Isolated Word Recognition

Team No. 47
Vinamra Benara
Dhrumil Patel
Deep Jahan Grewal

Introduction:

- Speech Recognition of isolated words
- Use Mel frequency cepstral coefficients as features
- Use different models to train and test features like:
  - Hidden Markov Model
  - Gaussian Mixture Model
  - Vector Quantization
  - Recurrent Neural Networks

Research Papers

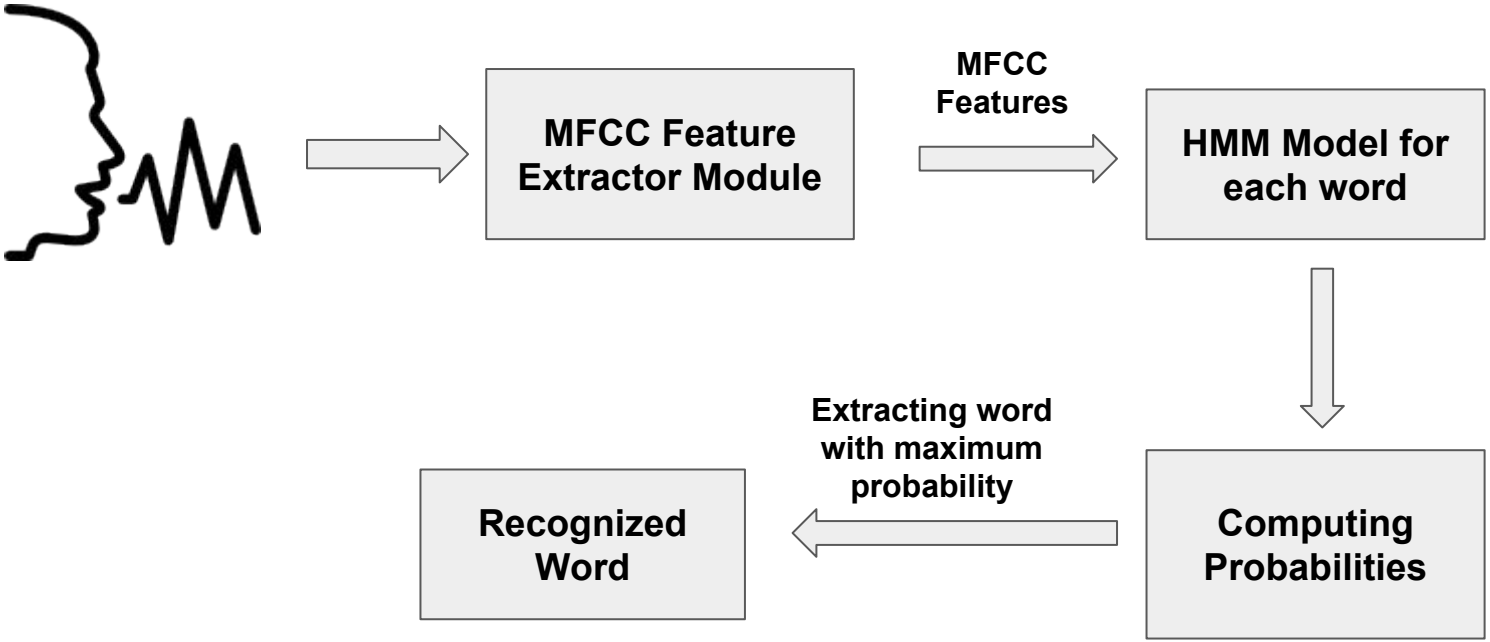
Speech-To-Text Conversion (STT) System Using Hidden Markov Model (HMM) http://www.ijstr.org/final-print/june2015/Speech-to-text-Conversion-stt-System-Using-Hidden-Markov-Model-hmm.pdf

Isolated Word Speech Recognition Using Vector Quantization (VQ)

http://www.ijarcsse.com/docs/papers/May2012/Volum2_issue5/V2I500451.pdf

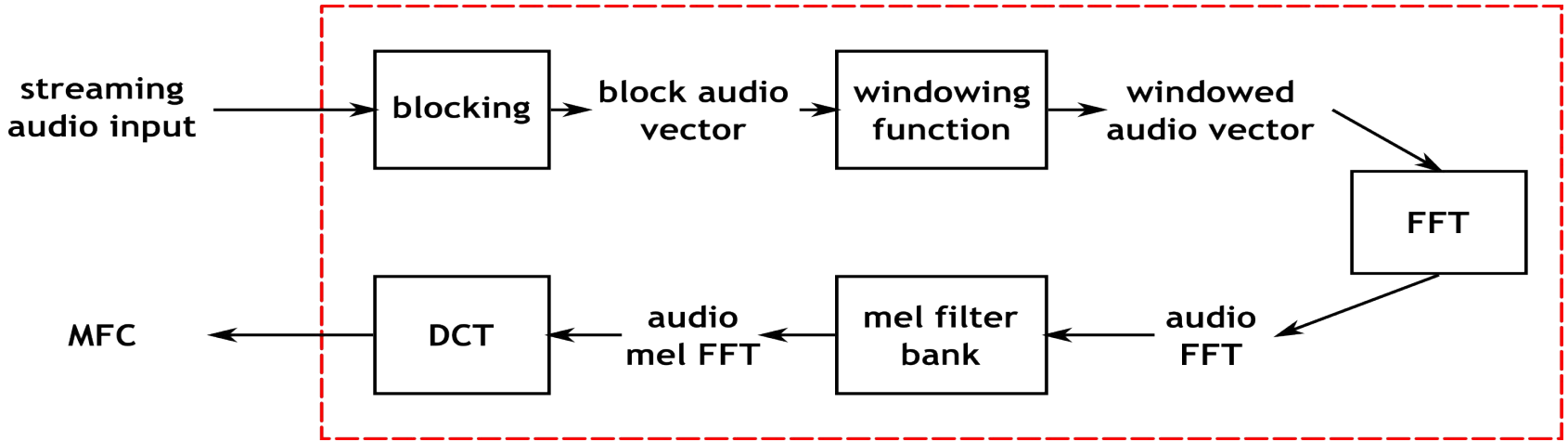
Speech recognition with recurrent neural network http://ieeexplore.ieee.org/Xplore/defdeny.jsp?url=http%3A%2F%2Fieeexplore.ieee.org%2Fstamp%2Fstamp.jsp%3Ftp%26arnumber%3D6638947&denyReason=-133&arnumber=6638947&productsMatched=null&userType=inst

Workflow...


MFCC Feature Extractor Module
MFCC Features
HMM Model for each word
Computing Probabilities
Extracting word with maximum probability
Recognized Word

Mel-frequency cepstral coefficients (MFCC)

- Mel Frequency Cepstral Coefficients (MFCCs) are the features widely used in automatic speech and speaker recognition.
- It identifies the components of the audio signal that are good for identifying the linguistic content and discarding all the other stuff which carries information like background noise, emotion etc.

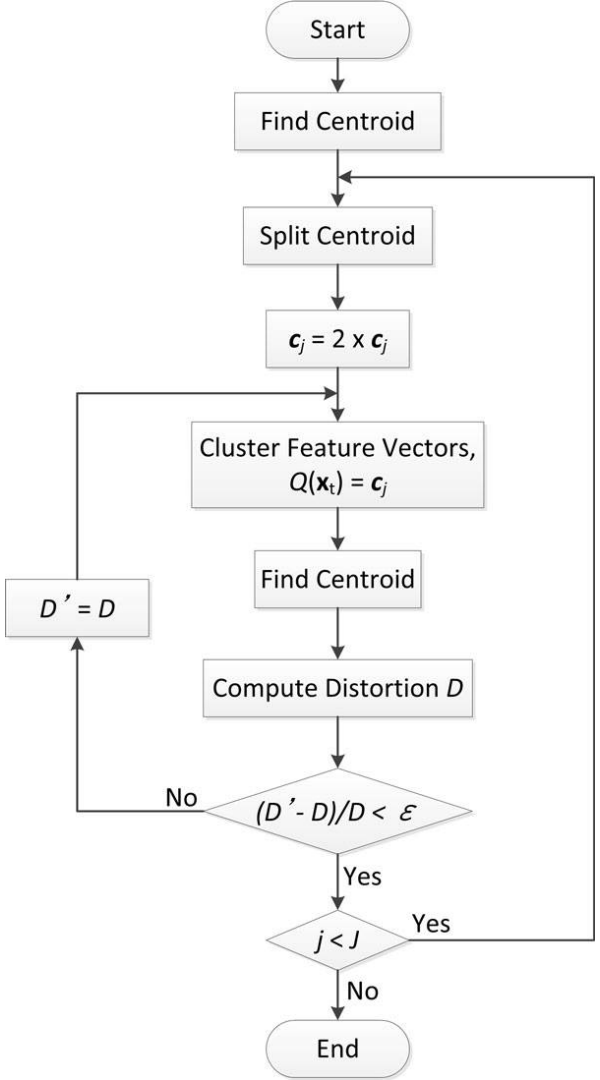


Vector Quantization

In this method , the data is compressed via lossy technique. A codebook is formed for each data which is a compressed version and is representative of that data. It is based on the principle of block coding.

A code-book is a reduced dataset version which can be achieved by many methods such as incremental K-Means Clustering, spectral clustering etc.

1 codebook is generated for each training dataset.

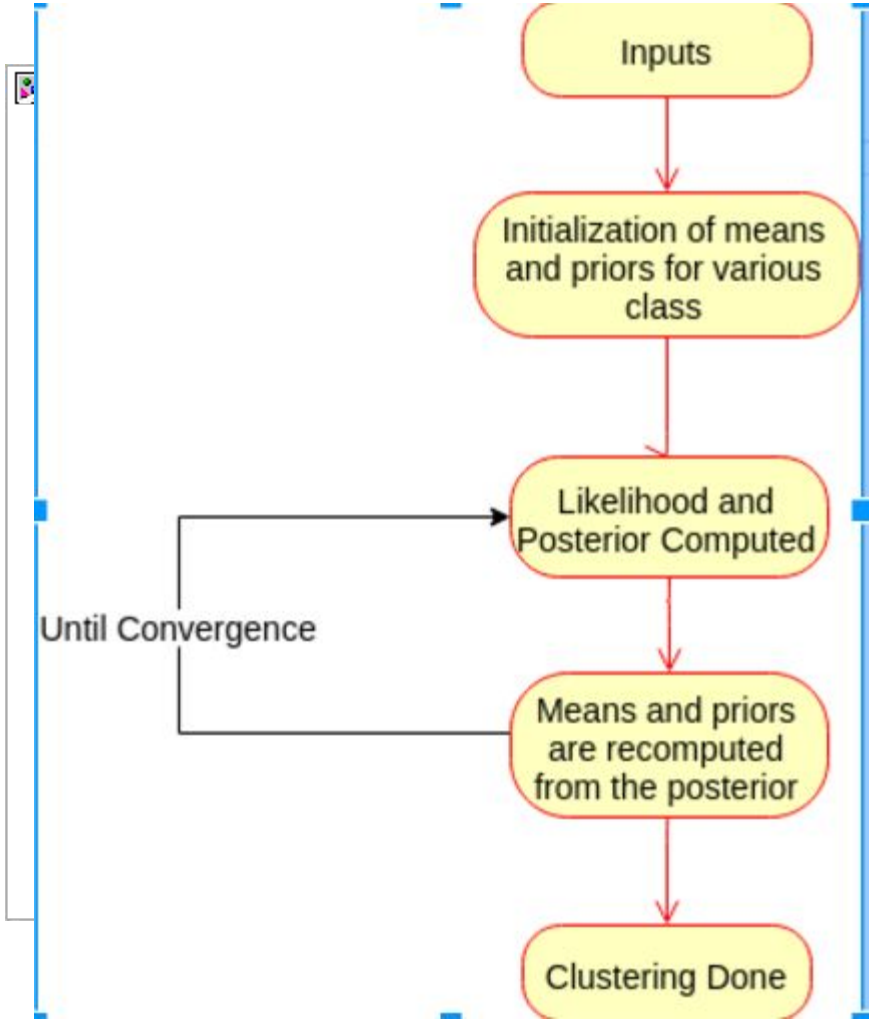

Start
Find Centroid
Split Centroid
$c_j = 2 \times c_j$
Cluster Feature Vectors, $Q(x_t) = c_j$
Find Centroid
Compute Distortion D
$(D' - D)/D < \varepsilon$
No
$D' = D$
Yes
$j < J$
Yes
No
End

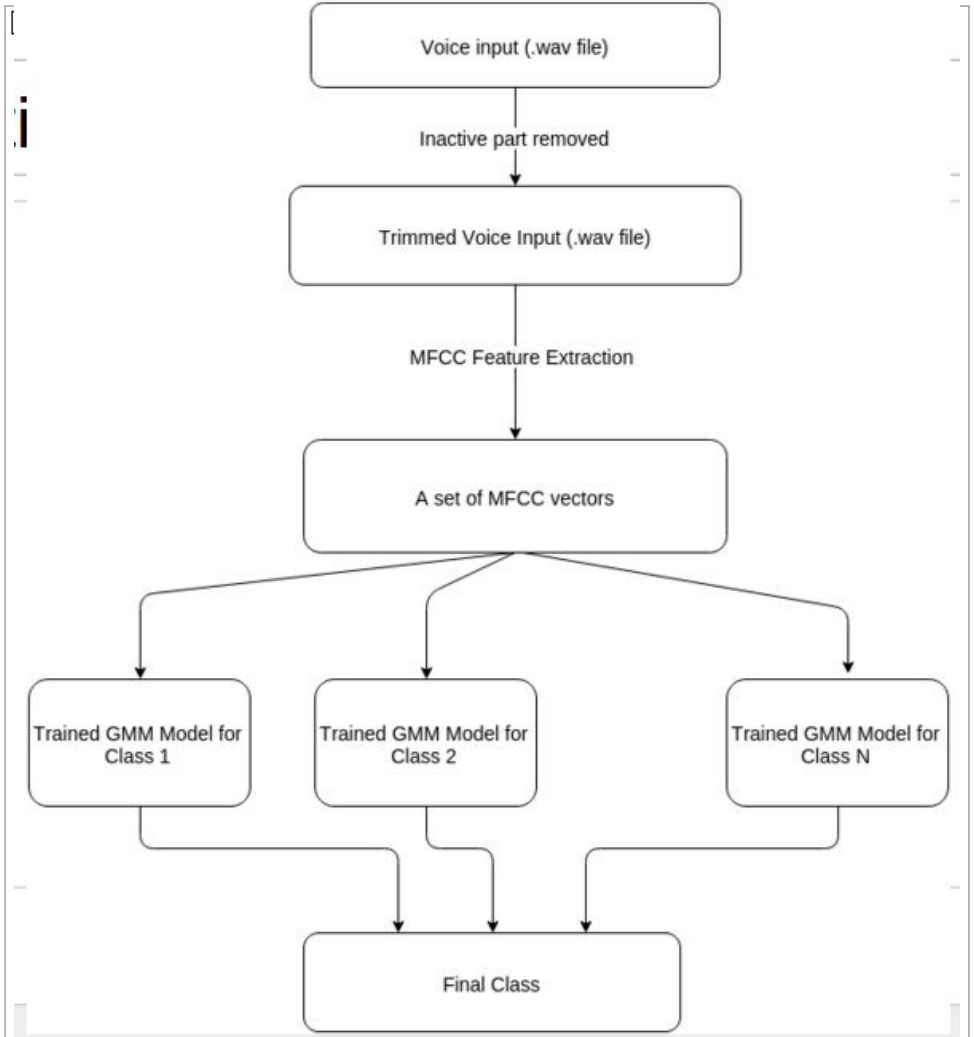
GMM

GMM is a probabilistic clustering method which does probabilistic soft clustering rather than conventional hard clustering.

It is represented as a weighted sum of Gaussian component densities. It is parametric problem where the means ,priors and variances are optimised using the EM algorithm Gmm is an approximate method wherein the underlying distribution is assumed to be gaussian.

Although this does not cover all the intricacies of the data , but it definitely makes the analysis easier and analytically tractable. They provide a smooth distribution of arbitrary densities. Initially the parameters are assumed to be a random value and then iteratively a local optima is achieved by EM , which is a disadvantage, hence the final result depends on initialization. So to get a global optima, various re-initializations may be required.


Inputs
Initialization of means and priors for various class
Likelihood and Posterior Computed
Until Convergence
Means and priors are recomputed from the posterior
Clustering Done

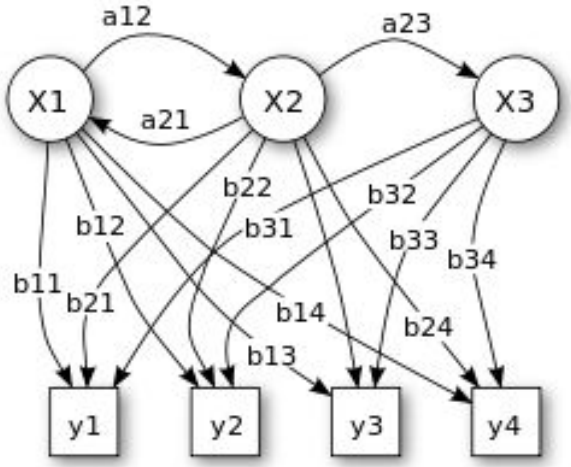

Voice input (.wav file)
Inactive part removed
Trimmed Voice Input (.wav file)
MFCC Feature Extraction
A set of MFCC vectors
Trained GMM Model for Class 1
Trained GMM Model for Class 2
Trained GMM Model for Class N
Final Class

Hidden Markov Model (HMM)

A **hidden Markov model** (**HMM**) is a statistical **Markov model** in which the system being modeled is assumed to be a **Markov process** with unobserved (**hidden**) states.

It models every outcome as a parametric random variable with certain probabilities under different situations.

Then the one with highest probability is selected.



Here: **X** = states; (**Not visible**)

a = transition probabilities;

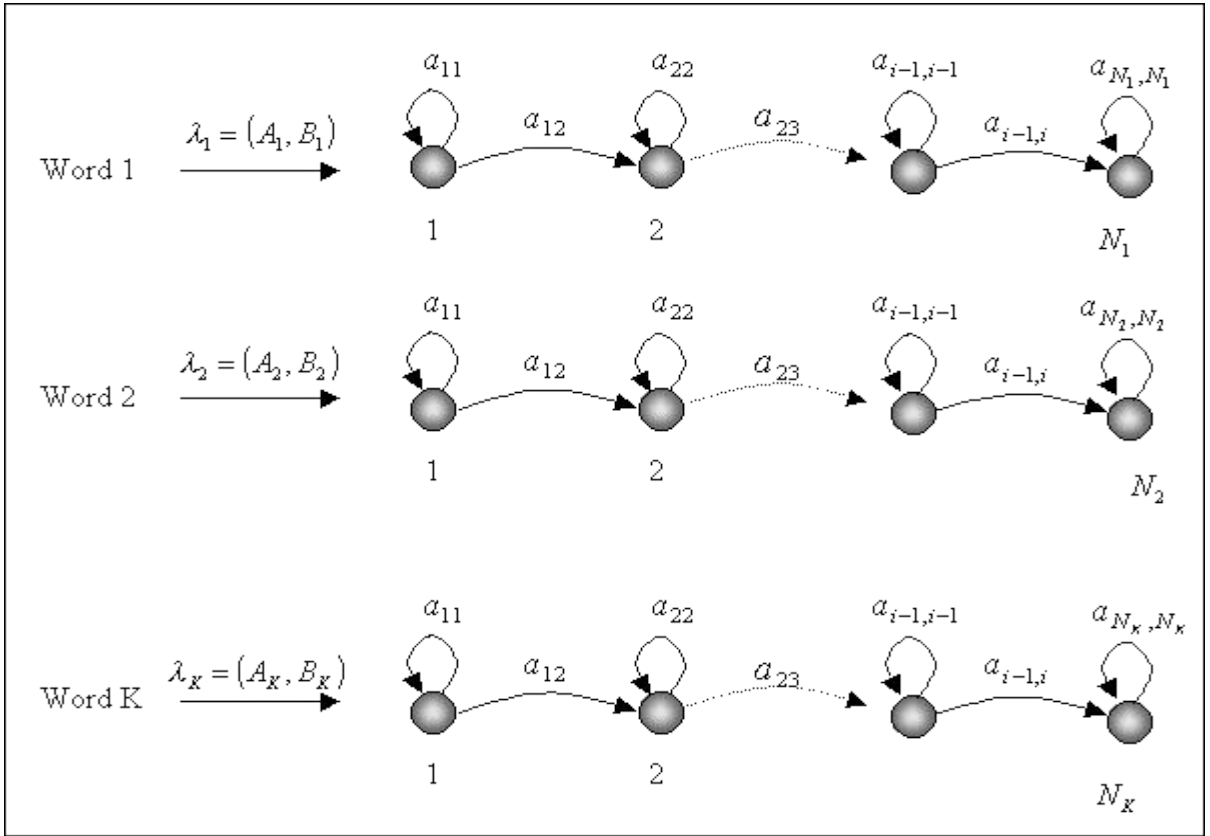
Y = possible observations; (**Visible**)

b = output probabilities;

HMM in Isolated Word Recognition

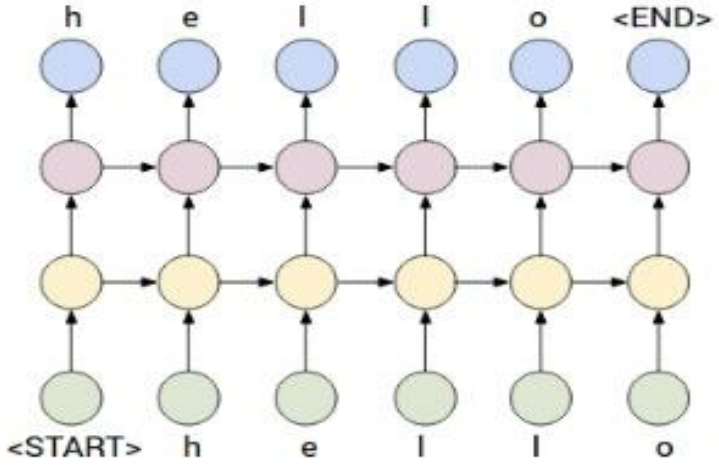
- Each word v modeled as distinct HMM H_v
- Training set of k occurrences per word
  - Each of which is an observation sequence
- Need to:
  - Estimate parameters for each Hv that maximize $P(O_1,\ldots,O_k \mid H_v)$ (i.e. **Baum-Welch Algorithm**)
  - Extract features $O = (O_1,\ldots,O_T)$ from unknown word
  - Calculate $P(O \mid H_v)$ for all v (**Forward Algorithm**), find v which maximizes

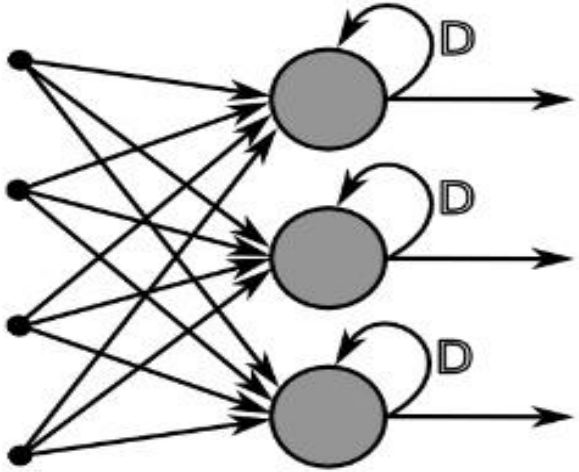
Its Interpretation

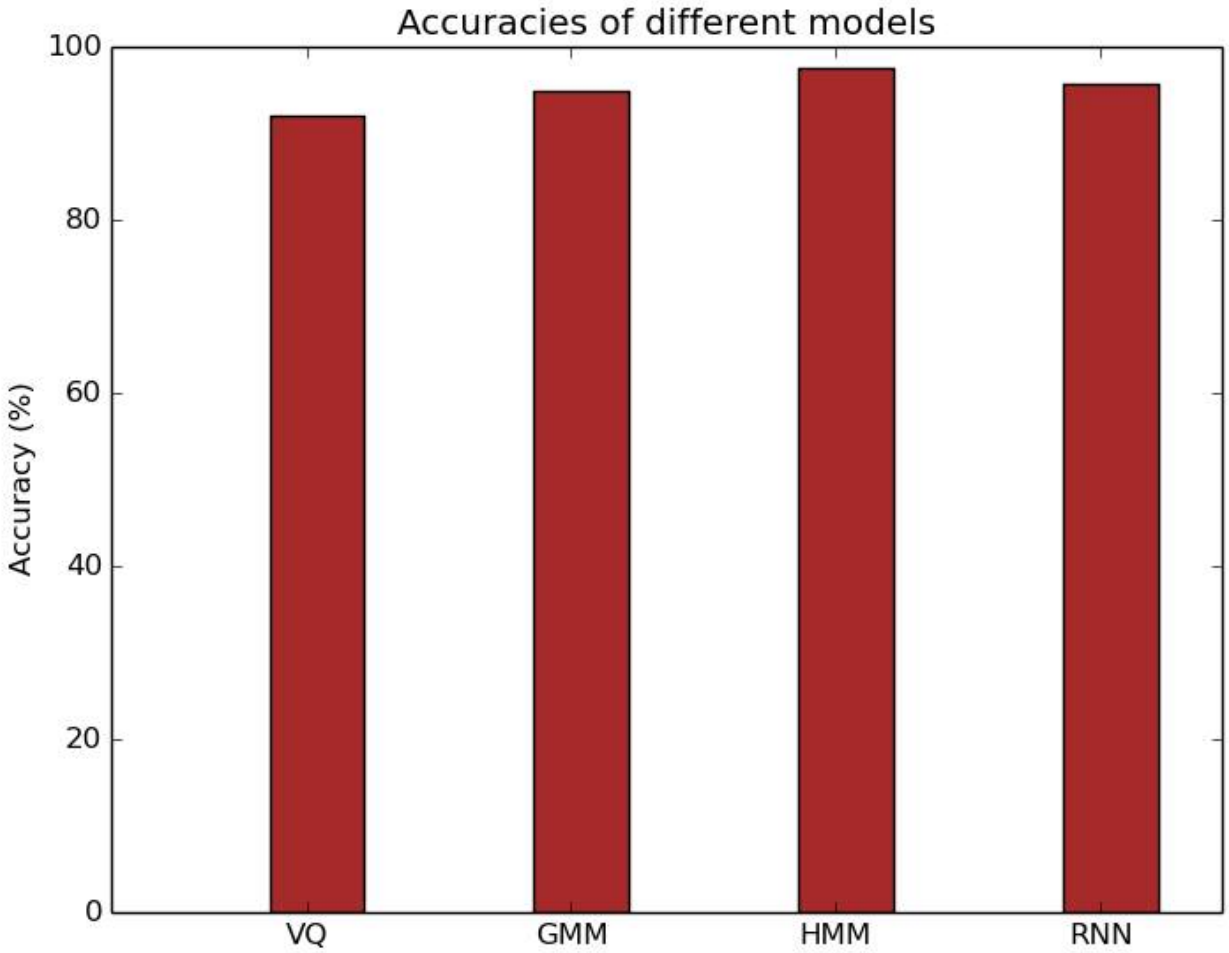


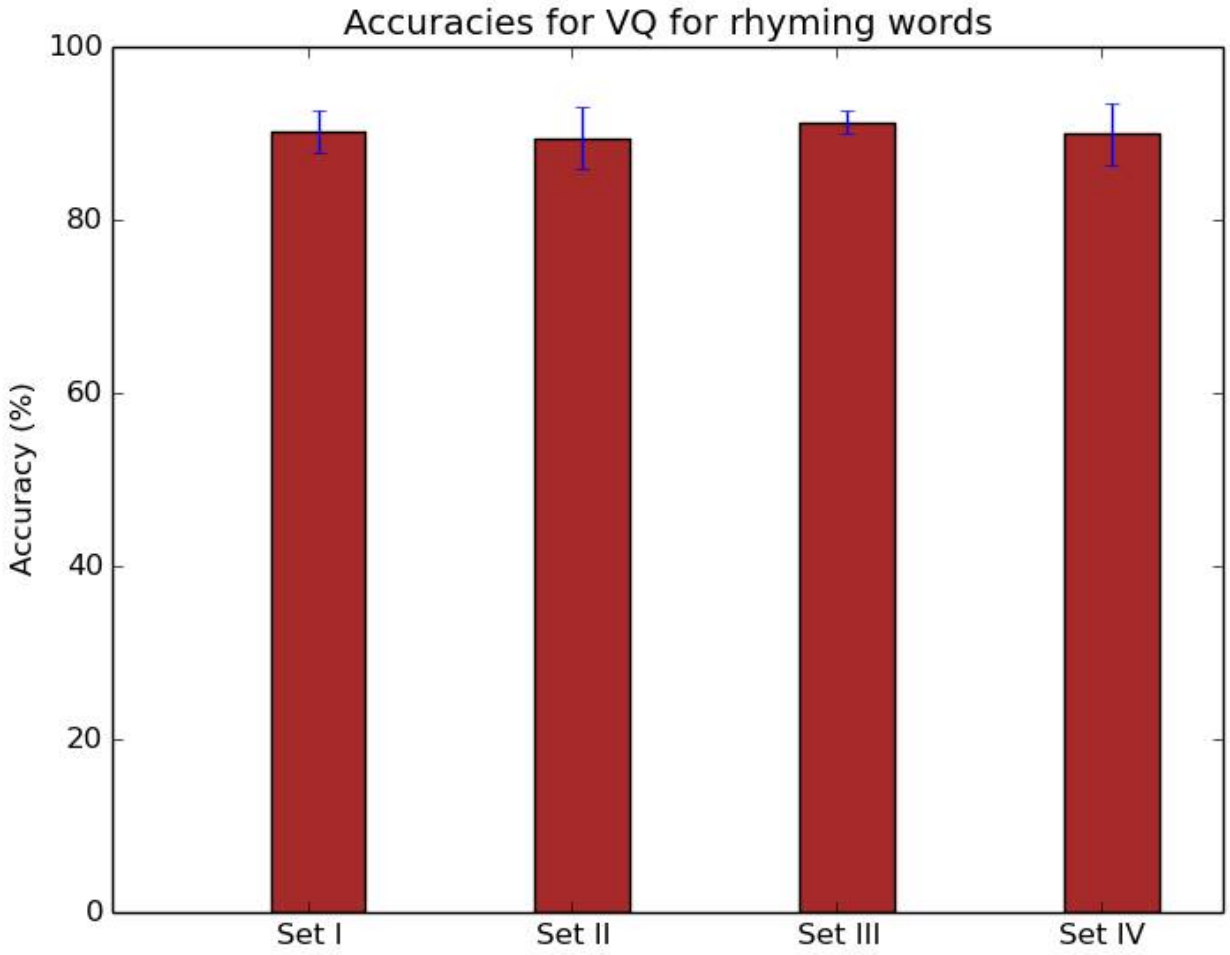
What are Recurrent Neural Networks

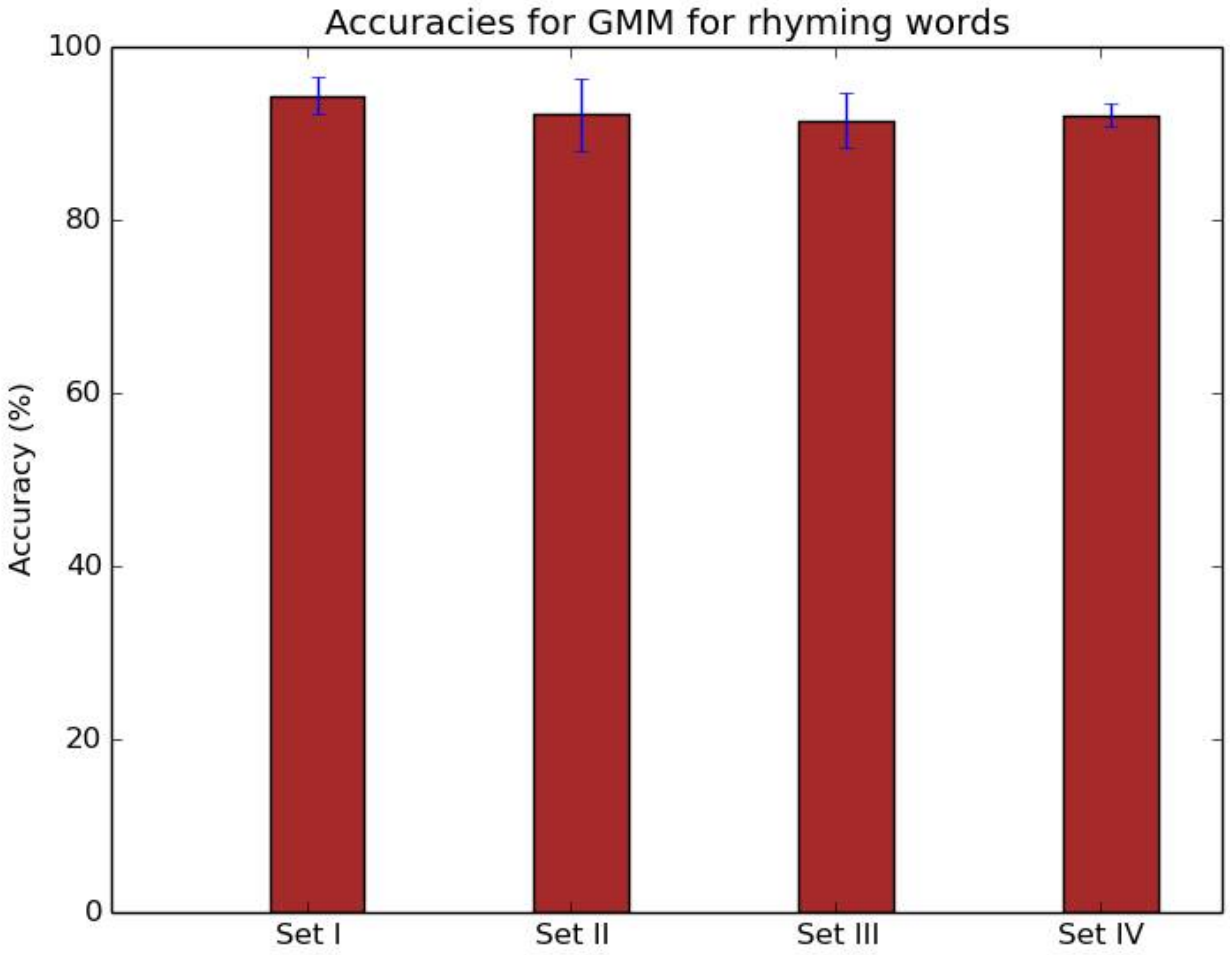
1. In their simplest form (RNNs), they are just Neural Networks with a feedback loop
2. The previous time step's hidden layer and final outputs are fed back into the network as part of the input to the next time step's hidden layers.

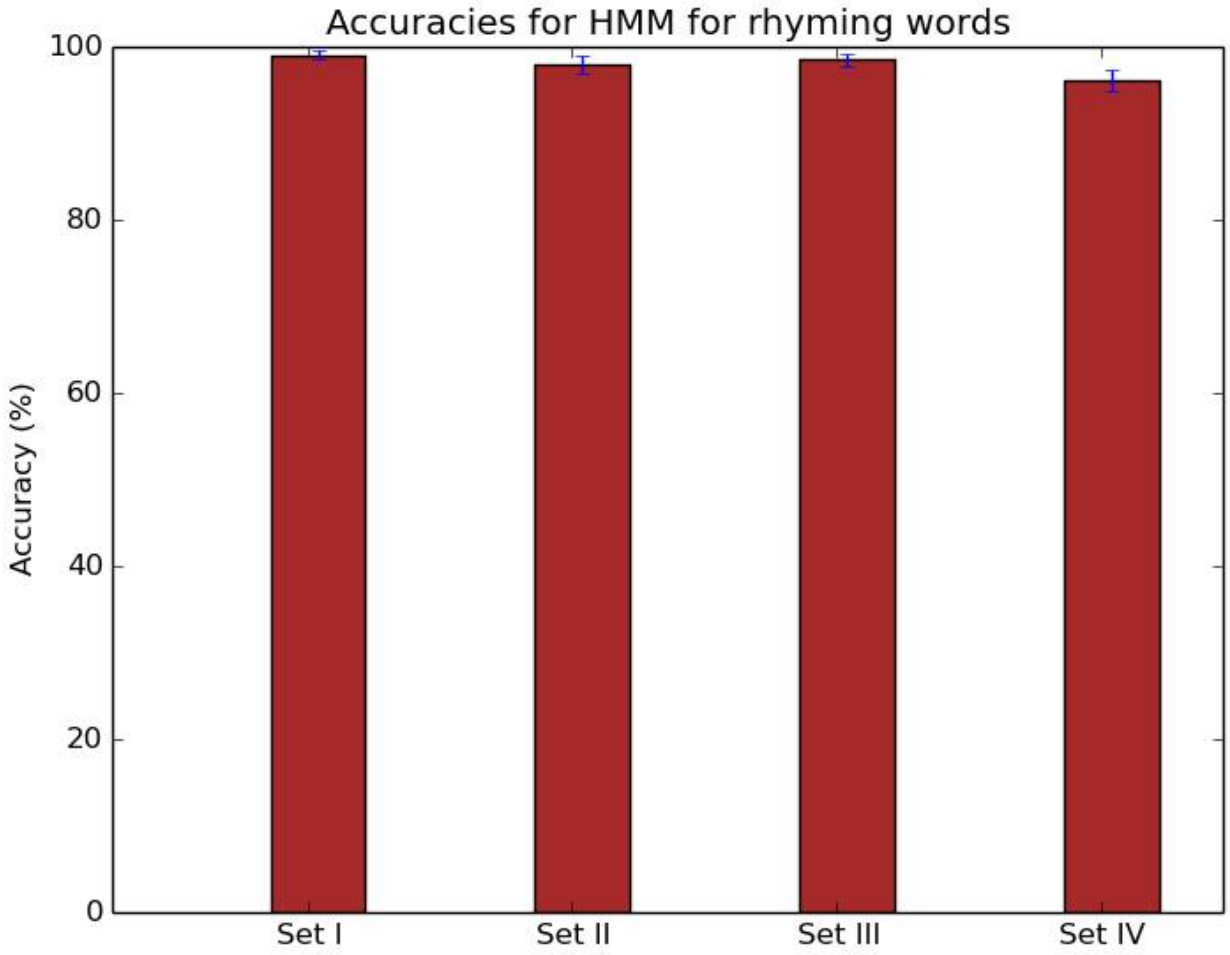


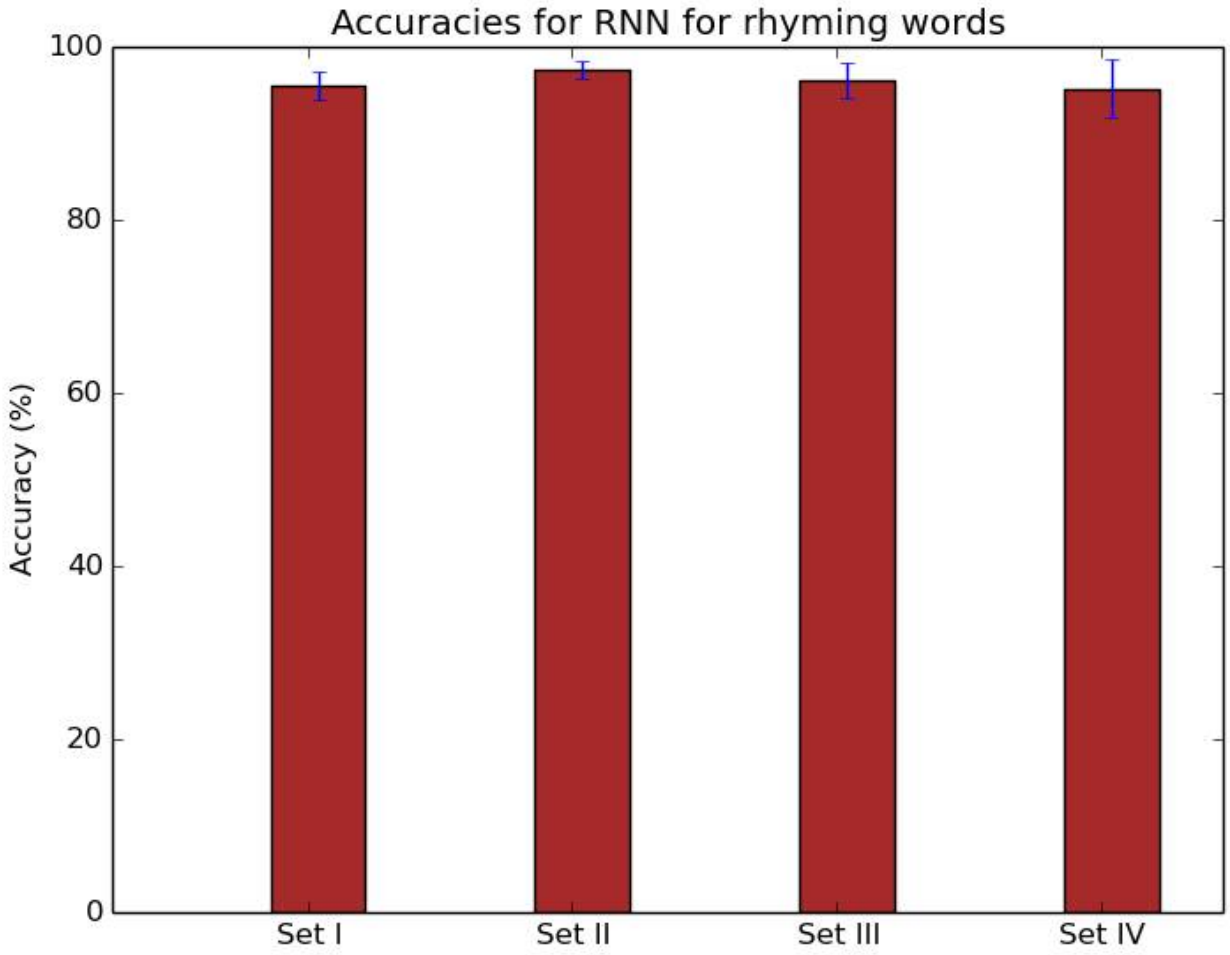


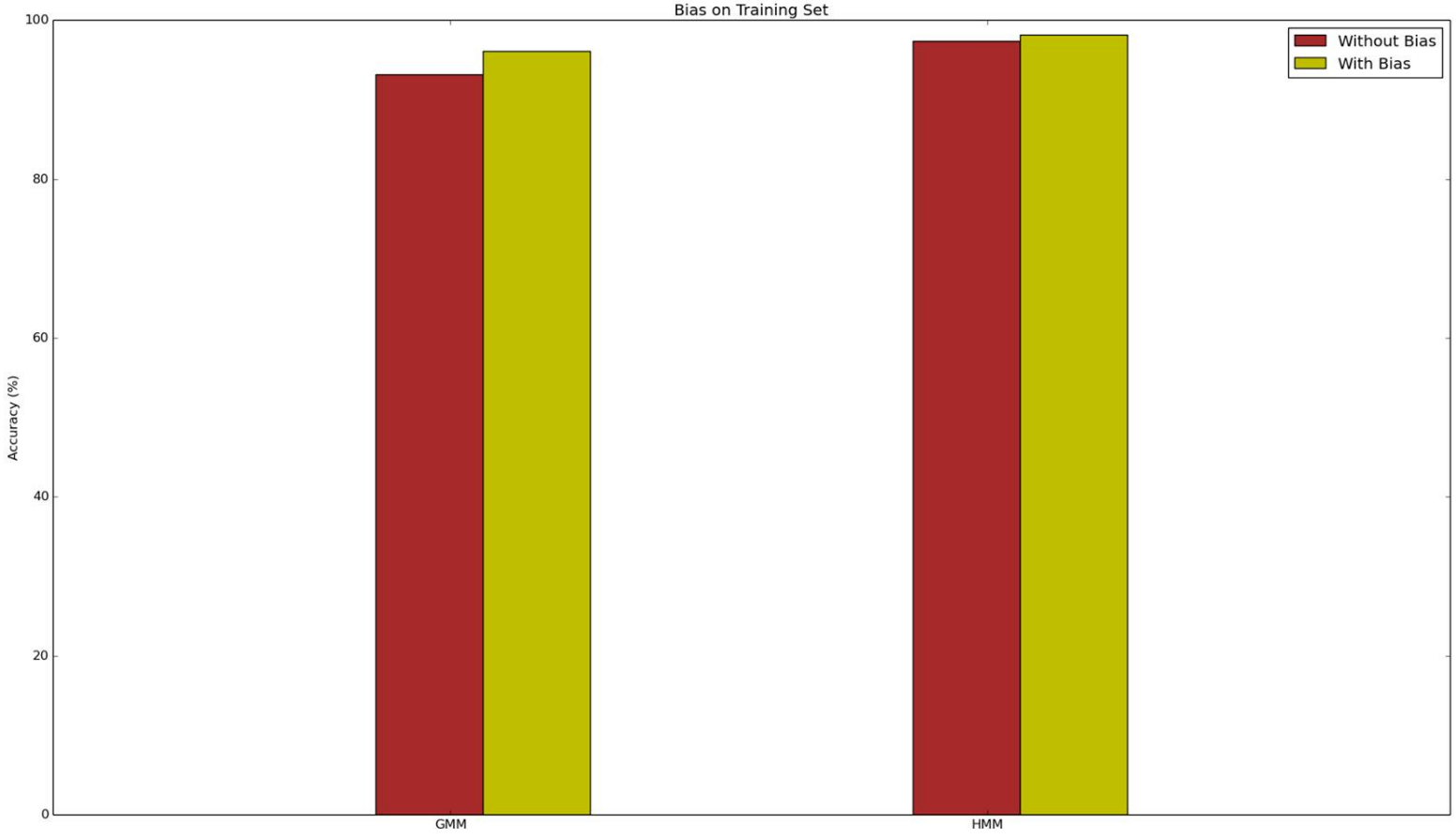

Accuracies of different models
Accuracy (%)
100
80
60
40
20
0
VQ
GMM
HMM
RNN


Accuracies for VQ for rhyming words
Accuracy (%)
100
80
60
40
20
0
Set I
Set II
Set III
Set IV


Accuracies for GMM for rhyming words
Accuracy (%)
100
80
60
40
20
0
Set I
Set II
Set III
Set IV


Accuracies for HMM for rhyming words
Accuracy (%)
100
80
60
40
20
0
Set I
Set II
Set III
Set IV


Accuracies for RNN for rhyming words
Accuracy (%)
100
80
60
40
20
0
Set I
Set II
Set III
Set IV


Bias on Training Set
Without Bias
With Bias
Accuracy (%)
100
80
60
40
20
0
GMM
HMM

THANK YOU