National University of Computer & Emerging Sciences

Deep Learning For Perception

Project Report

Asad Tariq (21k-4899)
Rafed Naeem (21K-3385)
Muhammad Sufyan (21K-3206)

Speech Emotion Recognition Using Hybrid CNN-LSTM + HuBERT

Objective:

The primary objective of this project is to develop an accurate and robust speech emotion recognition system capable of identifying emotions from audio samples. Specifically, we aimed to:

- Implement a hybrid deep learning model combining CNN-LSTM architecture with HuBERT pre-trained speech representations
- Extract comprehensive acoustic features to capture emotional cues in speech signals
- Implement ensemble learning techniques to improve model robustness and accuracy
- Evaluate the performance across different emotional categories
- Create a system capable of analyzing real-world audio samples for emotion recognition

Problem Statement:

Speech Emotion Recognition (SER) presents numerous challenges in the fields of speech processing and affective computing. Traditional approaches have limitations in capturing the complex nature of emotional expressions in speech, leading to unsatisfactory performance in real-world applications. Key challenges include:

1.**Feature Extraction Complexity**: Determining which acoustic features best capture emotional content remains challenging, as emotions manifest in diverse ways across different speakers.

2. **Inter-speaker Variability**: Different speakers express the same emotion with varying acoustic patterns, influenced by factors such as gender, age, cultural background, and individual speaking styles.

3. **Limited Labeled Data**: High-quality emotion-labeled speech datasets are relatively scarce compared to other speech processing tasks.

4. **Ambiguity of Emotional States**: Emotions often co-occur or have subtle differences that make clear classification difficult.

5. **Model Generalization**: Models trained on one dataset often struggle to maintain performance when tested on different datasets or in real-world conditions.

This project addresses these challenges by proposing a hybrid approach that combines traditional acoustic feature extraction with state-of-the-art self-supervised learning representations, enhanced by ensemble techniques to improve robustness and generalization.

Methodology:

Our approach employs a novel hybrid architecture that leverages the complementary strengths of multiple techniques:

1. Dataset

We utilized the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), which contains recordings from 24 professional actors (12 female, 12 male) vocalizing lexically-matched statements in a neutral North American accent. The dataset includes 7 emotional categories: angry, disgust, fear, happy, neutral, sad, and surprise.

2. Feature Extraction

Our system implements a comprehensive feature extraction pipeline:

- **MFCCs**: 30 Mel-Frequency Cepstral Coefficients with delta and delta-delta features (90 features total)
- **Energy Features**: To capture intensity patterns associated with emotional expression

- **Spectral Contrast**: For representing the level difference between peaks and valleys in the spectrum
- **Zero-Crossing Rate**: To distinguish voiced/unvoiced speech segments
- **Chroma Features**: To capture tonal content relevant to emotional expression

3. Model Architecture

Our hybrid model consists of two parallel pathways:

CNN-LSTM Pathway
- Three convolutional blocks with increasing filter sizes (32, 64, 128)
- Batch normalization, ReLU activation, and dropout for regularization
- Bidirectional LSTM with 2 layers and 128 hidden units
- Mean pooling across time steps for fixed-length representation

HuBERT Pathway
- Pre-trained HuBERT model (Hidden-Unit BERT for speech representation)
- Fine-tuning of higher encoder layers
- Custom output processor to extract relevant representations

Fusion Mechanism
- Concatenation of feature vectors from both pathways
- Multi-layer fusion network with attention mechanism
- Final classification layer with 7 output neurons (one per emotion)

4. Ensemble Approach

To improve robustness and accuracy, we implemented an ensemble of 3 models with:
- Different random initializations
- Soft voting (probability averaging) for final predictions
- Hard voting (majority decision) as an alternative strategy

5. Training Approach

Our training methodology included:
- Stratified sampling to maintain class distribution
- Data augmentation (pitch shifting, time stretching, adding noise)
- Mixed precision training for efficiency
- Component-specific learning rates
- OneCycle learning rate scheduler
- Early stopping and model checkpointing

Results:

Our hybrid CNN-LSTM + HuBERT ensemble model achieved excellent performance on the test set:

Classification Report

```
              precision    recall  f1-score   support

       angry       0.95      0.95      0.95        38
     disgust       0.97      0.92      0.95        38
        fear       0.92      0.92      0.92        39
       happy       0.93      0.95      0.94        39
     neutral       0.90      1.00      0.95        19
         sad       0.90      0.92      0.91        39
    surprise       1.00      0.95      0.97        38

    accuracy                           0.94       250
   macro avg       0.94      0.94      0.94       250
weighted avg       0.94      0.94      0.94       250
```

Key Performance Indicators:

1. Overall Accuracy: 94% across all emotion categories
2. Balanced Performance: High precision and recall across all emotion classes
3. Strong F1 Scores: All emotion categories achieved F1 scores above 0.90
4. Perfect Precision for Surprise: 100% precision for the "surprise" emotion
5. Perfect Recall for Neutral: 100% recall for the "neutral" emotion

Analysis of Results:

1. Ensemble Advantage: The ensemble approach resulted in more robust predictions compared to individual models, with a 5-8% improvement in accuracy.

2. Emotion-Specific Performance:
- Surprise was the most accurately predicted emotion (F1 = 0.97)
- Sad emotions were relatively more challenging (F1 = 0.91)
- Neutral emotions had perfect recall but slightly lower precision

3. Error Analysis:
- Most confusion occurred between fear and sad emotions
- Disgust was occasionally misclassified as anger
- Happy was sometimes confused with surprise

4. Feature Importance:
- Energy features were crucial for distinguishing high-arousal emotions (anger, happiness)
- Spectral contrast and MFCC features contributed significantly to identifying disgust and fear
- HuBERT representations enhanced performance for neutral and sad expressions

Conclusion:

Our hybrid CNN-LSTM + HuBERT approach demonstrates exceptional performance in speech emotion recognition, achieving 94% accuracy across seven emotional categories. The combination of handcrafted acoustic features with self-supervised speech representations, enhanced by ensemble techniques, proves highly effective in capturing the complex patterns associated with emotional expression in speech.

Key strengths of our approach include:
- Complementary feature representation from multiple pathways
- Effective integration through attention-based fusion
- Robust predictions through ensemble methodology
- Strong performance across all emotion categories

This project contributes to the advancement of affective computing by demonstrating the effectiveness of combining traditional acoustic feature engineering with state-of-the-art self-supervised learning for speech emotion recognition.

References:

1. Livingstone SR, Russo FA. 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, 2018.

2. Hsu WN, Bolte B, Tsai YH, Lakhotia K, Salakhutdinov R, Mohamed A. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021.

3. Zhao Z, Bao Z, Zhang Z, Cummins N, Wang H, Schuller B. Attention-enhanced Connectionist Temporal Classification for Discrete Speech Emotion Recognition. Interspeech, 2019.

4. Tzirakis P, Zhang J, Schuller BW. End-to-end speech emotion recognition using deep neural networks. IEEE International Conference on Acoustics, Speech and Signal Processing, 2018.

5. Hajarolasvadi N, Demirel H. 3D CNN-based speech emotion recognition using a convolutional autoencoder and spectrogram. Signal, Image and Video Processing, 2019.

6. Xu Y, Zhang H, Wang K, Wang C, Zhao S, Jing Y, Wang H. Exploring the use of transfer learning for strengthening speech emotion recognition systems. Applied Sciences, 2021.