Simple Emotions Detection in Speech using FastDTW

Tianyao He, Sara Iqbal Bavan

CS 5800: Algorithms, Summer 2024

Project Overview

Description:

- The goal of this project is to develop a simple voice-based emotion recognition system.
- Using Dynamic Time Warping (DTW) algorithm
- Enhanced with FastDTW and neural networks

Importance:

- Assists those who have difficulty in understanding emotions.
- Benefits individuals who struggle to hear or interpret emotional cues.
- Improves human-computer interaction.

Agenda

- Introduction to the Dataset
- Data Preparation
- Algorithm Implementation
- Neural Network Model
- Experimental Results
- Future Directions
- Q&A Session

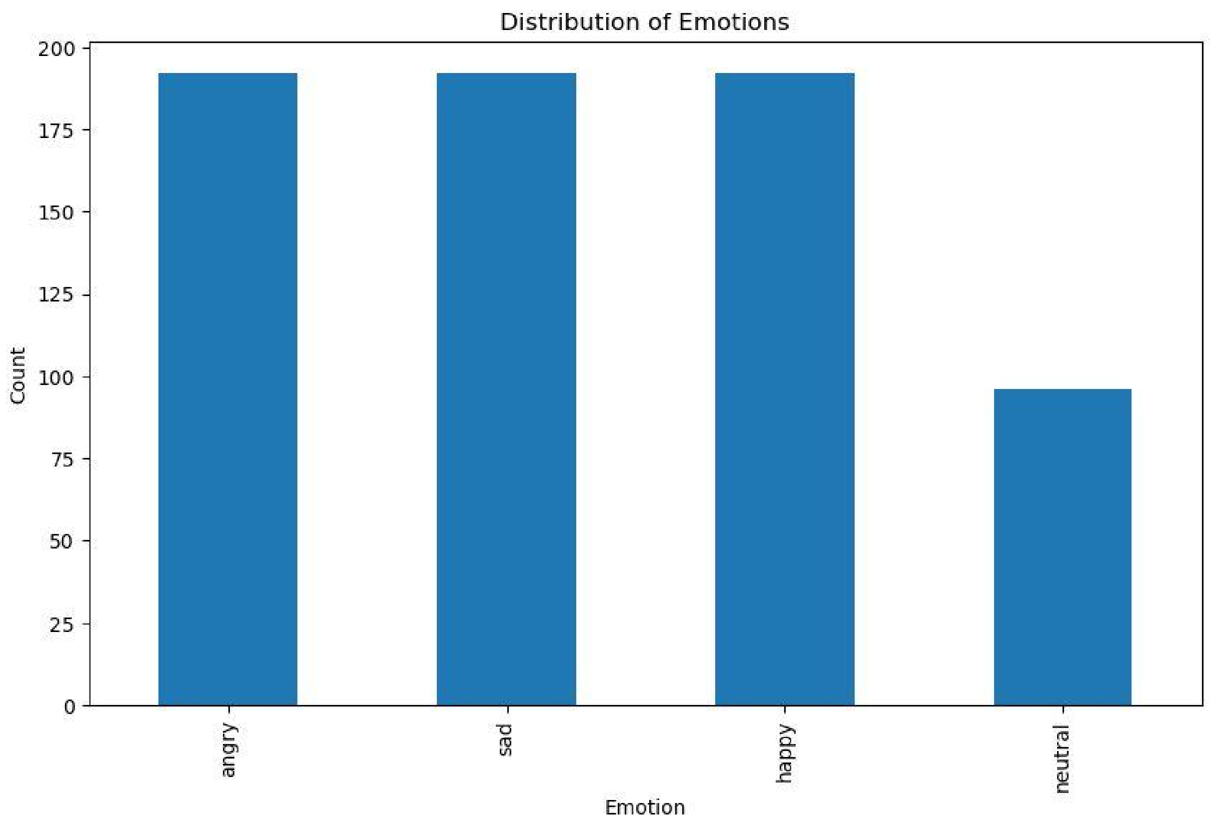
The Dataset

RAVDESS Dataset

- Ryerson Audio-Visual Database of Emotional Speech and Song
- 24 professional actors (12 female, 12 male)
- 8 emotions: neutral, calm, happy, sad, angry, fearful, disgust, surprised

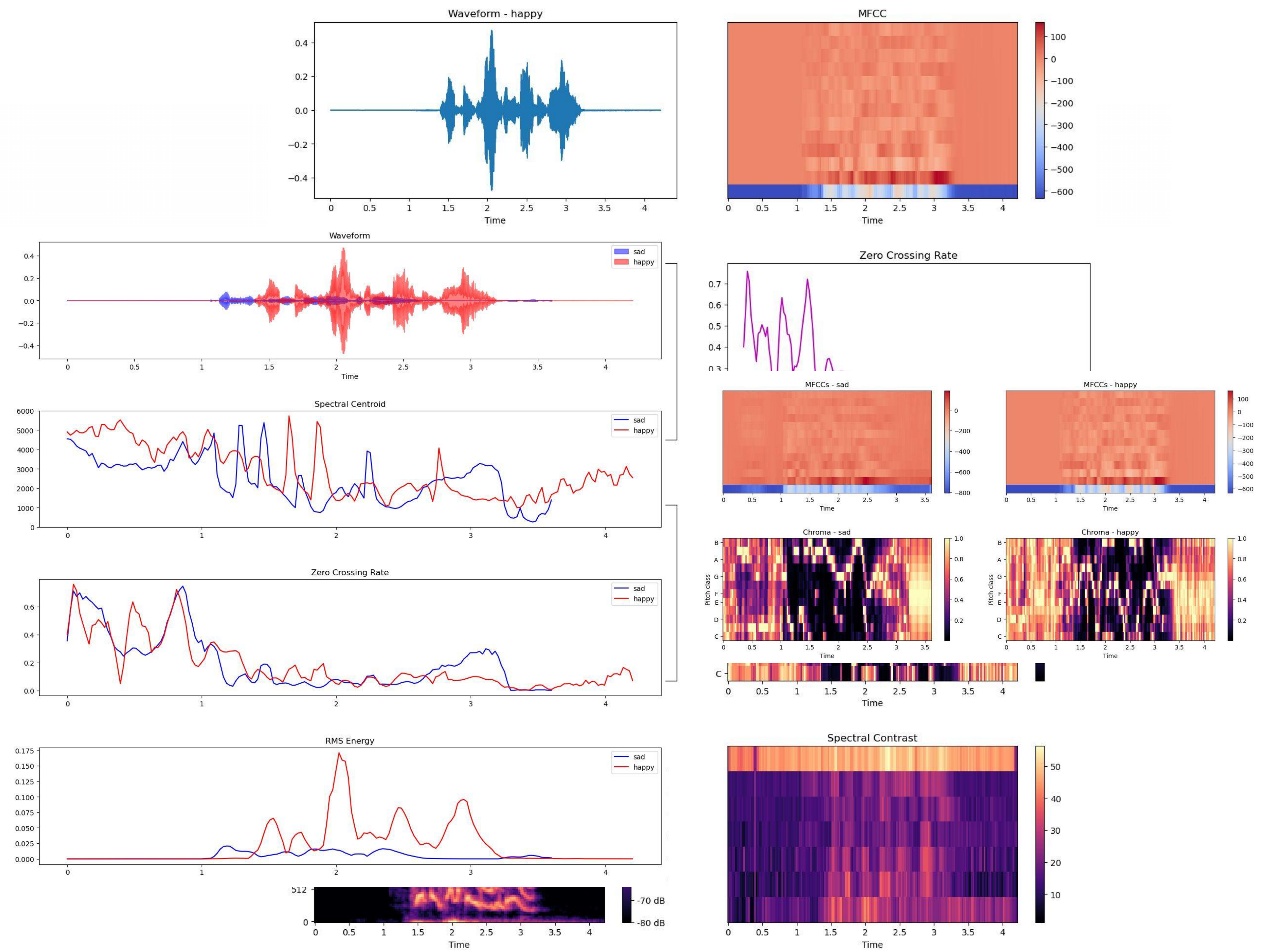
Key Features

- Audio-only files
- 16-bit, 48kHz .wav format
- Speech recordings (no songs)
- Focused on 4 emotions: neutral, happy, sad, angry



Data Preparation

1. Filter relevant audio files
2. Extract audio features:
   - MFCCs
   - Spectral Centroid
   - Zero Crossing Rate
   - Chroma
   - RMS Energy



Let’s refresh some concepts

Recap: Classical DTW

- Measures similarity between two temporal sequences
- Allows for non-linear 'warping' of time axis

Limitations of Classical DTW

- Time complexity: $O(n^2)$
- Space complexity: $O(n^2)$
- Impractical for large datasets or real-time applications

Transition to FastDTW

Introduction to FastDTW

- Approximation of DTW
- Linear time and space complexity: O(n)
- Enables scaling to larger datasets

Advantages of FastDTW

- Significantly faster computation
- Reduced memory usage
- Maintains accuracy comparable to classical DTW
- Suitable for real-time mood detection

Algorithm Implementation

Fast DTW algorithm uses a multilevel approach with three key operations:

1. **Coarsening** - Shrinks time series into smaller time series with fewer data points.
2. **Projection** - Find minimum path at a lower resolution, using that path as an initial guess for higher resolution minimum path.
3. **Refinement:** Refine the alignment using the original data.

Experiment: Compared classical DTW vs FastDTW

Objectives

1. Select best algorithm (DTW vs FastDTW)
2. Determine initial parameters for model training
3. Assess potential for real-time mood detection

DTW Classification Results

Accuracy Comparison

Method	Best Score
Classical DTW	25.31%
FastDTW	21.59%

Speed Comparison

Method	Runtime (50 iterations)
Classical DTW	78m 32.5s
FastDTW	11m 58.4s

- FastDTW is significantly faster (≈6.5x)
- Slight accuracy trade-off (3.72% lower)
- FastDTW enables real-time processing
- Both methods under-perform random guessing (25%)

Neural Network Model

Input Features

- MFCCs (Mel-frequency cepstral coefficients)
- Statistical features:
  - Spectral centroid
  - Zero crossing rate
  - Chroma
  - RMS energy

Output

- 4 emotion classes:
  - Neutral
  - Happy
  - Sad
  - Angry

```
Input MFCCs (batch_size, sequence_length, mfcc_dim)
    |
    V
+------------+
|    LSTM    |  -> Final hidden state (batch_size,
hidden_size)
+------------+
    |
    V
Concatenate with static features (batch_size, stat_dim)
    |
    V
Combined feature vector (batch_size, hidden_size +
stat_dim)
    |
    V
+------------+
|    fc1     | -> (batch_size, 64)
+------------+
    |
    V
ReLU Activation
    |
    V
+------------+
|    fc2     | -> (batch_size, num_classes)
+------------+
    |
    V
Output class scores (batch_size, num_classes)
```

Deep Learning Results Comparison

Accuracy Comparison

Method	Validation Accuracy
Classical DTW + NN	14.29% (1 epoch)
FastDTW + NN	53.11% (50 epochs)

Speed Comparison

Method	Runtime (50 iterations)
Classical DTW + NN	7m 20.6s (1 epoch)
FastDTW + NN	11m 58.4s (50 epochs)

- Significant accuracy improvement with FastDTW + NN
- FastDTW enables more epochs in less time
- Outperforms random guessing (25%)
- Complex model captures nuanced patterns

Future Directions

Expand Emotion Range

- Include more emotions (e.g., fear, disgust, surprise)
- Explore multi-label classification for mixed emotions

Enhance Model Architecture

- Experiment with different neural network structures
- Incorporate attention mechanisms
- Explore transfer learning from pre-trained audio models

Improve Data Processing

- Investigate advanced audio preprocessing techniques
- Explore real-time feature extraction for live applications

Real-World Applications

- Develop user-friendly interfaces for various platforms
- Explore integration with virtual assistants and call centers
- Investigate applications in mental health monitoring

Reference

- Salvador, S., & Chan, P. (2007). Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5), 561-580. https://doi.org/10.3233/IDA-2007-11508

Q & A

Thank you so much for listening!