Speaker Independent Isolated Word Recognition System

Anirudh B H, Rithwik Udayagiri, Shivam Kumar

National Institue of Techology Karnataka

November 28, 2018

Outline

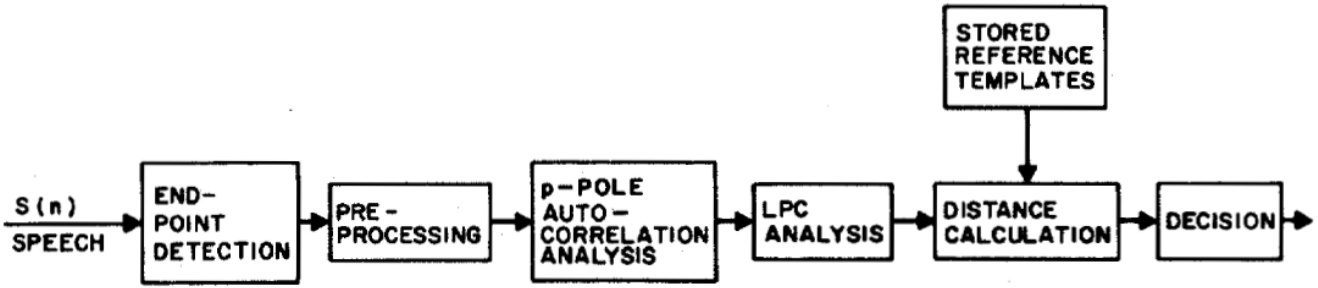
Introduction

- A word recognition system are used in everyday applications such as smart phones and other smart devices.
- With the evolution of IoT and smart devices, the feature of voice commands is now a necessity.
- In this presentation we will be going through one such system that accepts an user input and computes the result as to which word was spoken.

Overview of the Steps Involved

The implementation consists of 4 crucial steps

- Pre-Emphasis
- Feature Extraction : Linear Prediction Coefficients
- Feature Matching : Dynamic Time Warping
- Classification : K Nearest Neighbours



Pre-Emphasis and Windowing

- Pre-Emphasis reduces the variance during distance calculation for the word recognition system
- Filter is given by

$$H(z) = 1 - az^{-1} \tag{1}$$

  where a = 0.96 in our simulations
- Window used is Blackman window
- The l^{th} Pre-Emphasized windowed output is given by

$$x_l[n] = x[l * S + n] * w[n] \qquad 0 \leq n \leq N-1, 0 \leq l \leq L-1 \tag{2}$$

  where S is the Shift, w is the window, x is the speech input

Feature Extraction

- Linear Prediction Coefficients(LPC) are used as features
- The analysis filter is a FIR filter
- Predicts current value from past values
- A p^{th} order system is given

$$\hat{s}[n] = a_1 * s[n-1] + a_2 * s[n-2] + \cdots + a_p * s[n-p] \qquad (3)$$

  where $a_1, a_2, \ldots, a_p$ are the prediction coefficients
- p = 8 was chosen in the simulation

LPC Extraction

- The Autocorrelation function $R(k)$ of the windowed signal is calculated

$$R(k) = \sum_{-\infty}^{+\infty} x_l[n] * x_l[n-k] \qquad 0 \leq k \leq p \tag{4}$$

- Toplitz Matrix is calculated

$$\begin{bmatrix} R(0) & R(1) & \dots & R(p-1) \\ R(1) & R(2) & \dots & R(p-2) \\ \vdots & \vdots & & \vdots \\ R(p-1) & R(p-2) & \dots & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ \vdots \\ a_p \end{bmatrix}$$

Feature Matching

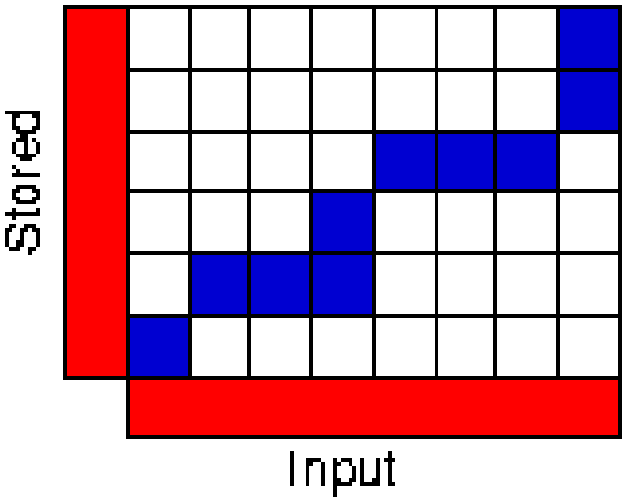
- Features are matched to templates
- The lenght of input and output need not be same
- Hence the Dynamic Time Warping algorithm is used
- If input is of length n and output of length m

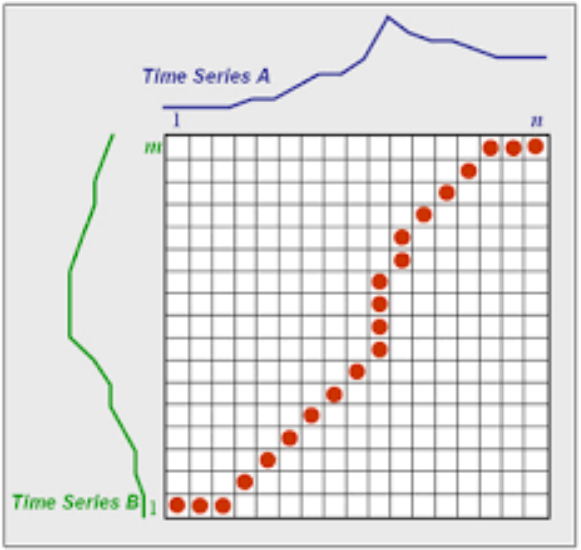
$$D(i,j) = d(i,j) + min(D(i-1,j-1), D(i,1,j), D(i,j-1)) \quad (5)$$
$$where \quad 0 \leq i \leq n-1, 0 \leq j \leq m-1$$

$d(i,j)$ is the distance between the feature vectors of the i^{th} frame of input and j^{th} frame of the template.
$D(i,j)$ is the Distance matrix. The final element is the DTW Distance

Dynamic Time Warping(DTW)

- Greedy algorithm
- Warps the time scale to compensate for difference in speed and time





Nearest Neighbours

- The K Nearest Neighbours Rule is used for classification
- The KNN rule is given by

$$r_j = \frac{1}{K} \sum_{k=1}^{K} D[x, x_{[k]}^{(j)}] \tag{6}$$

where $D[x, x_{[k]}^{(j)}]$ is the distance between input x and j^{th} word.

j^* is the recognized word, given by

$$r_{j^*} \leq r_j \qquad j = 1, 2, \ldots, J \tag{7}$$

Clustering

- The LPC varies from one speaker to another
- Large number of templates needed for better matching

But this increases computation time
So clustering is performed

- In large datasets, not all speaker's LPC are distinct
- They can be clustered into many groups
- One representative is chosen from each cluster

K-Mean Cluster

- Random points are chosen as centers
- For each point is colsest center is chosen by distance computation
- Each point is assigned to the same culster as the closest center
- The mean for each cluster is computed. This is the new center
- Steps 2, 3 and 4 are repeated for a fixed number of iteration or till error decreases below a lower bound

- An alternate version of this using the median instead of mean was also implemented

Simulation Results

Measure	K = 1	K = 2	K = 5	K = 7
Mean for C = 15	46.48	47.52	47.90	45.33
Median for C = 15	52.10	53.80	57.90	57.24
Mean for C = 25	45.90	49.81	50.19	50.19
Median for C = 25	52.29	57.52	60.00	59.90

K is the number of nearest neighbours
C is the number of cluster

- Median clusters gave better reults
- K = 5 neighbours gave the highest accuracy
- An increse in accuracy was observed by increasing C

Acknowledgements

We deeply express our gratitude towards Dr.Aparna P for giving us this wonderful opportunity to work on this project. This project gave us the opportunity to learn new and important concepts in the field of Audio and Speech Processing.

Thank You