Project Proposal

Title:

Speaker-Independent Speech Emotion Recognition: Limitations, Insights, and Real-World Impact

Background & Motivation

Recognizing emotions from speech has the potential to transform human-computer interaction, mental health support, and adaptive learning systems. Yet, while recent advances in machine learning have driven impressive results in controlled environments, most models still struggle when facing new voices or diverse real-world conditions. This gap is especially important in applications that demand robustness and fairness across people with different backgrounds, ages, and accents.

The TESS dataset, though widely used, includes just two speakers. While this makes it easy to reach high scores in typical train-test splits, it raises a crucial question: can a model trained in this way truly recognize emotions when it encounters a completely new speaker?
Our project aims to honestly explore this challenge and shed light on the real capabilities and limitations of current emotion recognition models.

Objectives

- **Develop a machine learning pipeline** to classify emotions from audio recordings using the TESS dataset.
- **Evaluate the model's performance** in both random train-test splits and, more importantly, in a speaker-independent scenario training on one speaker and testing on the other.
- **Identify the gap** between standard and realistic evaluation protocols, and discuss its implications for real-world deployment.
- **Lay the groundwork** for future work by highlighting practical limitations and proposing concrete next steps for more generalizable solutions.

Methodology

1. **Data Preparation**
    - ○ Audio files from TESS are processed to extract meaningful features (such as MFCCs).
    - ○ Data is cleaned, labeled, and structured into a consistent tabular format.
2. **Model Development**
    - ○ Multiple machine learning algorithms (Random Forest, XGBoost, SVM) are tested to classify the seven emotion categories.
    - ○ Feature scaling and label encoding are applied to ensure fair model comparisons.
3. **Evaluation**
    - ○ **Random split:** Traditional train-test split, where both speakers appear in each set.
    - ○ **Speaker-independent split:** The true test training on one speaker, testing on the other to simulate real-world use.
4. **Analysis**
    - ○ Report accuracy, precision, recall, and F1-score for both splits.
    - ○ Visualize results using confusion matrices and F1-score bar plots.
    - ○ Discuss discrepancies and real-world significance.

Expected Outcomes

- **A transparent benchmark** for how well common models generalize to new speakers on the TESS dataset.
- **Visual and quantitative evidence** showing the dramatic drop in performance when models are exposed to an unfamiliar speaker.
- **Actionable insights** into why robust, speaker-independent evaluation is essential for emotion recognition research and applications.

Potential Impact

This project will help bridge the gap between laboratory results and real-world impact in emotion AI. By openly discussing limitations and avoiding overhyped claims, we hope to encourage the development of fairer, more inclusive and reliable speech emotion recognition systems. Our findings can guide other researchers, developers, and industry teams to adopt stronger evaluation protocols and to prioritize diversity and robustness from the very start.

Future Directions

- Expanding to larger, multi-speaker datasets such as CREMA-D and RAVDESS.
- Exploring deep learning models and data augmentation strategies to improve generalization.
- Testing models under noisy, natural conditions and across different languages and demographics.
- Building open benchmarks and tools for honest, reproducible emotion AI research.

Prepared by:
Rashin Gholijani Farahani
For the Dubai Prototypes for Humanity Competition