On speech problems, tasks and challenges

Islam Faisal (Only Presenting)

Agenda

- Speech Signal and Speech Processing
  - Spectrograms
  - Source Separation
- Common Speech Tasks
  - Automatic Speech Recognition
  - Speaker Recognition, Verification, and Diarization
  - Language Identification
  - Translation
- Models Used
- Features

How do we speak?

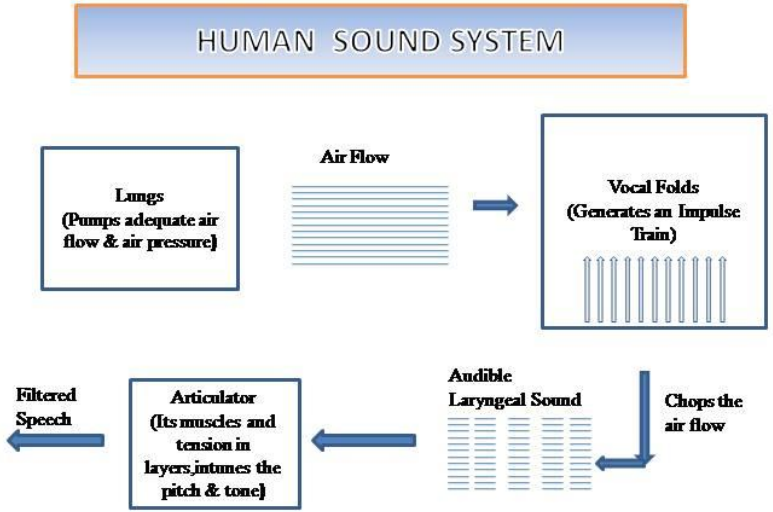


Image Credit: Ananya Panja: https://www.projectrhea.org/rhea/index.php/Male_vs._Female_Voice_characterestics

Memories from your DSP class

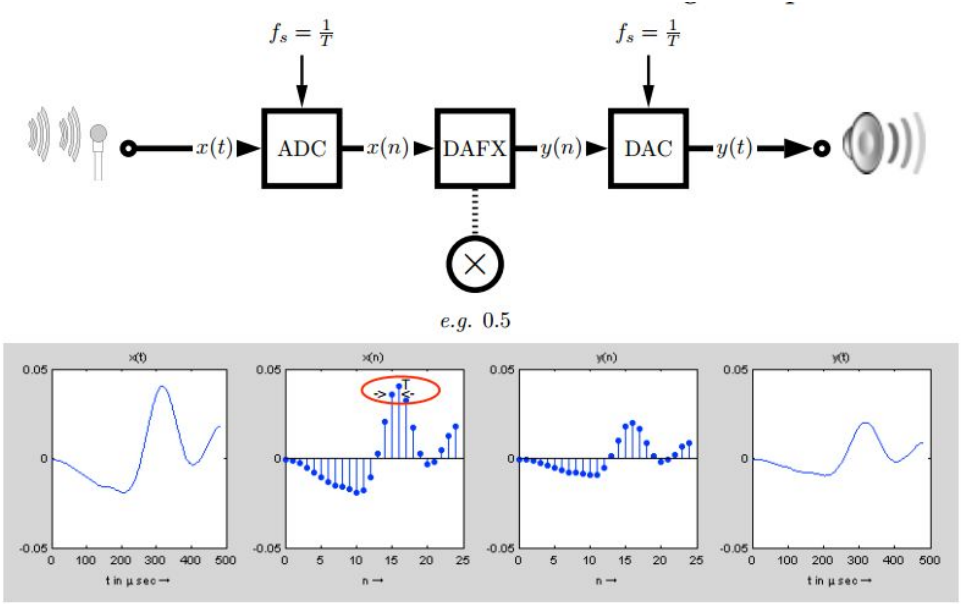
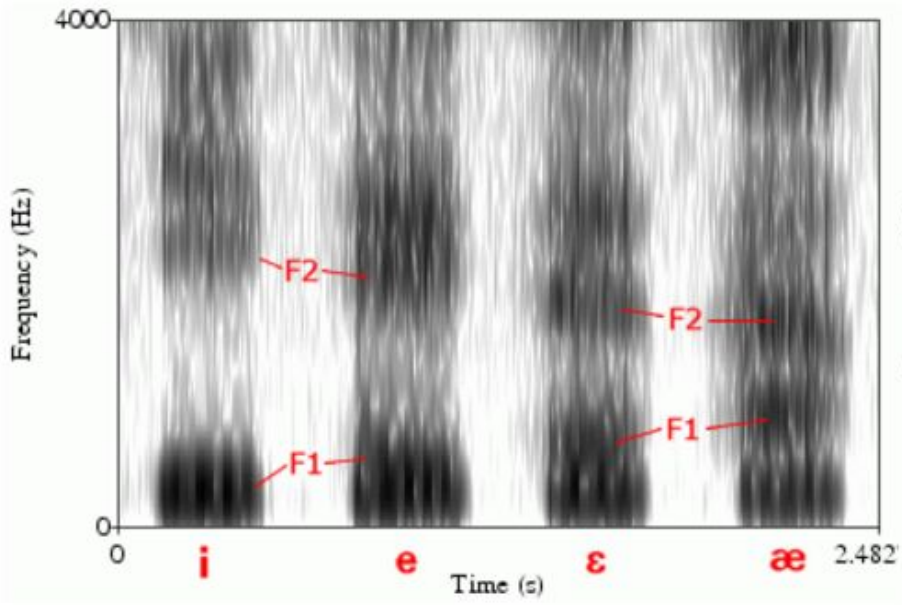


Image Credit: MATLAB, DSP, Graphics Module No: CM0268. Prof David Marshall, Dr Yukun Lai, Cardiff School of Computer Science

Spectrograms (voiceprints)



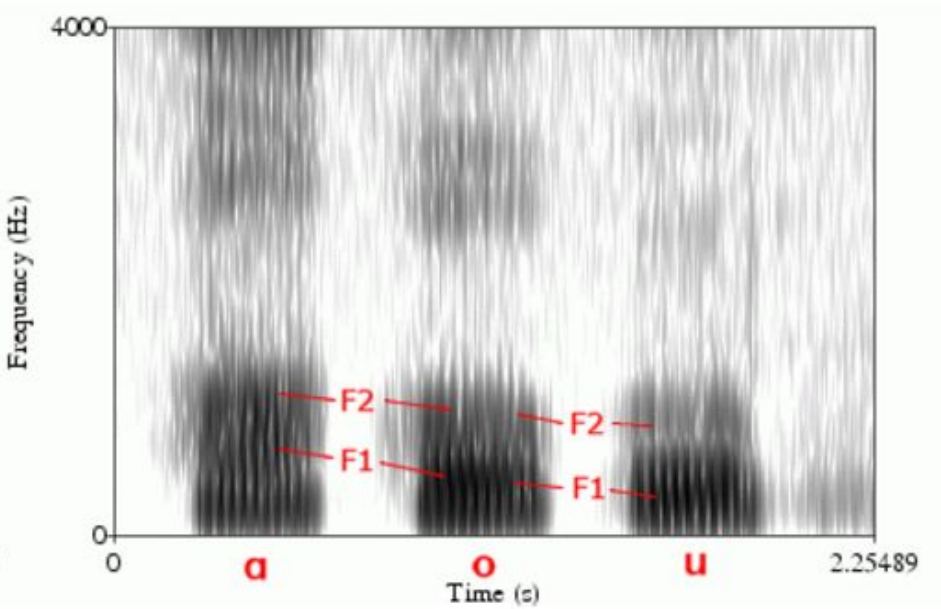


Image Credit: Identifying sounds in spectrograms, Kevin Russell.

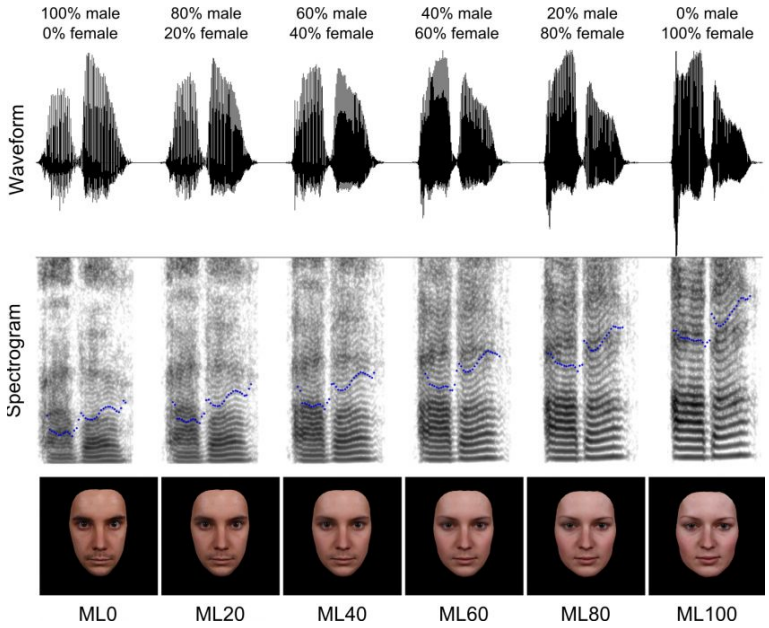


Image Credit: Stefan R. Schweinberger and Verena G. Skuk - Voice Perception: Basic Parameters (2012-2015 Project)

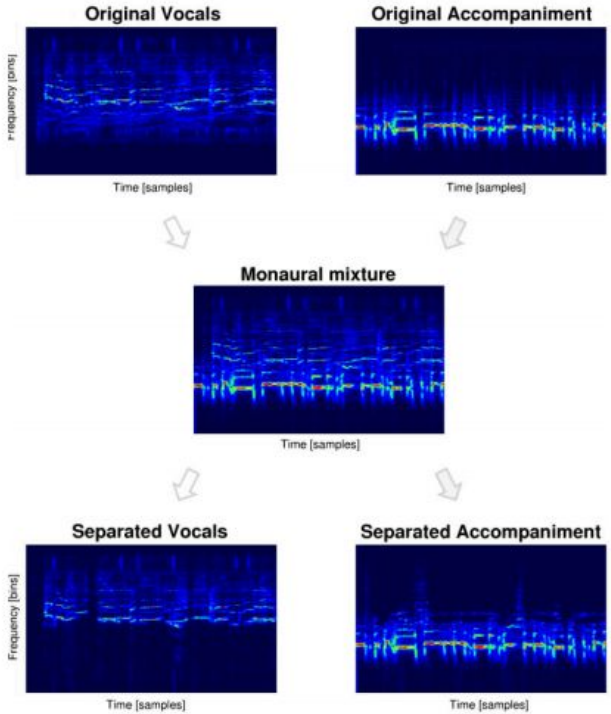
Waveform and spectrogram along a male-female voice morph continuum of the utterance /aba/, and face-morphs along a male-female morph continuum. All continua are morphed from male to female in steps of 20% (morph level).

Source Separation

- Multiple Superimposed Audio Sources
- Need to separate sources or eliminate Noise
- Independent Component Analysis, Singular Value Decomposition and Machine Learning

http://research.ics.aalto.fi/ica/cocktail/cocktail_en.cgi

Source Separation using Neural Networks



Separation of vocal sounds from musical mixtures using a probabilistic convolutional deep neural network. The upper pair of spectrograms plot a ~1.5-second excerpt from a typical song [..] illustrating the original monaural audio for the voice and non-voice (i.e., accompaniment) sources respectively. The middle spectrogram plots the monaural mixture (i.e., the ensemble music). The lower pair of spectrograms plot the respective separated channels (α = 0.5). Note the frequency axis represents the range 0 – 22 kHz on a logarithmic axis.

Simpson, A. J., Roma, G., & Plumbley, M. D. (2015, August). Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network. In International Conference on Latent Variable Analysis and Signal Separation (pp. 429-436). Springer International Publishing.

Automatic Speech Recognition (ASR)

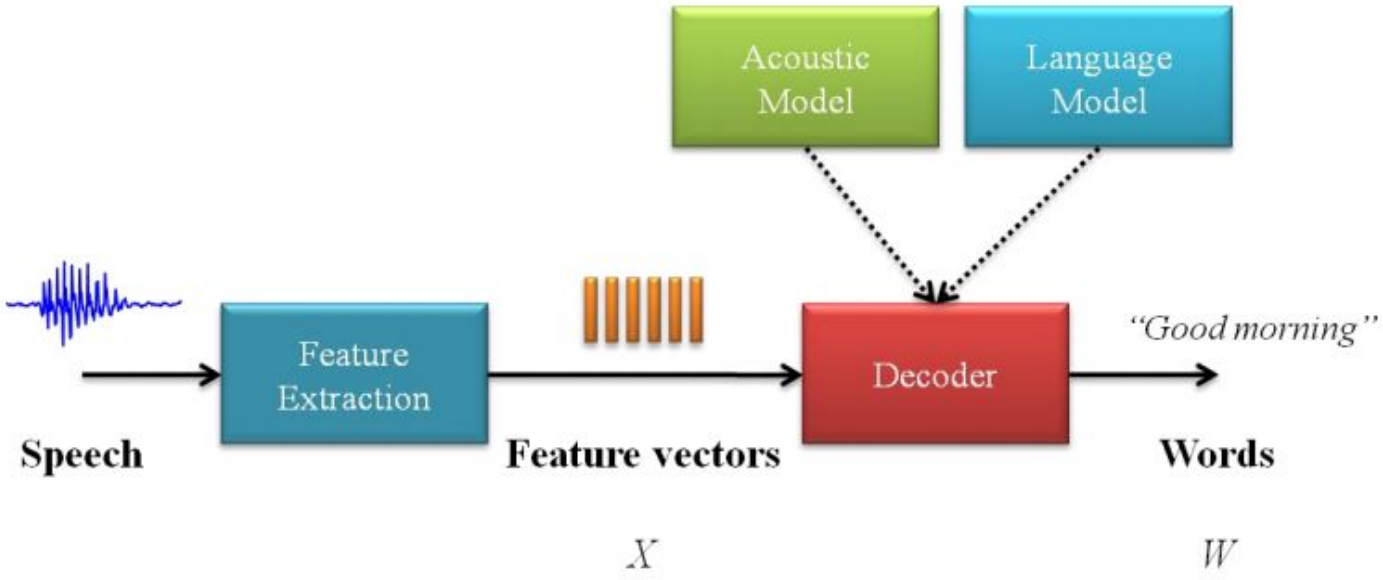


Fig. 3. Architecture of an ASR system.

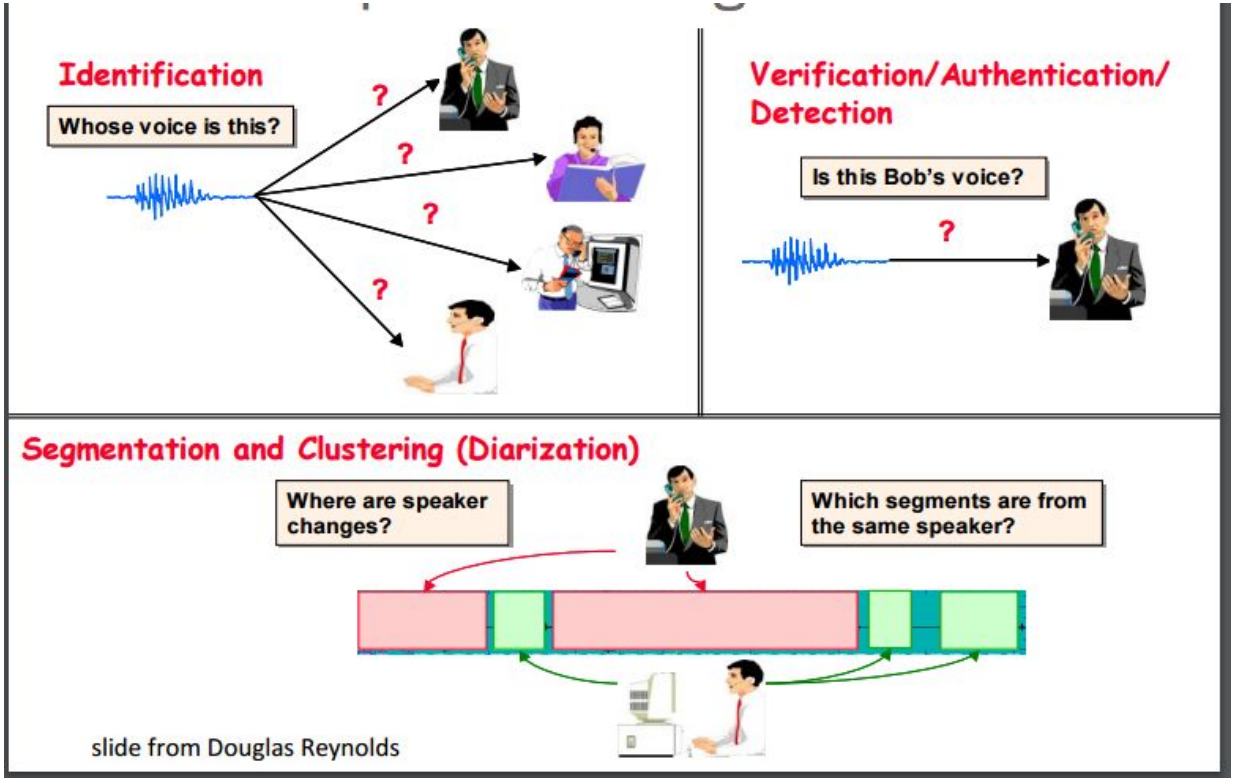
Image Credit: Automatic Speech Recognition: From the Beginning to the Portuguese Language by André Gustavo Adami

Features (More coming)

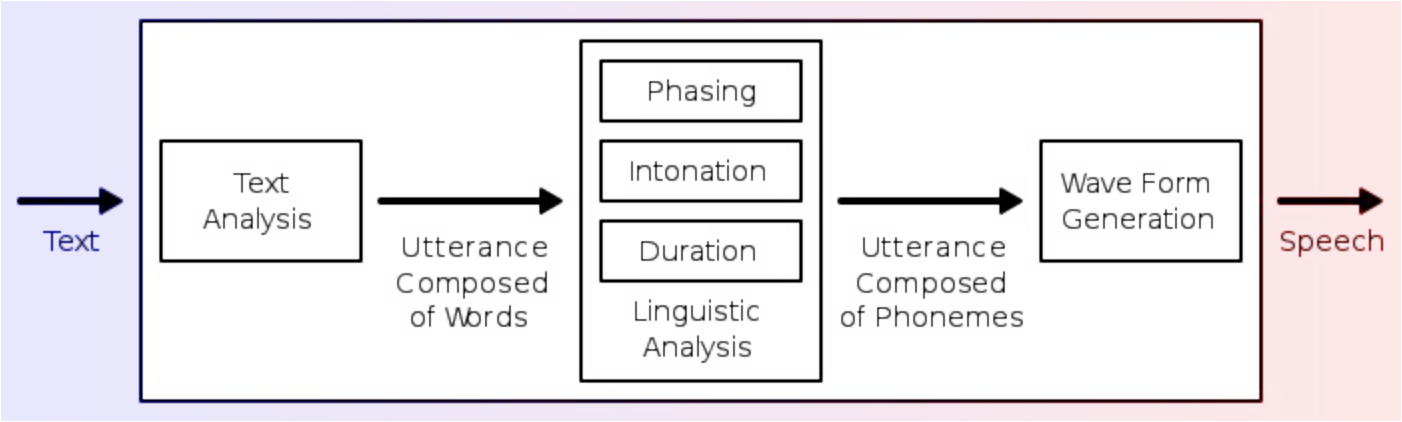
Fig. 19. Diagram of the Mel-Frequency Cepstrum Coefficients estimation.

Image Credit: Automatic Speech Recognition: From the Beginning to the Portuguese Language by André Gustavo Adami

Speaker Recognition, Verification, and Diarization



Speech Synthesis



Overview of a typical TTS system. Wikipedia by Andy0101, public domain

Speech-Language Pathology

- Receptive Language Disorders
- Expressive Language Disorders
- Use voice-quality features such as jitter and shimmer

More...

- Live Translation
- Robot Assistants
- Screen Reading

Emotion Recognition!

Generative and Discriminative Methods

Generative Models

- Probabilistic Models For Class (Phenomenon)
- Decision Boundary:
  - Which class is more “likely”?
- Can Model Unlabeled Data

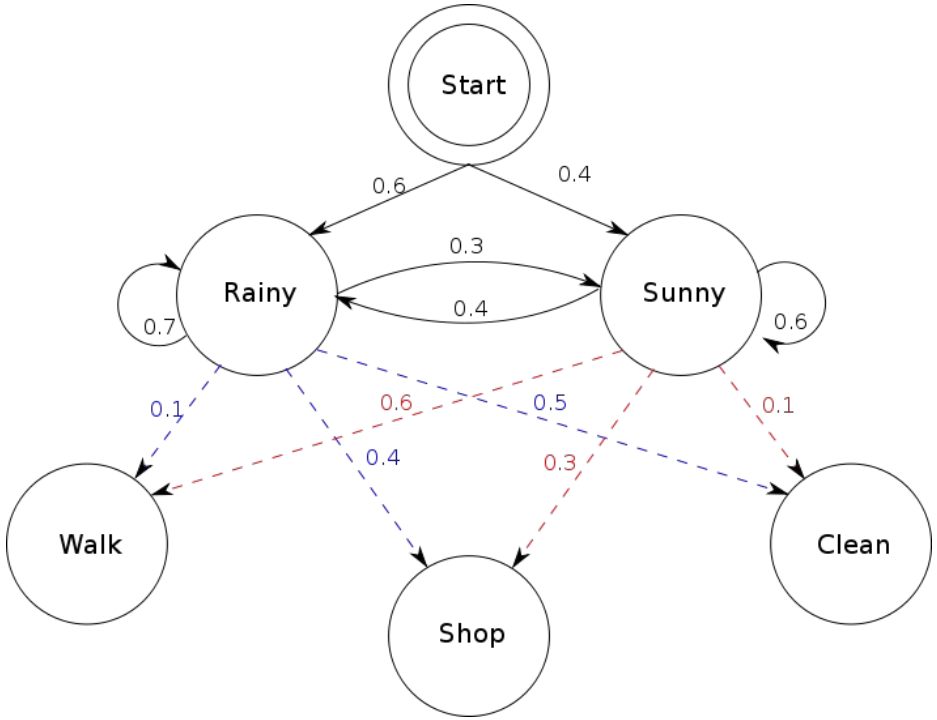
Discriminative Models

- Focus on the parameters of the decision boundary
- More Data → More Power

Hidden Markov Models

- **Memorylessness:** future states of the process depend only on the present state

- Observations → Hidden Output



Hidden Markov Models for ASR

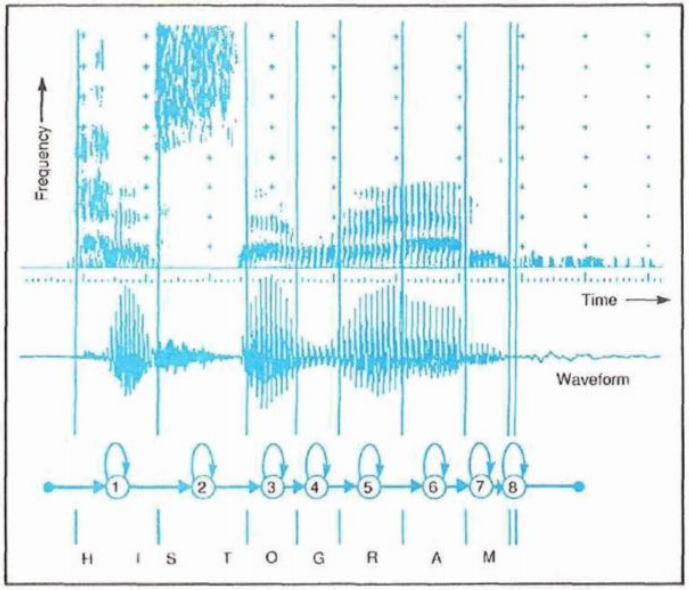


Fig. 3—Viterbi decoder alignment for the word "histogram."

Speech Recognition Using Hidden Markov Models, D.B. Paul
The Lincooln Laboratory Journal, Volume 3, Number 1 (l990)