Mobile Sensing: Voice Sensing

Partly based on: "Speech Recognition — GMM, HMM" by Jonathan Hui and "CS 224S / LINGUIST 285 Spoken Language Processing" by Andrew Maas

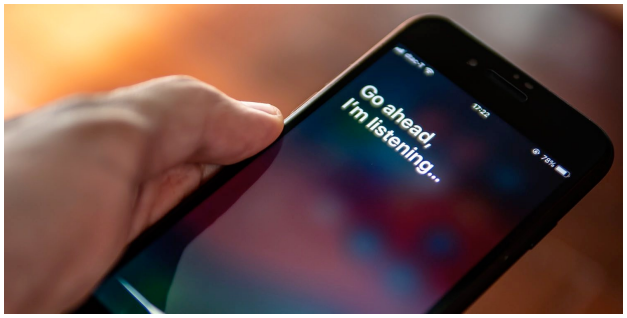
Master studies 2021/2022

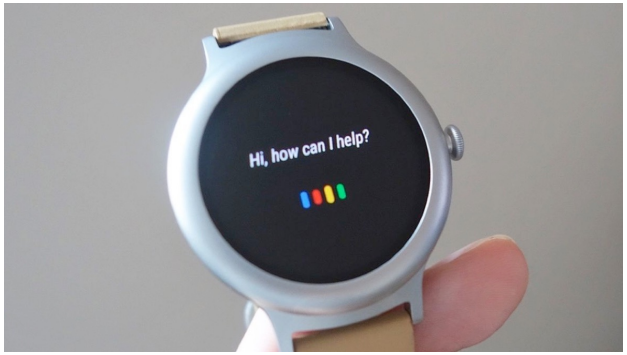
Dr Veljko Pejović
Veljko.Pejovic@fri.uni-lj.si

University *of Ljubljana*
Faculty *of Computer and Information Science*

Sound Sensing Opportunities

- Microphones included with
  - Smarpthones
  - Wearables
  - Home IoT devices
- Microphone array (multiple mics on a device)
  - Noise cancellation – the one close to the speaker records sound that is subtracted
  - Beamforming – signal processing allows you to “focus” mics in a particular direction





Sound Sensing Applications

- Environment sensing and citizen science
  - Noise pollution monitoring
  - Bird sound recording
- Speech recognition
  - Device interaction, dictation
  - Home automation system

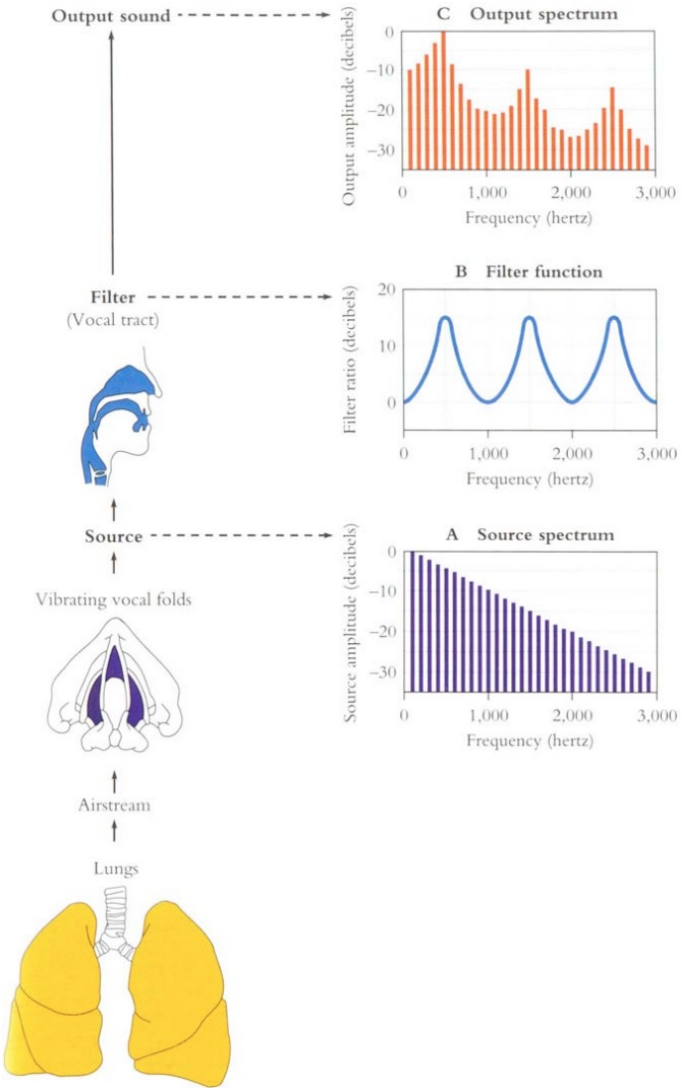
Sound Sensing Applications

- Speaker recognition
  - Identification – who is talking?
  - Verification – is it user X that is talking?
  - Passphrase verification – was “X” said?
- Voice property analysis
  - Identify speaker demographics
    - Gender
    - Age
  - Health-related voice analysis:
    - Smoker
    - Parkinson’s disease
    - Stress

Can be used for continuous authentication

Speech Processing

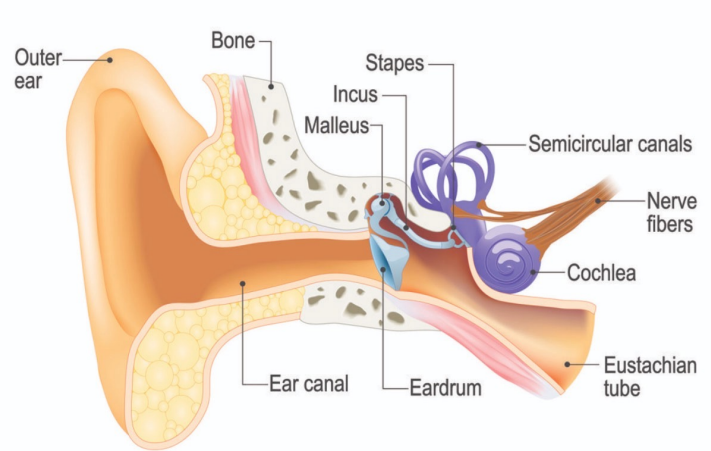
- Speech creation:
  - Lungs push air through the trachea
  - Vocal folds open and close creating vibrations (basic vibration 125 Hz male, 210 Hz female voice)
  - Vocal tract shapes the vibrating sound to produce different phonemes (composed of phones)

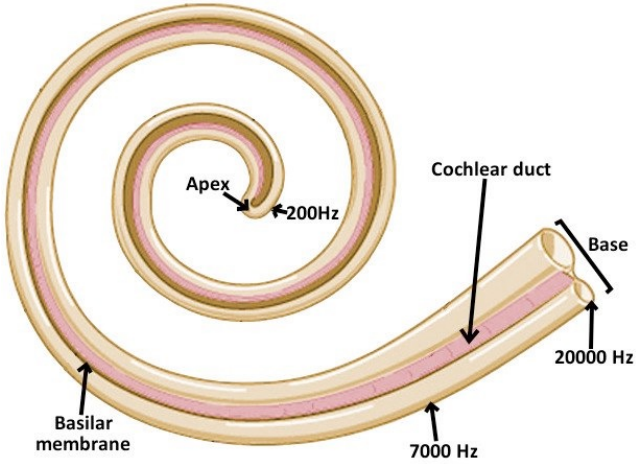


"Speech Recognition — Phonetics" by Jonathan Hui

Speech Processing

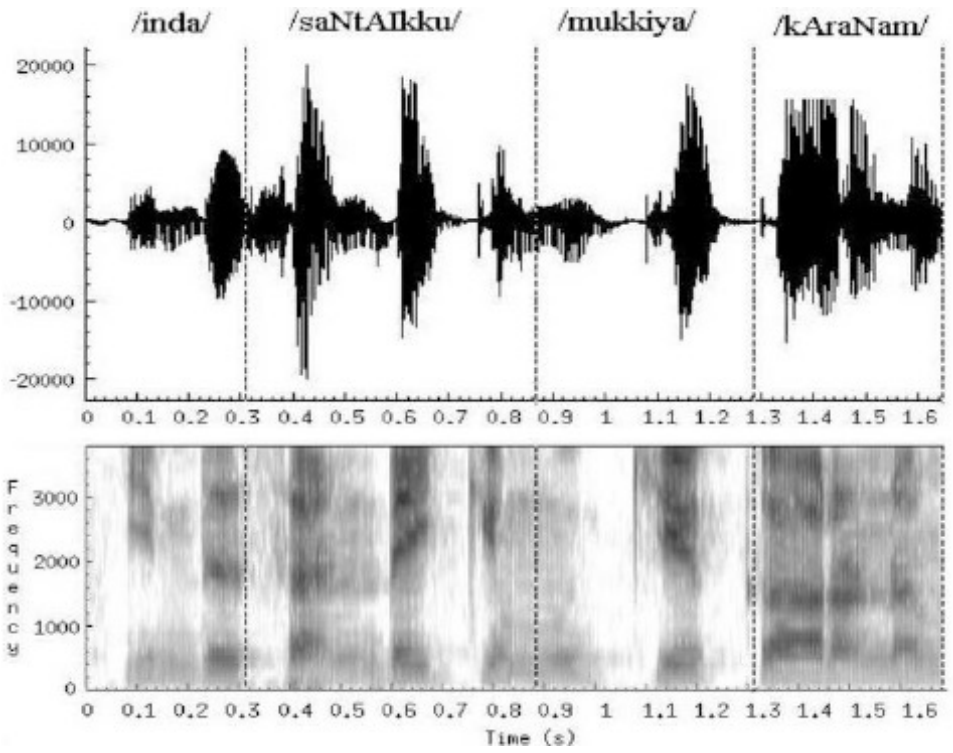
- Speech detection:
  - Eardrum transports vibrations via the small three bones to cochlea
  - Cochlea contains fluid that vibrates when excited by the eardrum vibration
  - Hair cells within cochlea send electrical impulses via the hearing nerve, yet different hair cells react to different fluid vibration frequencies





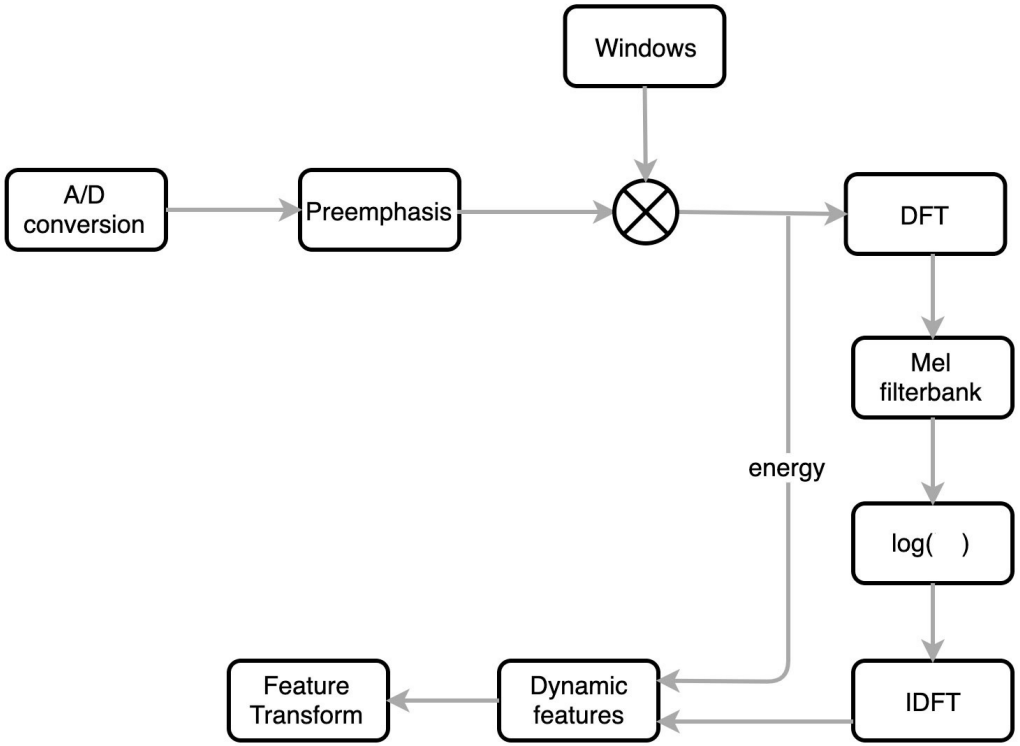
Speech Processing

- Speech is composed of phonemes, which last for about 30 – 200 ms
- Different frequencies are prominent with each phoneme/phone
- We want to pinpoint these frequencies!



Speech Features – MFCC

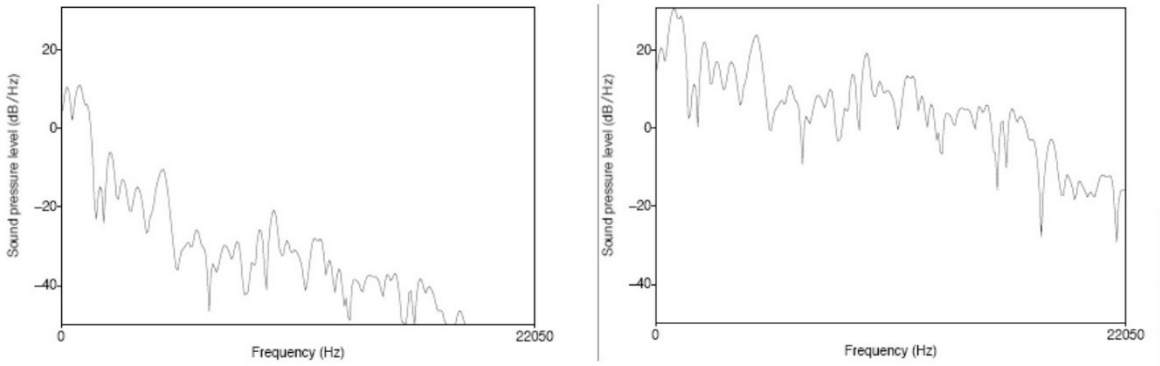
- Mel-frequency cepstral coefficients (MFCCs)



Speech Features – MFCC

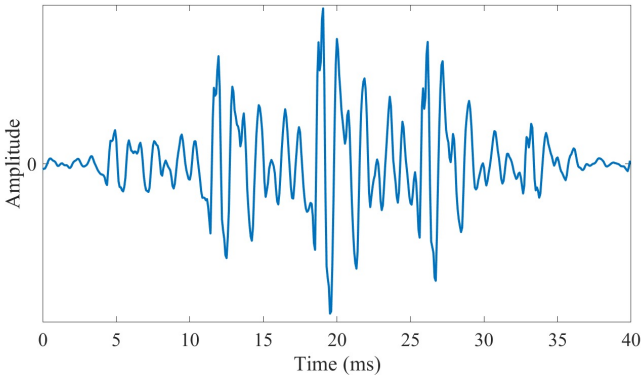
- Sampling and framing
  - Usually 16 kHz sampling rate, 16b samples
  - Frames of 20 ms to 40 ms (compare to phoneme)
  - Offset 10 ms between successive frames
- Preemphasis
  - Boost high frequencies, as these are pronounced with a lower intensity

Performed as a time domain filter



Speech Features – MFCC

- Windowing
  - Hamming window to avoid sharp amplitude changes
- Discrete Fourier Transform (DFT)
  - To obtain a frequency view of the signal
  - Calculate the intensity at different frequencies as cochlea's hair cells capture vibrations at certain frequencies



FFT from Sampling and Filtering lecture

Speech Features – MFCC

- Mel filter bank
  - Cochlea cannot discern the difference between two closely spaced frequencies
  - This effect becomes more pronounced as the frequencies increase

Cat filter bank

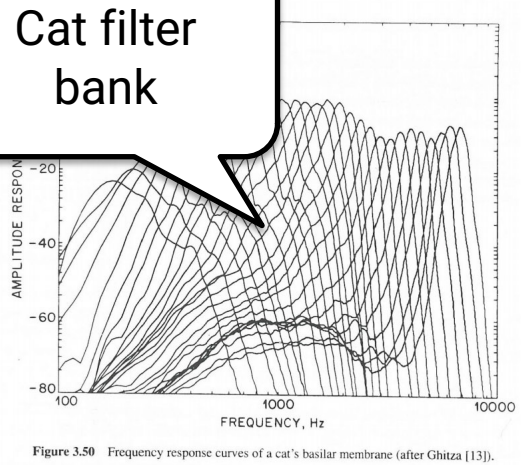
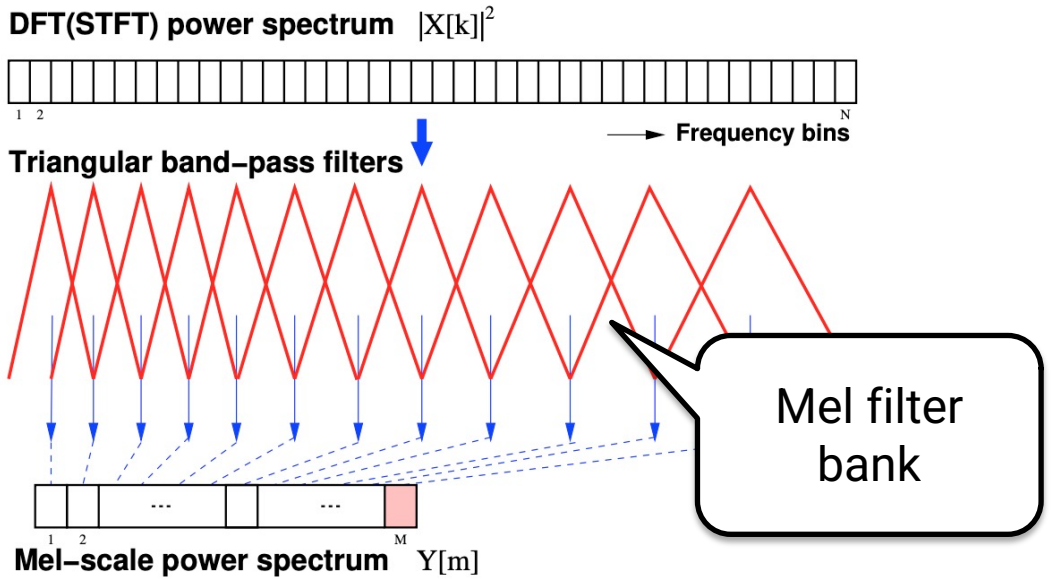


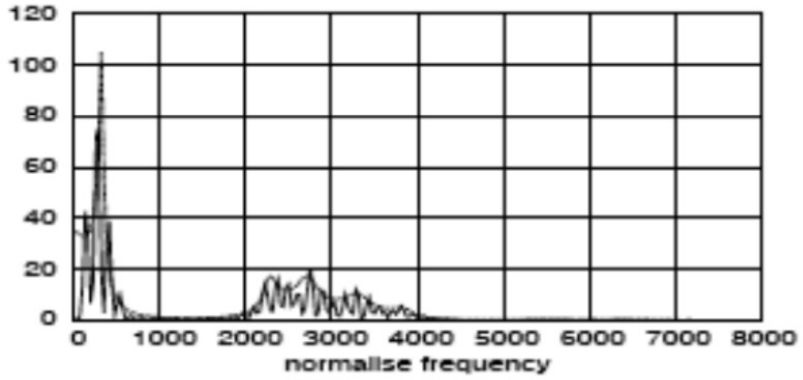
Figure 3.50 Frequency response curves of a cat's basilar membrane (after Ghitza [13]).

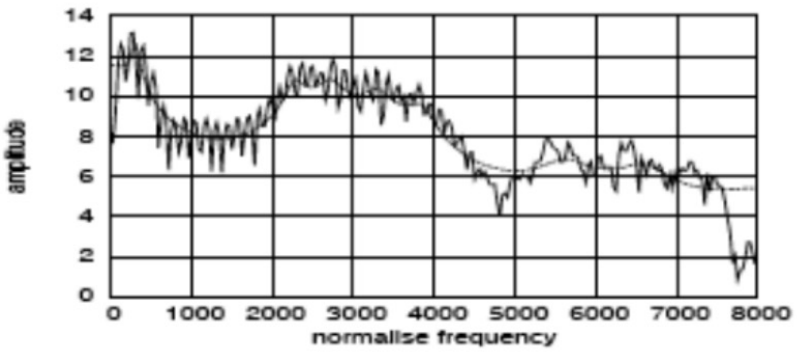


Mel filter bank

Speech Features – MFCC

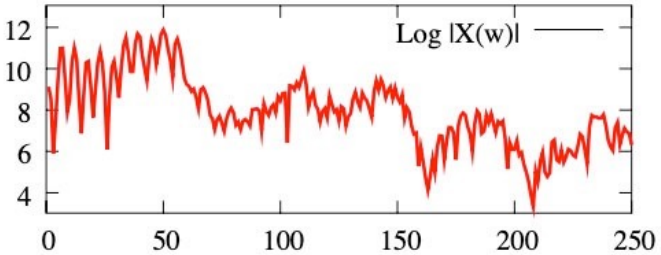
- Log scaling
  - We are less sensitive to small change at high energy than small changes at a low energy level
  - Log scaling levels it out



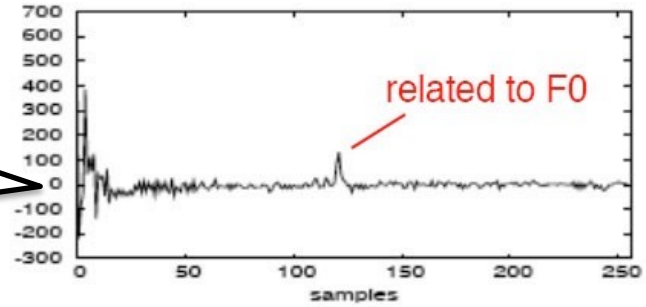


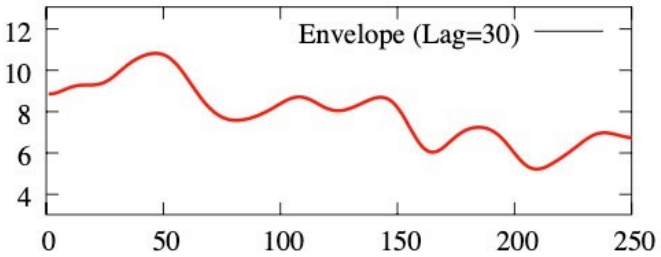
Speech Features – MFCC

- Inverse Discrete Fourier Transform (IDFT)
  - The spectrum picture we have contains info about the phoneme and about the pitch (F0)
  - IDFT, in this case discrete cosine transform (DCT), lets us single out phoneme-related information



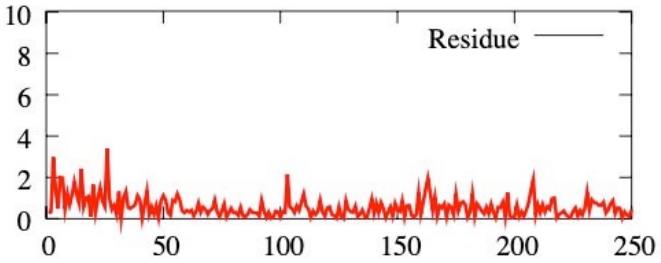
Cepstrum coefficient (keep only the first 13 or so)

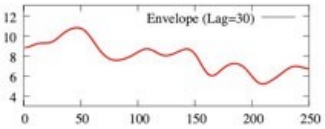




(after IDFT)

phone informations





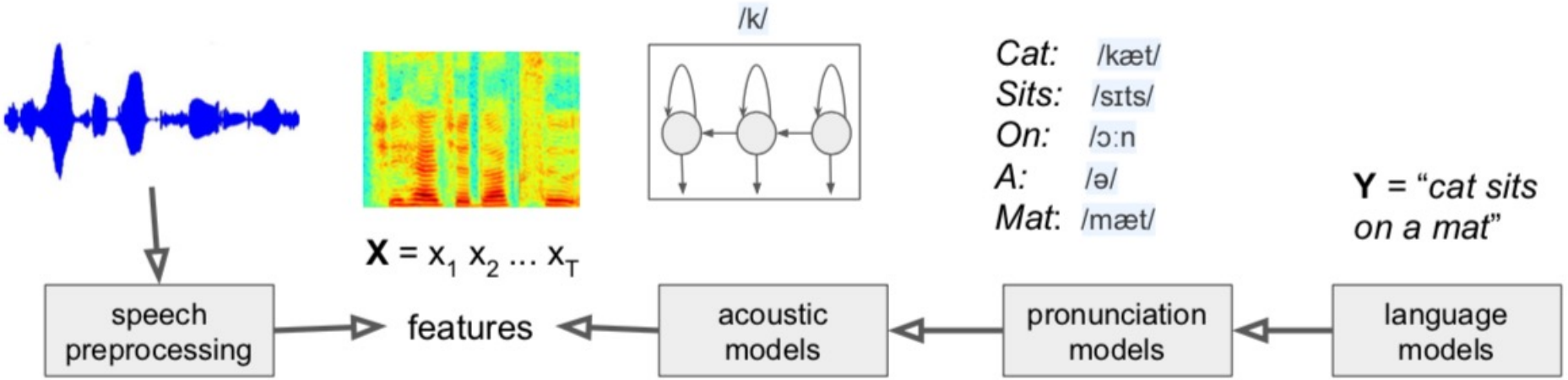
Speech Features – MFCC

- Dynamic MFCC-related features
  - Besides the original MFCC features extract
    - First derivative of the features
    - Second derivative of the features
- Final feature vector contains about 39 real numbers per each 20ms-40ms window
- How do we go from these features to speech recognition?

Speech Recognition Models

- Hidden Markov Model (HMM) with Gausian Mixture Models (GMM)
  - Based on the MFCC observations explains how the underlying phonemes are transitioned
  - The existing model is easilly re-trained for a particular person, emotional expression, etc.
- Deep neural networks
  - No need for feature (MFCC) extraction, although it might help
  - Nowadays better performance than HMM-GMM
  - Might be difficult to re-train

HMM with GMM



$$W^* = \arg\max_W P(W|X)$$

$$W^* = \arg\max_W P(X|W)\ P(W) / P(X) \quad \text{Apply Bayes' Theorem}$$

$$W^* = \arg\max_W P(X|W)\ P(W) \quad \text{Remove } P(X) \text{ because } X \text{ does not change w.r.t. } W$$

word sequence — acoustic model — language model

HMM+GMM

Hidden Markov Model (HMM)

- A Markov chain where the states are not directly observable
  - However, some other variables are observable, and their values depend on the underlying state
- Transition probability: the probability of transiting from one internal state to another
- Emission probability: the probability of observing an observable given an internal state

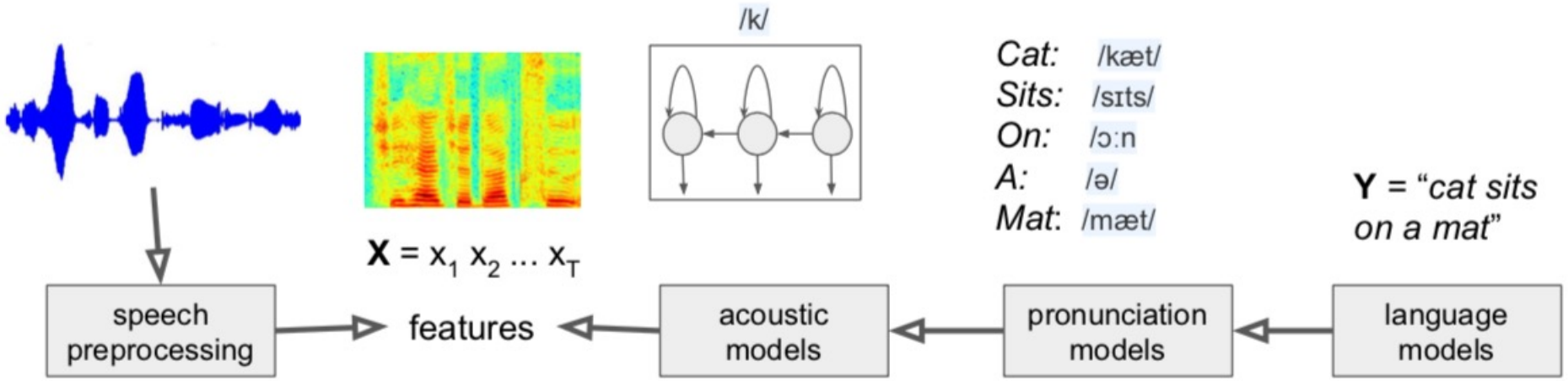
Hidden Markov Model (HMM)

- In speech recognition
  - Observables are audio properties (MFCC)
  - States are phonemes
- We aim to infer the sequence of phonemes, i.e. words according to:
  - The probability of the word
  - The probability of the sequence of observations reflecting the sequence of phonemes representing that word
- Training HMM and inferring on it:
  - Baum–Welch and Viterbi algorithms

Gaussian Mixture Model (GMM)

- GMM is a weighted combination of multiple Gaussian distributions
  - GMM that is built upon uncorrelated dimensions is computationally light
- Acoustic model – how is a phoneme represented in terms of the MFCC features
  - GMM is a good representation of speech pronunciation
  - MFCC features as dimensions
  - Multiple components
  - One GMM per each phoneme

HMM with GMM – Full Pipeline



- Extract MFCC features and train GMM
- Add the language model to train HMM
- Observe a sequence of observations (MFCC features)
- Use Viterbi to find the most likely sequence of phonemes, i.e. the most likely word

Deep Neural Networks

- Different parts of the pipeline can be adapted to neural network processing
  - Language model with a neural network
  - Acoustic model with a DNN+HMM or LSTM+HMM
  - Features extracted using neural network approaches
- Furthermore, a full end-to-end inference can be performed with deep learning
  - Example in the labs this week!

Beyond Speech Recognition

- Mental health-related applications might require additional speech features to be extracted:
  - Rate of speech [slow, rapid]
  - Flow of speech [hesitant, long pauses, stuttering]
  - Intensity of speech [loud, soft]
  - Clarity [clear, slurred]
  - Liveliness [pressured, monotonous, explosive]
  - Quality [verbose, scant]
- Example application – EmotionSense

The “other” paper for this week