SPEECH UNDERSTANDING

Emotion Detection in Speech
A Comparative Study of State-of-the-Art Approaches

PRANJAL MALIK (M24CSA021) GitHub
JYOTISHMAN DAS (M24CSA013) GitHub

Introduction

Task Definition. Emotion detection in speech aims to classify spoken audio into predefined emotional categories such as happy, sad, angry, calm, surprised, etc. This task is essential in human-computer interaction, healthcare, customer service, and security applications.

Real-World Importance

- Virtual Assistants: Enhances user experience by adapting responses based on emotional state.
- Healthcare: Supports mental health diagnosis and emotional well-being monitoring.
- Customer Service: Detects customer sentiment for improved interactions.
- Security & Surveillance: Identifies stress or aggression in security applications.

State-of-the-Art Models for Emotion Detection

Traditional Machine Learning Approaches

- Support Vector Machines (SVM), Random Forest, and k-NN models trained on hand-crafted features such as MFCCs, spectral, and prosodic features.
- Strengths: Computationally efficient and interpretable.
- Limitations: Requires extensive feature engineering and lacks robustness for real-world applications.

Deep Learning-Based Approaches

- CNN-Based Models
  - Model Example: CNN+LSTM Hybrid
  - Method: Uses spectrograms as input to convolutional layers followed by LSTMs for temporal feature extraction.
  - Strengths: Captures both spatial and temporal dependencies.
  - Limitations: Requires large amounts of labeled data.
- Recurrent Neural Networks (RNNs) & Transformers
  - Model Example: BiLSTM, GRUs, Wav2Vec2.0
  - Method: Processes raw audio features or MFCCs using LSTMs or transformers for sequential learning.
  - Strengths: Handles sequential dependencies efficiently.
  - Limitations: Prone to overfitting and computationally expensive.
- Self-Supervised Learning (SSL) Approaches
  - Model Example: Wav2Vec2.0 (Meta), HuBERT, Whisper
  - Method: Learns representations from raw audio without labeled data and fine-tunes on emotion datasets.
  - Strengths: Requires less labeled data, generalizes better.
  - Limitations: Large-scale models demand high computational resources.

Comparative Analysis of SOTA Models

Model	**Feature Extraction**	**Strengths**	**Limitations**
SVM	MFCCs, Spectral	Simple, interpretable	Needs manual features
CNN+LSTM	Spectrograms	Captures spatial & temporal info	High data requirement

BiLSTM	MFCCs, Raw Audio	Sequential modeling	Overfitting risk
Wav2Vec2.0	Raw Waveforms	Generalization, less labeling needed	Computationally expensive
HuBERT	Self-supervised	High accuracy, unsupervised learning	Needs fine-tuning
Whisper	Large-Scale Transformer	Robust multilingual support	Large model size

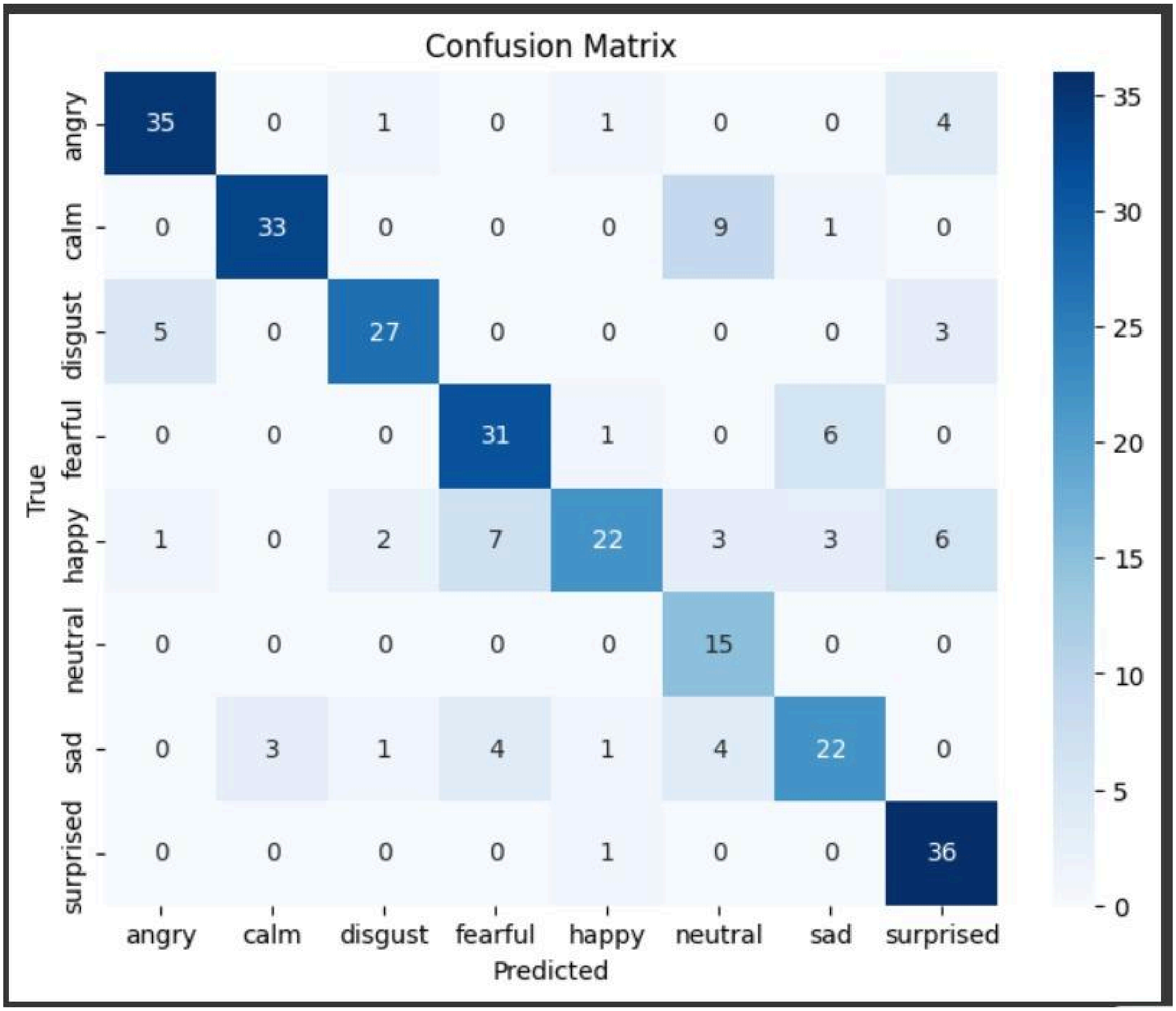
Evaluation Metrics & Results

Metrics Used

- Accuracy: Measures the percentage of correctly classified samples.
- Precision, Recall, and F1-score: Evaluates per-class performance, balancing false positives/negatives.
- Confusion Matrix: Visualizes model misclassifications.

Observations from Results

- Overall Accuracy: 76.7% using Wav2Vec2.0-based classification.
- Best Classified Emotions: Angry (85% Precision, 85% Recall) and Surprised (73% Precision, 97% Recall).
- Poorly Classified Emotions: Neutral (48% Precision, 100% Recall) and Happy (85% Precision, 50% Recall) indicate high false negatives.
- Misclassification Trends: Some emotions (e.g., happy & neutral) are often confused due to acoustic similarities.



	angry	calm	disgust	fearful	happy	neutral	sad	surprised
angry	35	0	1	0	1	0	0	4
calm	0	33	0	0	0	9	1	0
disgust	5	0	27	0	0	0	0	3
fearful	0	0	0	31	1	0	6	0
happy	1	0	2	7	22	3	3	6
neutral	0	0	0	0	0	15	0	0
sad	0	3	1	4	1	4	22	0
surprised	0	0	0	0	1	0	0	36

Open Problems and Future Directions

Open Challenges

- Data Scarcity: Emotional datasets are limited and imbalanced.
- Ambiguity in Emotions: Overlapping emotions (e.g., happy vs. neutral) cause misclassifications.
- Domain Adaptation: Models trained on one dataset often fail in real-world scenarios.
- Multilingual & Noisy Environments: Current models struggle with diverse accents and background noise.

Opportunities & Future Work

- Self-Supervised Learning (SSL): More models like Wav2Vec2.0 to reduce reliance on labeled data.
- Cross-Language Emotion Models: Develop robust multilingual speech emotion recognition.
- Explainability & Interpretability: Improve model transparency to understand misclassification trends.
- Multimodal Emotion Detection: Combining facial expressions & speech for enhanced accuracy.

Conclusion

- The SVM-based model achieved 76.7% accuracy, showing competitive results.
- Deep learning models like Wav2Vec2.0 and HuBERT outperform traditional methods but require large datasets and high computational resources.
- Future research should focus on self-supervised learning, multimodal approaches, and multilingual emotion recognition to enhance robustness.

Final Recommendation

For practical use cases, Wav2Vec2.0 or HuBERT-based models are preferred due to their robustness, while SVM remains a lightweight alternative for low-resource scenarios.