Speech Recognition Using Linear Predictive Coding and Support Vector Machines

David McNeil

Rose-Hulman Institute of Technology

mcneilde@rose-hulman.edu

Overview

Objective

Successfully be able to distinguish samples of speech.

- Feature Extraction– Extracting a unique set of features from the audio sample
- Classification– Comparing the extracted features to the features from a known database of samples

Create a Training Database

I created a program to record audio samples in order to build up a training database of audio samples

- 44100Hz sample rate
- 16 bits per sample
- Single channel of audio
- 2 second duration
- Relatively low level of background noise

Linear Predictive Coding (LPC)

- General filter difference equation:

$$y(n) = \sum_{j=1}^{N} a_j \cdot y(n-j) + \sum_{j=0}^{M} b_j \cdot x(n-j)$$

- $y(n)$ is predicted from past outputs and present and past inputs
- Use all pole filter model
  - $M = 0$
  - Prediction based only on past outputs and the current input

Linear Predictive Coding (LPC)

- Estimation of output:

$$\hat{y}(n) = \sum_{j=1}^{N} a_j \cdot y(n-j))$$

- Prediction error: $e(n) = y(n) - \hat{y}(n)$

Linear Predictive Coding (LPC)

- Solve for $y(n)$ and plug in definition for $\hat{y}(n)$:

$$y(n) = \sum_{j=1}^{N} a_j \cdot y(n-j) + e(n)$$

- All pole filter model:

$$y(n) = \sum_{j=1}^{N} a_j \cdot y(n-j) + b_0 \cdot x(n)$$

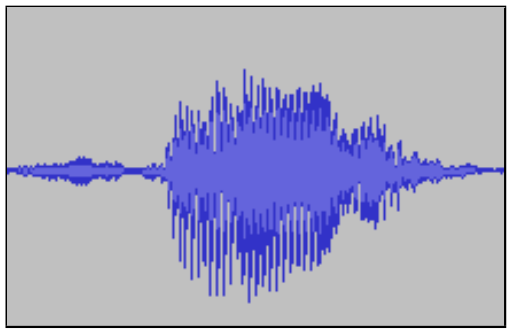
- By comparison $e(n) = b_0 \cdot x(n)$
- The prediction error is a result of the input (filter excitation source)

Linear Predictive Coding (LPC)

- System which produces N time-varying filter coefficients
- The coefficients act as a compressed version of the audio
  - The filter can be excited and an approximation of the signal produced
- The coefficients also uniquely represent the audio
  - Similar signals will have similar coefficients
- These filter coefficients make up our feature space

Example LPC Coefficients

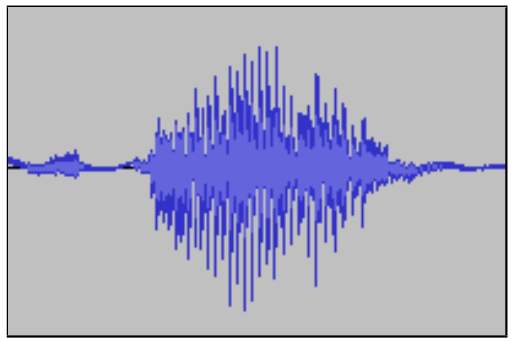
"start"



Coefficients

- -1.6538
- 1.0859
- -0.8826
- 0.3327
- 0.1480

"stop"



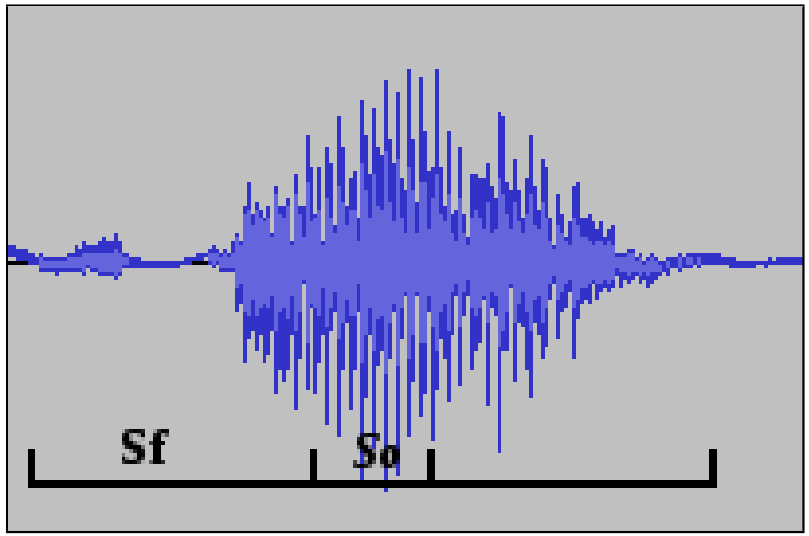
Coefficients

- -1.9805
- 1.7414
- -1.3340
- 0.5479
- 0.0415

LPC Coefficients Extraction Method

Parameters

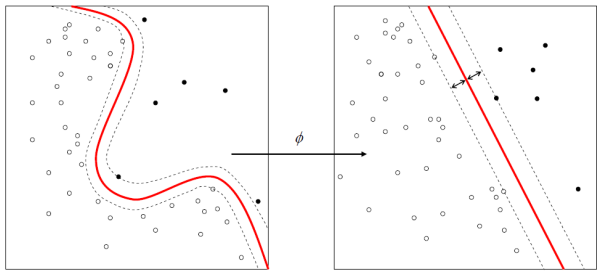
- N - Number of desired coefficients per frame
- S_f - Frame size in milliseconds
- S_o - The overlap of frames in milliseconds
- Simply concatenate the coefficients from each frame together

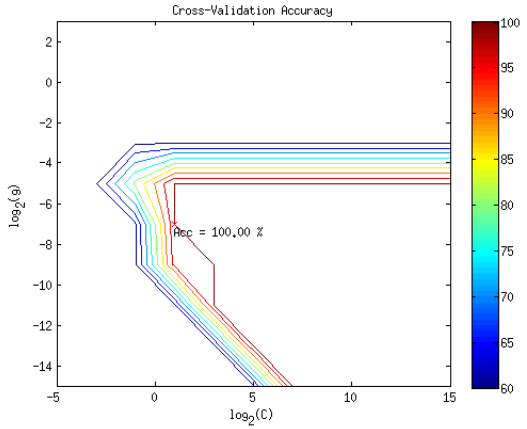


Train Classifier

Support Vector Machines (SVM)

- Linearly separates complex feature space by projecting to a higher dimension
- Use radial basis kernel function $K(x, x') = \exp(\gamma||x - x'||^2))$
- 5 fold cross validation
- Grid search for cost function (C) and γ





Real Time Prediction

Continuously Loop

- Record 2 seconds of audio
- Check for silence
- Calculate the LPC filter coefficients
- Pass the coefficient feature vector through the SVM classifier
- Output identified class

Commands

Simple commands for controlling a remote control vehicle.

Vocabulary

- Start
- Stop
- Left
- Right

SVM

- $N = 100$
- $S_f = 60\text{ms}$
- $S_o = 0\text{ms}$
- 48 Training Samples
- 32 Verification Samples
- Cross Validation Accuracy $= 93.75\%$
- Verification Accuracy $= 96.875\%$

Piano

The notes of the C major scale played on a piano.

Vocabulary

- C4
- D4
- E4
- F4
- G4
- A4
- B4
- C5

SVM

- $N = 100$
- $S_f = 30\text{ms}$
- $S_o = 15\text{ms}$
- 96 Training Samples
- 64 Verification Samples
- Cross Validation Accuracy = 100%
- Verification Accuracy = 100%

Speaker Recognition

The spoken word "test" for two speakers.

Vocabulary

- Speaker 1
- Speaker 2

SVM

- $N = 100$
- $S_f = 60$ms
- $S_o = 15$ms
- 24 Training Samples
- 16 Verification Samples
- Cross Validation Accuracy = 95.83%
- Verification Accuracy = 100%

Simple Arithmetic

The symbols necessary for simple arithmetic.

Vocabulary

- 1, 2, 3, 4, 5, 6, 7, 8, 9, 0
- $+, -, /, \times, =$

SVM

- $N = 100$
- $S_f = 60$ms
- $S_o = 30$ms
- 300 Training Samples
- 60 Verification Samples
- Cross Validation Accuracy $= 96.67\%$
- Verification Accuracy $= 98.33\%$

Evaluation of Results and Future Work

- Very susceptible to changes in environment
  - Background noise
  - Changes in location
  - Changes in speaker
- Significantly increase training database size
- Use other features in classification
  - Fourier Transform
  - Wavelet Transform
- Create better real time prediction environment

References

[1] C.Senthilkumar. "Maximizing the Speech Recognition Accuracy Using Linear predictive Coding" [Online] Available: http://www.academia.edu/4874338/Maximizing_the_Speech_Recognition_Accuracy_Using_Linear_predictive_Coding_Guided_by_Maximizing_the_Speech_Recognition_Accuracy_Using_Linear_predictive_Coding

[2] E. Doering. "Linear Prediction and Cross Synthesis" [Online] Available: https://legacy.cnx.org/content/m15478/latest/

[3] T. Wijoyo, S. Wijoyo. "Speech Recognition Using Linear Predictive Coding and Artificial Neural Network for Controlling Movement of Mobile Robot" [Online] Available: www.ipcsit.com/vol6/36-E091.pdf

[4] U. Shrawankar. "Techniques for Feature Extraction in Speech Rcognition System : A Comparative Study" [Online] Available: http://arxiv.org/ftp/arxiv/papers/1305/1305.1145.pdf

Questions?