Speech Emotion Recognition using Classical Machine Learning

WIDS (2025-26)

Sriya Akepati[1]

[1] Mentors: Krishna Kukreja and Garvit Meena

Motivation

- Human speech encodes emotion through tone, pitch, and energy
- Emotion-aware systems improve human–computer interaction
- Goal: understand how far classical ML can go in Speech Emotion Recognition

Project Objective

- Build an end-to-end Speech Emotion Recognition (SER) pipeline
- Extract meaningful acoustic features from raw audio
- Evaluate classical ML models and analyze their limitations

Dataset

RAVDESS: Ryerson Audio-Visual Database of Emotional Speech and Song.

- Speech part of the data set has been used for the project
- Studio-recorded emotional speech
- Multiple speakers(12 Male+12 Female), balanced emotion classes(8)
- Emotions: Angry, Happy, Sad, Neutral, Calm, Fearful, Disgust, Surprised
- Total of 1440 audio clips

Overall Pipeline

1. Raw audio input (.wav)
2. Acoustic feature extraction
3. Feature scaling
4. Model training and validation
5. Emotion prediction with confidence scores

Week 3: Feature Engineering

- Extracted **81 features** per audio sample
- Features included:
  - MFCCs
  - Chroma features
  - Spectral contrast
  - Tonnetz
- Captures spectral and harmonic characteristics of speech

Feature Matrix and Data Split

- Feature matrix shape: **(1440, 81)**
- Dataset split:
  - Training: 1008 samples
  - Validation: 216 samples
  - Test: 216 samples
- Fixed splits ensured consistent evaluation

Week 4: Baseline Models

- Feature normalization using **StandardScaler**
- Scaler saved for consistent inference
- Models evaluated:
    - Support Vector Machine (SVM)
    - Random Forest (RF)

Baseline Performance

Model	Validation Accuracy
SVM	62.96%
Random Forest	57.40%

- SVM performed better on handcrafted features
- Margin-based learning suited the feature space

Week 5: Hyperparameter Tuning

- Used GridSearchCV with 5-fold cross-validation
- Focused on SVM with RBF kernel
- Parameters explored:
    - Regularization parameter C
    - Kernel coefficient γ

Best Tuned Configuration

- Kernel: RBF
- $C = 10$
- $\gamma = 0.001$
- Best CV score: 0.4886

Tuned Model Performance

- Tuned SVM validation accuracy: **59.26%**
- Performance dropped compared to baseline
- Indicates overfitting during cross-validation

Key Insight

- Hyperparameter tuning did not improve performance
- Feature representation dominated model behavior
- Classical SER is limited by handcrafted features

Inference on Unseen Audio

- Model tested on unseen RAVDESS samples
- Outputs emotion probabilities, not just labels
- Enables qualitative analysis of predictions

Sample Predictions

- Angry: 84.74% confidence
- Happy: 77.87% confidence
- Surprised: 65.01% confidence
- Strong emotions classified confidently
- Subtle emotions showed uncertainty

Observed Confusions

- Neutral, Calm, and Sad frequently overlapped
- Indicates acoustic similarity between low-arousal emotions
- Limitation of static spectral features

Limitations

- No temporal modeling of speech dynamics
- Limited dataset size for fine-grained emotions
- Speaker variability affects neutral classes

Future Work

- CNN/LSTM models on spectrograms
- Data augmentation and noise robustness
- Speaker normalization
- Emotion-wise confusion analysis

Conclusion

- Built a complete SER pipeline from raw audio to inference
- Baseline models outperformed tuned models, needs more work to be put
- Project emphasized understanding over metric chasing
- Currently working on live speech emotion recognition and Improving the model accuracies

— Thank you —