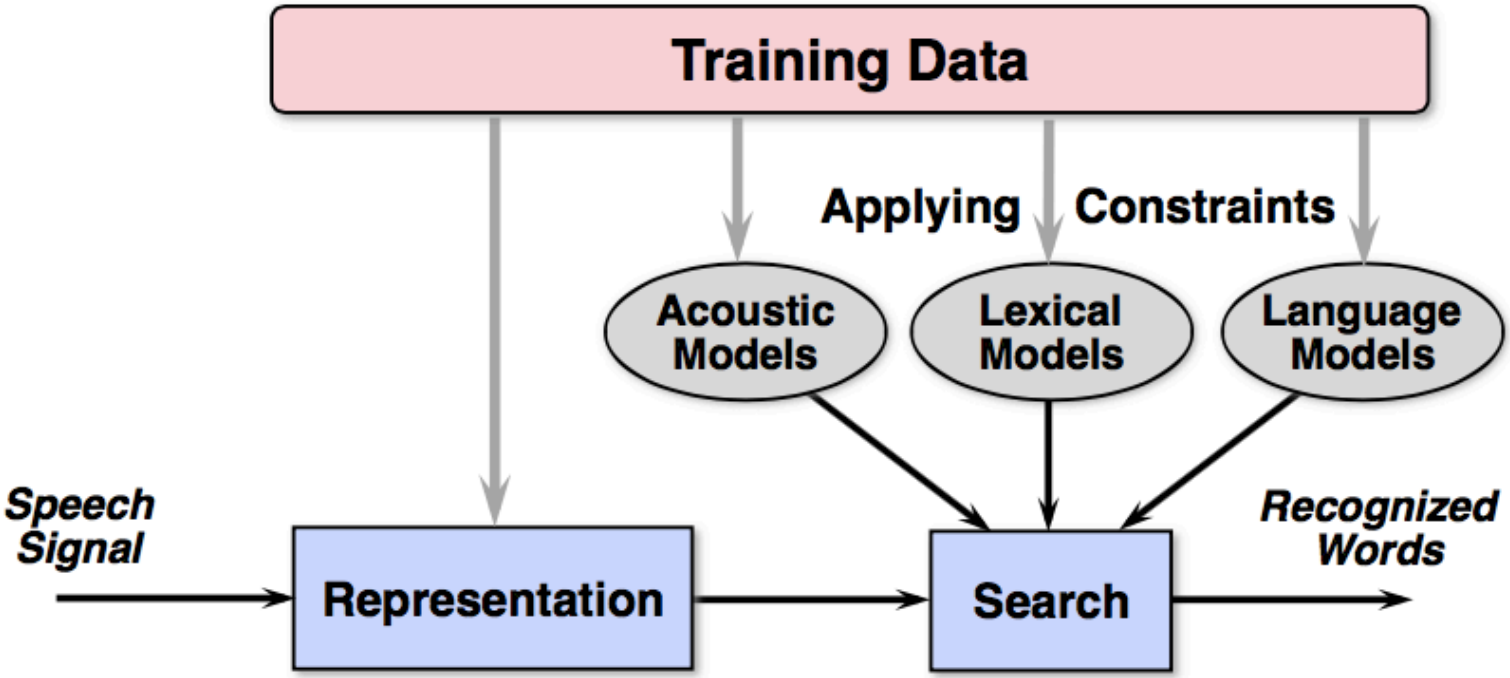
ASR

Many slides courtesy of James Glass

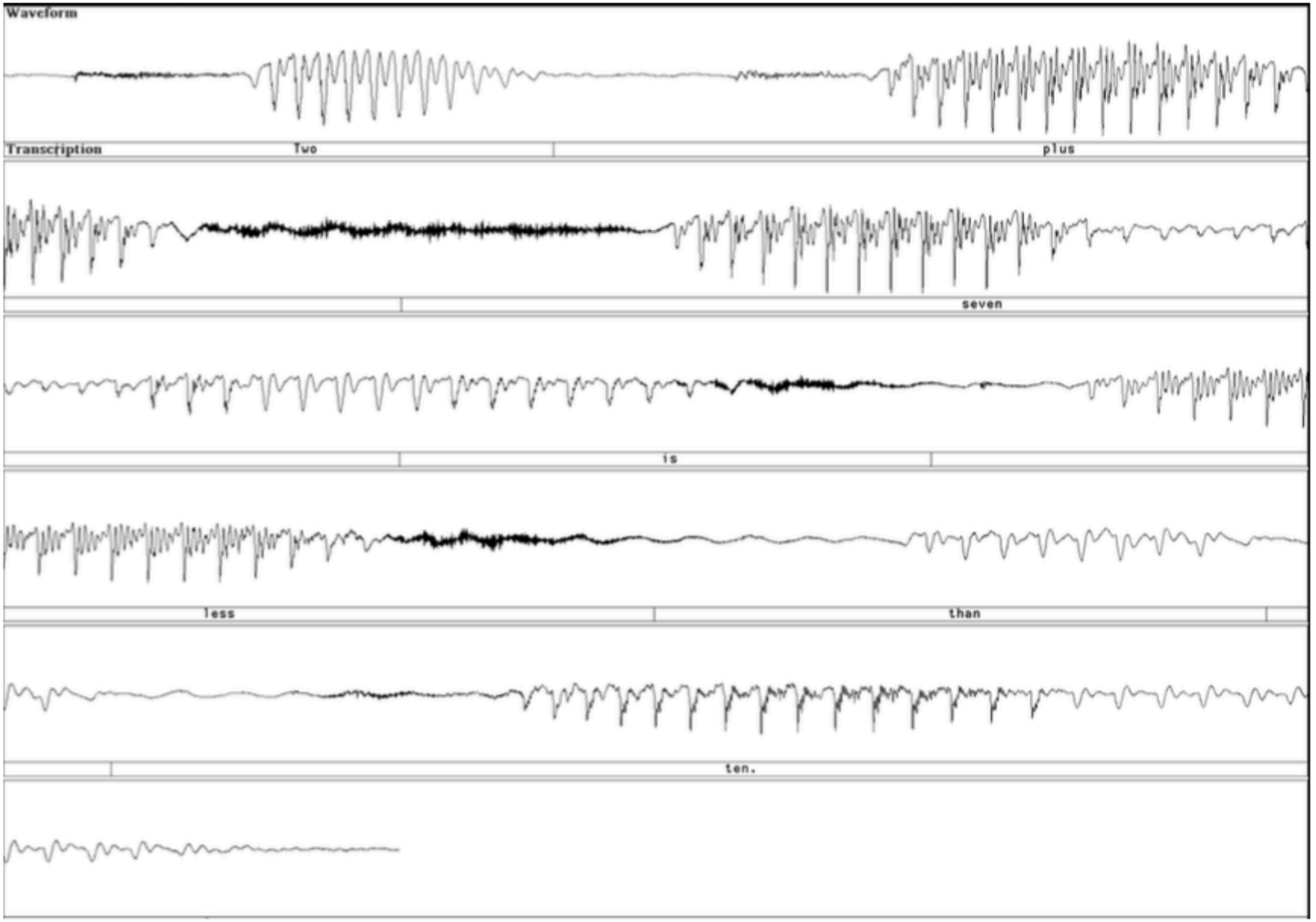
Automatic Speech Recognition (ASR)

- Components of a speech recognizer

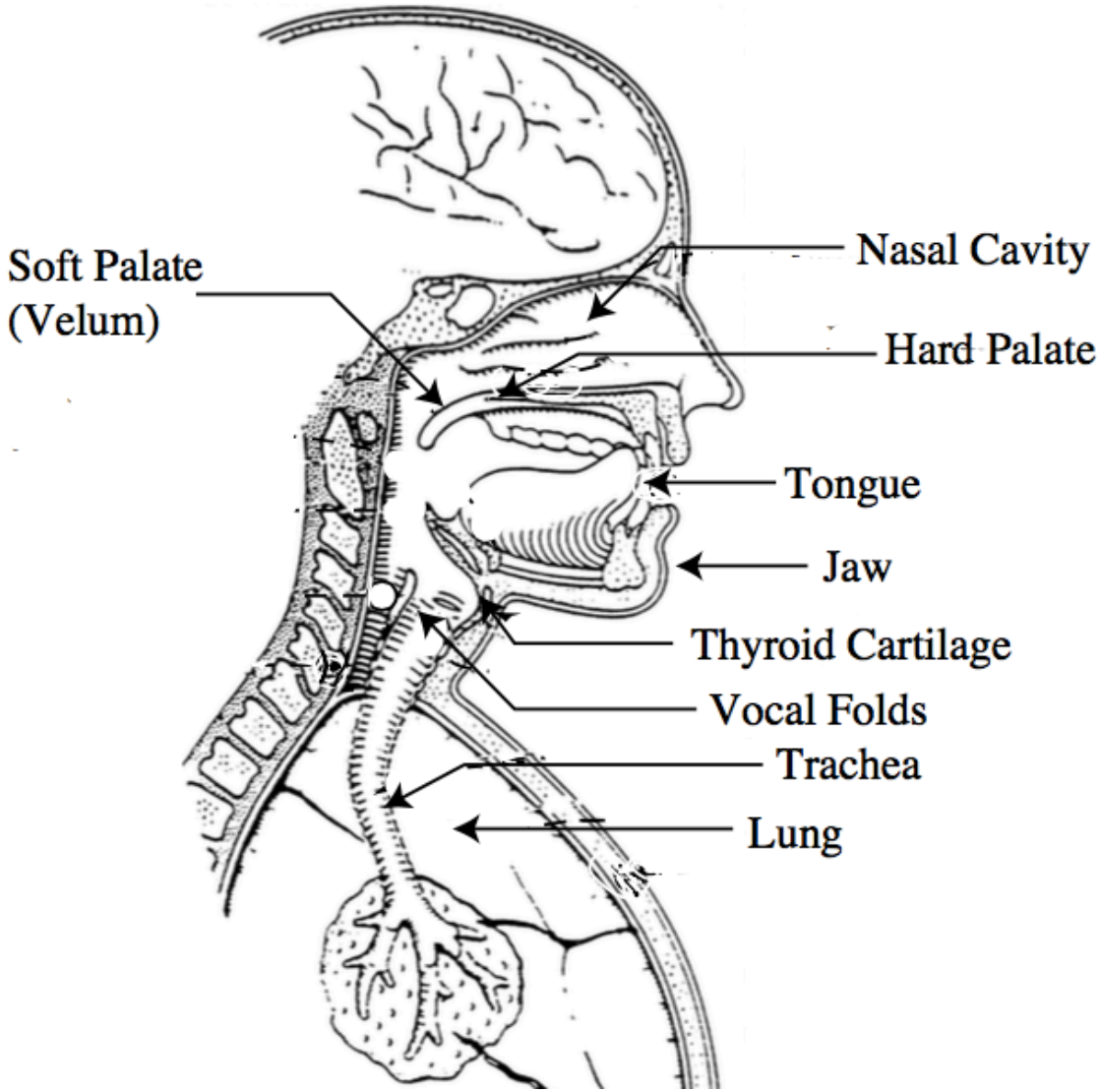


3 Components
How to represent the signal
How to model the constraints
How to search for the optimal answer

Speech signal

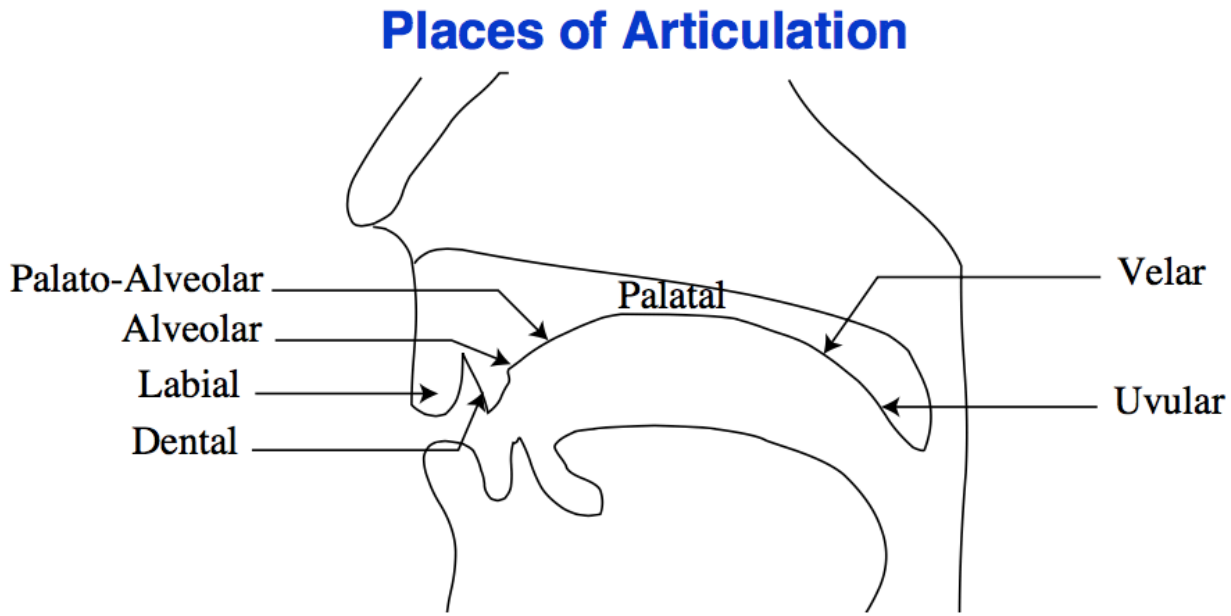


Speech production

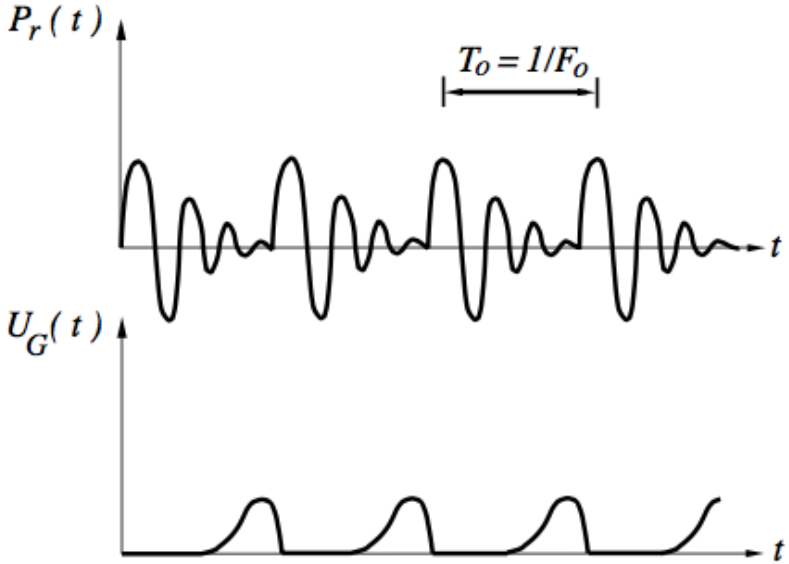


Types of sound

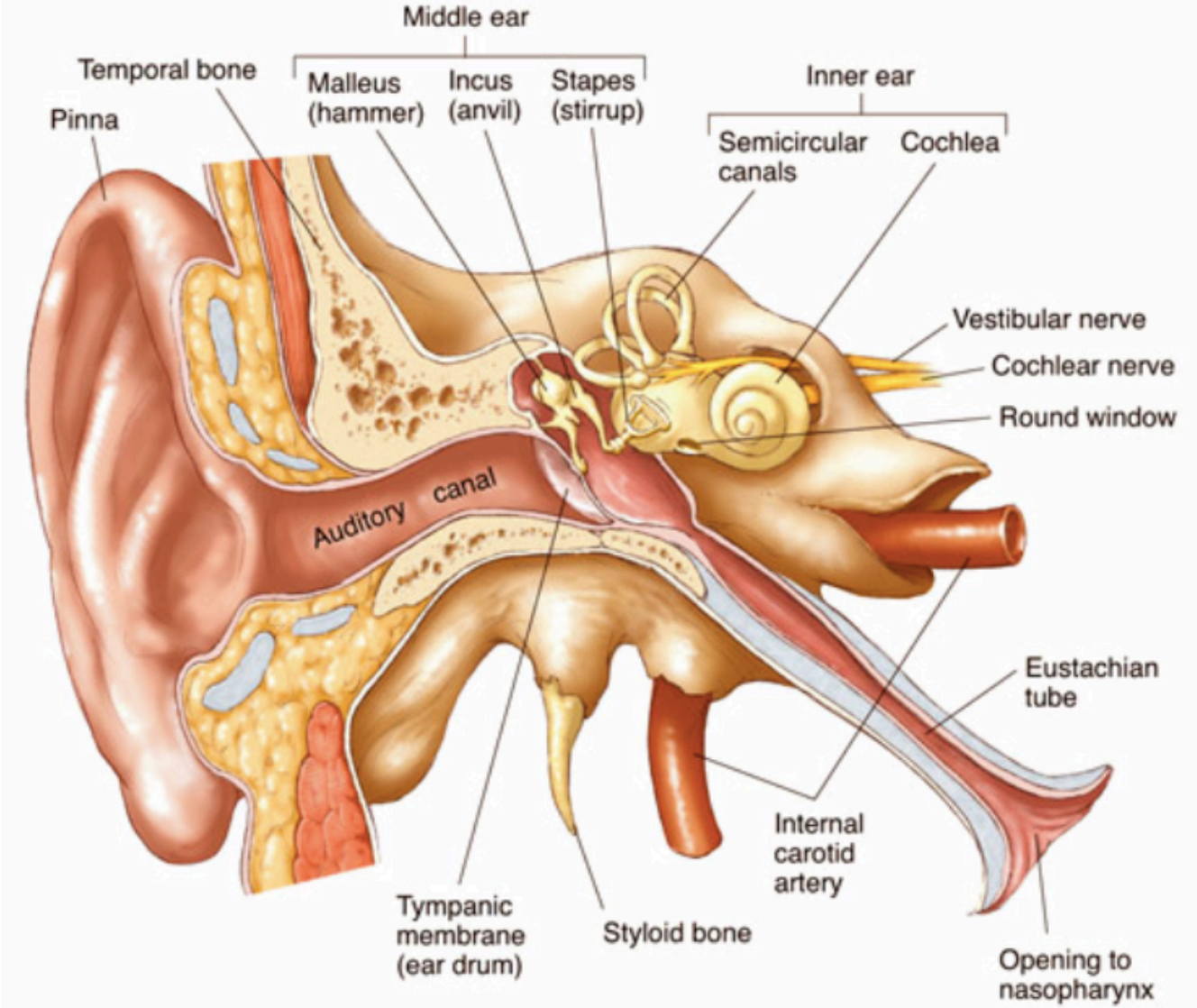
- Speech articulation categorized by manner and place
  - Vowels: No significant constriction in the vocal tract
  - Fricatives: Turbulence produced at a narrow constriction
  - Stops: complete closure of the vocal tract; pressure build up
  - Nasals: velum lowering results in airflow through the nasal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



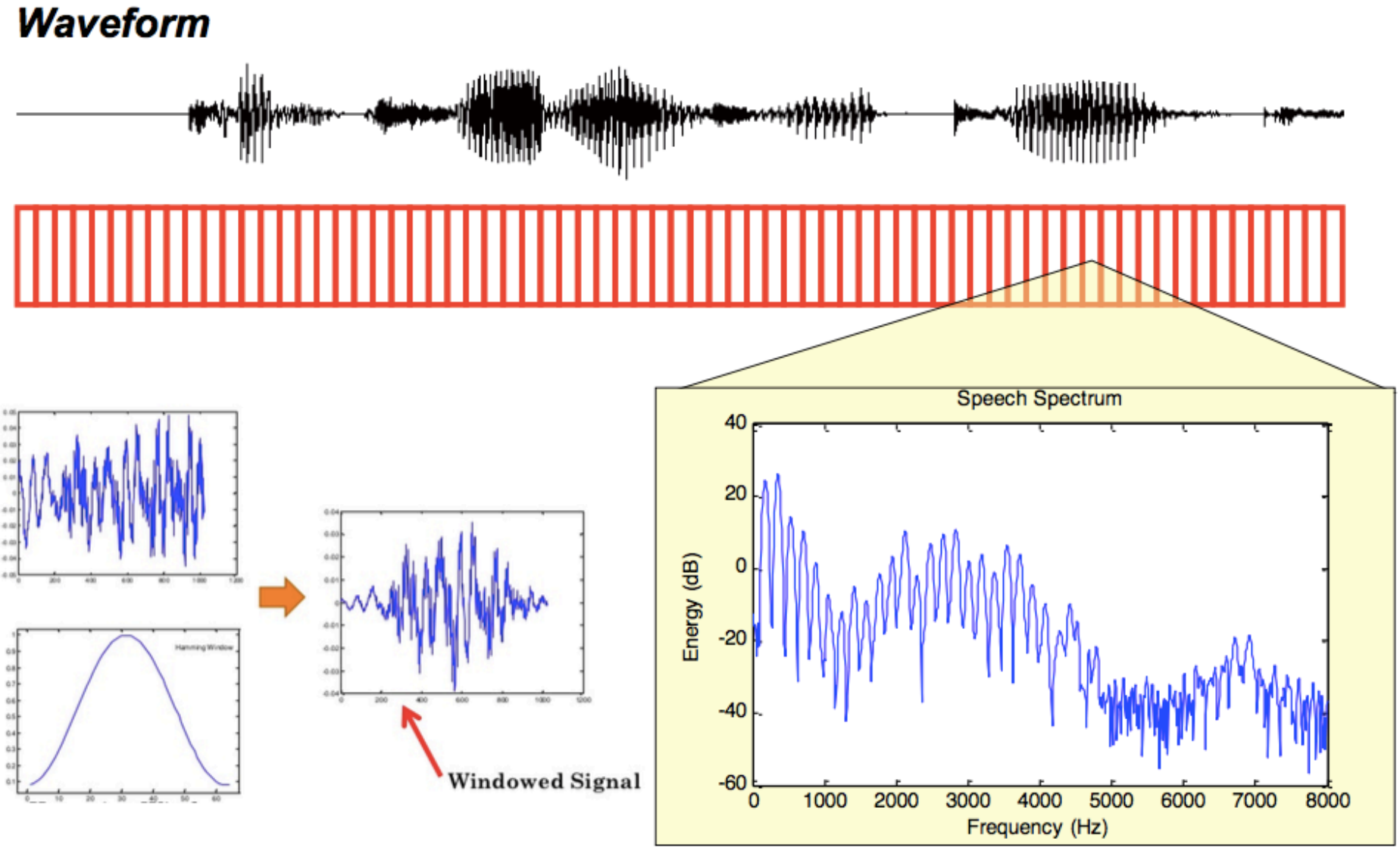
Vocal fold vibration



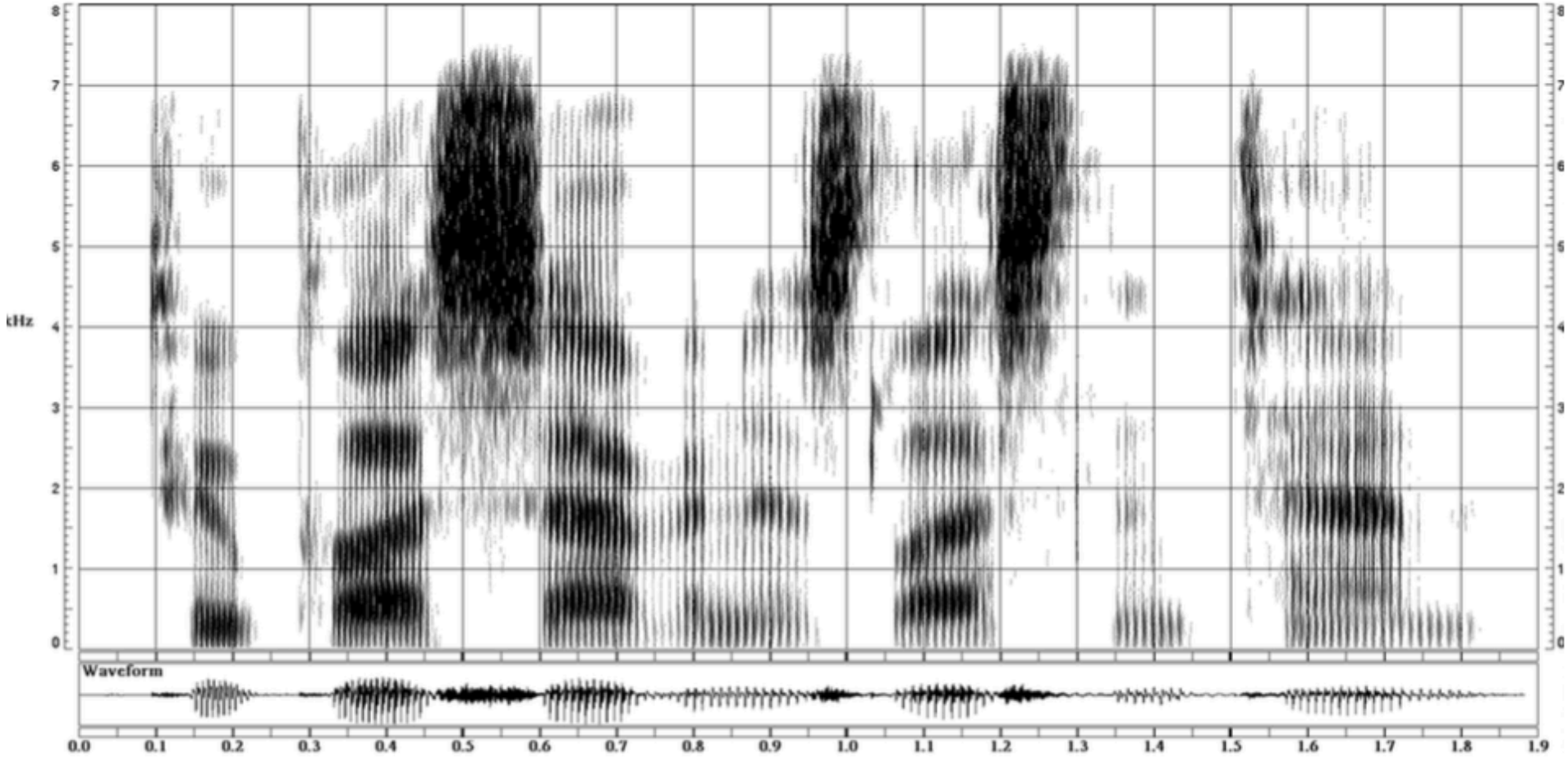
Human ear


Middle ear
Temporal bone
Pinna
Malleus (hammer)
Incus (anvil)
Stapes (stirrup)
Inner ear
Semicircular canals
Cochlea
Vestibular nerve
Cochlear nerve
Round window
Auditory canal
Eustachian tube
Internal carotid artery
Tympanic membrane (ear drum)
Styloid bone
Opening to nasopharynx

Speech processing

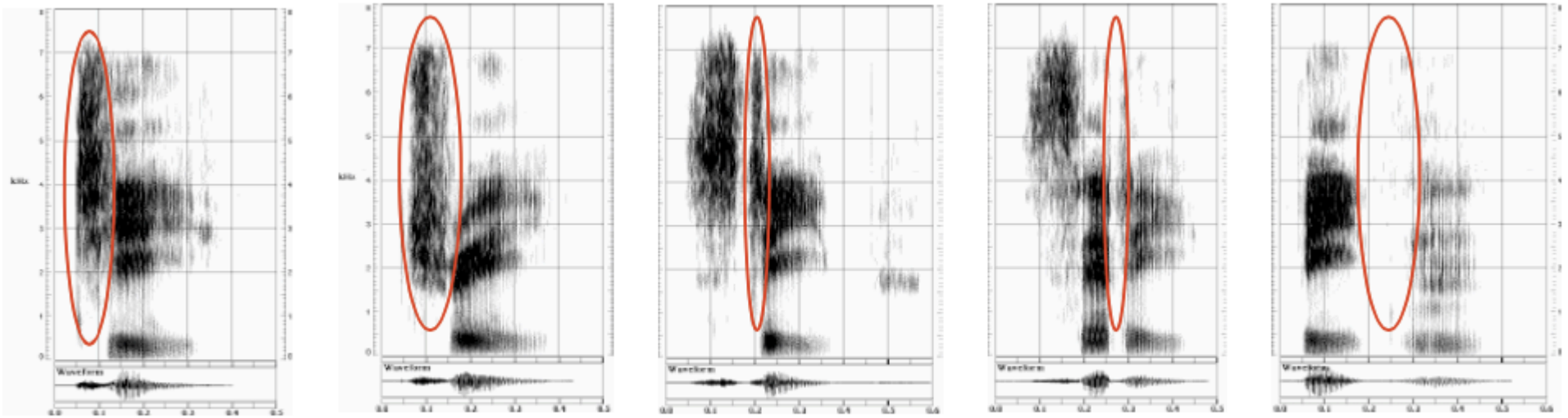


Spectrogram

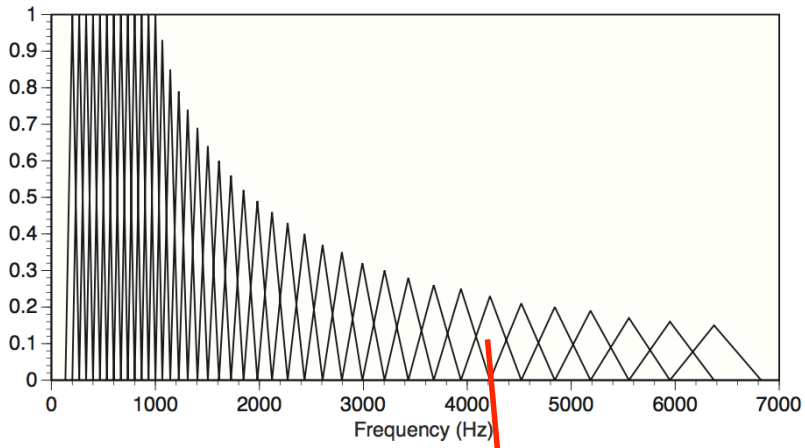


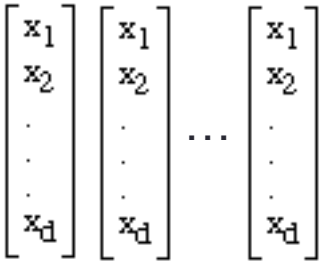
Variation of a phoneme

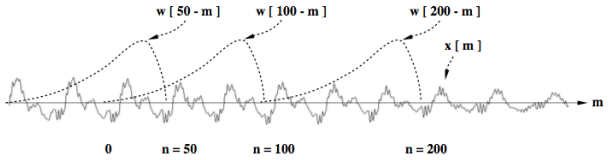
- The most basic sound unit “phoneme”
- The acoustic realization of a phoneme depends on the context



Speech feature extraction







Window and Fourier transform

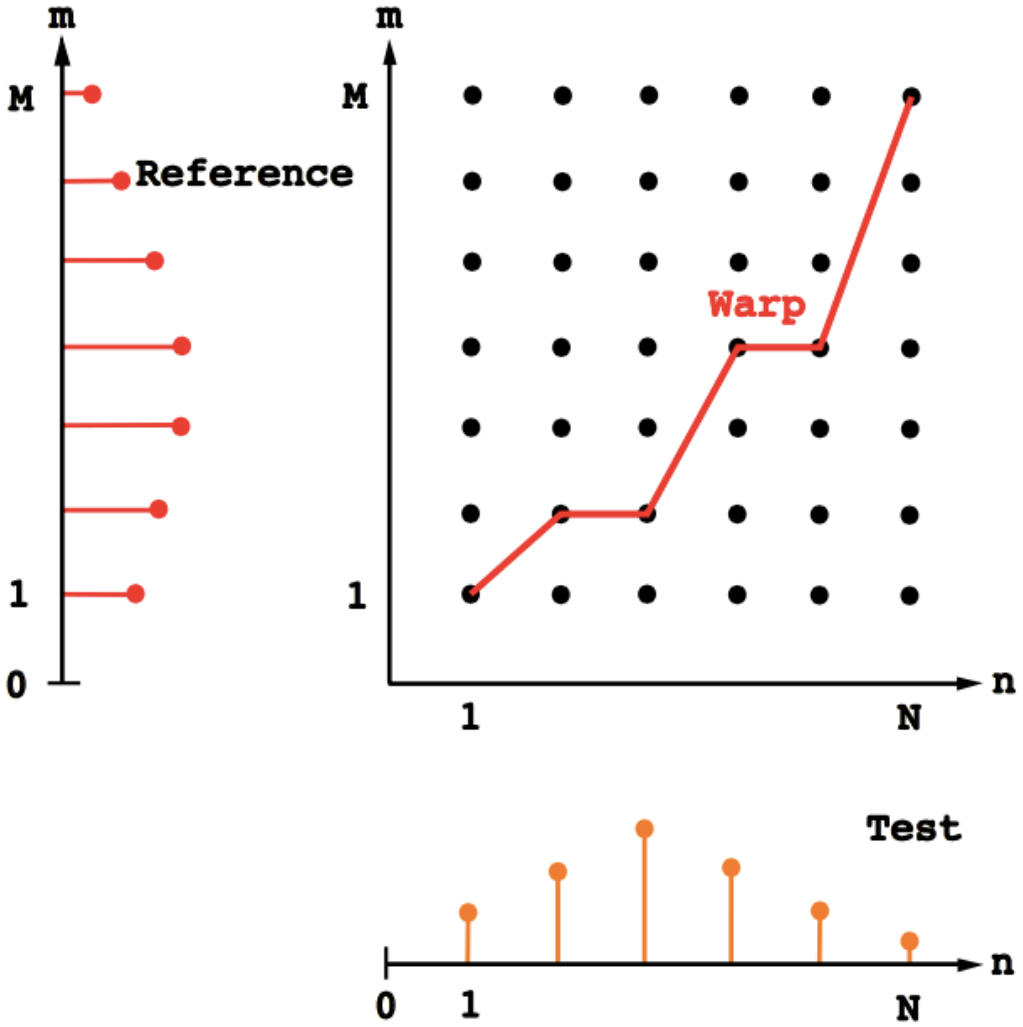
Mel filterbank

log

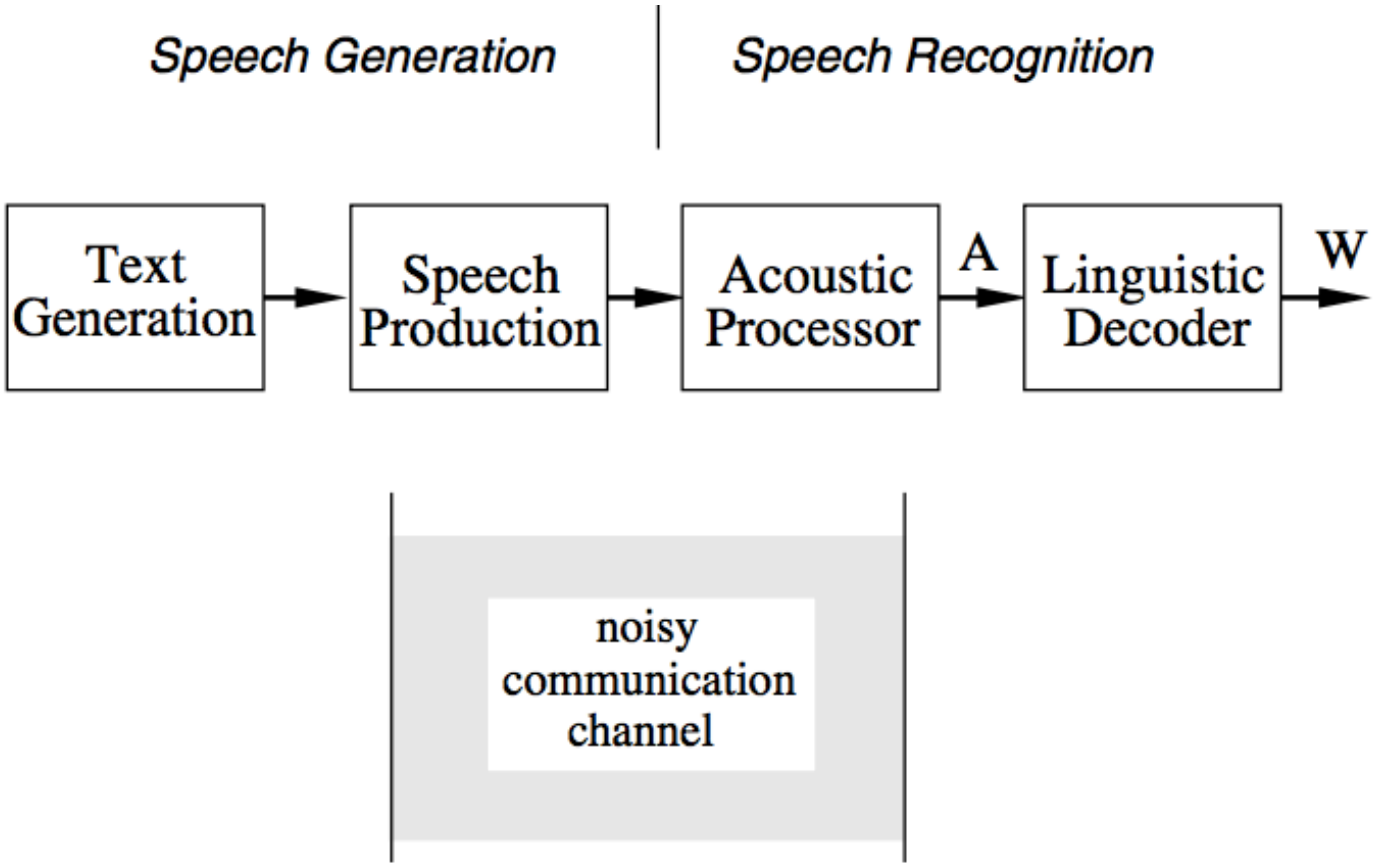
DCT

Dynamic Time Warping (DTW)

- A kind of dynamic programming for aligning things of different length



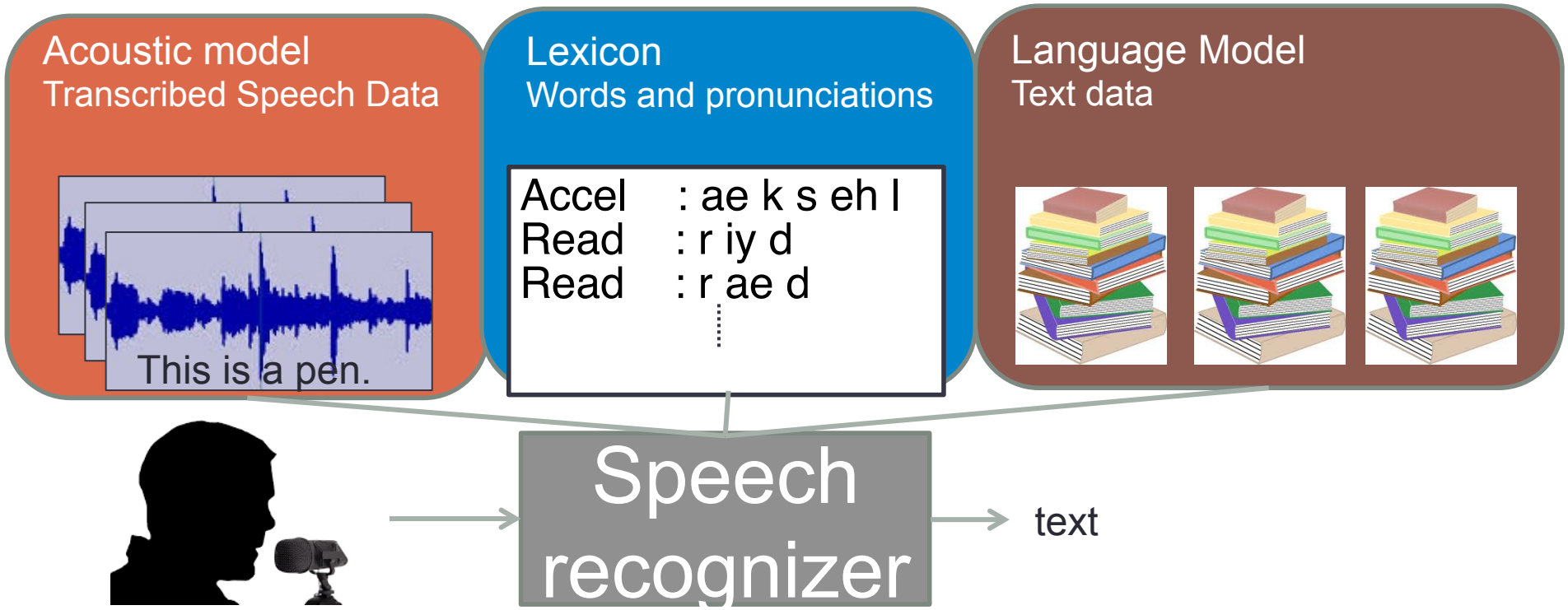
Information theoretic formulation



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

Probabilistic ASR formulation

- A search on the space of all possible words, W, and their pronunciation, L.
- Seek best path using dynamic programing or other graph search strategies



The ASR Equation

X - waveform, L - pronunciation, W - words

$$W^* = argmax_W P(W \mid X)$$

$$= argmax_W \frac{P(X \mid W)P(W)}{P(X)}$$

$$= argmax_W P(X \mid W)P(W)$$

$$P(X|W) = \sum_L P(X, L \mid W)$$

$$= \sum_L P(X \mid W, L)P(L \mid W)$$

$$= \sum_L P(X \mid L)P(L \mid W)$$

$$= argmax_W \sum_L P(X \mid L)P(L \mid W)P(W)$$

$$= argmax_{W,L} P(X \mid L)P(L \mid W)P(W)$$

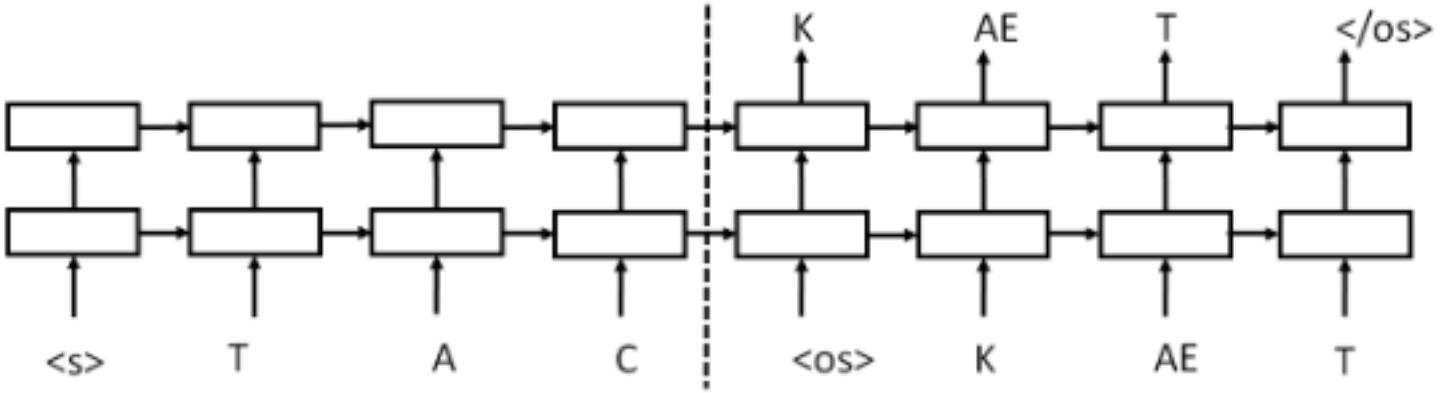
Lexical modeling (dictionary)

- How a word can be pronounced
- Describes words in terms of phonemes
- Provided by linguists, but can be learned by models (G2P or L2S)
  - Example: sequence to sequence models

```
กฎ          k o t^
กฎหมาย      k o t^ m aa j^
กฎหมายอาญา      k o t^ m aa j^ z aa j aa
กฎเกณฑ์     k o t^ k ee n^
กด          k o t^
กดขี่       k o t^ kh ii
กดดัน       k o t^ d a n^
กตัญญู      k a t a n^ j uu
กติกา       k a t i k aa
กทม.        k @@ th @@ m @@
กบฏ         k a b o t^
กมล         k a m o n^
กรกฎาคม     k a r a k^ k a d aa kh o m^
กรกฎาคม     k a r a k a d aa kh o m^
```

Seq2Seq G2P

- Machine translation-based approach
- Input character order is reversed (standard for MT)
- Use beamsearch on decoding side

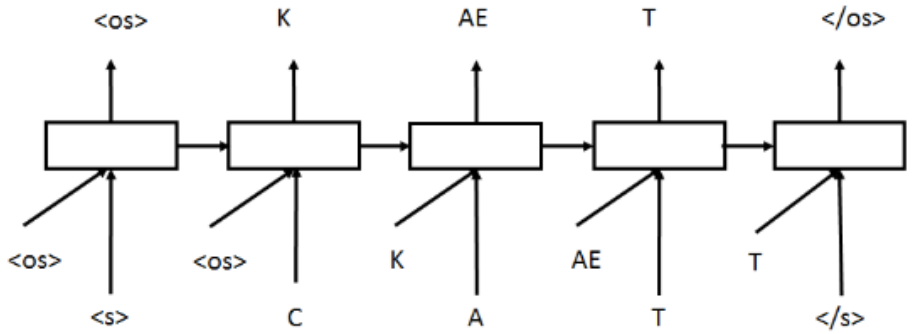


Encoder-decoder G2P

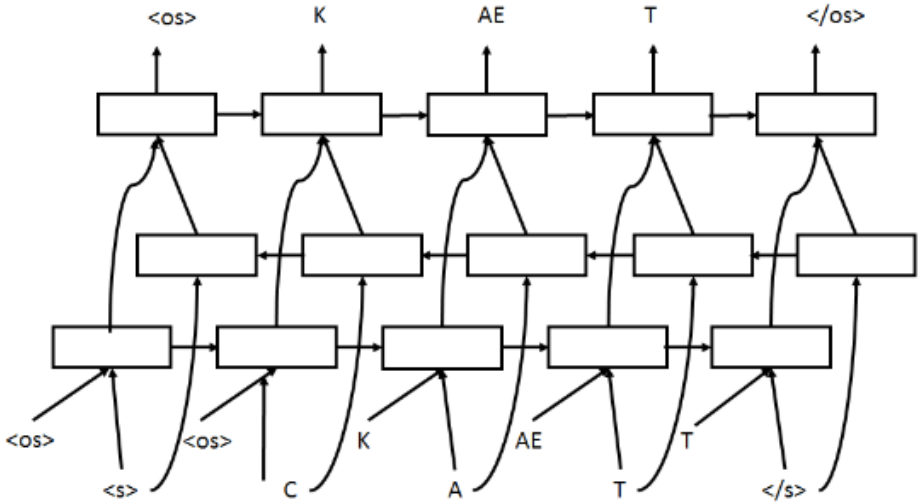
https://arxiv.org/pdf/1506.00196.pdf

Seq2Seq G2P

- Direct translation models



Uni-directional



Bi-directional

https://arxiv.org/pdf/1506.00196.pdf

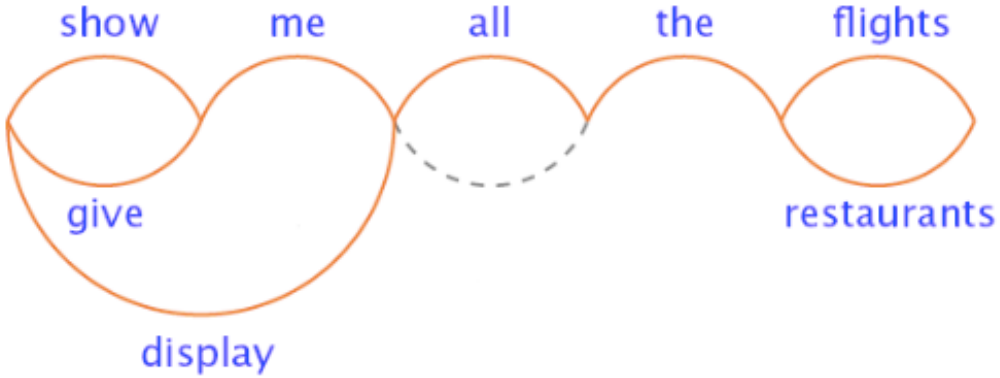
Results of G2P

Method	PER (%)	WER (%)
encoder-decoder LSTM	7.53	29.21
encoder-decoder LSTM (2 layers)	7.63	28.61
uni-directional LSTM	8.22	32.64
uni-directional LSTM (window size 6)	6.58	28.56
bi-directional LSTM	5.98	25.72
bi-directional LSTM (2 layers)	5.84	25.02
bi-directional LSTM (3 layers)	5.45	23.55

https://arxiv.org/pdf/1506.00196.pdf

Language Model

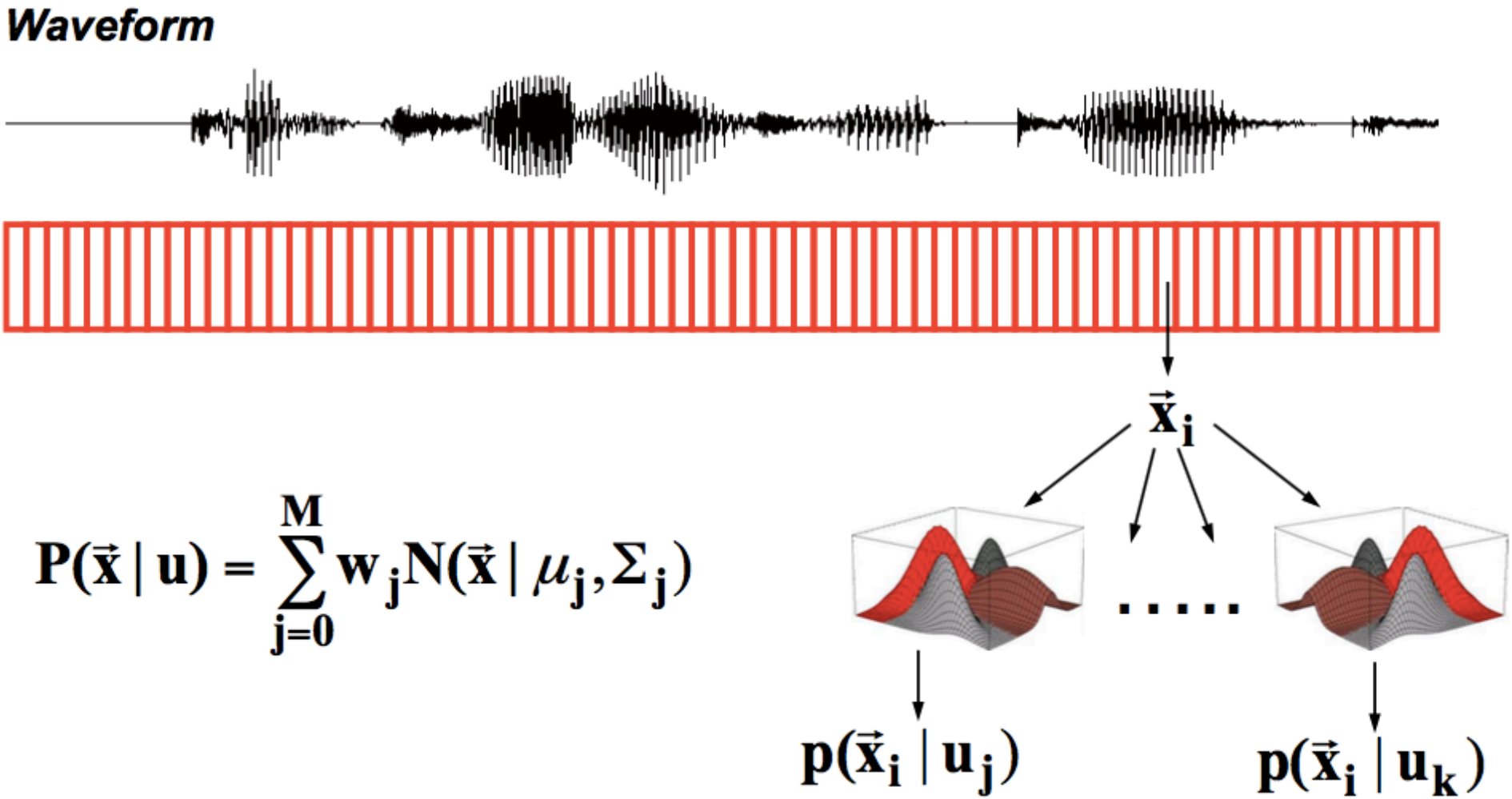
- Very important for ASR
  - “Please write a letter right now to Mrs. Wright. Tell her that two is too many to buy.”
  - “How to wreck a nice beach” vs “How to recognize speech”
- Can use n-grams (tri-gram is most popular)
- Or CFG for simple tasks
- Recent work have incorporate neural LM (used for post-processing)



0.036 the
0.026 a
0.018 of
…
0.007 good
…
0.005 day
…
0.003 morning
…

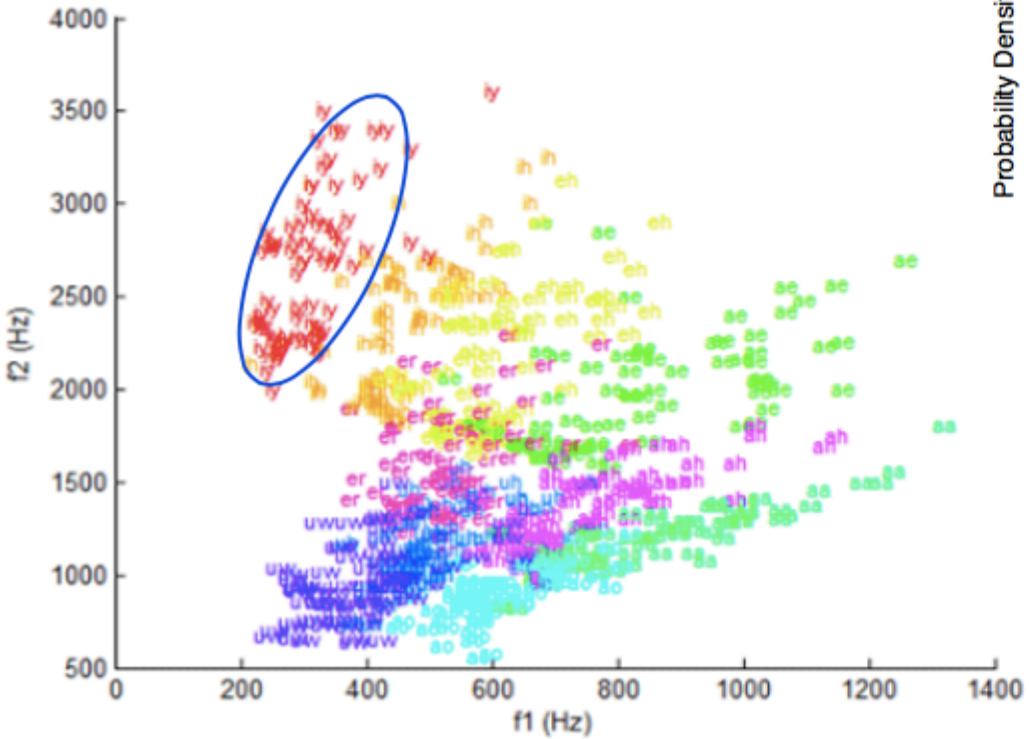
…
0.011 a good
0.003 a morning
…
0.086 good morning
0.026 good day
…
0.149 of a
0.057 of day
…

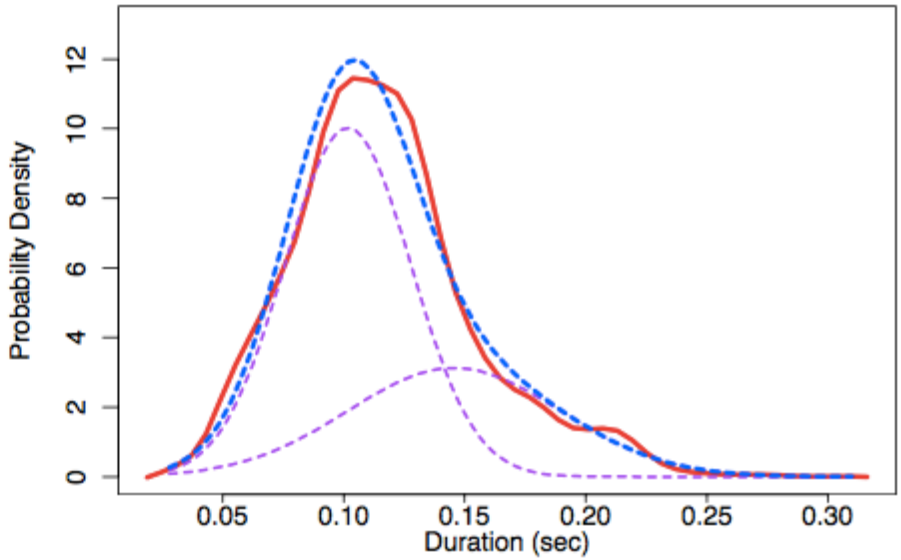
Acoustic modeling

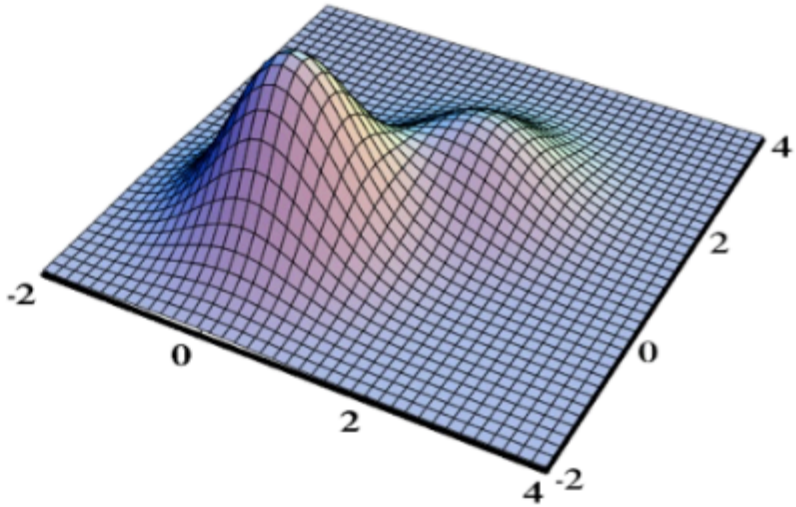


$$P(\vec{x} \mid u) = \sum_{j=0}^{M} w_j N(\vec{x} \mid \mu_j, \Sigma_j)$$

Gaussian Mixture Models

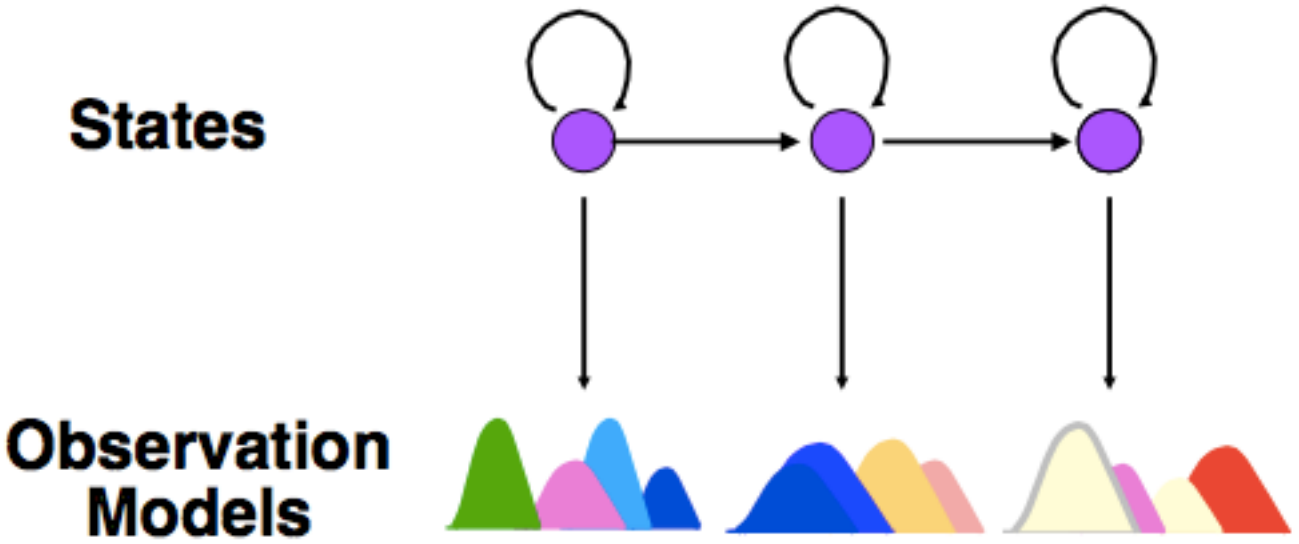




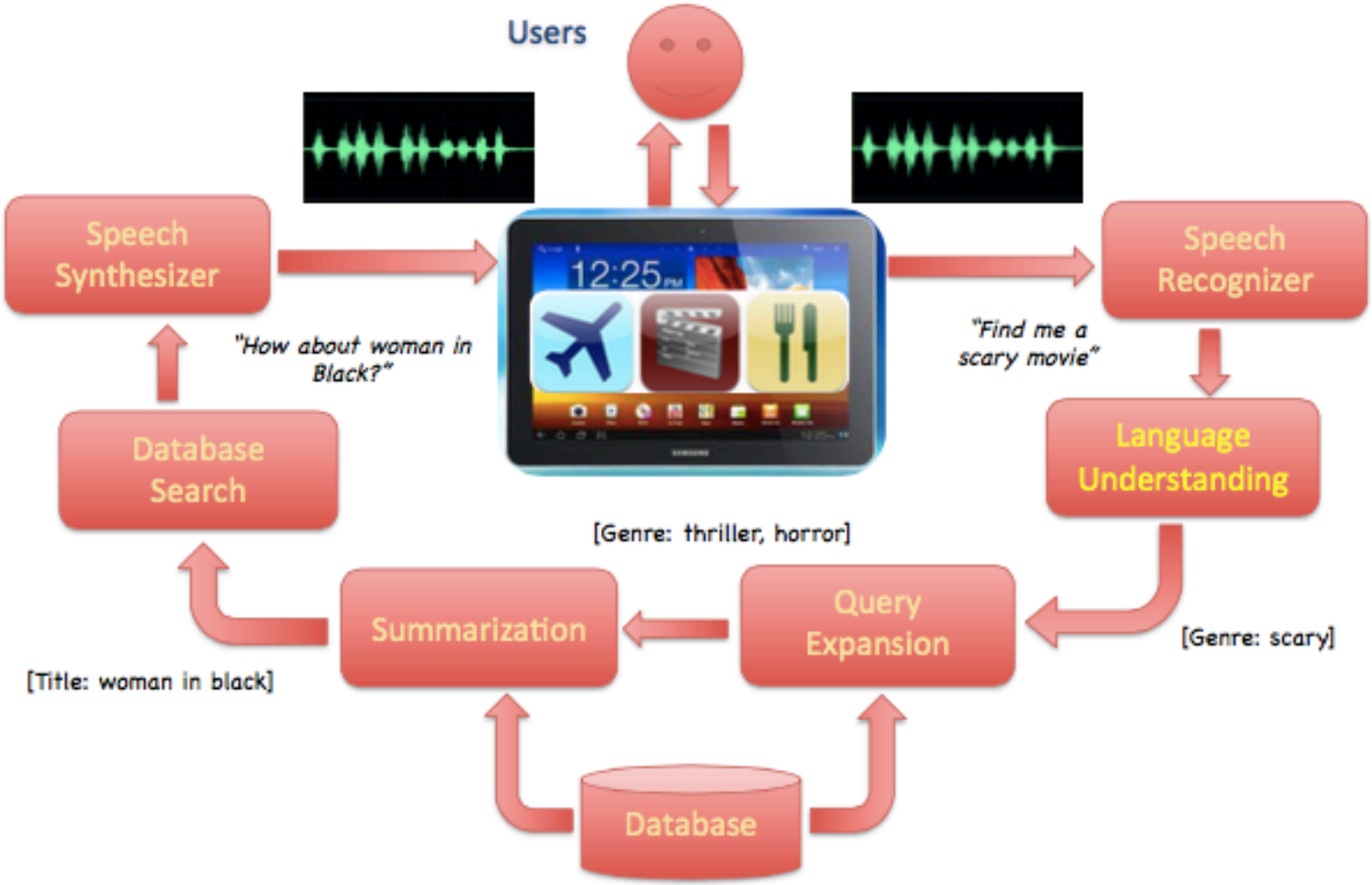


Hidden Markov Models (HMM)

- Dominant framework for ASR
- Model phonemes as hidden states
- Outputs are MFCCs observations
- Unlike PoS tag, the emission probability is continuous rather than discrete

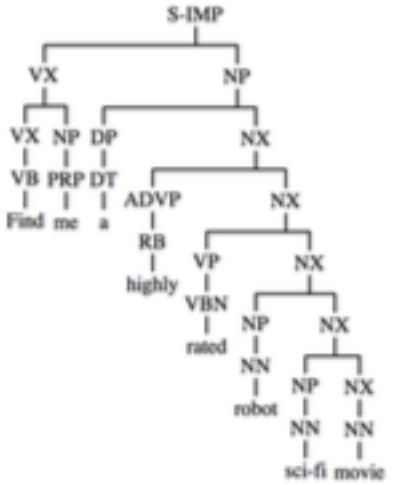


Spoken dialogue system



Language understanding

- **Syntactic Understanding**
  - Hierarchical parse tree
    - E.g., "Find me a highly rated robot sci-fi movie"



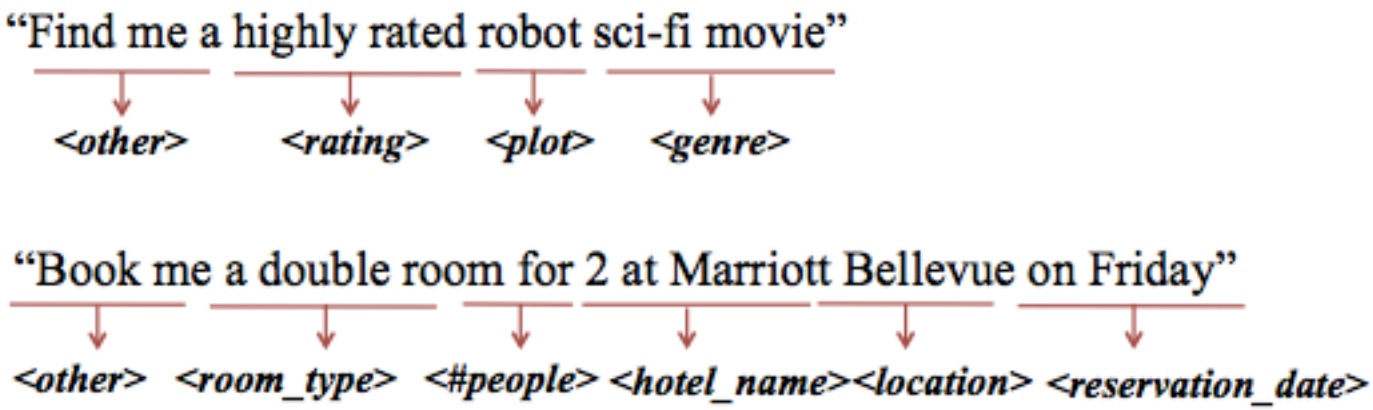
- **Semantic Understanding**
  - A sequence labeling task
  - Map a sequence of words to a sequence of concepts

"Find me a highly rated robot sci-fi movie"

<other> *<rating>* *<plot>* *<genre>*

Semantic tagging

- Joint segmentation/classification problem
  - Segment query constituents and classify the segments into semantic classes



CRF cookbook for language understanding

- Key ingredients
  - Domain
    - E.g., movie, flight, restaurant, weather…
  - Semantic classes

Domain	Semantic classes
Flight	General city, General date, General time, Departure city, Departure date, Departure time, Arrival city, Arrival date, Arrival time, Return date, Return time, Transit city, Airline
Restaurant	Goal, Restaurant name, Amenity, Cuisine, Dish, Hours, Location, Price, Rating
Movie	Title, Viewers' rating, Year, Genre, Director, MPAA rating, Plot, Actor, Trailer, Song, Review, Character

CRF cookbook for language understanding

- Key ingredients
  - Models
    - CRFs
    - Semi-CRFs
  - Features
    - Transit features
    - Lexical features (e.g., *n*-grams in training data)
    - Regular expression features (e.g., time, date, numbers)
    - Semantic features with lexicons (e.g., list of restaurants, movie titles, cities)
    - Linguistic features (e.g., segment length, POS tagging)

CRF cookbook for language understanding

- Key ingredients
  - Data
    - Natural language queries
    - Semantic labels

Domain	Movie	Flight
Query	what is the 1959 american thriller film directed by alfred hitchcock and starring cary grant and eva marie saint	a flight to covington leaving next monday around 2 p m from pittsburgh i prefer no red eye flights
Labels	what is the | Other 1959 | Release Year american thriller | Genre film directed by | Other alfred hitchcock | Director and starring | Other cary grant | Actor and | Other eva marie saint | Actor	a flight to | Other covington | Arrival City leaving | Other next Monday | Departure Date around 2 p m | Departure Time from | Other pittsburgh | Departure City i prefer | Other no red eye flights | Preference