CS4320 Final Presentation: Speaker Identification using Logistic Regression and Gradient Descent

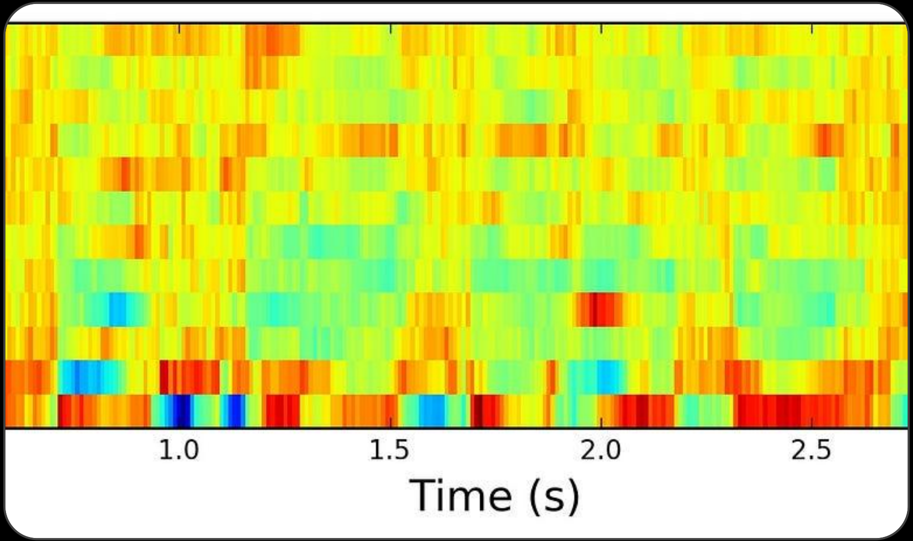
By: Evan Kim

Problem Description

- Standard speech-to-text systems transcribe audio without distinguishing between speakers.
- For multi-speaker conversations, this produces ambiguous transcripts that lack important context about who said what.
- Speaker identification solves this by automatically labeling each segment with the corresponding speaker.
- Use case: Meeting transcripts, interviews, and podcasts become searchable and analyzable by individual speaker contributions.

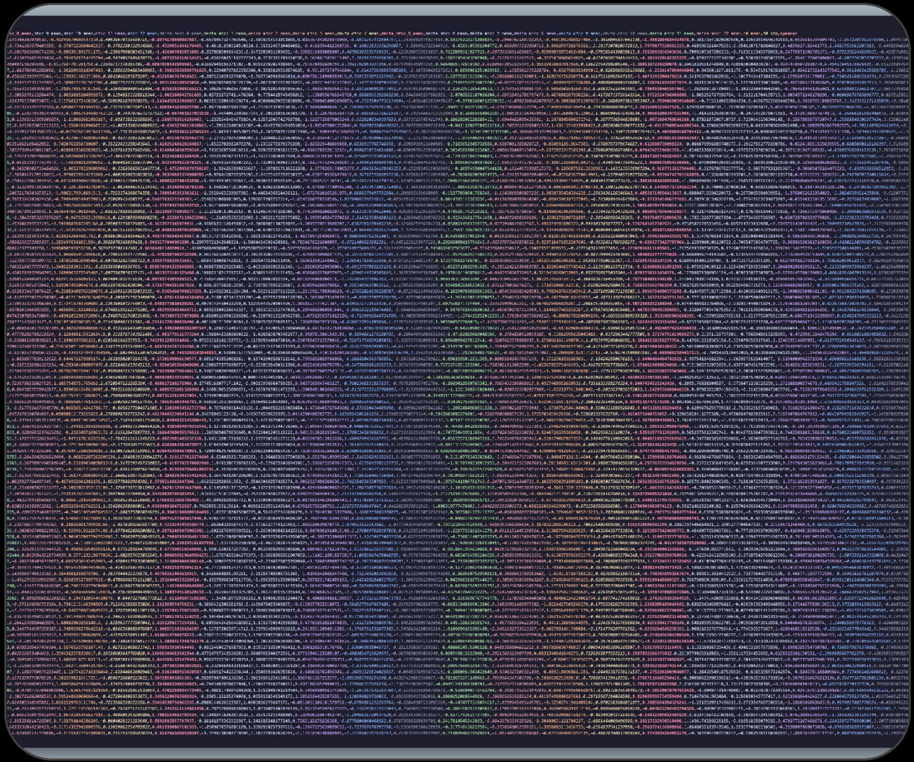
Data

- Data Collection Protocol:
  - Training data: The model is trained using short recordings of participants reading The Rainbow Passage.
- Preprocessing pipeline:
  - Remove silent segments using voice activity detection (VAD) to isolate active speech
  - Segment continuous speech into 5-second windows.
- Feature Extraction:
  - Mel-Frequency Cepstral Coefficients (MFCCs):
    - Mean MFCC values across frequency bands (captures spectral envelope and vocal tract shape.
    - Mean Delta MFCCs mean of the derivative of MFCC
  - Pitch Features:
    - Mean fundamental frequency (F0)
    - Pitch standard deviation
  - All of the above are then standardized



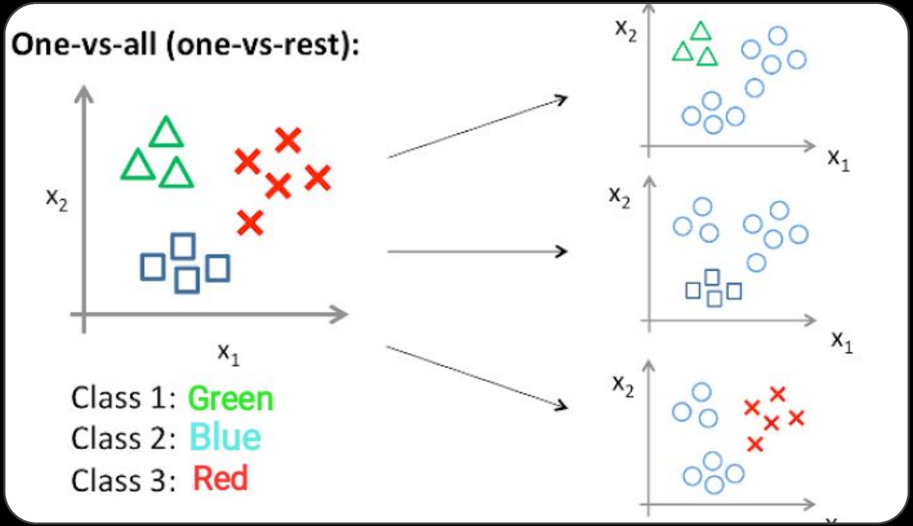
Data Description Summary

- Features: 29 total columns
  - 13 MFCC mean values (mfcc_0_mean through mfcc_12_mean)
  - 13 Delta MFCC mean values (delta_mfcc_0_mean through delta_mfcc_12_mean)
  - 2 pitch features (f0_mean, f0_std)
  - 1 target variable: speaker (categorical: mz, ek, vl, mb)
- Rows: 114 data samples
- Speakers: 4 speakers in the dataset



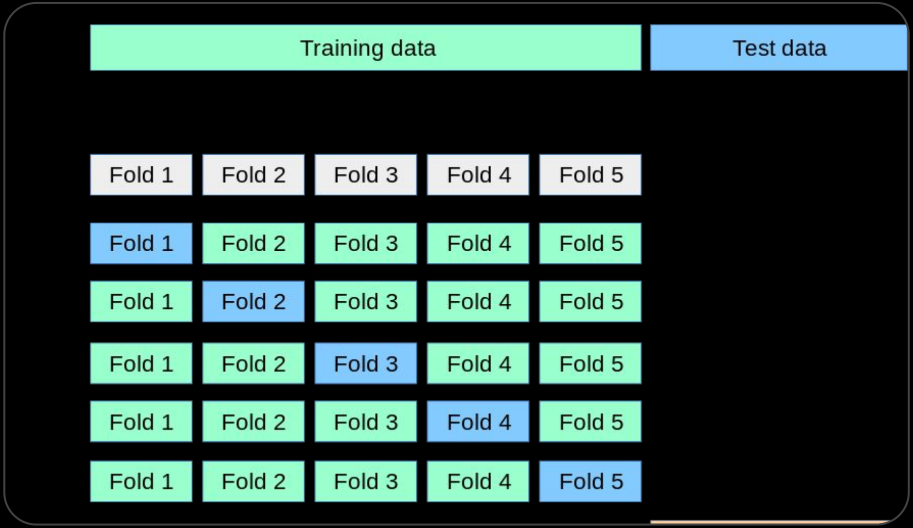
Algorithms

- I used Logistic Regression with Gradient Descent for speaker classification:
  - One-vs-All (OvA) approach for multi-class classification:
    - One binary classifier will be trained for each speaker.
    - Each classifier learns to distinguish one speaker from all other.
- Prediction process:
  - When a new audio sample is received, it is passed through all trained classifiers
  - Each classifier outputs a confidence score (probability between 0 and 1)
  - The speaker corresponding to the classifier with the highest confidence score is selected as the prediction



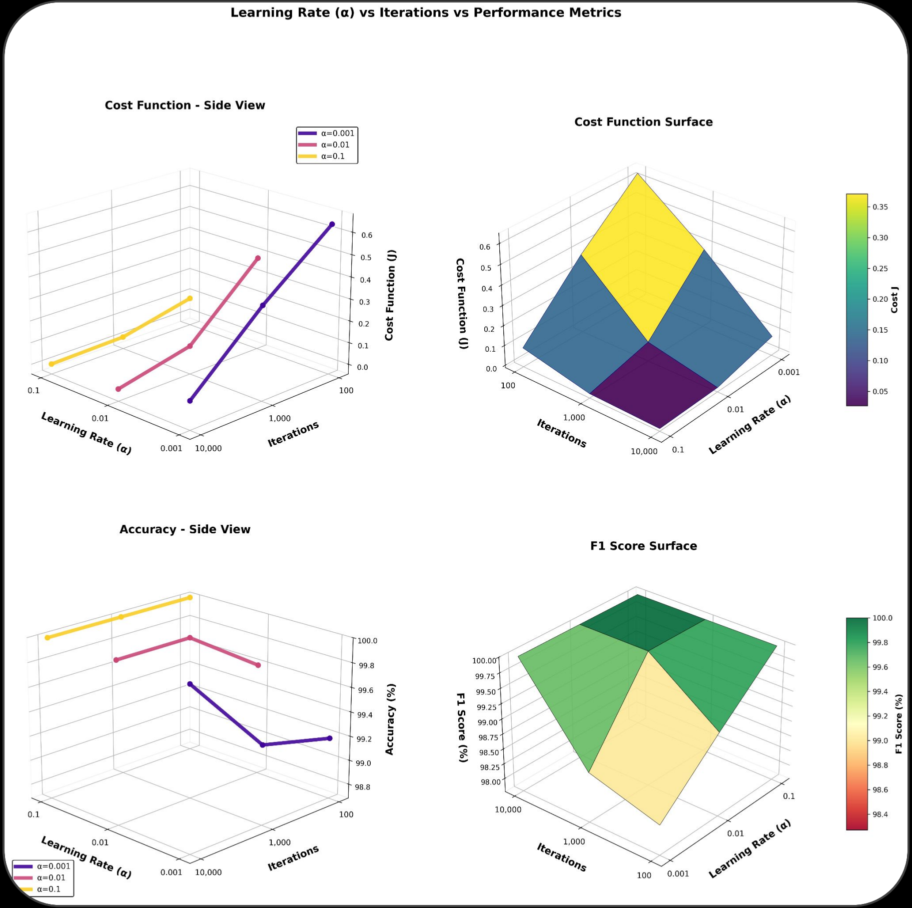
Training, Validation, and Testing Methods

- Cross-Validation Strategy:
  - Stratified K-Fold Cross-Validation with K = 5 folds
  - Stratified sampling maintains balanced speaker representation across all folds
  - Data split: approximately 80% training, 20% validation per fold
- Training Process
  - Independent training of each classifier using gradient descent
  - Binary classification approach: target speaker labeled as positive (1), all other speakers as negative (0)
  - Hyperparameter tuning:
    - Learning rate (α): 0.001, 0.01, 0.1
    - Training iterations: 100, 1,000, 10,000
- Validation and Testing
  - The validation set from each fold serves as the test set for that iteration
- Evaluation Metrics:
  - Accuracy
  - Precision
  - Recall
  - F1-score
  - Confusion matrix components: True Positives (TP), True Negatives (TN), False Positives (FP), False Negatives (FN)



Evaluating

- How I Evaluated the Results
  - Focused on F1-score and accuracy metrics across all tested hyperparameter combinations to gauge and optimize model performance/generalization.
  - Compared performance across different learning rates (0.001, 0.01, 0.1) and iteration counts (100, 1,000, 10,000)
- Key Findings
  - Models converged to 100% accuracy with as few as 1,000 iterations, with diminishing returns beyond this point
  - Gender-based patterns: Female voices consistently showed lower cost function (J) values compared to male voices
- What I Learned
  - Voices are unique and given the right parameters are easily distinguishable.
  - Overfitting is a Key Challenge when choosing features
    - Initially used 169 features, which required significant reduction to prevent overfitting
  - Environmental Factors Have Major Impact
    - Microphone variations caused substantial accuracy degradation
    - Background speakers severely impacted classification accuracy
  - Controlled recording conditions are essential for robust speaker recognition
- Anecdotal Observations
  - Pitch alteration: Changing voice pitch did not consistently fool the model
    - MFCC focuses on the vocal tract shape
  - Speaking style: Minimal difference observed between scripted reading and natural speech patterns.



Questions?