
Speech Recognition

Mitch Marcus

A Sample of Speech Recognition

Today's class is about:

First, **Weiss** speech recognition is difficult. As you'll see, the impression we have speech is like beads on a string is just wrong.

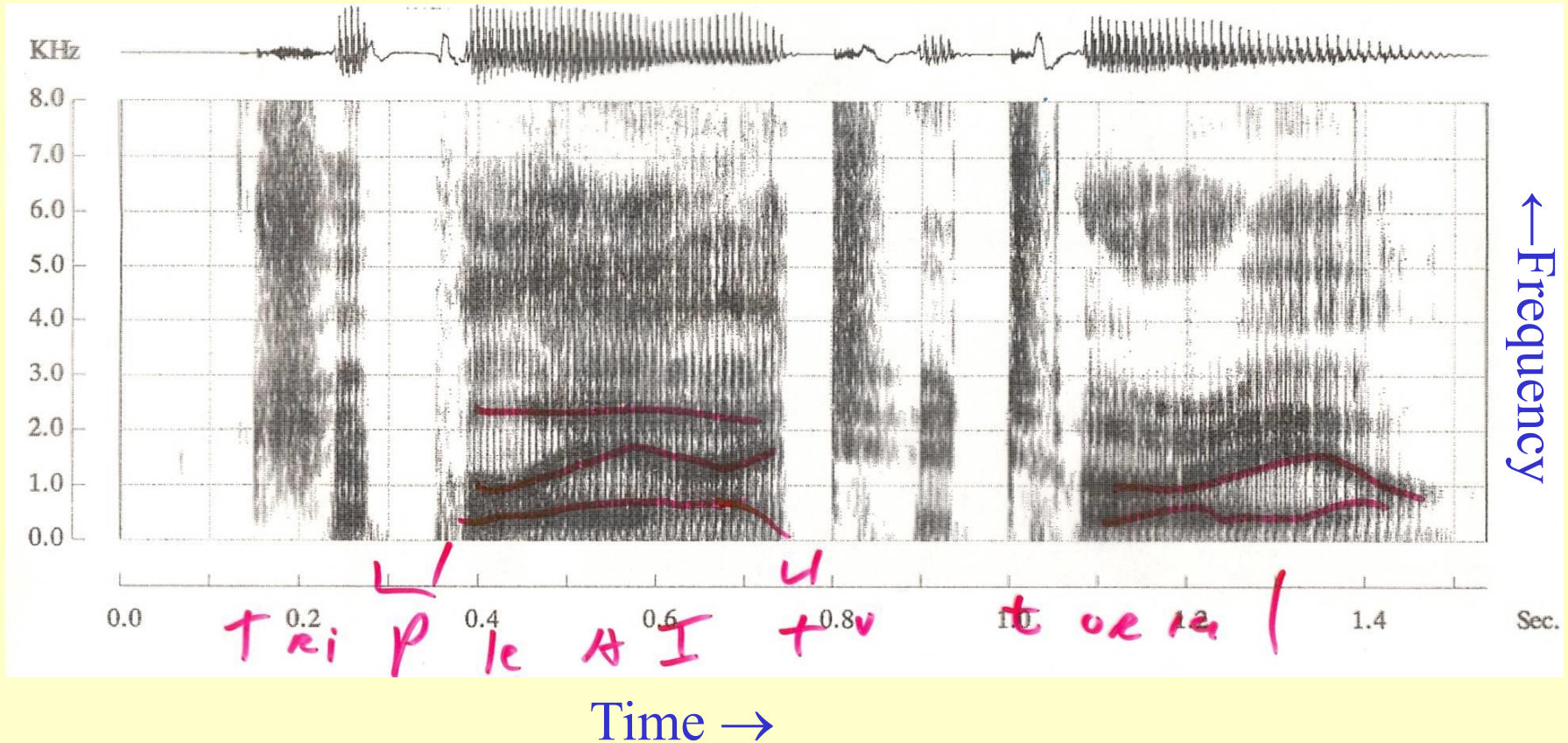
Second we will look at how hidden Markov models are used to do speech recognition.

And finally, we will look at how the speech dialogue technology behind systems like Siri might be configured.

This was dictated **one** November 11, 2016, into the email app on my iPhone.

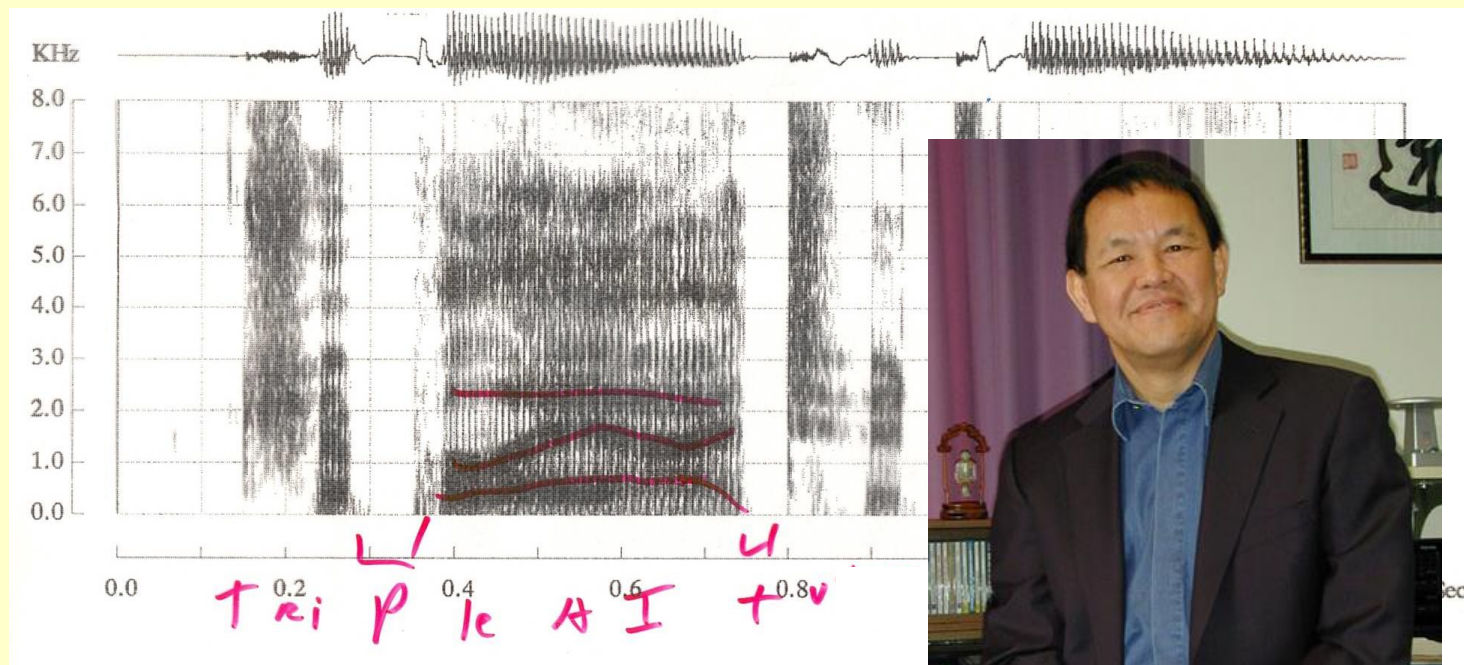
I. Why is Speech Recognition Hard??

A Speech Spectrogram



- Represents the varying short term amplitude spectra of the speech waveform
- Darkness represents amplitude at that time & frequency.

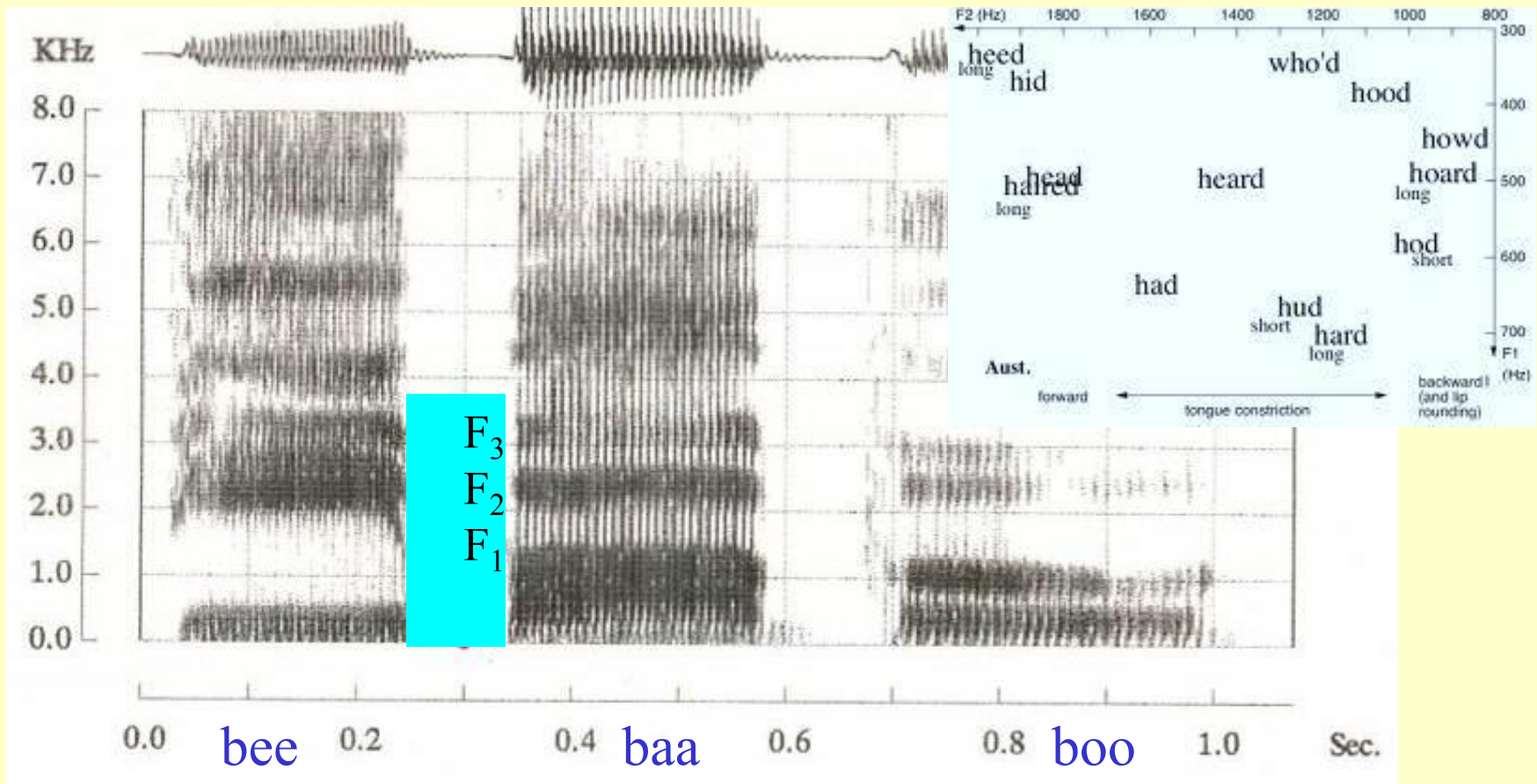
A trained person can “read” a spectrogram



Therefore, the spectrogram contains all the information a machine needs as well....

Prof. Victor Zue, MIT

Vowels are determined by their *formants*

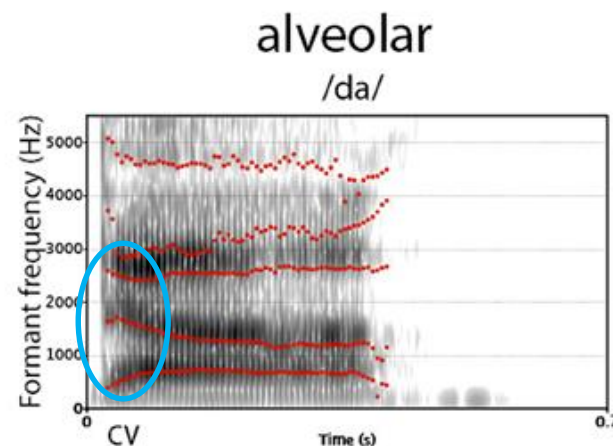
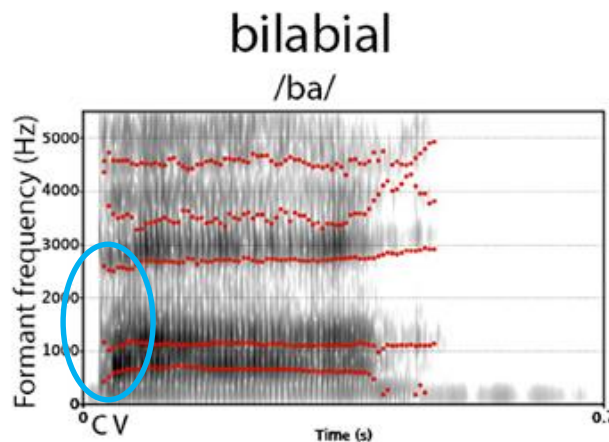


The frequencies of F_1 , F_2 , and F_3 – the first three resonances of the vocal tract – largely determine the perceived vowel

Consonants are determined by (*inter alia*):

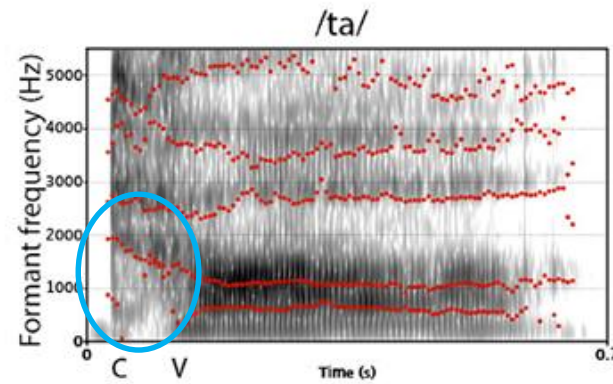
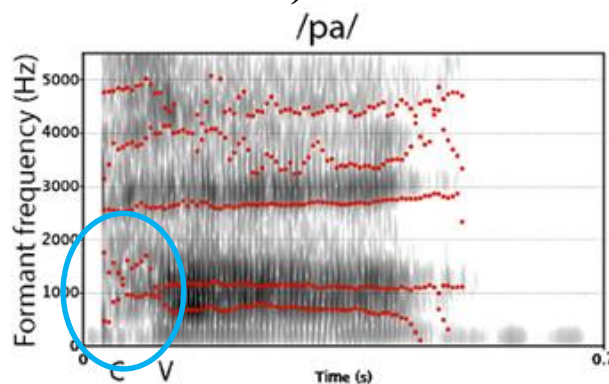
Formant motion

short VOT



Length of Silence (“Voice Onset Time”)

long VOT

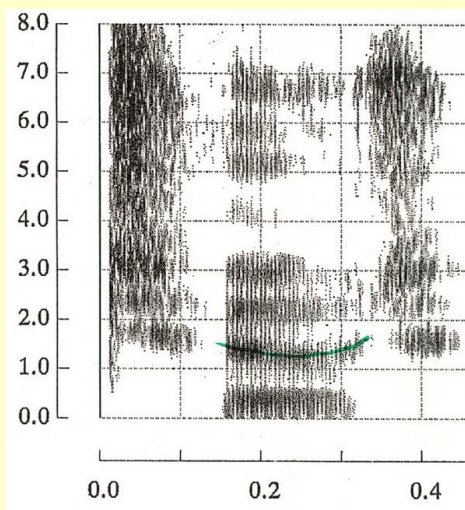


http://www.frontiersin.org/files/Articles/76444/fpsyg-05-00549-HTML/image_m/fpsyg-05-00549-g001.jpg

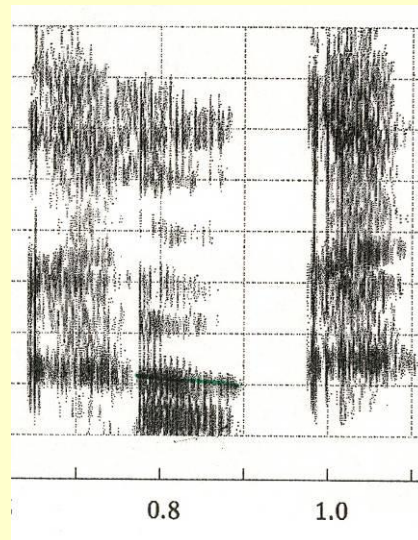
Coarticulation

- The same abstract phoneme can be realized very differently in different phonetic contexts: *coarticulation*
- F_2 in the vowel /u/, crucial to its identification, varies significantly due to surrounding consonants in the syllables:

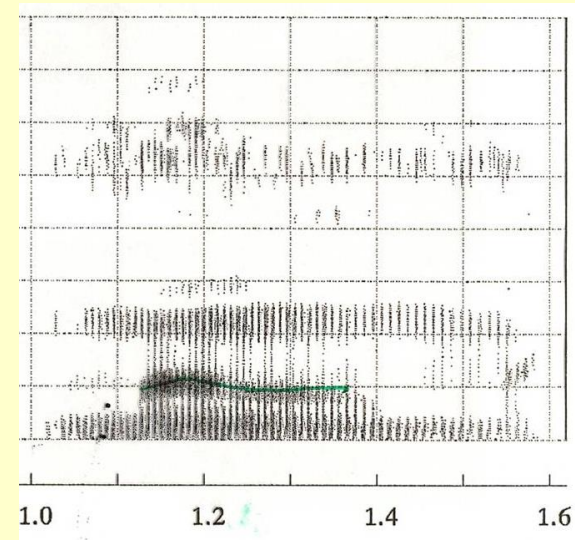
Context	F_2 (kHz)
“kook”	1.0
“moom”	0.8-1.0
“toot”	1.2



Toot



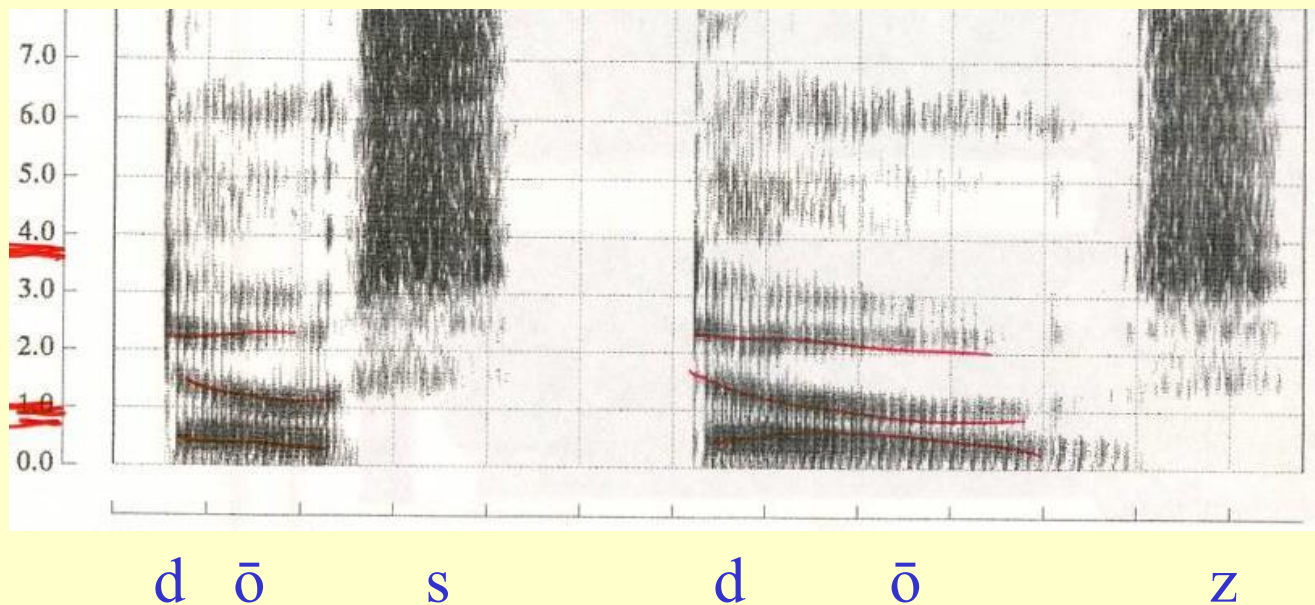
Kook



Moom

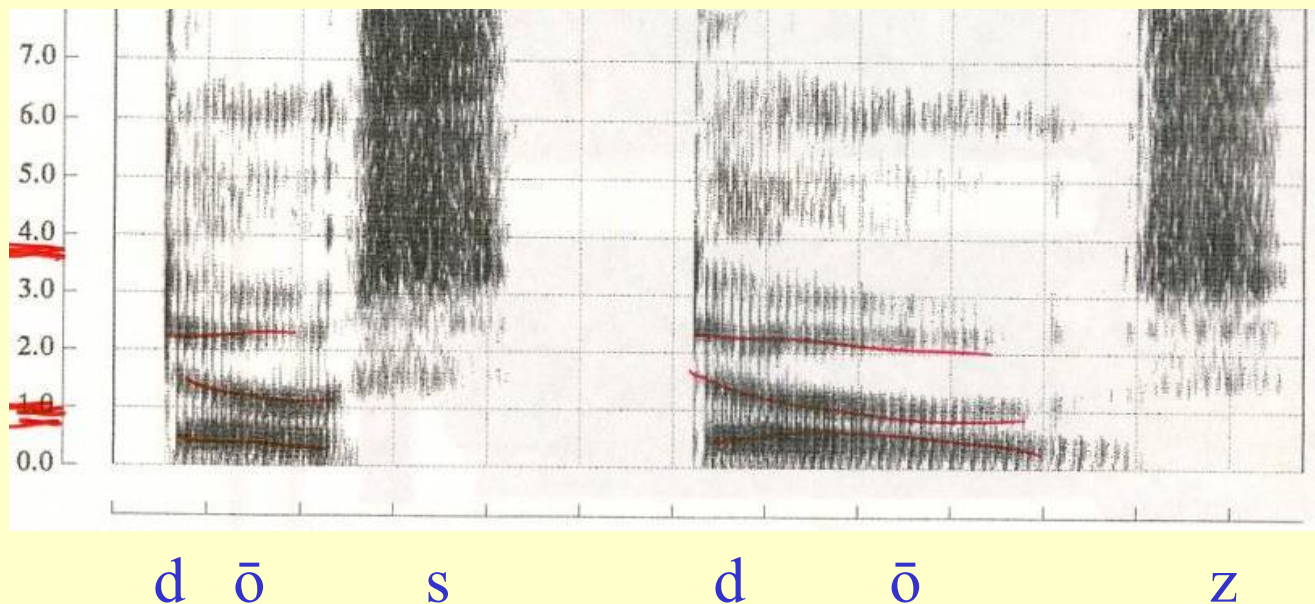
Speech Information is not local

- The identity of speech units, *phones*, cannot be determined independently of context.
- Sometimes two phones can best be distinguished by examining properties of neighboring phones:



Speech Information is not local

- /s/ and /z/ are often acoustically identical...
- They are differentiated by the length of the preceding vowel:

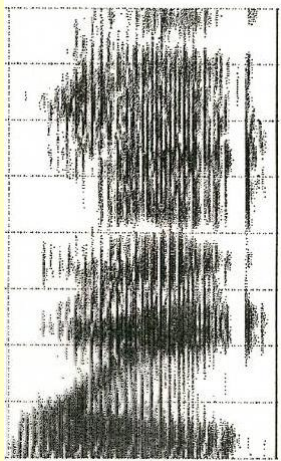


Words are constant, but utterances aren't

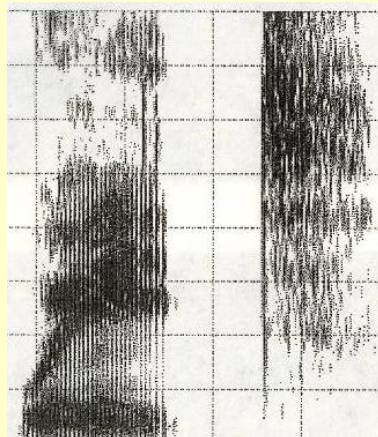
Spectrograms of *similar* words pronounced by the *same* speaker

may be more alike than

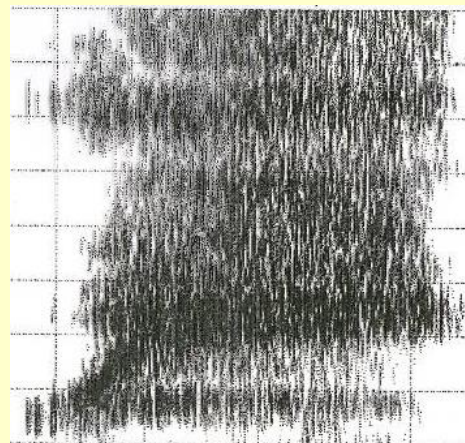
Spectrograms of the *same* word pronounced by *different* speakers.



"wait" – MM (m)



"wait" – JH (f)

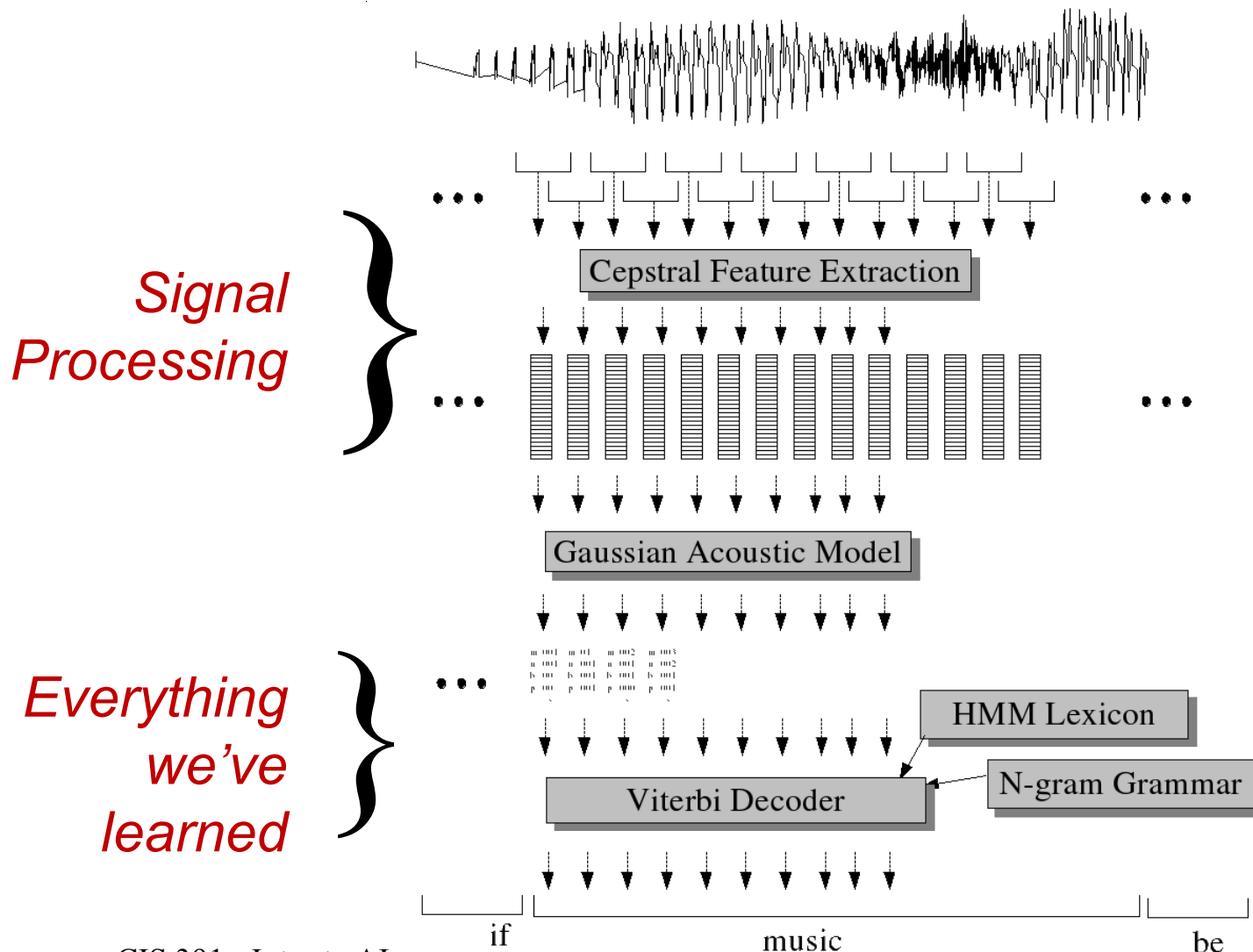


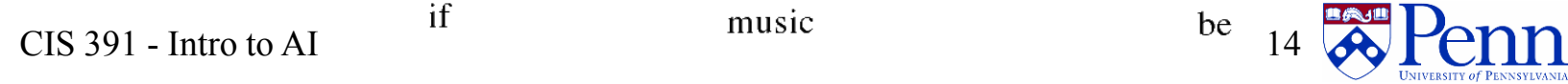
"wait" – whispered(MM)

II. HMMs for Speech Recognition

**(Illustrations in II from Chapter 9,
Jurafsky & Martin)**

Speech Recognition Architecture

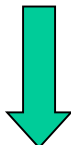
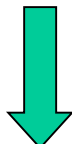




Speech recognition via Bayes Rule!

$$\hat{W} = \arg \max_{W \in L} P(\textit{Signal} | W) P(W)$$

likelihood prior

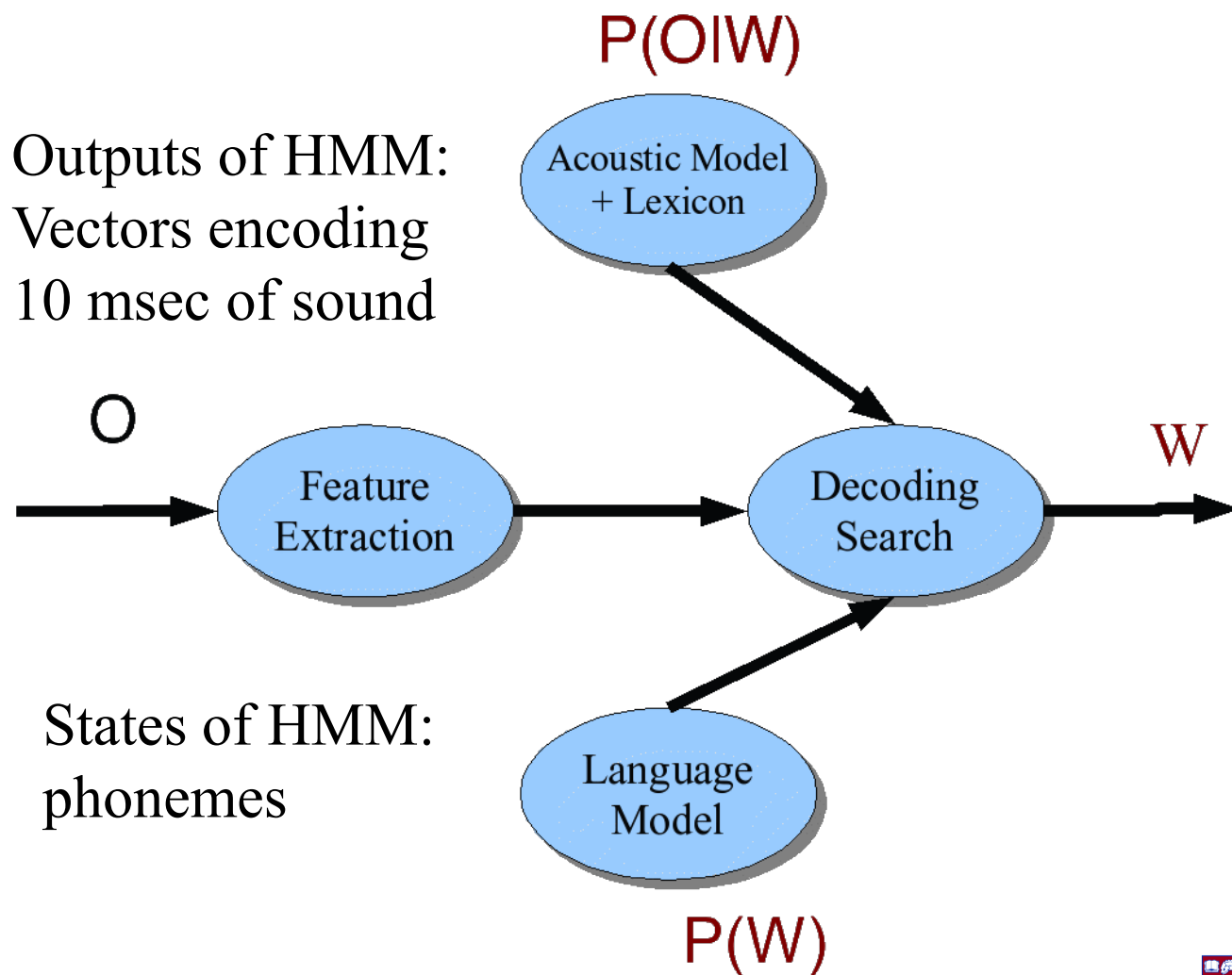
Where: W is a (text) string from a Source
 \textit{Signal} is the (speech) output from a
“noisy channel”

The noisy channel model: another view of HMMs

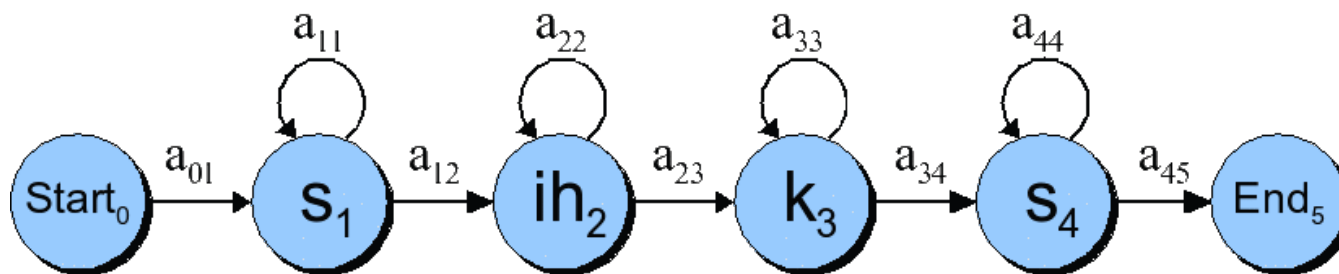
- Ignoring the denominator leaves us with two factors: $P(\text{Source})$ and $P(\text{Signal}|\text{Source})$



Speech Architecture meets Noisy Channel

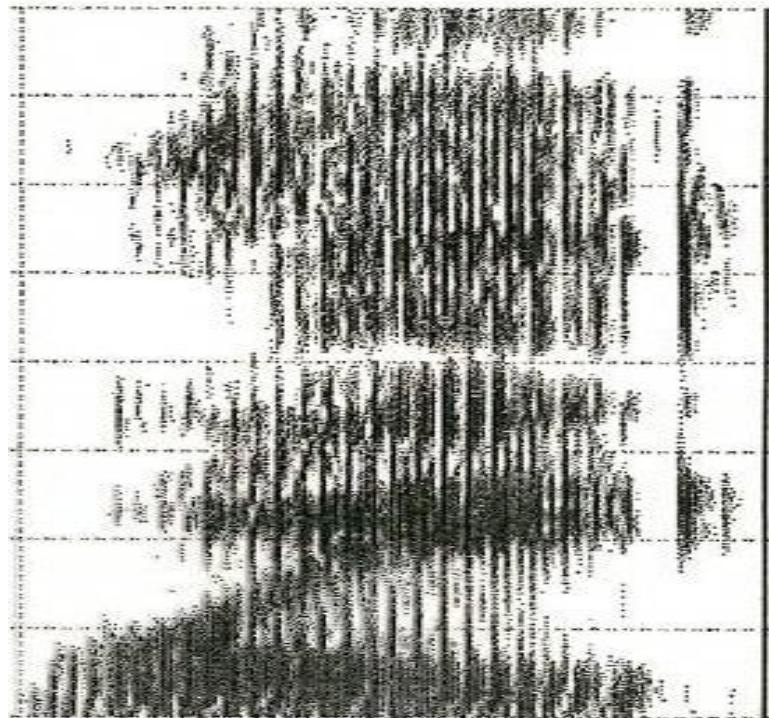


Schematic HMM for the word *six*



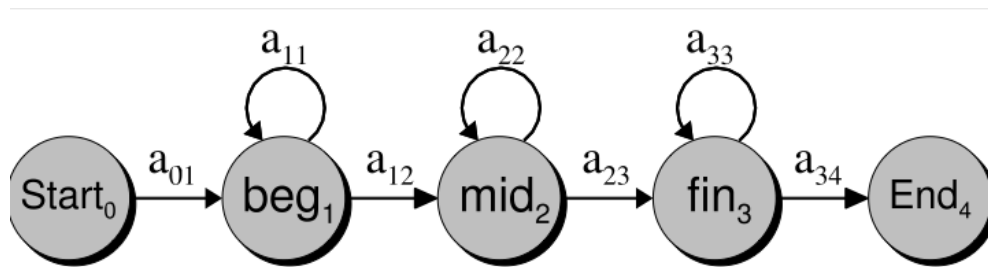
- Simple one state per phone model
- Left to right topology with self loops and no skips
- Start and End states with no emissions
- States output 10 msec spectral slices or DNN vectors

Phones have dynamic structure

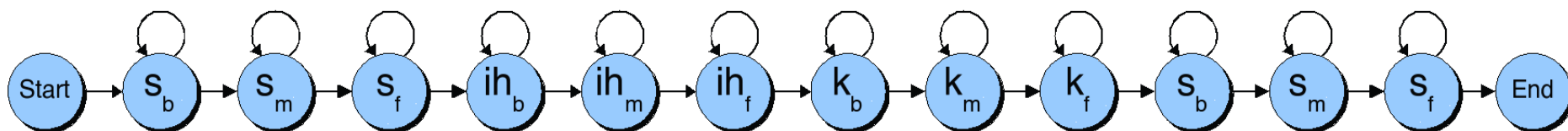


- ***Wait*** (said by Mitch Marcus), pronounced [w ey t]
- The formants of the diphthong ey move continually
- ***T*** consists of (a) a silence, (b) a burst

A 3-state HMM phone model

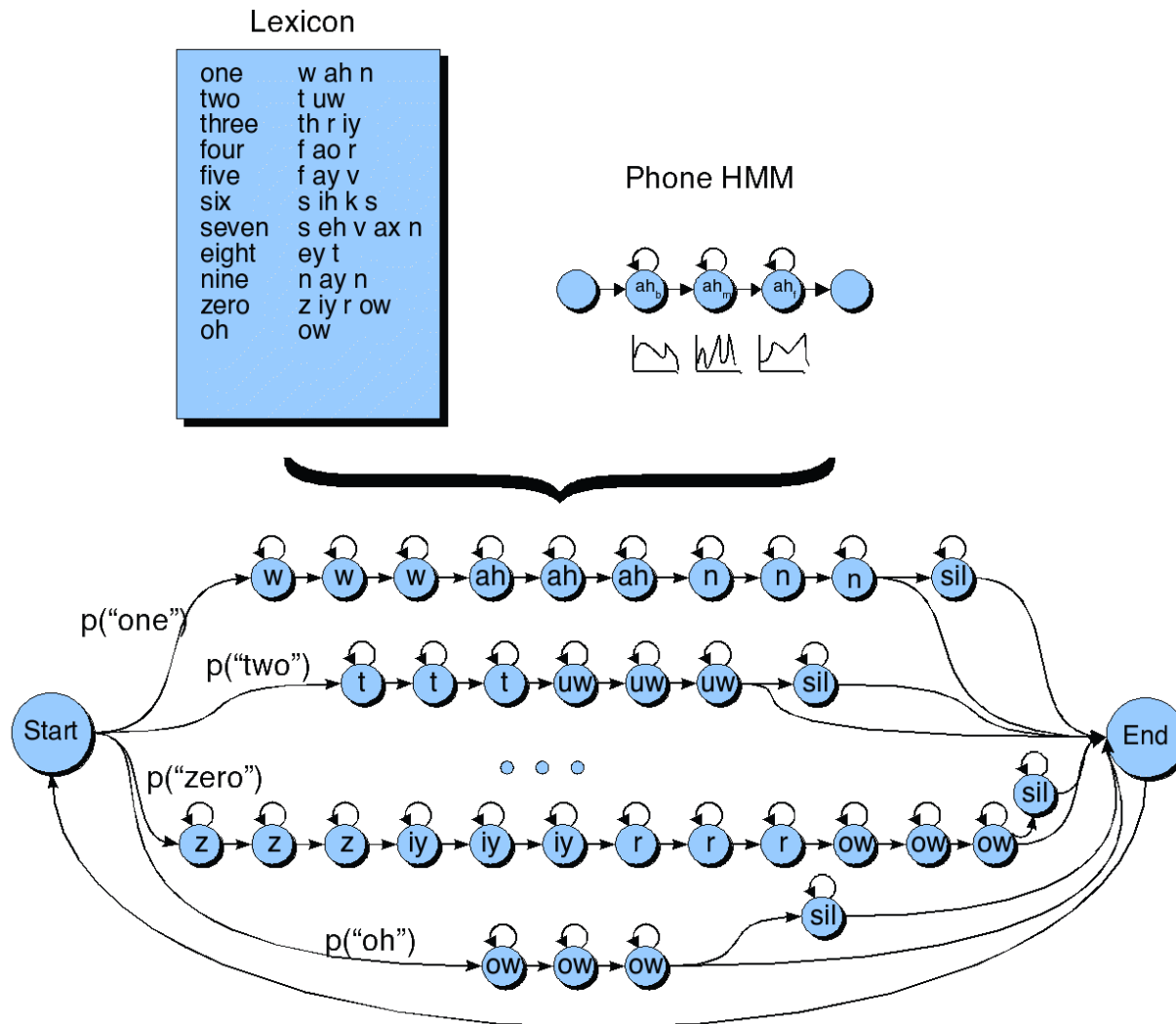


- Three emitting states
- Two non-emitting states
- Usually includes skip states



- The word *six* [siks] using 3-state HMM phone models

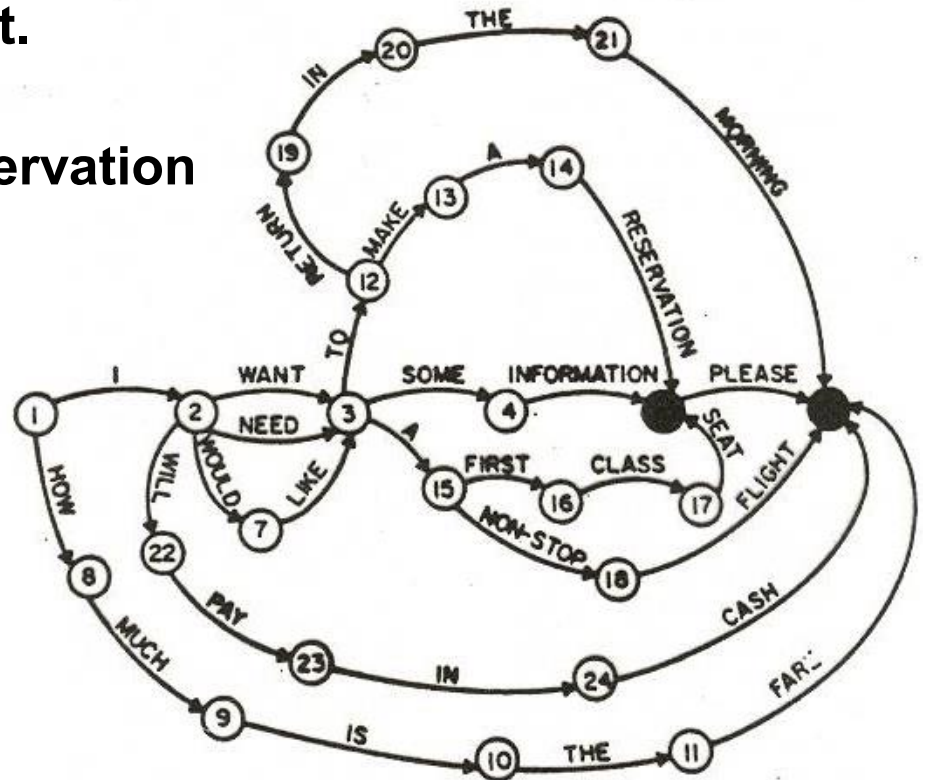
A simple full HMM for digit recognition



III. Speech Dialogue Understanding

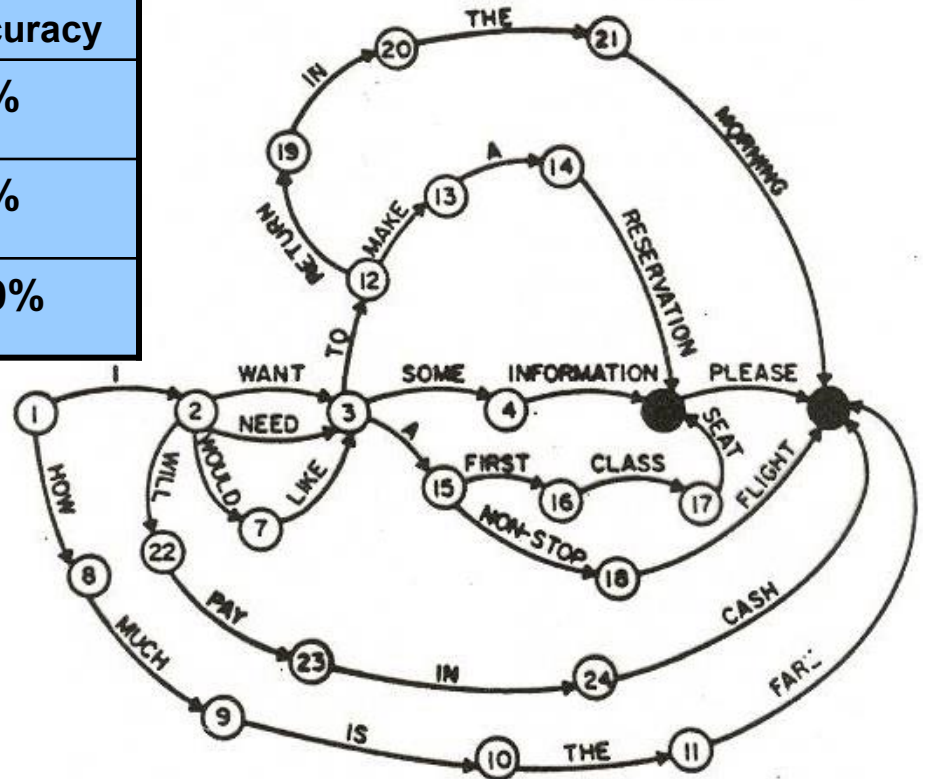
Multiple knowledge sources provide redundancy

- **Grammatical, semantic and pragmatic information can be used to make recognition robust.**
- **A first experiment:
AT&T Bell Labs airline reservation system
(Levinson-1977)**



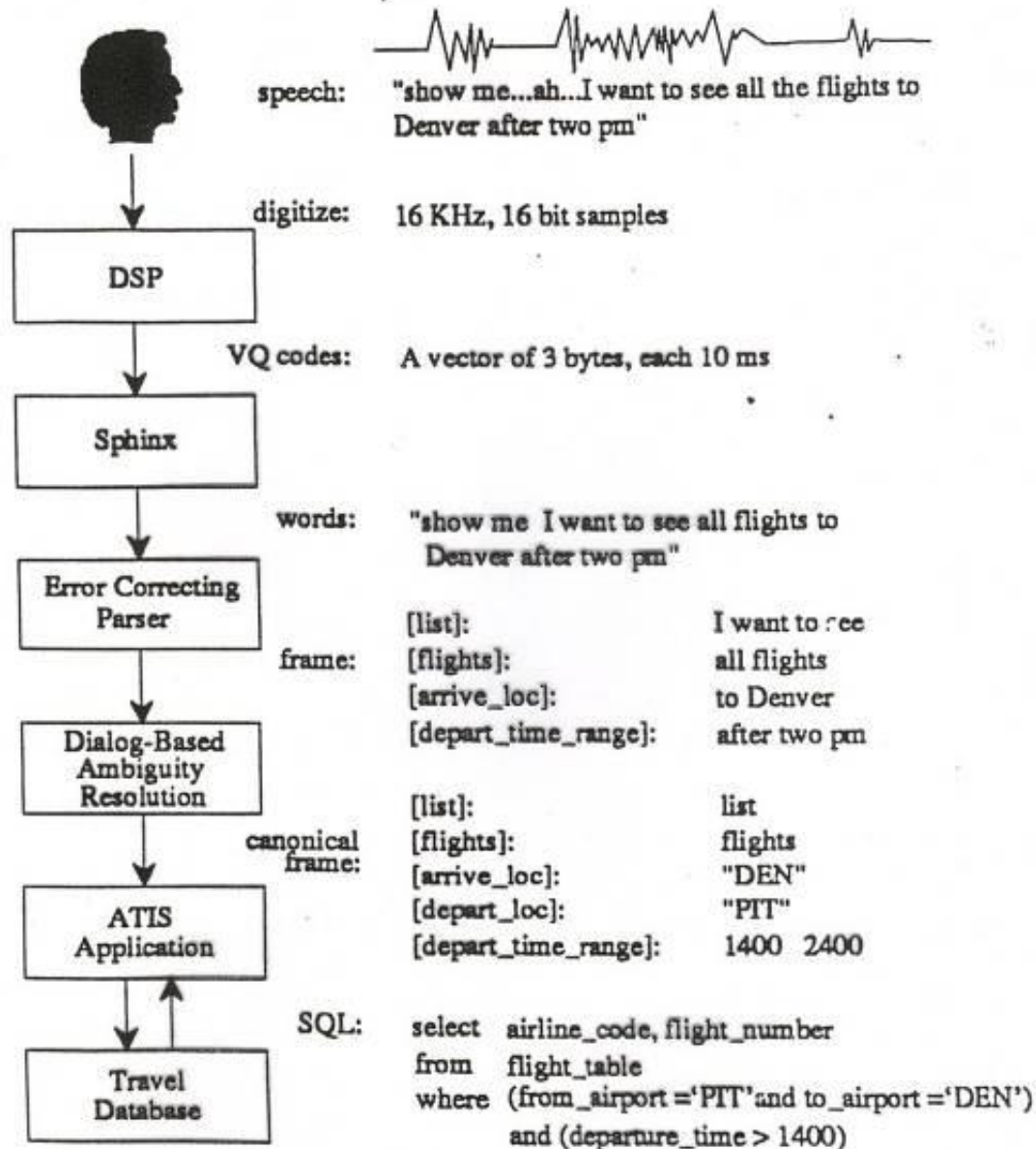
redundancy

Results for 351 test sentences			
Processing level	Sentences correct	Errors detected	Word Accuracy
Acoustic	Na	0	88%
Syntactic	330	0	99%
Pragmatic	345	6	>99%



Structure of Phoenix

A Spoken Language Understanding System



Speech Recognition: Task Dimensions

- **Speaker Dependent, Independent, Adaptive**
 - Speaker dependent: System trained for current speaker
 - Speaker independent: No modification per speaker
 - Speaker Adaptive: adapts an initial model to speaker
- **Read vs. dictation vs. conversational**
- **Quiet Conditions vs. various noise conditions**
- **Known microphone vs. unknown microphone**
- **Perplexity level**
 - Low perplexity: Average expected branching factor of grammar < 10 -20
 - High perplexity: Average expected branching factor of grammar > 100

Perplexity (average branching factor of LM): Why it matters

- **Experiment (1992): read speech, Three tasks**

- Mammography transcription (*perplexity 60*)
“There are scattered calcifications with the right breast”
“These too have increased very slightly”
- General radiology (*perplexity 140*)
“This is somewhat diffuse in nature”
“There is no evidence of esophageal or gastric perforation”
- Encyclopedia dictation (*perplexity 430*)
“Czechoslovakia is known internationally in music and film”
“Many large sulphur deposits are found at or near the earth’s surface”

Task	Vocabulary	Perplexity	Word error
Mammography	837	66	3.4%
Radiology	4447	141	5.8%
Encyclopedia	3021	433	14.6%