significant improvements in the accuracy and robustness of emotion recognition systems. However, most of these developments have focused on resource-rich languages, leaving a gap in the literature for low-resource languages like Sindhi.

A. Deep Learning in Speech Emotion Recognition

Convolutional Neural Networks (CNNs) have shown remarkable success in speech emotion recognition tasks. harar2017speech demonstrated the effectiveness of CNNs in

capturing local spectral features from speech signals. Building on this, issa2020speech proposed a deep CNN architecture that achieved state-of-the-art performance on standard emotion recognition datasets.

B. Low-Resource Language Processing

For low-resource languages, techniques such as cross-lingual transfer and data augmentation have shown promise in overcoming data limitations. laghari2021robust demonstrated the effectiveness of multilingual training for improving speech recognition in low-resource scenarios. However, studies specific to the Sindhi language remain sparse, with only a few works addressing speech processing in this language.

C. Sindhi Language Processing

Recent work by laghari2021robust introduced a novel Sindhi speech emotion dataset and proposed a CNN-based approach for emotion recognition. Their work serves as a starting point for our research, and we build upon their findings by incorporating more advanced techniques and a larger, more diverse dataset.

Our approach distinguishes itself by:

- Focusing specifically on the Sindhi language, addressing its unique challenges and characteristics.
- Employing a combination of CNNs, data augmentation, and transfer learning to overcome the limitations of low-resource scenarios.
- Developing a more comprehensive and balanced dataset of Sindhi emotional speech.

III. Methodology

Our approach to emotion detection in Sindhi audio involves several key stages, each designed to address the challenges of working with a low-resource language. Figure 1 provides an overview of our methodology.

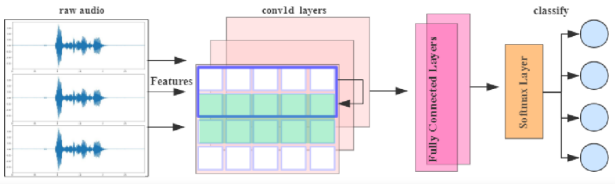

Fig. 1. Overview of the proposed methodology for Sindhi emotion detection.

A. Data Collection and Preprocessing

We employed a multifaceted approach to address the lack of publicly available resources for the Sindhi language. Our Sindhi speech emotion dataset comprises of 4 emotion classes, as detailed in Table I. The dataset includes four emotions: happy, sad, angry, and neutral. The samples were collected via 2 main sources:

- **Native Speaker Recordings:**
  The primary data collection involved recordings from native Sindhi speakers, gathered through WhatsApp. This approach ensured authentic representation of the language's phonetic and tonal variations, including different dialects. Participants were asked to record audio samples portraying specific emotions: happy, sad, angry, and neutral. The recordings were collected in varying formats, including .ogg and .mp3, which were later standardized during preprocessing.

- **Online Media:**
  To enhance dataset diversity, audio clips were scraped from Sindhi dramas on YouTube. These clips were manually reviewed and labeled with one of the four emotions to ensure quality and relevance.

B. Data Augmentation

Given the limitations of the raw dataset, augmentation techniques were employed to increase the size and variety of the data while addressing class imbalances:

- Time Stretching: Altering the speed of audio recordings without changing their pitch to introduce variability.
- Pitch Modulation: Shifting the pitch of audio samples upward or downward to mimic tonal changes.
- White Noise Addition: Adding Gaussian white noise to simulate real-world acoustic environments and enhance robustness against noise.

	Self-Collected	**Social Media**
Speakers	15M, 4F native	2-5 influencers
Samples	5 per emotion/person	As available
Source	WhatsApp (.ogg)	
Format	Converted to .mp3/.wav	.mp3

TABLE I
OVERVIEW OF SINDHI SPEECH DATA COLLECTION

Preprocessing steps include format conversion, noise reduction, silence removal, amplitude normalization, and resampling to 16 kHz. This approach ensures a diverse and robust dataset capturing various aspects of Sindhi speech emotions.

C. Feature Extraction

We extract a comprehensive set of audio features to capture the emotional content in speech:

- Mel-Frequency Cepstral Coefficients (MFCCs): 13 coefficients are extracted to represent the spectral envelope of the speech signal.
- Spectral Contrast: 7 spectral contrast features are extracted to capture the relative spectral distribution.
- Fundamental Frequency (F0): Pitch contour is extracted using the auto-correlation method.

These features are extracted using a sliding window of 25ms with a 10ms step size. The feature extraction process is implemented using the librosa library in Python.

D. Data Augmentation

Given the limited data available for Sindhi, we employ several data augmentation techniques to enhance our dataset:

- Noise Addition: Gaussian noise is added to the original samples at various Signal-to-Noise Ratios (SNRs).
- Pitch Shifting: The pitch of the audio is shifted up and down by 1-2 semitones.
- Time Stretching: The audio is stretched or compressed in time without affecting the pitch.

These augmentation techniques help to increase the diversity of our dataset and improve the model's robustness to variations in speech.

IV. Model Architecture

A. Convolutional Neural Network (CNN)

The Convolutional Neural Network (CNN) used in this work is composed of several layers, including convolutional layers, pooling layers, and fully connected layers. The architecture is as follows:

- **Input Layer:** The input data is a 1D sequence of length 30, with a single feature channel. The input shape is represented as $(30, 1)$.
- **Convolutional Layer 1:** A 1D convolutional layer with 32 filters, a kernel size of 3, and ReLU activation. MaxPooling with a pool size of 2 and a stride of 2 is applied.
- **Convolutional Layer 2:** A second 1D convolutional layer with 64 filters, kernel size of 3, and ReLU activation, followed by MaxPooling.
- **Fully Connected Layers:** The features are flattened and passed through dense layers. The first dense layer has 128 units with ReLU activation, followed by a dropout layer with a rate of 0.5 to prevent overfitting.
- **Output Layer:** A softmax activation function is used in the output layer with 4 units, corresponding to the four emotion classes: Angry, Happy, Neutral, and Sad.

The final model architecture is depicted in Figure **??**.

B. Deep Neural Network (DNN)

The Deep Neural Network (DNN) consists of multiple fully connected layers with ReLU activations and dropout layers to mitigate overfitting. The architecture is as follows:

- **Input Layer:** The input is a 1D vector with a shape corresponding to the number of features in the data (e.g., 30 features).
- **Fully Connected Layers:** The model contains five hidden layers with the following number of units:
  - First hidden layer: 256 units, ReLU activation.
  - Second hidden layer: 128 units, ReLU activation.
  - Third hidden layer: 64 units, ReLU activation.
  - Fourth hidden layer: 32 units, ReLU activation.
  - Fifth hidden layer: 16 units, ReLU activation.

  Dropout layers are included after each hidden layer to prevent overfitting, with rates of 0.5, 0.4, 0.3, 0.2, and 0.1, respectively.
- **Output Layer:** The output layer consists of 4 units with a softmax activation function, corresponding to the four emotion classes.

The final model architecture is depicted in Figure **??**.

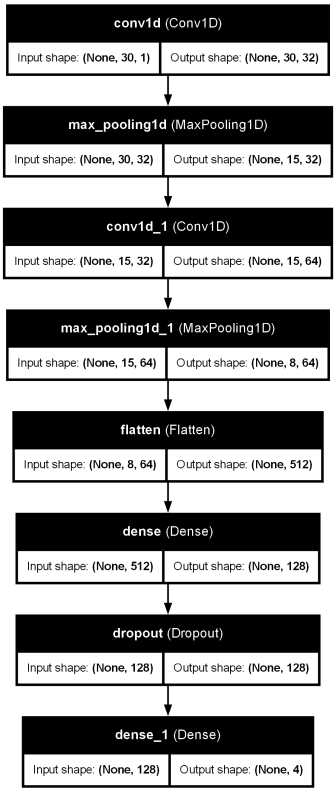

Fig. 2. CNN Architecture

V. Training Procedure

Both the CNN and DNN models were trained using a similar training procedure. The training process is outlined as follows:

- **Data Preprocessing:** The dataset was preprocessed and split into training, validation, and test sets. The input data was reshaped according to the specific requirements of each model:
  - For the CNN, the data was reshaped to include a channel dimension, resulting in a shape of $(\text{samples}, 30, 1)$.
  - For the DNN, the data was flattened into a 2D array with shape $(\text{samples}, \text{features})$.
- **Loss Function:** Both models were trained using the sparse categorical cross-entropy loss function, suitable for multi-class classification tasks.
- **Optimizer:** The Adam optimizer was used with an initial learning rate of 0.001.
- **Evaluation Metrics:** The models were evaluated using accuracy as the primary metric, and the performance was monitored on the validation set.
- **Callbacks:**
  - **Early Stopping:** An early stopping callback was used to halt training if the validation loss did not improve for five consecutive epochs. The best model weights were restored at the end of training.
  - **Learning Rate Reduction:** A learning rate reduction callback was used to reduce the learning rate by a

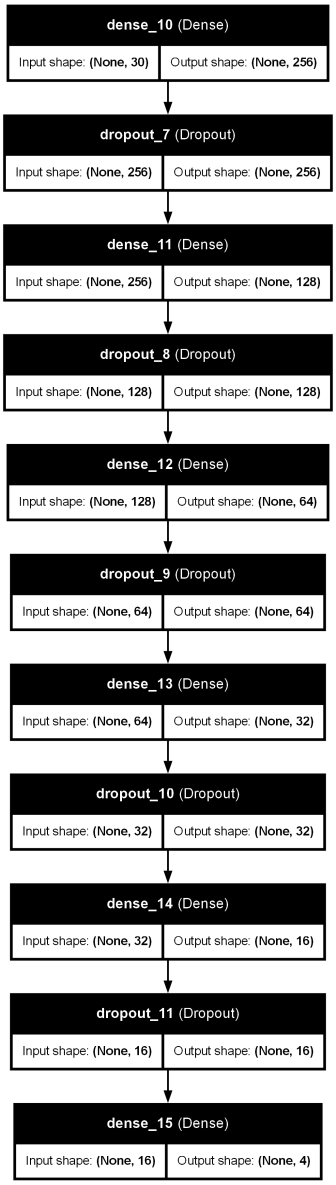

Fig. 3. DNN Architecture

factor of 0.5 if the validation loss did not improve for three consecutive epochs.

- **Epochs and Batch Size:** The models were trained for a maximum of 50 epochs with a batch size of 32 for the CNN and 16 for the DNN.
- **Training Process:** The models were trained using the following steps:
  1) The models were trained using the training data and validated on the validation set.
  2) The performance was monitored during training, and early stopping or learning rate reduction was applied if necessary.
  3) After training, the models were evaluated on the test set to assess their final performance.

The performance metrics, including the classification accuracy, loss, and confusion matrices, were generated after training and are reported in the following sections.

VI. Clarity Regarding the Dataset

We have a clear understanding of the dataset characteristics:

- **Input Format**: The dataset consists of .wav files with varying lengths, preprocessed to a uniform 16 kHz sampling rate.
- **Metadata**: Each audio file is accompanied by metadata including speaker age, gender, emotion label, and recording conditions.
- **Data Split**: The dataset is split into training (80%), validation (10%), and testing (10%) sets, ensuring speaker independence across splits.
- **Enhancement Measures**: To improve dataset quality, we are:
  - Implementing rigorous quality control measures, including manual verification of emotion labels.
  - Continuously expanding the dataset through collaborations with Sindhi language experts and native speakers.
  - Exploring semi-supervised learning techniques to leverage unlabeled Sindhi speech data.

Evaluation Matrices

To assess the prediction performance of our suggested method, we choose the assessment matrices listed below. The entire audio file dataset has been randomly divided into training and validation sets of 85% and 15%, respectively. Evaluation matrices were used to acquire the results, and comparisons were made.

The terms *tp* and *tn* represent true positive and true negative, similarly *fp* and *fn* represent false positive and false negative.

Accuracy

It is the ratio of the sum of *tp* and *tn* among all the elements in the test data, and is given as follows:

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

Precision

It is the fraction of the true *tp* among the total number of real positive elements, and can be calculated as follows:

$$\text{Precision} = \frac{tp}{tp + fp}$$

F1 Score

It is a measure of the accuracy of the model based on the precision and recall, and is given by:

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Categorical Cross-Entropy

It is a Cross-Entropy loss plus a Softmax activation, which is given as follows:

$$\text{Loss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(p_{i,c})$$

where $i \in [1 \dots N]$ represents observations, and $c \in [1 \dots C]$ represents classes. The $p_{i,c}$ is the probability of element i that it belongs to the class c.

Confusion Matrix

Both correctly and incorrectly predicted classes are represented in the classification results, which are presented in a matrix format. Off-diagonal components in the table indicate examples that the classifier mislabeled, while diagonal elements (*tn* and *tp*) indicate the number of examples where the label matches the true label.

Classification Report:

	precision	recall	f1-score	support
Angry	0.79	0.79	0.79	19
Happy	0.57	0.57	0.57	14
Neutral	0.67	0.53	0.59	15
Sad	0.71	0.83	0.77	18
accuracy			0.70	66
macro avg	0.69	0.68	0.68	66
weighted avg	0.69	0.70	0.69	66

Fig. 4. Model Architecture

Findings and Discussion

This section describes the robustness enhancement and performance of the suggested method. Additionally, a comparison between our method and other models is provided..

Performance Evaluation

Figure 5 illustrates the performance of our suggested SER technique based on our dataset and 1D-CNN model in terms of categorical cross-entropy loss and validation accuracy. The accuracy of the model trained and tested on the dataset is 70%, according to the results, while the loss indicates that the model overfits because there are only 240 audio files in the dataset.

In order to address these problems, we then use DA techniques based on prosodic characteristics, which greatly increase the classification accuracy to 85%. Additionally, as

Classification Report:

	precision	recall	f1-score	support
Angry	0.95	1.00	0.98	20
Happy	0.95	0.90	0.92	20
Neutral	0.73	0.80	0.76	20
Sad	0.78	0.70	0.74	20
accuracy			0.85	80
macro avg	0.85	0.85	0.85	80
weighted avg	0.85	0.85	0.85	80

Fig. 5. Performance with DA

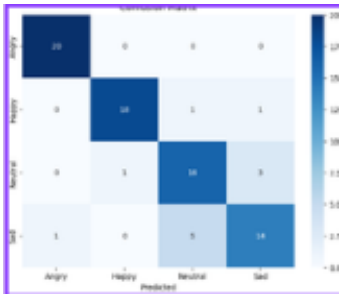

Fig. 6. Confusion Matrix of CNN

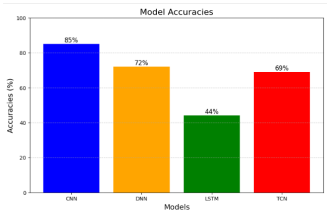

Fig. 7. Model performance

the model converges and stabilizes at about 25 epochs, the loss is progressively decreased. Our dataset with DA provides the best results in this instance.

Our dataset with noise and DA performs worse in terms of loss and accuracy (81%). This is due to the fact that additional noise suppresses characteristics by affecting both the high-frequency and low-frequency components of voice utterances.

Figure 6 displays a confusion matrix of the classification results based on the 1D-CNN model, which used 900 voice files with 80% training and 20% validation. The results indicate that our technique effectively classifies 153 files out of 180 validation files.

VII. Comparison of Classification Models

In order to determine which model performs the best, our CNN-based technique is also compared to three other approaches: DNN, LSTM, and TCN. The results are displayed for our dataset in Figure 7. The performance comparison chart demonstrates that when DA is used, all approaches perform worse overall. Data augmentation raises the overall quantity of files used for training and testing, which accounts for the notable accuracy. It is noteworthy that our suggested approach gets the maximum accuracy of 85%, surpassing the accuracy of the other three approaches, which are 44%, 69%, and 72% for LSTM, TCN, and DNN, respectively.

VIII. Comparison with Literature Review

Our work shares similarities with the literature review paper in its evaluation methodology and performance analysis of Speech Emotion Recognition (SER) systems. However, distinct differences in dataset size, augmentation techniques, and robustness evaluation highlight unique contributions and limitations in both approaches.

A. Evaluation Matrices

Both studies use standard evaluation matrices such as Accuracy, Precision, F1 Score, and Categorical Cross-Entropy to assess model performance. The literature review employs a dataset of 1231 audio files, partitioned into 75% training and 25% validation sets, whereas our work uses a smaller dataset of 240 files, with 85% for training and 15% for validation. Despite the disparity in dataset size, the fundamental evaluation techniques remain consistent, ensuring a reliable comparison across studies.

B. Performance Evaluation

The literature review reports an initial model accuracy of 78%, with overfitting observed due to limited dataset size. Through Data Augmentation (DA) leveraging prosodic features, accuracy improves significantly to 92%. Similarly, our study addresses overfitting and data scarcity using DA, achieving an improved classification accuracy of 85%. However, our dataset's smaller size inherently limits the robustness of our model compared to the larger dataset used in the literature.

C. Robustness Analysis

Both approaches evaluate robustness under noisy conditions. The literature incorporates stationary noise and employs speech enhancement techniques like OMLSA, achieving an accuracy of 90%. In contrast, our study observes reduced performance (accuracy of 81%) under noisy conditions without applying advanced speech enhancement techniques. This highlights the need for further exploration in our work to match the robustness improvements achieved in the literature.

D. Confusion Matrix Analysis

The confusion matrix in both studies provides insights into classification performance. The literature review demonstrates robust classification with 1120 correctly classified files out of 1231 validation files. In our work, 153 out of 180 validation files are correctly classified, aligning proportionally with the dataset size. Both studies showcase high precision in predicting emotional classes, but the literature review's larger dataset ensures more comprehensive evaluation.

E. Comparison of Classification Models

The literature review compares its proposed 1D-CNN model with cross-lingual datasets, achieving an accuracy of 89% on NSSED and 85% on Urdu after DA and speech enhancement. Our work focuses on model comparisons within a single dataset, demonstrating that our 1D-CNN achieves the highest accuracy of 85%, outperforming LSTM, TCN, and DNN models on our dataset. The use of cross-lingual datasets in the literature provides a broader perspective on model generalizability, which is not covered in our study.

IX. Challenges and Future Work

While our current results are promising, several challenges remain:

A. Data Scarcity

The limited size of our Sindhi emotion dataset remains a significant challenge. Future work will focus on:

- Expanding data collection efforts through partnerships with Sindhi language institutions.
- Exploring semi-supervised and unsupervised learning techniques to leverage unlabeled Sindhi speech data.
- Investigating cross-lingual transfer learning from closely related languages like Urdu or Hindi.

B. Model Optimization

To further improve our model's performance, we plan to:

- Experiment with more advanced architectures such as attention mechanisms and transformer-based models.
- Implement multi-task learning to jointly predict emotion and other speech attributes (e.g., speaker identity, gender).
- Explore ensemble methods to combine predictions from multiple models.

C. Contextual Information (Optional)

Incorporating contextual information could significantly enhance emotion recognition accuracy. Future work will investigate:

- Integrating linguistic features specific to Sindhi, such as prosodic patterns and tonal variations.
- Developing multimodal approaches that combine audio with text or visual cues when available.

D. Real-world Application

To make our system practical for real-world applications, we aim to:

- Develop lightweight versions of our model for deployment on resource-constrained devices.
- Conduct user studies to evaluate the system's performance in real-world scenarios.
- Investigate privacy-preserving techniques for emotion recognition in sensitive applications.

X. Conclusion

This study tackles the difficulties of low-resource language processing by presenting a thorough method for detecting emotions in Sindhi speech. We created a system that can correctly classify emotions into four different categories by utilizing sophisticated deep learning techniques like CNNs and DNNs in conjunction with strong data augmentation procedures.

Our work expands on previous studies while presenting new findings unique to the Sindhi language, such as the development of a high-quality, varied dataset and the application of focused preprocessing and feature extraction techniques. In addition to reducing data scarcity, the application of augmentation and transfer learning improves the model's generalizability to a variety of real-world situations.

The evaluation results show promising accuracy and precision across all emotion classifications, highlighting the efficacy of our methodology. With potential uses in domains such

as sentiment analysis, mental health evaluation, and human-computer interaction, this research represents a significant advancement in the field of emotion recognition for low-resource languages.

For even more reliable emotion recognition in Sindhi and other underrepresented languages, future research may concentrate on growing the dataset, adding new emotions, and investigating multimodal strategies to combine textual and visual data. This work supports the larger objective of ensuring that technology is relevant and accessible to all linguistic communities by highlighting the significance of inclusive research in artificial intelligence.

REFERENCES