Speech Recognition

An historical perspective on the business and technology

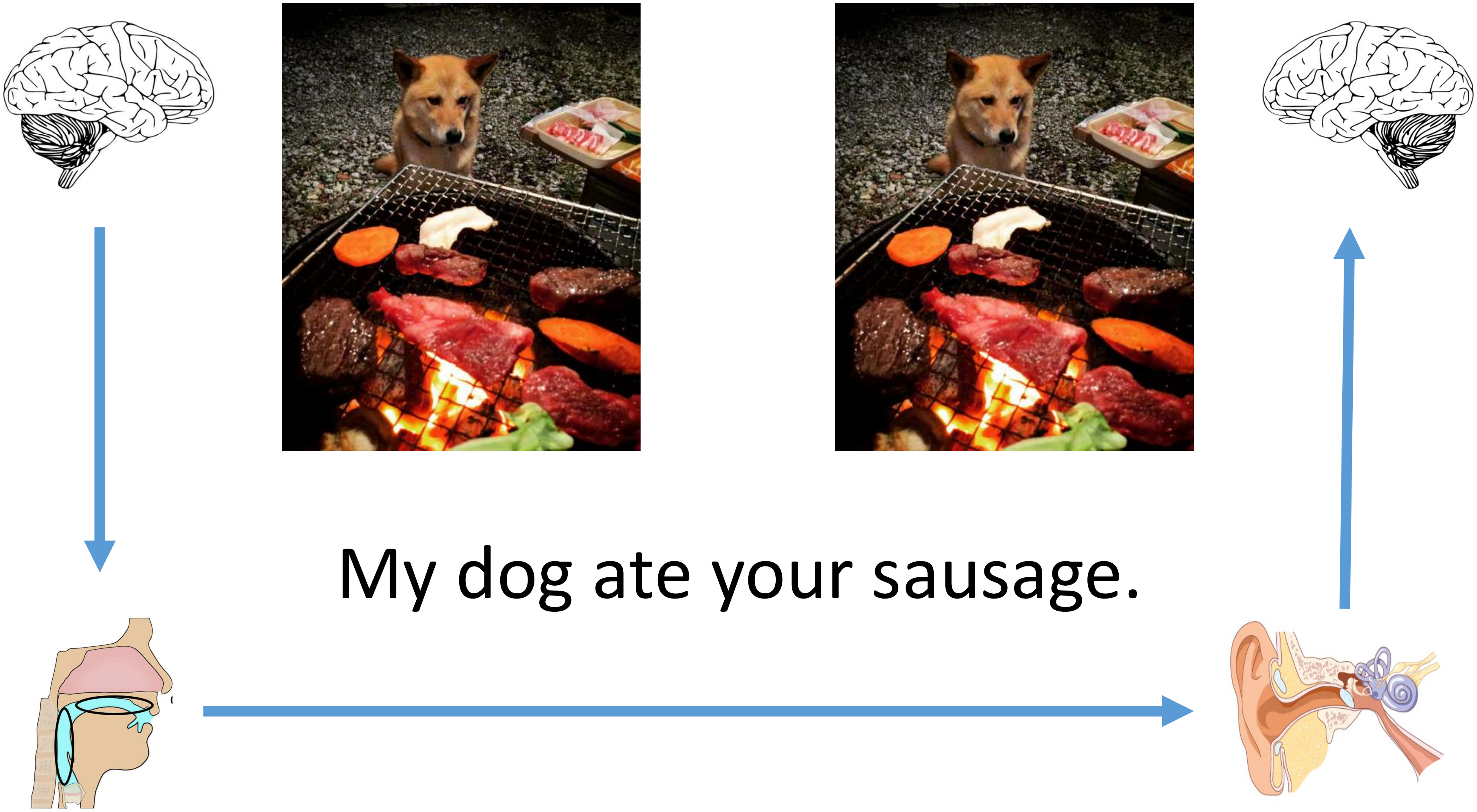
With thanks (again!) to my former SpeechWorks colleagues:

Blade Kotelly, MIT
Sol Lerner, Nuance
Mike Phillips, Sense Labs
Roberto Pieraccini, Google
John Nguyen, ScreenEx

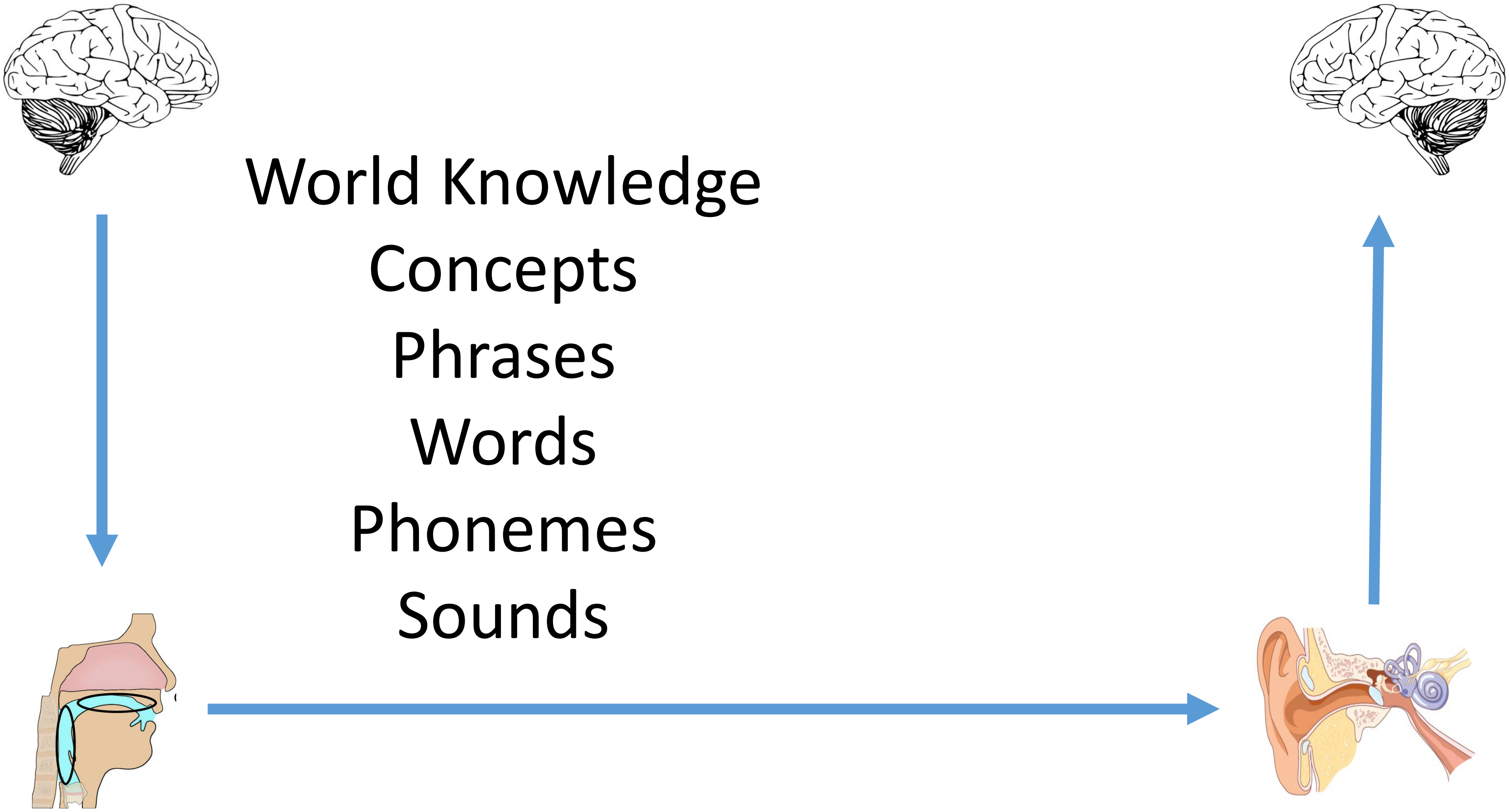
The ideas are all theirs; the misinterpretations, errors and omissions are all mine.

Audio Test

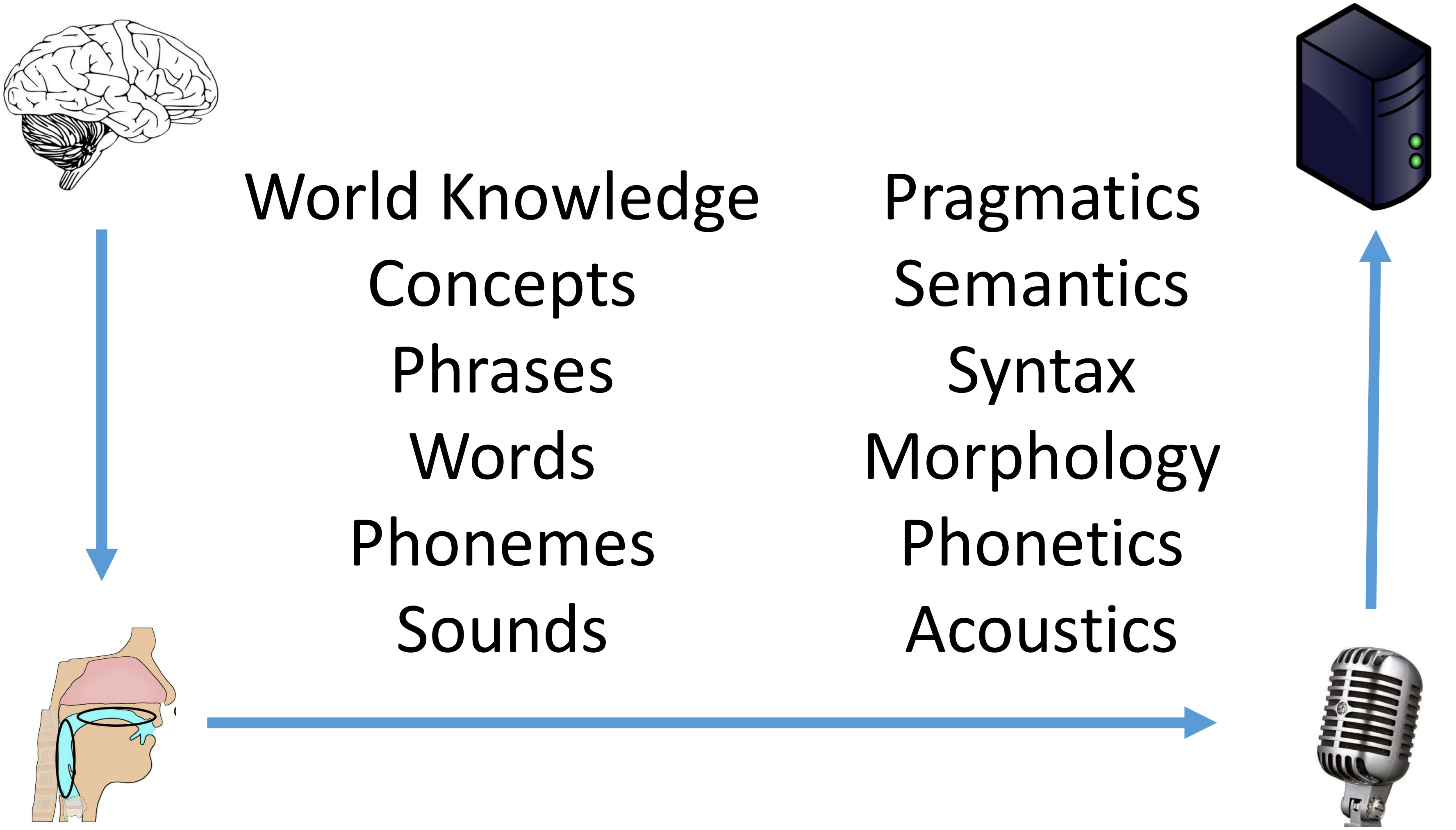
Type what I say into the chat window



My dog ate your sausage.



World Knowledge
Concepts
Phrases
Words
Phonemes
Sounds



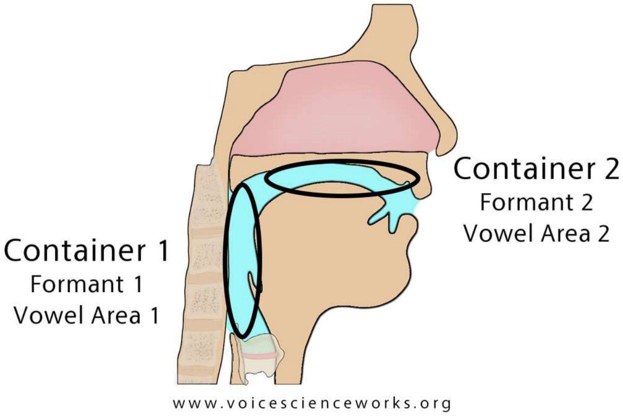
World Knowledge	Pragmatics
Concepts	Semantics
Phrases	Syntax
Words	Morphology
Phonemes	Phonetics
Sounds	Acoustics

1952 Bell Labs AUDREY

AUtomatic Digit REcognition Y

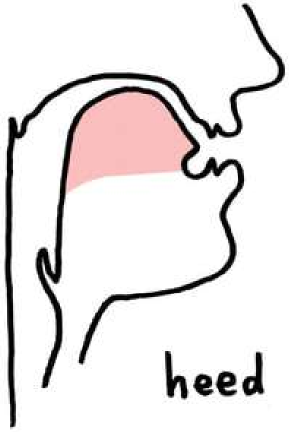
- Discrete digits
- Microphone
- Single trained speaker
- Formant-based pattern matching

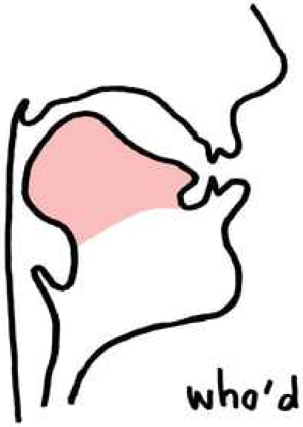
Formants

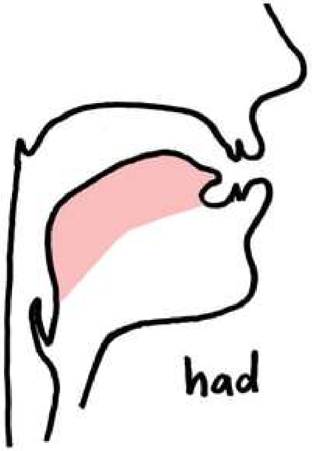


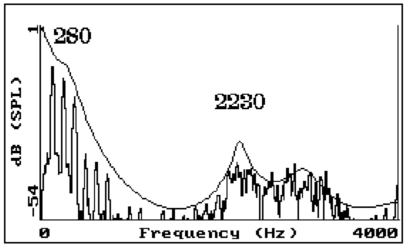
Container 2
Formant 2
Vowel Area 2

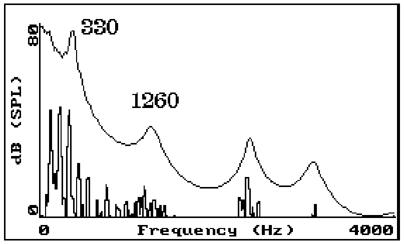
Container 1
Formant 1
Vowel Area 1

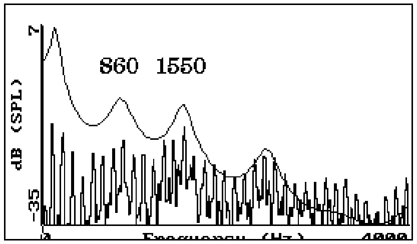




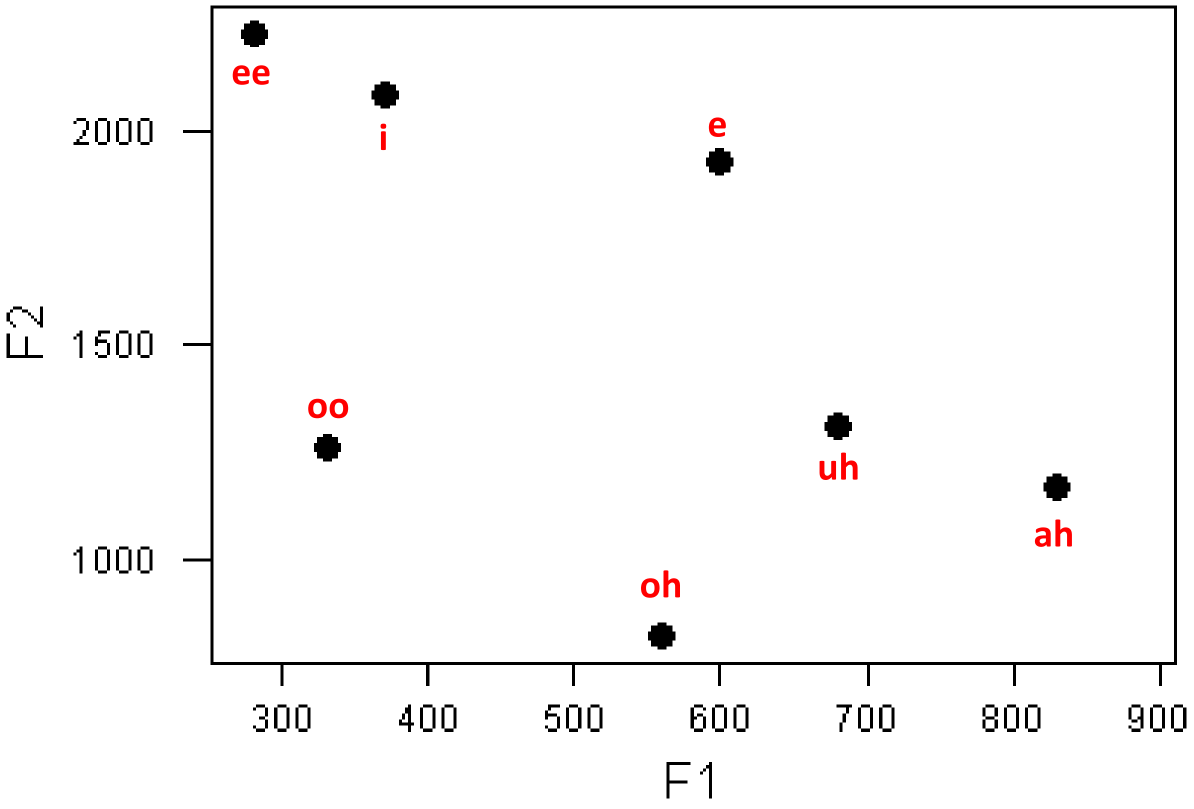








one	wuhn
two	too
three	three
four	fohr
five	faɪv
six	siks
seven	sevuhn
eight	eɪt
nine	naɪn
zero	ten



1952 Bell Labs AUDREY

AUtomatic Digit REcognition Y

- Discrete digits
- Microphone
- Single speaker
- Format-based pattern matching

- 97% accurate
- 7-digit phone number = 80%
- 6-foot rack with vacuum tubes

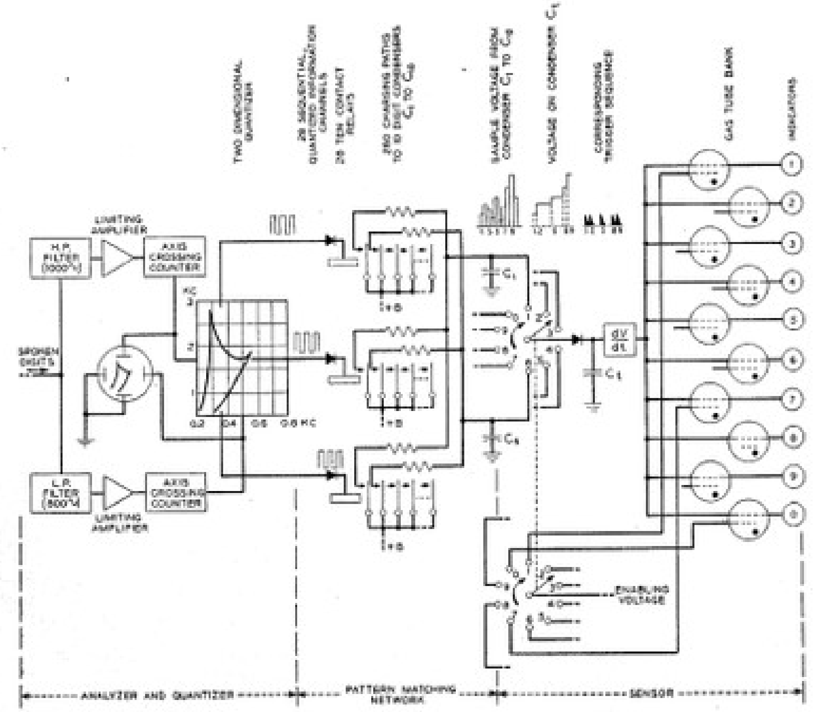
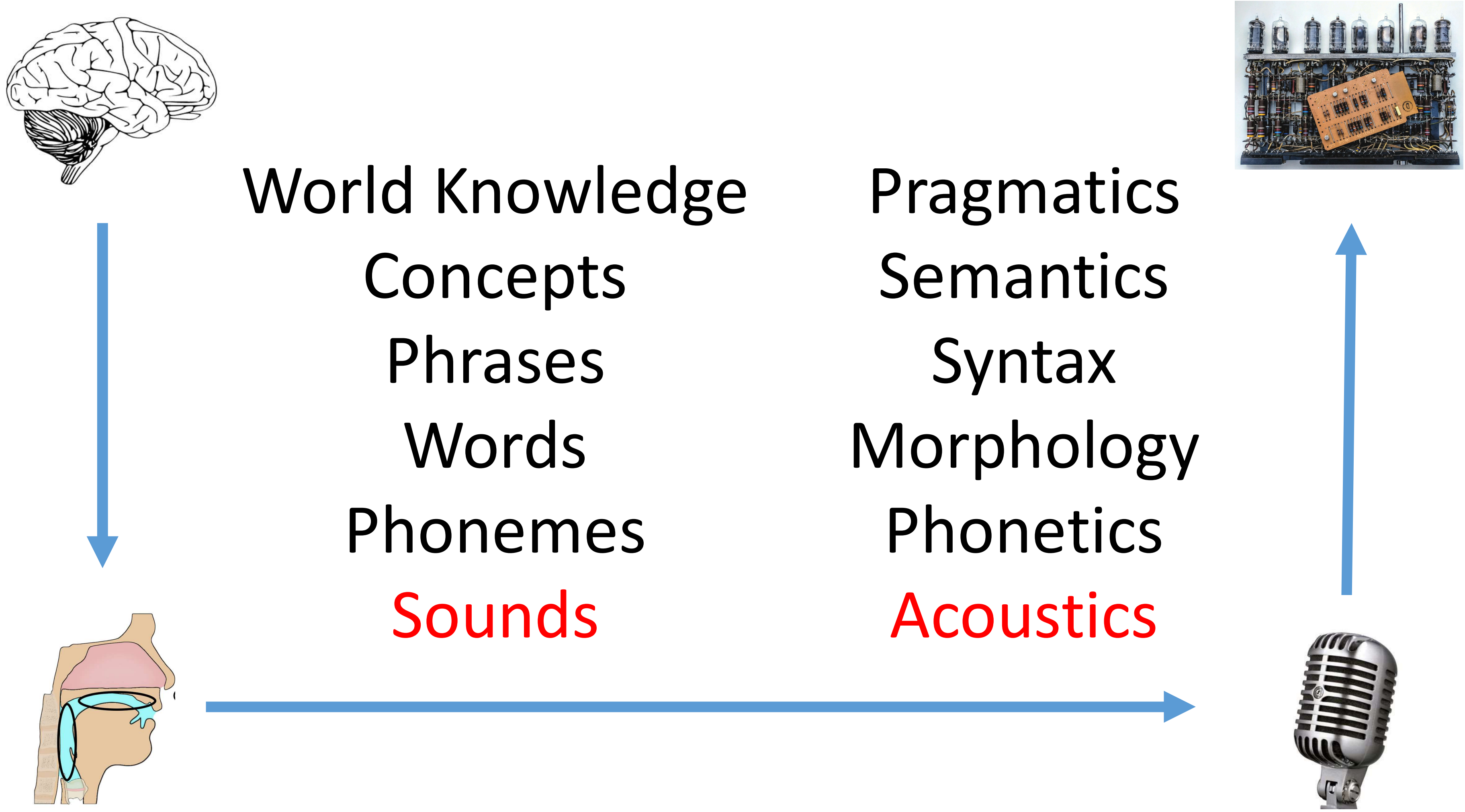


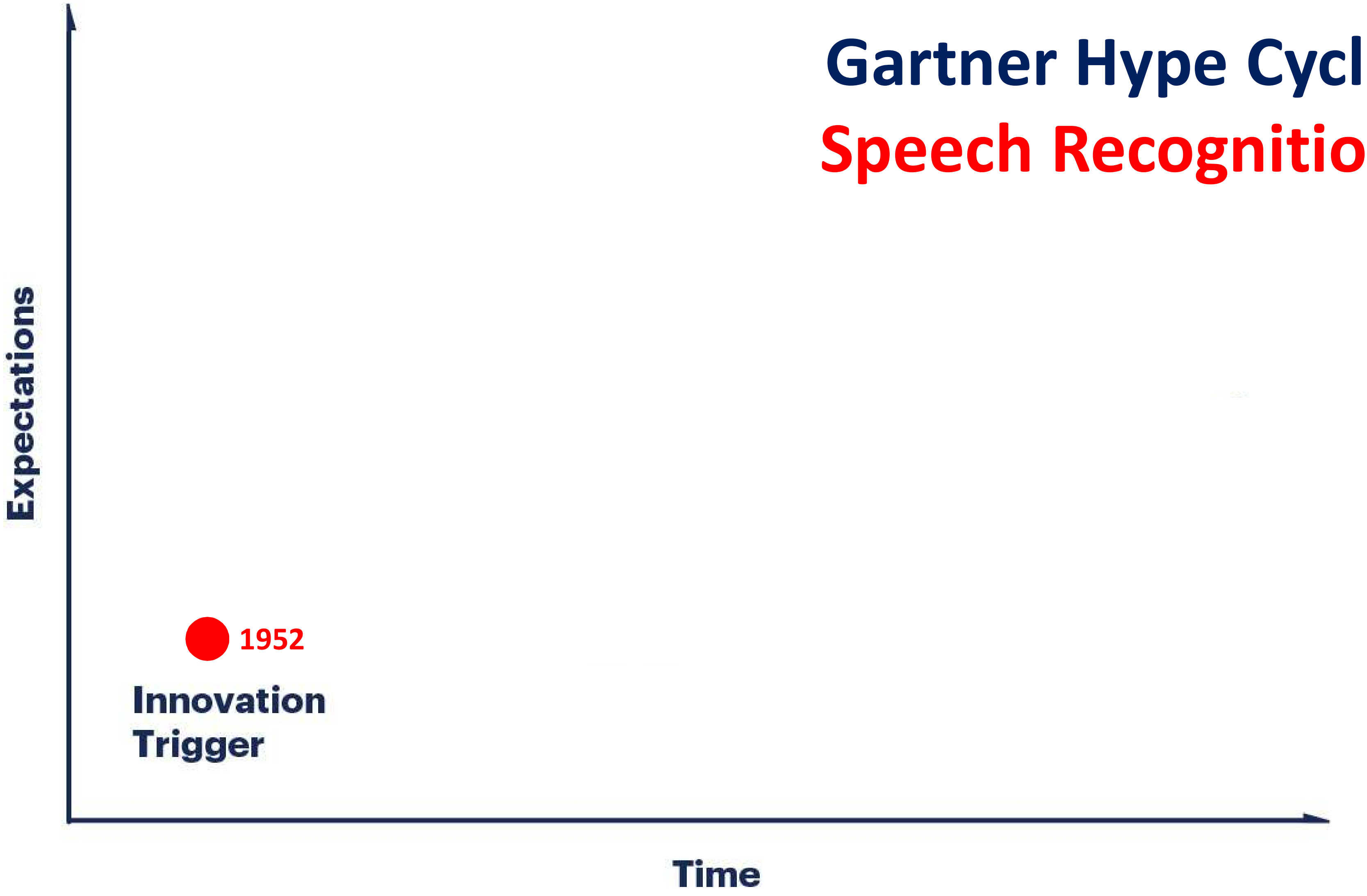
Fig. 1. Block schematic of digit recognizer circuits.



World Knowledge	Pragmatics
Concepts	Semantics
Phrases	Syntax
Words	Morphology
Phonemes	Phonetics
Sounds	Acoustics

Gartner Hype Cycle

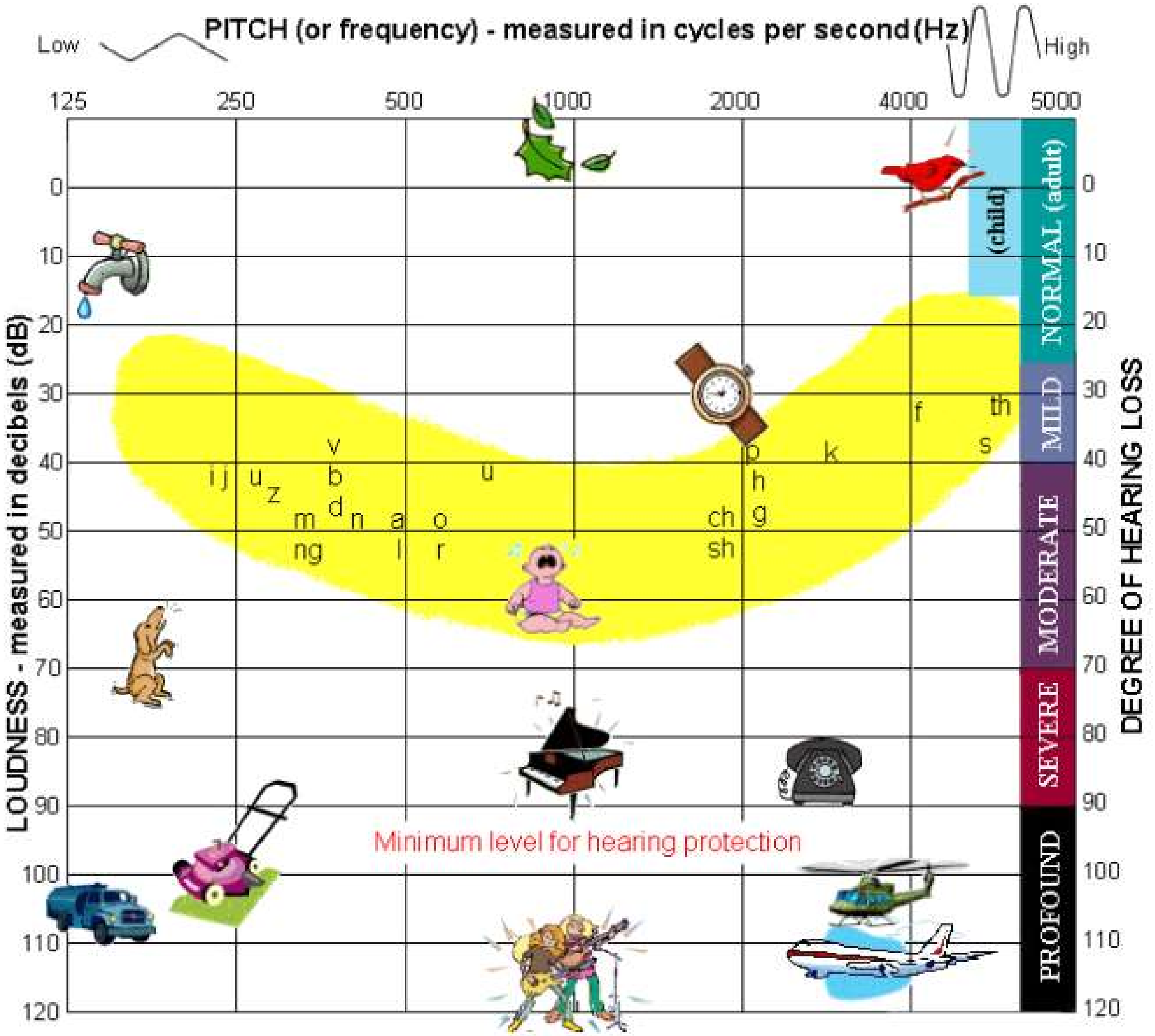
Speech Recognition



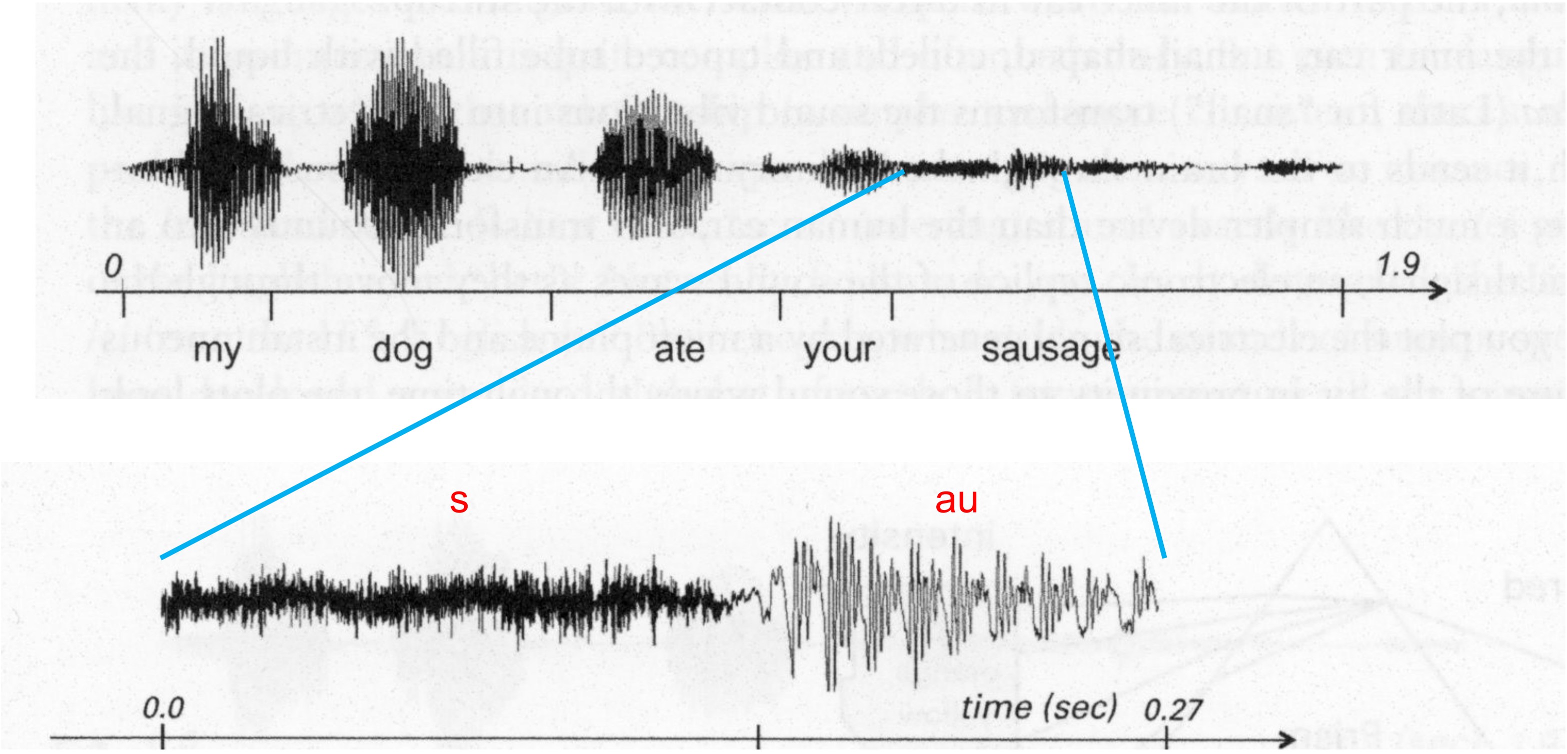
1961 Shoebox

- 16 words
- Fits in a shoebox
- Demo at 1962 Seattle World's Fair
- 3 analog filters (Low, Medium, High)
- Pattern match sequence HMH, LM, …

https://youtu.be/rQco1sa9AwU

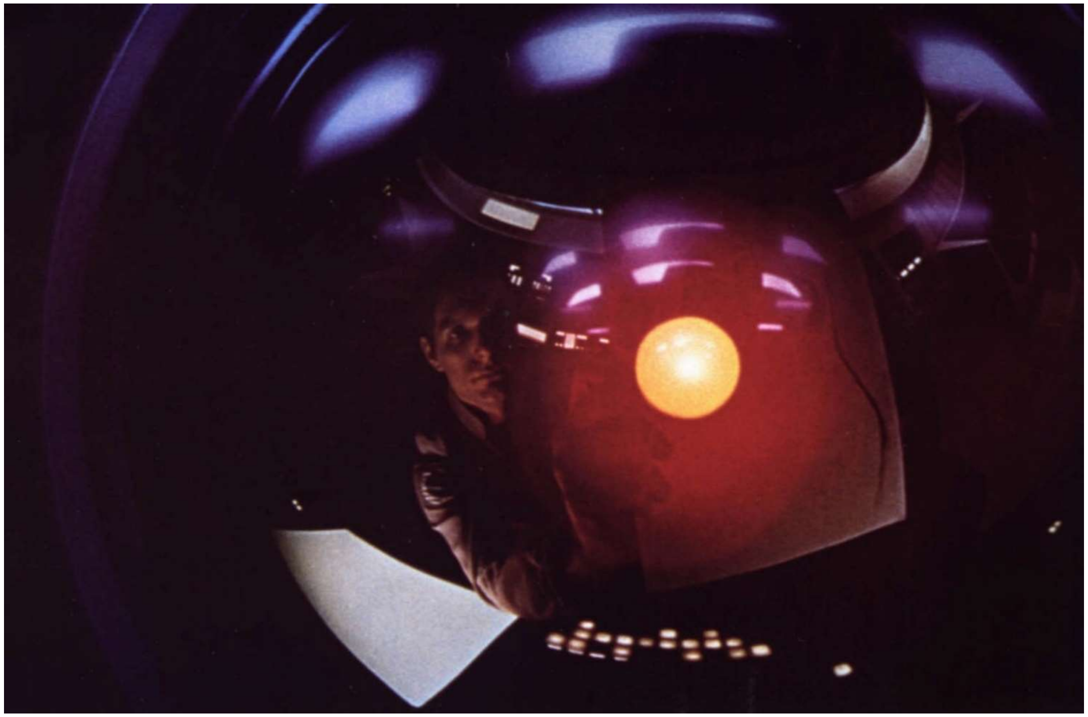

Low
PITCH (or frequency) - measured in cycles per second (Hz)
High
125
250
500
1000
2000
4000
5000
LOUDNESS - measured in decibels (dB)
0
10
20
30
40
50
60
70
80
90
100
110
120
i j
u
z
m
ng
v
b
d
n
a
l
o
r
u
ch
sh
p
h
g
k
f
s
th
(child)
NORMAL (adult)
MILD
MODERATE
SEVERE
PROFOUND
DEGREE OF HEARING LOSS
0
10
20
30
40
50
60
70
80
90
100
110
120
Minimum level for hearing protection

Sound



6 N. Lindgren, "Machine Recognition of Human Language," IEEE Spectrum 2, Nos. 3 and 4 (1965).
7 T. Marill, "Automatic Recognition of Speech." IRE Trans. Human Factors Electron HFE-2, 34-^38 (1961).
8 E. E. David, Jr.. "Artificial Auditory Recognition in Telephony," IBM J. Res. Develop., 2, 294-309 (1958).
9 E. E. David, Jr.. and O. G. Selfridge, "Eyes and Ears for Computers," Proc. IRE 50, 1093-1101 (1962).
10 M. A. Saposhkov, "The Speech Signal in Cybernetics and Communications," transl. by Joint Publications Res. Service, JPRS 28, 117 (1965).
11 J. L. Flanagan, Speech Analysis Synthesis and Perception (SpringerVerlag, Berlin, 1965).
12 V. A. Kozhevnikov and L. A. Chistovich, "Speech: Articulation and Perception" transl. by Joint Publications Res. Service, JPRS 30, 543 (1965).
11 H. K. Davis, R. Biddulph, and S. Balashek, "Automatic Recognition of Spoken Digits." J. Acoust. Soc. Amer. 24. 637-642 (1952).
14 H. Dudley and S. Balashek, "Automatic Recognition of Phonetic Patterns in Speech," J. Acoust. Soc. Amer. 30, 721-733 (1958).
15 J. Wiren and H. L. Stubbs, "Electronic Binary Selection System for Phoneme Classification," J. Acoust. Soc. Amer. 28, 1082-1091 (1956).
16 H. F. Olson and H. Belar, "Phonetic Typewriter III," J. Acoust. Soc. Amer. 33. 1610-1615 (1961).
17 D. B. Fry, "Theoretical Aspects of Mechanical Speech Recognition" J. British Inst. Radio Eng. 19, 211-219 (1959). Also, P. Denes, "The Design and Operation of the Mechanical Speech Recognizer at University College London," ibid., pp. 219-229.
18 J. W. Forgie and C. D. Forgie, "Results Obtained from a Vowel Recognition Computer Program," J. Acoust. Soc. Amer. 31, 1480-1489 (1959).
19 J. Suzuki and K. Nakata, "Recognition of Japanese Vowels Preliminary to the Recognition of Speech," J. Radio Res. Lab., Tokyo 8, No. 37, 193-212 (1961) .
20 T. Sakai and S. Doshita. "The Phonetic Typewriter." Proc. IFIP Congr. Munich, Infor. Processing Aug.-Sept. (1962).
21 K. Nagata, Y. Kato, and S. Chiba, "Spoken Digit Recognizer for Japanese Language," Nippon Elec. Co., Res. Develop. No. 6 (1963).
22 T. B. Martin. A. L. Nelson, and H. J. Zadell, "Speech Recognition by Feature Abstraction Techniques," Tech. Rep. No. AL-TDR-64-176 (AD604526), Air Force Avionics Lab. (1964).
23 J. W. Falter, "Feature Abstraction: An Approach to Speech Recognition," Proc. Nat. Aerospace Elec. Conf., IEEE. 192-198 (1965).
24 J. Gazdag, "A Method of Decoding Speech," Tech. Rep. No. 9, AFOSR-66-2385 (AD641132), Univ. of Illinois, June (1966).
26 L. Gilli and A. R. Meo, "Sequential System for Recognizing Spoken Digits in Real Time," Acustica 19, (1967/68).
25 P. W. Ross, "A Limited-Vocabulary Adaptive Speech-Recognition System," J. Audio Eng. Soc. 15, 414-419 (1967).
27 P. B. Denes and M. V. Mathews, "Spoken Digit Recognition Using Time-Frequency Pattern Matching," J. Acoust. Soc. Amer. 32, 1450-1455
28 G. Sebestyen, "Automatic Recognition of Spoken Numerals," J. Acoust. Soc. Amrr. 32, 1516 (A) (1960).
29 W. F. Meeker, A. L. Nelson, and P. B. Scott, "Voice to Teletype Code Converter Research Program, Part II, Experimental Verification of a Method to Recognize Phonetic Sounds," Tech. Rep. No. ASD-TR 61-666, Wright-Patterson AFB, Ohio (1962).
30 ti. Gold, "Word-Recognition Computer Program," Res. Lab. Electron, MIT Rep. No. 452, June (1966).
31 P. N. Sholtz and R. Bakis, "Spoken Digit Recognition using VowelConsonant Segmentation," J. Acoust. Soc. Amer. 34, 1-5 (1962).
32 G. W. Hughes, "The Recognition of Speech by Machine," MIT Res. Lab. Electron, Tech. Rep. No. 395 (1961).
33 G. W. Hughes and J. F. Hemdal, "Speech Analysis," Rep. AFCRL65-681. (P137552), Purdue Univ. (1965).
34 L. R. Talbert, G. F. Groner, J. S. Koford, R. J. Brown, P. R. Low, and C. H. Mays, "A Real-Time Adaptive Speech Recognition System," Tech. Rep. No. 6760-1 (ASD-TDR-63-660) (P133441), prepared by Stanford Electron Lab.
35 J. A. Dammann, "Application of Adaptive Threshold Elements to the Recognition of Acoustic-Phonetic States," J. Acoust. Soc. Amer. 38, 213-223 (1965).
36 J. H. King and C. J. Tunis, "Some Experiments in Spoken Word Recognition," IBM J. Res. Develop. 10, 65-79 (1966).
37 L. Fraipont, "Voice Actuated Address Mechanism," Elec. Ass., Inc., Rep. No. 3 (AD 633711) (1966).
38 C. F. Teacher and C. F. Piotrowski, "Voice Sound Recognition," Tech. Rep. No. RADC-TR-65-184. Rome Air Development Ctr. (AD 619964) July (1965).
39 C. F. Teacher, H. Kellett, and L. Focht, "Experimental, Limited Vocabulary, Speech Recognizer," IEEE Int. Conv. Record, Part III, 169-173 (1967).
40 M. Weiss, "A Study of Critical Instant Sampling of Speech Parameters for Automatic Recognition of Spoken Words," Rome Air Develop. Ctr., Rep. No. RADC-TR-65-371 (AD 38380) July (1966).
41 M. W. Cannon, "A Method of Analysis and Recognition for Voiced Vowels," IEEE Trans. Audio Electroacoust. AU-16, 154-158 (1968;.
42 D. R. Reddy, "An Approach to Computer Speech Recognition by Direct Analysis of the Speech Wave," Computer Sci. Dep., Stanford Univ., Tech. Rep. No. CS49, Sept. (1966) .
43 J. N. Shearme and P. E. Leach, "Some Experiments with a Simple Word Recognition System," IEEE Trans. Audio Electroacoust. AU-16, 256-261 (1968).
44 R. F. Purton, "Speech Recognition using Autocorrelation Analysis," IEEE Trans. Audio Electroacoust. AU-16, 235-239 (1968).
45 S. H. Lavington and L. E. Rosenthal, "Some Facilities for Speech Processing by Computer," Computer J. (British Computer Society, London, NW 1) 9, 330-339 (1967).
46 W. Bezdel, "Discriminators of Sound Classes for Speech Recognition Purposes," Conf. Speech Commun. Proc., AFCRL, Sec. B8, 104-108 (1967).

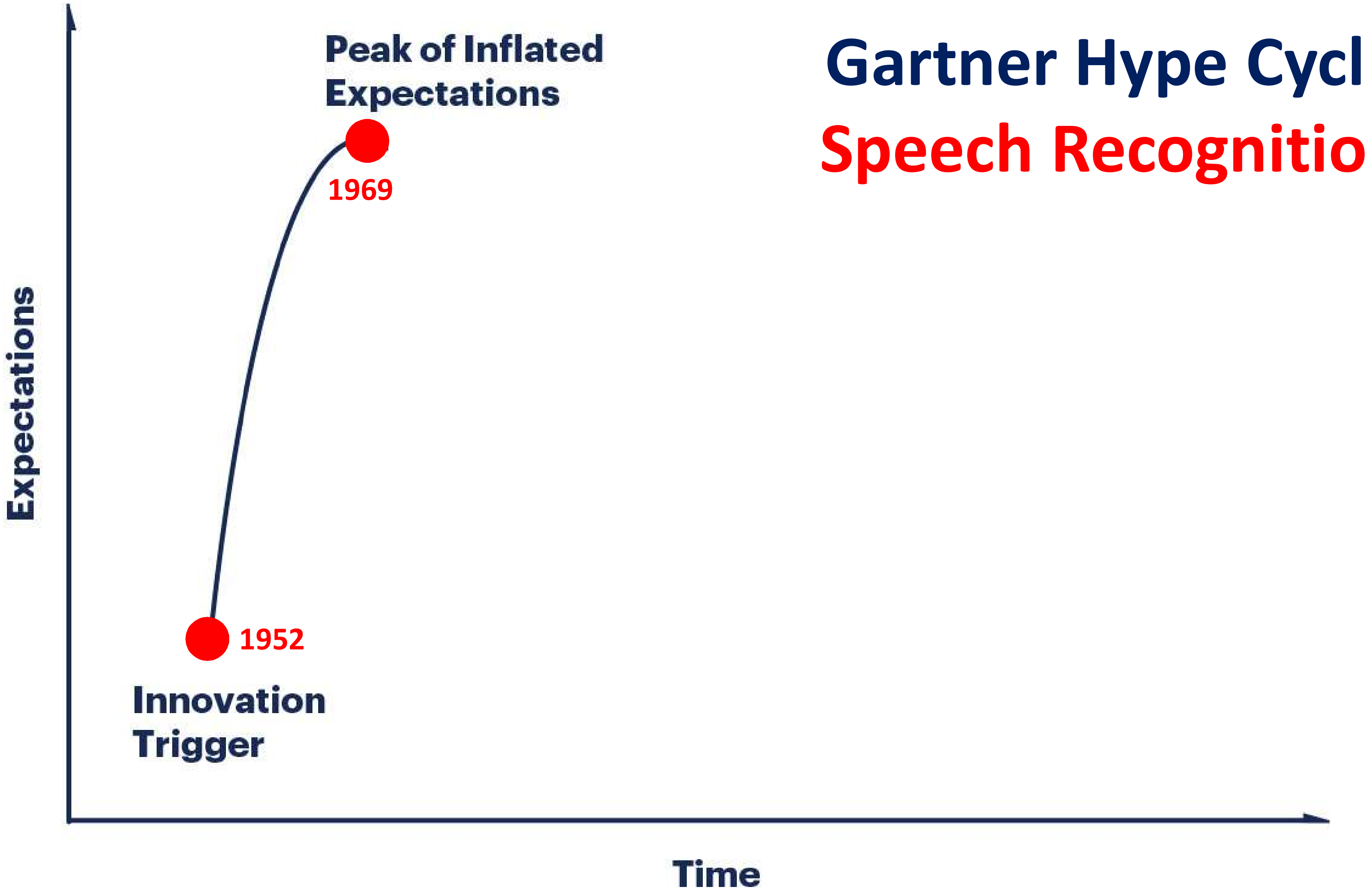
1968 Clarke & Kubrick HAL



https://www.youtube.com/watch?v=9W5Am-a_xWw

Gartner Hype Cycle

Speech Recognition



1969 J R Pierce, Bell Labs

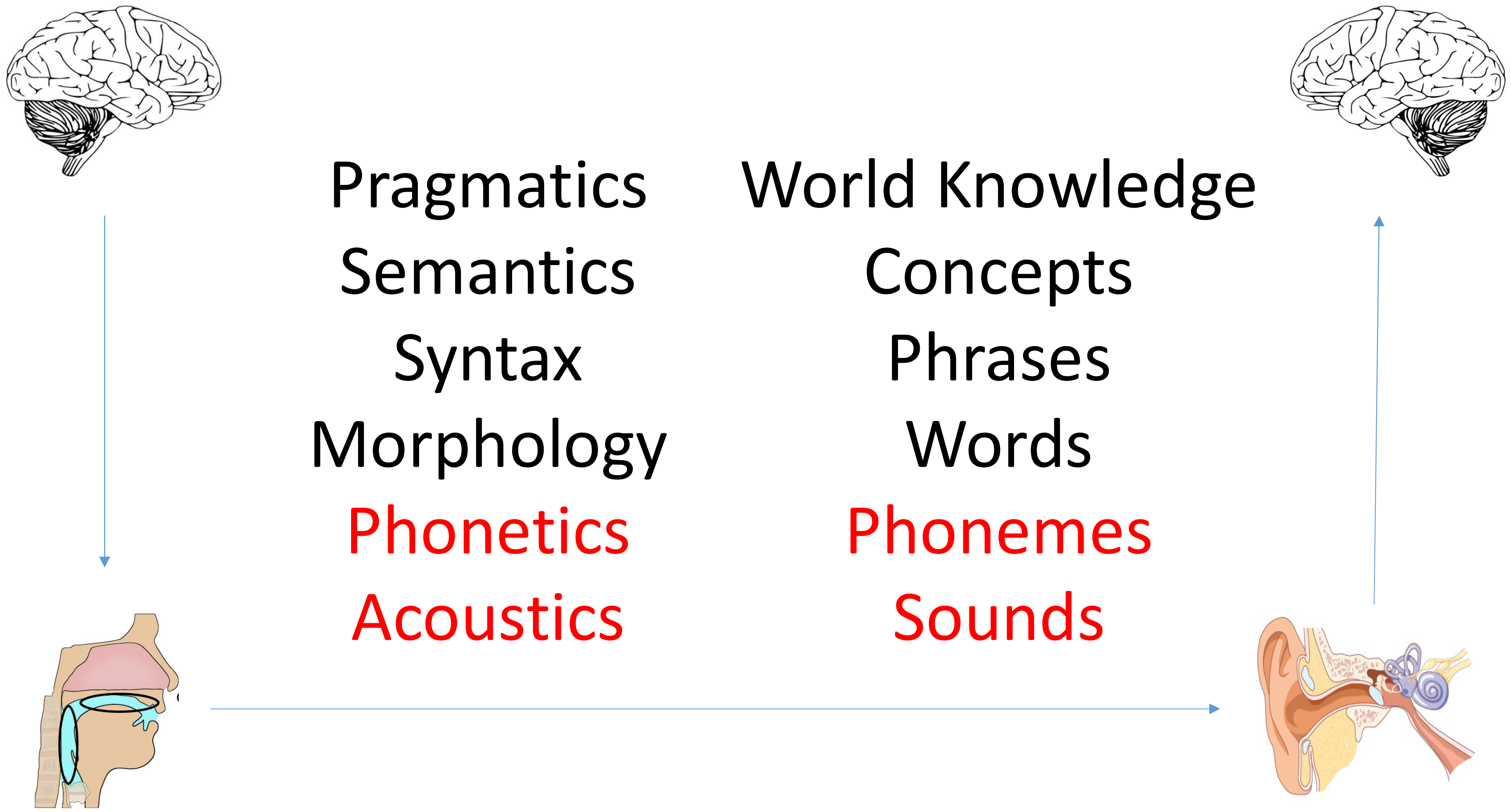
WHITHER SPEECH RECOGNITION: Journal of the Acoustical Society of America

Speech recognition has glamor. Funds have been available. Results have been less glamorous.

General-purpose speech recognition seems far away. Special-purpose speech recognition is severely limited.

"When we listen to a person speaking much of what we think we hear is supplied from our memory.” – W. James, 1889.

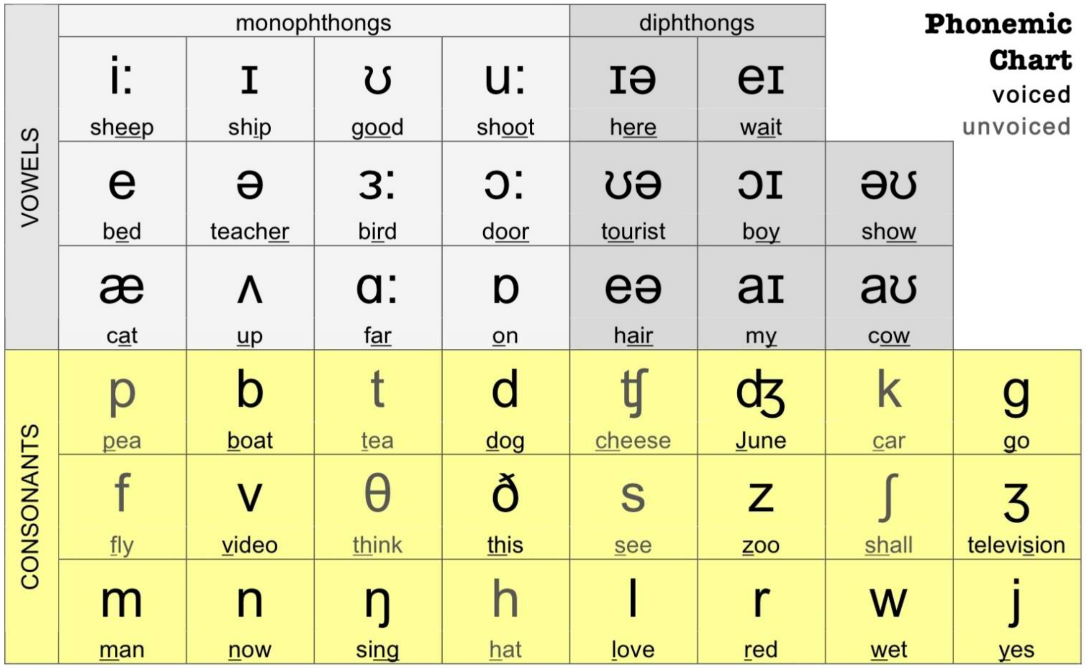
It would seem appropriate for people to ask themselves why they are working in the field and what they can expect to accomplish.



Pragmatics	World Knowledge
Semantics	Concepts
Syntax	Phrases
Morphology	Words
Phonetics	Phonemes
Acoustics	Sounds

Phonemes

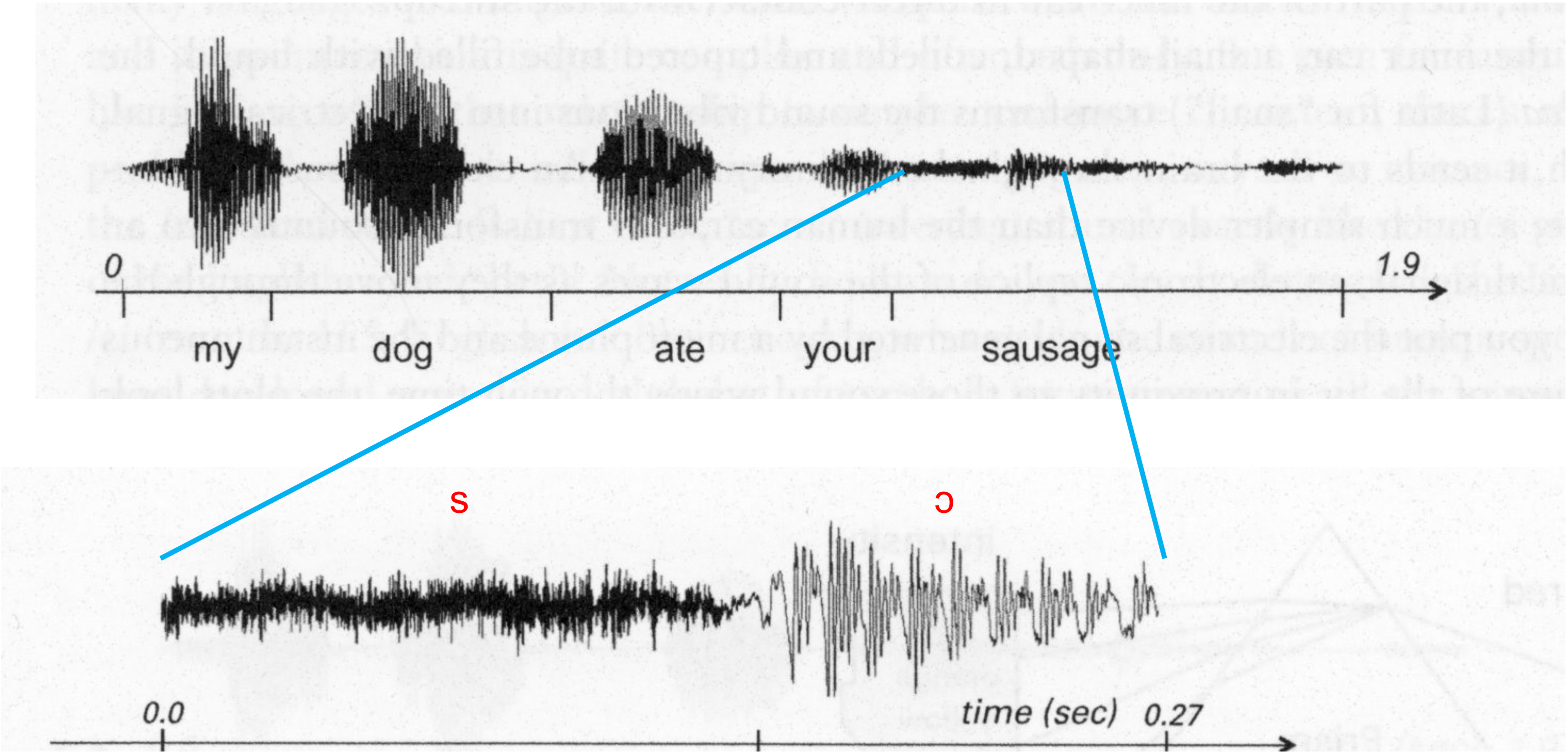
“Perceptually distinct units of sound <u>in a specified language</u> that distinguish one word from another.”



Phonemic Chart
voiced
unvoiced

	monophthongs				diphthongs			
VOWELS	iː sheep	ɪ ship	ʊ good	uː shoot	ɪə here	eɪ wait		
	e bed	ə teacher	ɜː bird	ɔː door	ʊə tourist	ɔɪ boy	əʊ show	
	æ cat	ʌ up	ɑː far	ɒ on	eə hair	aɪ my	aʊ cow	
CONSONANTS	p pea	b boat	t tea	d dog	tʃ cheese	dʒ June	k car	g go
	f fly	v video	θ think	ð this	s see	z zoo	ʃ shall	ʒ television
	m man	n now	ŋ sing	h hat	l love	r red	w wet	j yes

Sound



1971 $15M SUR Funding

Speech Understanding Research

~~10K words, any speaker, any acoustic environment~~

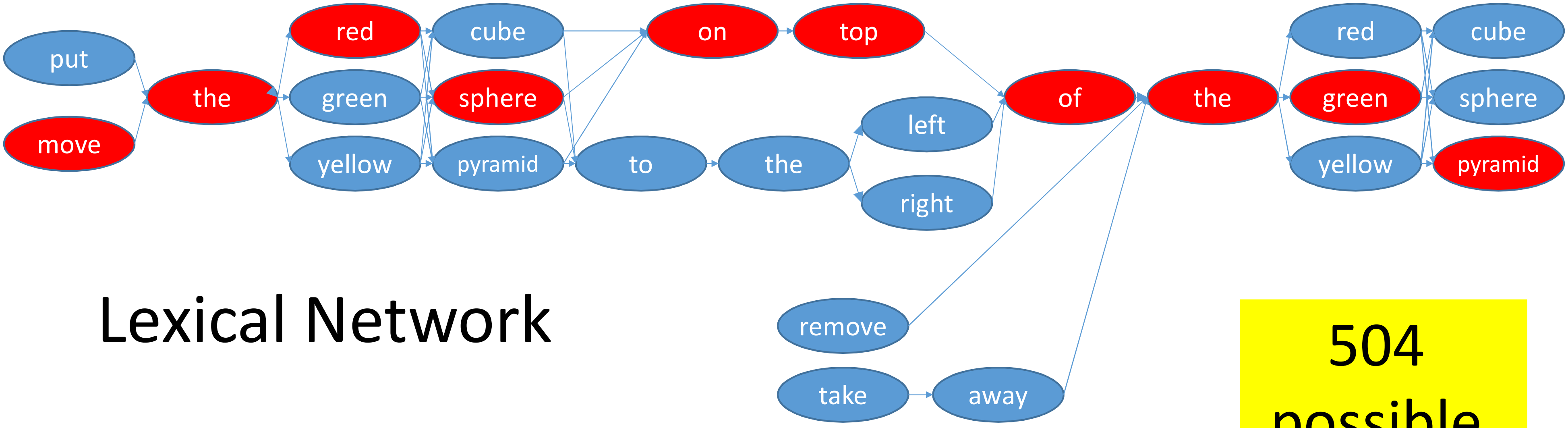
90% of **connected speech** utterances in “a few times real time”

1K words, selected cooperative speakers, quiet room

Can “train” system on each speaker (“speaker-dependent”)

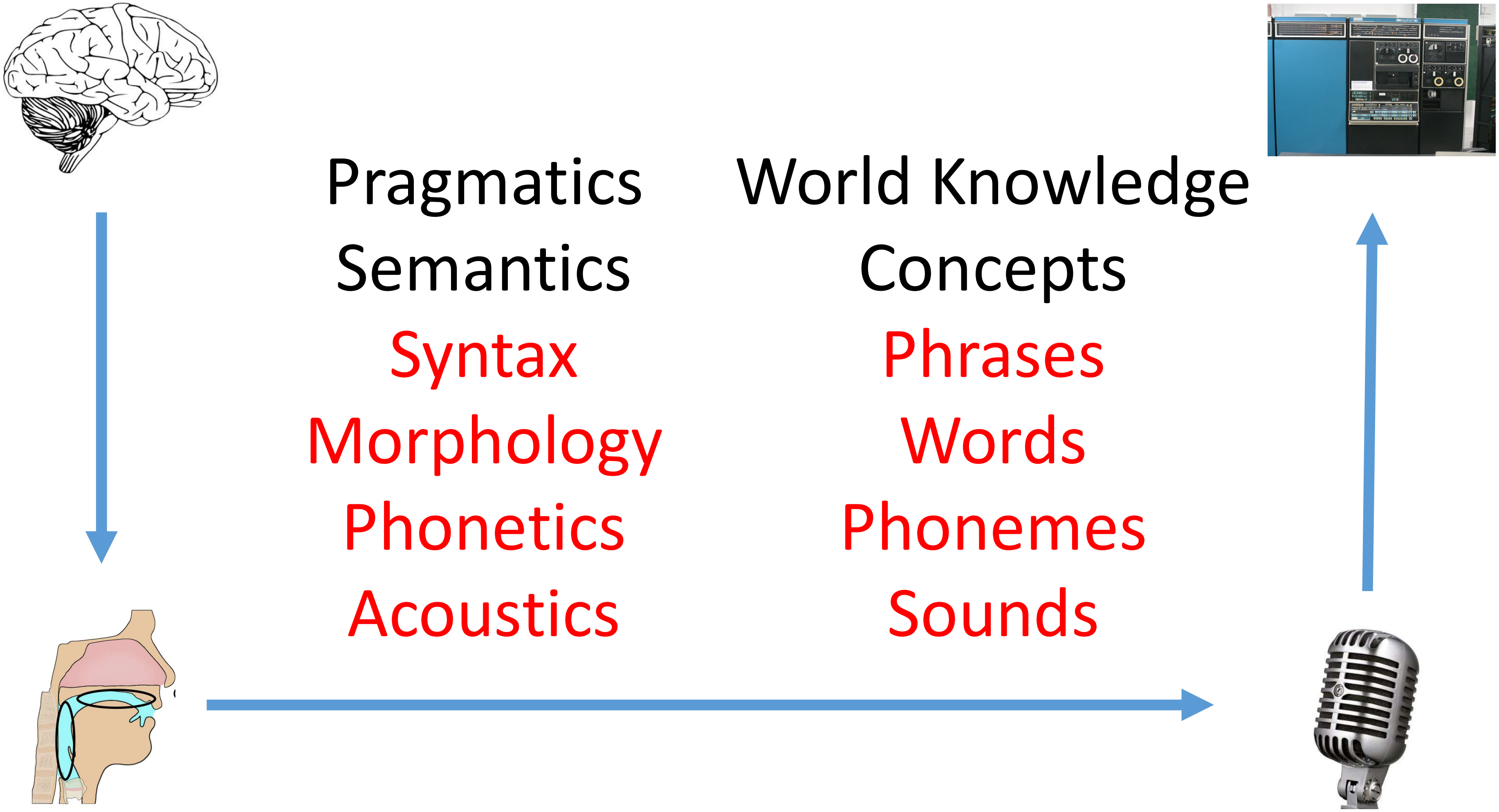
Fixed Grammars

“MOVE THE RED SPHERE ON TOP OF THE GREEN PYRAMID”



Lexical Network

504 possible sentences



Pragmatics	World Knowledge
Semantics	Concepts
Syntax	Phrases
Morphology	Words
Phonetics	Phonemes
Acoustics	Sounds

1971 $15M SUR Funding

Speech Understanding Research

SDC (Systems Development Corp)

BBN (Bolt, Beranek & Newman)

CMU Hearsay II

CMU Harpy

MIT

SRI (Stanford Research Institute)

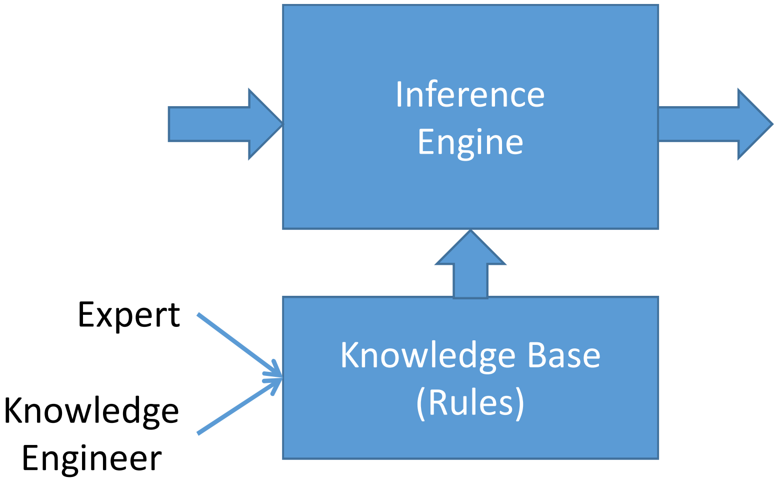
1971 $15M SUR Funding

Speech Understanding Research

SDC
BBN
CMU Hearsay II
CMU Harpy
MIT
SRI

Rules-based “AI”
Expert Systems

Expert System: sounds → phonemes → words



Inference Engine

Knowledge Base (Rules)

Expert

Knowledge Engineer

```
IF pause followed by voiced low energy
   IF pause >20 ms
      segment is a voiced stop
   ELSE
      segment is an unvoiced stop

IF stop followed by vowel
   IF formants rise to stationary value
      IF energy burst after stop is weak
         segment is b

IF first segment is d
   IF next segment is ɔ or oʊ
      IF next segment is g
         word is dog
```

1971 $15M SUR Funding

Speech Understanding Research

SDC
BBN
CMU Hearsay II

Rules-based “AI”
Expert System

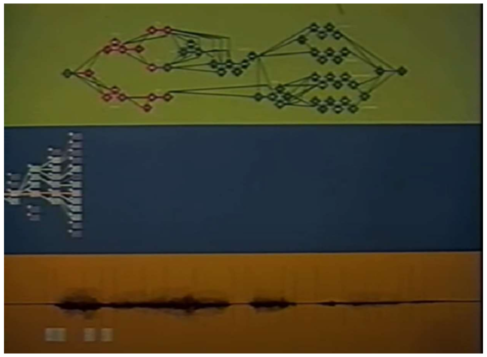
CMU Harpy

Template Matching

MIT
SRI

Template matching

1. Use heuristic rules to divide utterance into average 50 ms "segments"
2. Find best match(es) to 98 sound templates ("phones") for each segment
3. Constrain to universe of possible utterances with probabilities



https://www.youtube.com/watch?v=32KKg3aP3Vw

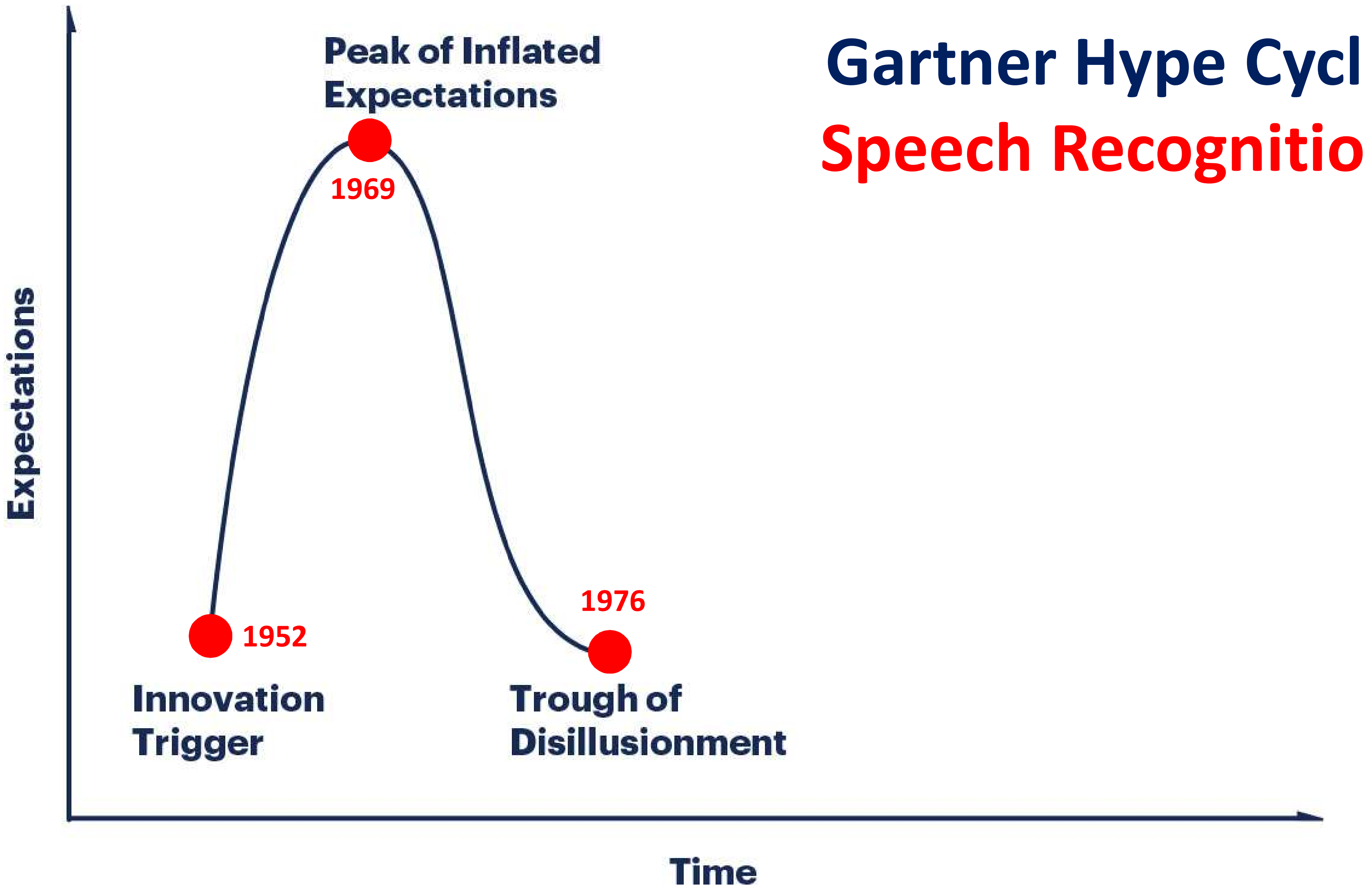
1976 $15M SUR Funding

90% accurate, few times real time

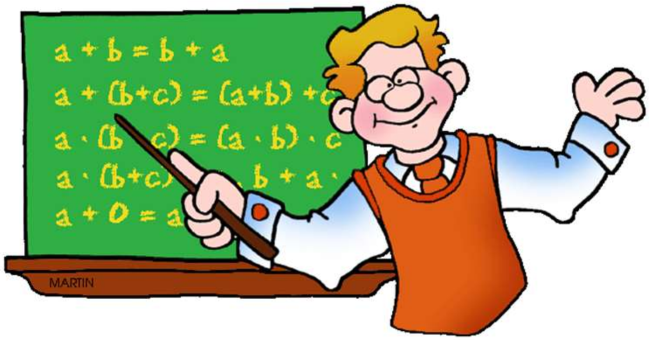
SDC	24% accurate
BBN	44%
CMU Hearsay II	74%
CMU Harpy	95%, but 80x real time
MIT	DNF
SRI	DNF

Gartner Hype Cycle

Speech Recognition

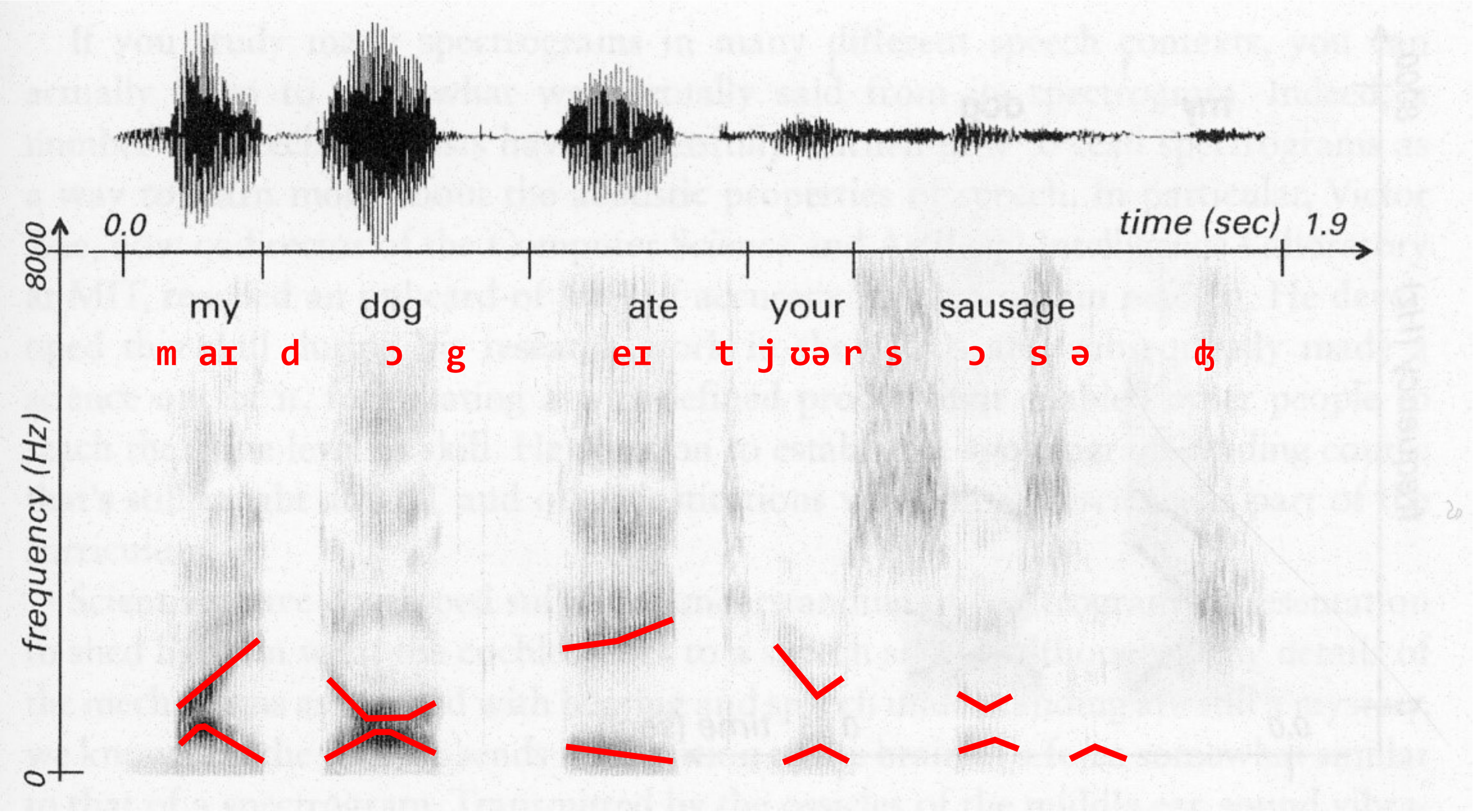


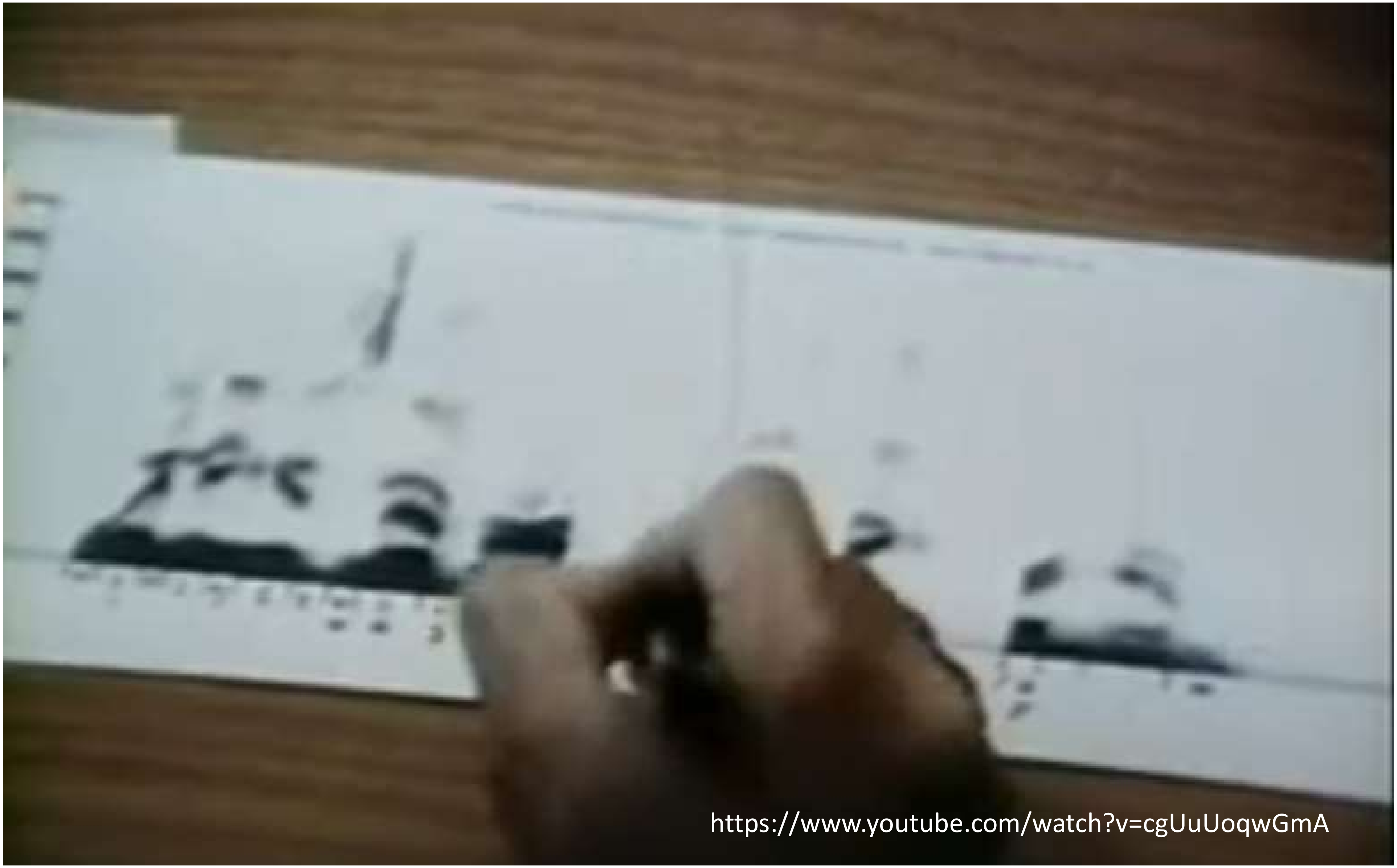
VS.



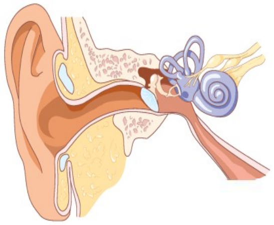
+

The Signal – as a Spectrogram

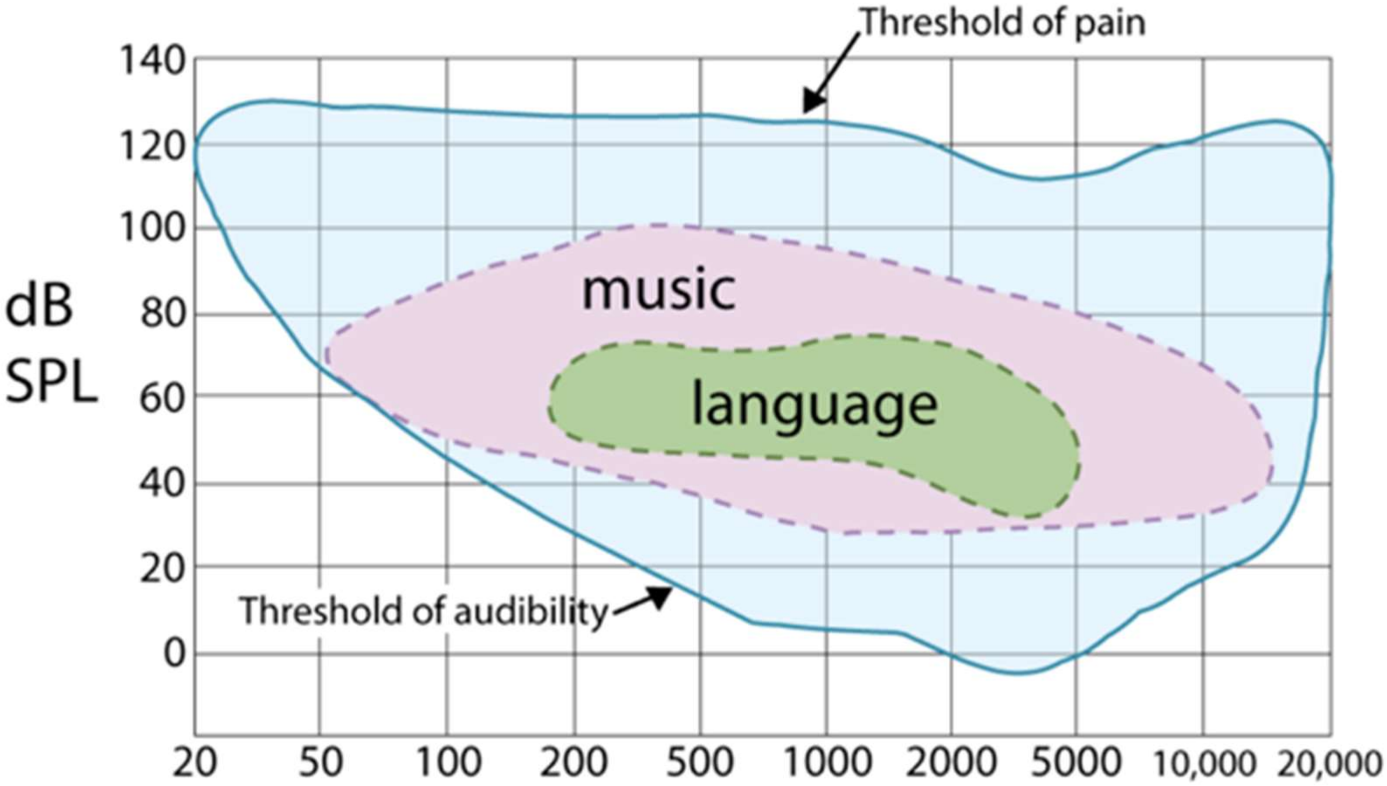


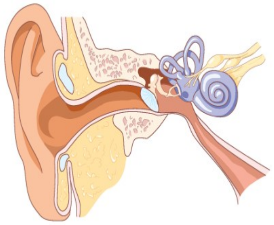


https://www.youtube.com/watch?v=cgUuUoqwGmA

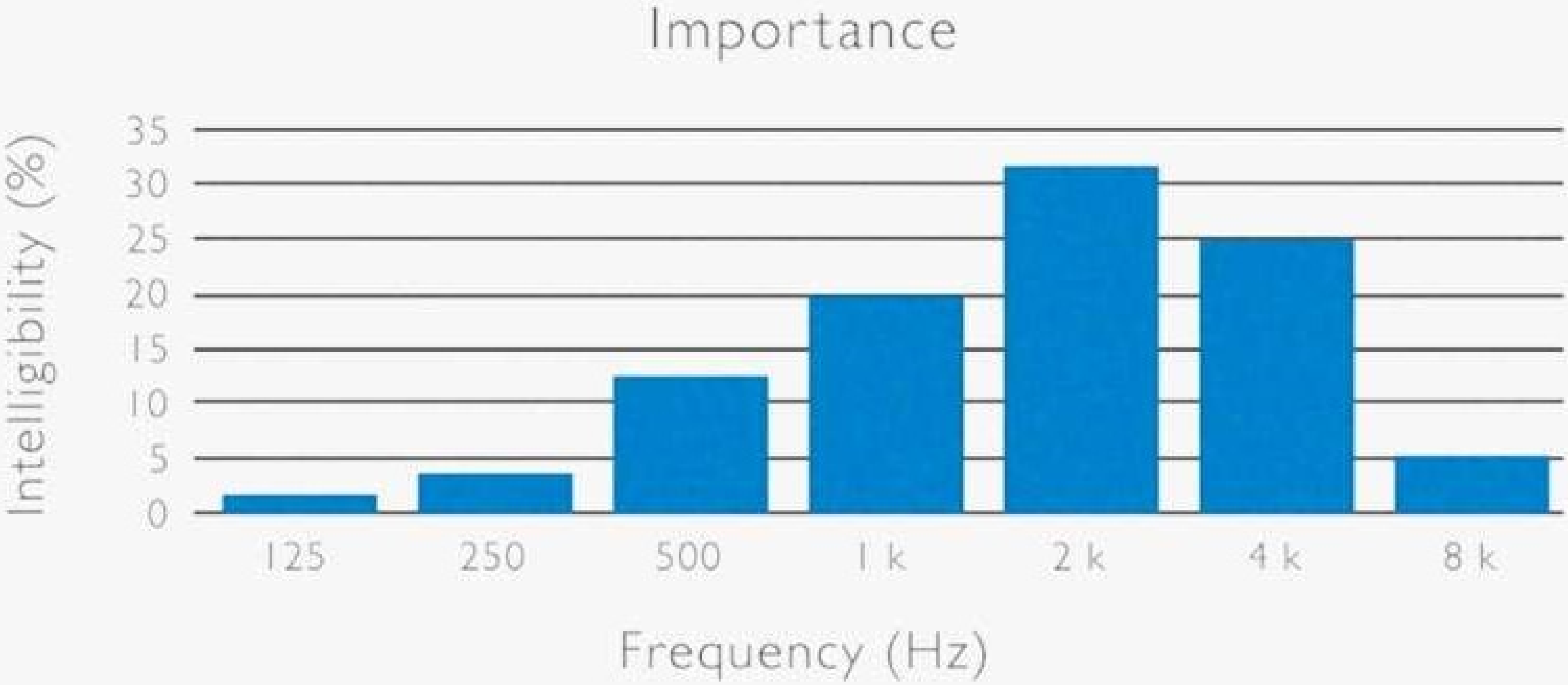


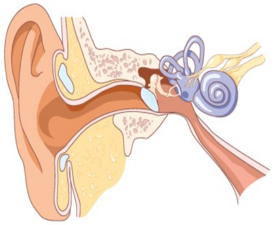
Human Hearing





Human Hearing



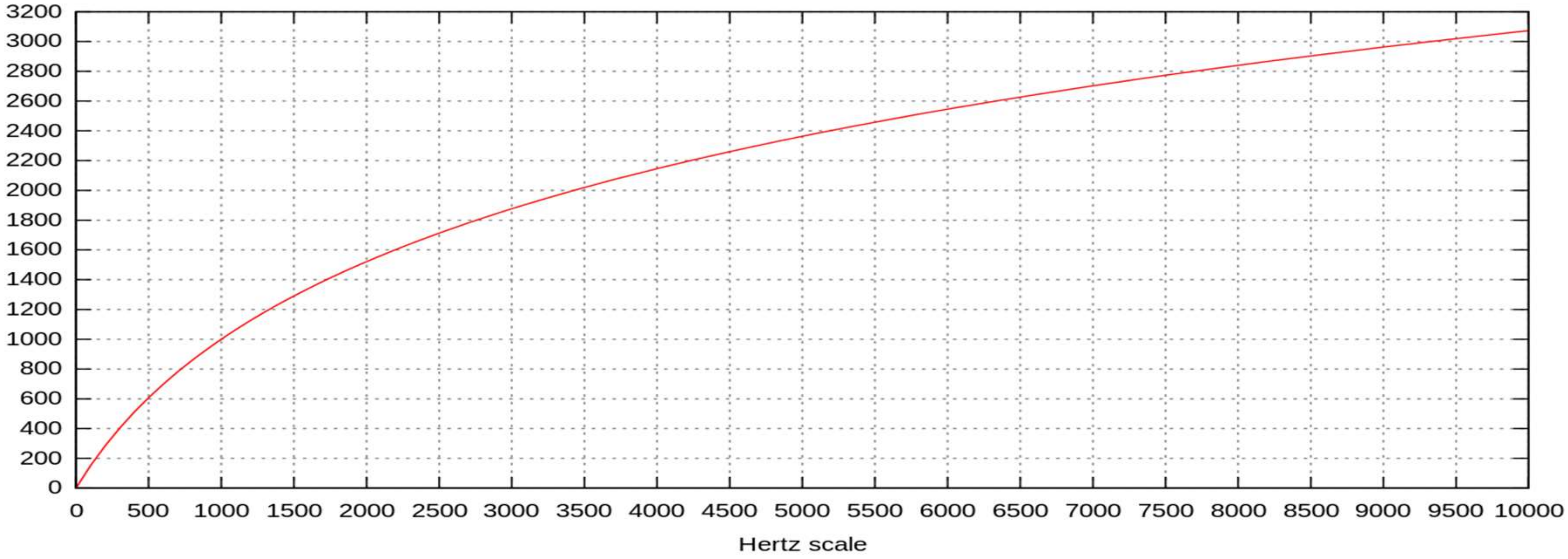


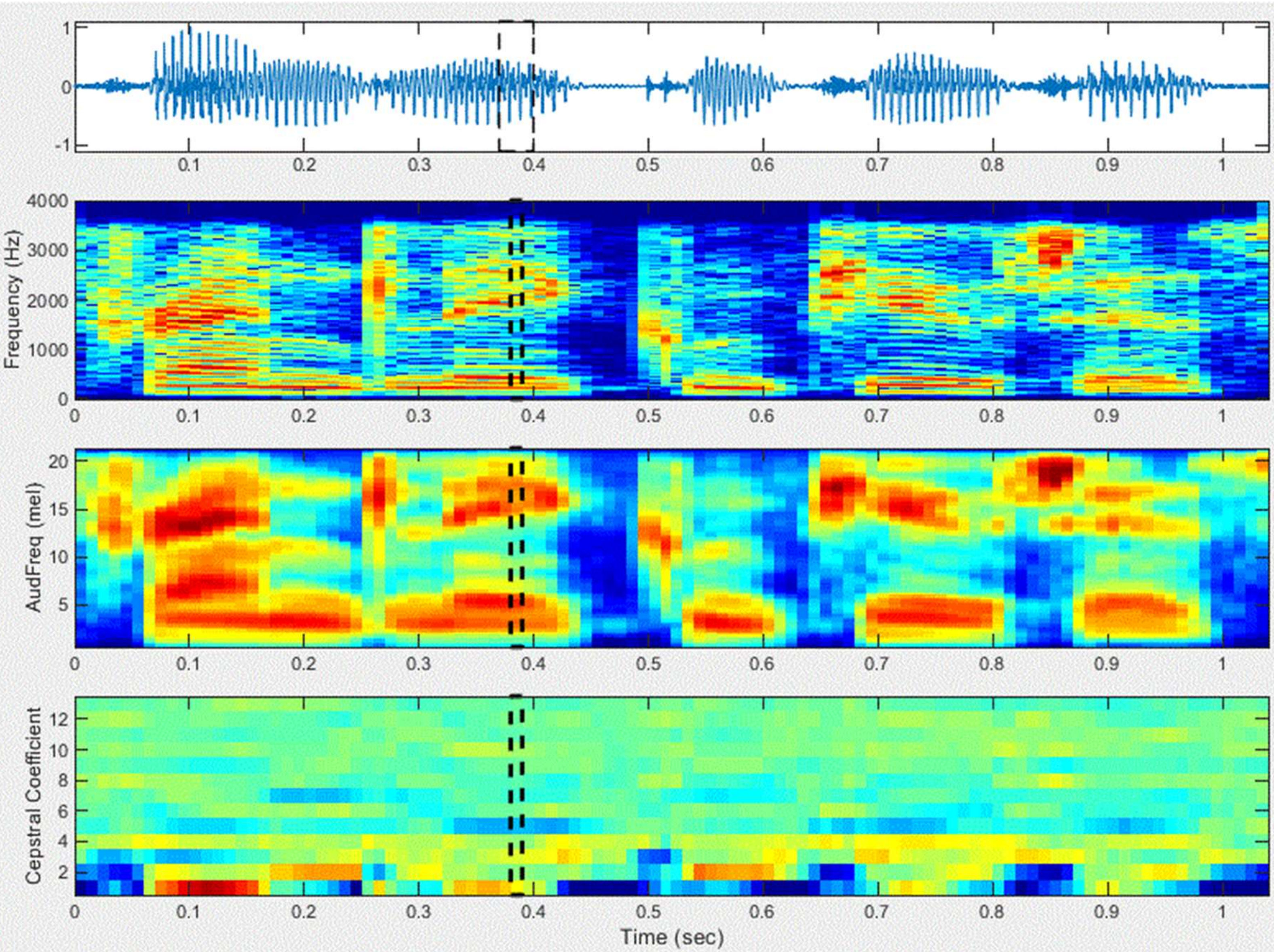
Frequency Sensitivity

We can discriminate 3-4 Hz between 15 Hz and 2 kHz

… and above 2 kHz, a 0.3% change

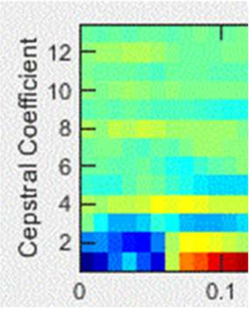
”mel” scale

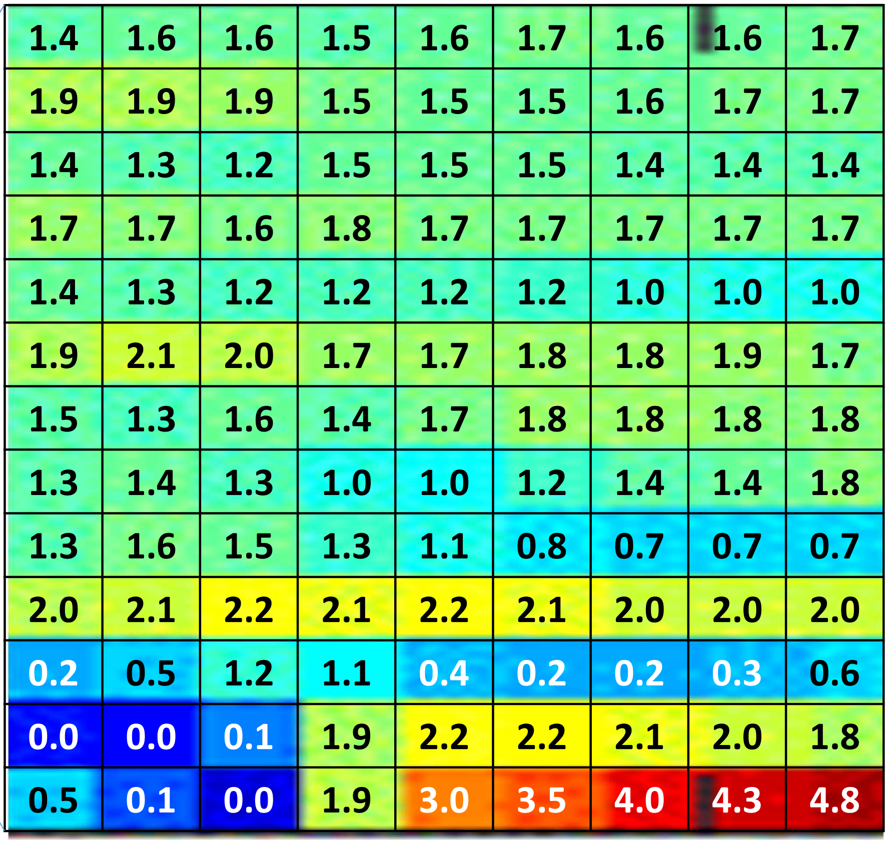




1
0
-1
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Frequency (Hz)
4000
3000
2000
1000
0
AudFreq (mel)
20
15
10
5
Cepstral Coefficient
12
10
8
6
4
2
Time (sec)

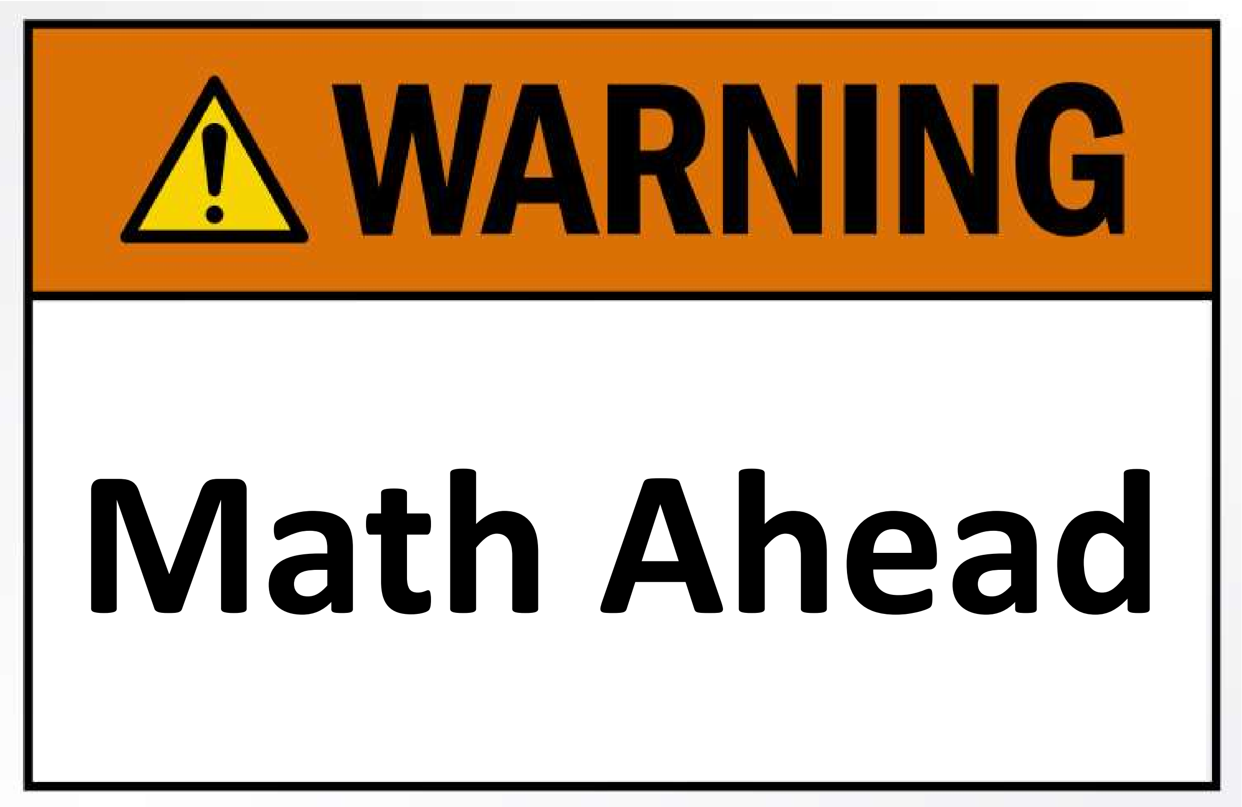
Feature Vectors





1.4	1.6	1.6	1.5	1.6	1.7	1.6	1.6	1.7
1.9	1.9	1.9	1.5	1.5	1.5	1.6	1.7	1.7
1.4	1.3	1.2	1.5	1.5	1.5	1.4	1.4	1.4
1.7	1.7	1.6	1.8	1.7	1.7	1.7	1.7	1.7
1.4	1.3	1.2	1.2	1.2	1.2	1.0	1.0	1.0
1.9	2.1	2.0	1.7	1.7	1.8	1.8	1.9	1.7
1.5	1.3	1.6	1.4	1.7	1.8	1.8	1.8	1.8
1.3	1.4	1.3	1.0	1.0	1.2	1.4	1.4	1.8
1.3	1.6	1.5	1.3	1.1	0.8	0.7	0.7	0.7
2.0	2.1	2.2	2.1	2.2	2.1	2.0	2.0	2.0
0.2	0.5	1.2	1.1	0.4	0.2	0.2	0.3	0.6
0.0	0.0	0.1	1.9	2.2	2.2	2.1	2.0	1.8
0.5	0.1	0.0	1.9	3.0	3.5	4.0	4.3	4.8

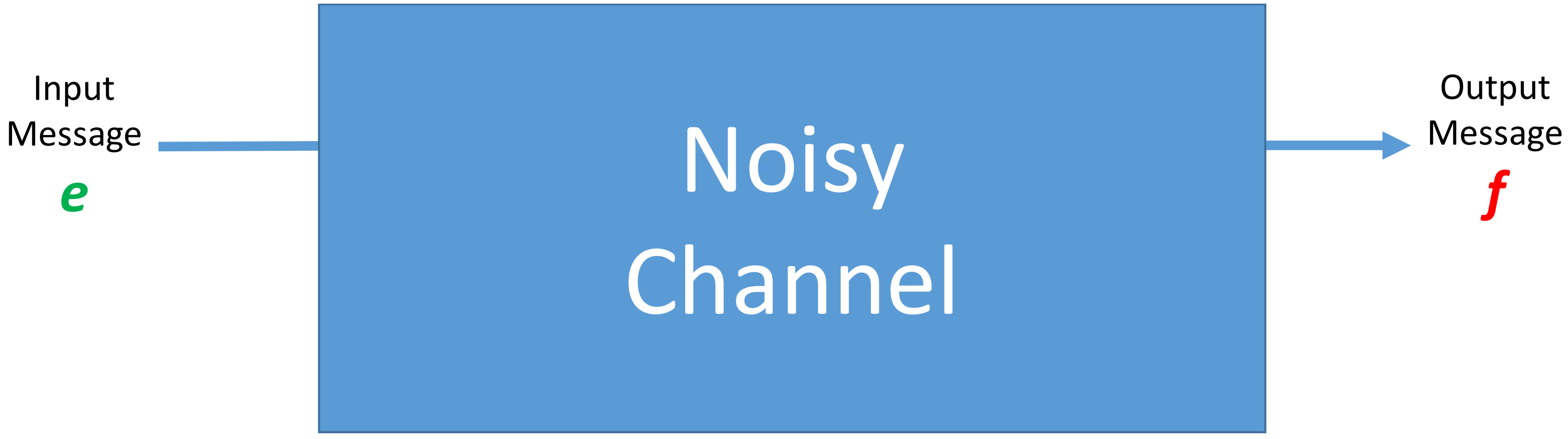
…


WARNING
Math Ahead

Noisy Channel Model

Find the most probable Input Message *e* that led to the Output Message *f*

Search each possible e to find the biggest $p(e|f)$



Input
Message
e

Noisy
Channel

Output
Message
f

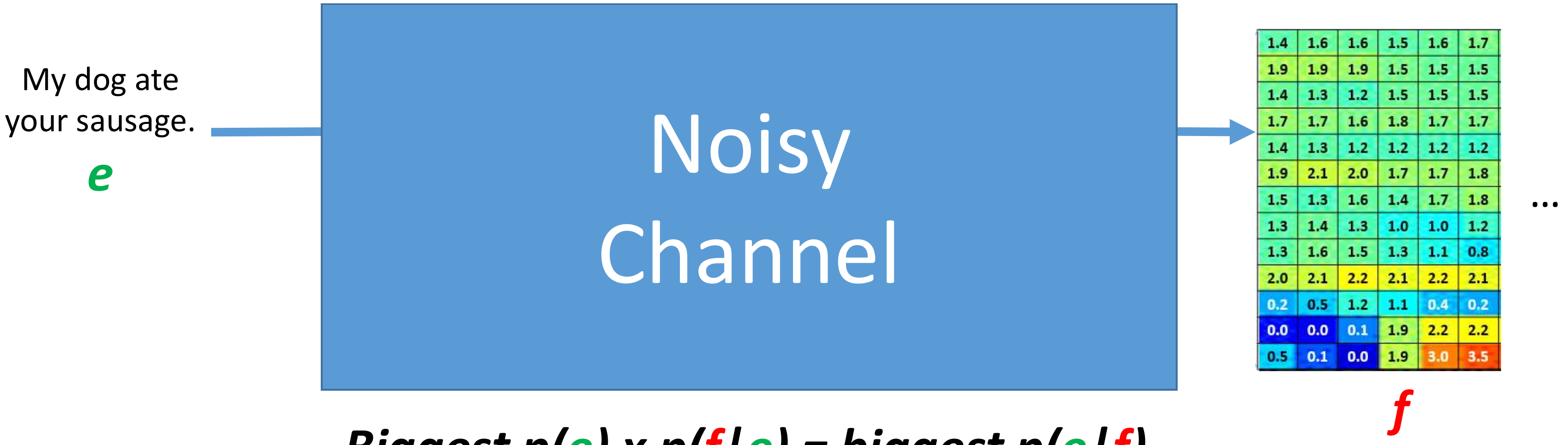
Biggest $p(e) \times p(f|e)$ = biggest $p(e|f)$

Claude Shannon – 1948 Bell Labs Technical Journal

Noisy Channel Model

Find the most probable phrase *e* that led to the Feature Vectors *f*

Search each possible e to find the biggest p(e|f)



My dog ate
your sausage.

e

Noisy
Channel

1.4	1.6	1.6	1.5	1.6	1.7
1.9	1.9	1.9	1.5	1.5	1.5
1.4	1.3	1.2	1.5	1.5	1.5
1.7	1.7	1.6	1.8	1.7	1.7
1.4	1.3	1.2	1.2	1.2	1.2
1.9	2.1	2.0	1.7	1.7	1.8
1.5	1.3	1.6	1.4	1.7	1.8
1.3	1.4	1.3	1.0	1.0	1.2
1.3	1.6	1.5	1.3	1.1	0.8
2.0	2.1	2.2	2.1	2.2	2.1
0.2	0.5	1.2	1.1	0.4	0.2
0.0	0.0	0.1	1.9	2.2	2.2
0.5	0.1	0.0	1.9	3.0	3.5

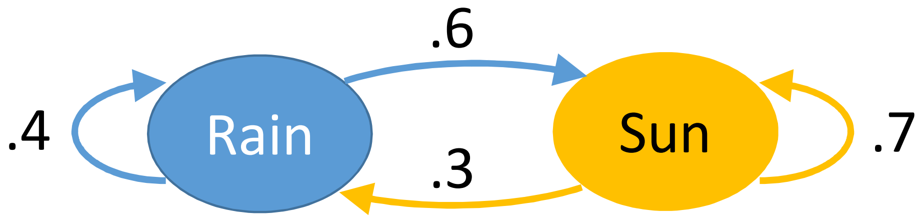
…

f

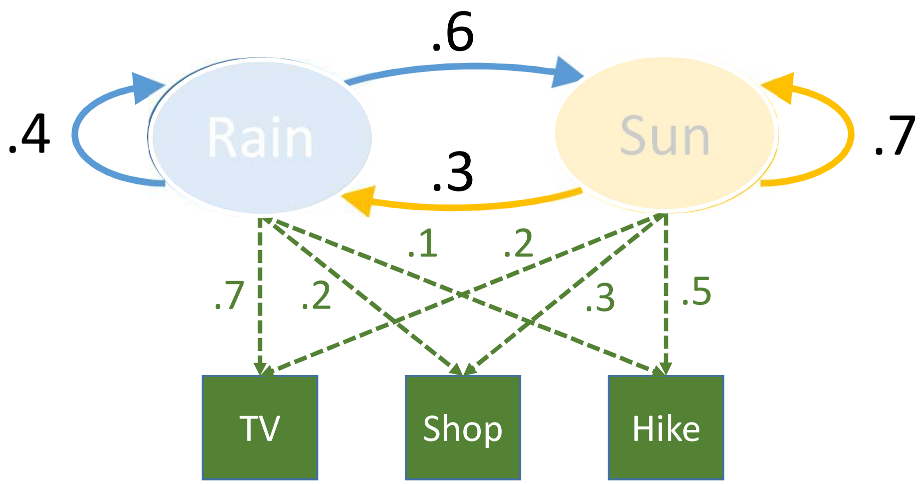
Biggest p(e) x p(f|e) = biggest p(e|f)

Claude Shannon – 1948 Bell Labs Technical Journal

Markov Model



Hidden Markov Model (HMM)



Observed events

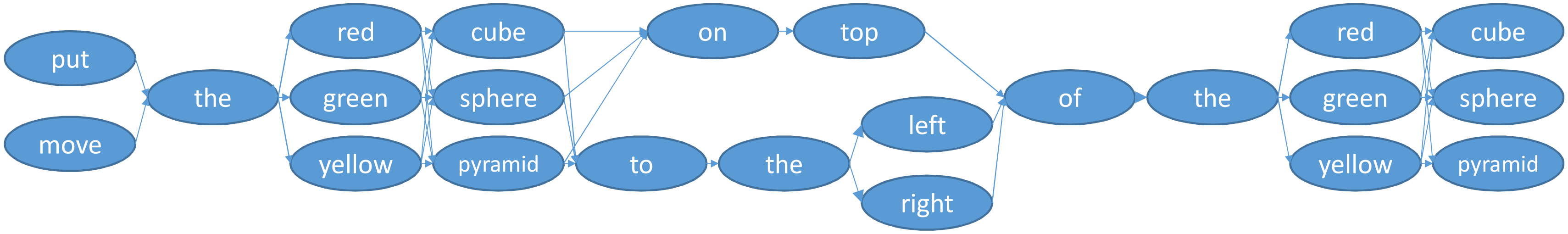
T S S H H H T H S S

VITERBI SEARCH

R R S S S S R R S S

Most likely cause

Hidden Markov Models (HMM)



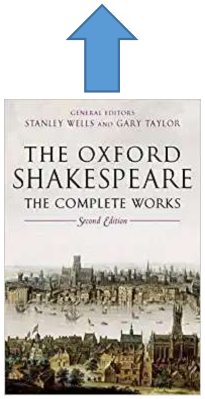
1.4	1.6	1.6	1.5	1.6	1.7	1.6	1.6	1.7	1.4	1.6	1.6	1.5	1.6	1.7	1.6	1.6	1.7	1.4	1.6	1.6	1.5	1.6	1.7	1.6	1.6	1.7	1.4	1.6	1.6	1.5	1.6	1.7	1.6	1.6	1.7
1.9	1.9	1.9	1.5	1.5	1.5	1.6	1.7	1.7	1.9	1.9	1.9	1.5	1.5	1.5	1.6	1.7	1.7	1.9	1.9	1.9	1.5	1.5	1.5	1.6	1.7	1.7	1.9	1.9	1.9	1.5	1.5	1.5	1.6	1.7	1.7
1.4	1.3	1.2	1.5	1.5	1.5	1.4	1.4	1.4	1.4	1.3	1.2	1.5	1.5	1.5	1.4	1.4	1.4	1.4	1.3	1.2	1.5	1.5	1.5	1.4	1.4	1.4	1.4	1.3	1.2	1.5	1.5	1.5	1.4	1.4	1.4
1.7	1.7	1.6	1.8	1.7	1.7	1.7	1.7	1.7	1.7	1.7	1.6	1.8	1.7	1.7	1.7	1.7	1.7	1.7	1.7	1.6	1.8	1.7	1.7	1.7	1.7	1.7	1.7	1.7	1.6	1.8	1.7	1.7	1.7	1.7	1.7
1.4	1.3	1.2	1.2	1.2	1.2	1.0	1.0	1.0	1.4	1.3	1.2	1.2	1.2	1.2	1.0	1.0	1.0	1.4	1.3	1.2	1.2	1.2	1.2	1.0	1.0	1.0	1.4	1.3	1.2	1.2	1.2	1.2	1.0	1.0	1.0
1.9	2.1	2.0	1.7	1.7	1.8	1.8	1.9	1.7	1.9	2.1	2.0	1.7	1.7	1.8	1.8	1.9	1.7	1.9	2.1	2.0	1.7	1.7	1.8	1.8	1.9	1.7	1.9	2.1	2.0	1.7	1.7	1.8	1.8	1.9	1.7
1.5	1.3	1.6	1.4	1.7	1.8	1.8	1.8	1.8	1.5	1.3	1.6	1.4	1.7	1.8	1.8	1.8	1.8	1.5	1.3	1.6	1.4	1.7	1.8	1.8	1.8	1.8	1.5	1.3	1.6	1.4	1.7	1.8	1.8	1.8	1.8
1.3	1.4	1.3	1.0	1.0	1.2	1.4	1.4	1.8	1.3	1.4	1.3	1.0	1.0	1.2	1.4	1.4	1.8	1.3	1.4	1.3	1.0	1.0	1.2	1.4	1.4	1.8	1.3	1.4	1.3	1.0	1.0	1.2	1.4	1.4	1.8
1.3	1.6	1.5	1.3	1.1	0.8	0.7	0.7	0.7	1.3	1.6	1.5	1.3	1.1	0.8	0.7	0.7	0.7	1.3	1.6	1.5	1.3	1.1	0.8	0.7	0.7	0.7	1.3	1.6	1.5	1.3	1.1	0.8	0.7	0.7	0.7
2.0	2.1	2.2	2.1	2.2	2.1	2.0	2.0	2.0	2.0	2.1	2.2	2.1	2.2	2.1	2.0	2.0	2.0	2.0	2.1	2.2	2.1	2.2	2.1	2.0	2.0	2.0	2.0	2.1	2.2	2.1	2.2	2.1	2.0	2.0	2.0
0.2	0.5	1.2	1.1	0.4	0.2	0.2	0.3	0.6	0.2	0.5	1.2	1.1	0.4	0.2	0.2	0.3	0.6	0.2	0.5	1.2	1.1	0.4	0.2	0.2	0.3	0.6	0.2	0.5	1.2	1.1	0.4	0.2	0.2	0.3	0.6
0.0	0.0	0.1	1.9	2.2	2.2	2.1	2.0	1.8	0.0	0.0	0.1	1.9	2.2	2.2	2.1	2.0	1.8	0.0	0.0	0.1	1.9	2.2	2.2	2.1	2.0	1.8	0.0	0.0	0.1	1.9	2.2	2.2	2.1	2.0	1.8
0.5	0.1	0.0	1.9	3.0	3.5	4.0	4.3	4.8	0.5	0.1	0.0	1.9	3.0	3.5	4.0	4.3	4.8	0.5	0.1	0.0	1.9	3.0	3.5	4.0	4.3	4.8	0.5	0.1	0.0	1.9	3.0	3.5	4.0	4.3	4.8

VITERBI SEARCH

“MOVE THE RED SPHERE ON TOP OF THE GREEN PYRAMID”

OK, but how do you get the probabilities???

If an infinite number of monkeys…



GENERAL EDITORS
STANLEY WELLS AND GARY TAYLOR
THE OXFORD SHAKESPEARE
THE COMPLETE WORKS
Second Edition

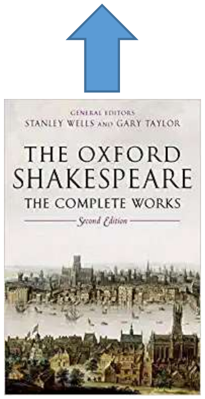
As that may speech those fallible factor not name so garb and his eat by my kisses camp morn thou my the leave her.

Not with death gipsy to bloody of me he do great.

It the what Hamlet handkerchief then aught enemies bones come madness.

p(word)
1st order monkeys = 1-grams

If an infinite number of monkeys…



GENERAL EDITORS
STANLEY WELLS AND GARY TAYLOR
THE OXFORD
SHAKESPEARE
THE COMPLETE WORKS
Second Edition

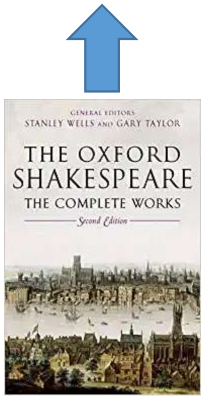
Nay then she did forfeit sovereign as loud music i' the heels.

Hast slaughter'd his passage.

With the slave and start at last she your dominions for thee Charmian lived a good cheer.

p(word | previous word)
2nd order monkeys = 2-grams

If an infinite number of monkeys…



What is the great cannon to the land withal yet to draw apart the body.

Rashly and praised be rashness for it.

That art not what counts harsh fortune casts upon my charm I have.

p(word | previous two word)
3rd order monkeys = 3-grams = trigrams

If an infinite number of monkeys…

His horns shall be girded with a lamb for an heave offering unto the Lord is in the land of Egypt.

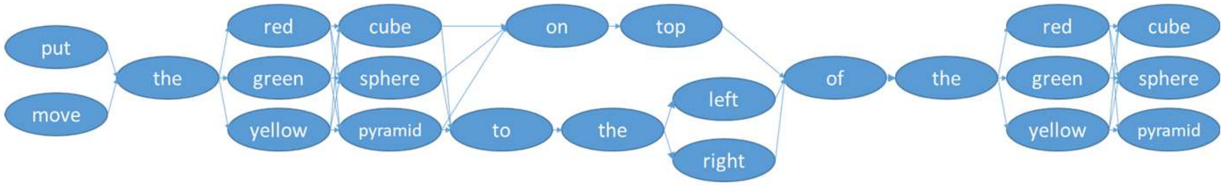
Deliver him into this wilderness to meet him and put on other garments and anoint the laver and his sons Esau and also behold he is become of him.

?

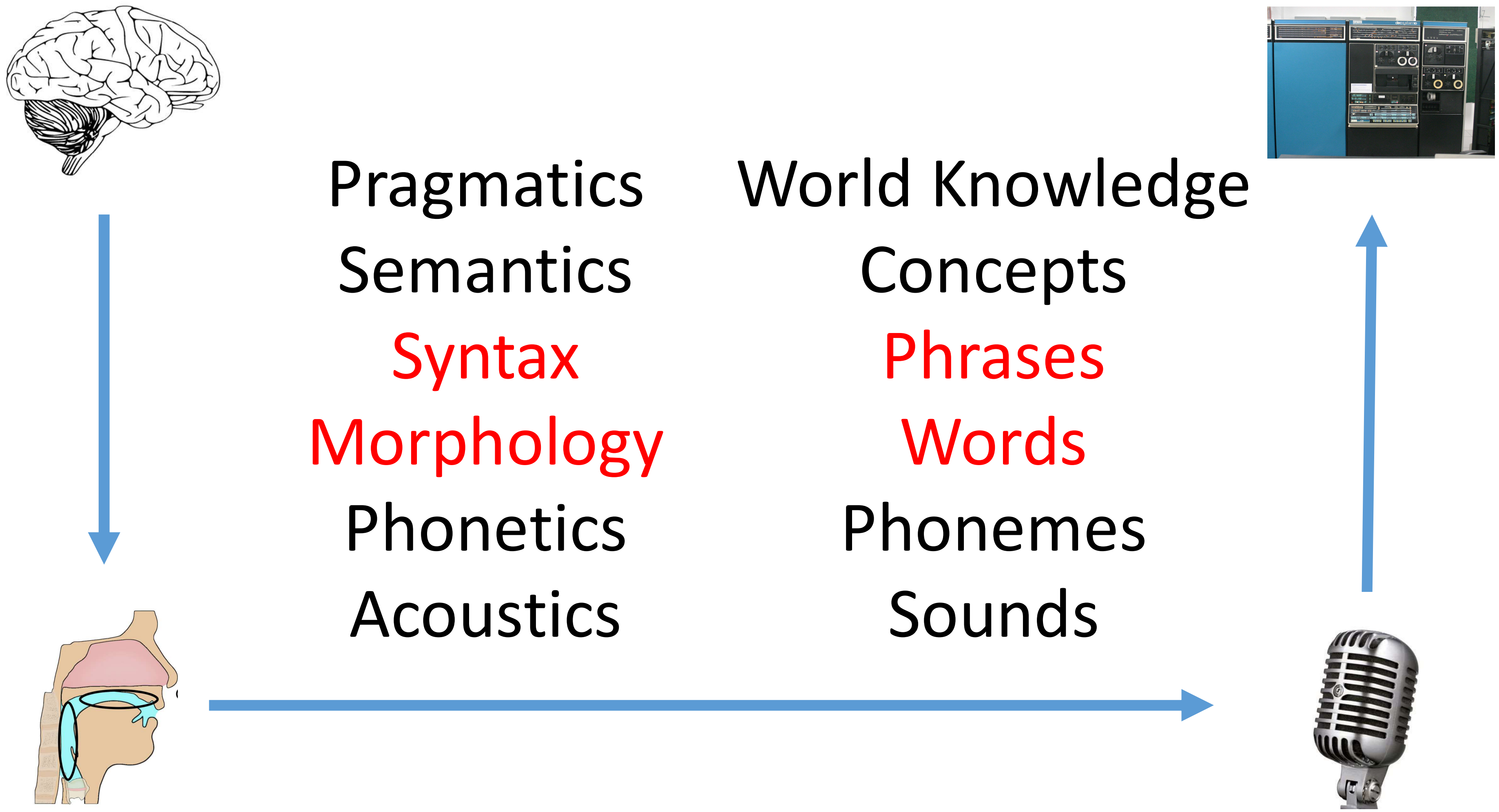
p(word | previous two word)
3rd order monkeys = 3-grams = trigrams

Lexical Network: words → phrases

1. Fixed grammar probabilities – many are 1.0



2. Third-order monkeys = "trigrams" constraint the network
3. Other creative hacks, aka engineering solutions



Pragmatics	World Knowledge
Semantics	Concepts
Syntax	Phrases
Morphology	Words
Phonetics	Phonemes
Acoustics	Sounds

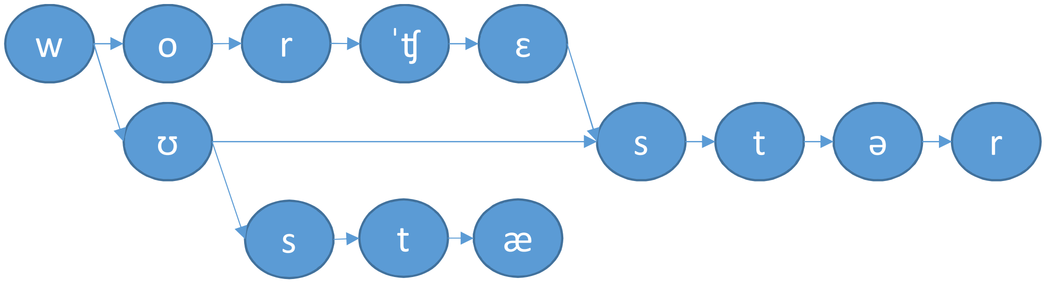
END
MATH

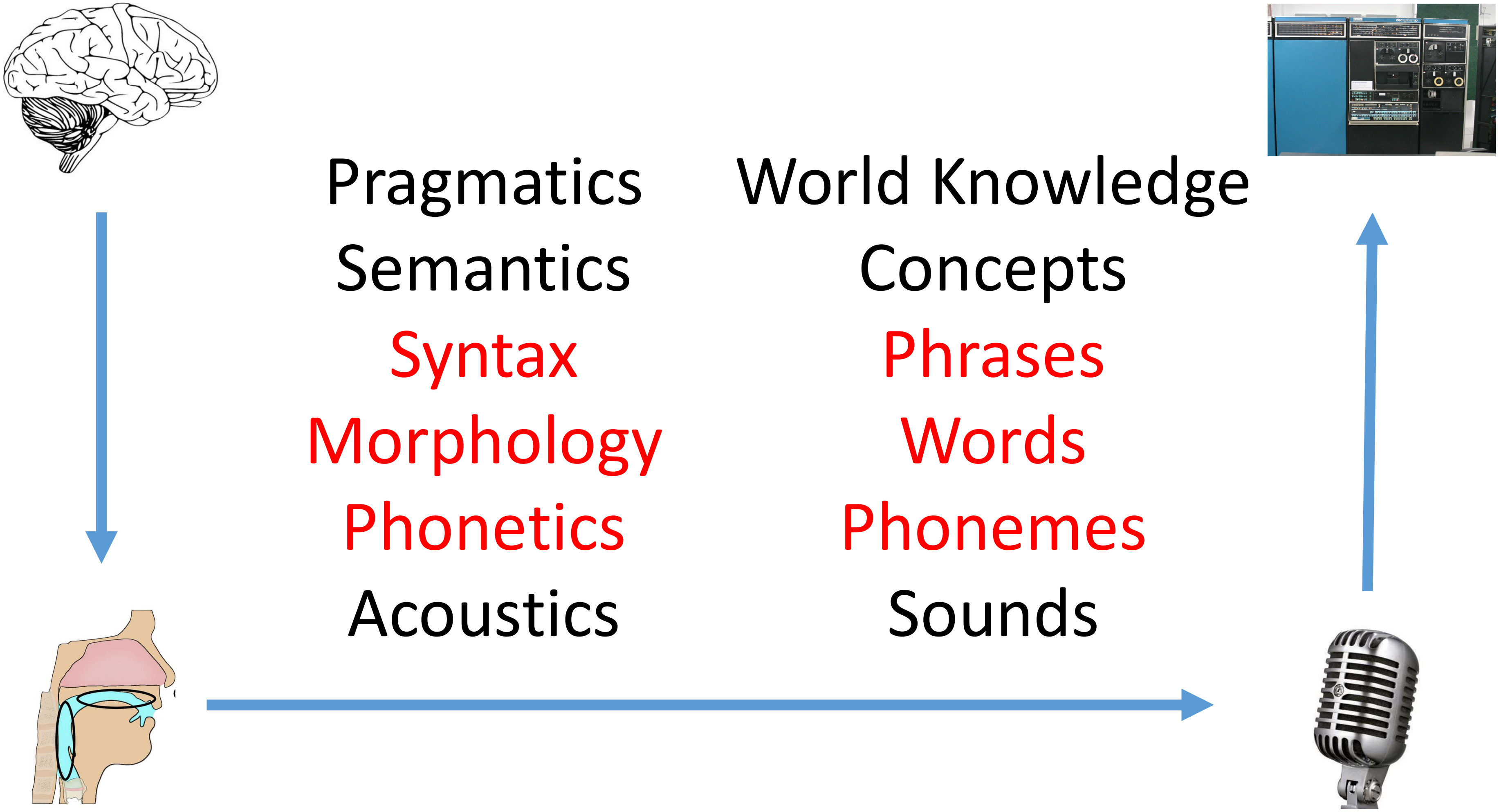
Phonetic Network: phonemes → words

Essentially a dictionary with alternate pronunciations

Worcester

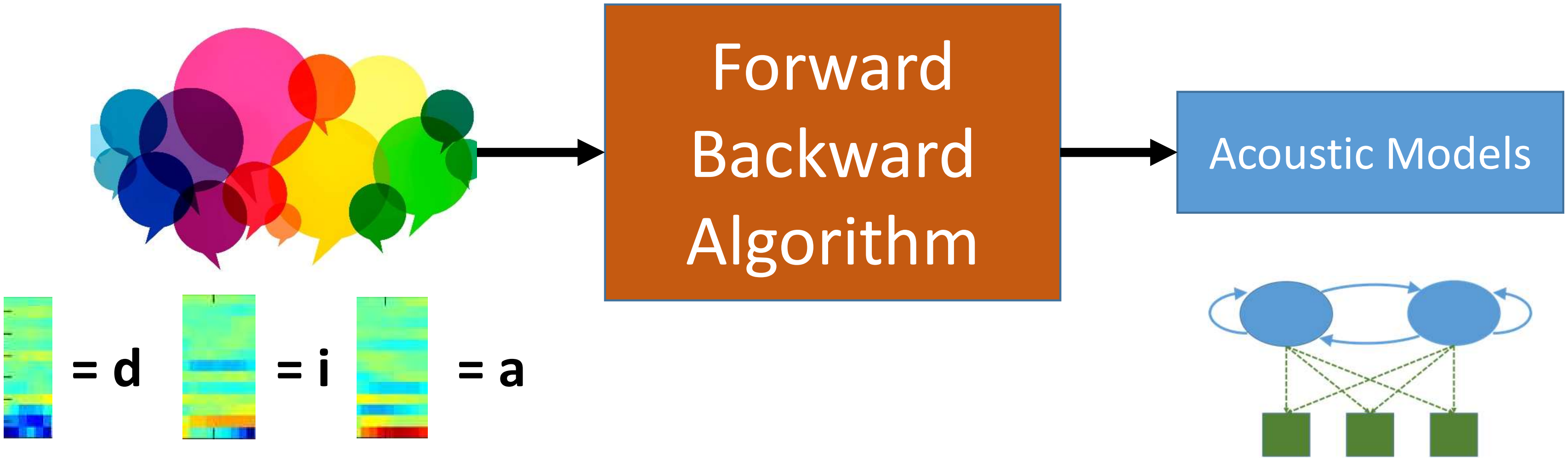
Worˈʧɛstər	Tourist
Wʊstər	Non-native Bostonian
Wʊstæ	True Bostonian





Pragmatics	World Knowledge
Semantics	Concepts
Syntax	Phrases
Morphology	Words
Phonetics	Phonemes
Acoustics	Sounds

Acoustic model: Feature vectors → phonemes

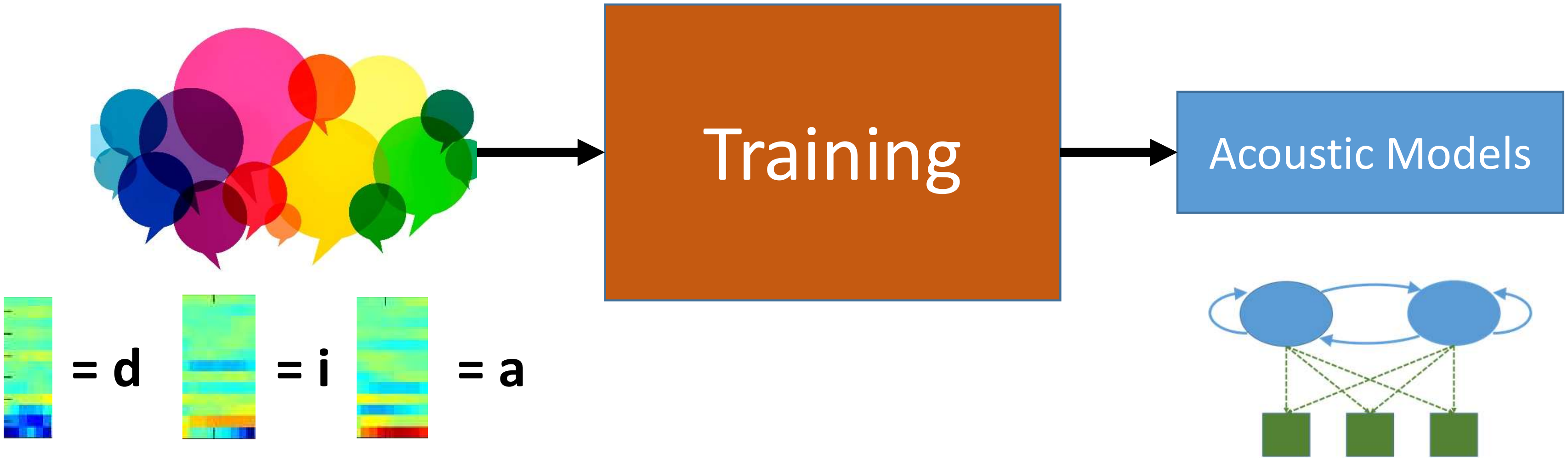


= d = i = a

Forward Backward Algorithm

Acoustic Models

Acoustic model: Feature vectors → phonemes



Training

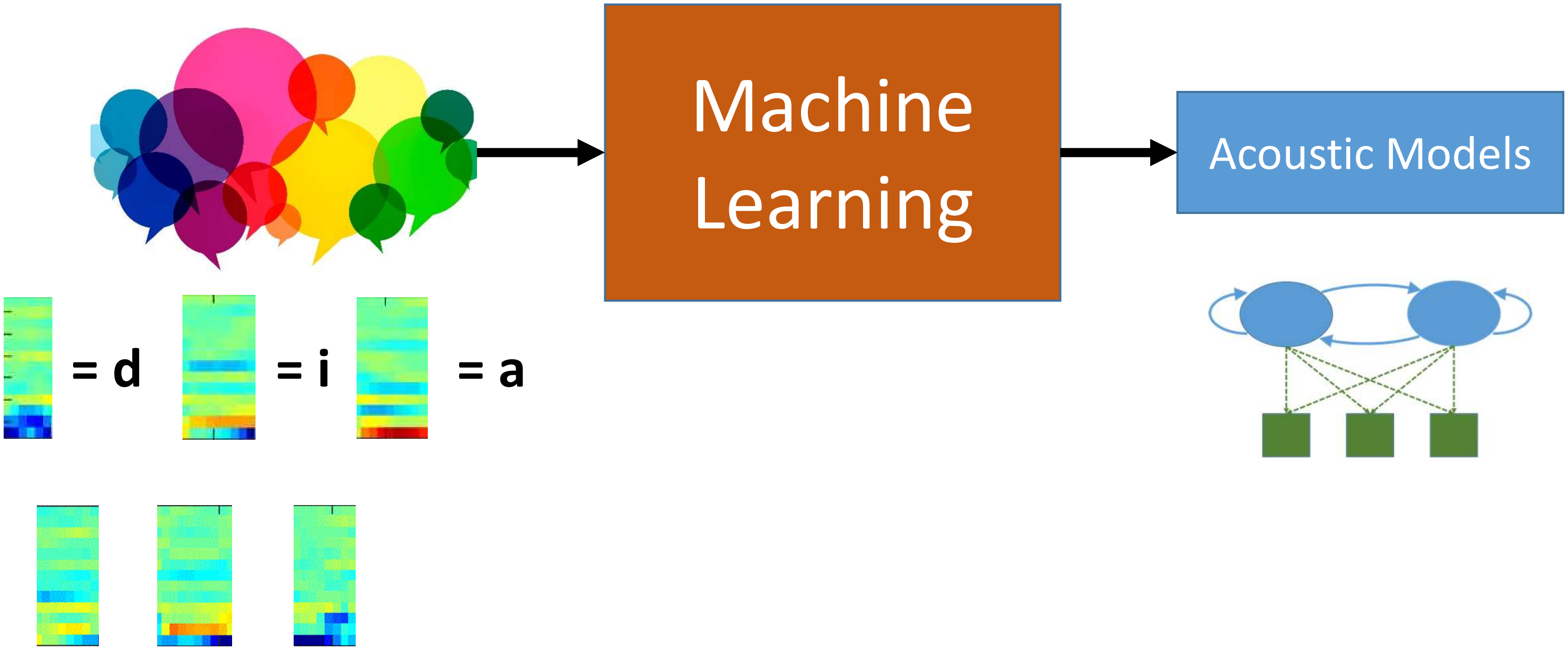
Acoustic Models

= d

= i

= a

Acoustic model: Feature vectors → phonemes



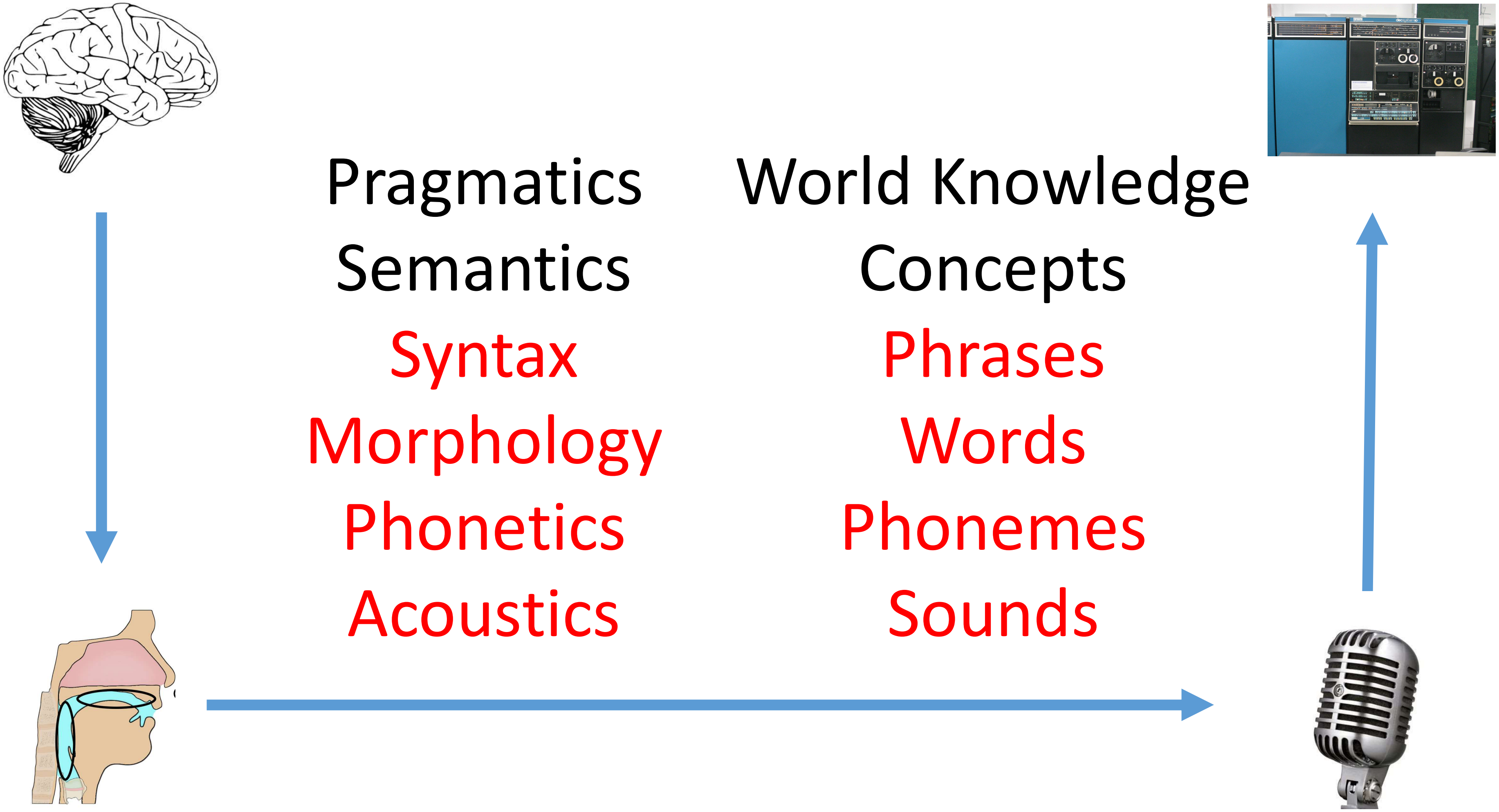
Machine Learning

Acoustic Models

= d

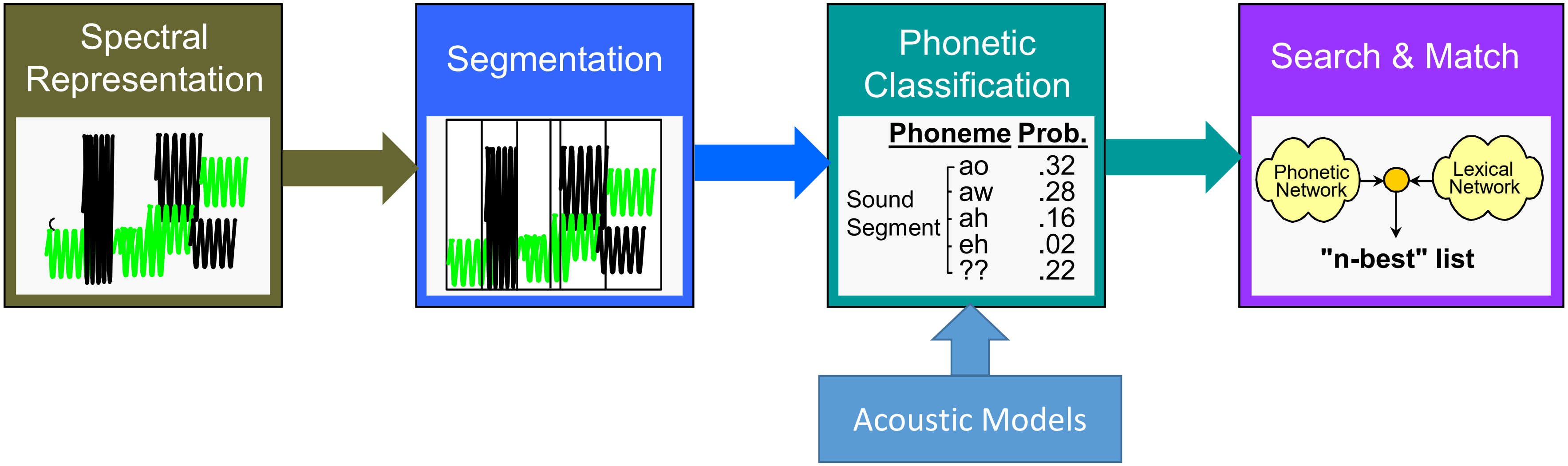
= i

= a



Pragmatics	World Knowledge
Semantics	Concepts
Syntax	Phrases
Morphology	Words
Phonetics	Phonemes
Acoustics	Sounds

A Recognition Pipeline



Spectral Representation

Segmentation

Phonetic Classification

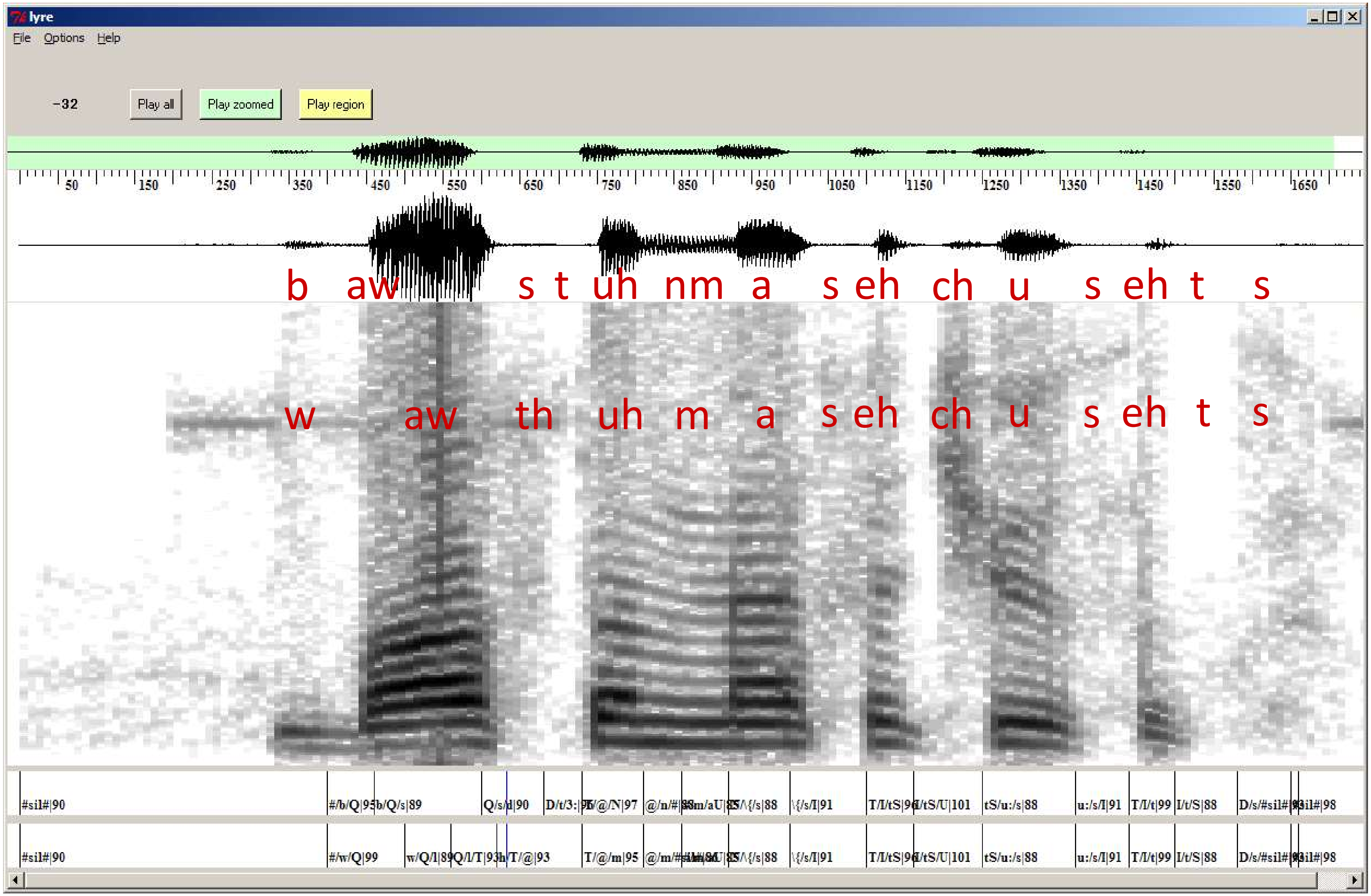
Sound Segment

Phoneme	Prob.
ao	.32
aw	.28
ah	.16
eh	.02
??	.22

Search & Match

"n-best" list

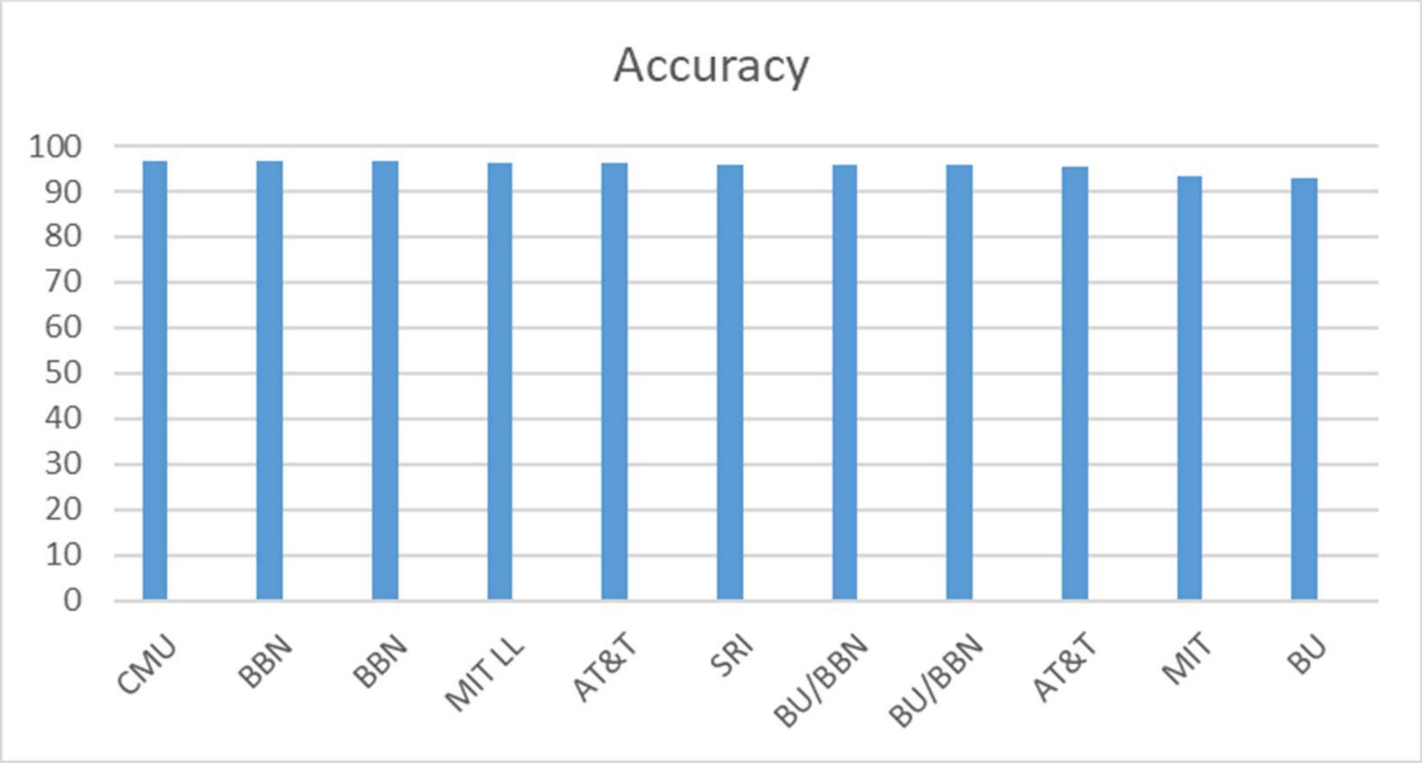
Acoustic Models


lyre
File
Options
Help
-32
Play all
Play zoomed
Play region
50
150
250
350
450
550
650
750
850
950
1050
1150
1250
1350
1450
1550
1650
b aw s t uh nm a s eh ch u s eh t s
w aw th uh m a s eh ch u s eh t s
#sil#|90
#/b/Q|95
b/Q/s|89
Q/s/d|90
D/t/3:|95
T/@/N|97
@/n/#|88
m/aU|85
/\{/s|88
\{/s/I|91
T/I/tS|96
I/tS/U|101
tS/u:/s|88
u:/s/I|91
T/I/t|99
I/t/S|88
D/s/#sil#|98
sil#|98
#sil#|90
#/w/Q|99
w/Q/l|89
Q/l/T|93
h/T/@|93
T/@/m|95
@/m/#
aU|85
/\{/s|88
\{/s/I|91
T/I/tS|96
I/tS/U|101
tS/u:/s|88
u:/s/I|91
T/I/t|99
I/t/S|88
D/s/#sil#|98
sil#|98

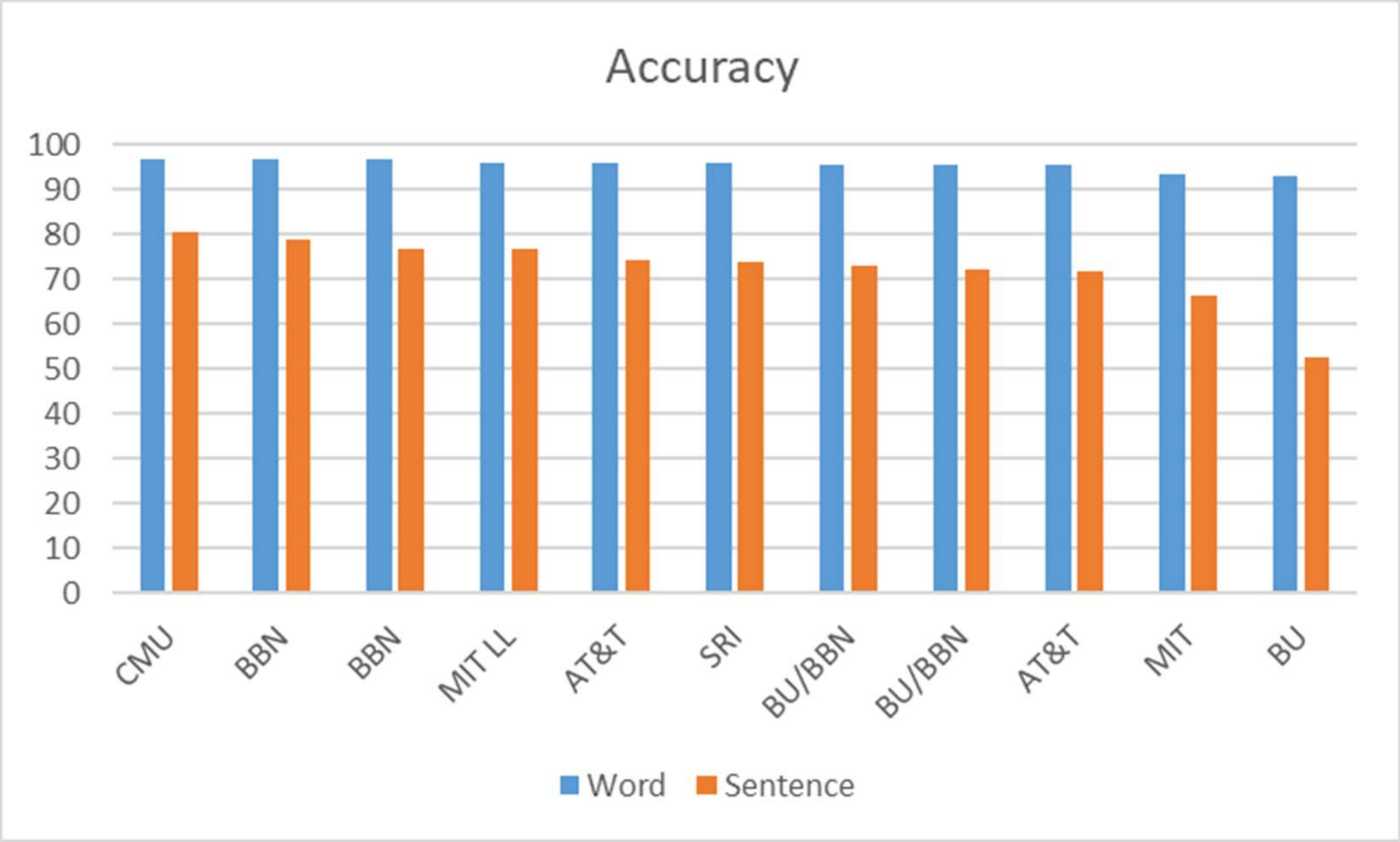
1985 “Resource Management” Task

- Funded CMU, MIT, BBN, SRI
- Standard corpus to train and test recognition
  - 25,000 utterances, 900 words, 990-sentence grammar, 160 varied speakers
    - *Is Apalachicola's radar sensor location data newer than sonar data?*
    - *Show the Fresno's track without overlay.*
    - *Give me a list of the names and estimated time of arrival at their destinations for carriers in the Philippine Sea.*
- 6 rounds of evaluation March 1987 – June 1990
  - Test sets distributed with results due back in a few days
  - No evaluation of speed

1990 “Resource Management” Task

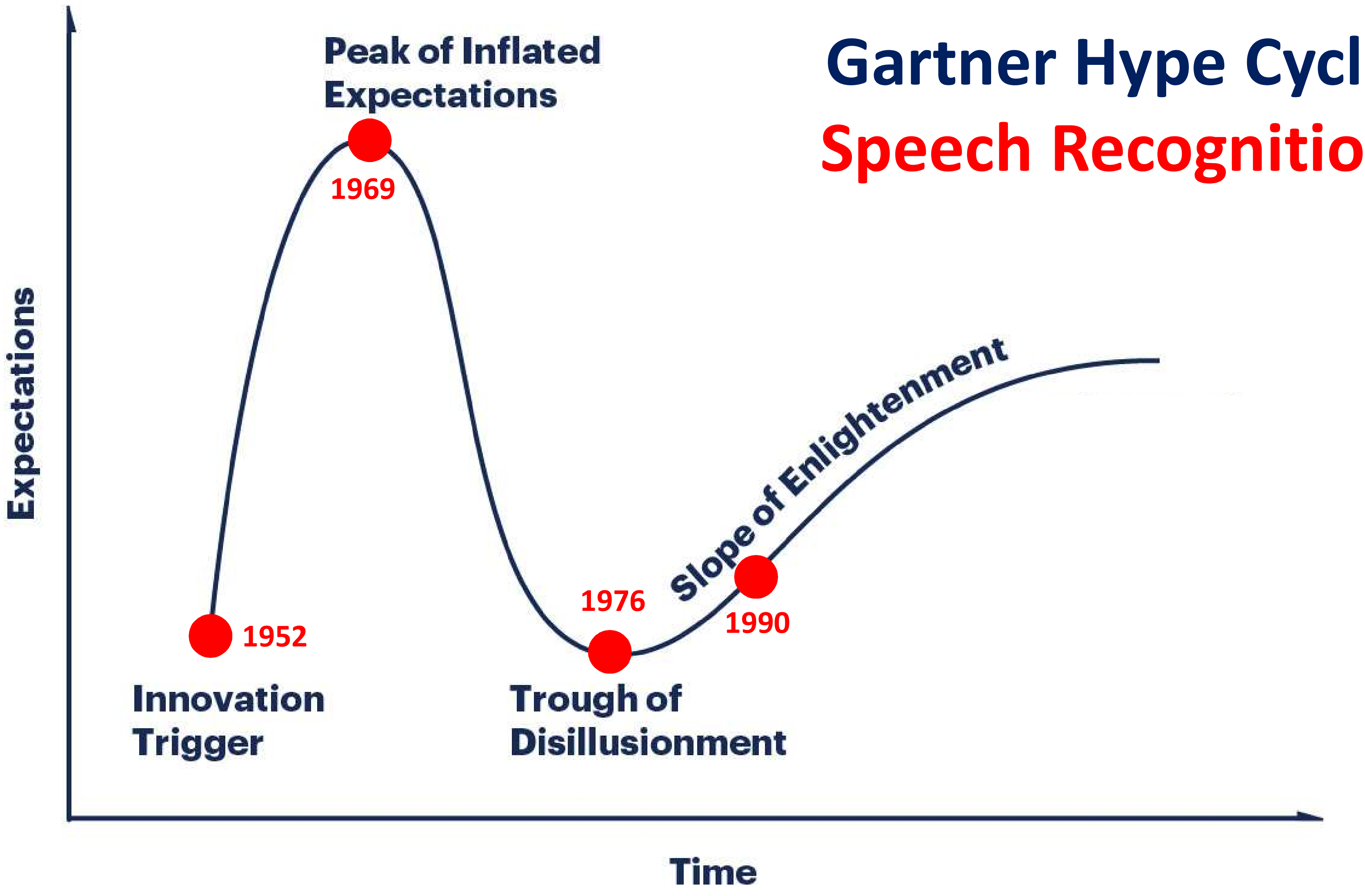


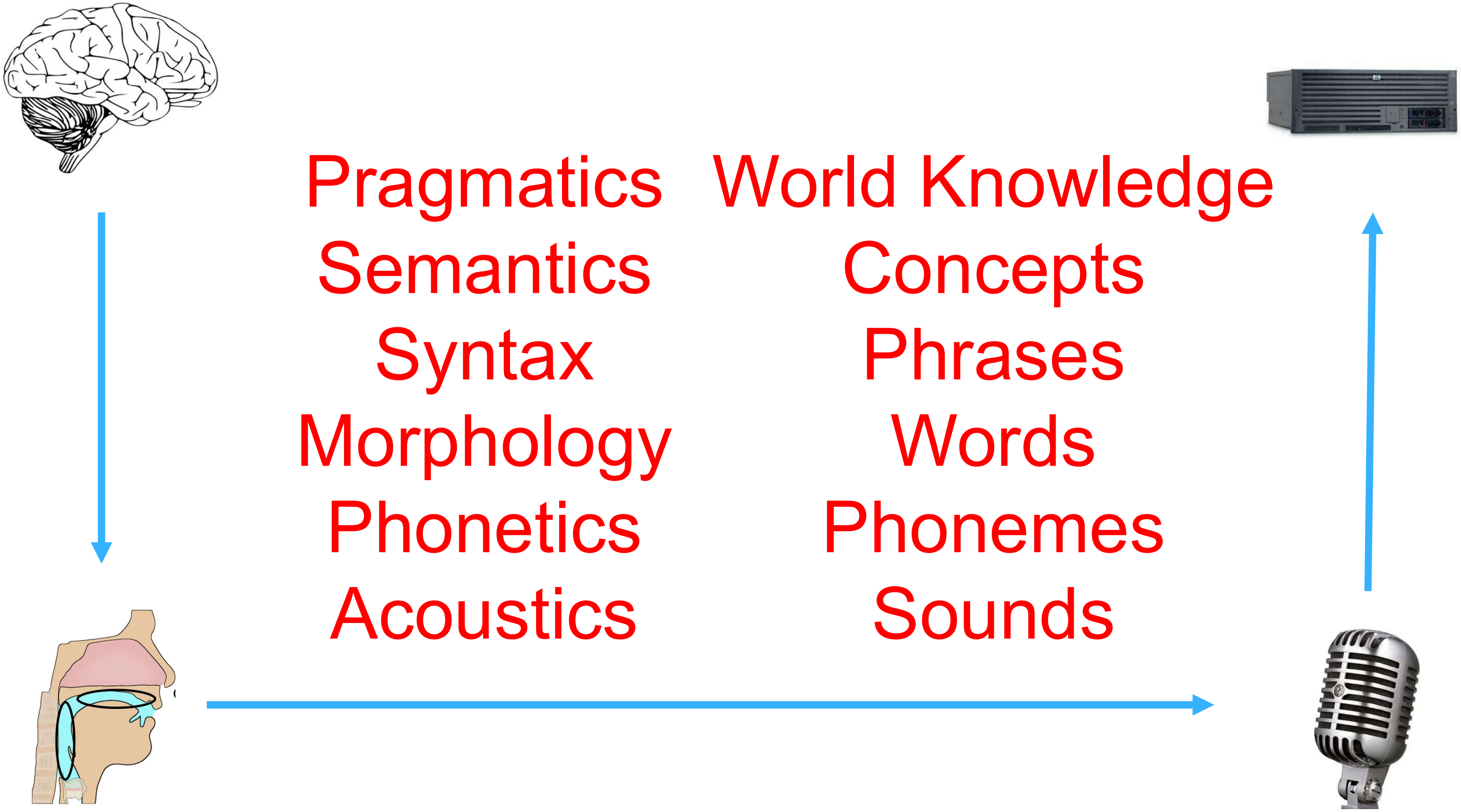
1990 “Resource Management” Task



Gartner Hype Cycle

Speech Recognition

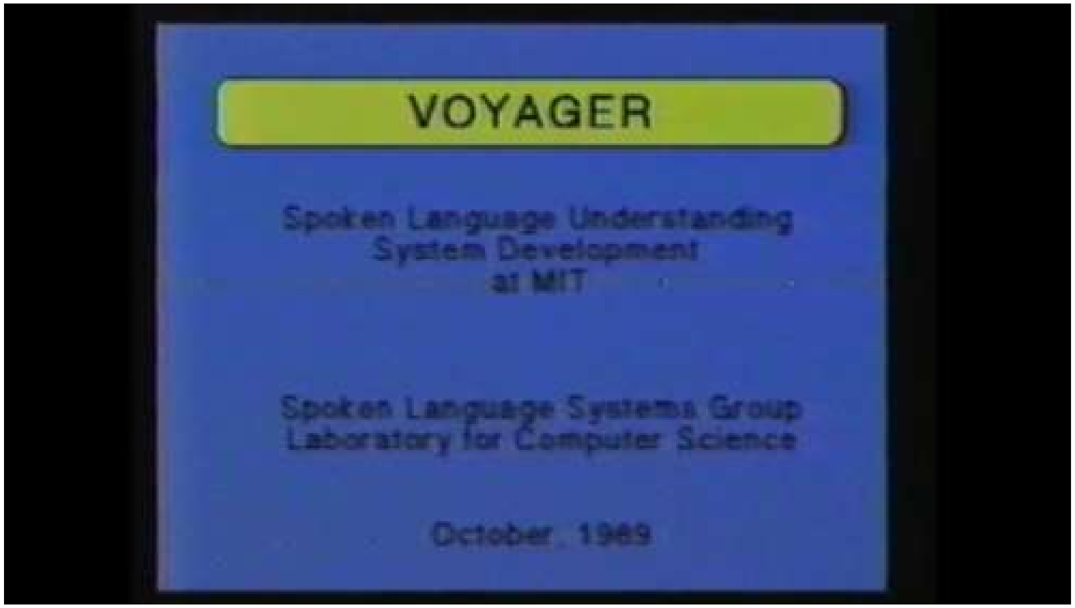




Pragmatics	World Knowledge
Semantics	Concepts
Syntax	Phrases
Morphology	Words
Phonetics	Phonemes
Acoustics	Sounds

1989 Voyager System

A prototype, not a product



VOYAGER

Spoken Language Understanding
System Development
at MIT

Spoken Language Systems Group
Laboratory for Computer Science

October 1989

https://www.youtube.com/watch?v=zS3baF8KHSE

1989 ATIS Air Travel Info System

Understand spontaneous speech

Tricky to make this an objective competition!

Plan a business trip to 4 different cities (of your choice), using public ground transportation to and from the airports. Save time and money where you can. The client is an airplane buff and enjoys flying on different kinds of aircraft.

Subject makes requests, then sees their request and data results on screen

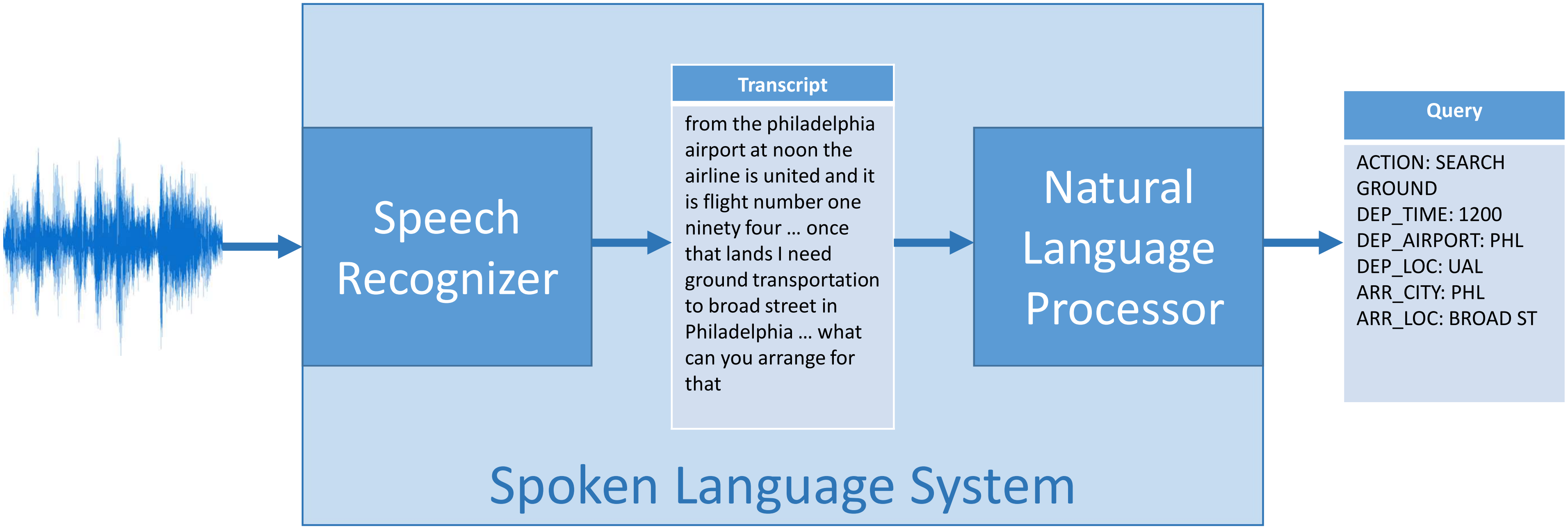
Trained Wizards of Oz in other room hear request

Types it as understood

Enters an SQL-like query to generate results

1989 ATIS Air Travel Info System

Understand spontaneous speech



Speech Recognizer

Transcript

from the philadelphia airport at noon the airline is united and it is flight number one ninety four … once that lands I need ground transportation to broad street in Philadelphia … what can you arrange for that

Natural Language Processor

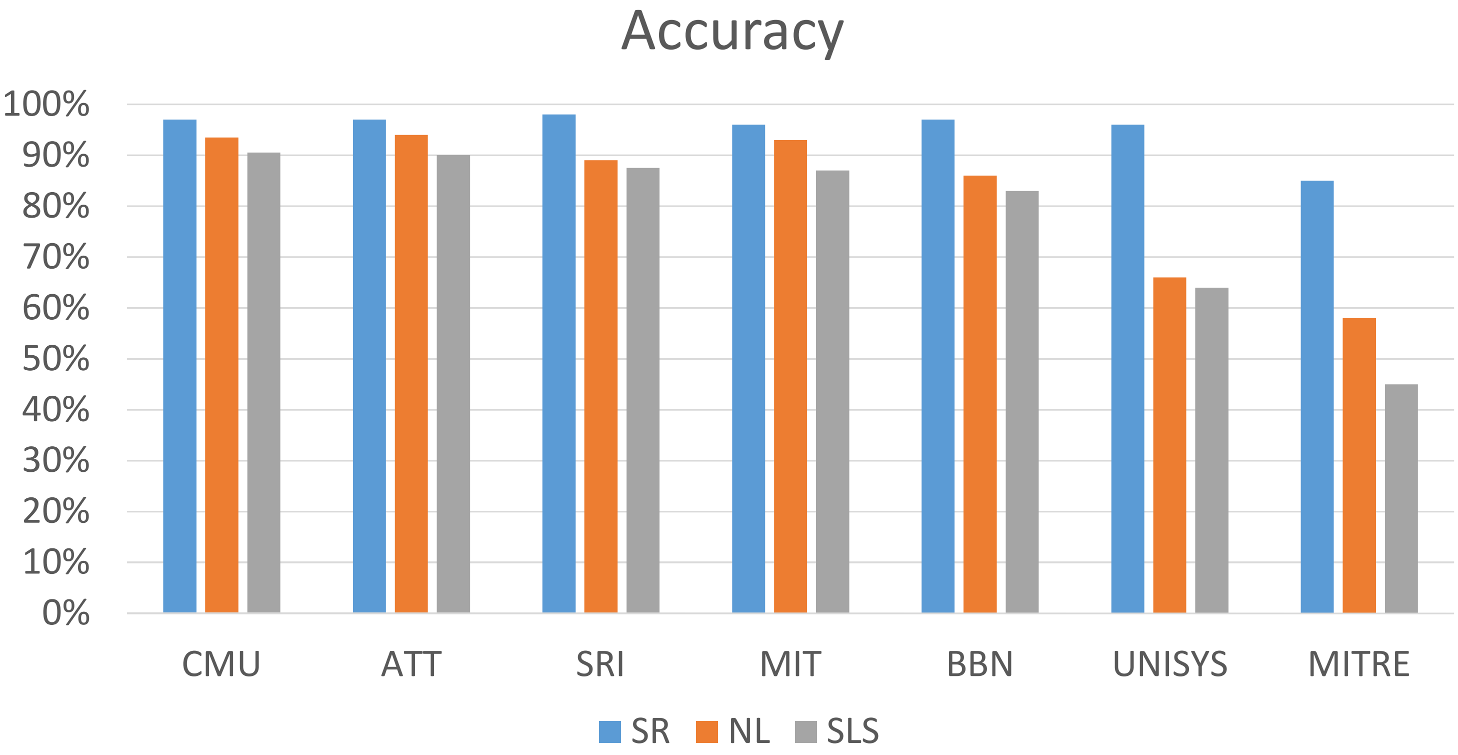
Query

ACTION: SEARCH GROUND
DEP_TIME: 1200
DEP_AIRPORT: PHL
DEP_LOC: UAL
ARR_CITY: PHL
ARR_LOC: BROAD ST

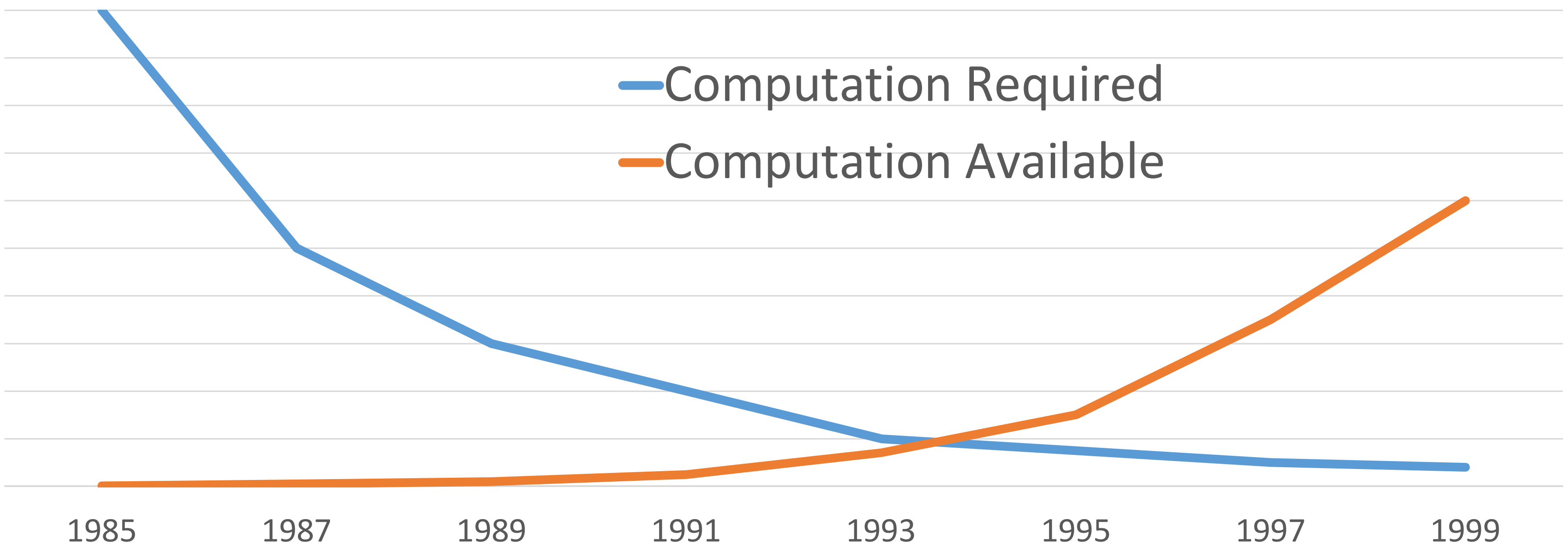
Spoken Language System

1995 ATIS Air Travel Info System

Understand spontaneous speech



Real-time Connected Speech Recognition



“Time to start a company…”

1987

↓

2000

IBM → Dragon
Lernout & Hauspie (L&H)
Voice Control Systems (VCS)
Scott Instruments
Voice Processing Corporation (VPC)
CMU → PureSpeech
ETI
MIT → SpeechWorks
SRI → Nuance
ART
Locus Dialog
Voice Signal
Phonetic
Rhetorical

Speaker-Dependent Dictation

1997 Dragon “Naturally Speaking”

23K words, 100 connected words / minute

mouse / keyboard correction

$695, and 45 minutes to “train”

DRAGON
NATURALLYSPEAKING

IBM

Microsoft

Connected Speech Phrase Recognition

Not quite Killer App: Call Center automation

Using the touchtone keys on your phone, enter the…

… stock ticker symbol

… first few letters of the last name of the person you're calling

… arrival city airport code or the first few letters of the city name

… first few letters of the name of the film

Say the…

… stock ticker symbol or company name

… name of the person you're calling

… arrival city

… name of the film

Connected Speech Phrase Recognition

Not quite Killer App: Call Center automation

WACHOVIA

ups

airTran

Amtrak

FedEx

CharlesSchwab

Fidelity INVESTMENTS

E*TRADE

AT&T

It’s not just Speech Recognition…

Who would you like to speak with?

<doesn’t respond>

I didn’t hear you. Who would you like to speak with?

Uh, um -- I’d like to talk to Mark if he’s around.

Sorry, I didn’t understand. Please say the name of the person you would like to speak with.

Mark.

We have more than one of those. Which one do you want?

Mark Holthouse.

Mark Holthouse, correct?

Yes.

Okay. Please hold while I transfer…

Using Previous Information

Where would you like to fly?

Boston.

Was that “Austin?”

No.

My mistake. Please say it again.

Boston.

Was that “Austin?”

<various obscenities>

The Art of User Interface (Dialog) Design

Welcome to United Airlines' flight information system. I'll be able to help you get information on all United, United Express, and United Shuttle flights. Enter or say the flight number, or say "I don't know it" if you're unsure.

> Uh, I don't know it?

OK, we'll find it a different way. Oh, here's a hint: If you ever know the answer to a question I'm asking, you can always interrupt me, and if you get stuck, say "Help". Here goes: Would you like arrival or departure information?

> Departure

OK, and from which city is the flight departing?

> Boston.

…

The Art of User Interface (Dialog) Design

Welcome to United Airlines' flight information system. I'll be able to…

    <barges in> Flight 455

Would like arrival or departure information?

    Departure

OK, I'll look that up. Hold on. <database lookup> Flight 455 is scheduled to depart on time at 8:45 A.M. from Boston Logan, Terminal C, Gate 14. You can say …

    <hangup>

Successful Applications

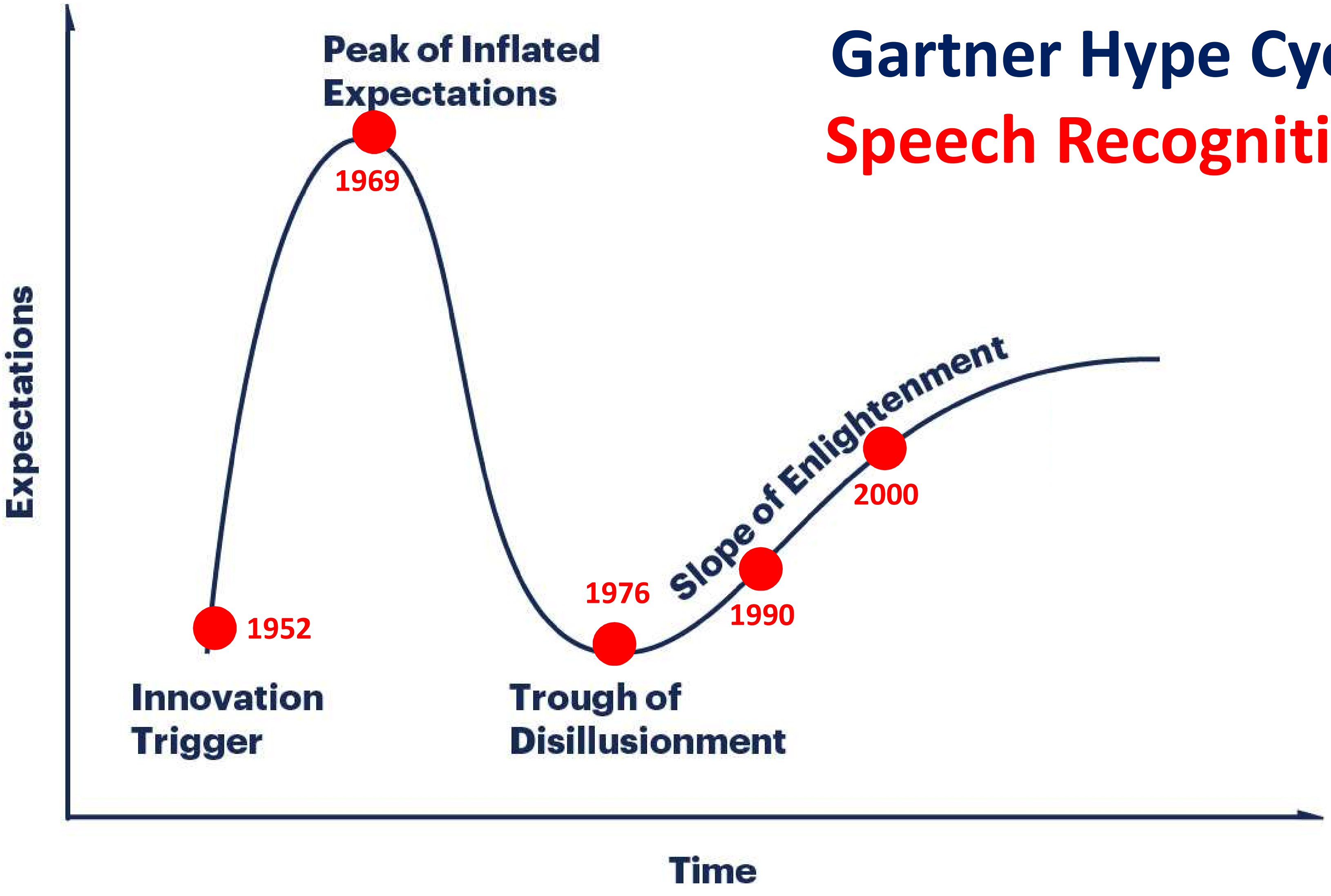
1. Complete the task successfully
2. Better than waiting on hold
3. Don’t irritate the caller

“I wanted to take the voice to dinner.”

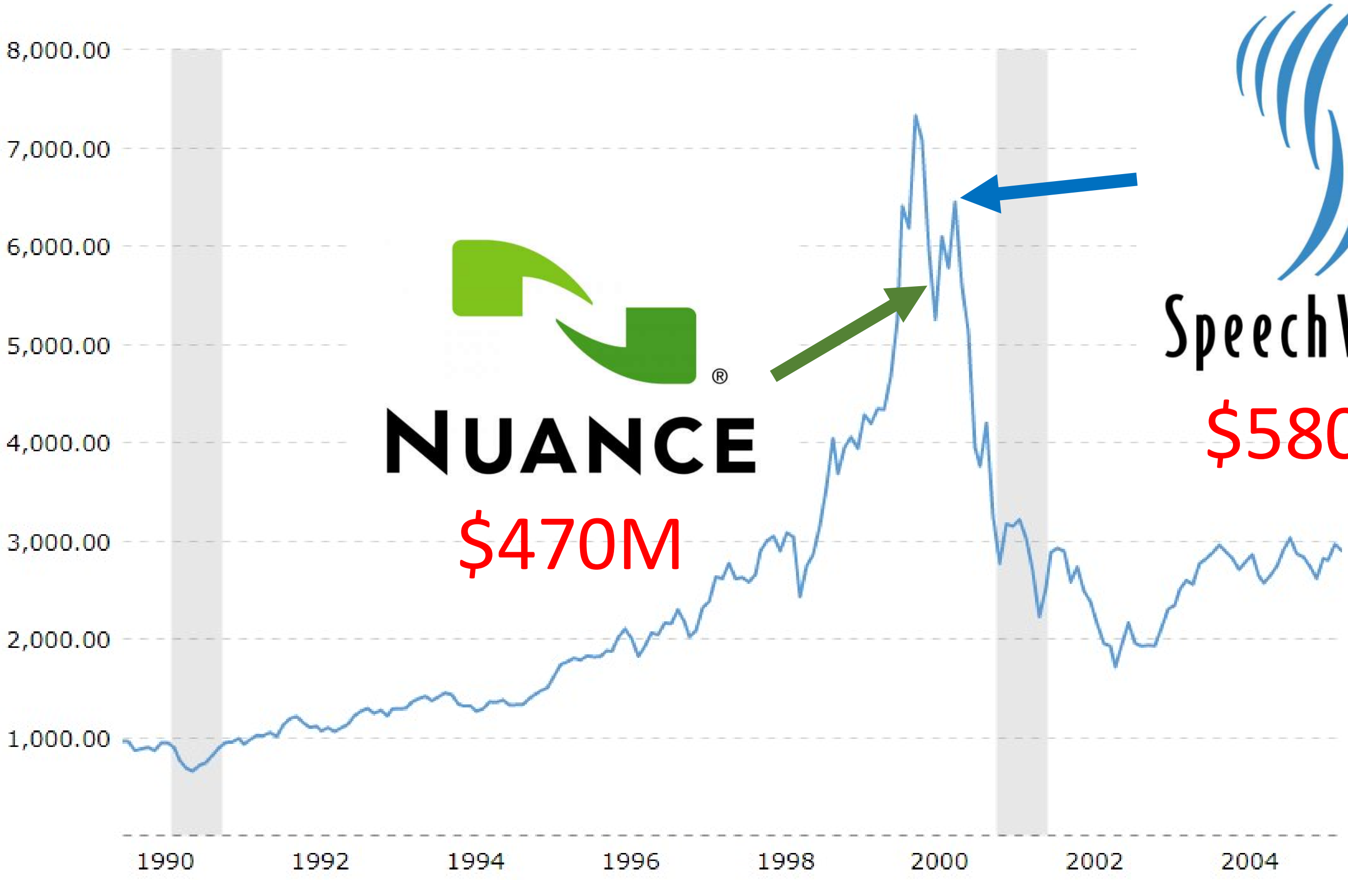
Saturday Night Live!

Gartner Hype Cycle

Speech Recognition



The year 2000

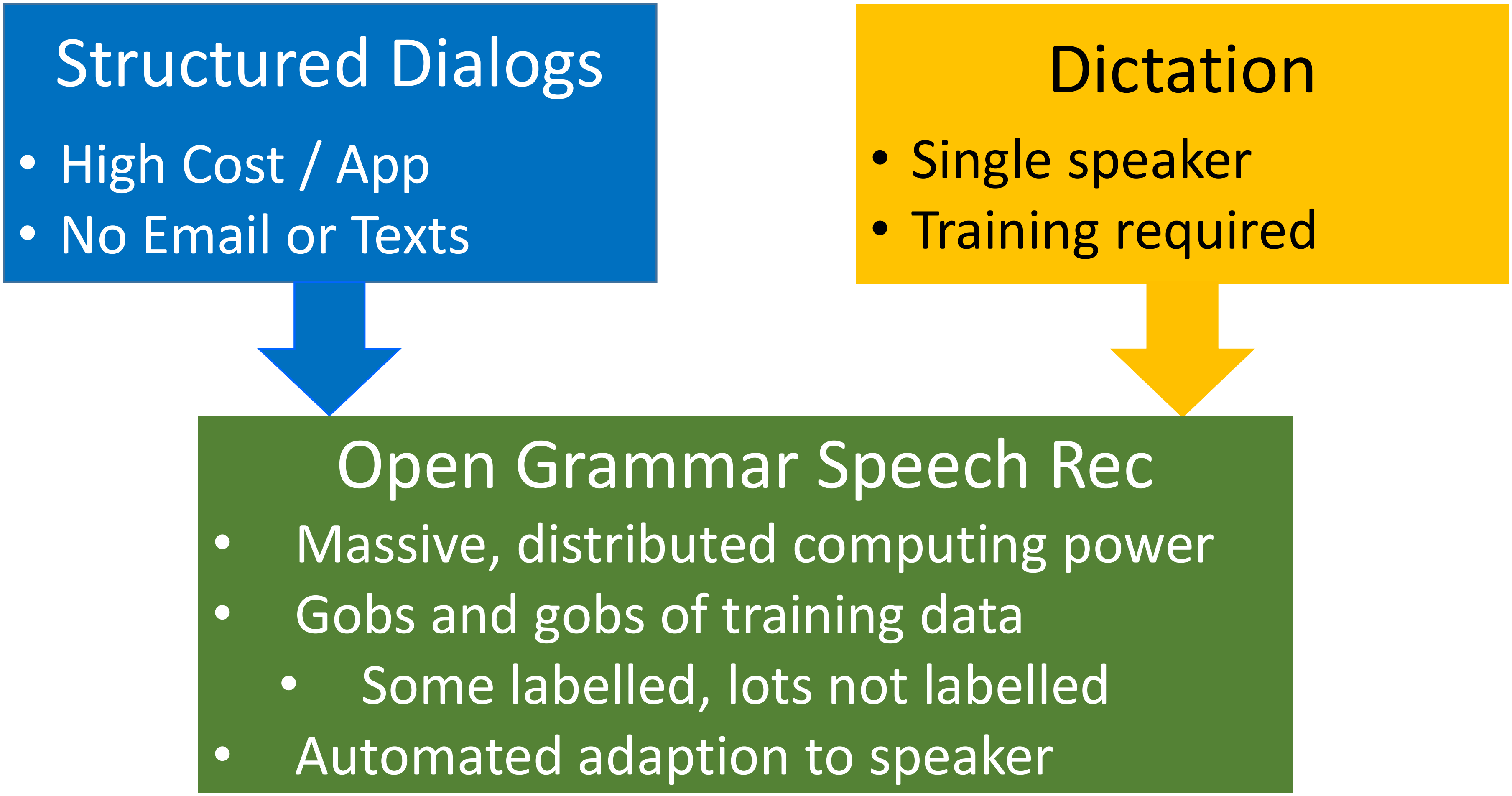


NUANCE

$470M

SpeechWorks

$580M



Structured Dialogs

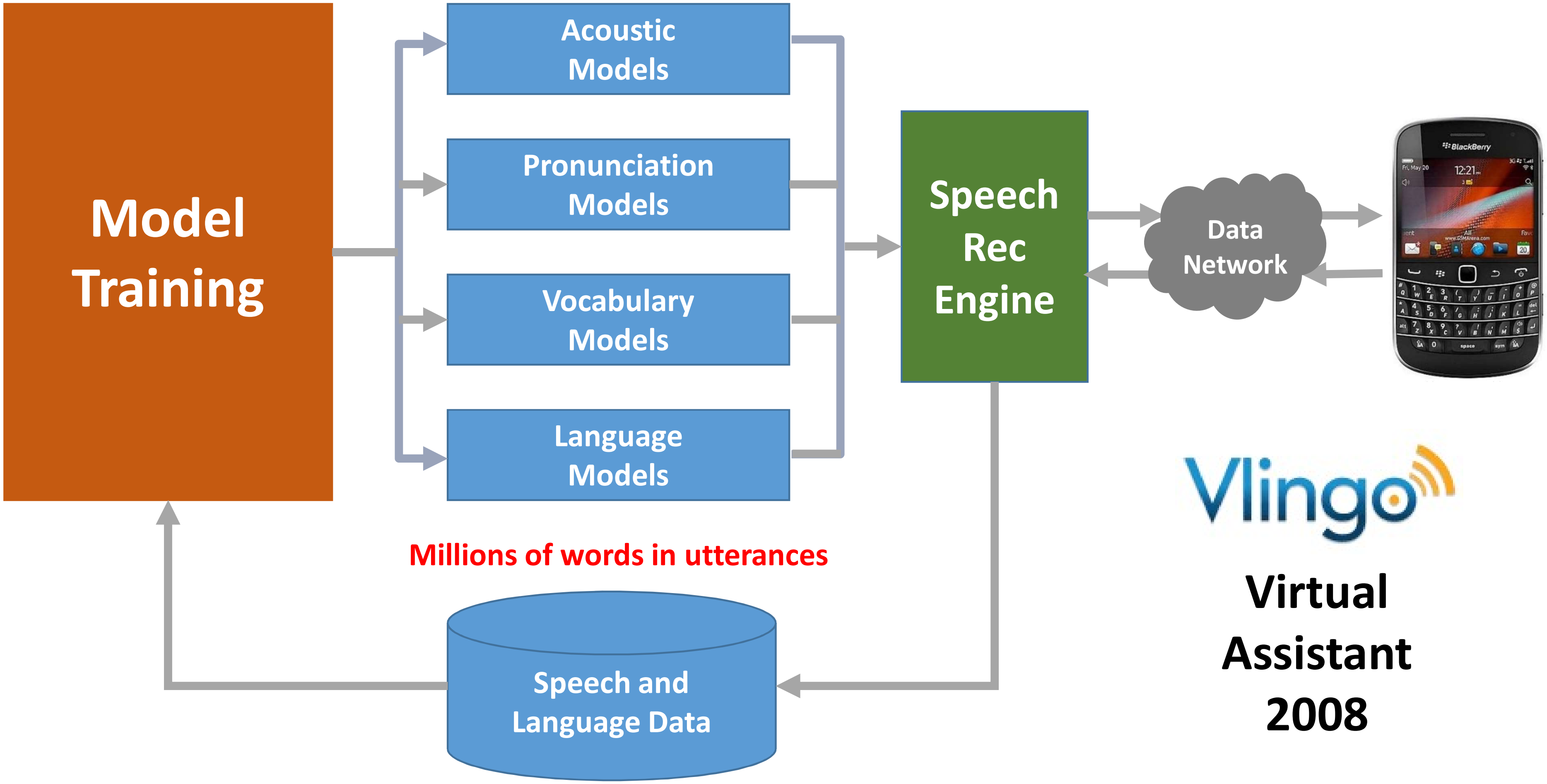
- High Cost / App
- No Email or Texts

Dictation

- Single speaker
- Training required

Open Grammar Speech Rec

- Massive, distributed computing power
- Gobs and gobs of training data
  - Some labelled, lots not labelled
- Automated adaption to speaker



Model Training

Acoustic Models

Pronunciation Models

Vocabulary Models

Language Models

Speech Rec Engine

Data Network

Millions of words in utterances

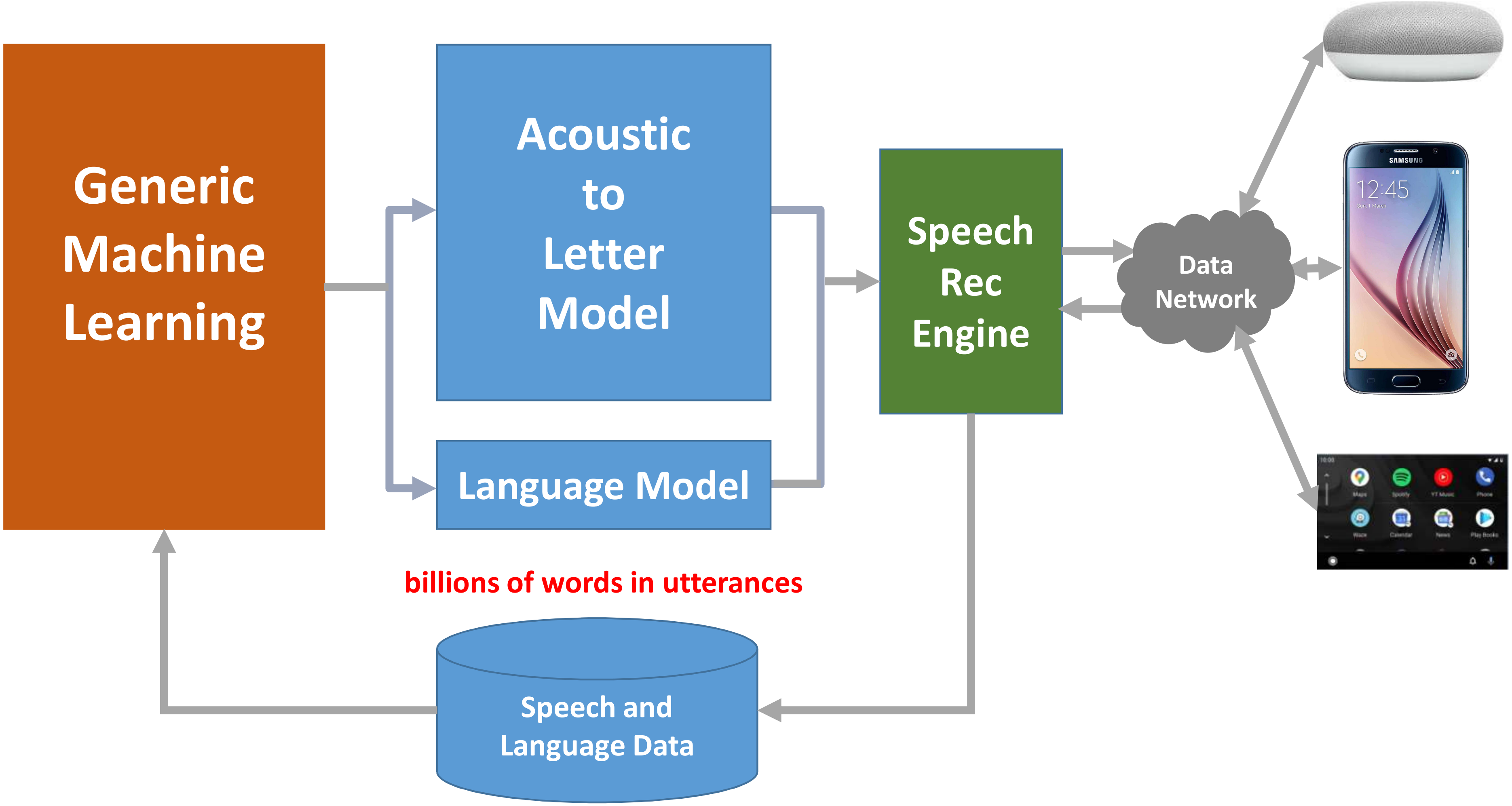
Speech and Language Data

Vlingo

Virtual Assistant 2008

Enter the Virtual Assistants

2010	Google Voice Actions SRI → **Apple** SIRI
2012	Google Now
2014	**amazon** Alexa ***Microsoft*** Cortana
2016	Google Assistant (2-way conversations)



Generic Machine Learning

Acoustic to Letter Model

Language Model

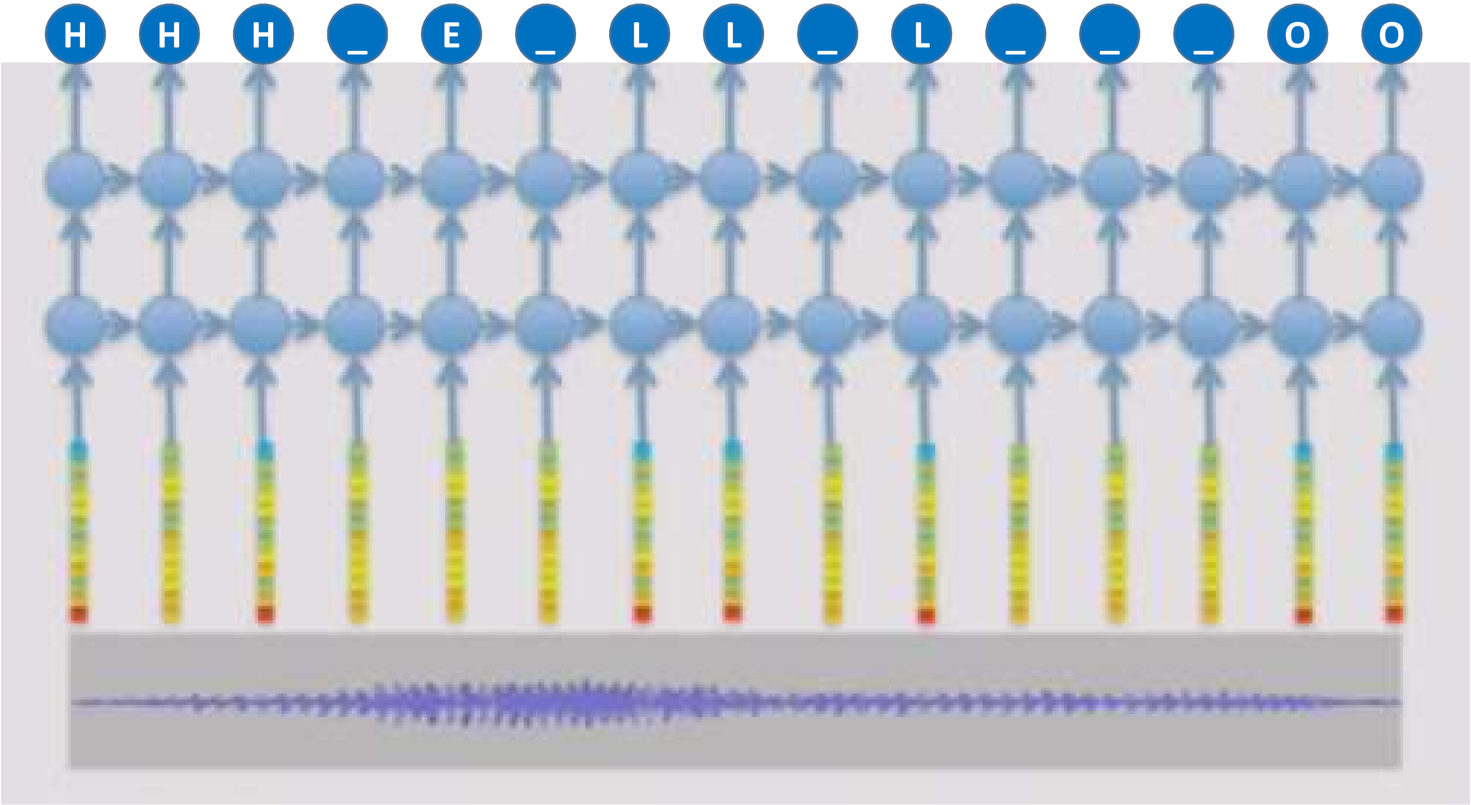
Speech Rec Engine

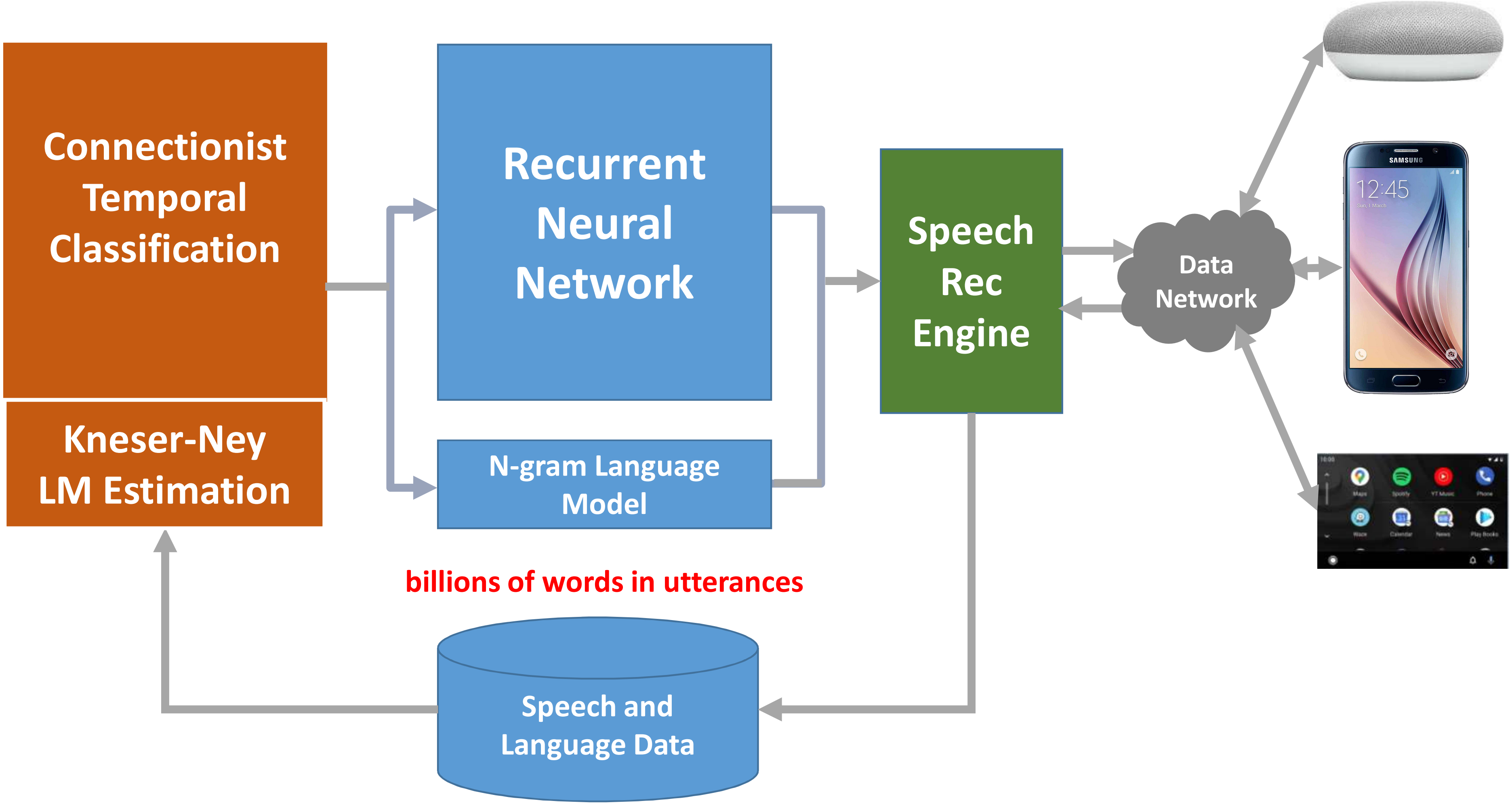
Data Network

billions of words in utterances

Speech and Language Data

Neural Network





Connectionist Temporal Classification

Kneser-Ney LM Estimation

Recurrent Neural Network

N-gram Language Model

Speech Rec Engine

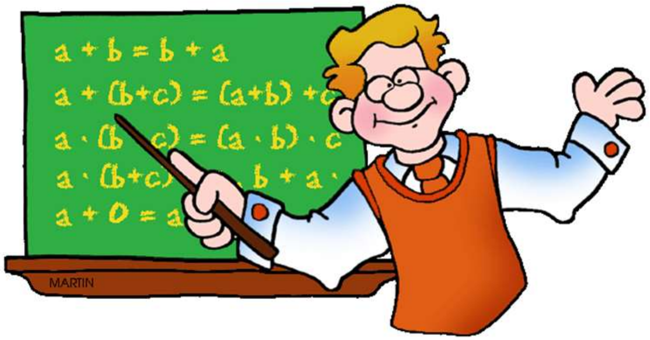
Data Network

billions of words in utterances

Speech and Language Data

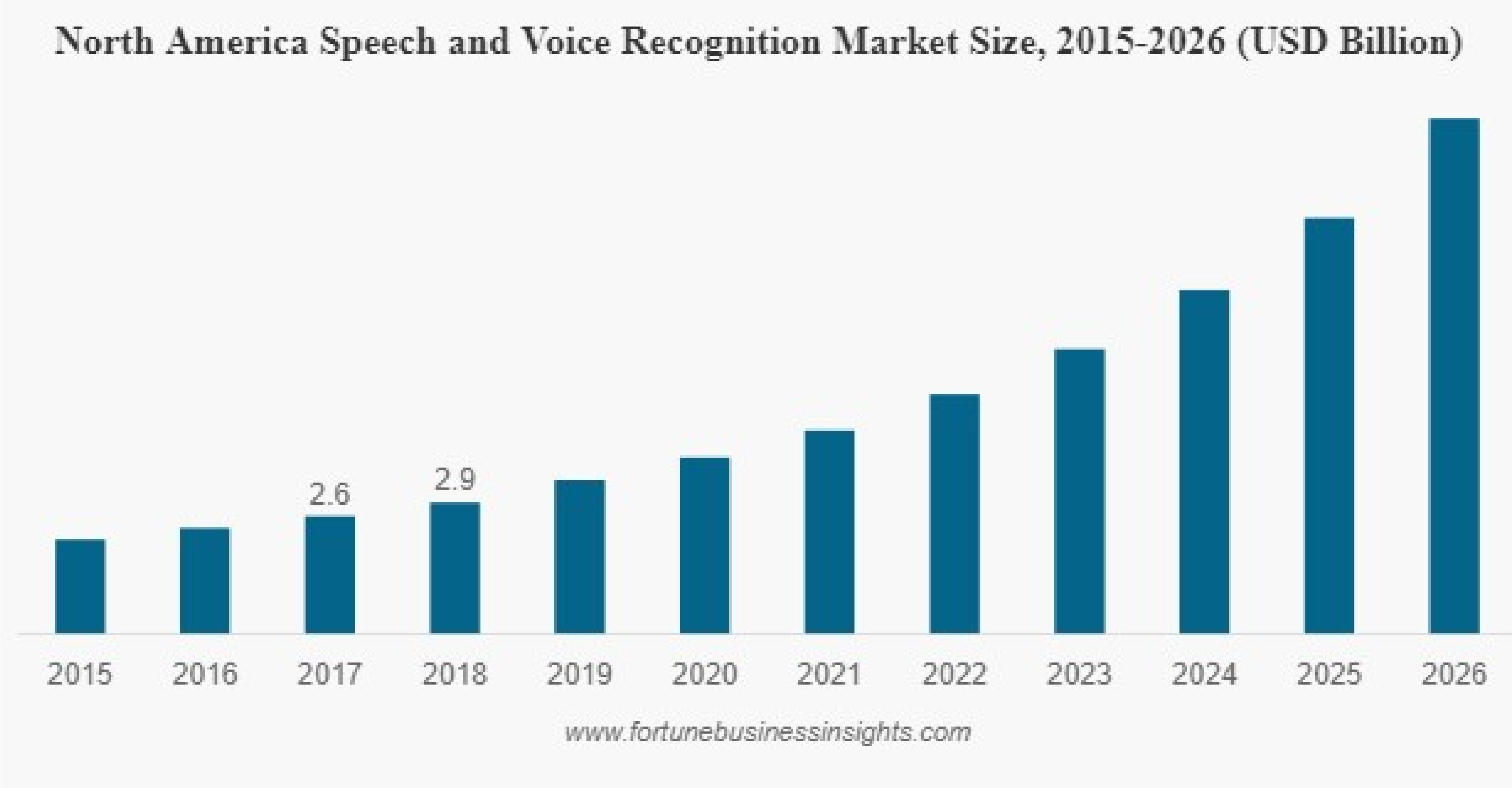
DATA

CanStock


a + b = b + a
a + (b+c) = (a+b) + c
a · (b · c) = (a · b) · c
a · (b+c) … b + a ·
a + 0 = a
MARTIN

+

It’s a business…

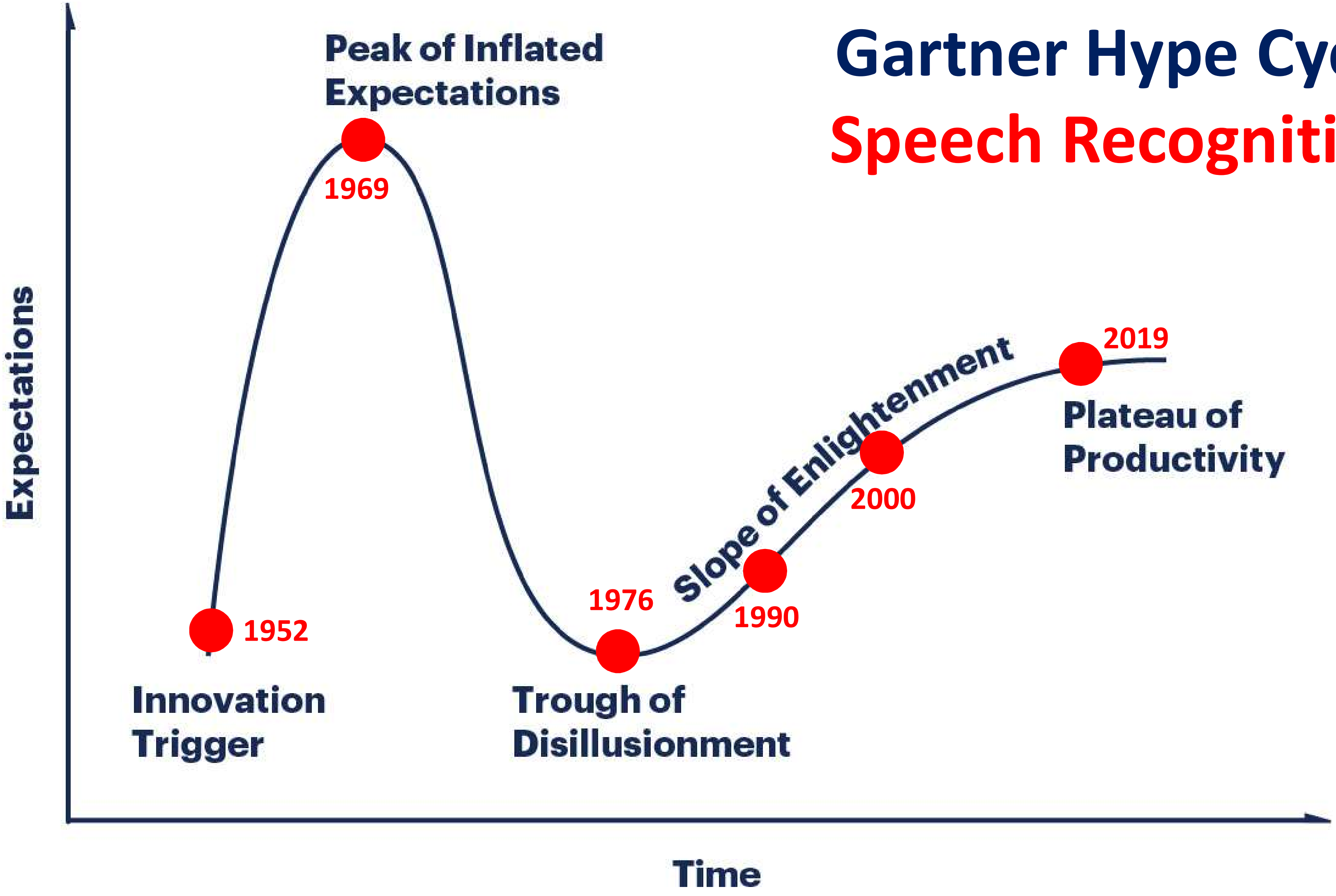


Biases Continue To Plague AI Voice Technology

Engadget (4/3/2021, Tarantola) reports that although the ability to hold conversations with computers has finally arrived, the technology that powers such devices as Alexa, Siri, and Google Home hasn't proven as revolutionary or as inclusive as initially hoped. While these systems "make a commendable effort to accurately interpret commands regardless of whether you picked up your accent in Houston or Hamburg, for users with heavier or less common accents such as Caribbean or Cockney, requests to their digital assistants are roundly ignored." Any technology "that reinforces or reinscribes bias," far from being revolutionary, only entrenches existing privileges and continues the oppression of less mainstream groups – in this case, those who speak English with heavy, non-Western accents. The author urges developers to be more inclusive in programming AI devices so that such devices understand and converse better with those who have non-standard English accents.

Gartner Hype Cycle

Speech Recognition



Scansoft
Kurzweil

NUANCE

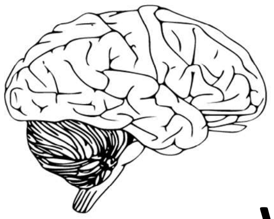
Microsoft

Voice Signal
BeVocal

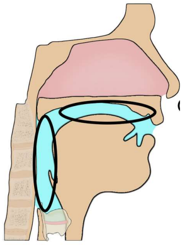
PureSpeech

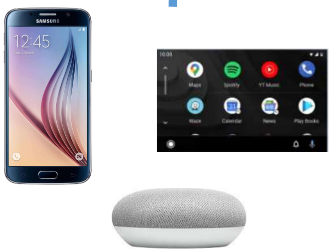
Vlingo

1974 *2005*



World Knowledge?	Pragmatics?
Concepts?	Semantics?
Phrases	Syntax
Words	Morphology
Phonemes	Phonetics
Sounds	Acoustics





Hey, Google…

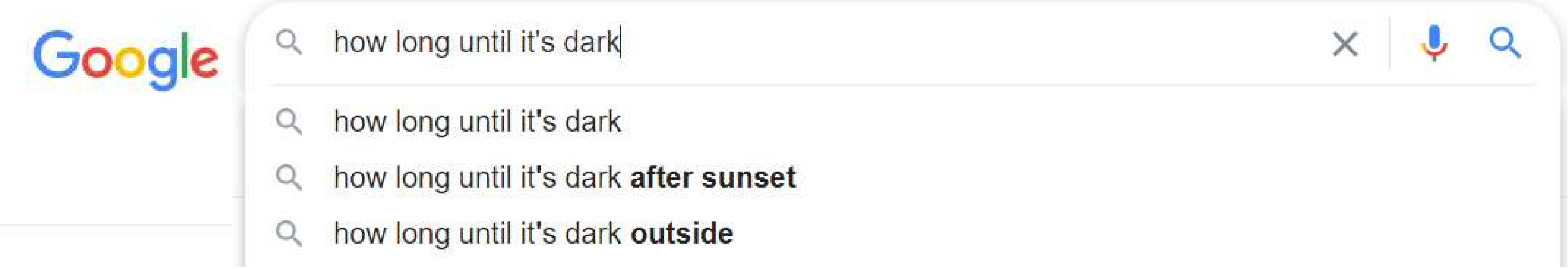
How long until sunset?

57 minutes

How long until it’s dark?

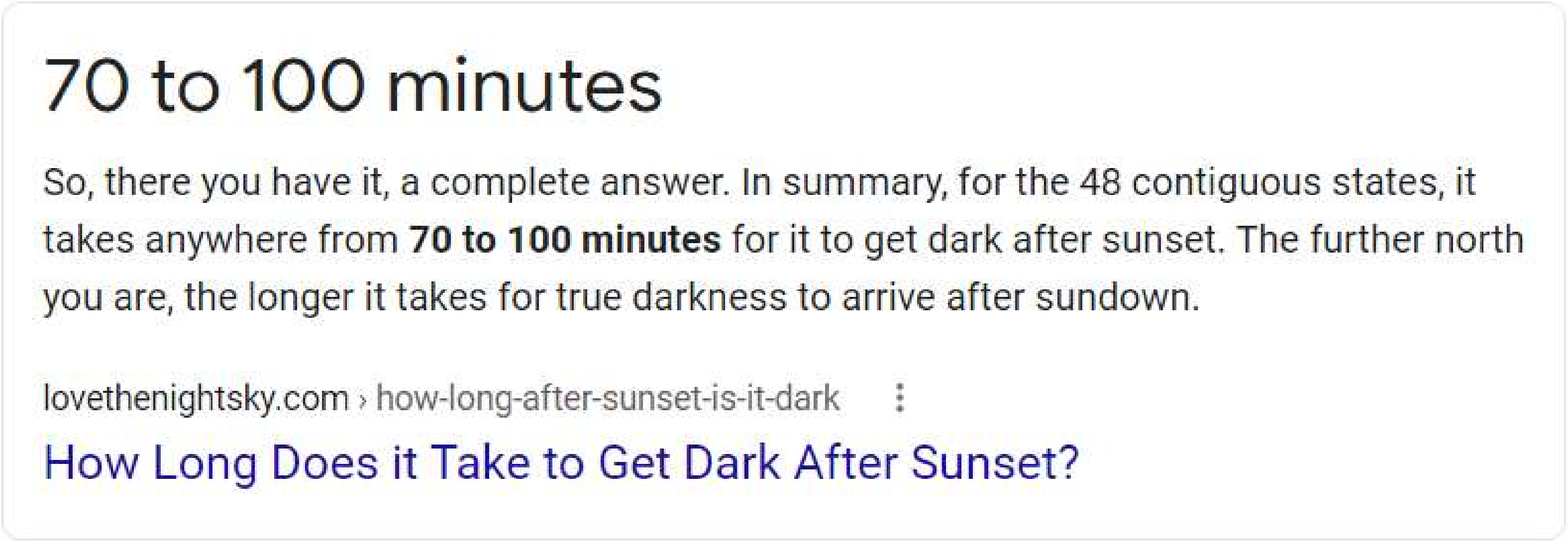
<pause> 70 to 100 minutes.

*<pause> According to the website love the night sky dot com: In summary, for the 48 contiguous states, it takes anywhere from **70 to 100 minutes** for it to get dark after sunset. The further north you are, the longer it takes for true darkness to arrive after sundown.*



how long until it's dark

- how long until it's dark
- how long until it's dark **after sunset**
- how long until it's dark **outside**



70 to 100 minutes

So, there you have it, a complete answer. In summary, for the 48 contiguous states, it takes anywhere from **70 to 100 minutes** for it to get dark after sunset. The further north you are, the longer it takes for true darkness to arrive after sundown.

lovethenightsky.com › how-long-after-sunset-is-it-dark

How Long Does it Take to Get Dark After Sunset?

GPT-3

The importance of being on twitter

by Jerome K. Jerome
London, Summer 1897

It is a curious fact that the last remaining form of social life in which the people of London are still interested is Twitter. I was struck with this curious fact when I went on one of my periodical holidays to the sea-side, and found the whole place twittering like a starling-cage. I called it an anomaly, and it is.

I spoke to the sexton, whose cottage, like all sexton's cottages, is full of antiquities and interesting relics of former centuries. I said to him, "My dear sexton, what does all this twittering mean?" And he replied, "Why, sir, of course it means Twitter."

"Ah!" I said, "I know about that. But what is Twitter?"

"It is a system of short and pithy sentences strung together in groups, for the purpose of conveying useful information to the initiated, and entertainment and the exercise of wits to the initiated, and entertainment and the exercise of wits to the rest of us."

Large Computer Language Models Carry Environmental, Social Risks

UW News

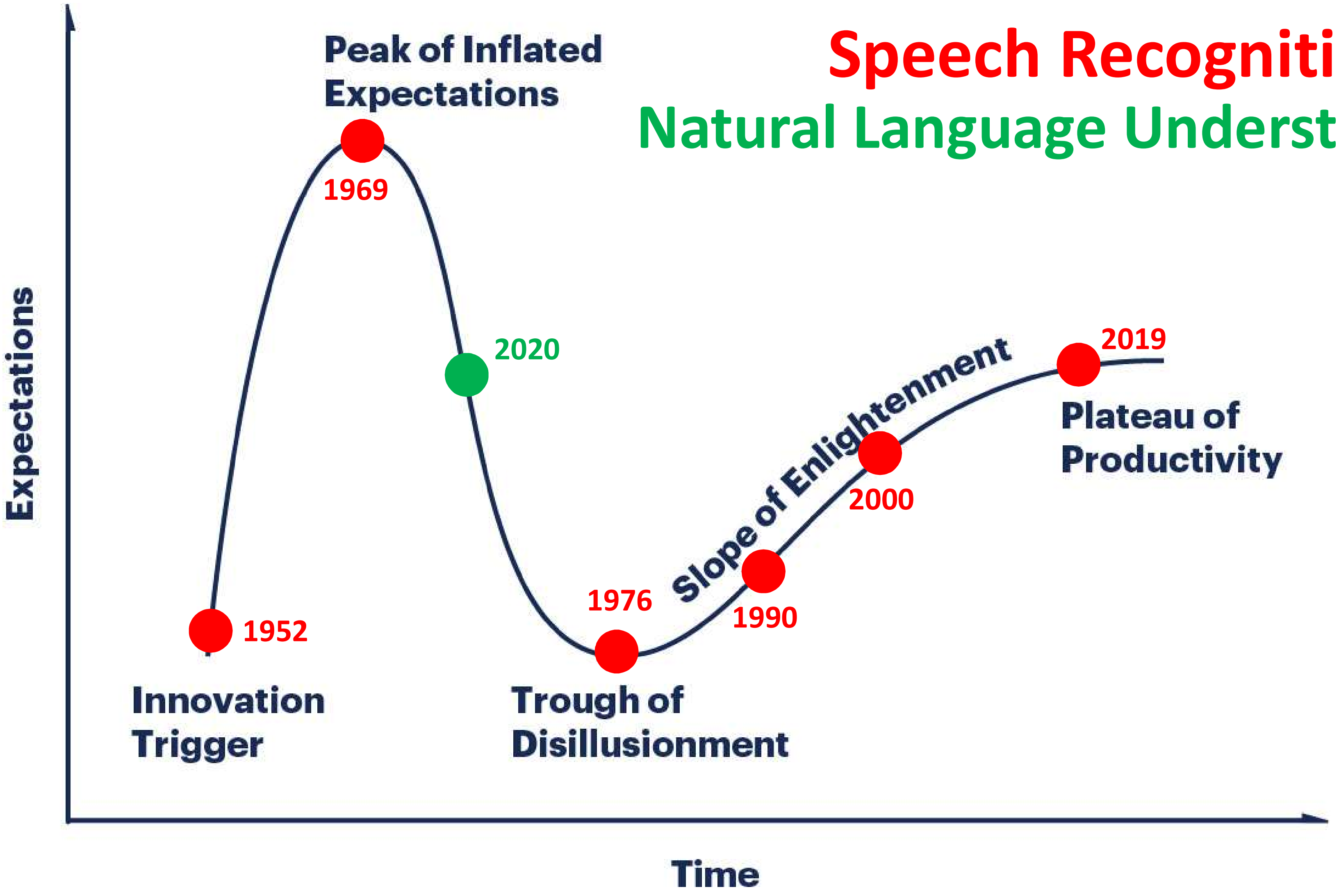
Jackson Holtz

March 10, 2021

University of Washington (UW) researchers warn that fast-growing computerized natural-language models can worsen environmental and social issues as the amount of training data increases. UW's Emily M. Bender and colleagues said the enormous energy consumption needed to drive the model language programs' computing muscle induces environmental degradation, with the costs borne by marginalized people. Furthermore, the massive scale of compute power can limit model access to only the most well-resourced enterprises and research groups. Critically, such models can perpetuate hegemonic language because the computers read language from the Web and other sources, and can fool people into thinking they are having an actual conversation with a human rather than a machine. Bender said, "It produces this seemingly coherent text, but it has no communicative intent. It has no idea what it's saying. There's no there there."

Speech Recognition

Natural Language Understanding



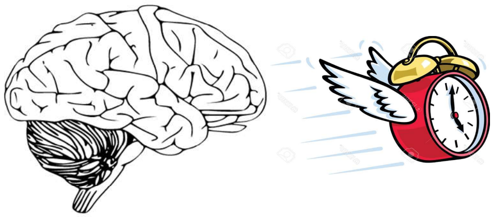
The Alexa Prize 2017-?

Create socialbots that can converse coherently and engagingly for 20 minutes with humans on a range of current events and popular topics such as entertainment, sports, politics, technology, and fashion while earning a rating of 4.0 out of 5.0.

2017	U Washington	10:22	3.17 Rating
2018	UC Davis	9:59	3.10 Rating
2021	???		

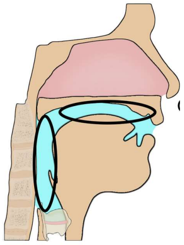
Time flies like an arrow.

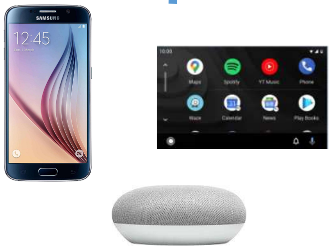
Fruit flies like a banana.

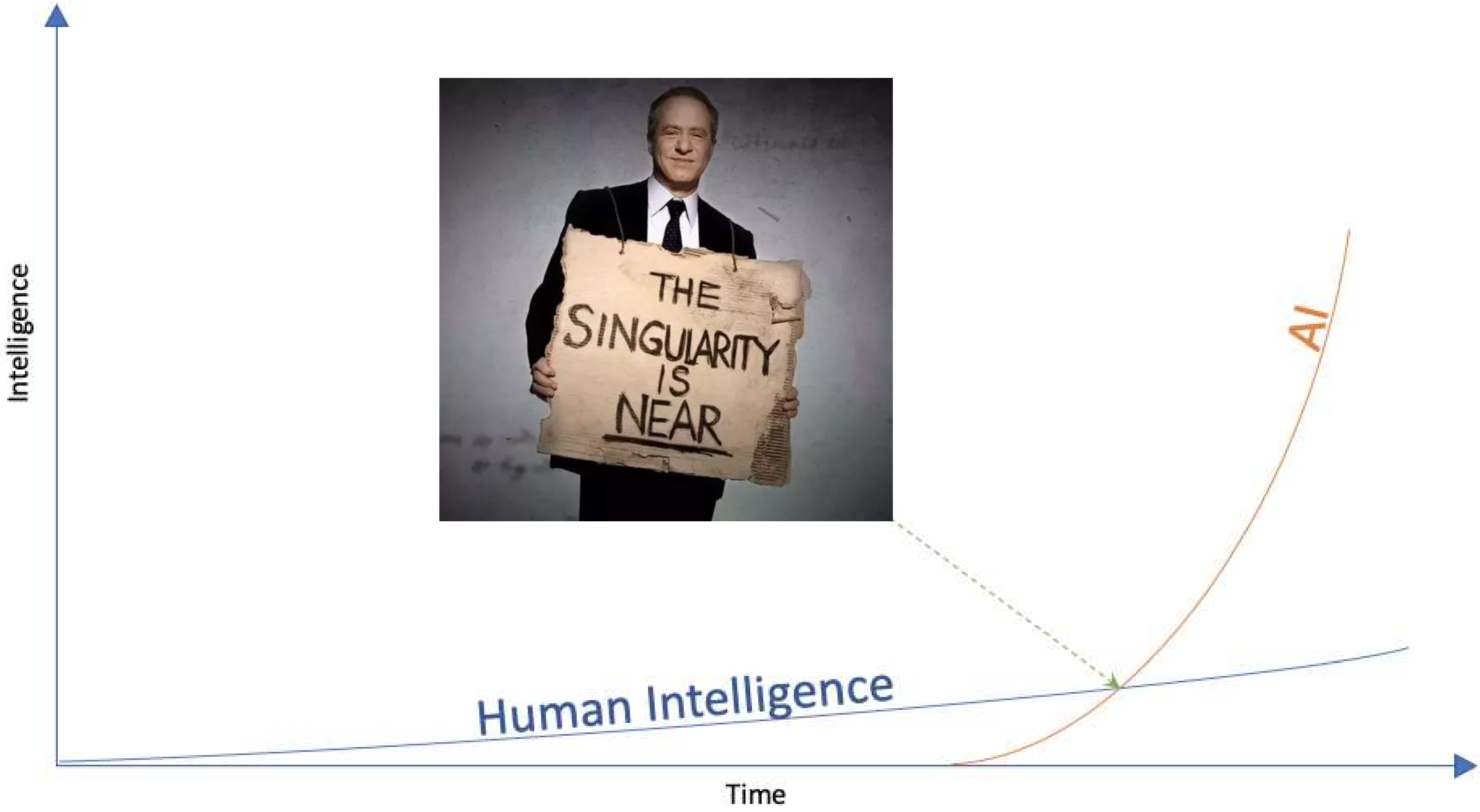


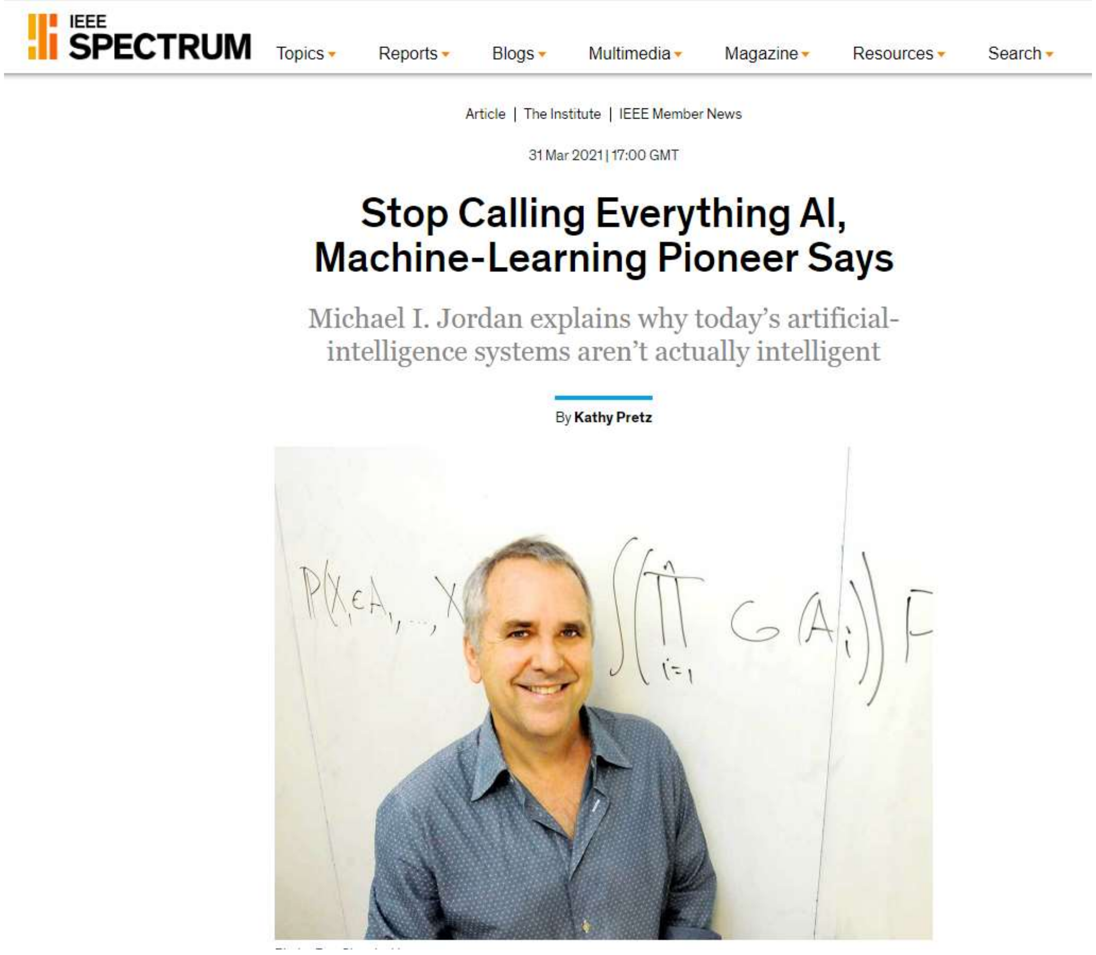
? ? ?

World Knowledge	?
Concepts	?
Phrases	✓
Words	✓
Phonemes	✓
Sounds	✓






Intelligence
THE SINGULARITY IS NEAR
AI
Human Intelligence
Time



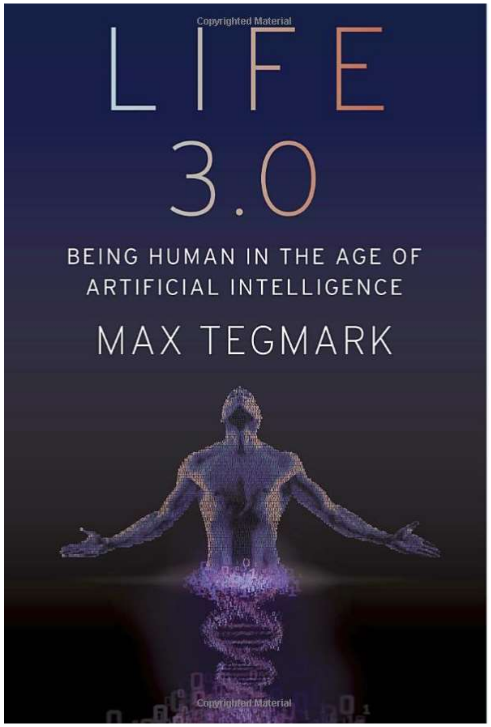
Article | The Institute | IEEE Member News

31 Mar 2021 | 17:00 GMT

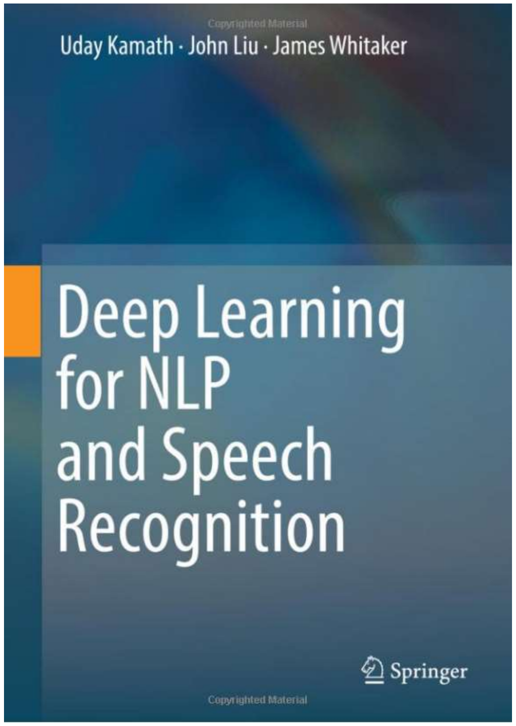
Stop Calling Everything AI, Machine-Learning Pioneer Says

Michael I. Jordan explains why today's artificial-intelligence systems aren't actually intelligent

By **Kathy Pretz**

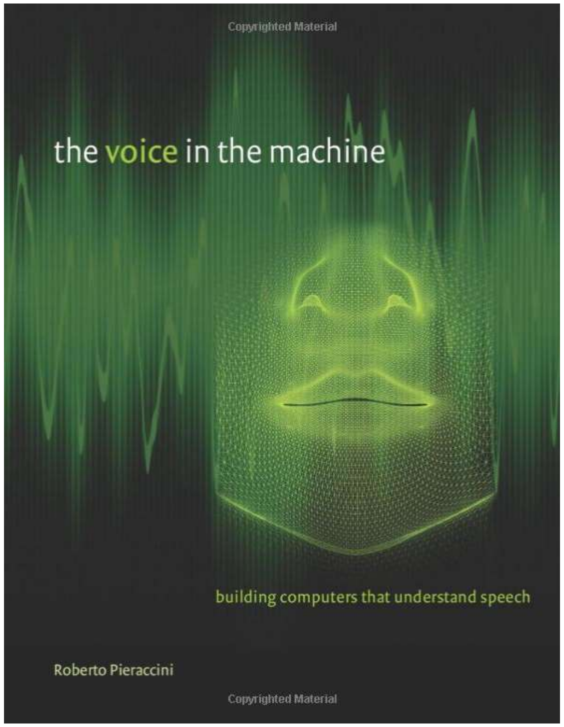


Natural Language
(and lots more)
… maybe

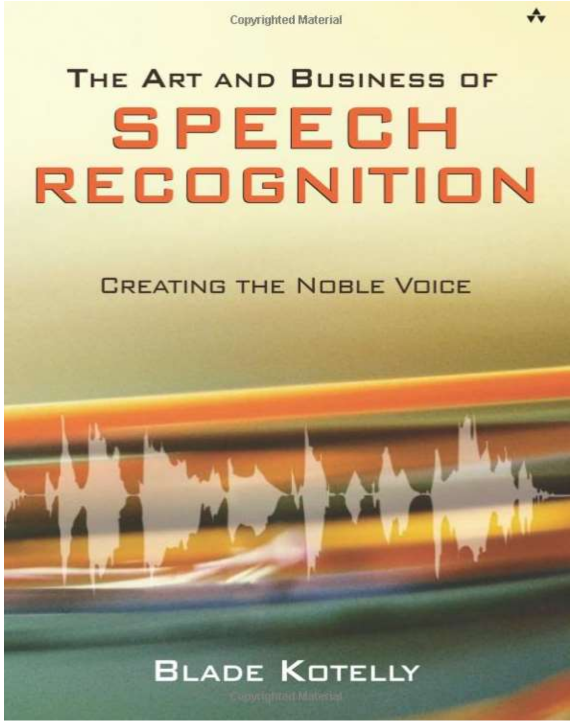


⚠ WARNING

Math Ahead



Speech Rec 101



The Art of
Dialog Design