Multimedia Class 2016 TA Class

Slide 3:

Introduction to Speech Processing using MATLAB

Preface, what do we learn?

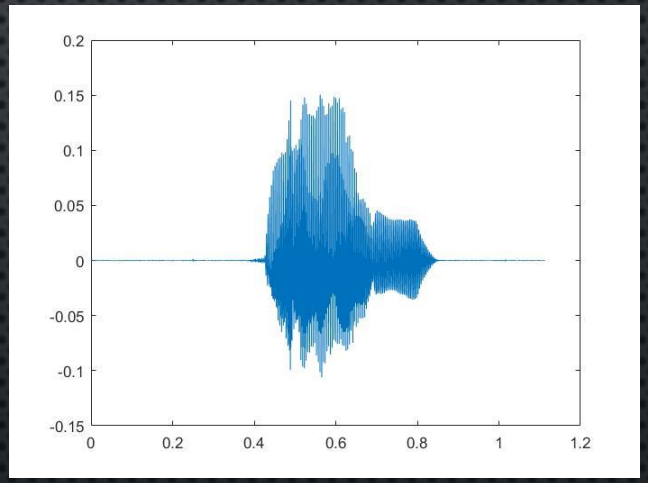
- In this chapter we will learn a bit more about speech processing in MATLAB
- You know the basics from your course
- Some the things we will learn in this session:
  - All about MFCC, what it is and how it can be used
  - A bit more about advanced Automatic Speech Recognition (ASR) and Robust Speech Processing (RSP)
  - Take it down a notch: A simple 2-class speech recognition + implementational details (guess what, new homework is up!)
  - A bit more about GMMs (Gaussian Mixture Models), and how we can use it

What do we want to achieve?

- We should be able to interact with computer using only our beautiful voices
- Simplest form: we should be able to answer the computer with Yes/No
  - Like: Would you like me to order you pizza for lunch, Yes, OK order has been places
- How can computer understand what we just said?
  - Just like a baby, using probabilistic frameworks
  - Show it a couple hundreds of examples, it will find the common denominator
  - At least it learn faster than you! (how old were you when you learned to understand what other people said?)

What is audio signal in MATLAB?

- What happens when we read an audio file in MATLAB?
  - They become and array (1-D for mono and 2-D for stereo) of Doubles
  - Each number is relative to sound pressure of microphone at the time of recording
  - We can plot it easily, but it is a lot of data
  - You have to perform recognition based on an 8908-D feature dimension
    - Huge amount of data for probabilistic approaches




```
>> whos y
  Name          Size                Bytes  Class     Attributes

  y          8908x1                 71264  double
```

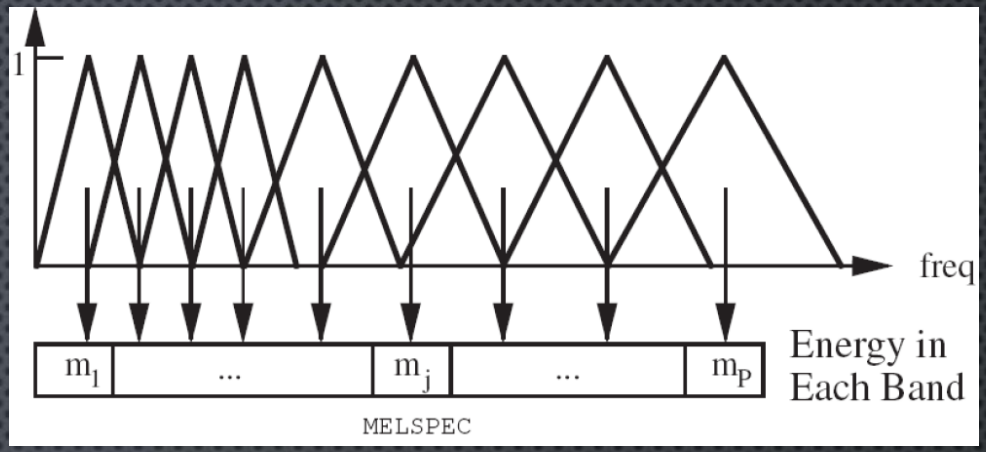
Feature Extraction (FE)

- OK, so we said that 8K+ dimension array with variable length is really hard to perform recognition on
- So, how does Google do millions of millions of them on a daily basis?
- Feature Extraction is the ART of extracting what really matters from all the data you have
- It was done until recently based on hand-crafted features that made sense
- Lately by the evolution of CNNs, it is designed automatically for you (it is called end-to-end learning, just google it)

FE, continued

- Here comes the FE we are going to use: MFCC
- It stands for Mel-Frequency Cepstral Coefficients
  - Wait a minute, Mel sound familiar, where have we heard it?
  - You have heard all about it from Dr. Seyedin, in Chapter3: Speech Processing
  - In case you need to remember it again, take a look at you notes or handbooks

MFCC



- What is it?
  - Well it says how much energy is there is there in each given frequency band
  - Bands are derived from human perception of audio
  - In case you forgot the formula for bands, get back to Dr. Seyedin presentations
- How it is done (performed on each signal window) [Wikipedia]:

  1- Take the fourier of the windowed signal

  2- use triangular overlapping windows to map the frequencies to Mel-Scale

  3- Take the logs of powers

  4- take DCT of the logs of powers (don't worry, you will read all about it in DSP)

  5- MFCC is the amplitude of the resulting spectrum
- Wait a minute, you said power??? What about the Phase???
  - Don't worry, as I said before phase is not very important in Speech Processing

MFCC, continued

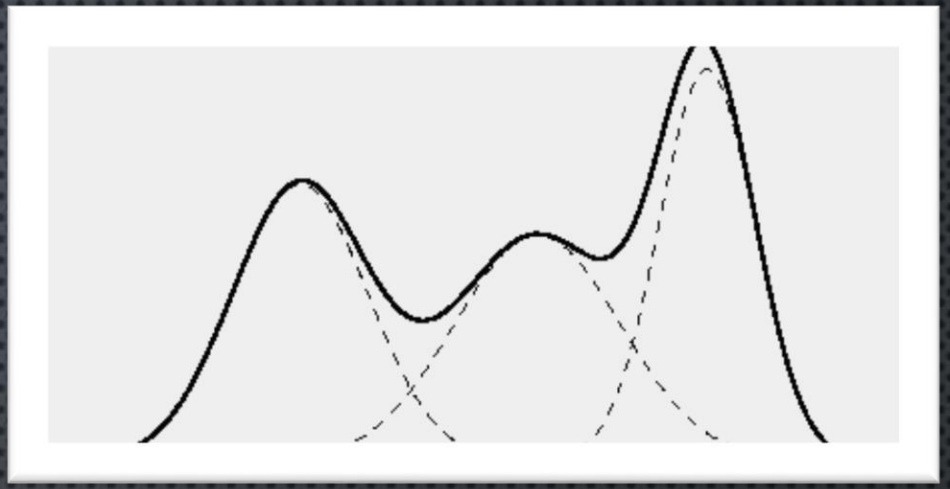
- OK, I don't really understand, you said simple!
  - It is OK, that's where libraries come useful
  - Rastamat is a really useful library made for speech processing in MATLAB
    - Check the end of this tutorial for links
  - We will let Rastamat do the heavy lifting for us, we will just use its results
- So what do you need to do?
  - Be patient, GMM is up

MFCC IN MATLAB

- Here is the code:
- [mm,~] = melfcc(y, Fs, 'maxfreq', Fs/2, 'numcep', 13 ...
- , 'nbands', 13, 'fbtype', 'fcmel', 'dcttype', 1, ...
- 'usecmp', 1, 'wintime', 0.025, 'hoptime', 0.010, ...
- 'preemph', 0, 'dither', 1);

- Now you can ask about the parameters…

Next up, GMM



- So what is GMM?
  - Gaussian Mixture Model is a way to easily model a PDF, by saying it is the addition of some gaussians
  - Look at the picture above, solid line is the original PDF, and dotted lines are the gaussians added to build it
  - It is very useful in Audio and Speech Processing, since it models the data well
  - A complete PDF resolves to some means and variances, so it becomes much easier to keep track of thousand of model for different words
- I added a tutorial in PDF to your Git, take a look at it

Complete Simple Speech Recognition, Training

1. Segment each audio file into windows
2. Perform MFCC feature extraction on each window
    1. In this section you have lots of lots of 13-D arrays for each word model
3. Fit a GMM model (3-4 Mixtures) to each class
4. Save your models and Done!

Complete Simple Speech Recognition, Testing

Perform the Following for each audio file:

1. Perform steps 1 and 2 in training to get the MFCCs
2. get the likelihood of belonging to each class for each window (value in [0-1])
3. Take the log of likelihoods
4. Sum them up for each class (results in 2 numbers, 1 for class "Yes" and another for class "No")
5. The word with the highest likelihood is your estimation

Finishing

- Do what you learned for all the audio files in the testing sequence, report your recognition rate
- Tune the parameters to get better results
- Try to write a fast code
- Optional Features:
  - VAD: Voice Activity Detector (remove noise at the beginning and end of each file)
  - Feature Normalization (CMVN: Cepstral Mean and Variance Normalization)
  - Add derivatives of features to FE.m and make your results get better
- OK, let's see what we've got in MATLAB

Acknowledgements

- Library for feature extraction (MFCC and all):
  - Rastamat Library by Daniel Ellis
    - Available at: http://www.ee.columbia.edu/ln/rosa/matlab/rastamat/
  - HMM Toolkit by Kevin Murphy
    - Available at: http://www.ai.mit.edu/~murphyk/Software/HMM.html
- Some of the audio files belong the AURA dataset, they have been given to me by SPRL (Signal and speech Processing Research Lab) under supervision of Dr. Ahadi and Dr. Seyedin
- Other stuff (uses of the following library) have been written by me a couple of years age