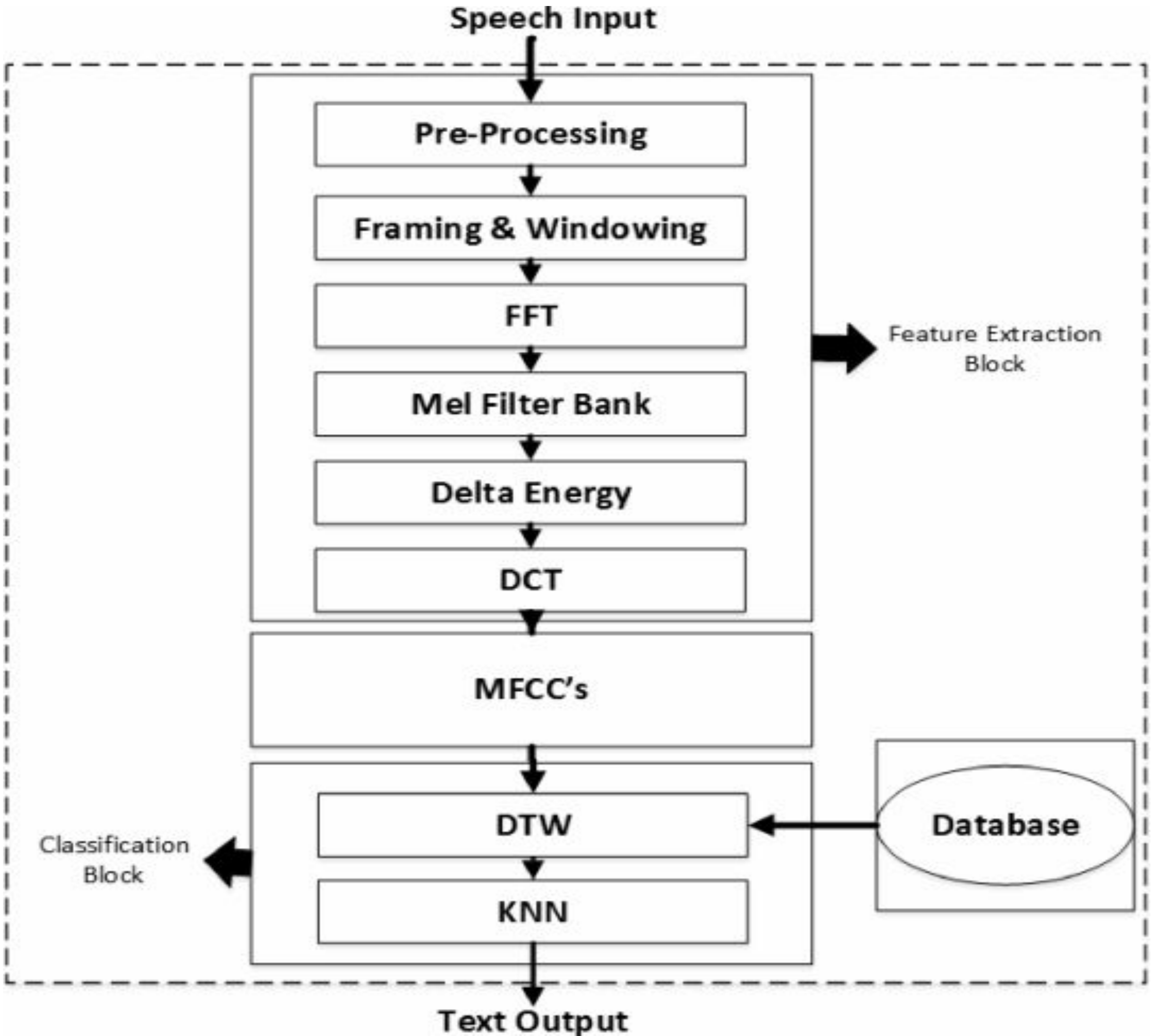
Home Automation

On Embedded System
By Isolated Word Recognition
Using Machine Learning

Introduction

- Approach of ASR system based on isolated word structure using Mel-Frequency Cepstral Coefficients (MFCC's), Dynamic Time Wrapping (DTW) and K-Nearest Neighbor (KNN) techniques
- The Mel-Frequency scale used to capture the significant characteristics of the speech signals; features of speech are extracted using MFCC's
- DTW is applied for speech feature matching.
- KNN is employed as a classifier.


Speech Input
Pre-Processing
Framing & Windowing
FFT
Mel Filter Bank
Delta Energy
DCT
Feature Extraction Block
MFCC's
DTW
KNN
Database
Classification Block
Text Output

Algorithm

- Creating a training dataset
- Computing MFCC
- Calculating Distance of each word using DTW library by https://github.com/pierre-rouanet/dtw
- Applying KNN classifier to entire cost matrix
- Testing
- Computing MFCC
- Map the MFCC into the classifier to get the predicted output

Going through code

Speech Recognition Methods

1. Time Domain
- Involves observing zero crossing rates of signal
- Short Time Energy of signal
- Amplitude Variations
- Variation in speed

2. Frequency Domain

- Extracting MFCC features

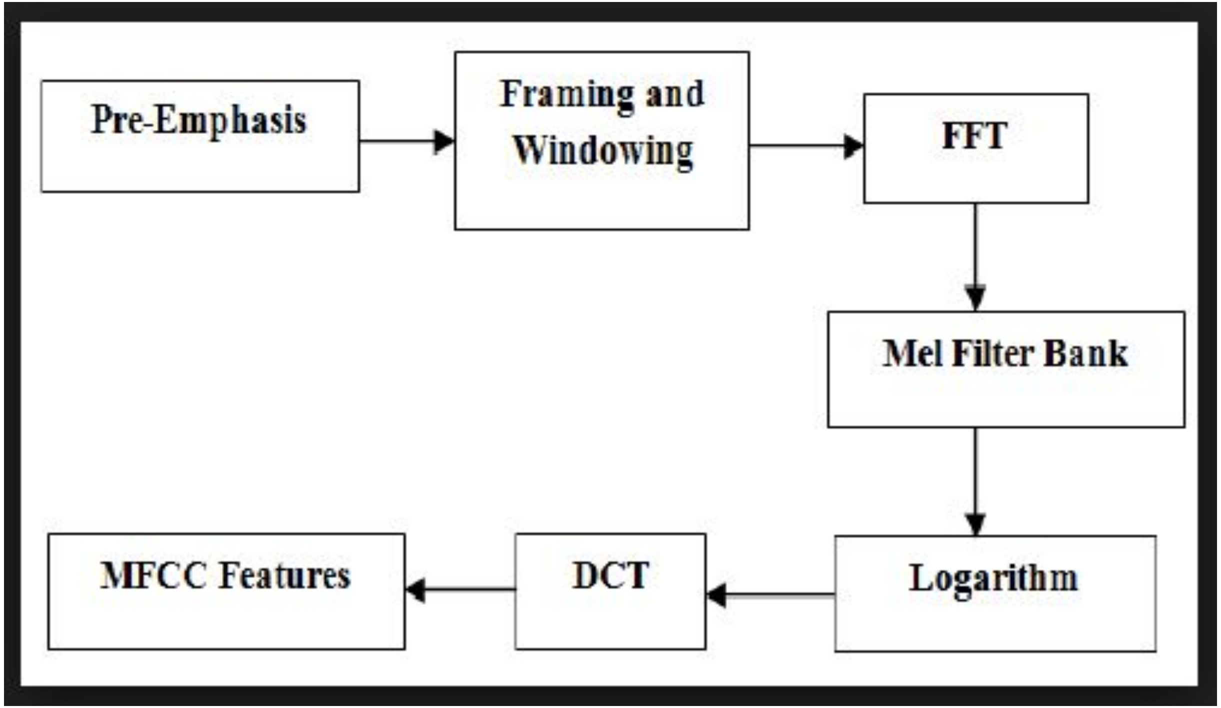
Why Frequency Domain??

- Distinguishing features can be recognized.
- Frequency-domain analysis shows how the signal's energy is distributed over a range of frequencies
- A frequency-domain representation also includes information on the phase shift that must be applied to each frequency component in order to recover the original time signal with a combination of all the individual frequency components.
- Frequency-domain analysis becomes useful when you are looking for cyclic behavior of a signal.

MFCC

- The shape of the vocal tract manifests itself in the envelope of the short time power spectrum, and the job of MFCCs is to accurately represent this envelope.
- Mel Frequency Cepstral Coefficients (MFCCs) are a feature widely used in automatic speech and speaker recognition.
- The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale.
- This approximates the human auditory system response more closely than the linearly-spaced frequency bands used in the normal cepstrum.
- This frequency warping can allow for better representation of sound
- That is because MFCC can better describe the nonlinear relation that humans ear feels the frequency of speech signal.

Overview of MFCC

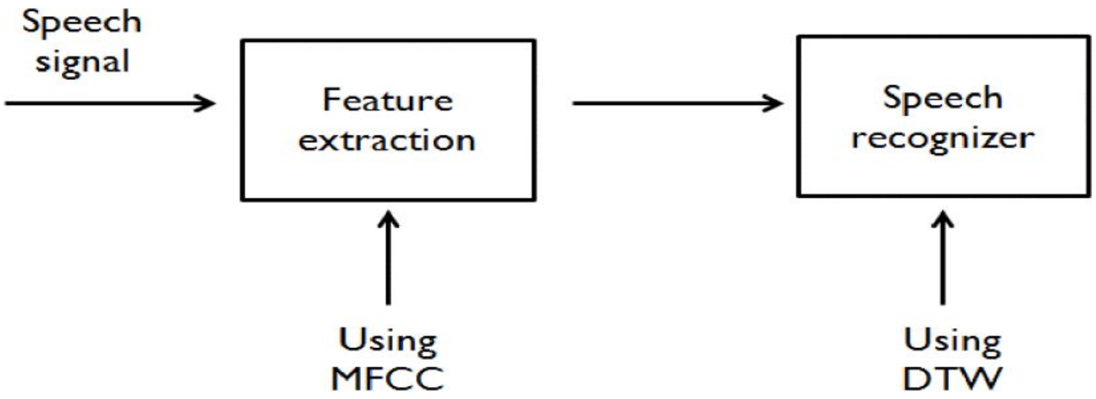


Dynamic Time Warping

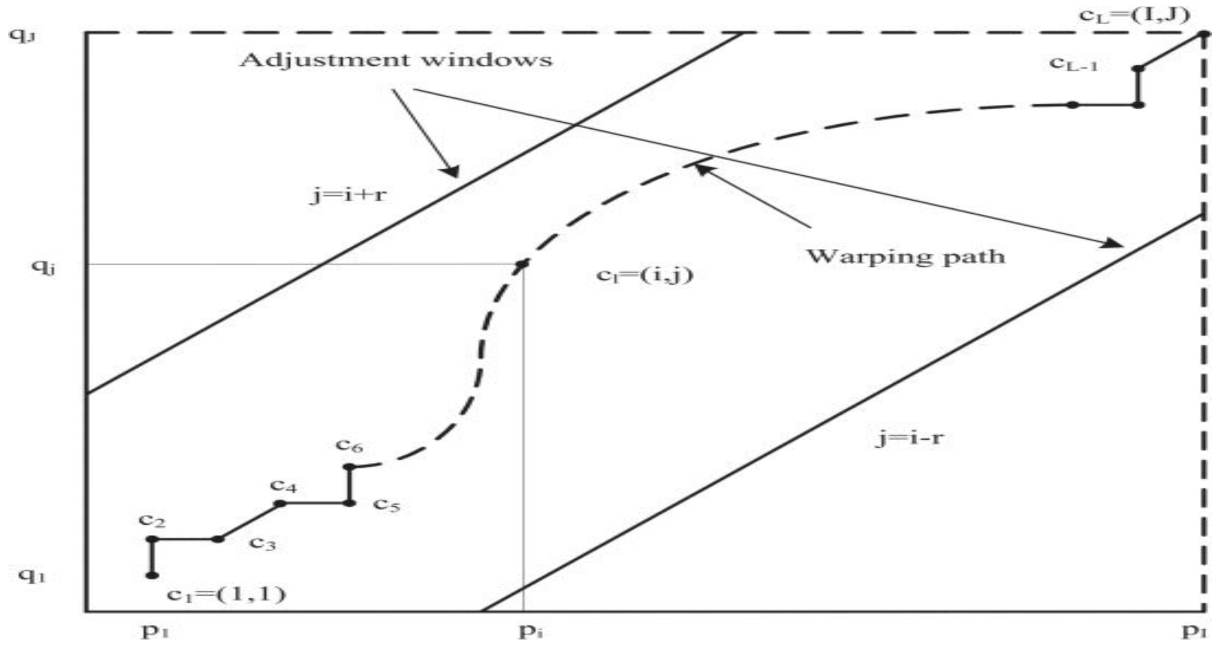
- Dynamic time warping (DTW) is one of the algorithms for measuring similarity between two sequences, which may vary in speed
- DTW has been applied to temporal sequences of video, audio, and graphics data — indeed, any data that can be turned into a linear sequence can be analyzed with DTW
- A well known application has been automatic speech recognition, to cope with different speaking speeds.
- Applications include speaker recognition and online signature recognition.

Why DTW

There are two main techniques in speech recognition. One is hidden markov model (HMM), the other is DTW. Although HMM is a very popular technique in speech recognition, DTW is still used in the small-scale embedded systems (e.g. cell phones, mobile applications) because of simplicity of its hardware implementation, straightforwardness and speed of the training procedure. The Fig shows a simple speech recognition system using DTW.



The objective of DTW is to warp two speech templates P=(p1,p2,···,pI) and Q=(q1,q2,···,qJ) in the time dimension as represented in Fig. 3. Each pi and qj is a vector of parameters (MFCC).



These two speech templates are of the same category, the timing differences between them can be depicted by a sequence of points $c = (i, j)$:

$$C = c(1), c(2), \cdots, c(L) \tag{1}$$

▸ View Source

where

$$c(l) = (i(l), j(l)) \tag{2}$$

▸ View Source

This sequence can be considered to represent a warping path which approximately realizes a mapping from the time axis of template P onto that of template Q. As a measure the difference between two speech vectors p_i and q_j. a distance $d(i, j)$ is defined.

$$d(c) = d(i, j) = \|a_i - b_j\| \tag{3}$$

▸ View Source

We will compute the distance between the starting point (1, 1) and the end point (I, J) from left to right $D(I, J)$.

$$D(C) = \sum_{l=1}^{L} d(c(l)) \tag{4}$$

Since there are X possible paths from (1, 1) to (I, J), We will identify the smallest accumulated distances from (1, 1) to (I, J) among all possible, and the path which has the minimum $D(I, J)$ is the optimal path between P and Q.

KNN Algorithm

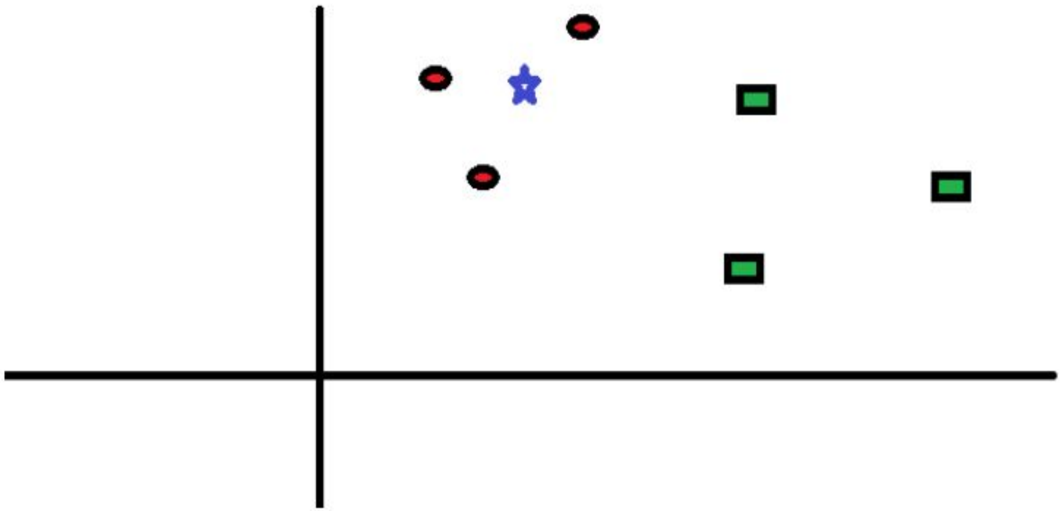
- In pattern recognition, the *k*-nearest neighbors algorithm (*k*-NN) is a non-parametric method used for classification and regression.[1]

- In both cases, the input consists of the *k* closest training examples in the feature space.

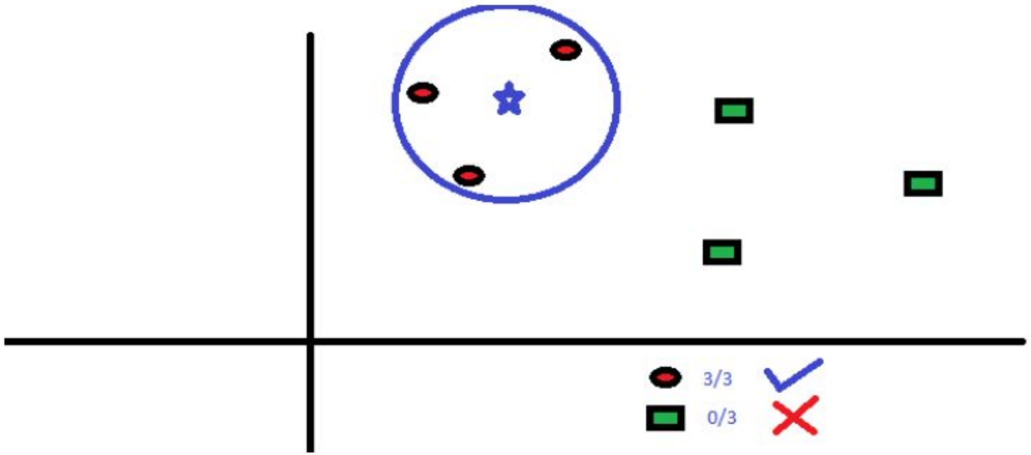
Why KNN??

	Logistic Regression	CART	Random Forest	KNN
1. Ease to interpret output	2	3	1	3
2. Calculation time	3	2	1	3
3. Predictive Power	2	2	3	2

How does the KNN algorithm work?



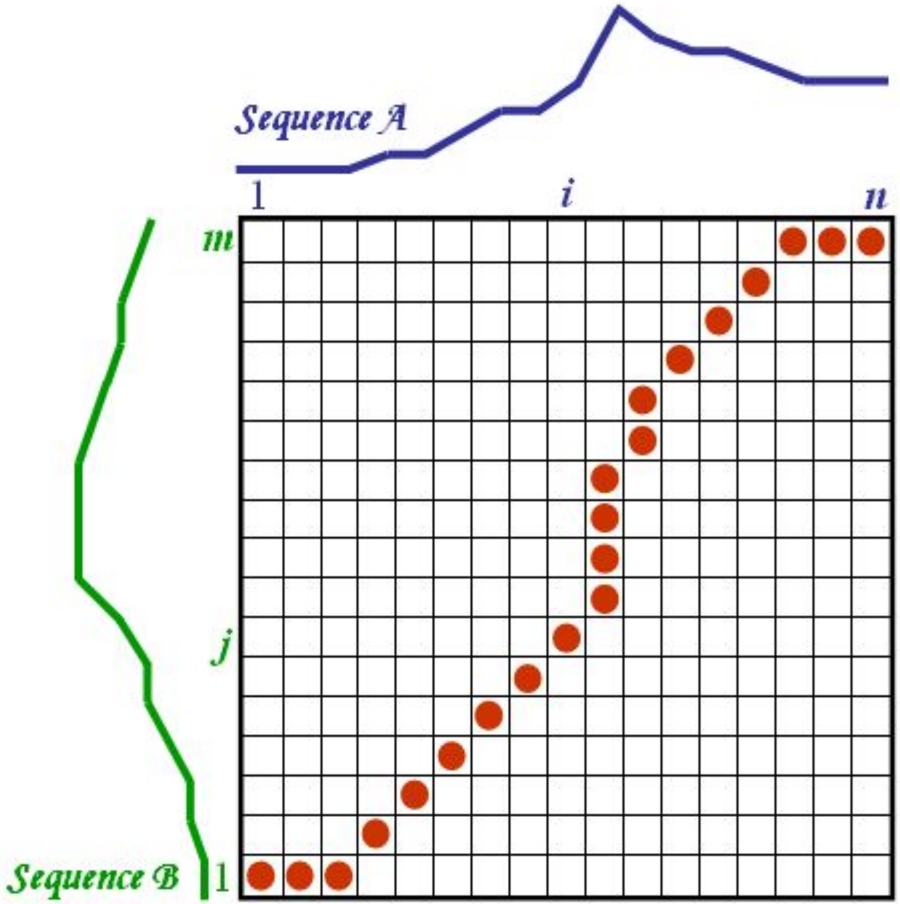
You intend to find out the class of the blue star (BS) . BS can either be RC or GS and nothing else. Let's say K = 3. Hence, we will now make a circle with BS as center just as big as to enclose only three datapoints on the plane.



Testing Phase..

- Recording for the whole word
- Calculate the MFCC's for each frame. (recording of the whole word, not just part of it.)
- Calculate the distance between recording and each of the templates in database.
- (In case of DTW) Calculate the cost between each frames (simple distance metric/norm, i.e. Euclidean, Manhattan, etc.).
- Once the DTW algorithm is finished, we will end up with the distance value between your test sample and each of the templates.

- The last step is to make a decision: to which class the test sample actually belongs to.
- One method is do it by picking the class of template with the minimum DTW distance.
- Better method is using the k-Nearest Neighbours for that.


Sequence A
1
i
n
m
j
Sequence B
1