Speech Recognition

Jim Glass / MIT 6.806-6.864 / Spring 2021

1

Today's ASR Topics

- Spoken language processing
- Speech production
- Signal representation
- Probabilistic formulation for ASR
- Lexicons, language models, acoustic models
- Search space representation via FSTs
- Corpora, evaluations, progress over time

2

Virtues of Spoken Language

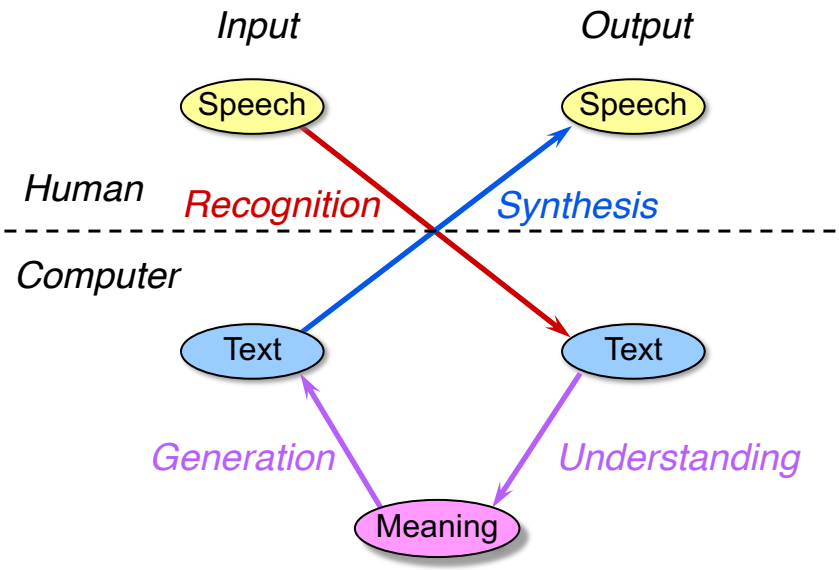
- Natural – No special training
- Flexible – Hands/eye free interaction
- Efficient – High input rate

Ideal for information access and management:

- Broad, complex information space
- Technically naïve users

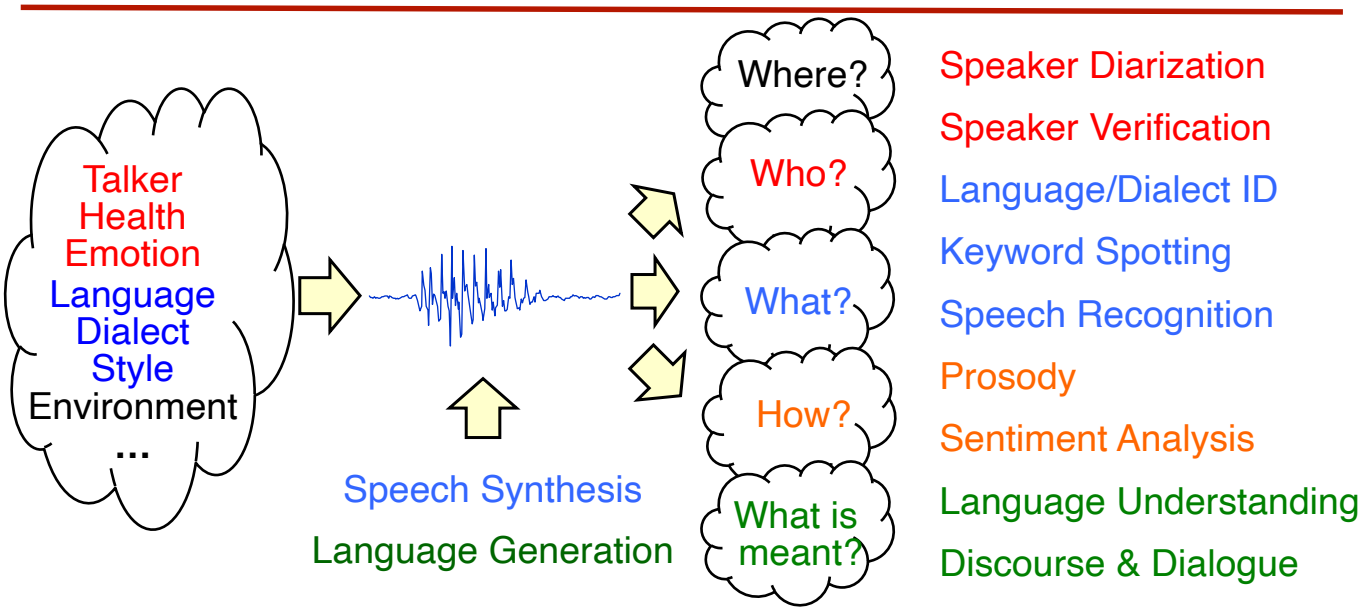
3

Spoken Communication w. Machines



4

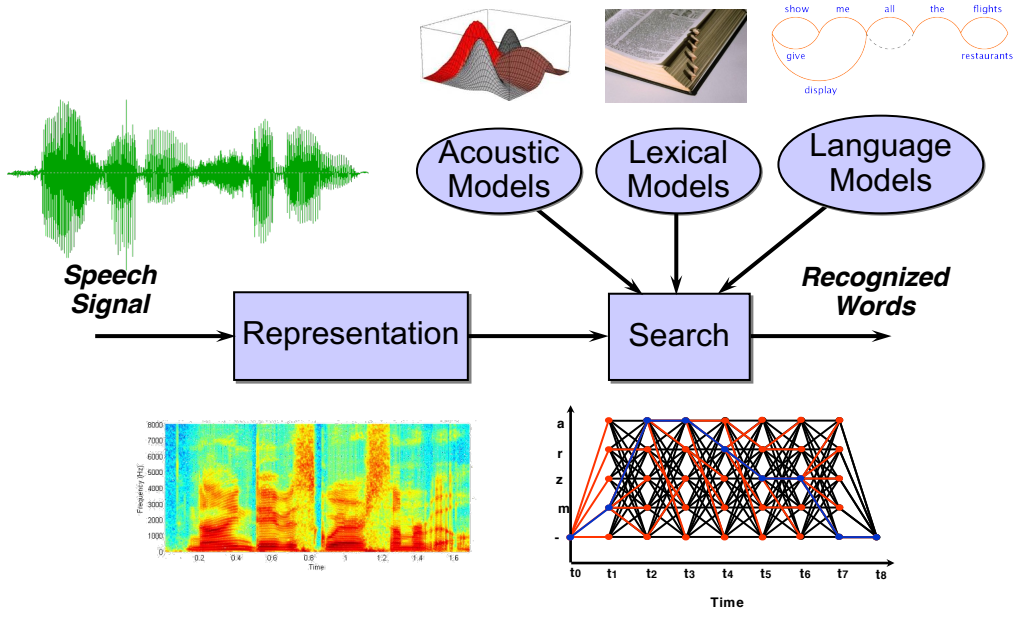
Spoken Language Processing



- For interfaces, content, language learning, health, …

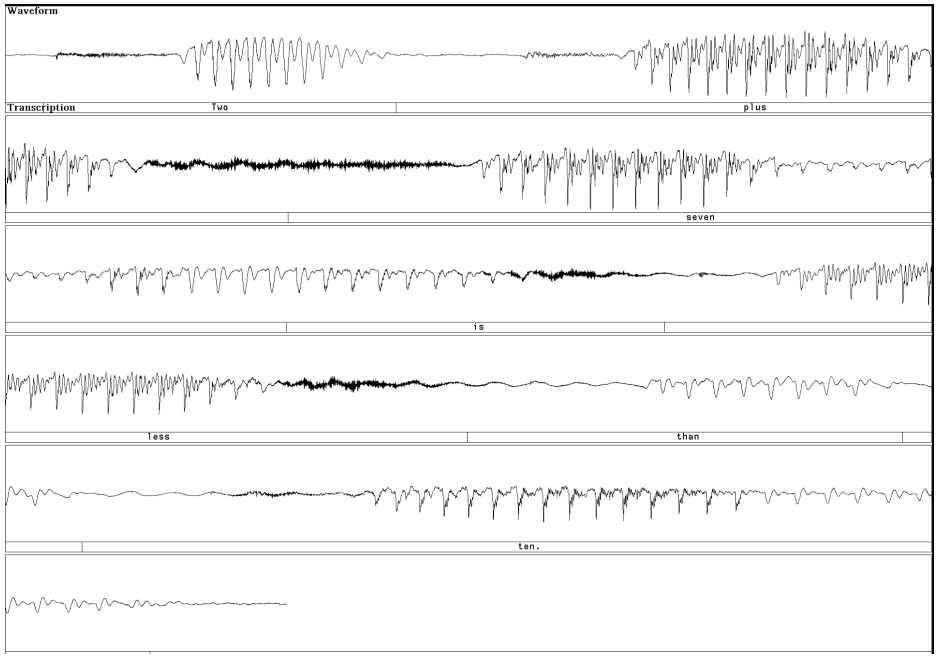
5

Speech Recognition Architecture



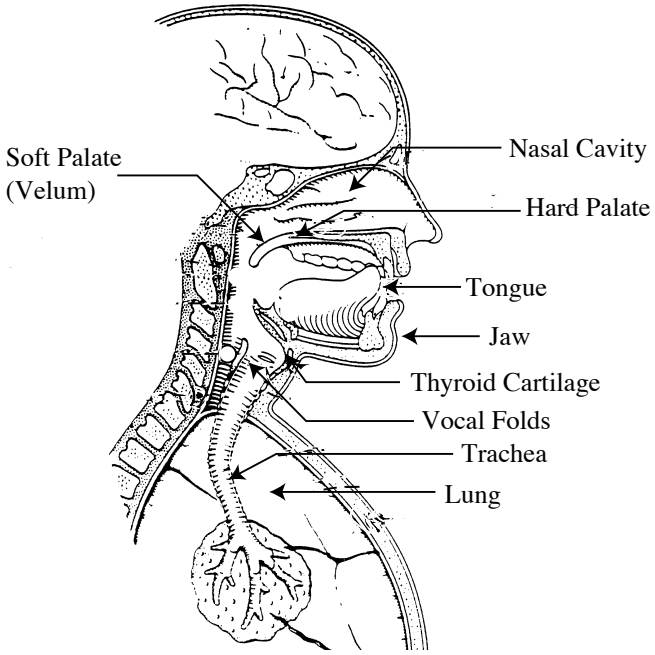
6

A Speech Waveform



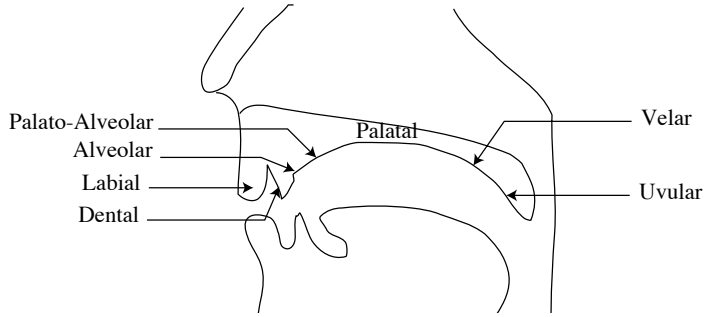
7

Anatomical Structures for Speech Production



8

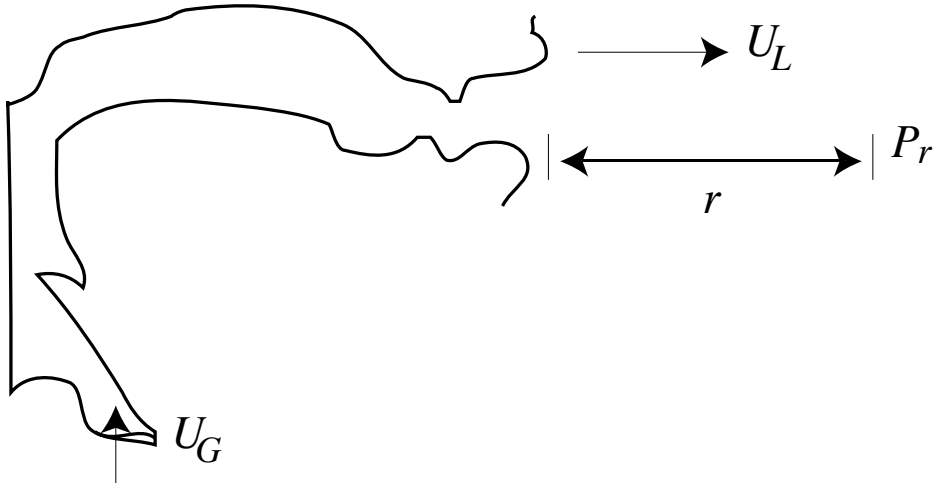
Manner and Places of Articulation



- Speech articulation characterized by manner and place
  - Vowels: No significant constriction in the vocal tract; usually voiced
  - Fricatives: turbulence produced at a narrow constriction
  - Stops: complete closure in the vocal tract; pressure build up
  - Nasals: velum lowering results in airflow through nasal cavity
  - Semivowels: constriction in vocal tract, no turbulence

9

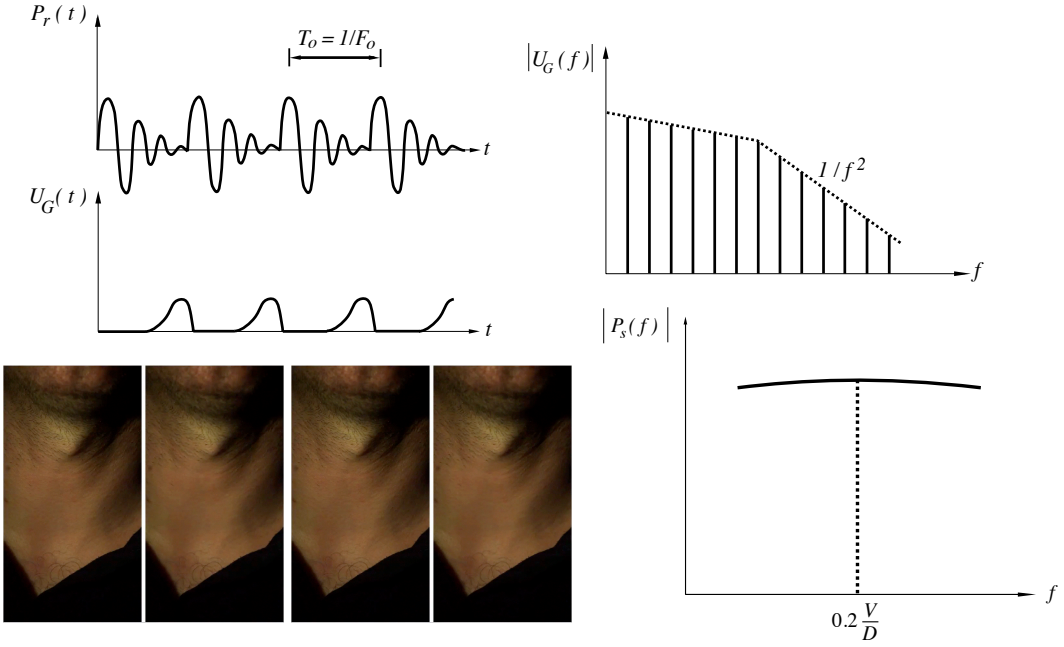
Acoustic Theory of Speech Production



- Speech produced via coordinated movement of articulators
- Spectral characteristics of speech influenced by source, vocal tract shape, and radiation characteristics

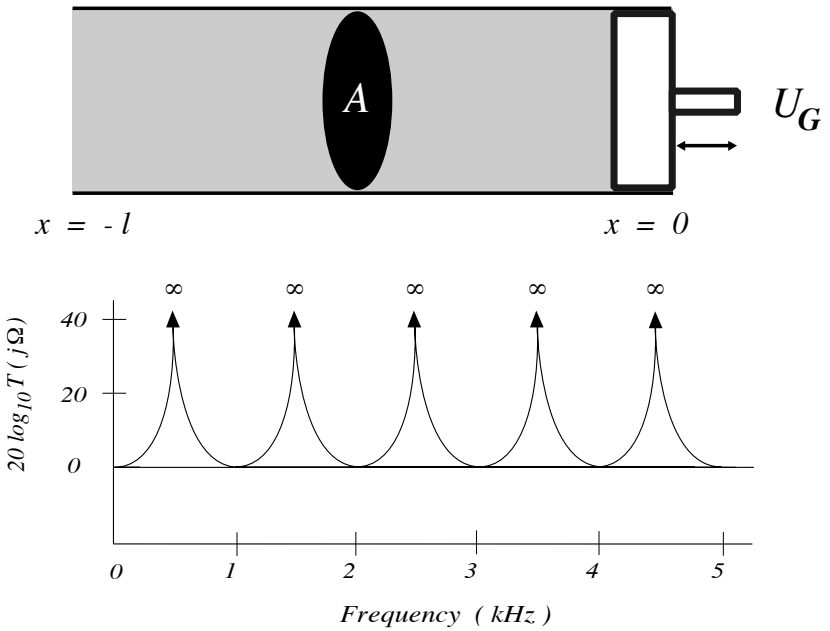
10

Vocal Fold Vibration & Turbulence



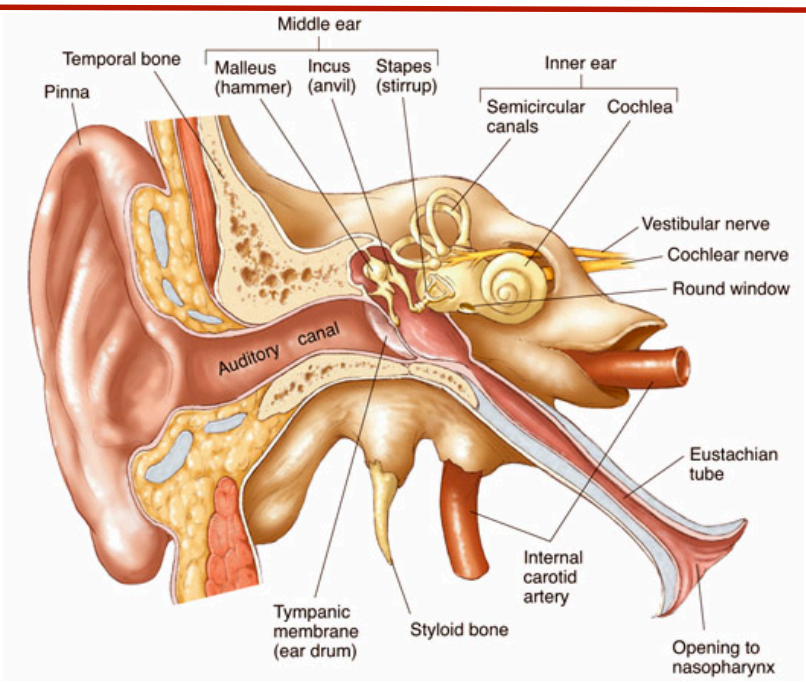
11

Propagation of Sound in a Uniform Tube



12

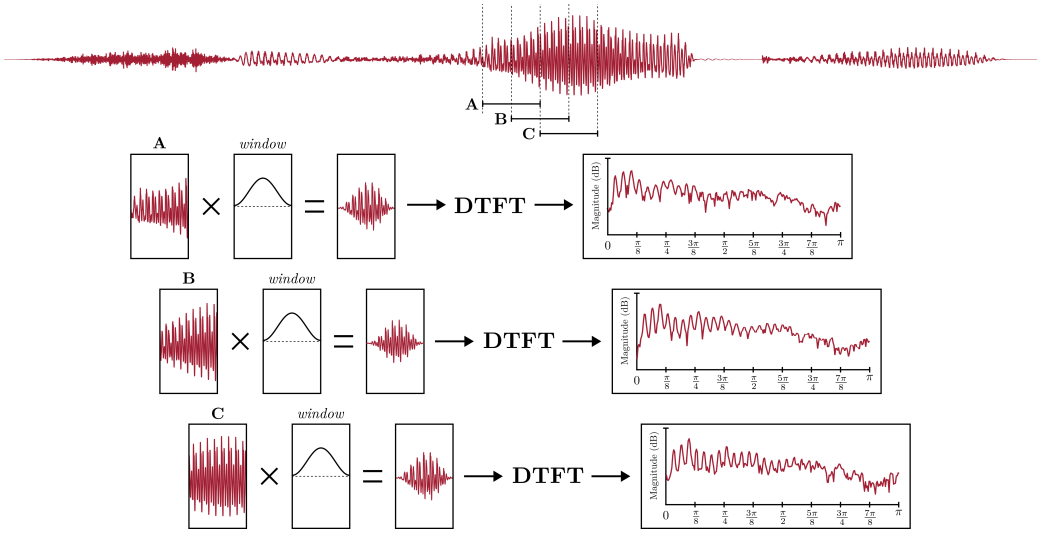
Anatomy of the Human Ear



13

Digital Speech Processing

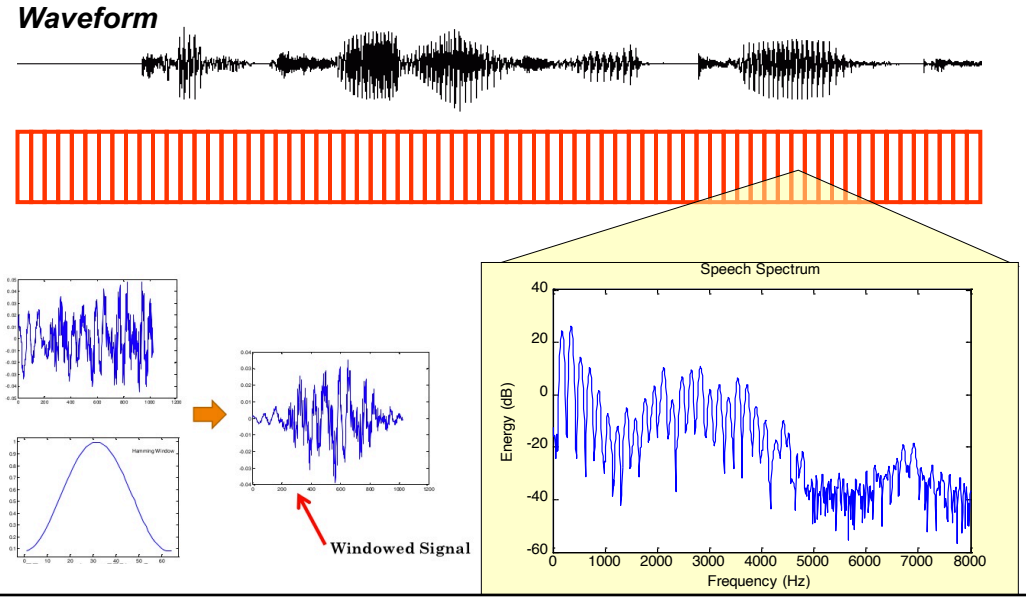
- Speech is typically represented by a sequence of Discrete-Time Fourier Transforms computed over a windowed snippet of speech



14

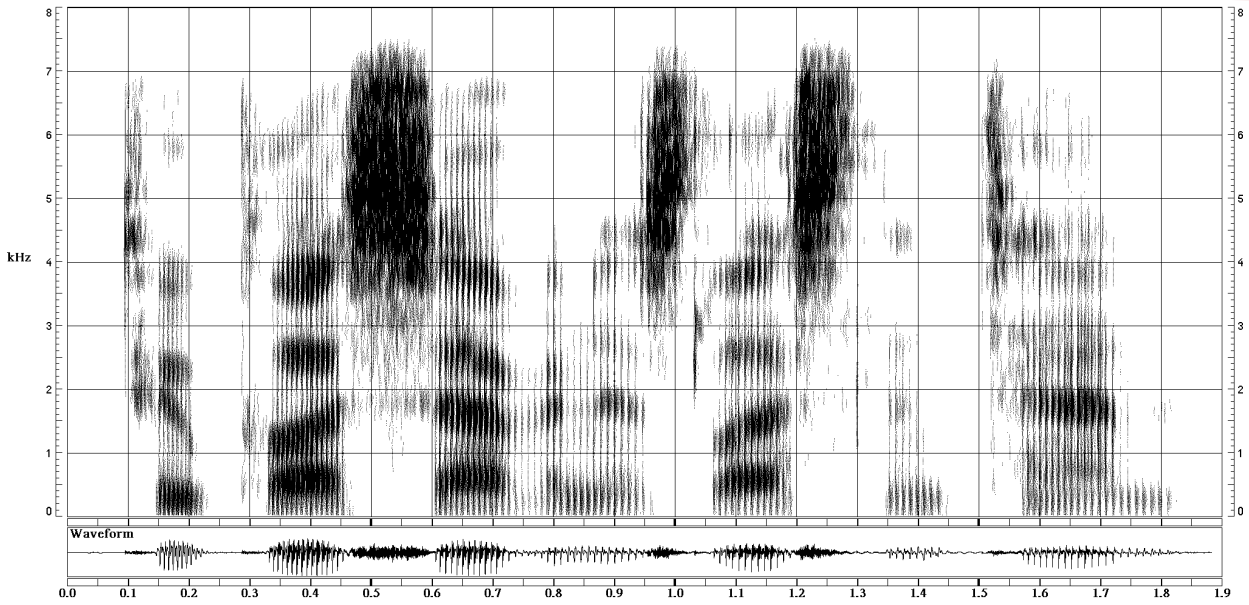
Speech Processing

Spectral energies (frames) are typically computed at regular (e.g., 10ms) intervals



15

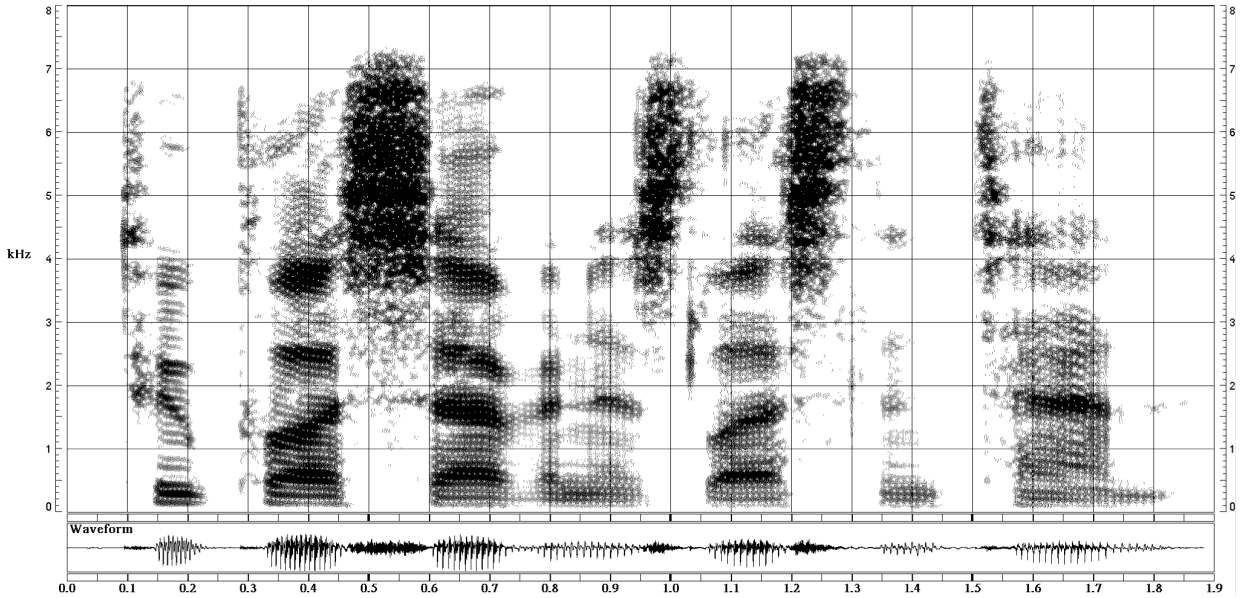
The Speech Spectrogram



Spectrograms visualize spectral-temporal characteristics of speech

16

A Narrowband Spectrogram

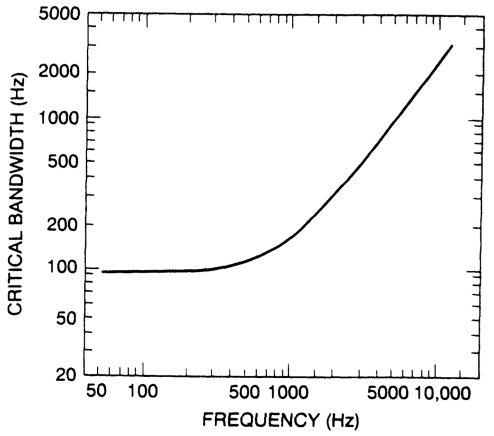


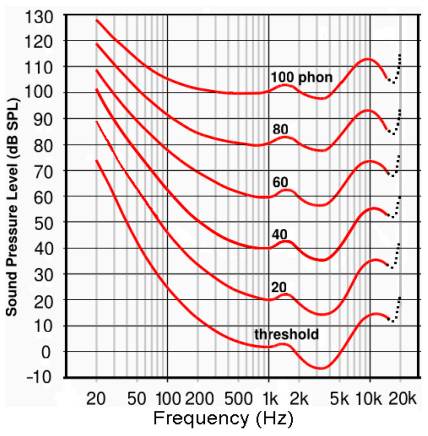
ASR spectra are typically computed with a ~25ms Hamming window

17

Auditory Representations

- Perceptually relevant properties of the human auditory system can be incorporated into the spectral estimation process



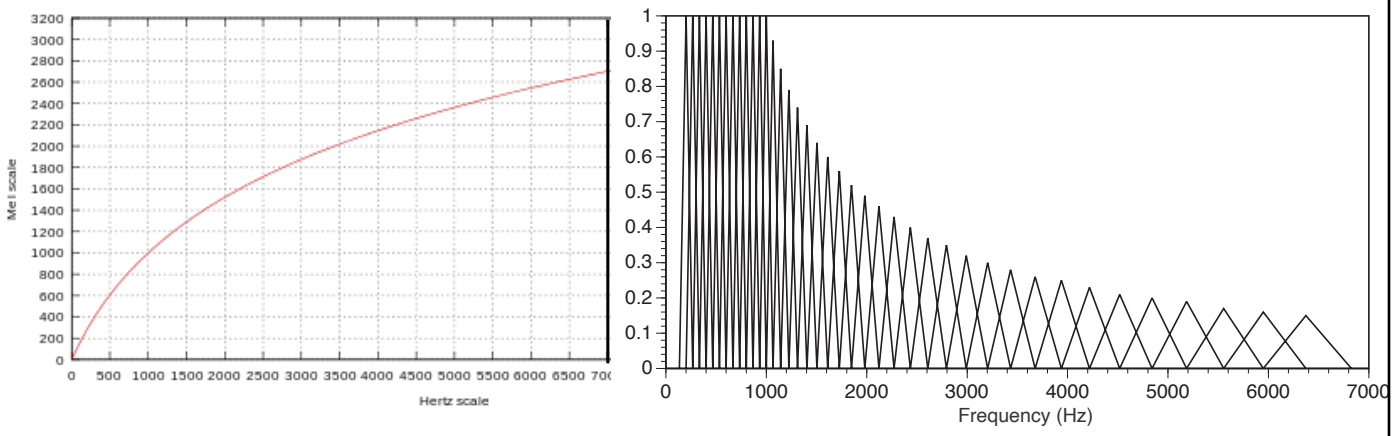


- A variety of models have been explored in the past but simple models based on Mel-frequency warping have worked well

18

Mel-Frequency Spectral Coefficients (MFSCs)

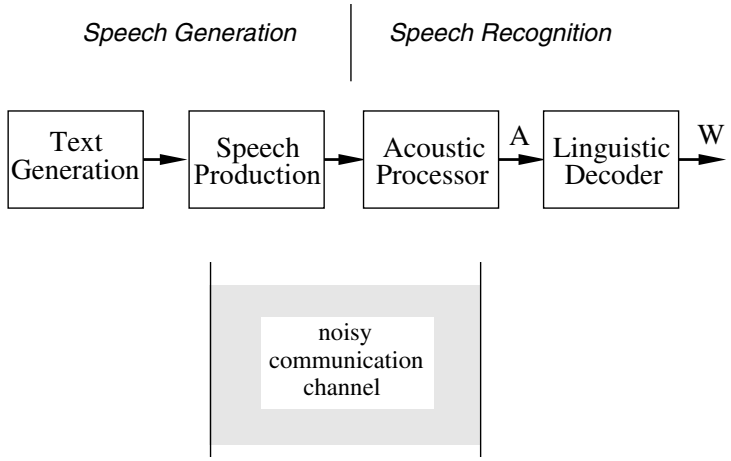
- Mel-frequency scale originally based on pitch perception
  - Spectral filters are spaced evenly on Mel-scale



- Mel-frequency cepstral coefficients (MFCCs) are popular for ASR
  - MFCCs are typically computed via the discrete cosine transform

19

Information Theoretic Formulation of ASR



$$W^* = \arg\max_W P(W \mid A) \qquad P(W \mid A) = \frac{P(A \mid W)P(W)}{P(A)}$$

20

Probabilistic ASR Formulation

- A full search considers all possible segmentations, *S*, and units, *U*, for each hypothesized word sequence, *W*

$$W^* = \arg\max_W P(W \mid A) = \arg\max_W \sum_S \sum_U P(WUS \mid A)$$

- Can seek best path to simplify search using dynamic programming (e.g., Viterbi) or graph-searches (e.g., *A**)

$$W^*, U^*, S^* \approx \arg\max_{W,U,S} P(WUS \mid A)$$

- The modified Bayes decomposition has four terms:

$$P(WUS \mid A) = \frac{P(A \mid SUW)P(S \mid UW)P(U \mid W)P(W)}{P(A)}$$

acoustic model | duration model | lexical model | language model

21

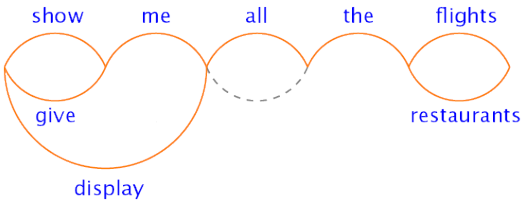
Lexical Modeling

- Pronunciation dictionaries are typically used to describe the realizations of words in terms of sub-word units
  - English is comprised of ~40 phonemes
  - Pronunciations and phonological rules can be supplied or learned from data
- For languages with regular letter-to-sound transformations, *graphemic* approaches are effective
  - Many end-to-end neural network-based ASR models output letter sequences

a	(ax | ey)
...	
beach	b iy ch
...	
nice	n (iy | ay) s
...	
recognize	r eh k ax gd n ay z
...	
speech	s p- iy ch
...	
stata	s t- (ey | aa) tf ax
...	
tomato	t ax m (ey | aa) tf ow
...	
wreck	r eh kd
...	

22

Language Modeling



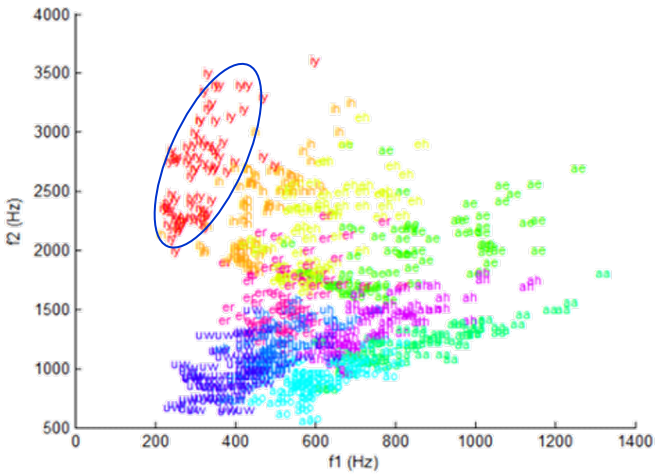
- ASR systems constrain possible word combinations by way of simple, but powerful, language models
  - Trigram is the dominant language model for ASR:

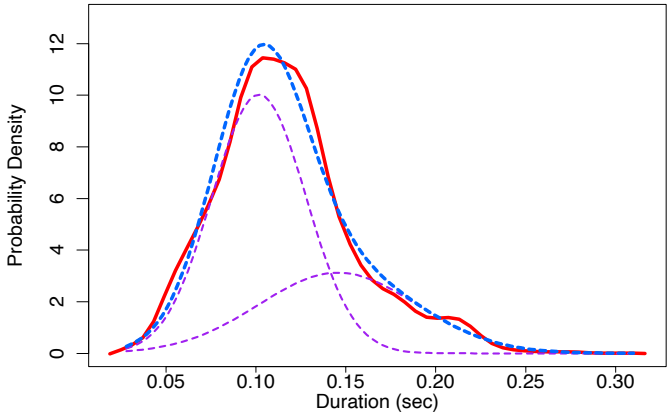
$$P(w_i | w_{i-2}, w_{i-1})$$

  - Much effort has gone into smoothing techniques for sparse data
- RNN-based LMs have shown improved performance
  - Generally used as a post-process to re-rank ASR hypotheses
- Task difficulty is measured by perplexity

23

Acoustic Modeling



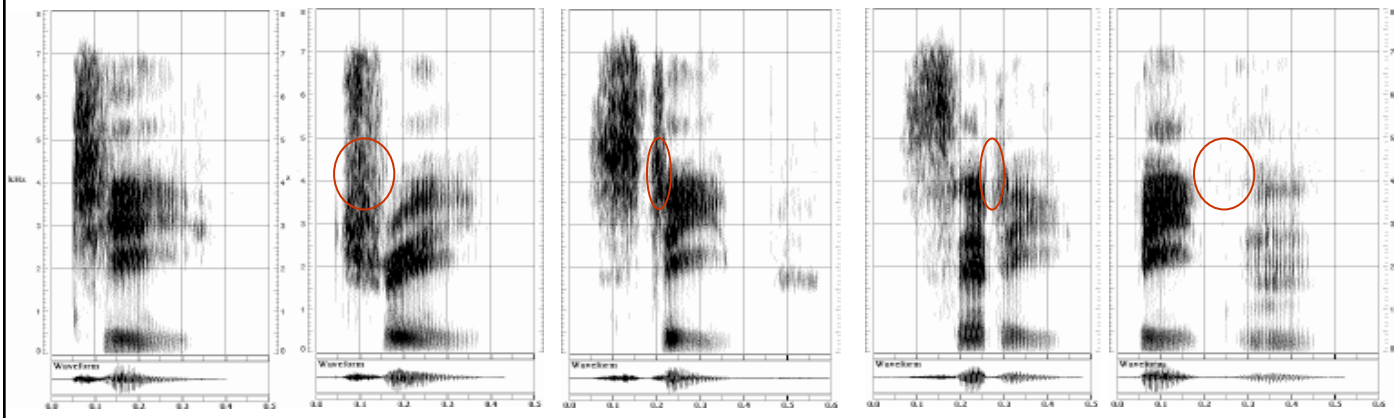


Acoustic models must capture the variation in the physical realization of an underlying linguistic unit due to speaker, style, environmental factors etc.

24

Coarticulation and Phonological Variation

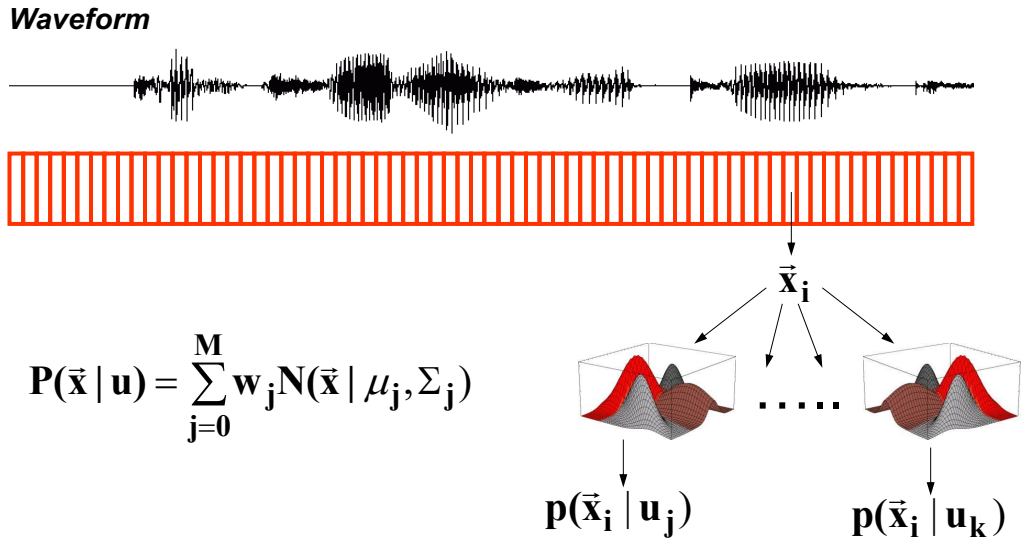
- The acoustic realization of a phoneme depends strongly on context
- Context-dependent models are often used to model coarticulation



25

Frame-based Acoustic Modeling

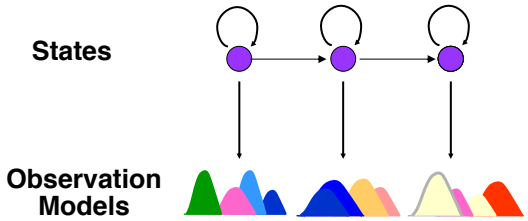
Acoustic models (e.g., GMMs, ANNs) are typically applied to each spectral frame



26

Hidden Markov Models (HMMs)

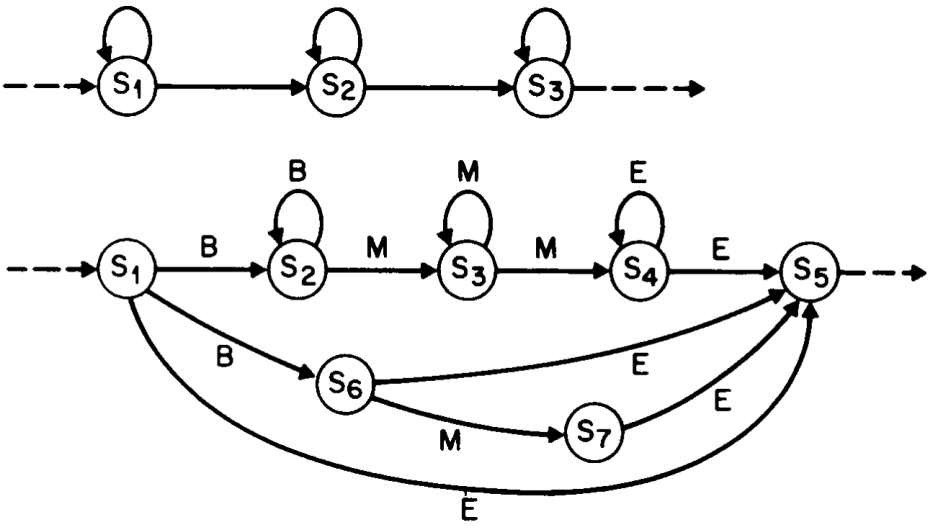
- Dominant modeling framework used for speech recognition
- Generative model that predicts likelihood of observation sequence being generated by an underlying state sequence
  - Either discrete or continuous observation models can be used



- HMMs can model words or sub-words (e.g., phones)
  - Sub-word HMMs concatenated to create larger word-based HMMs

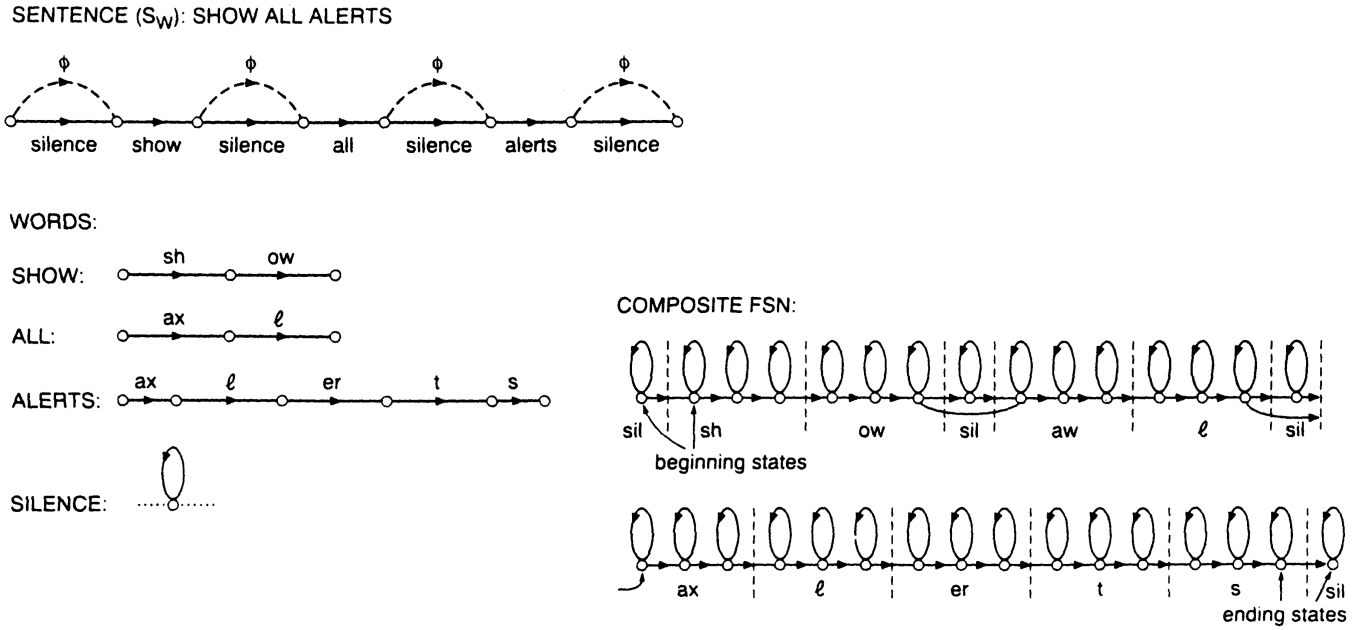
27

HMM Sub-Word Topologies



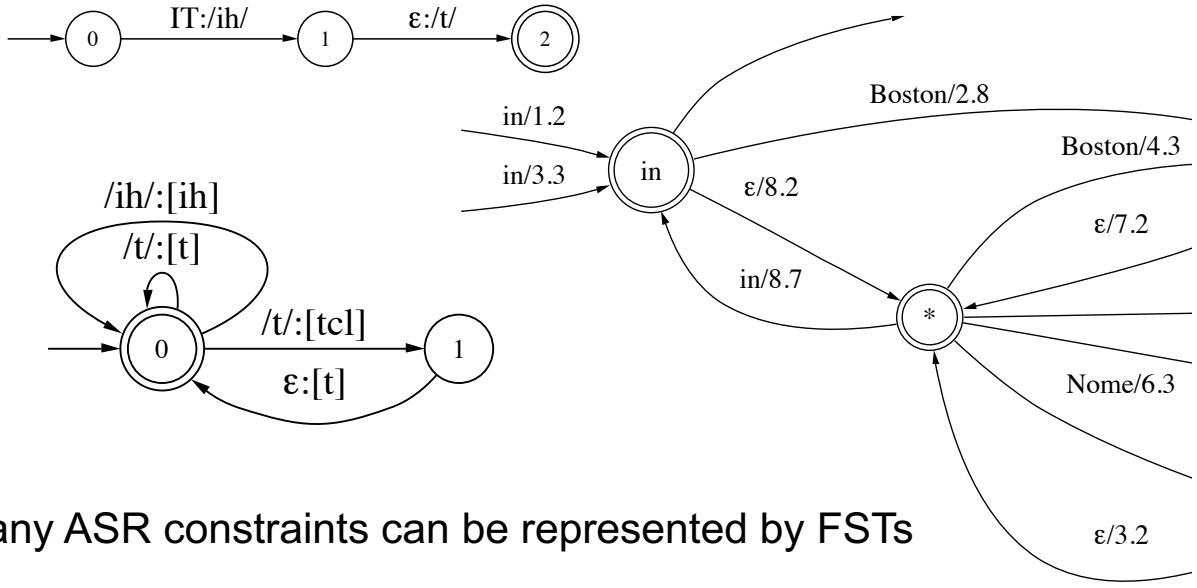
28

HMM ASR Formulation



29

Finite State Transducer Representations for ASR



Many ASR constraints can be represented by FSTs

30

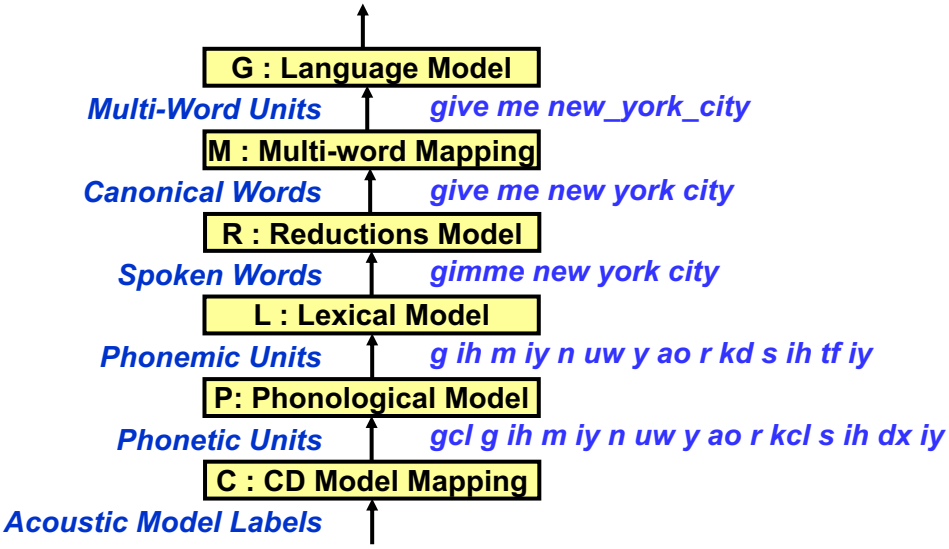
Speech Recognition as Cascade of FSTs

- ASR as a cascade of FSTs:
  $$O \circ (M \circ P \circ L \circ G)$$
  - G: language model (weighted words ← words)
  - L: lexicon (phonemes ← words)
  - P: phonological rule application (phones ← phonemes)
  - M: model topology (e.g., HMM) (states ← phones)
  - O: observations with acoustic model scores
- (M ∘ P ∘ L ∘ G) is *single* FST seen by search
- Viterbi search performs composition of O with (M ∘ P ∘ L ∘ G)
- Gives great flexibility in how components are combined

31

Expanded FST Representation

- FST representation can be expanded for more efficient representation of lexical variation



32

Evaluating Speech Recognition

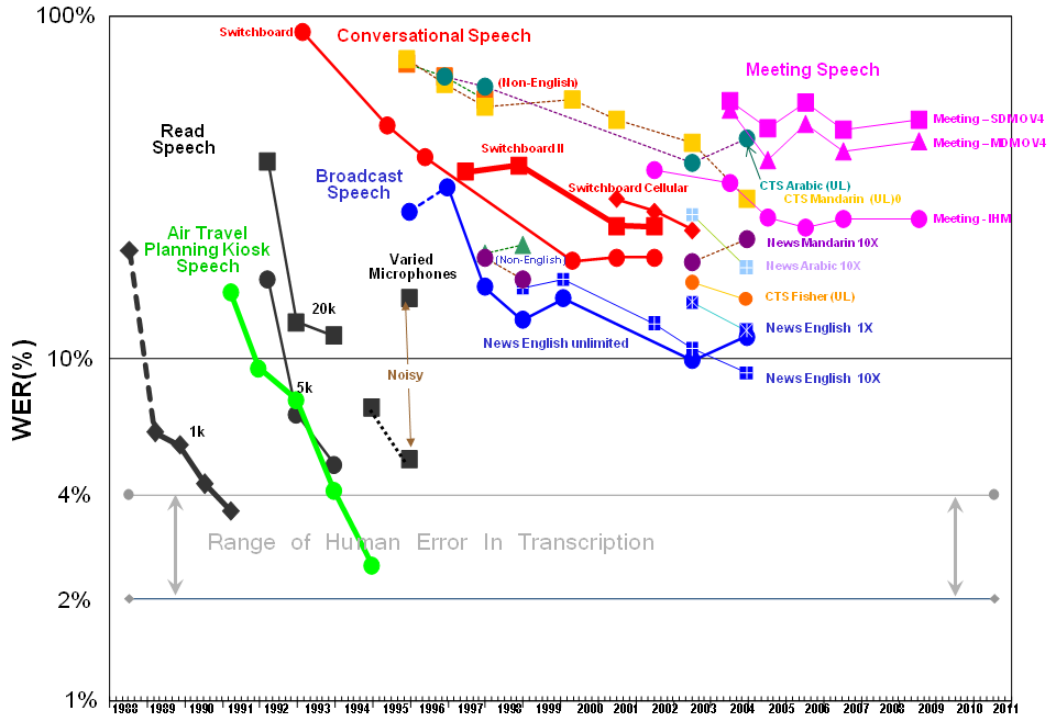
- The standard evaluation metric for ASR is word error rate (WER)
  - Based on a string alignment between hypothesis and reference text
  - Errors can include insertions, deletions, and substitutions

REF:	i	***	**	UM	the	PHONE	IS		i	LEFT	THE	portable	****	PHONE	UPSTAIRS	last	night
HYP:	i	GOT	IT	TO	the	*****	FULLEST		i	LOVE	TO	portable	FORM	OF	STORES	last	night
Eval:		I	I	S		D	S			S	S		I	S	S		

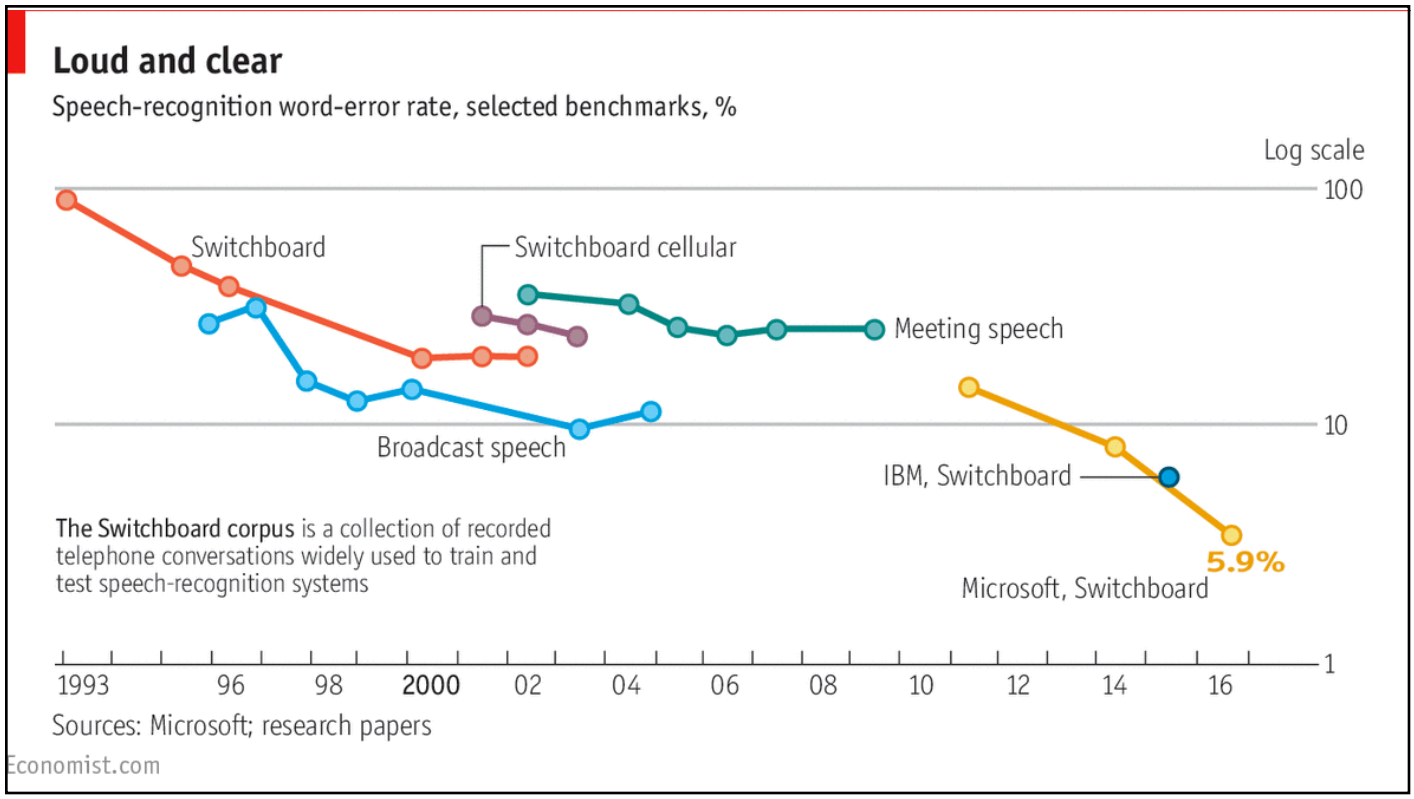
$$WER = 100 \text{ x } \frac{Insertions + Substitutions + Deletions}{Total\ Words\ in\ Correct\ Transcript}$$

- The same metric can be applied to character error rate (CER) etc.
- String edit distance tends to underestimate errors at high WERs
- Standard NIST software (sclite) is available to measure WERs

33

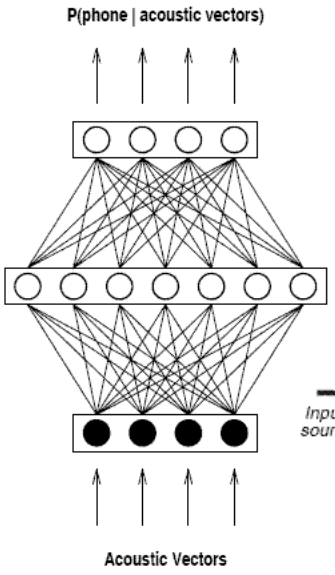


34

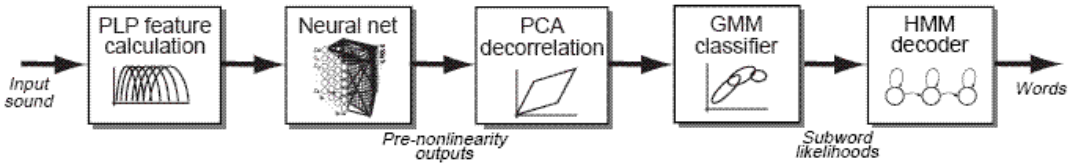


35

Neural Network-based ASR



- Artificial neural network models were first explored for ASR in 1990s
- Over the last decade ANNs have replaced GMMs as the dominant acoustic model for ASR
- ANN outputs can replace GMMs altogether (*Hybrid*), or be used as inputs to HMM-GMMs (*Tandem*)



- Connectionist Temporal Classification (CTC) and Seq2Seq models replace HMMs (next lecture!)

36

References

- Readings:
  - Jurafsky and Martin, “Speech and Language Processing”, Chp. 26

37