Singing Data Labeling Tool Milestone 2

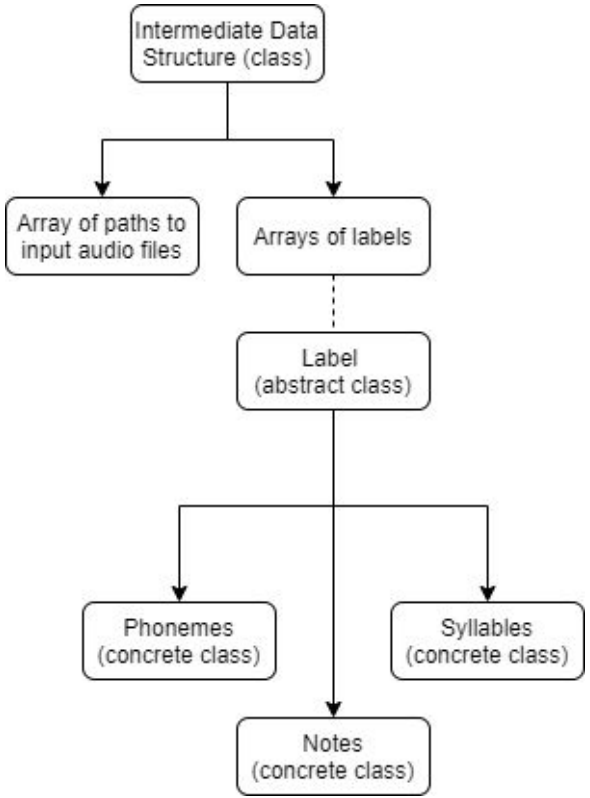
Avinash Persaud, Nandith Narayan, Carlos Cepeda

Renamed Program

- Singing Data Labeling Tool -> Singing Phoneme Egonomic LabeLer (SPELL)

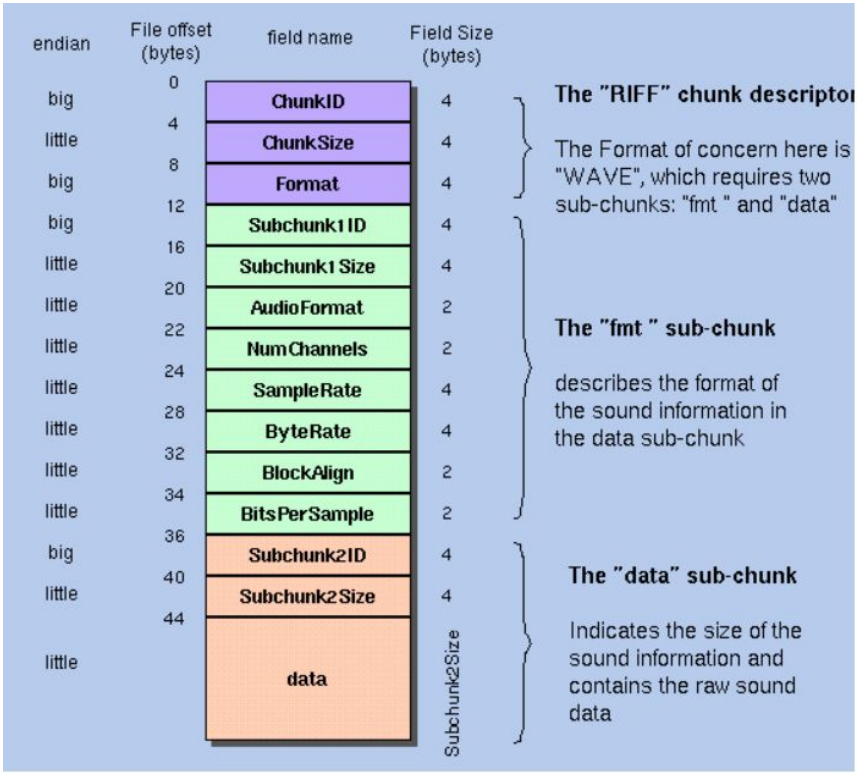
Intermediate Data Structure

- Implemented the Intermediate Data structure
- Loading of audio files into the data structure
- Checks on audio file validity
- Added Antlr v4 runtime as a sub library of the parser
- Generated parser from grammar



Intermediate Data Structure - Loading Audio Data

- WAV files
- 44 Byte header
- Samples normalized to [-1,+1]



Build System

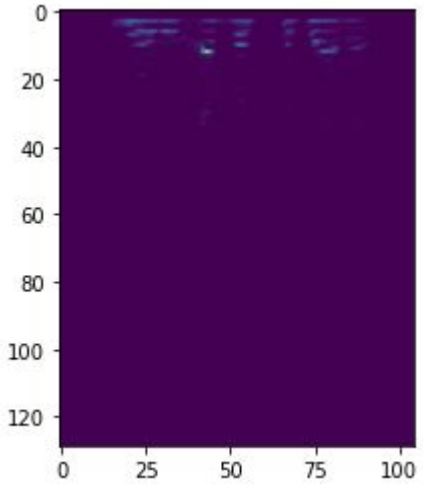
- Cmake
- Nested projects
- Auto install translations

Phoneme Detection

- Attempted to create singing specific model for CMU Sphinx
  - Scripts failed to operate correctly
- SHIRO is missing documentation on it’s inputs
- Further research indicates that phoneme detection on singing is not very accurate in general
- Changing goals boundary detection and language tools
  - Onset detection of phonemes
  - Utility conversions between lyrics and phonemes
  - Conversions between phonemes and syllables

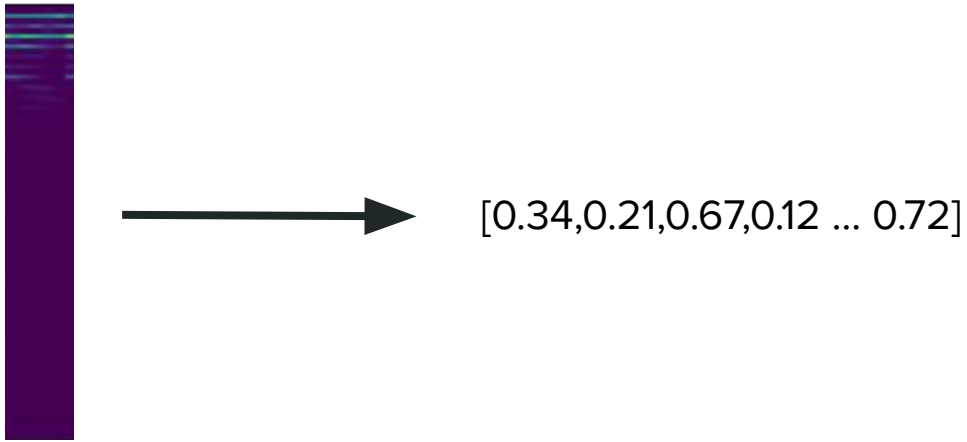
Phoneme Classification

- Process of identifying which phoneme is present in a given audio sample.
- Converted raw audio data into a 2D image representation.



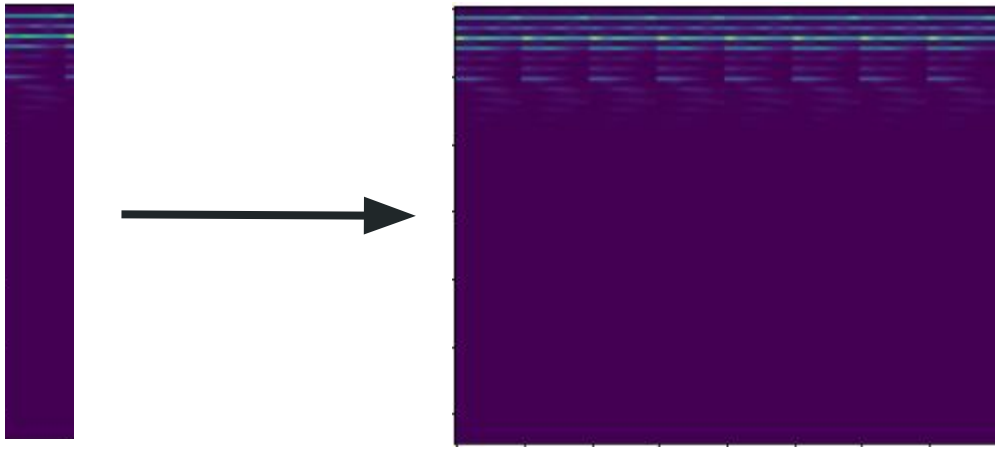
Phoneme Classification - Fully Connected Network

- Tried to use a fully connected neural network.



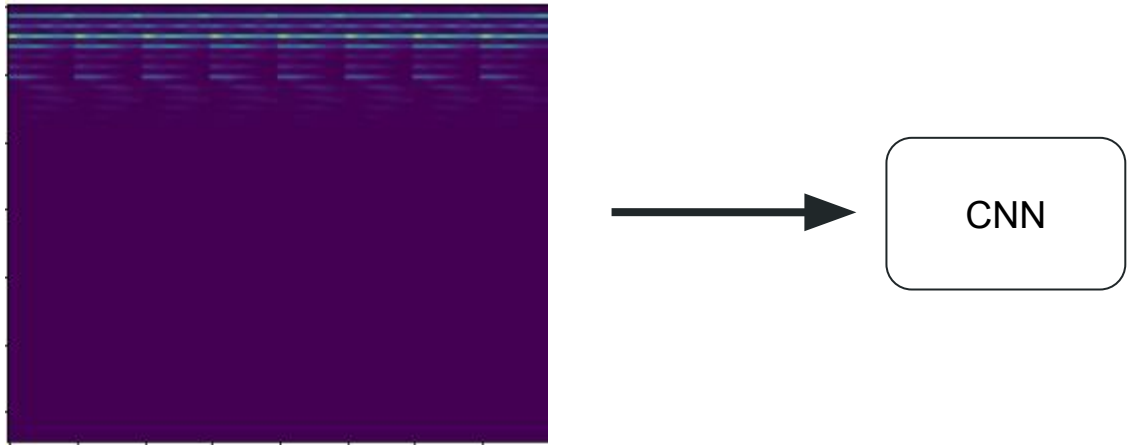
Phoneme Classification - Data augmentation

- Repeated spectrogram horizontally to create a larger image.



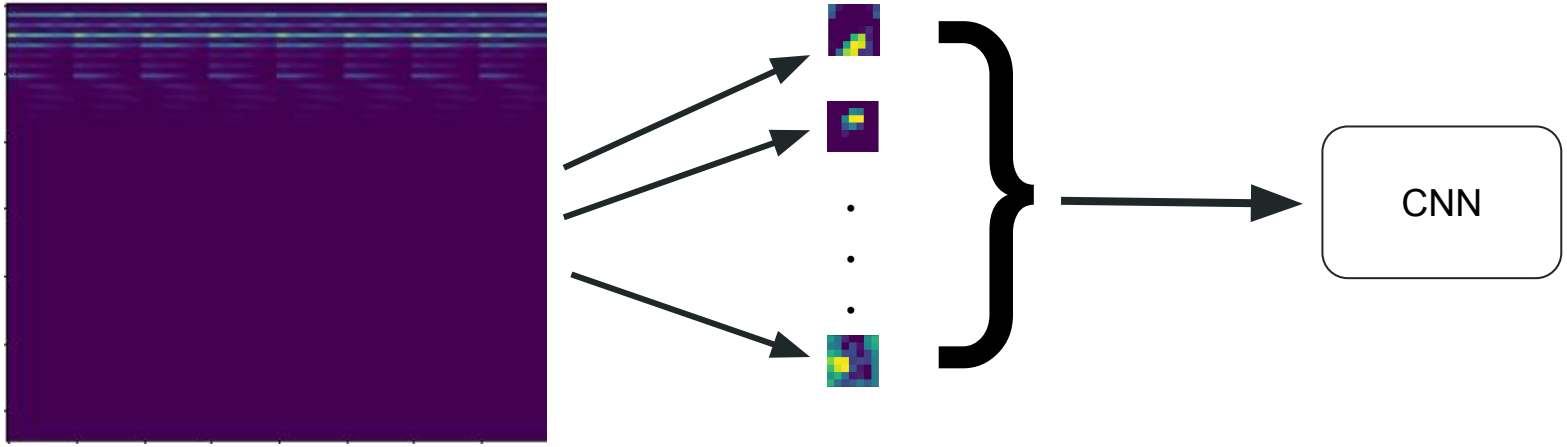
Phoneme Classification - CNN

- Tried to use a Convolutional Neural Network to classify phonemes.



Phoneme Classification - Transfer Learning

- Tried to use VGG16 to extract Features

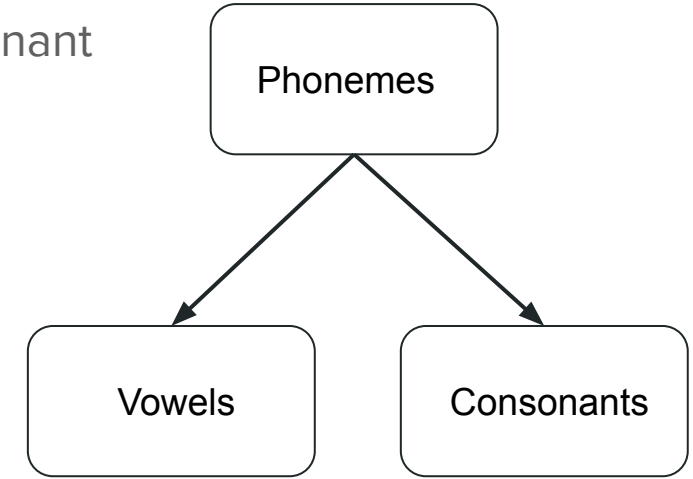


Phoneme Classification - Results

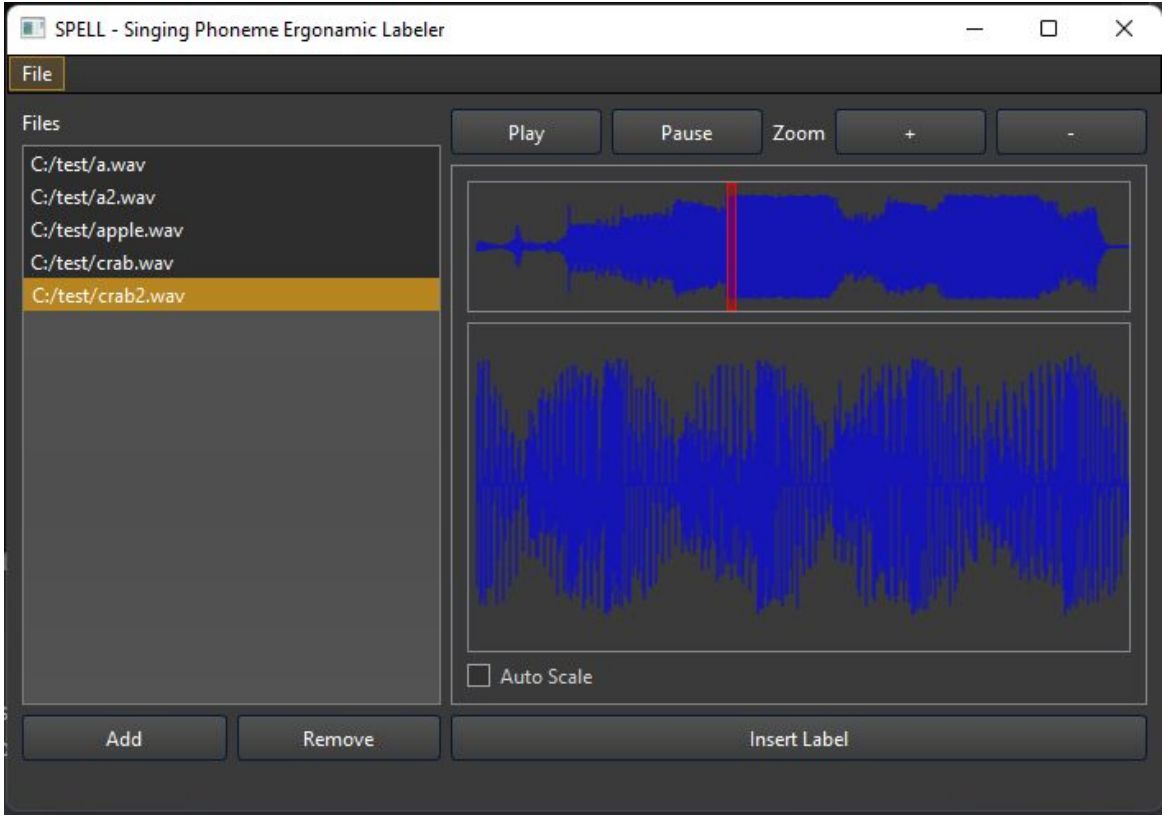
- Fully connected Network:
  - After 100 Epochs: 20% Accuracy
  - After 48,000 Epochs: 55% Accuracy
- CNN:
  - After 10 Epochs: 43% Accuracy
  - After 100 Epochs: 58% Accuracy
  - Still requires more training and tuning
- Transfer learning with VGG16 + CNN:
  - After 10 Epochs: 12% Accuracy
  - After 100 Epochs: Overfit with 70% training accuracy and only 38% test accuracy.

Phoneme Classification - Our New Approach

- Subdivide the problem
- Classify if Phoneme is Vowel or Consonant
- Classify each sub category



GUI


SPELL - Singing Phoneme Ergonamic Labeler
File
Files
C:/test/a.wav
C:/test/a2.wav
C:/test/apple.wav
C:/test/crab.wav
C:/test/crab2.wav
Play
Pause
Zoom
+
-
Auto Scale
Add
Remove
Insert Label

Demo

https://youtu.be/nZLh7dSm5uE

Milestone 3 Tasks

- Continue searching and testing models for boundary detection
- Refine Phoneme Classification model
- Allow audio file input from the user
- Display Spectrogram and add ability to add markers
- Allow User to manually add phonemes
- Save output of parser into intermediate data structure
- Create output template window and integrate with parser

Questions?