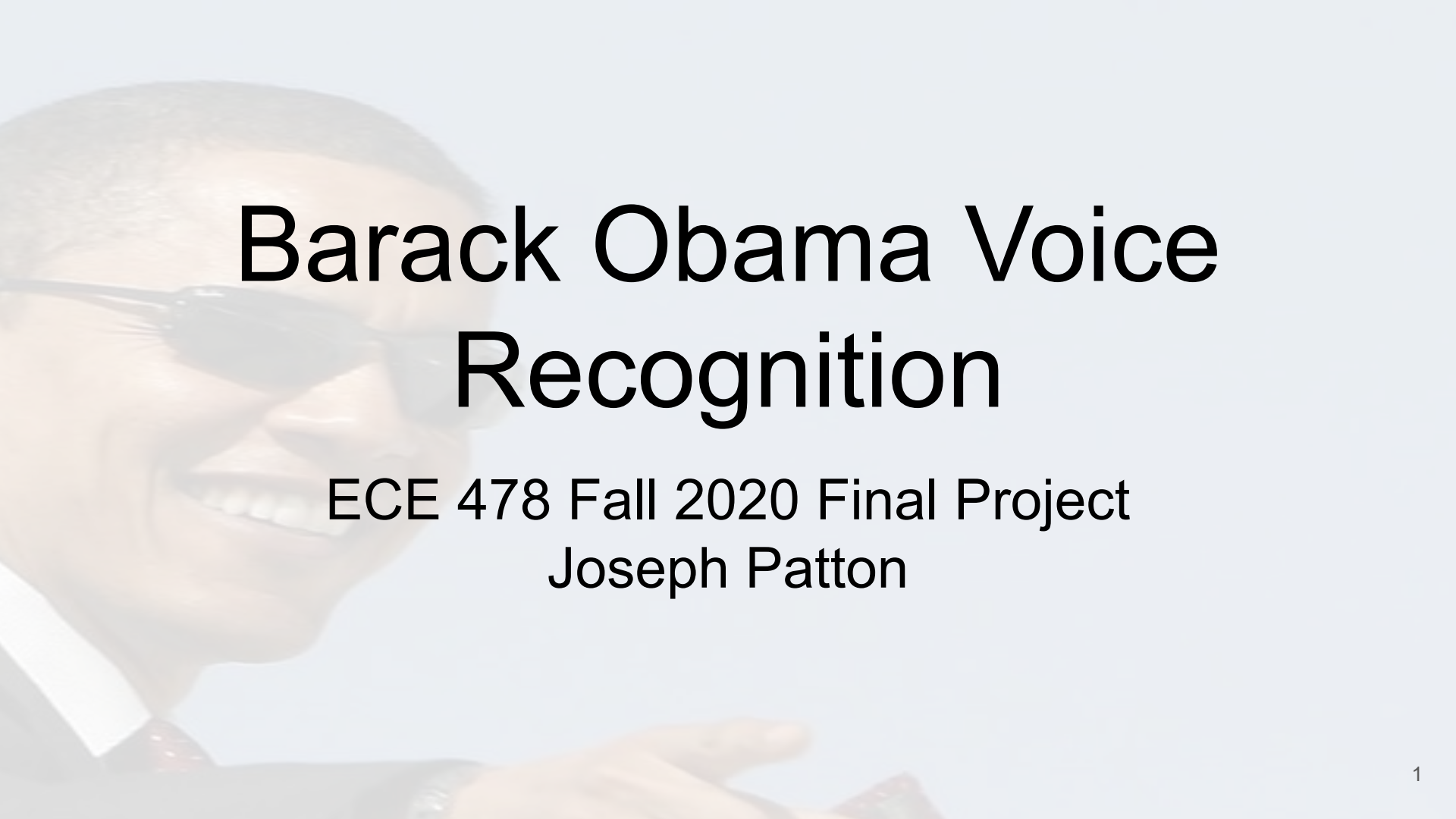


Barack Obama Voice Recognition

ECE 478 Fall 2020 Final Project

Joseph Patton

Presentation Overview

- **Project Objective**
- **Neural Network Training Methods**
  - **Finding and Conditioning Data**
  - **Extracting Data Features**
  - **Training the Network**
- **Testing Methods**
- **Results**
- **Code Demonstration**

Objective: Train a radial basis function neural network to recognize the voice of Barack Obama.

Training Methods

Finding Data

Two Types of Required Data

- **Voices of people who are Barack Obama**
- **Voices of people who are not Barack Obama**

Ideally, the voice recordings are isolated and contain little to no noise.

Solution: Take Audio Files from Cable-Satellite Public Affairs Network (C-SPAN)

- **Speaker's Voice Isolated**
- **Consistent Room Noise**

Conditioning Data

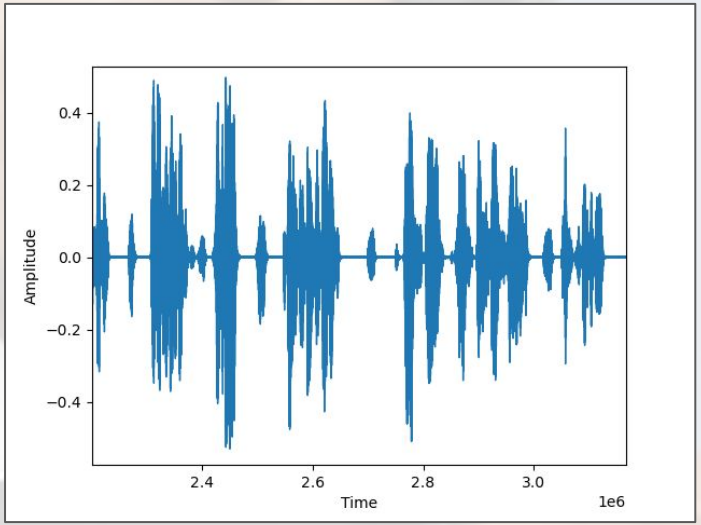
Data should not contain silences:

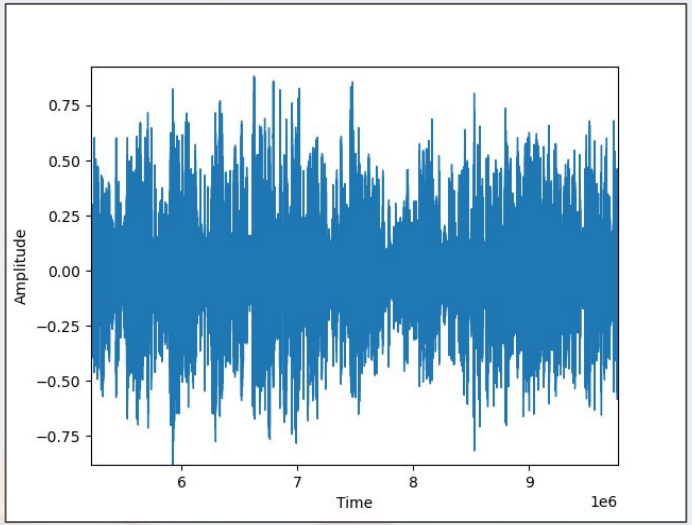
- **Silence sounds the same for all speakers**

Software to remove silences:

- **SoX, "the Swiss Army knife of sound processing programs."**

```
sox filename.wav filename_ns.wav silence 1 0.1 1% -1 0.1 1%
```





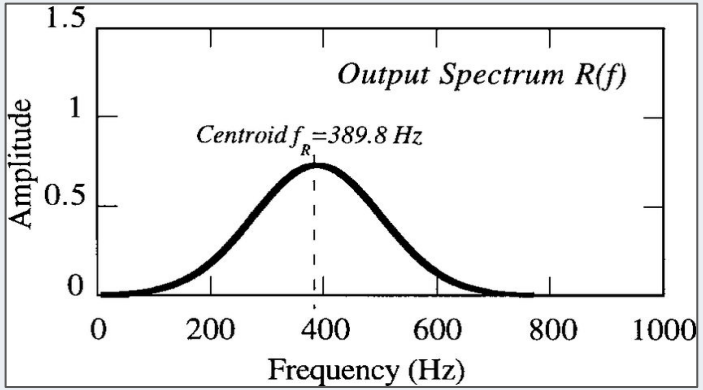
Left Image: Voice Recording with Silences included

Right Image: Voice Recording with Silences Removed

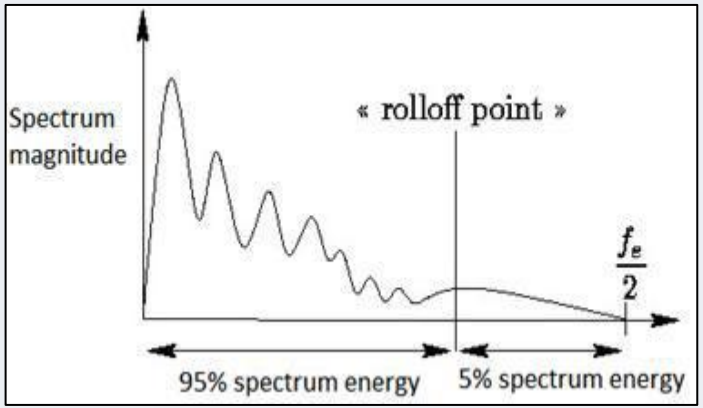
Extracting Features

Features Extracted from Conditioned Voice Data

- Spectral Centroid
- Spectral Roll-off
- Zero Crossing Rate
- Spectral Flux
- Mel-Frequency Cepstral Coefficients (first 13 coefs)



Spectral Centroid is the frequency that corresponds to the center of mass of the spectrum



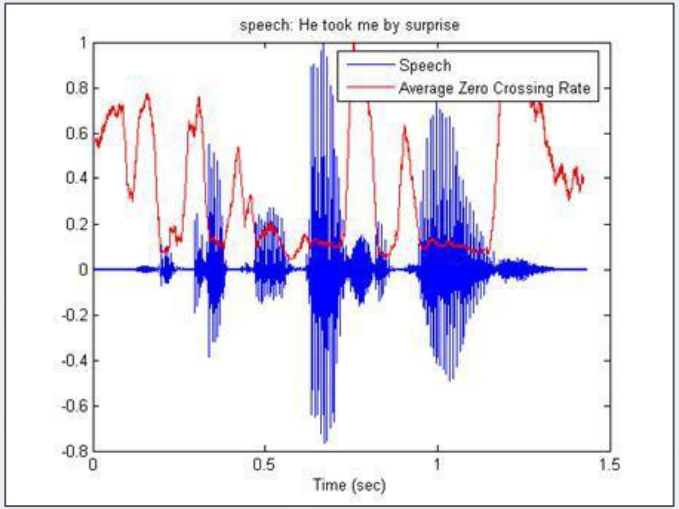
Spectral Roll-Off is the frequency at which c% of the energy of the signal is contained at or below that frequency

Images: https://sudonull.com/post/101170-Analysis-of-audio-analytics-algorithms

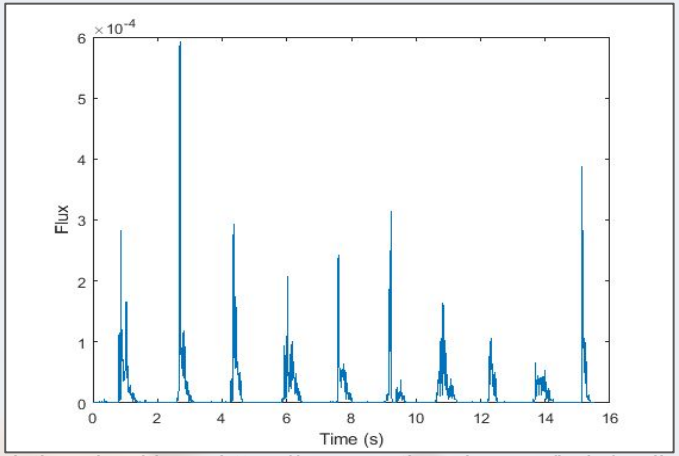
Extracting Features

Features Extracted from Conditioned Voice Data

- Spectral Centroid
- Spectral Roll-off
- Zero Crossing Rate
- Spectral Flux
- Mel-Frequency Cepstral Coefficients (first 13 coefs)



Zero Crossing Rate is the rate that the normalized audio crosses the x-axis



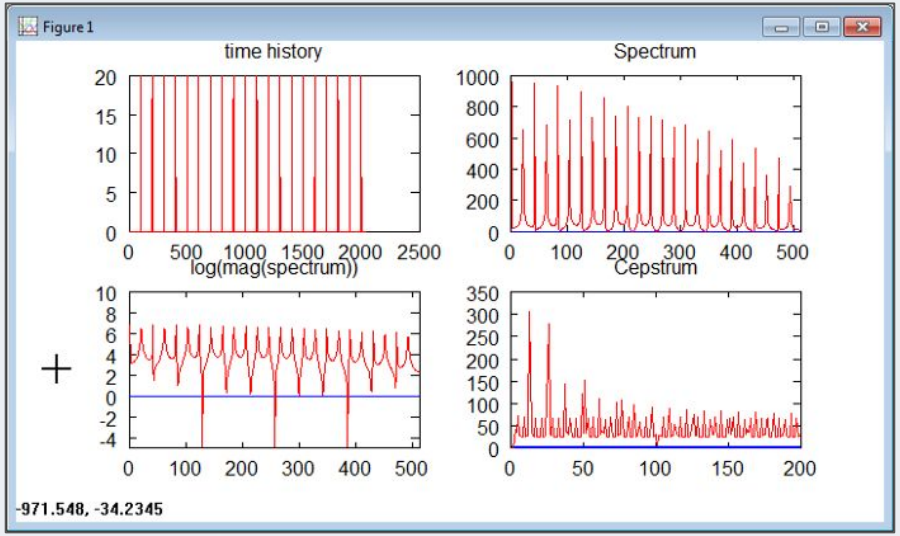
Spectral Flux is a measure of how quickly the power spectrum of a signal is changing over time

Images: https://sudonull.com/post/101170-Analysis-of-audio-analytics-algorithms, https://www.mathworks.com/help/audio/ref/spectralflux.html

Extracting Features

Features Extracted from Conditioned Voice Data

- Spectral Centroid
- Spectral Roll-off
- Zero Crossing Rate
- Spectral Flux
- Mel-Frequency Cepstral Coefficients (first 13 coefs)



Steps to find Mel-Frequency Cepstral Coefficients.

1. Take the Fourier transform of a signal.
2. Map the powers of the spectrum obtained above onto the mel scale.
3. Take the logs of the powers at each of the mel frequencies.
4. Take the discrete cosine transform of the list of mel log powers, as if it were a signal.
5. The MFCCs are the amplitudes of the resulting spectrum.

In this case, librosa.feature.mfcc library was used to find MFCC

$$\mathrm{Mel}(f) = 2595 \log\left(1 + \frac{f}{700}\right)$$

Frequency to Mel Scale Conversion Formula

Images: https://medium.com/prathena/the-dummys-guide-to-mfcc-aceab2450fd

Training Network

Input 17-Dimensional Data

↓

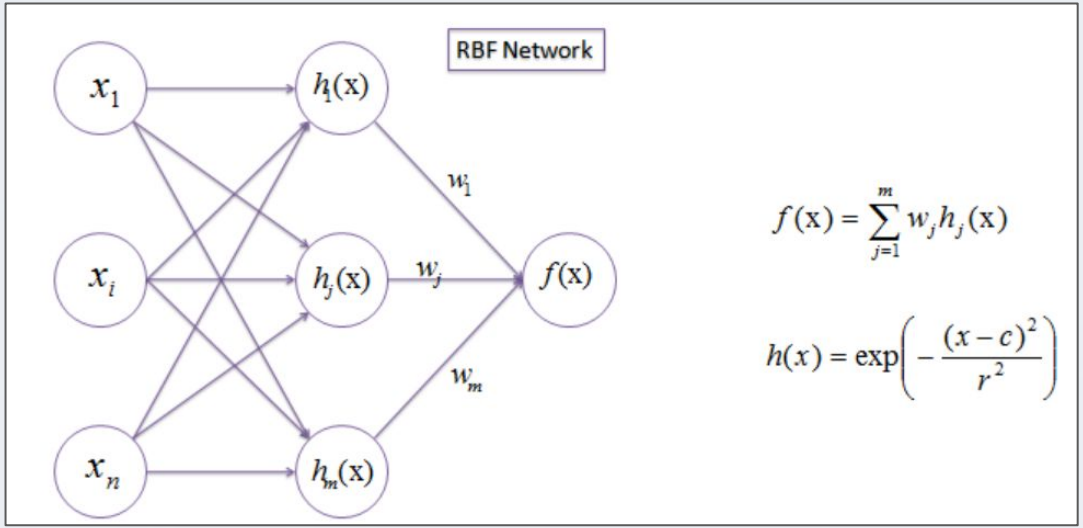
Calculate Activation Function Centers

↓

Calculate Activation Function Widths

↓

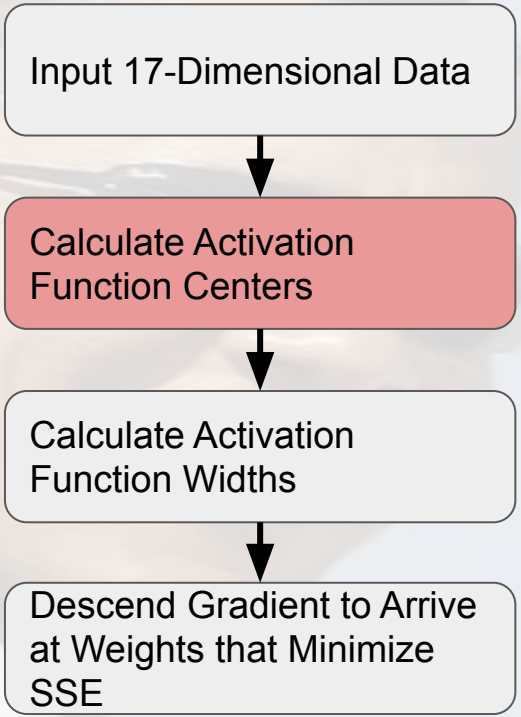
Descend Gradient to Arrive at Weights that Minimize SSE

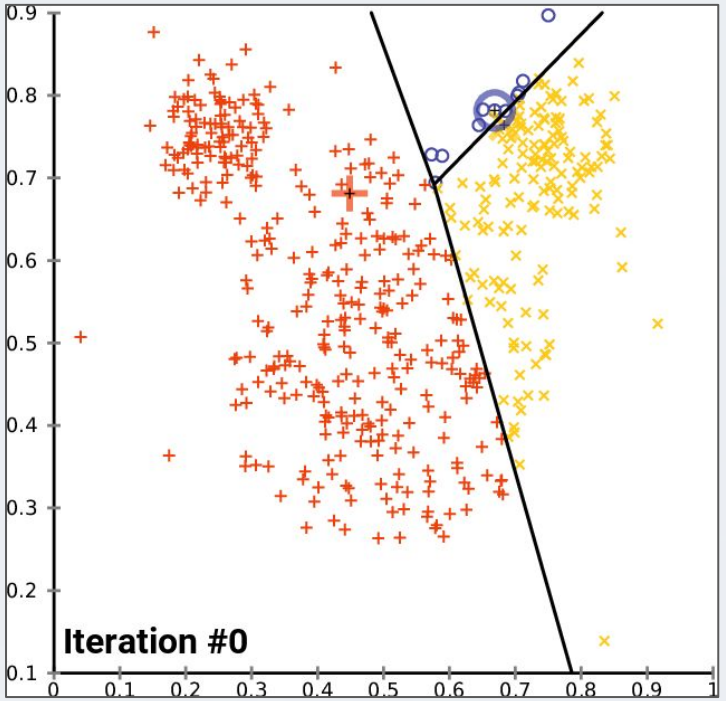


Radial Basis Function Neural Network

Images: https://www.saedsayad.com/artificial_neural_network_rbf.htm

Training Network

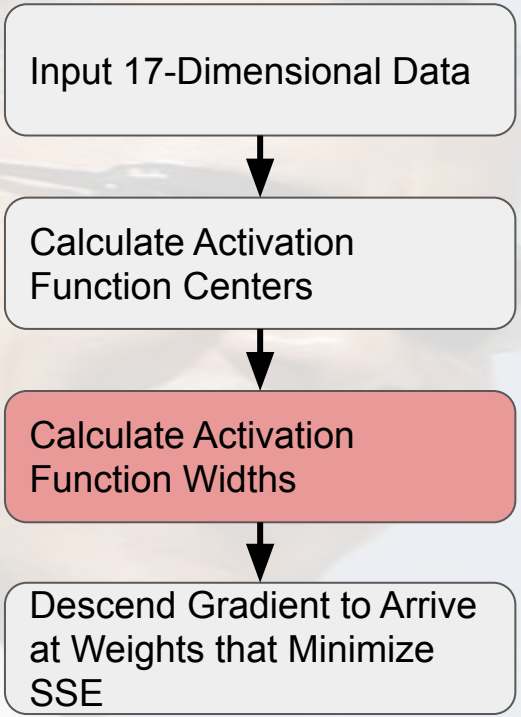




K-means Clustering Algorithm. The sklearn library k-means method was used.

Images: https://en.wikipedia.org/wiki/K-means_clustering

Training Network



$$r_j = \sqrt{\frac{\sum_{i=1}^{k} (c_j - c_i)^2}{k}}$$

K-Nearest Neighbor Equation to find the width of each activation function. K=2 was used.

Training Network

Input 17-Dimensional Data

↓

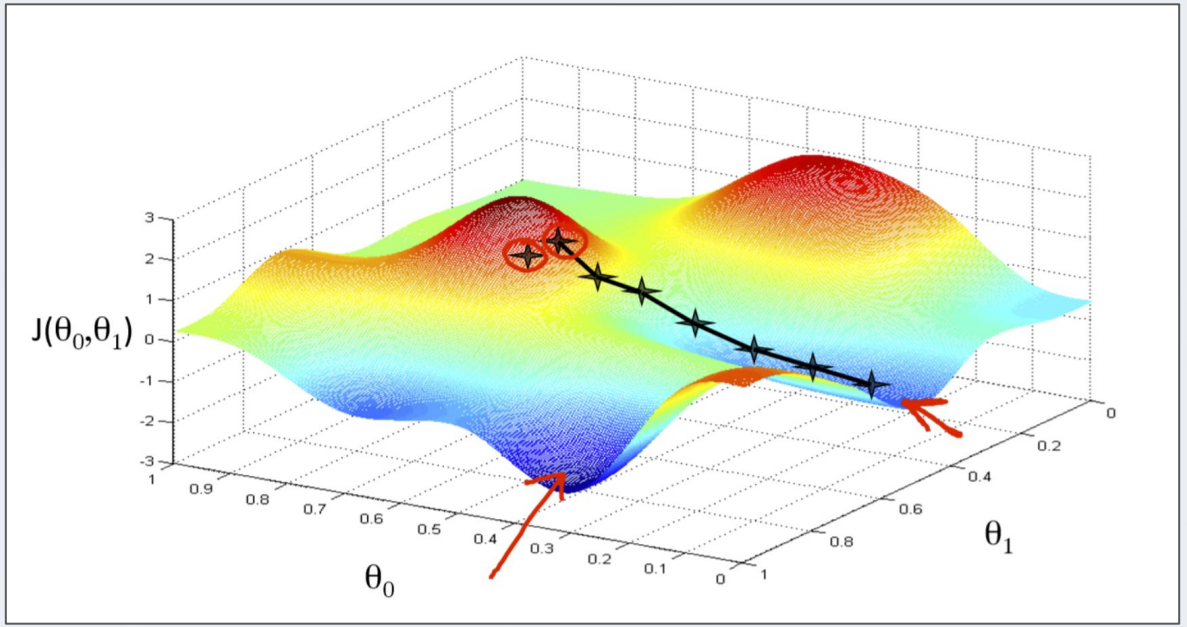
Calculate Activation Function Centers

↓

Calculate Activation Function Widths

↓

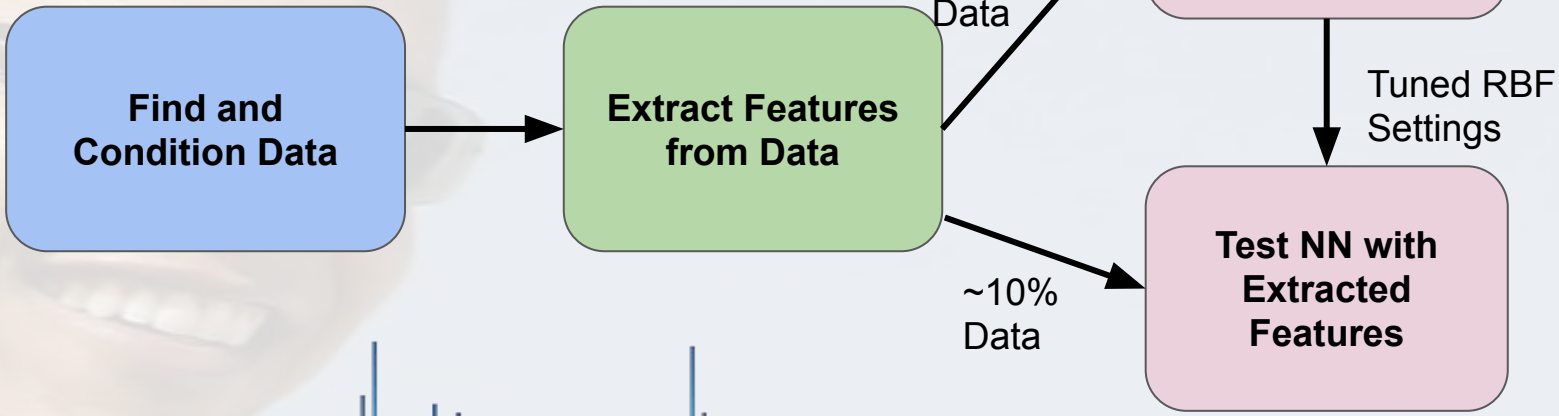
Descend Gradient to Arrive at Weights that Minimize SSE



Calculate sum squared error, calculate gradient, and take a small step down the gradient.

Images: https://towardsdatascience.com/an-intuitive-explanation-of-gradient-descent-83adf68c9c33

Testing Methods



Testing Methods

Input Data:

- Four Sound Clips
  - Two of Barack Obama, 15m18s, 19m44s
  - One of Mitt Romney, 8m30s
  - One of Bernie Sanders, 13m18s
  - Silences removed
- Sample rate 44.1 ksps, FFT size = 65536, window length = 1.48 sec
- 1371 data points overall, 87.5% randomly selected for training, other 12.5% used for testing

Results

Program Output in 100 Runs:

- Iterations Completed: >450000
- Avg Pos ID Success: 92.7618%
- Avg Neg ID Success: 73.6810%
- Avg Overall Success: 82.2924%