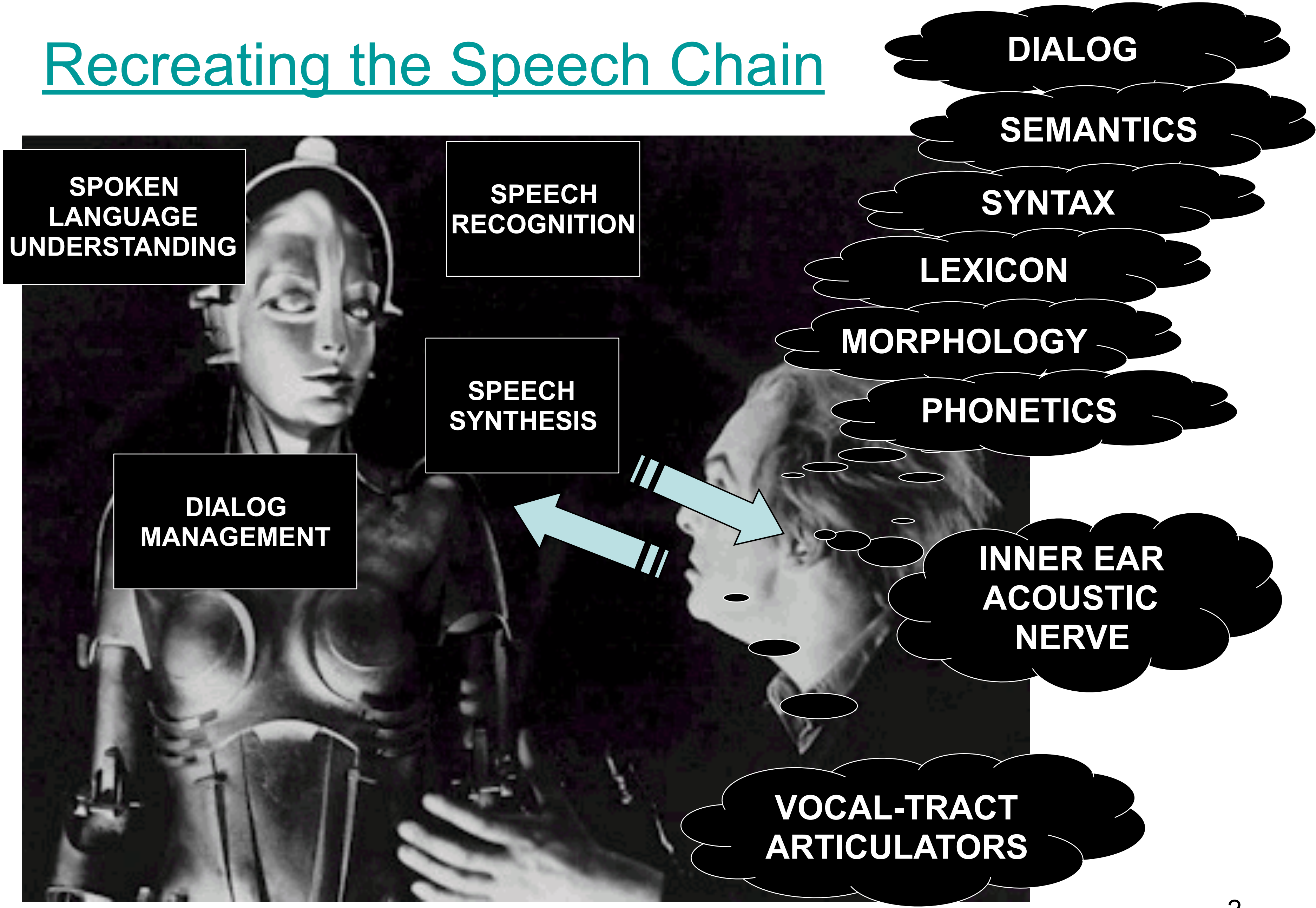
Automatic Speech Recognition

Instructor: Wei Xu
Ohio State University

Many Slides from Julia Hirschberg

Recreating the Speech Chain



SPOKEN LANGUAGE UNDERSTANDING

SPEECH RECOGNITION

SPEECH SYNTHESIS

DIALOG MANAGEMENT

DIALOG

SEMANTICS

SYNTAX

LEXICON

MORPHOLOGY

PHONETICS

INNER EAR ACOUSTIC NERVE

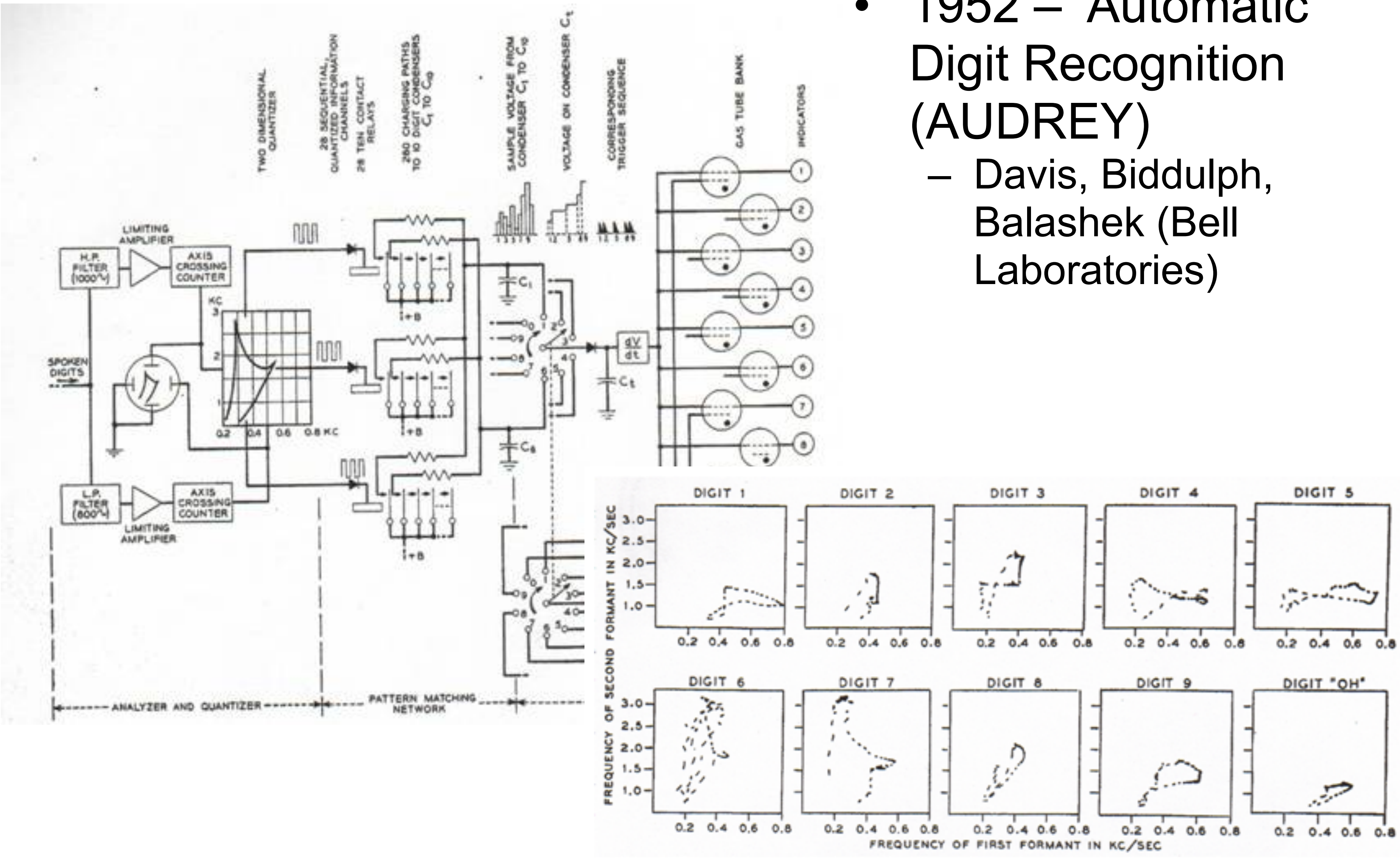
VOCAL-TRACT ARTICULATORS

Speech Recognition: the Early Years

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

Speech Recognition: the Early Years

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

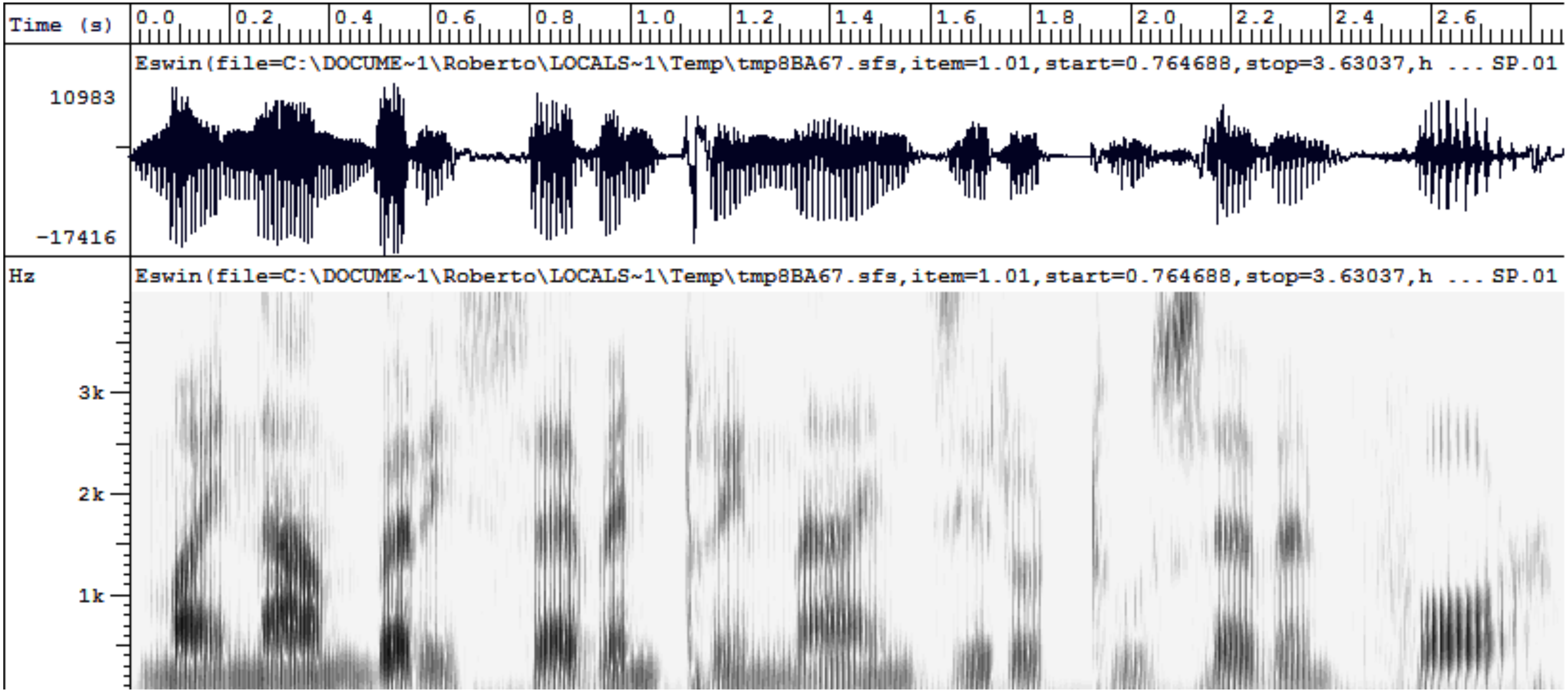


1960’s – Speech Processing and Digital Computers

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

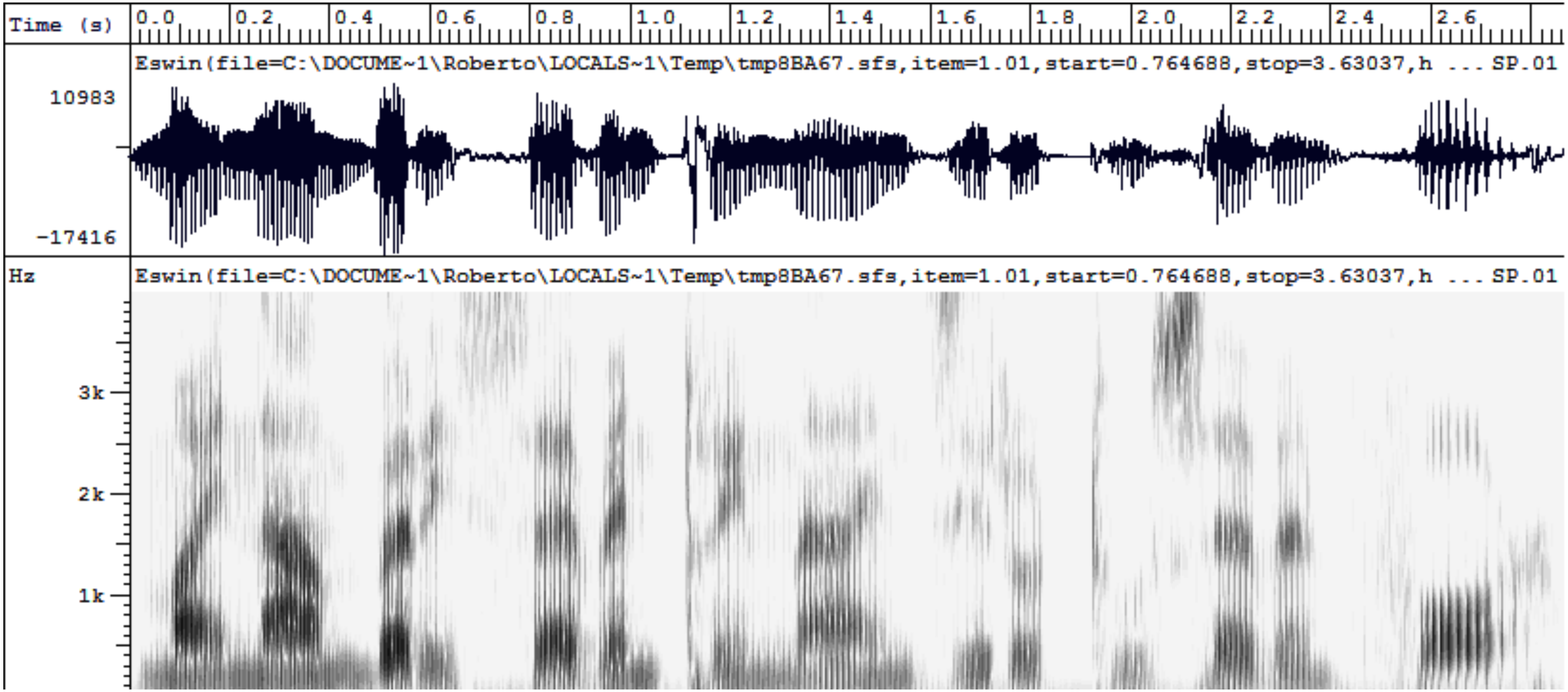
James Flanagan
Bell Laboratories

The Illusion of Segmentation... or...



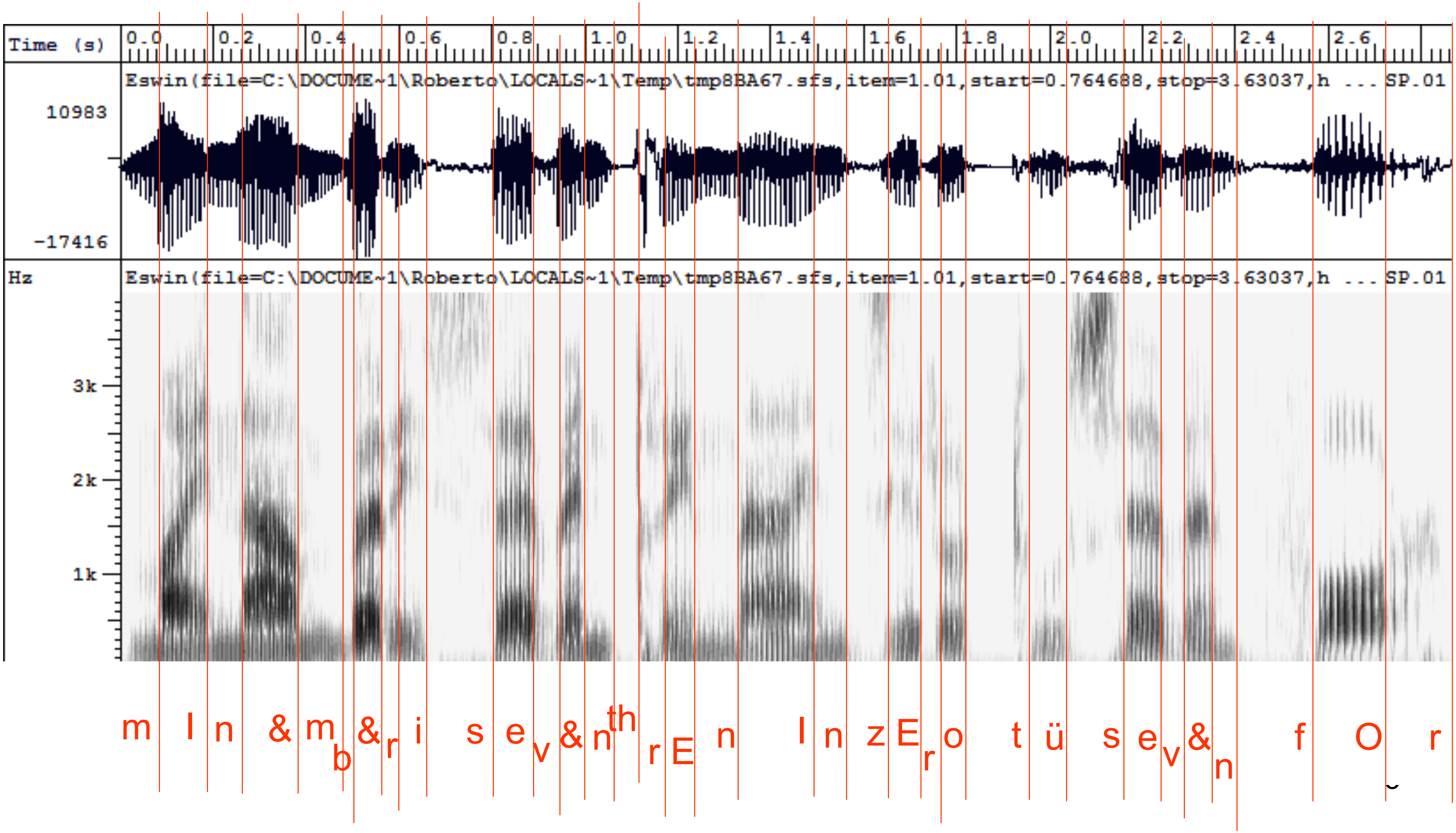
The Illusion of Segmentation... or...

Why Speech Recognition is so Difficult

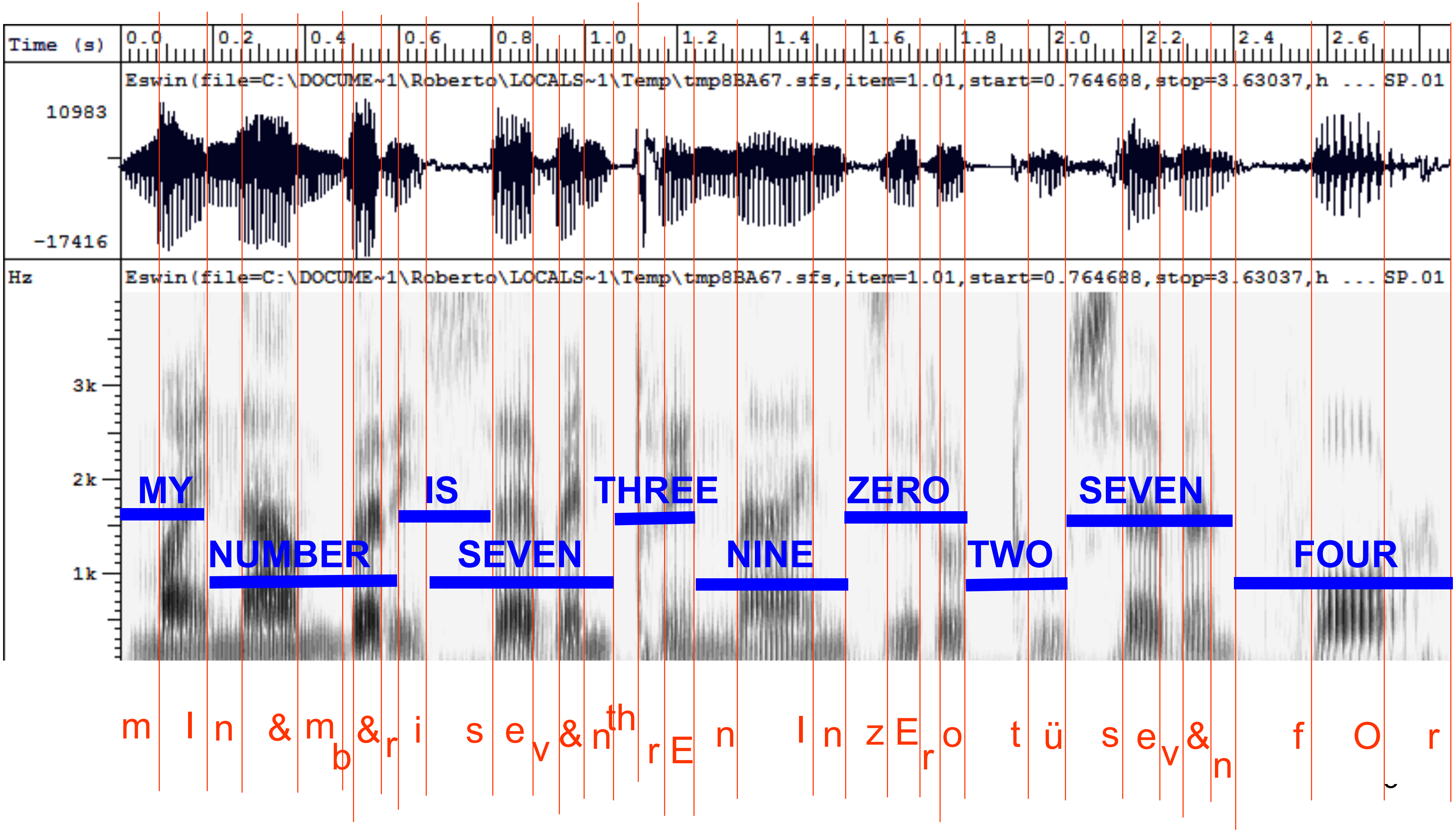


The Illusion of Segmentation... or...

Why Speech Recognition is so Difficult

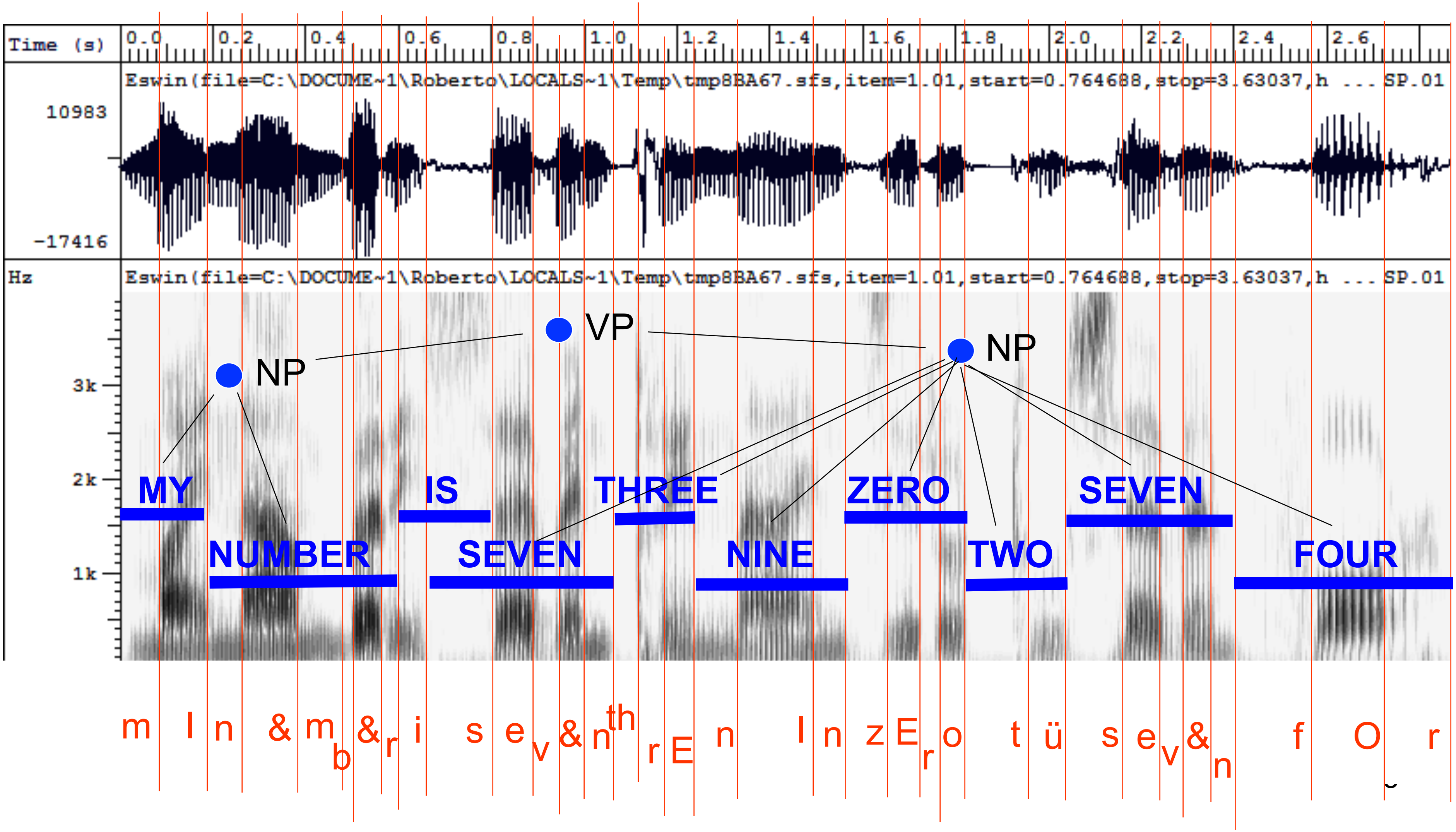


The Illusion of Segmentation... or... Why Speech Recognition is so Difficult



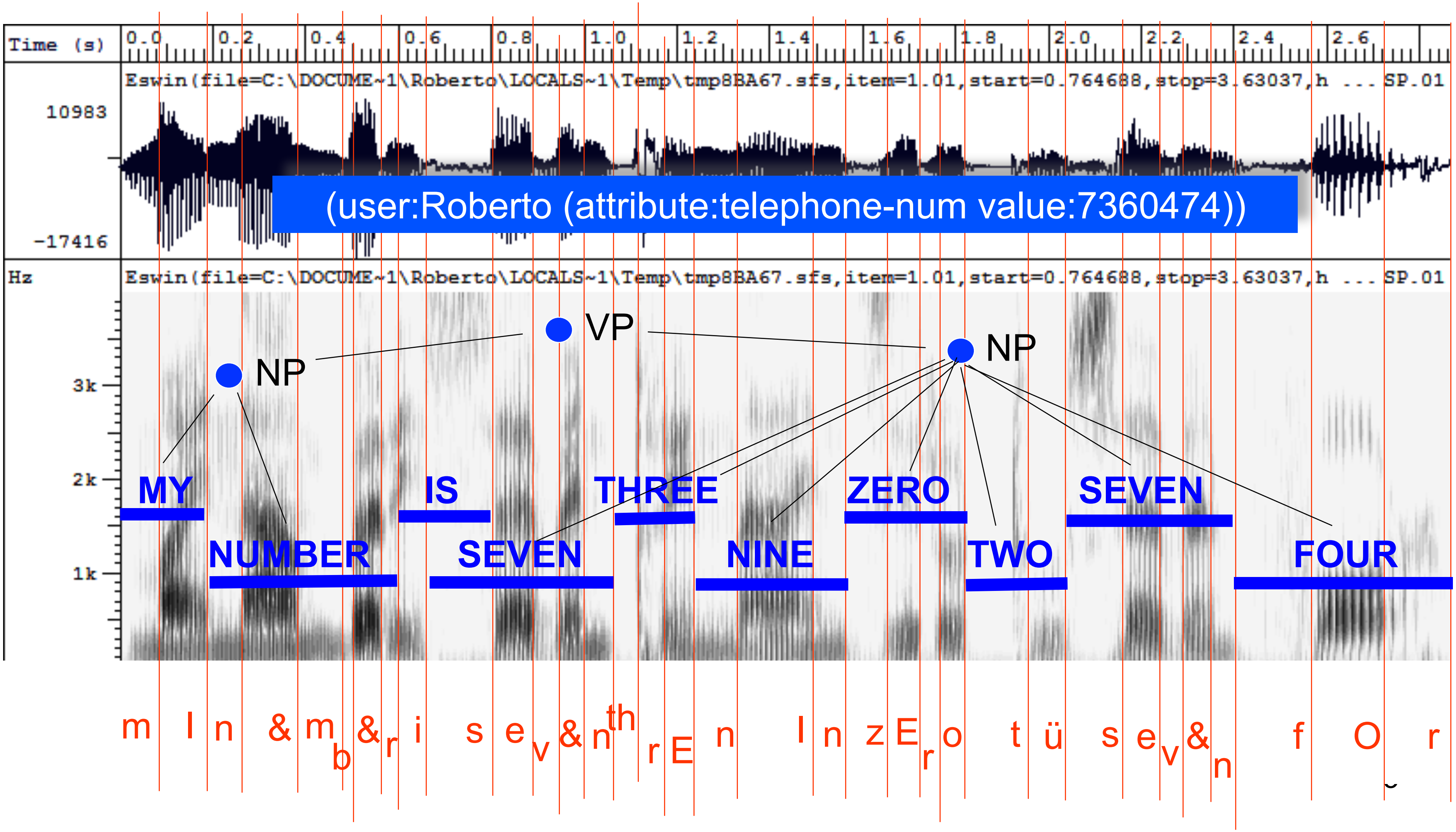
The Illusion of Segmentation... or...

Why Speech Recognition is so Difficult



The Illusion of Segmentation... or...

Why Speech Recognition is so Difficult

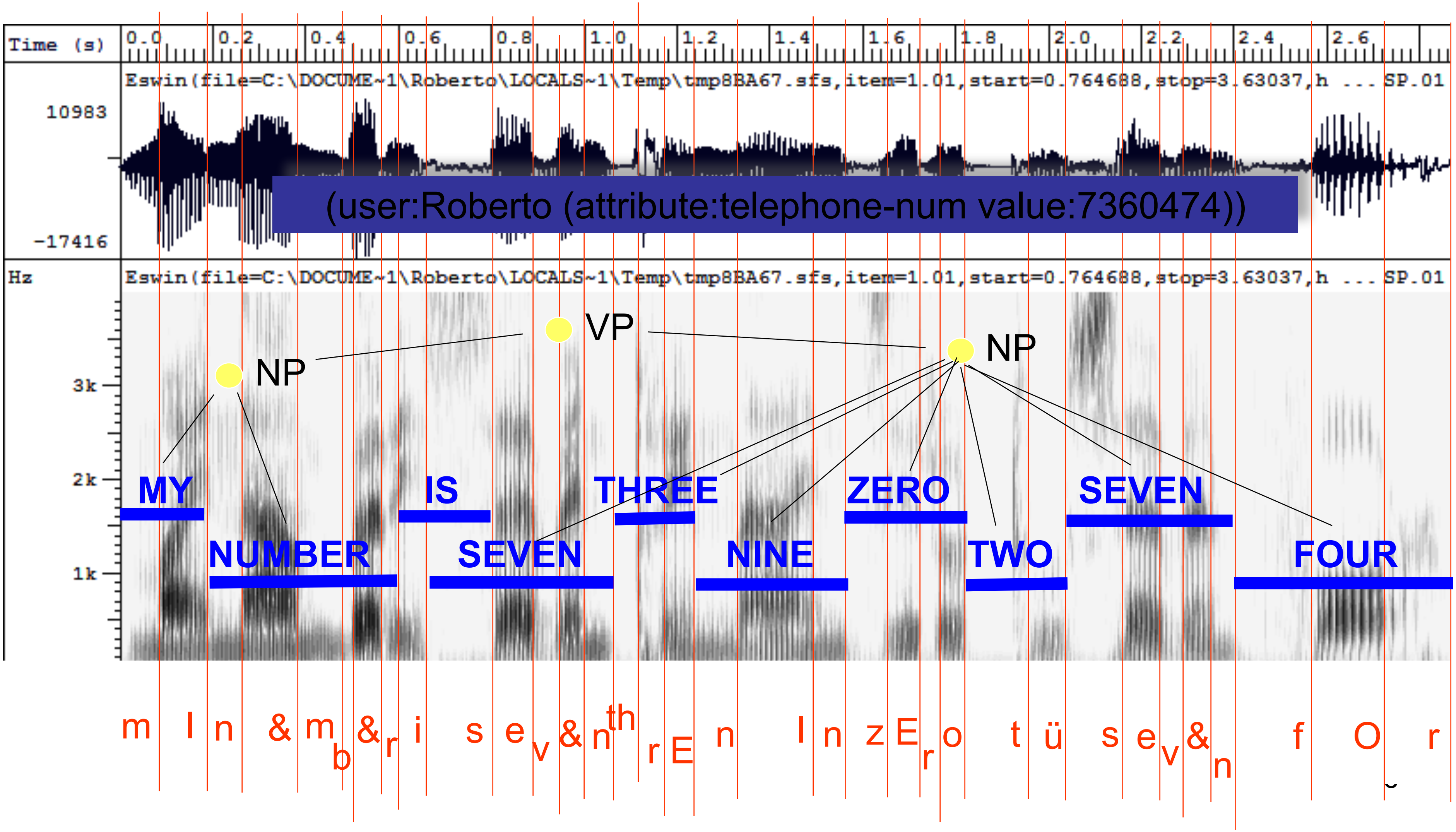


(user:Roberto (attribute:telephone-num value:7360474))

Time (s)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2
2.4
2.6
Eswin(file=C:\DOCUME~1\Roberto\LOCALS~1\Temp\tmp8BA67.sfs,item=1.01,start=0.764688,stop=3.63037,h ... SP.01
10983
-17416
Hz
Eswin(file=C:\DOCUME~1\Roberto\LOCALS~1\Temp\tmp8BA67.sfs,item=1.01,start=0.764688,stop=3.63037,h ... SP.01
3k
2k
1k
VP
NP
NP
MY
NUMBER
IS
SEVEN
THREE
NINE
ZERO
TWO
SEVEN
FOUR
m I n & m b & r i s e v & n th r E n I n z E r o t ü s e v & n f O r

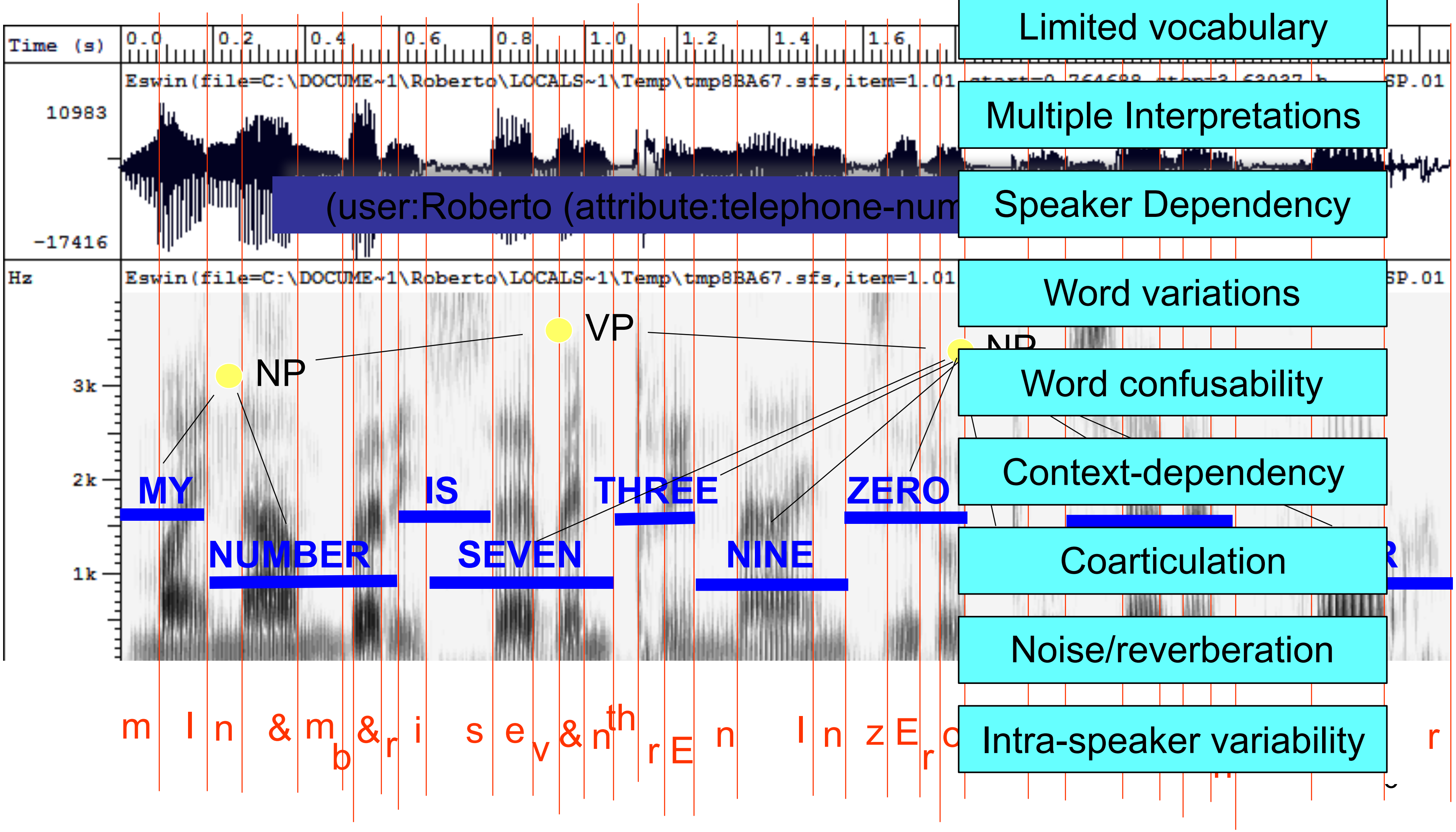
The Illusion of Segmentation... or...

Why Speech Recognition is so Difficult



Time (s)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2
2.4
2.6
Eswin(file=C:\DOCUME~1\Roberto\LOCALS~1\Temp\tmp8BA67.sfs,item=1.01,start=0.764688,stop=3.63037,h ... SP.01
10983
-17416
(user:Roberto (attribute:telephone-num value:7360474))
Hz
Eswin(file=C:\DOCUME~1\Roberto\LOCALS~1\Temp\tmp8BA67.sfs,item=1.01,start=0.764688,stop=3.63037,h ... SP.01
3k
2k
1k
VP
NP
NP
MY
NUMBER
IS
SEVEN
THREE
NINE
ZERO
TWO
SEVEN
FOUR
m I n & m b & r i s e v & n th r E n I n z E r o t ü s e v & n f O r

The Illusion of Segmentation... or... Why Speech Recognition is So Difficult



- Ellipses and Anaphors
- Limited vocabulary
- Multiple Interpretations
- Speaker Dependency
- Word variations
- Word confusability
- Context-dependency
- Coarticulation
- Noise/reverberation
- Intra-speaker variability

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

1969 – Whither Speech Recognition?

J. R. Pierce
Executive Director,
Bell Laboratories

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

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 the end of the program
  - 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 poss utterances*

Raj Reddy -- CMU

1971-1976: The ARPA SUR project

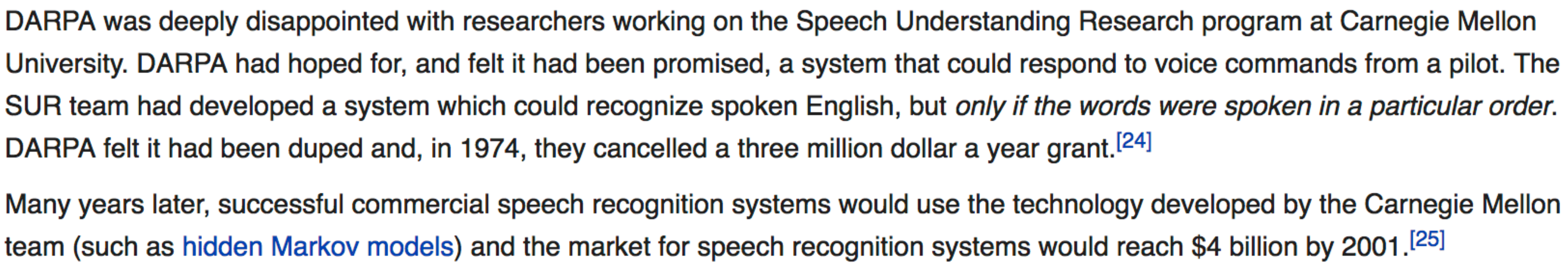
- Despite anti-speech recognition campaign led by *Pierce Commission* ARPA launches 5 year Spoken Understanding Research program
- Goal: 1000-word vocabulary, 90% understanding rate, near real time on 100 mips machine
- 4 Systems built by the end of the program
  - SDC (24%)
  - BBN's *HWIM (44%)*
  - *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 poss utterances*

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

Raj Reddy -- CMU

AI Winter (**1974**–80 and 1987–93)

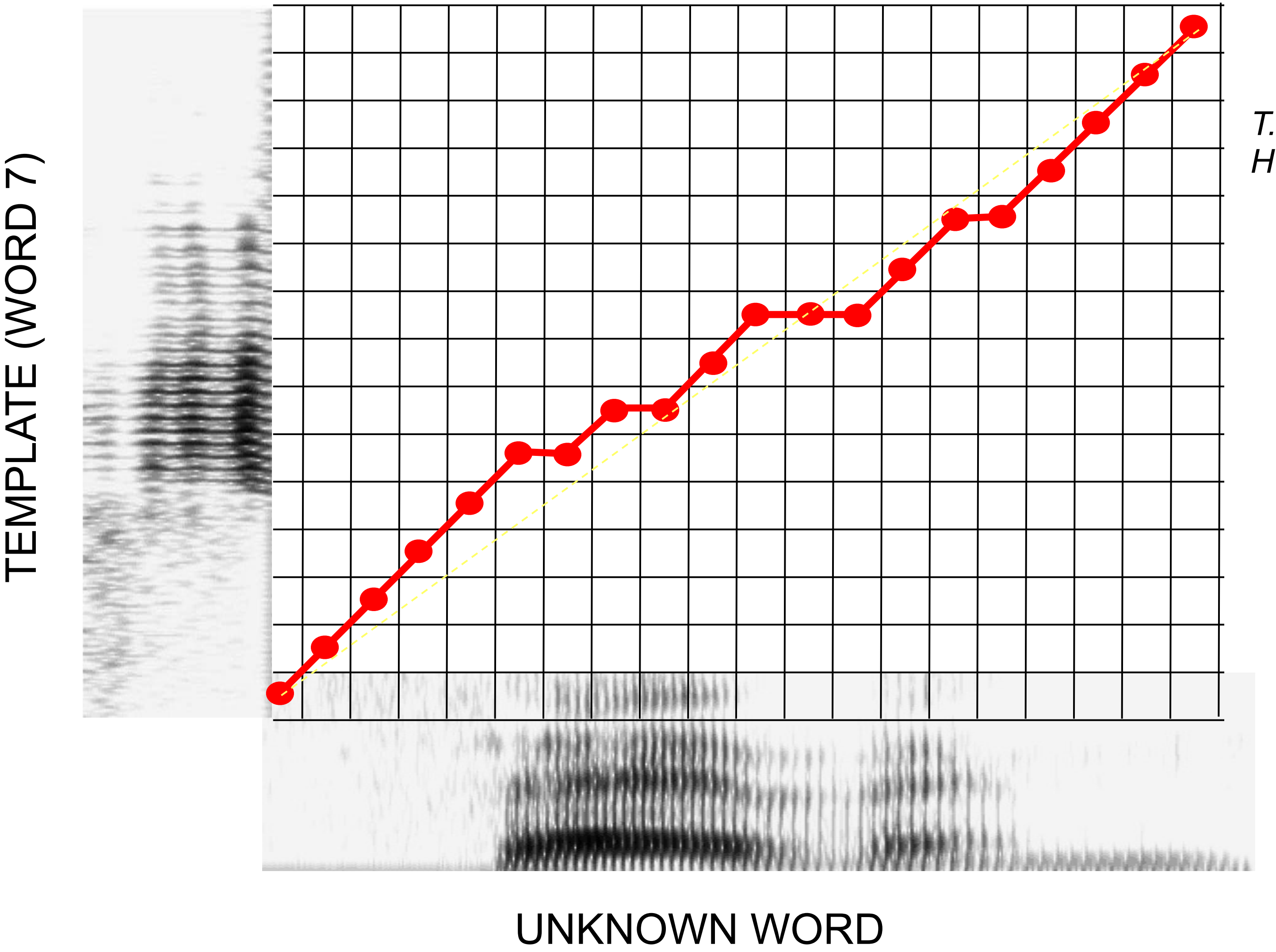
- Lack of clear evaluation criteria
  - ARPA felt systems had failed
  - Project not extended
- Speech Understanding: too early for its time
- Need a standard evaluation method



DARPA was deeply disappointed with researchers working on the Speech Understanding Research program at Carnegie Mellon University. DARPA had hoped for, and felt it had been promised, a system that could respond to voice commands from a pilot. The SUR team had developed a system which could recognize spoken English, but *only if the words were spoken in a particular order.* DARPA felt it had been duped and, in 1974, they cancelled a three million dollar a year grant.[24]

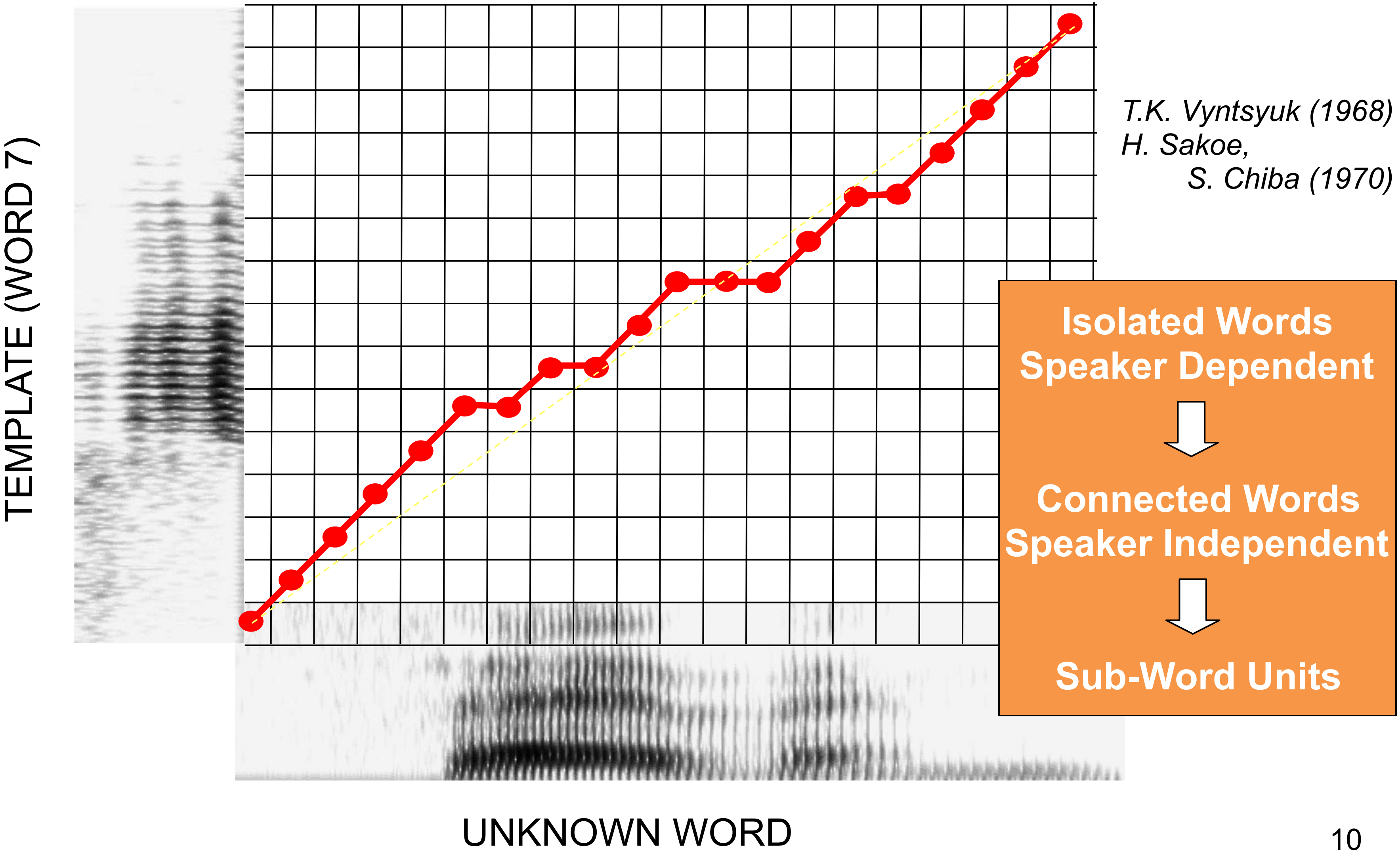
Many years later, successful commercial speech recognition systems would use the technology developed by the Carnegie Mellon team (such as hidden Markov models) and the market for speech recognition systems would reach $4 billion by 2001.[25]

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



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

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

⇩

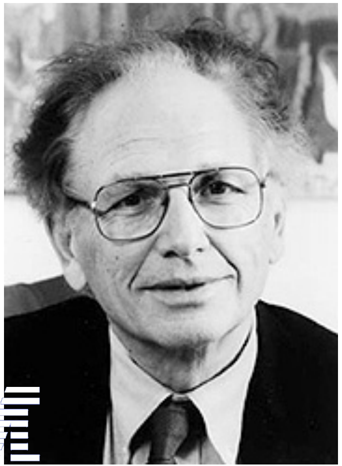
Connected Words
Speaker Independent

⇩

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
- Foundations of modern speech recognition engines

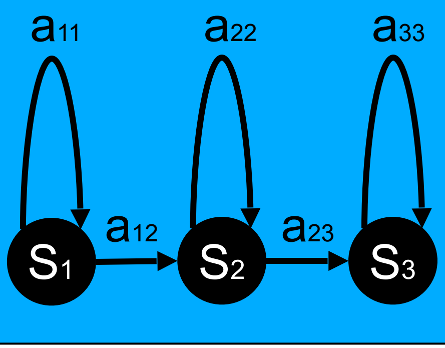


Fred Jelinek

$$\hat{W} = \underset{W}{\arg\max}\, P(A \mid W)\, P(W)$$

Jim Baker

Acoustic HMMs



Word Tri-grams

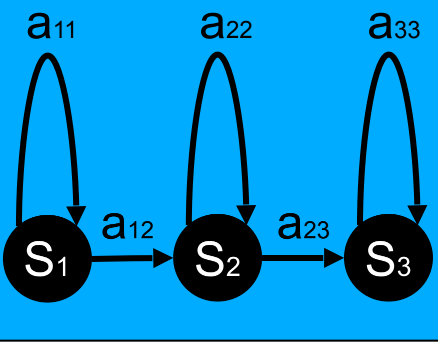
$$P(w_t \mid w_{t-1}, w_{t-2})$$

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
- Foundations of modern speech recognition engines

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

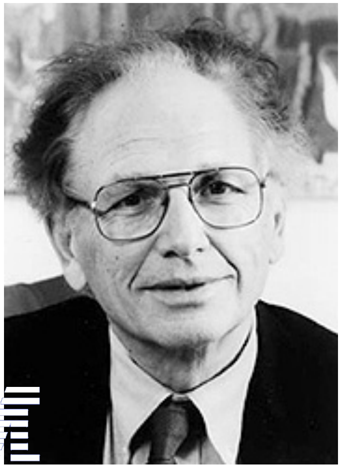
Acoustic HMMs



Word Tri-grams

$$P(w_t \mid w_{t-1}, w_{t-2})$$

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

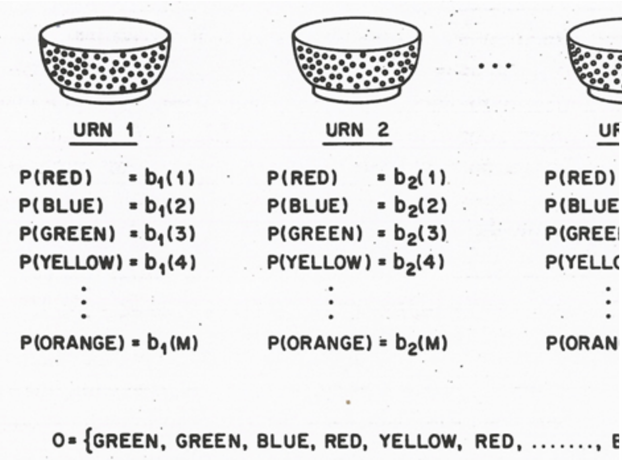


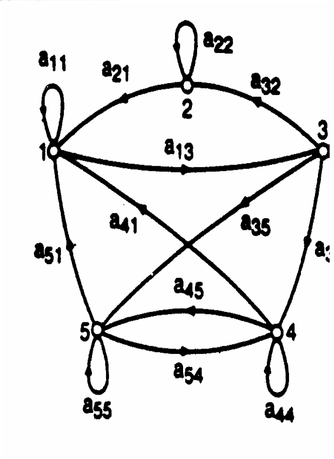
Fred Jelinek

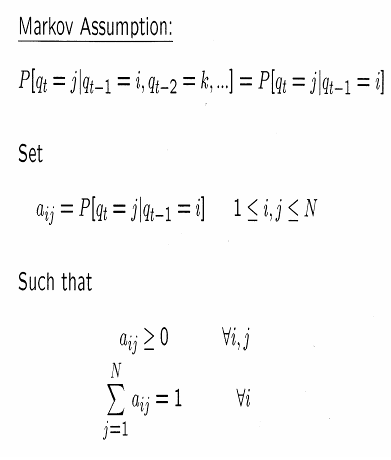
Jim Baker

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:

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

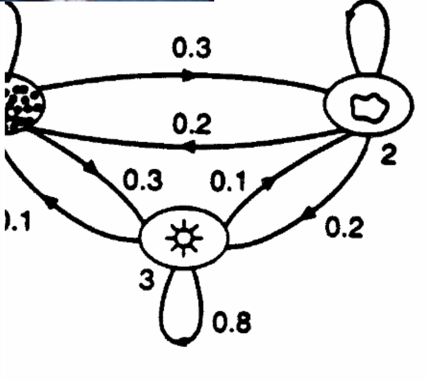
Set

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

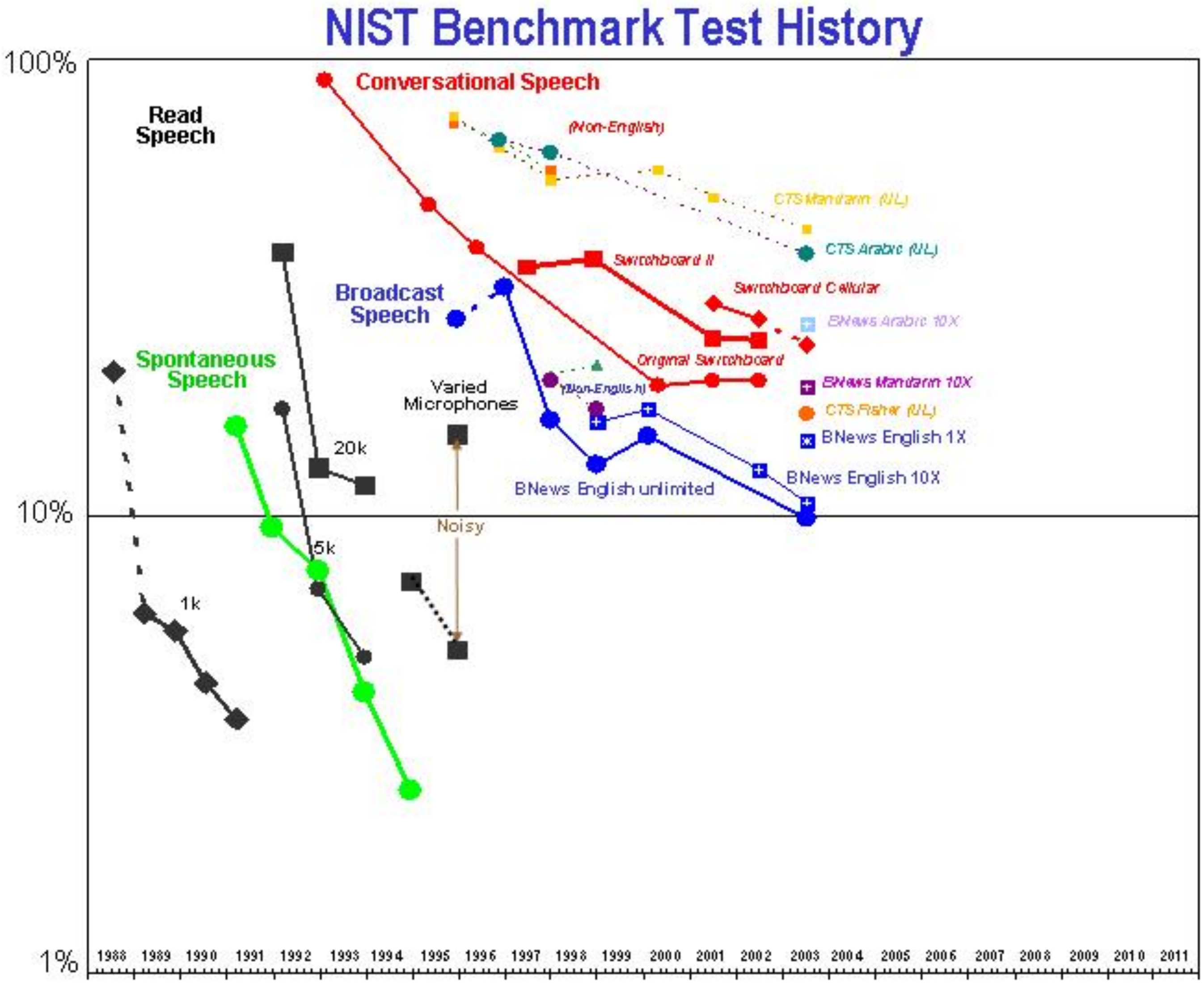
Such that

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

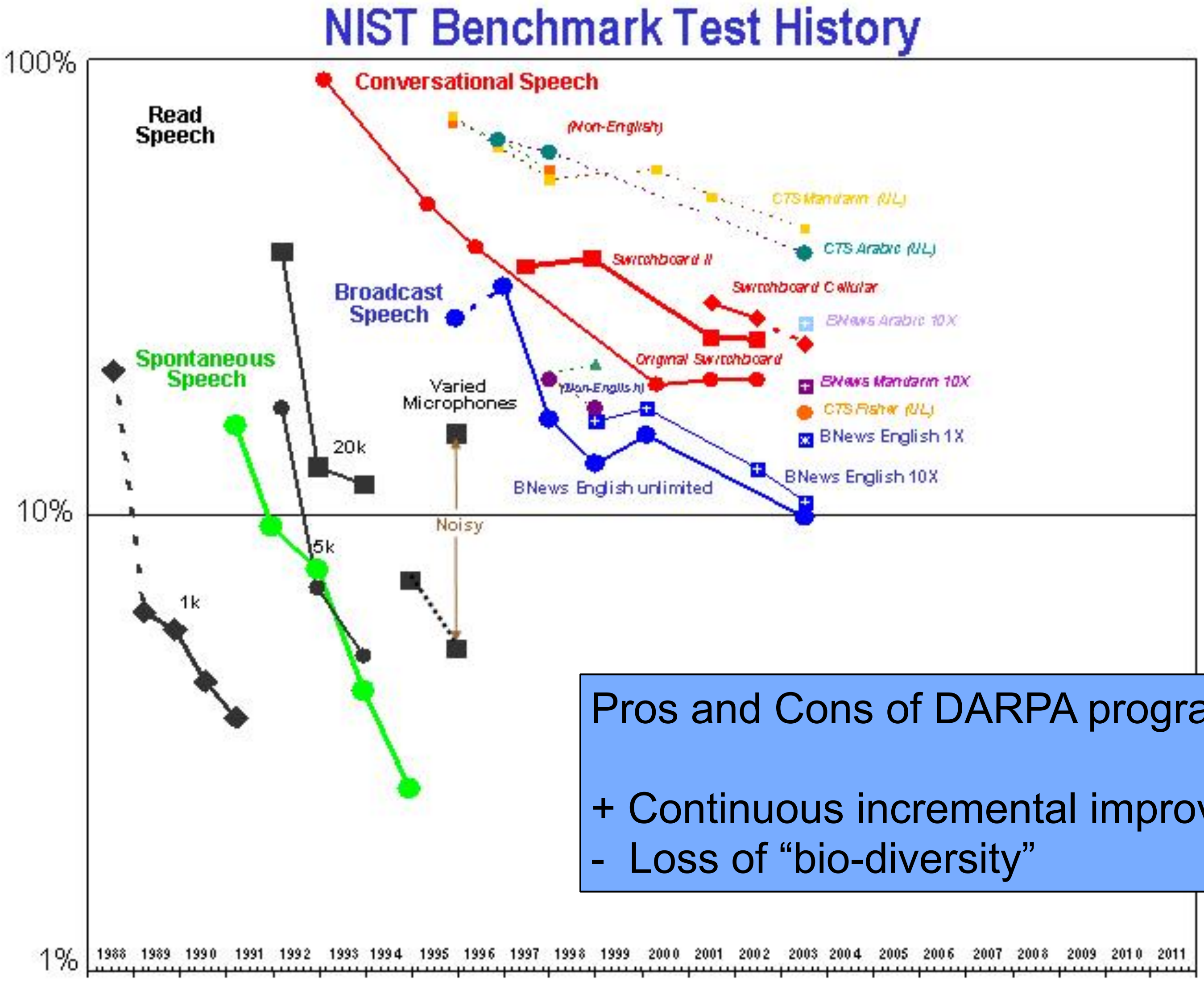
$$\sum_{j=1}^{N} a_{ij} = 1 \qquad \forall i$$



1980s-1990s – The Power of Evaluation



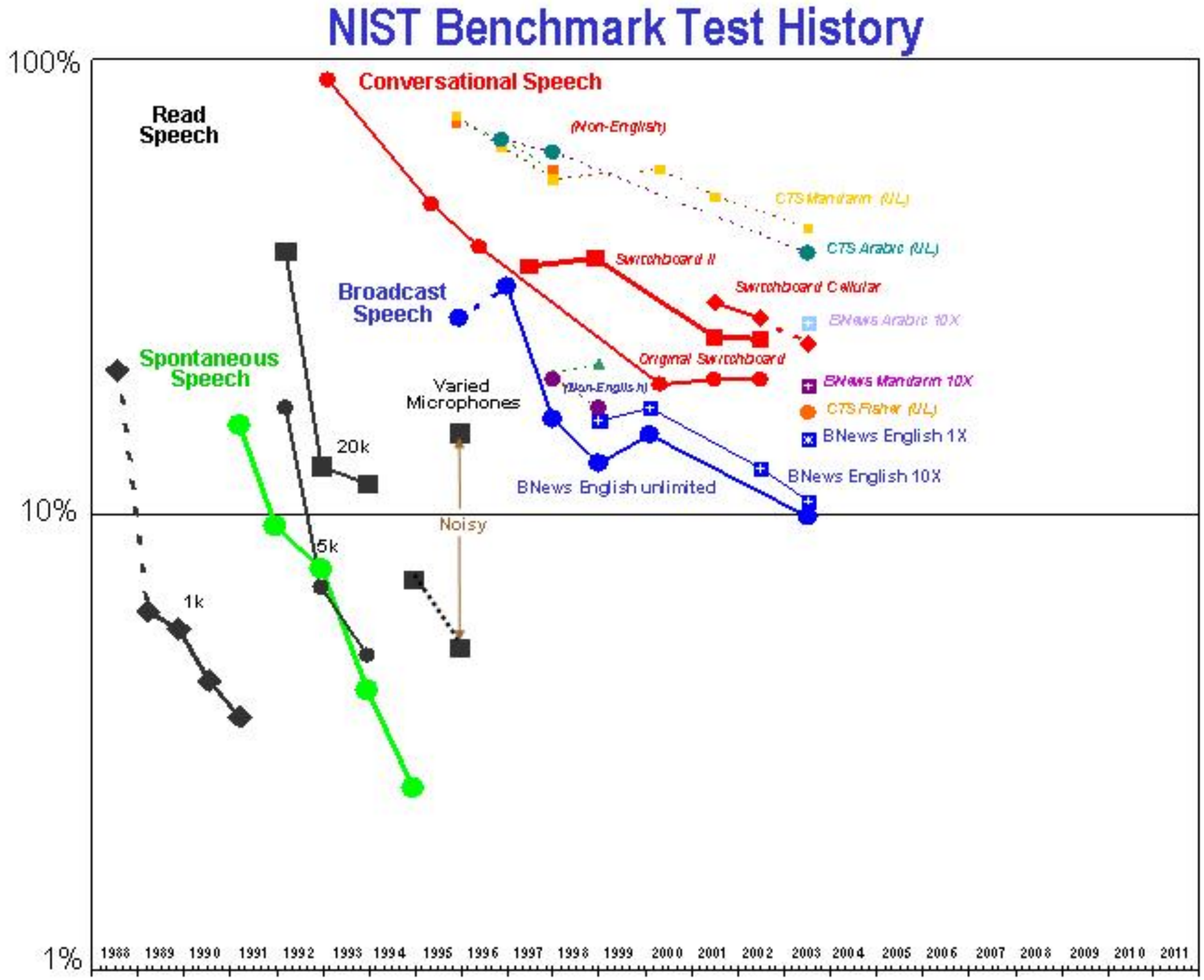
1980s-1990s – The Power of Evaluation



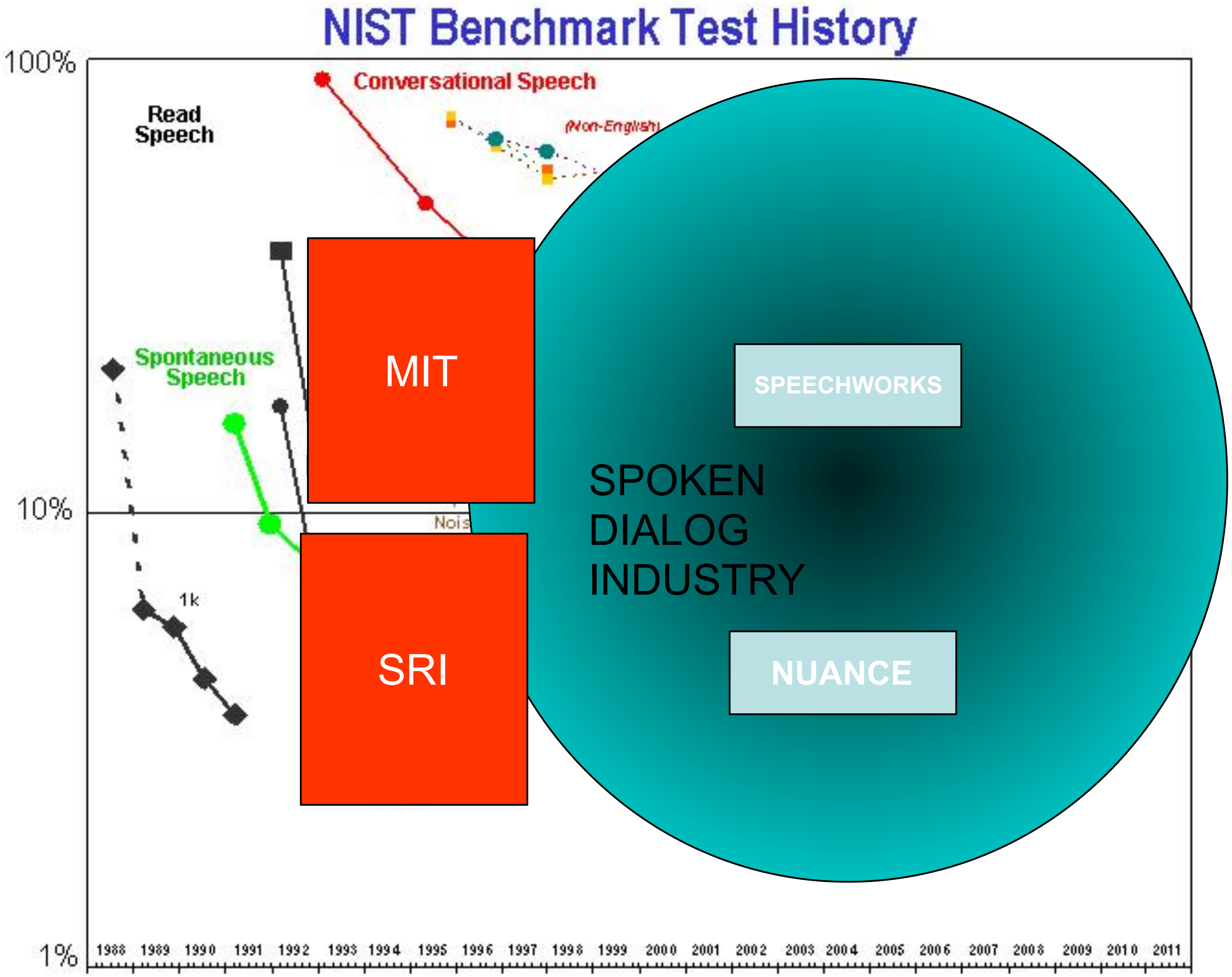
Pros and Cons of DARPA programs

+ Continuous incremental improvement
- Loss of “bio-diversity”

1980s-1990s – The Power of Evaluation



1980s-1990s – The Power of Evaluation



MIT

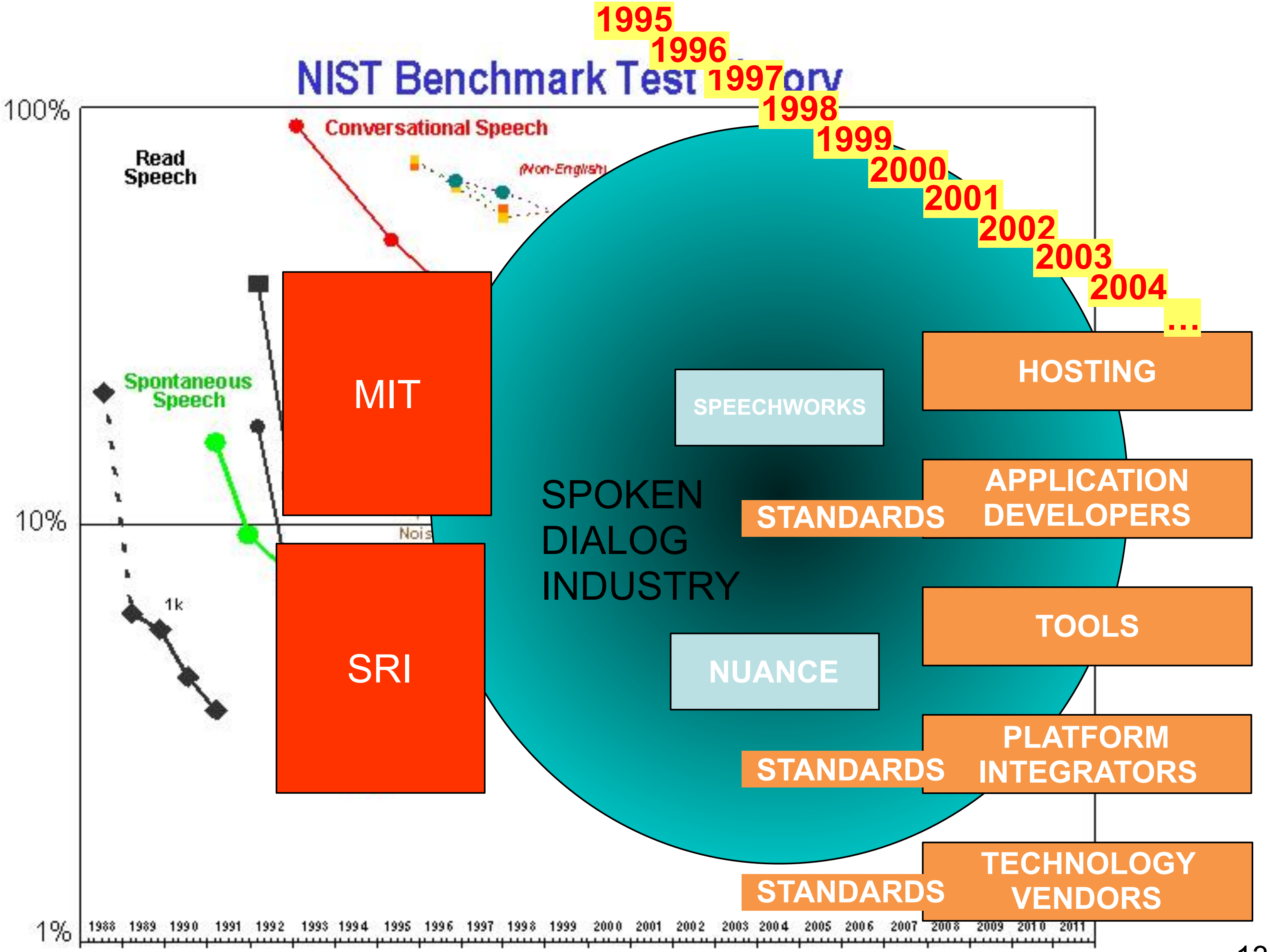
SRI

SPOKEN DIALOG INDUSTRY

SPEECHWORKS

NUANCE

1980s-1990s – The Power of Evaluation



1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
…

MIT

SRI

SPOKEN DIALOG INDUSTRY

SPEECHWORKS

NUANCE

STANDARDS

STANDARDS

STANDARDS

HOSTING

APPLICATION DEVELOPERS

TOOLS

PLATFORM INTEGRATORS

TECHNOLOGY VENDORS

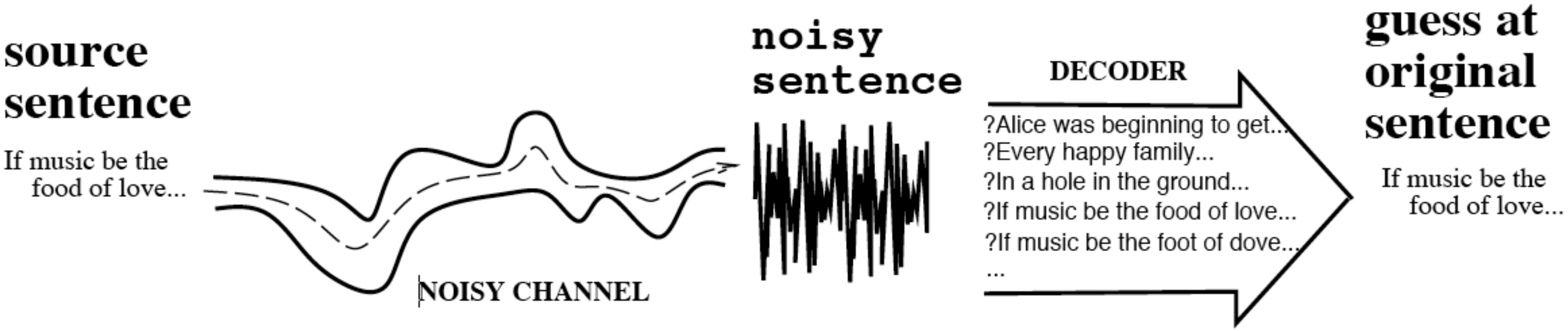
State of the Art before Deep Learning

- 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


source sentence
If music be the food of love...
NOISY CHANNEL
noisy sentence
DECODER
?Alice was beginning to get...
?Every happy family...
?In a hole in the ground...
?If music be the food of love...
?If music be the foot of dove...
...
guess at original sentence
If music be the food of love...

- 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:

$$\hat{W} = \underset{W \in L}{\arg\max}\, P(W \mid O)$$

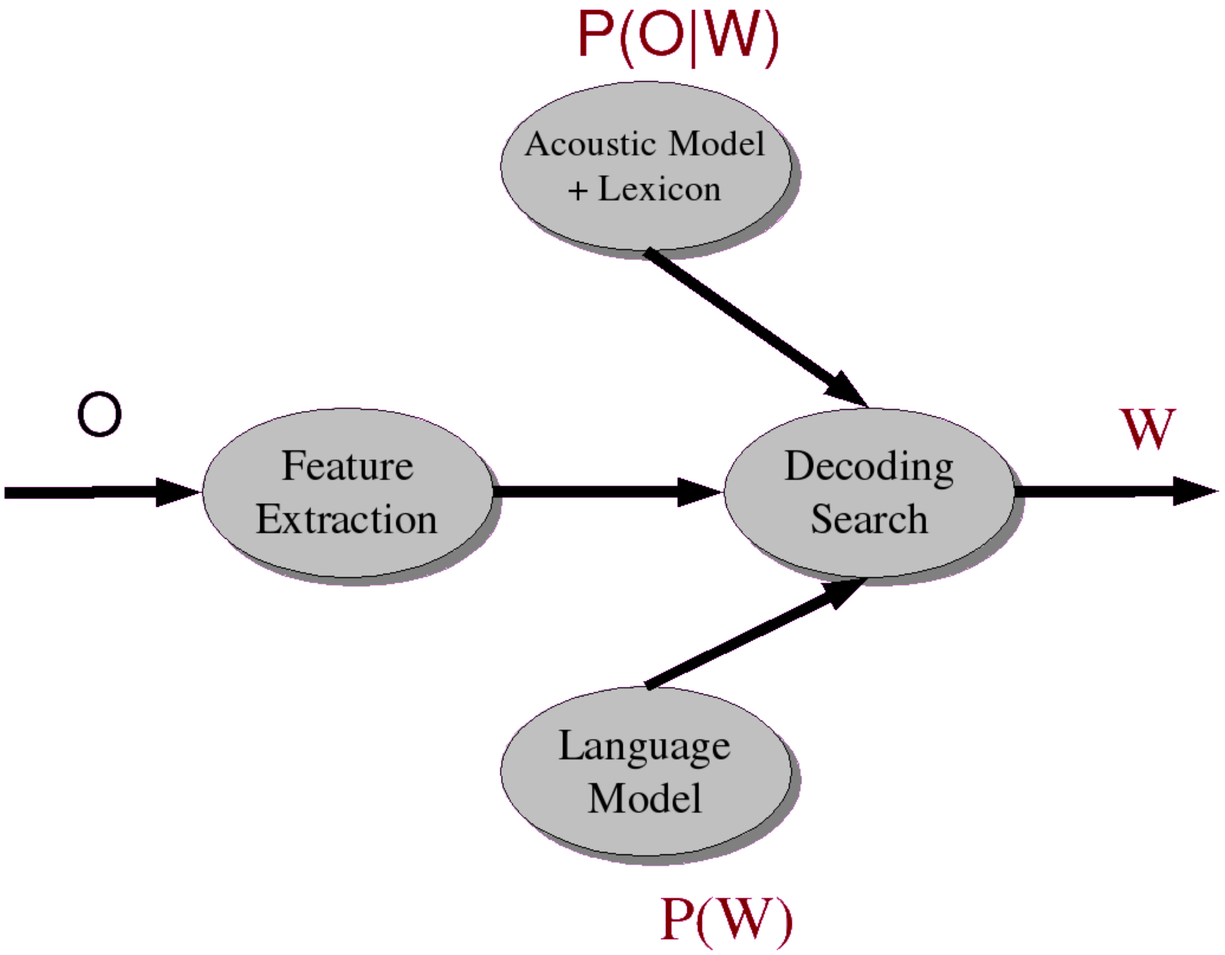
- We can use Bayes rule to rewrite this:

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

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

$$\hat{W} = \underset{W \in L}{\arg\max}\, P(O \mid W)P(W)$$

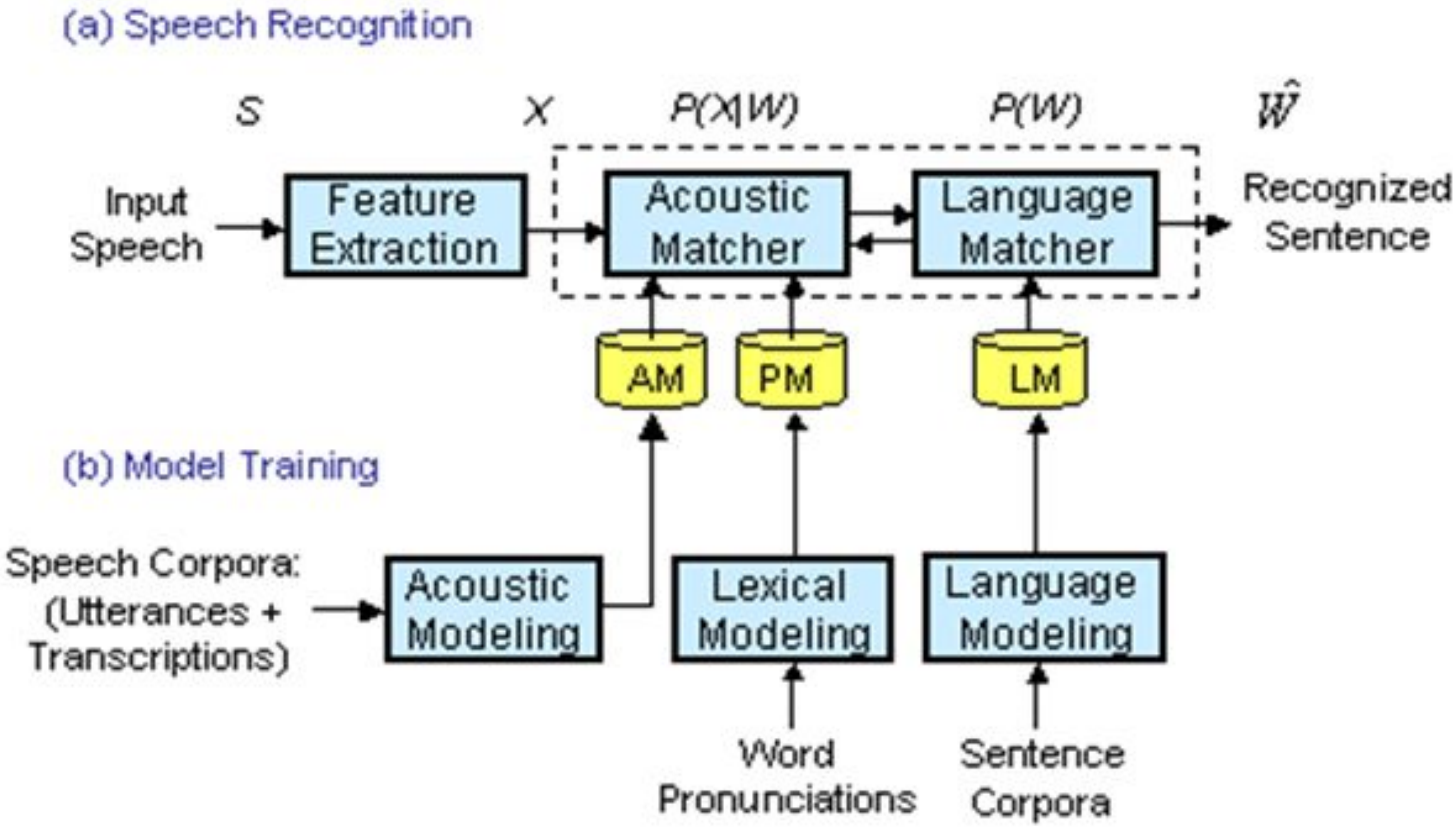
Speech Recognition Meets Noisy Channel: Acoustic Likelihoods and LM Priors



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

Speech Recognition (HMM+GMM)



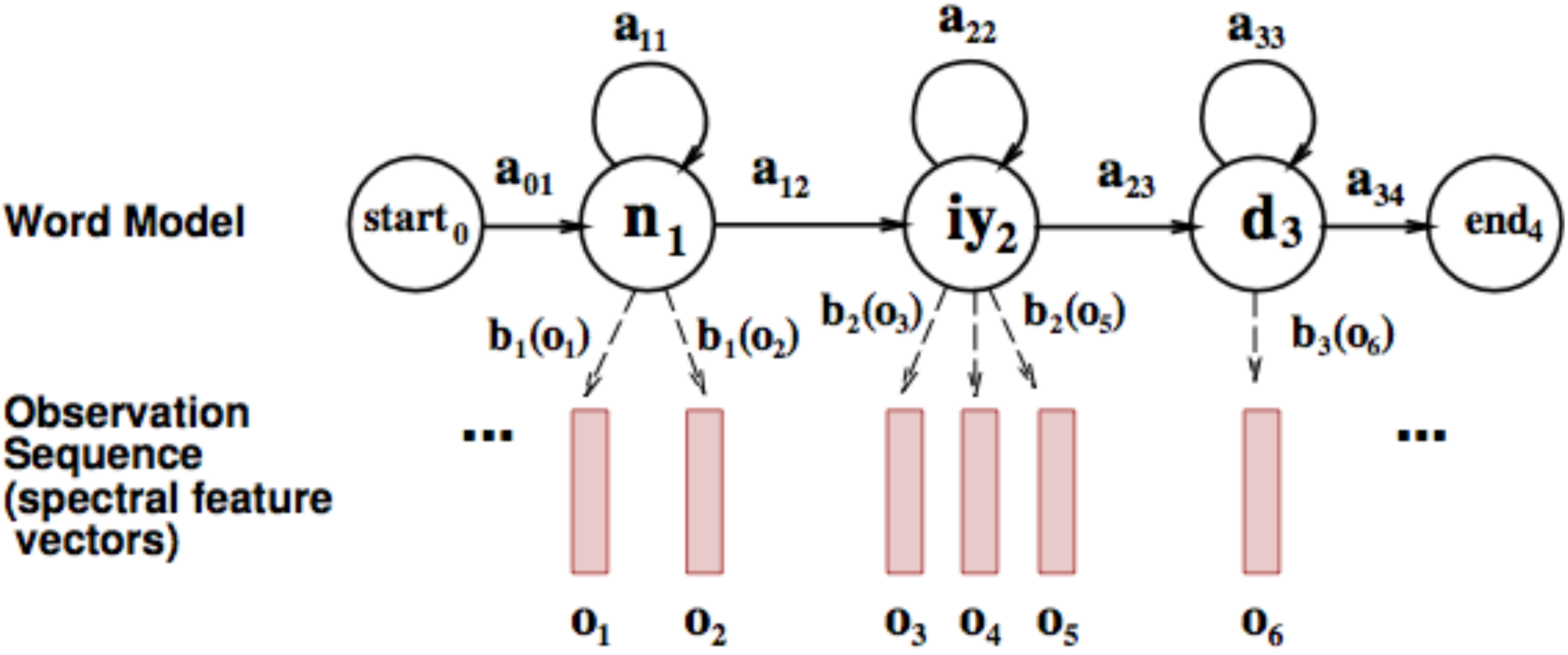
Training and Test Corpora

- 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



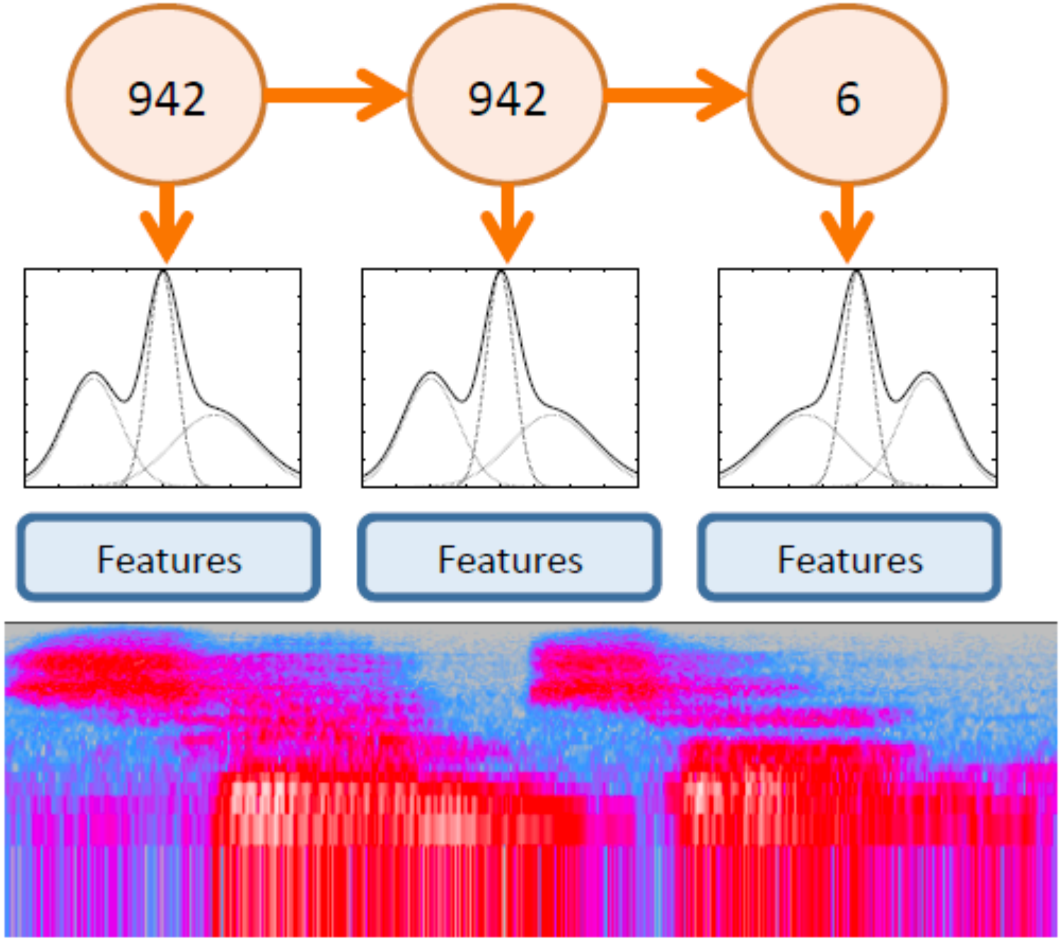
Word Model
Observation Sequence (spectral feature vectors)
start$_0$
a_{01}
a_{11}
n$_1$
a_{12}
a_{22}
iy$_2$
a_{23}
a_{33}
d$_3$
a_{34}
end$_4$
$b_1(o_1)$
$b_1(o_2)$
$b_2(o_3)$
$b_2(o_5)$
$b_3(o_6)$
...
...
o_1
o_2
o_3
o_4
o_5
o_6

Speech Recognition (HMM+GMM)

Transcription: Samson

Pronunciation: S – AE – M – S –AH – N

Sub-phones : 942 – 6 – 37 – 8006 – 4422 …



Hidden Markov Model (HMM):

Acoustic Model:

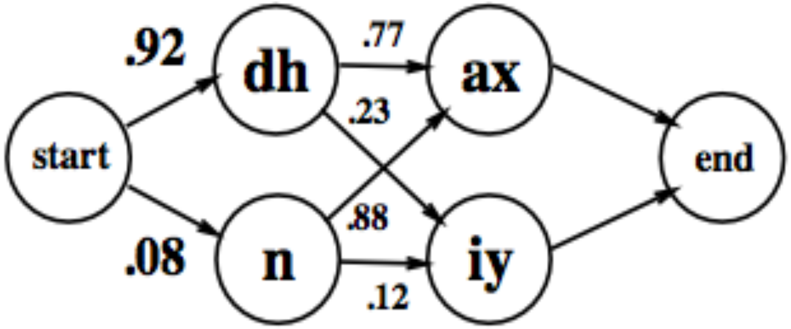
Audio Input:

GMM models:
P(x|s)
x: input features
s: HMM state

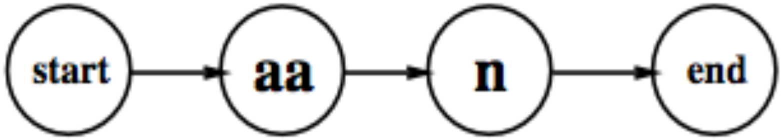
Building the Pronunciation Model

- 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

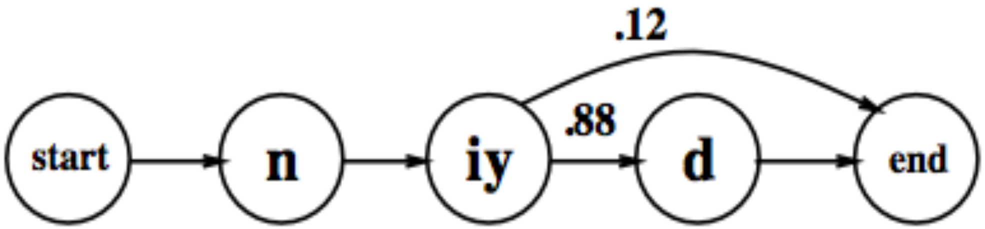
ASR Lexicon: Markov Models for Pronunciation



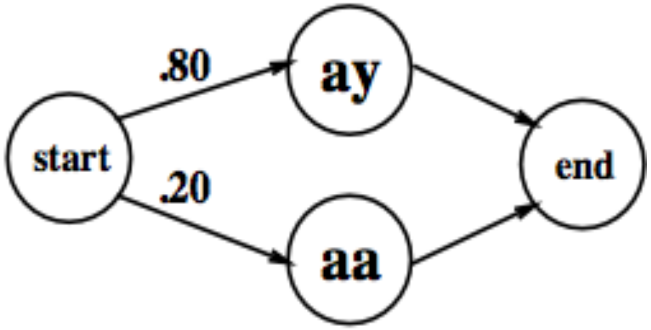
Word model for "the"



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
– 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?