Hidden Markov Model (HMM) based speech classification system

EQ2340 - Pattern Recognition

Lars Kuger & Navneet Agrawal

November 1, 2016

Self Introduction

Lars Kuger

- Exchange Student
- RWTH Aachen, Germany
- Wireless Systems program

Navneet Agrawal

- Masters student
- KTH, Stockholm, SE
- Wireless Systems program

Application

- Hands-free Calendar APP
- Based on word (speech) recognition
- Data: Audio recording of words used by four different people
- Words: *Hello*, *Bye*, *Yes*, *No*, *Monday*, *Tuesday* ... *Sunday*

DEMONSTRATION

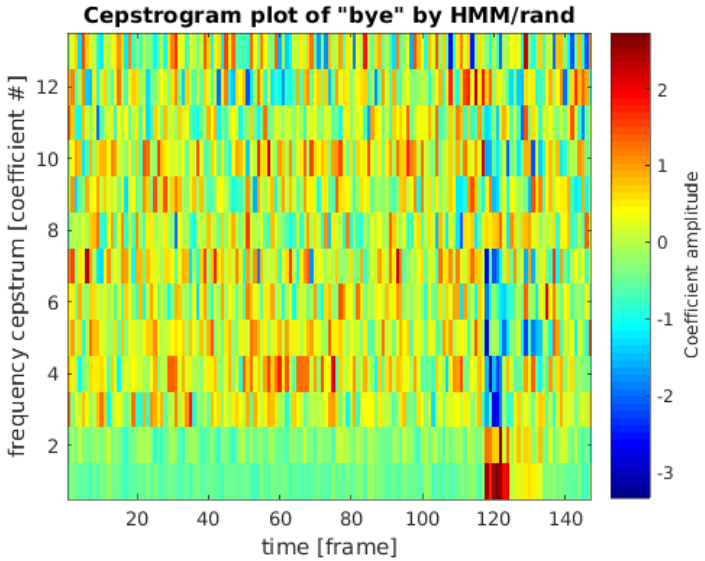
Data & Feature Extraction

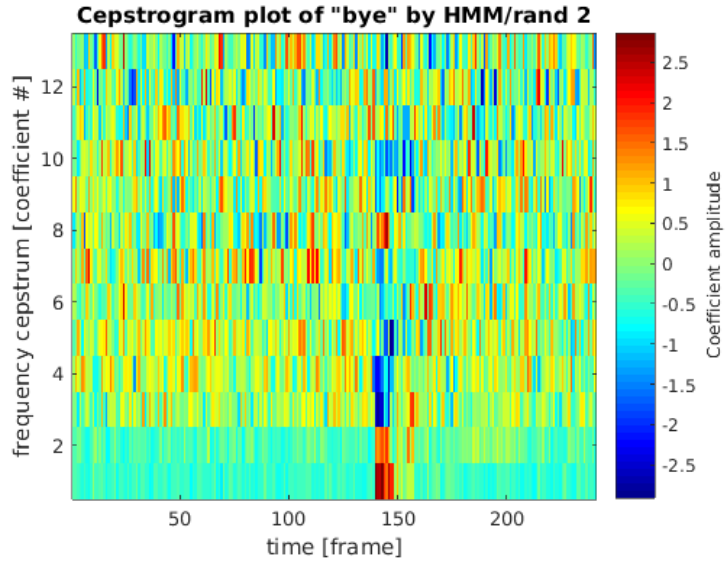
- Data characteristics
  - Audio data, 22,050 Hz bitrate, 16 bit, 1 channel
  - Speakers: 4 (2 Male, 2 Female)
  - Audio clips: 14 clips per speaker per word
  - Duration of clips: 2 sec
  - Window size: 0.02 sec
- Speech characteristics captured in features
  - Phonemes and Formants for recognizing the word
  - Rate of change of characteristics
  - Speaker independent recognition
- Speech data features used:
  - Mel-Frequency Cepstral coefficients (MFCC)
  - Delta of MFCC
  - Delta-Delta of MFCC
- Continuous scalar data
- Features are robust against
  - Time duration of the word
  - Pitch of different speakers (Male/Female)

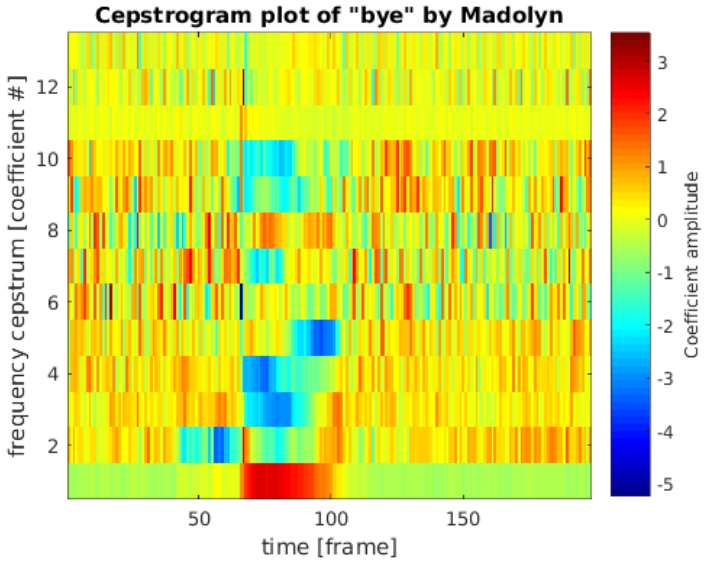
HMM design parameters

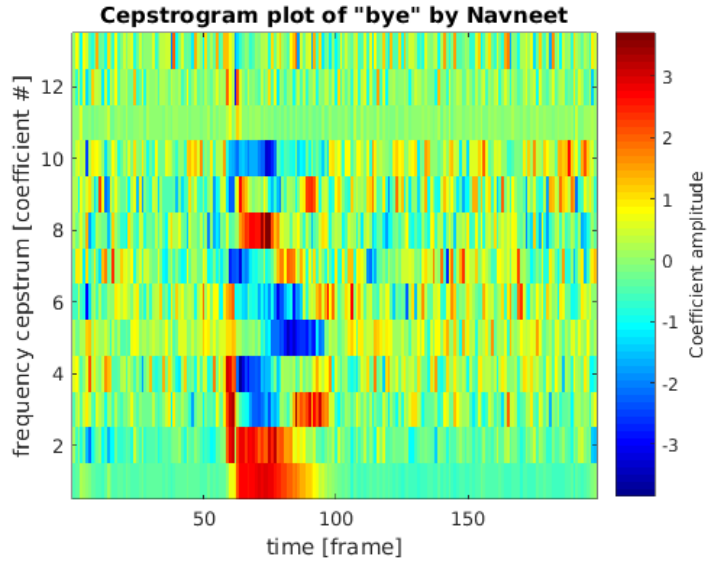
- Finite Left-Right HMM
- Size of training set: 9, Test set: 5
- Number of States: $2 \times (\text{Length of word}) + 2$
  - Each syllable in a word represented by 2 states
  - Noisy Silence represented 2 states
- Cepstral Coefficients used: 10
  - Higher frequency coefficients creates confusion
  - Most important speech data in lower coefficients
- Modeling of Observed Data using *GaussD* & *GaussMixD*
  - *GaussD* : Single speaker
  - *GaussMixD* : Multiple speakers (2 or 3 Gaussians)

Trained HMM sequences: Similarity and Differences









Classification Error

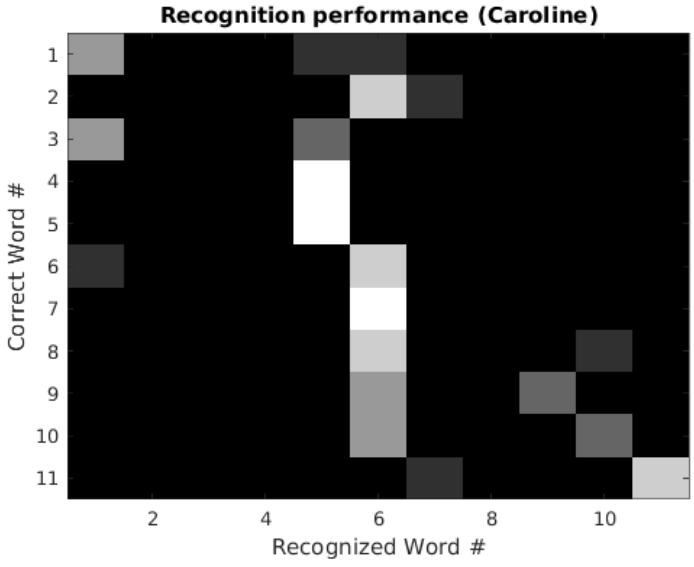
Table: Classification error of each speaker in different models

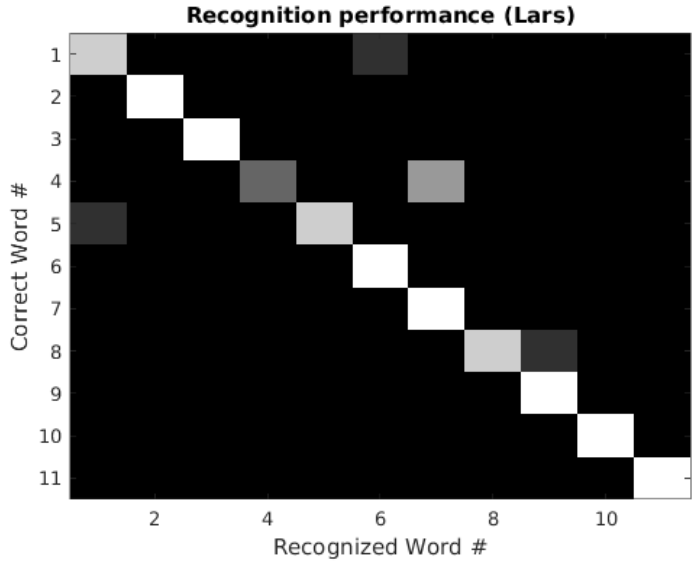
Model → Speaker ↓	**Single GaussD**	**All GaussD**	**GaussMixD 2Mix**	**GaussMixD 3Mix**
Caroline	32.73%	27.27%	21.82%	10.91%
Lars	10.91%	32.73%	32.73%	21.82%
Madolyn	38.18%	54.55%	60.00%	40.00%
Navneet	30.91%	49.09%	54.55%	29.09%
Total	-	40.91%	42.27%	25.45%

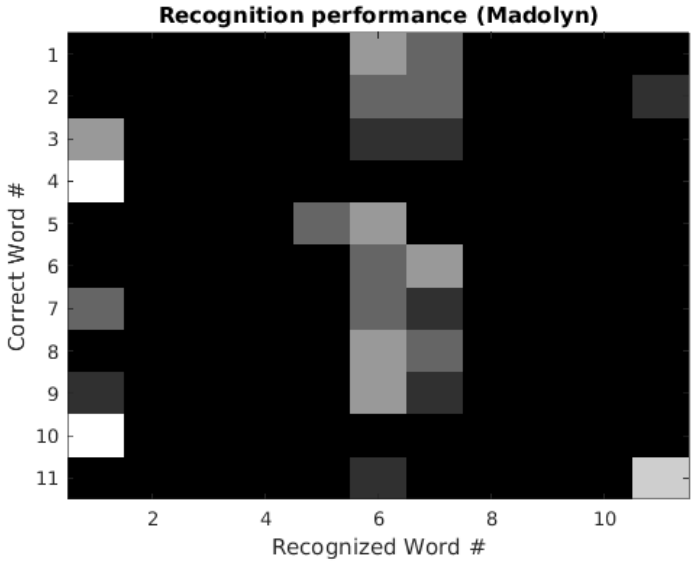
For the best performance model i.e. GMM (mix-3) model:

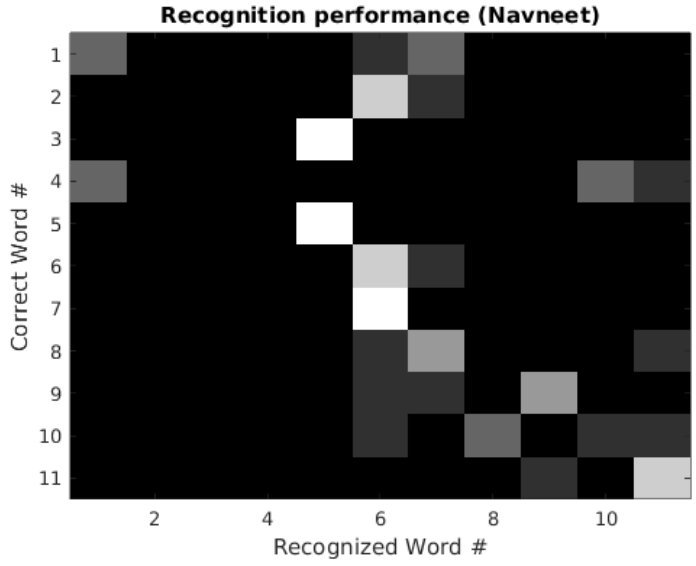
- Word Friday gives maximum error (19.64%) in classification
- Word Tuesday gives minimum error (0%)

Classification Error (Single Speaker GaussD)

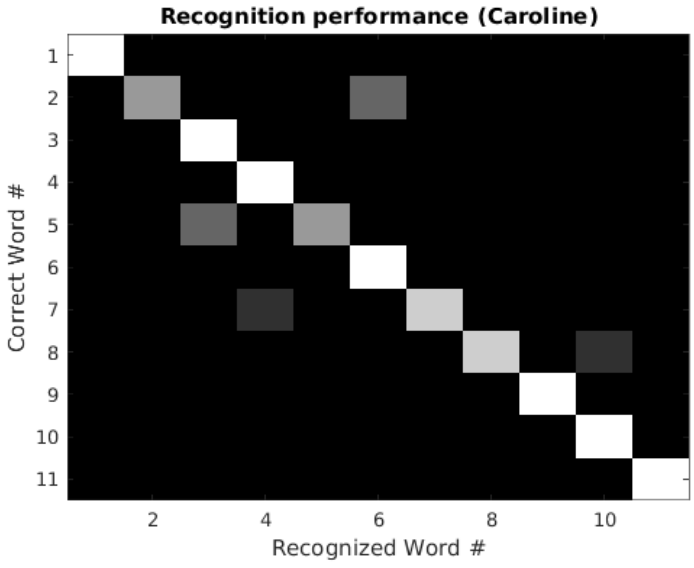


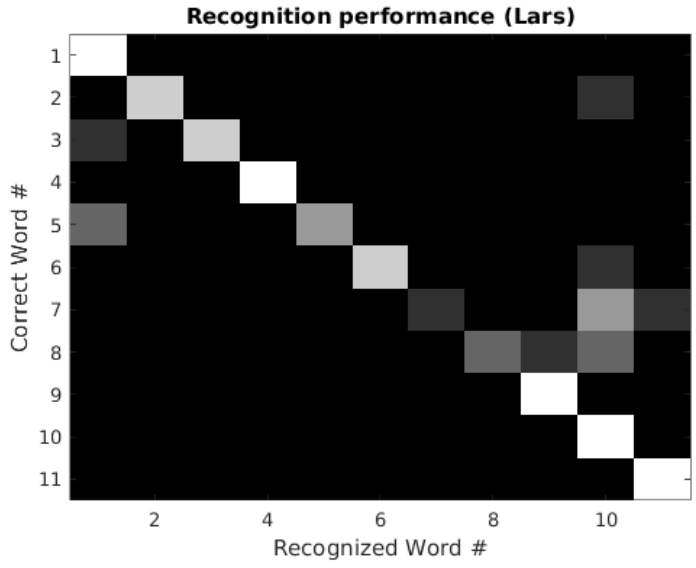


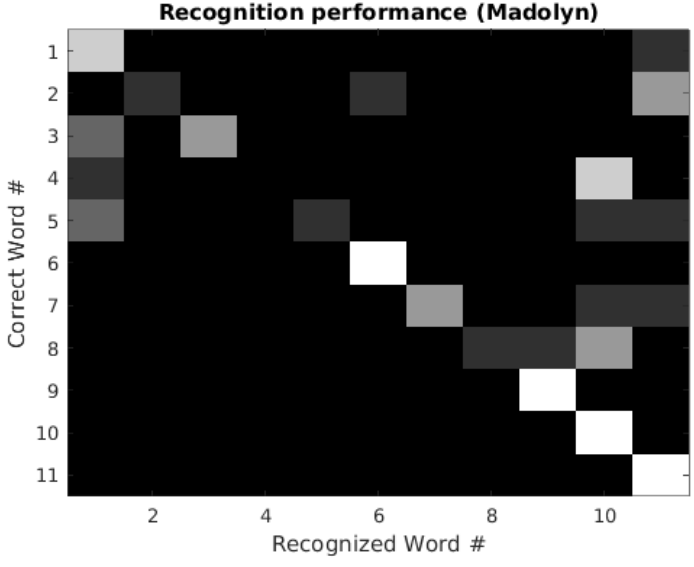


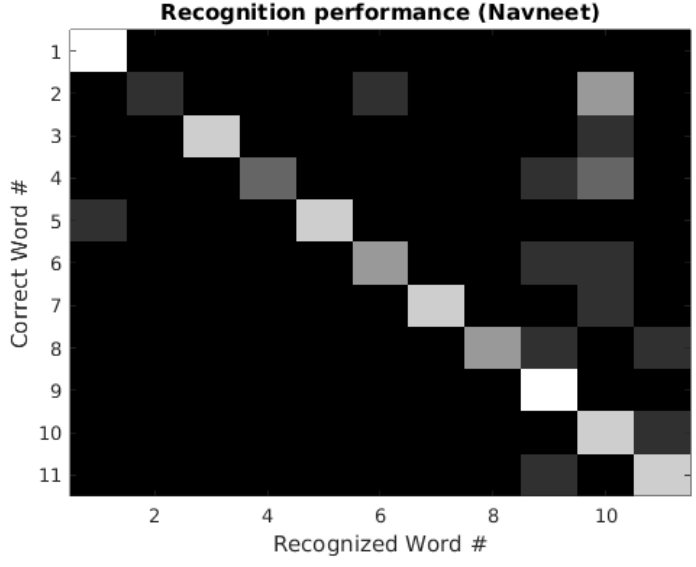


Classification Error (GMM mix:3)









Conclusion

- Works pretty well for single speaker
- GaussMixD model is better for multiple speakers support
- Some words are difficult to recognize than others
- System works fine with limited number of words
- Performance depends on various factors including amount of training data, complexity of the words and number of different speakers used for training
- Learned to implement HMM based classification scheme
- Got acquainted with features used in speech recognition
- Hands-on experience with designing HMM model parameters to fit data