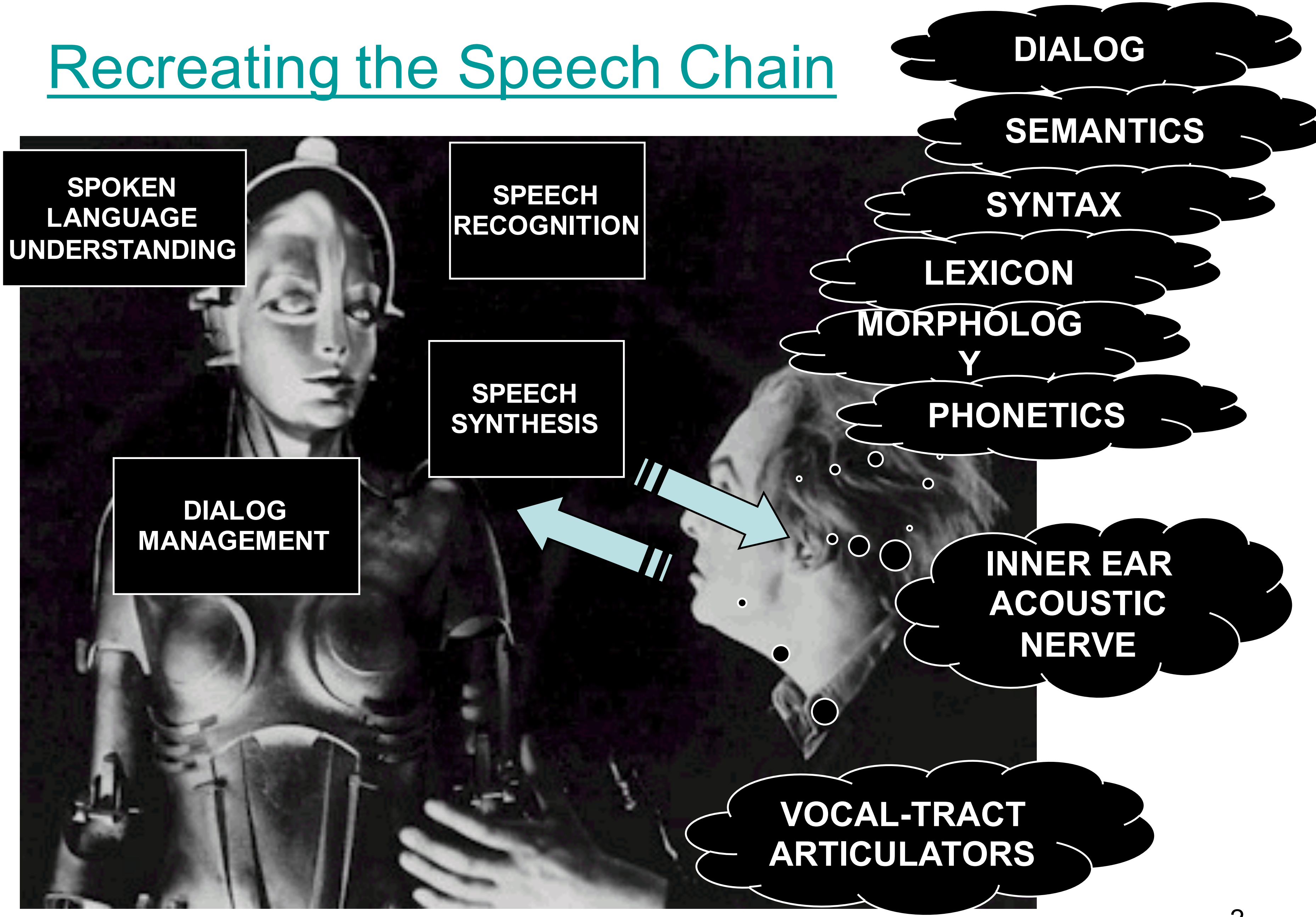
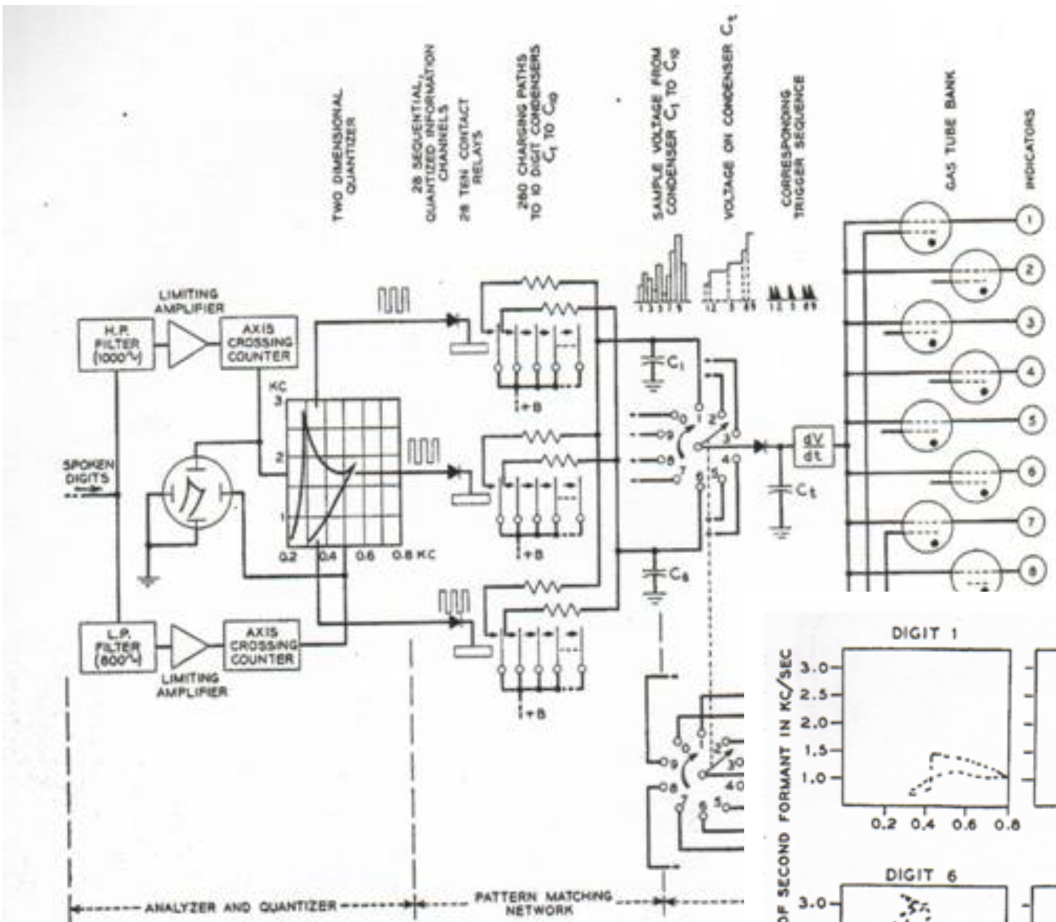
Automatic Speech Recognition

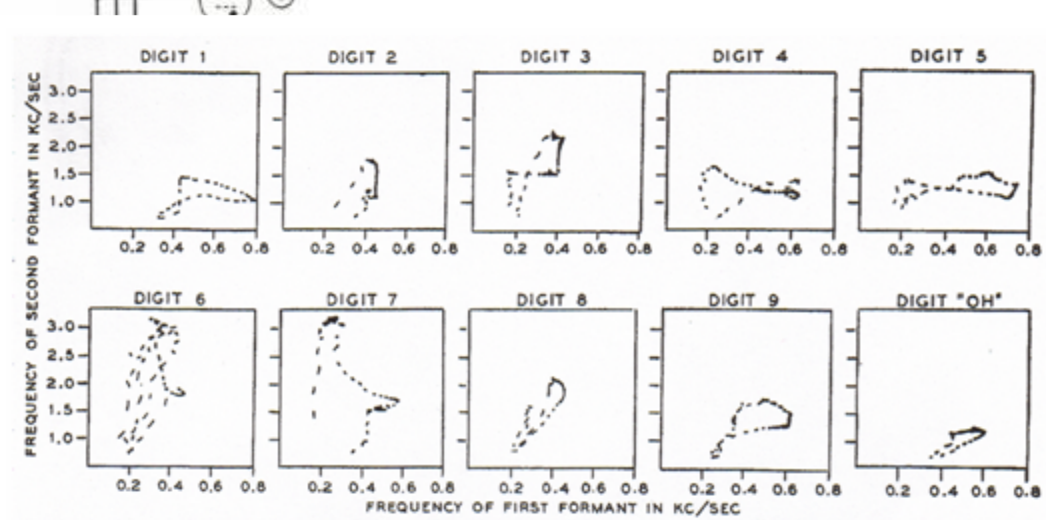
Many Slides from Julia Hirschberg

Recreating the Speech Chain



Speech Recognition: the Early Years





- 1952 – Automatic Digit Recognition (AUDREY)
  - Davis, Biddulph, Balashek (Bell Laboratories)

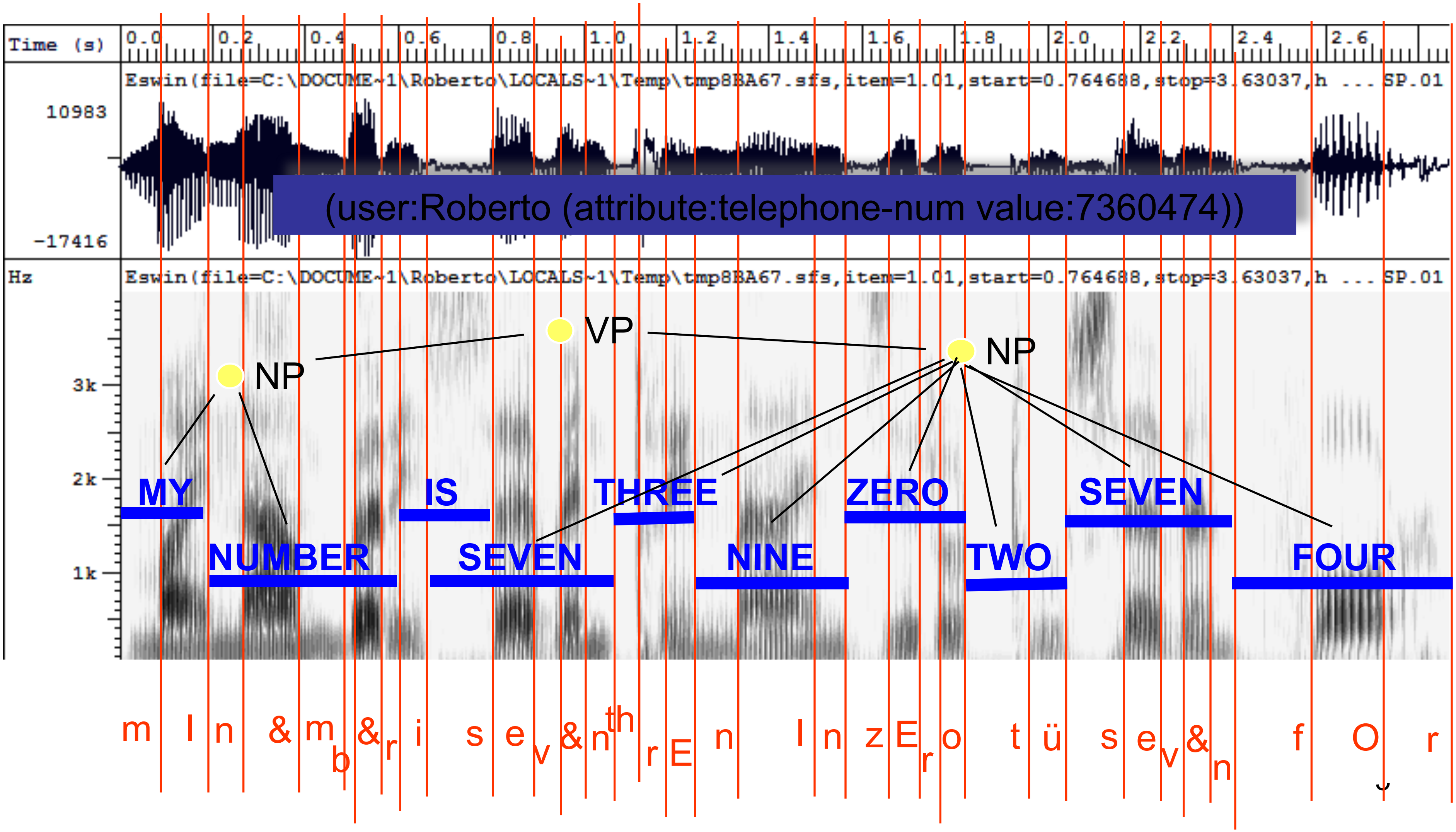
1960’s – Speech Processing and Digital Computers

- AD/DA converters and digital computers start appearing in the labs


James Flanagan
Bell Laboratories

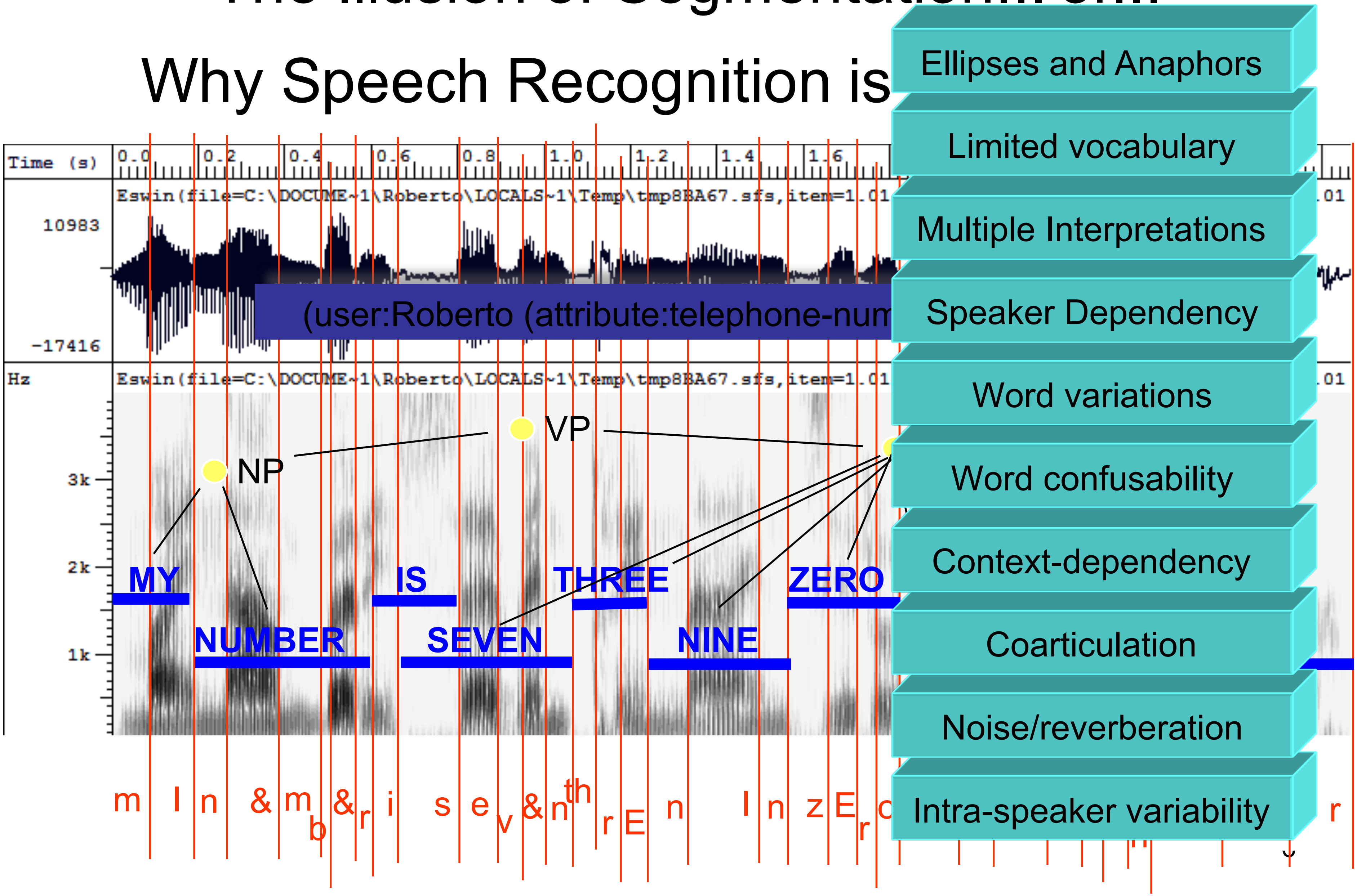
The Illusion of Segmentation... or...

Why Speech Recognition is so Difficult



The Illusion of Segmentation... or...

Why Speech Recognition is



1969 – Whither Speech Recognition?

General purpose speech recognition seems far away. Social-purpose speech recognition is severely limited. It would seem appropriate for people to ask themselves why they are working in the field and what they can expect to accomplish…

It would be too simple to say that work in speech recognition is carried out simply because one can get money for it. That is a necessary but not sufficient condition. We are safe in asserting that speech recognition is attractive to money. The attraction is perhaps similar to the attraction of schemes for turning water into gasoline, extracting gold from the sea, curing cancer, or going to the moon. One doesn't attract thoughtlessly given dollars by means of schemes for cutting the cost of soap by 10%. To sell suckers, one uses deceit and offers glamour…

Most recognizers behave, not like scientists, but like mad inventors or untrustworthy engineers. The typical recognizer gets it into his head that he can solve "the problem." The basis for this is either individual inspiration (the "mad inventor" source of knowledge) or acceptance of untested rules, schemes, or information (the untrustworthy engineer approach).

The Journal of the Acoustical Society of America, June 1969

J. R. Pierce
Executive Director,
Bell Laboratories

1971-1976: The ARPA SUR project

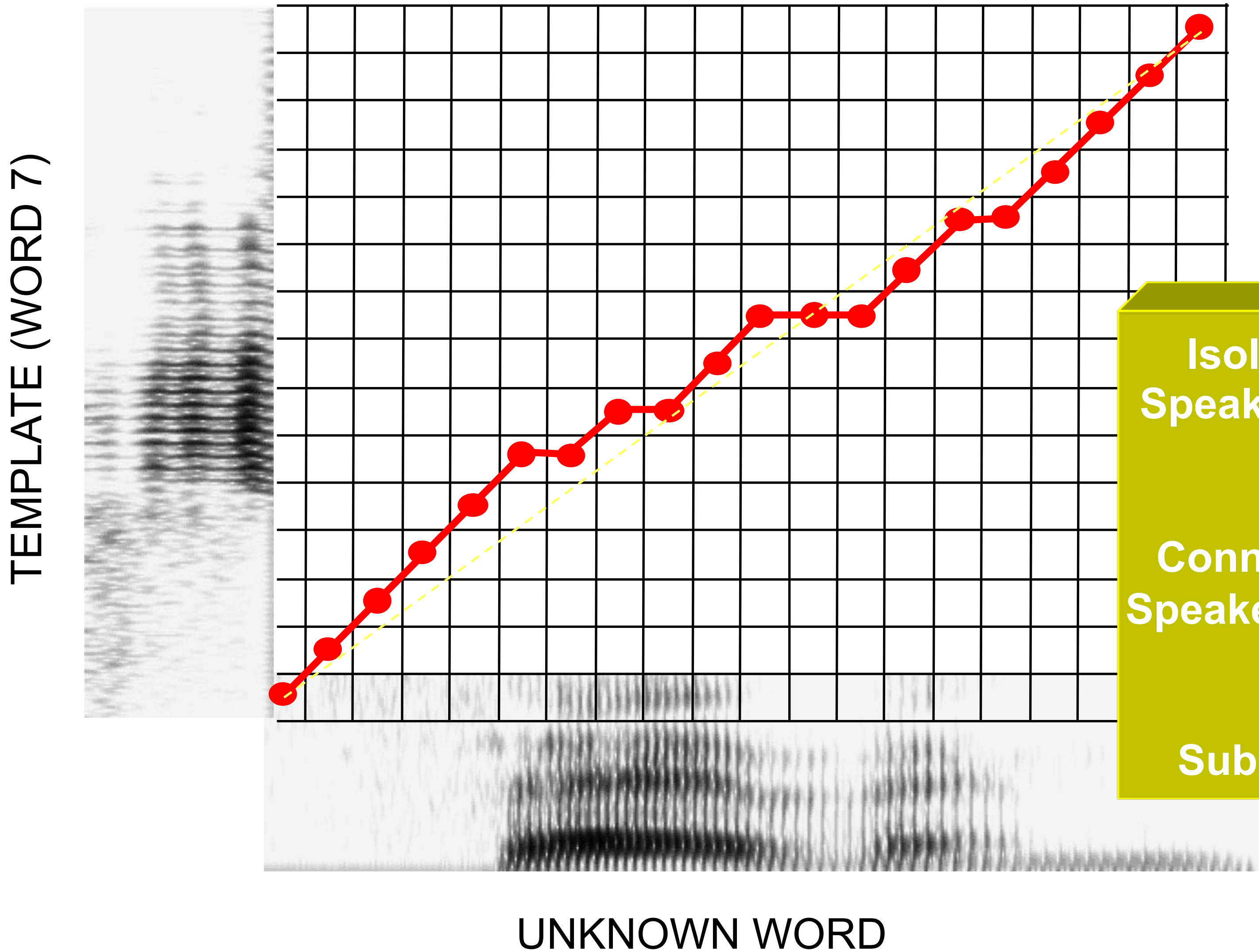
- Despite anti-speech recognition campaign led by *Pierce Commission* ARPA launches 5 year Spoken Understanding Research program
- Goal: 1000-word vocabulary, 90% understanding rate, near real time on 100 mips machine
- 4 Systems built by th
  - SDC (24%)
  - BBN's *HWIM (44%)*
  - *CMU's Hearsay II (74%)*
  - *CMU's HARPY (95% -- but 80 times real time!)*
- *Rule-based systems except for Harpy*
  - *Engineering approach: search network of all the possible utterances*

LESSON LEARNED:
Hand-built knowledge does not scale up
Need of a global "optimization" criterion

Raj Reddy -- CMU

- Lack of clear evaluation criteria
  - ARPA felt systems had failed
  - Project not extended
- Speech Understanding: too early for its time
- Need a standard evaluation method

1970's – Dynamic Time Warping
The Brute Force of the Engineering Approach



T.K. Vyntsyuk (1968)
H. Sakoe,
S. Chiba (1970)

Isolated Words
Speaker Dependent

⇩

Connected Words
Speaker Independent

⇩

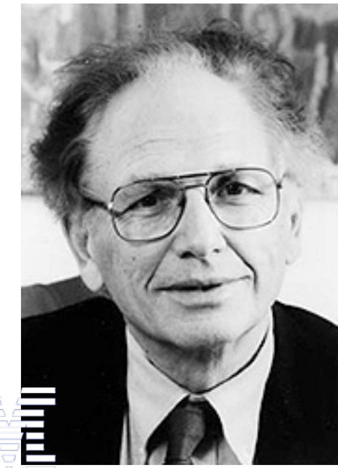
Sub-Word Units

1980s -- The Statistical Approach

- Based on work on Hidden Markov Models done by Leonard Baum at IDA, Princeton in the late 1960s
- Purely statistical approach pursued by Fred Jelinek and Jim Baker, IBM T.J.Watson Research

$$\hat{W} = \arg\max_{W} P(A \mid W) P(W)$$

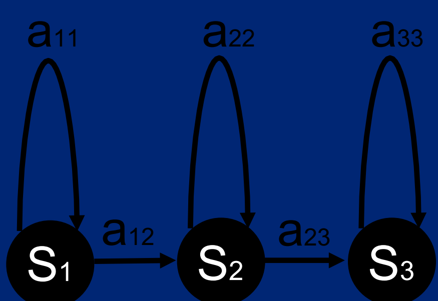
- Foundations of modern speech recognition engines



Fred Jelinek

Jim Baker

Acoustic HMMs

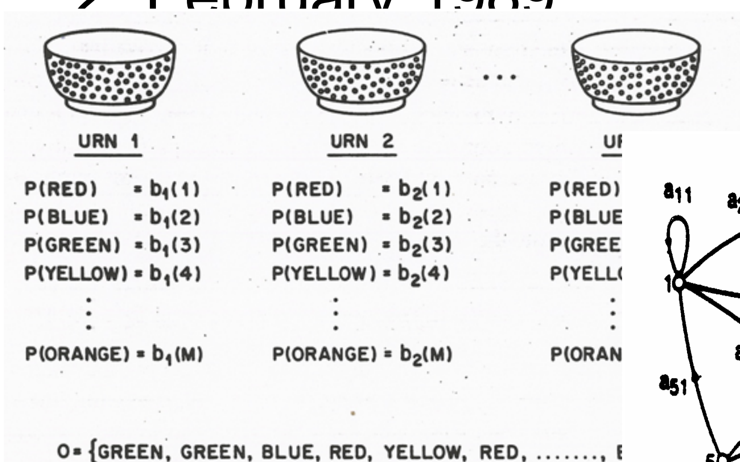


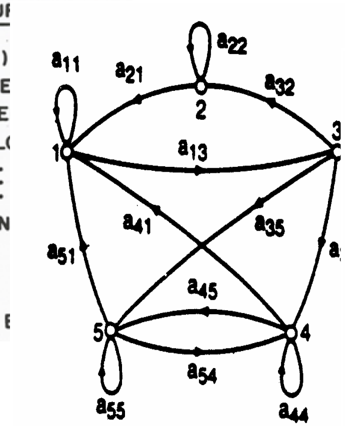
Word Tri-grams

- **No Data Like More Data**
- ***Whenever I fire a linguist, our system performance improves (1988)***
- ***Some of my best friends are linguists (2004)***

1980-1990 – Statistical approach becomes ubiquitous

- Lawrence Rabiner, *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,* Proceeding of the IEEE, Vol. 77, No. 2, February 1989





Markov Assumption:

$$P[q_t = j | q_{t-1} = i, q_{t-2} = k, \ldots] = P[q_t = j | q_{t-1} = i]$$

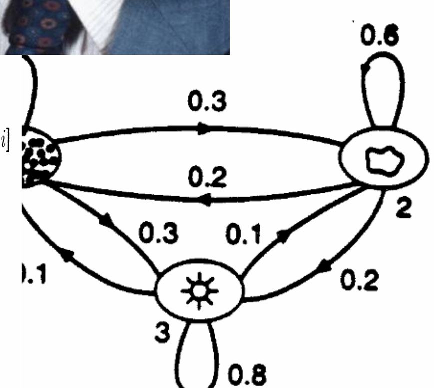
Set

$$a_{ij} = P[q_t = j | q_{t-1} = i] \quad 1 \le i, j \le N$$

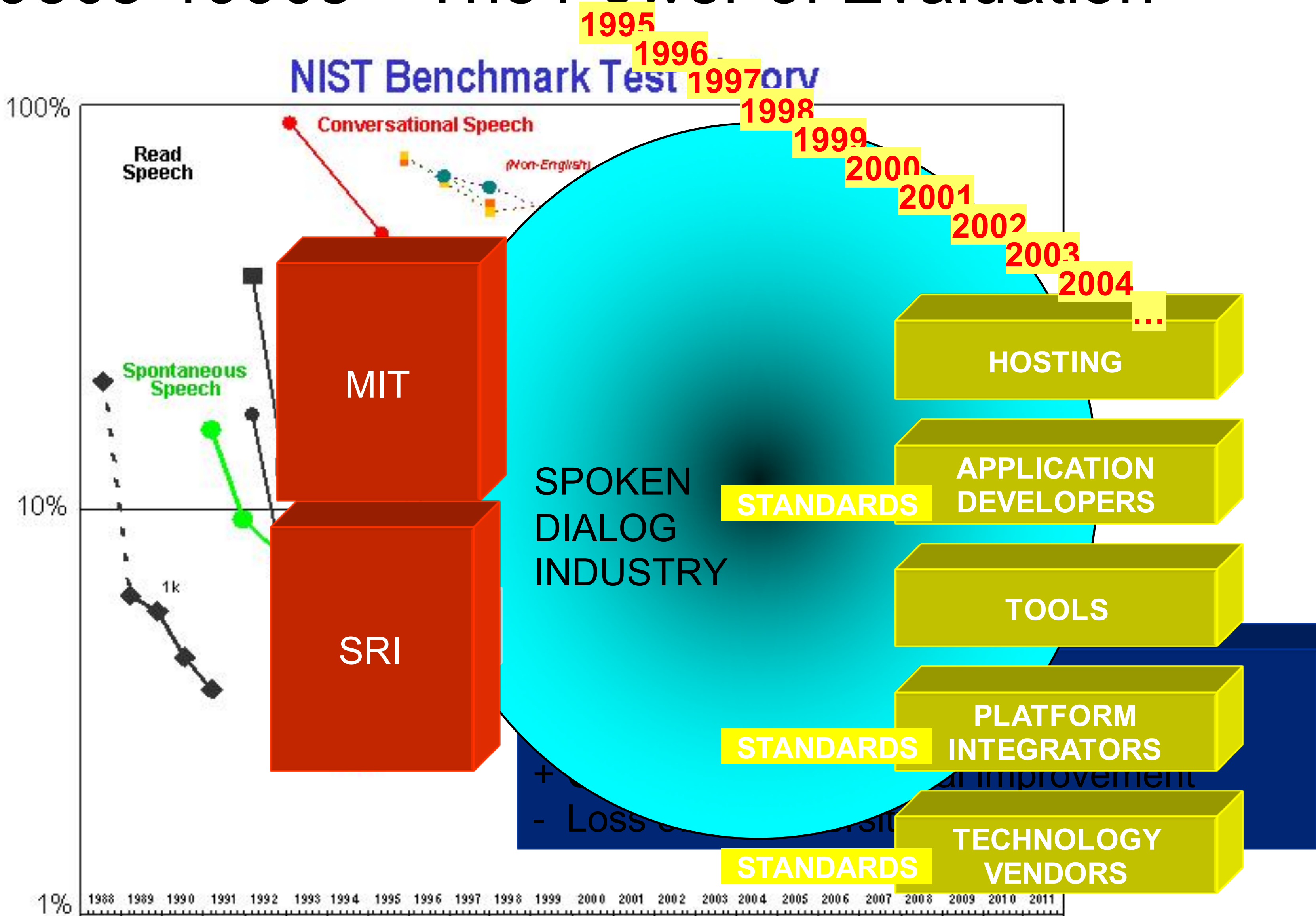
Such that

$$a_{ij} \ge 0 \quad \forall i, j$$

$$\sum_{j=1}^{N} a_{ij} = 1 \quad \forall i$$



1980s-1990s – The Power of Evaluation



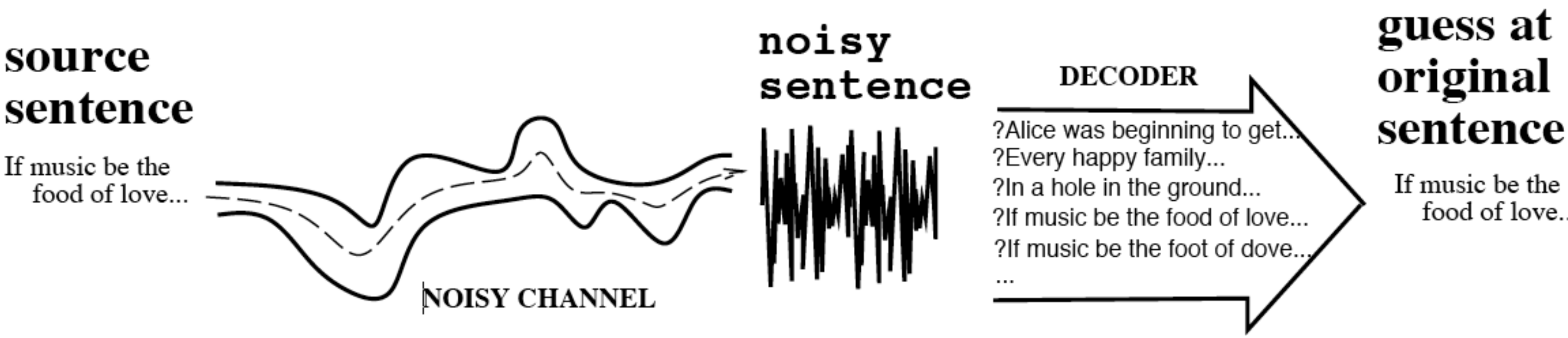
Today’s State of the Art

- Low noise conditions
- Large vocabulary
  - ~20,000-60,000 words or more…
- Speaker independent (vs. speaker-dependent)
- Continuous speech (vs isolated-word)
- Multilingual, conversational
- World’s best research systems:
  - Human-human speech: ~13-20% Word Error Rate (WER)
  - Human-machine or monologue speech: ~3-5% WER

Building an ASR System

- Build a statistical model of the speech-to-words process
  - Collect lots of speech and transcribe all the words
  - Train the model on the labeled speech
- Paradigm:
  - Supervised Machine Learning + Search
  - The Noisy Channel Model

The Noisy Channel Model



- Search through space of all possible sentences.
- Pick the one that is most probable given the waveform

The Noisy Channel Model (II)

- What is the most likely sentence out of all sentences in the language L, given some acoustic input O?
- Treat acoustic input O as sequence of individual acoustic observations
  - $O = o_1, o_2, o_3, \ldots, o_t$
- Define a sentence as a sequence of words:
  - $W = w_1, w_2, w_3, \ldots, w_n$

Noisy Channel Model (III)

- Probabilistic implication: Pick the highest probable sequence:

$$\hat{W} = \arg\max_{W \in L} P(W \mid O)$$

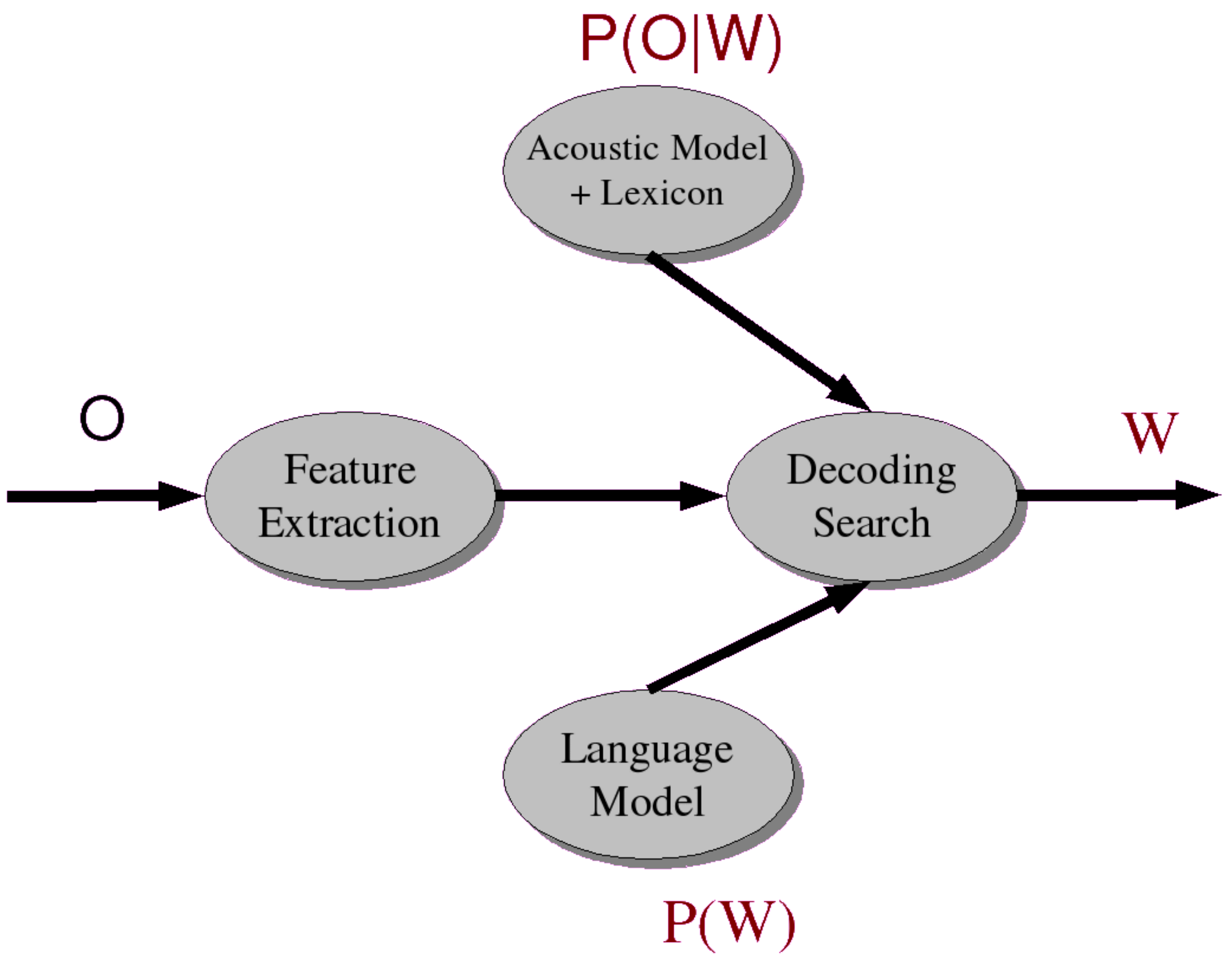
- We can use Bayes rule to rewrite this:

$$\hat{W} = \arg\max_{W \in L} \frac{P(O \mid W)P(W)}{P(O)}$$

- Since denominator is the same for each candidate sentence W, we can ignore it for the argmax:

$$\hat{W} = \arg\max_{W \in L} P(O \mid W)P(W)$$

Speech Recognition Meets Noisy Channel: Acoustic Likelihoods and LM Priors



Components of an ASR System

- Corpora for training and testing of components
- Representation for input and method of extracting
- Pronunciation Model
- Acoustic Model
- Language Model
- Feature extraction component
- Algorithms to search hypothesis space efficiently

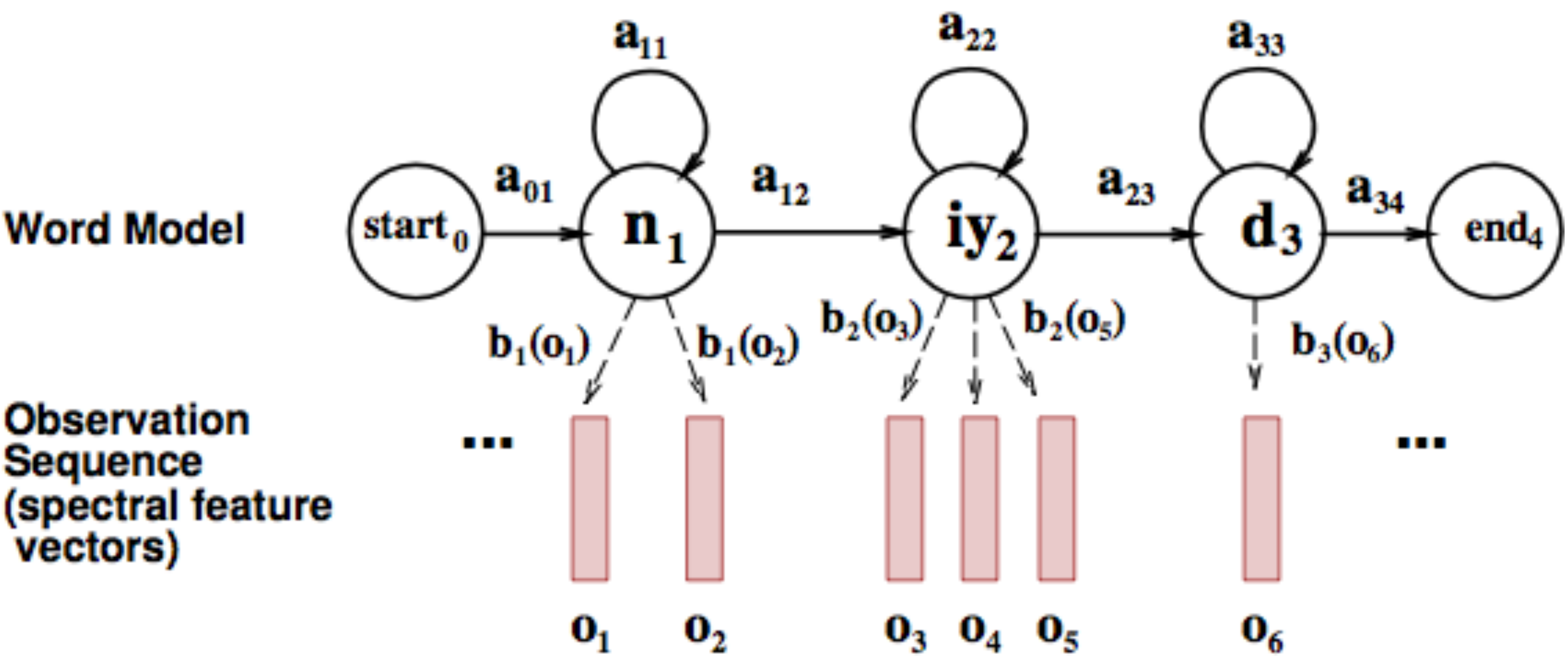
Training and Test Corpora

- Collect corpora appropriate for recognition task at hand
  - Small speech + phonetic transcription to associate sounds with symbols (Acoustic Model)
  - Large (>= 60 hrs) **s**peech + orthographic transcription to associate words with sounds (Acoustic Model)
  - Very large text corpus to identify ngram probabilities or build a grammar (Language Model)

Building the Acoustic Model

- Goal: Model likelihood of sounds given spectral features, pronunciation models, and prior context
- Usually represented as Hidden Markov Model
  - States represent phones or other subword units
  - Transition probabilities on states: how likely is it to see one sound after seeing another?
  - Observation/output likelihoods: how likely is spectral feature vector to be observed from phone state i, given phone state i-1?

Word HMM



Word Model
Observation Sequence (spectral feature vectors)
start$_0$
n$_1$
iy$_2$
d$_3$
end$_4$
a$_{01}$
a$_{11}$
a$_{12}$
a$_{22}$
a$_{23}$
a$_{33}$
a$_{34}$
b$_1$(o$_1$)
b$_1$(o$_2$)
b$_2$(o$_3$)
b$_2$(o$_5$)
b$_3$(o$_6$)
...
o$_1$
o$_2$
o$_3$
o$_4$
o$_5$
o$_6$
...

- Initial estimates from phonetically transcribed corpus or flat start
  - Transition probabilities between phone states
  - Observation probabilities associating phone states with acoustic features of windows of waveform
- Embedded training:
  - Re-estimate probabilities using initial phone HMMs + orthographically transcribed corpus + pronunciation lexicon to create whole sentence HMMs for each sentence in training corpus
  - Iteratively retrain transition and observation probabilities by running the training data through the model until convergence

Training the Acoustic Model

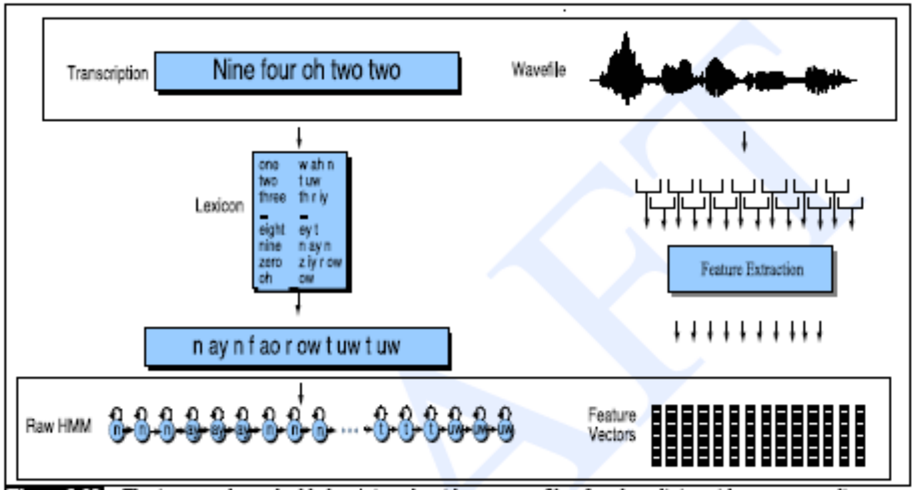
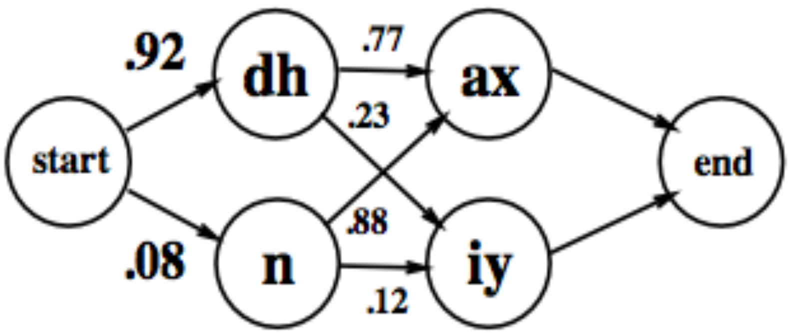


Figure 9.32 The input to the embedded training algorithm; a wavefile of spoken digits with a corresponding transcription. The transcription is converted into a raw HMM, ready to be aligned and trained against the cepstral features extracted from the wavefile.

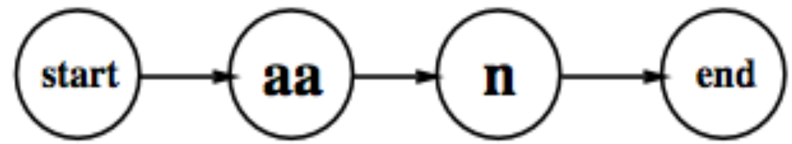
Building the Pronunciation Model

- Models likelihood of word given network of candidate phone hypotheses
  - Multiple pronunciations for each word
  - May be weighted automaton or simple dictionary
- Words come from all corpora (including text)
- Pronunciations come from pronouncing dictionary or TTS system

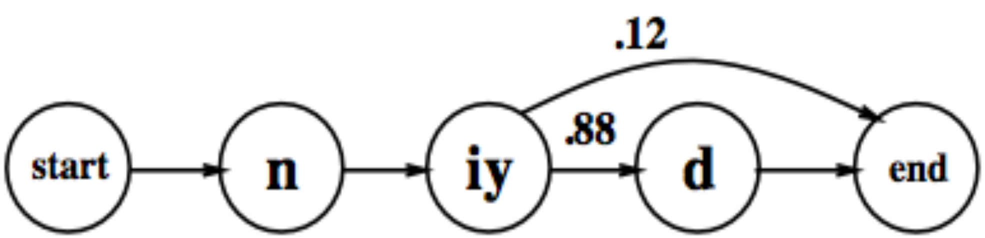
ASR Lexicon: Markov Models for Pronunciation



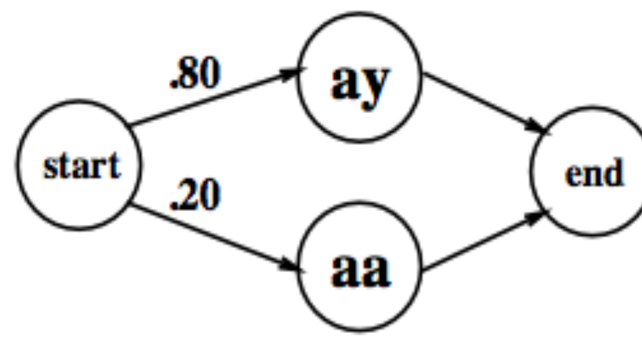
Word model for "the"



Word model for "on"



Word model for "need"



Word model for "I"

Building the Language Model

- Models likelihood of word given previous word(s)
- Ngram models:
  - Build the LM by calculating bigram or trigram probabilities from text training corpus: how likely is one word to follow another? To follow the two previous words?
  - Smoothing issues
- Grammars
  - Finite state grammar or Context Free Grammar (CFG) or semantic grammar
- Out of Vocabulary (OOV) problem

Search/Decoding

- Find the best hypothesis P(O|W) P(W) given
  - A sequence of acoustic feature vectors (O)
  - A trained HMM (AM)
  - Lexicon (PM)
  - Probabilities of word sequences (LM)
- For O
  - Calculate most likely state sequence in HMM given transition and observation probs
  - Trace back thru state sequence to assign words to states
  - N best vs. 1 best vs. lattice output
- Limiting search
  - Lattice minimization and determinization
  - Pruning: beam search

Evaluating Success

- Transcription
  - Low WER (Subst+Ins+Del)/N * 100

  Thesis test vs. This is a test. 75% WER

  Or That was the dentist calling. 125% WER
- Understanding
  - High concept accuracy
    - How many domain concepts were correctly recognized?

      I want to go from Boston to Baltimore on September 29

Domain concepts	Values

Domain concepts	Values
– source city	Boston
– target city	Baltimore
– travel date	September 29

- Score recognized string "Go from Boston to Washington on December 29" vs. "Go to Boston from Baltimore on September 29"
- (1/3 = 33% CA)

Summary

- ASR today
  - Combines many probabilistic phenomena: varying acoustic features of phones, likely pronunciations of words, likely sequences of words
  - Relies upon many approximate techniques to ‘translate’ a signal
  - Finite State Transducers
- ASR future
  - Can we include more language phenomena in the model?