Speech-Recognition

Erdos Institute
Data Science BootCamp
Spring 2023

Team 8 - Jacob Mashburn, Benjamin Warren, Suraj Singh Khurana

Introduction

- Speech recognition is an important part of smart homes, phones, voice assisted technologies.
- A challenging task is to build a model that transcribes speech in a noisy setting.
- **Aim-** a model that recognizes whether speech is present.
- Many datasets are available such as LibriSpeech, CHiME, Urban Sounds and AudioSet.

The Google AudioSet

A dataset of nearly 2,000,000 human-labeled 10-second YouTube video soundtracks.

- Each sound event human-labeled according to a hierarchical ontology.

  (Gemmeke, J. et. al., AudioSet: An ontology and human-labelled dataset for audio events, ICASSP 2017)

- Around 527 labels.
- For example:- Music, Vehicle, Bird, Drum, Traffic Noise, Finger snapping etc
- Has been used to train CNNs to identify the labels

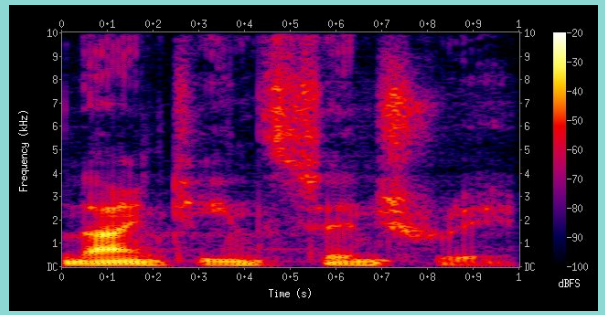
Temporally Strong Labels

- In 2021, Google released a set of temporally strong labels for 5% of the segments in its dataset (Hershey S., et. al., The Benefit Of Temporally-Strong Labels In Audio Event Classification, 2021)
- Labels that cover each second of audio, not just the whole 10 second segment
- We aim to use the temporally strong labels and a simplified classification task to get the best results we can on the data

Audio Features and .tfrecords

- Audioset stored in .tfrecord format - we converted to simple csv format, and filtered to find data with temporally strong labels
- Audio features:

Spectrogram

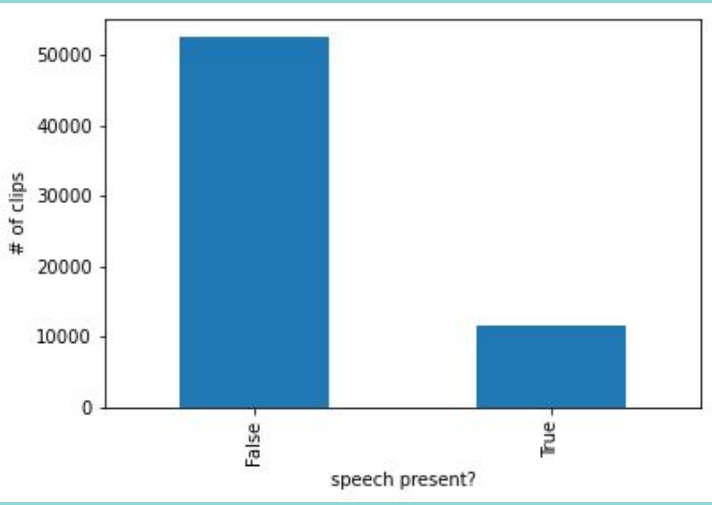


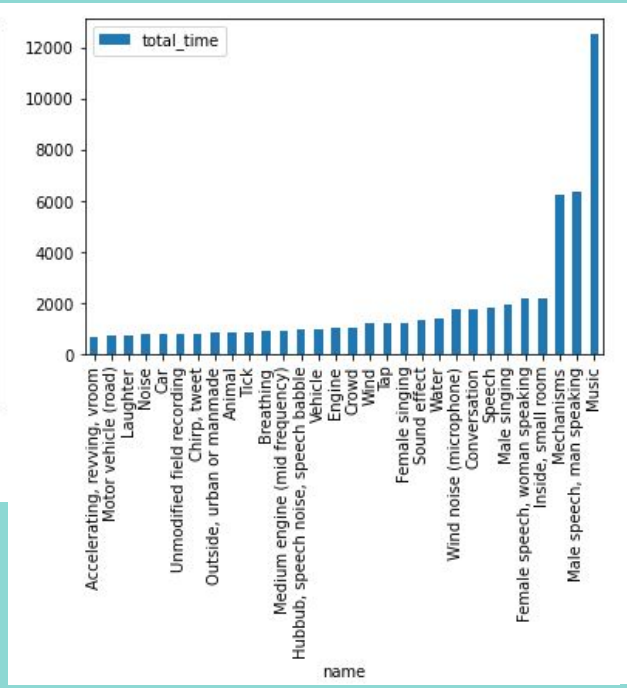
Spectrogram by Aquegg, wikipedia: https://commons.wikimedia.org/wiki/File:Spectrogram-19thC.png?uselang=en#Licensing

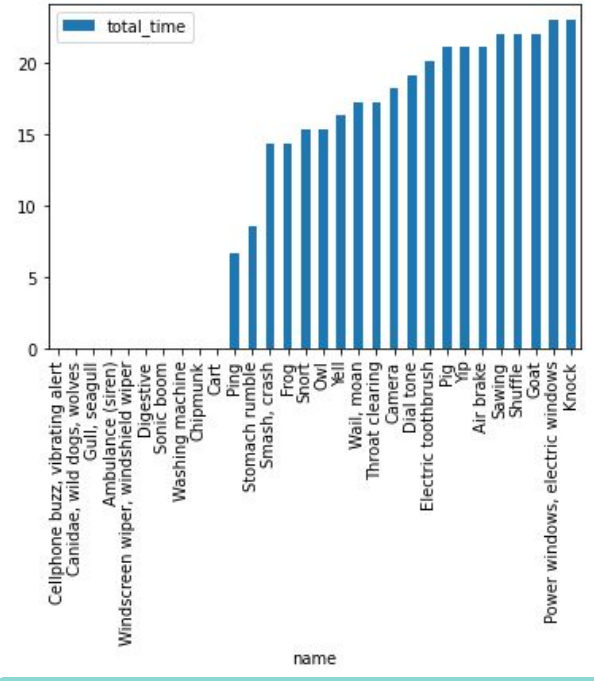
128-dimensional vector

[1, 72, 0, 255, 87, ...]

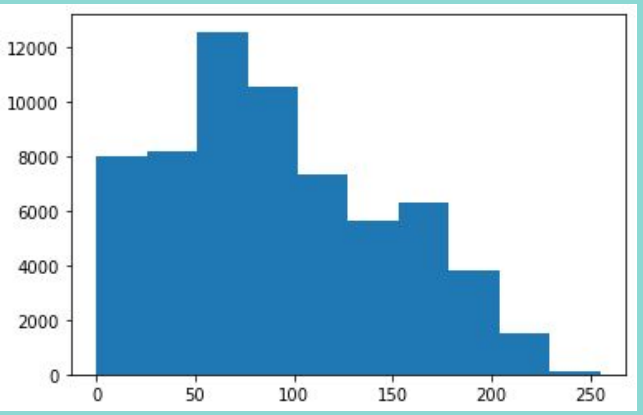
Visualizations



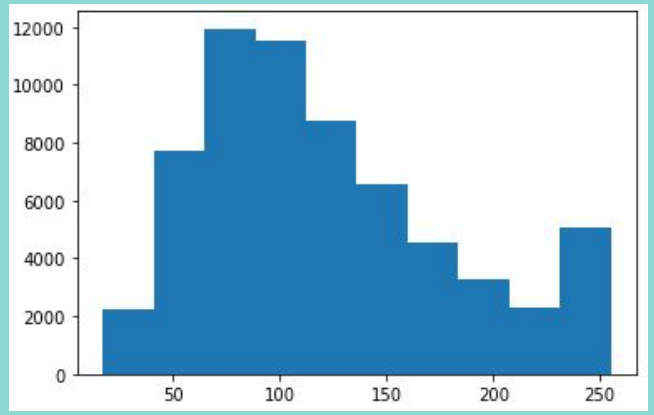




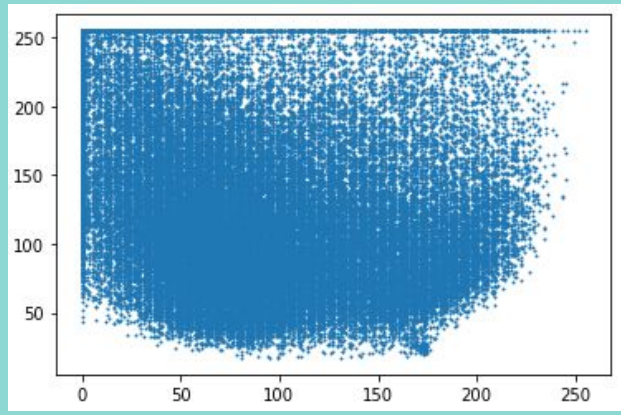
Visualizations (cont.)



First dimension



Second dimension



First (x-axis) vs. Second (y-axis)

Model: Simple Neural Network

Layer (type)	Output Shape	Param #
normalization (Normalization)	(None, 128)	0
dense_1 (Dense)	(None, 200)	25800
dense_2 (Dense)	(None, 160)	32160
dense_3 (Dense)	(None, 140)	22540
dense_4 (Dense)	(None, 100)	14100
dense_5 (Dense)	(None, 60)	6060
dense_6 (Dense)	(None, 1)	61

Total params: 100,721
Trainable params: 100,721
Non-trainable params: 0

Performance Metrics (Avg over 30 trials)

- Accuracy = # correct predictions / # 960ms excerpts

  Accuracy was 82.2%

- Recall = # true positives / (# true positives + # false negatives)

  Recall was 66.0%

- Precision = (# true pos.) / (# true pos. + # false pos.)

  Precision was 64.4%

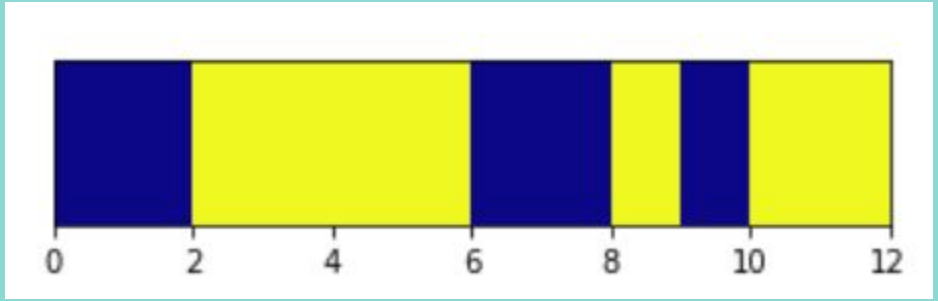
Conclusions and Future Directions

Confusion Matrix	Predicted Negative	Predicted Positive
Negative	**9361**	**864**
Positive	**746**	**1818**

- Good accuracy (over 80%), but precision and recall a bit lower
- Still promising result for our model
- More work to be done on recognizing human non-speech sounds

Demonstration

The model's prediction:



Yellow = Speech Found
Blue = No Speech Found