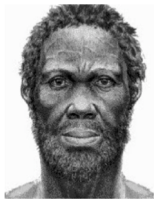
Speech

Pristerodon

Evolution of hearing

200 000 000 years



Homo sapiens

Evolution of speech

200 000 years

Environment (survival)

Hearing (communication)

We hear to survive

…. sensory neurons are adapted to the statistical properties of the signals to which they are exposed.

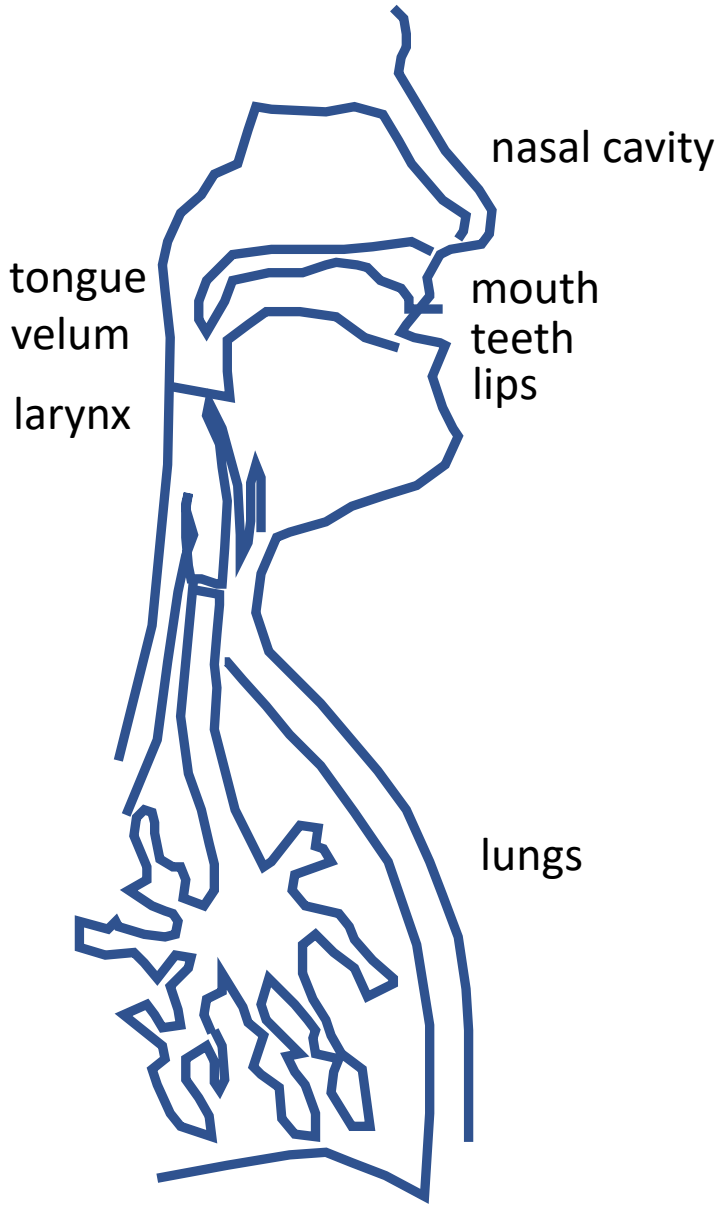
Simoncelli and Olshausen

We speak to hear

We speak in order to be heard and need to be heard in order to be understood.

Jakobson and Waugh p.95

Human speech evolved to fit properties of human hearing



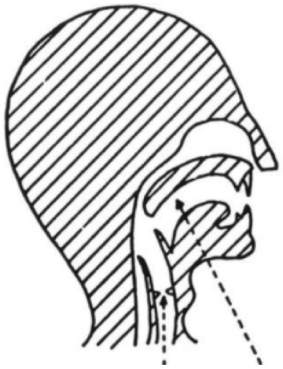
breathing
eating
biting

speaking?

Speech – sequence of short stationary (10-20 ms) signal segments

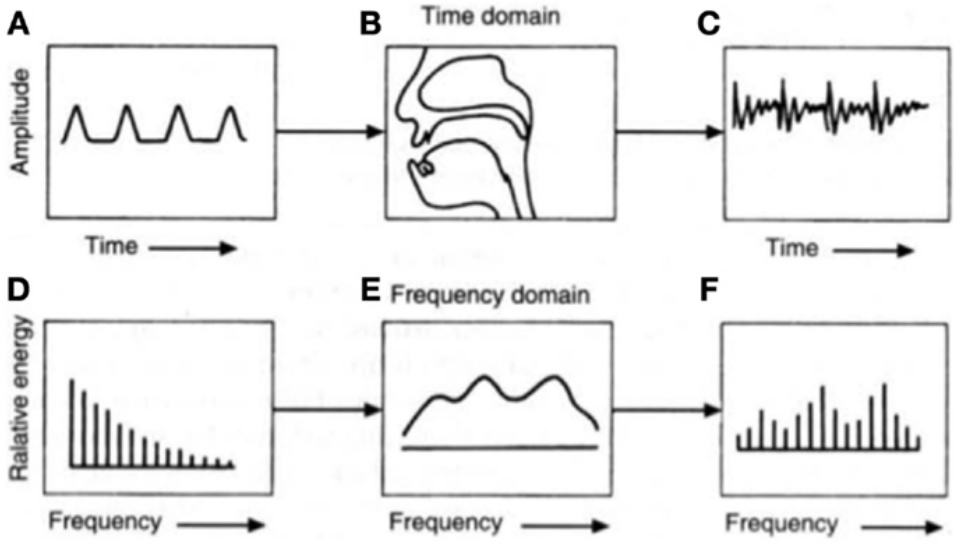
Linear model of speech production (Chiba and Kajiyama 194, Fant 1960, ..)

source → filter → filtered source signal

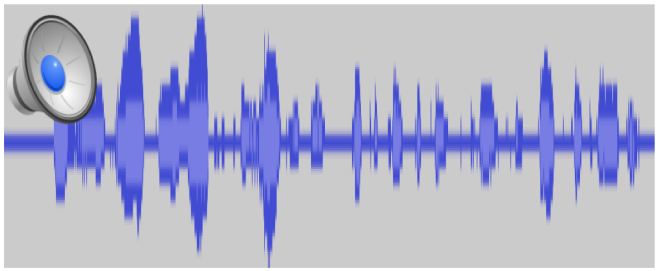


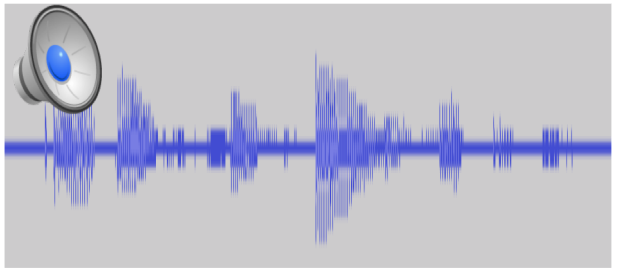
source filter

speech signal

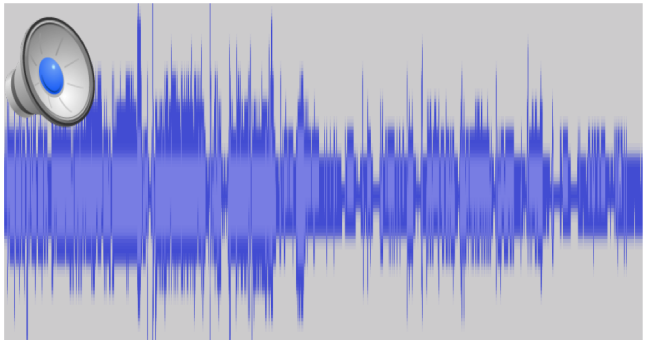


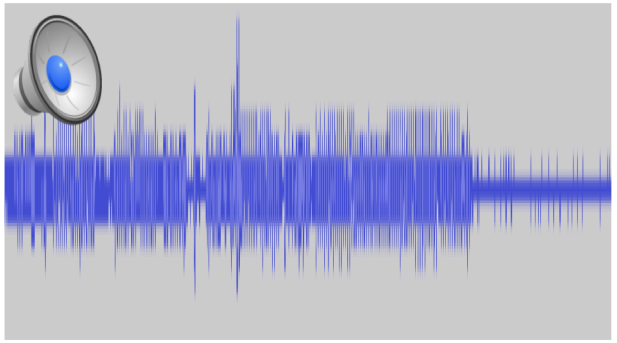
vocal source vocal tract speech signal



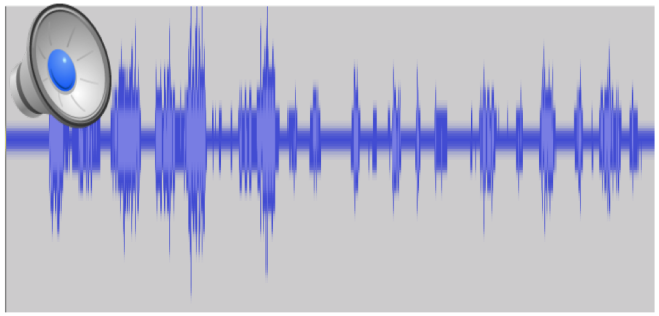


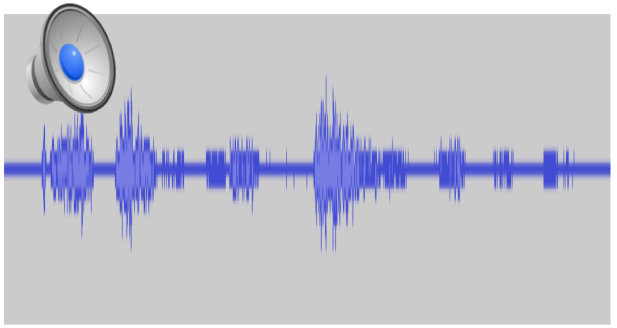
signal





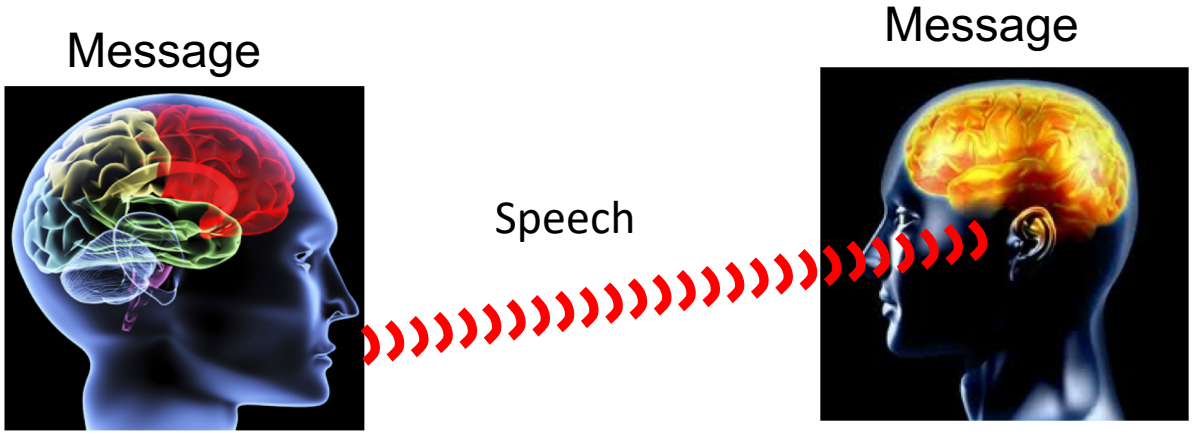
source





vocal tract

Human Speech



Messages

- Only a limited number of speech sounds can be produced and distinguished
- Many things need to be said

Composionality: meaning of a complex expression is determined by the meanings of its constituent expressions and the rules used to combine them (*Wikipedia)*

Create words as ordered sequences of speech sounds (phonemes).
file /fīl/
life /līf/

Create phrases as ordered sequences of words.
Tom chased horse.
Horse chased Tom.

Prior probabilities of different letters in English alphabet

Letter	Relative frequency	Letter	Relative frequency
e	12.702%	**m**	2.406%
t	9.056%	**w**	2.360%
a	8.167%	**f**	2.228%
o	7.507%	**g**	2.015%
i	6.966%	**y**	1.974%
n	6.749%	**p**	1.929%
s	6.327%	**b**	1.492%
h	6.094%	**v**	0.978%
r	5.987%	**k**	0.772%
d	4.253%	**j**	0.153%
l	4.025%	**x**	0.150%
c	2.782%	**q**	0.095%
u	2.758%	**z**	0.074%

Samuel Morse (self-portrait)

Morse code

e - single dot
z - dot and three dashes

In 1939, Ernest Vincent Wright published a 267-page novel, Gadsby, in which **no use is made of the letter E**. Here is a paragraph from the novel:

Upon this basis I am going to show you how a bunch of bright young folks did find a champion; a man with boys and girls of his own; a man of so dominating and happy individuality that Youth is drawn to him as is a fly to a sugar bowl. It is a story about a small town. It is not a gossipy yarn; nor is it a dry, monotonous account, full of such customary "fill-ins" as "romantic moonlight casting murky shadows down a long, winding country road." Nor will it say anything about tinklings lulling distant folds; robins carolling at twilight, nor any "warm glow of lamplight" from a cabin window. No. It is an account of up-and-doing activity; a vivid portrayal of Youth as it is today; and a practical discarding of that worn- out notion that "a child don't know anything."

How “efficient” is a given code?

Entropy

$$H(s) = -\sum_{i=1}^{n} p_i \cdot \log(p_i)$$

26 letters and space of English alphabet

$$H(s) = -\sum_{i=1}^{27} 1/27 \cdot \log(1/27)$$

$$= -\log(1/27) = 4.74 \text{ bit}$$

all letters are equally probable (zero order)

H(s)= 4.74 bit

Respecting relative frequencies of letters
(first order)

H(s)= 4.28 bit

Respecting relative frequencies of combinations of three letters (third order)

H(s)= 2.77 bit

Letters in real text (estimate)

H(s) ~ 0.6-1.3 bit

Shannon
Prediction and Entropy of Printed English
BSTJ 1951

The Relative Frequency of Phonemes in General-American English

Hayden 1950

Phoneme	Frequency Percentage					
	per cent					
ə	9.96	n	7.95	f	1.61	
I	9.75	t	7.59	y	1.20	
æ	3.09	r	7.10	g	1.14	
ɛ	2.03	s	4.89	h	1.11	
e	1.94	l	3.65	š	0.87	
a	1.80	đ	3.35	ŋ	0.80	
i	1.66	d	3.21	č	0.53	
u	1.52	k	2.98	ǰ	0.50	
O	1.49	m	2.87	θ	0.44	
a^i	1.46	z	2.36	w	0.37	
ɔ	1.02	v	2.33	ž	0.03	
U	0.99	p	2.25		62.6	
a^u	0.64	w	1.77			
o^i	0.06	b	1.65			
	37.4					

Phonemes

Perceptually distinct speech elements that could distinguish one words from another

Graphemes

Letters and combinations of letters representing speech sounds (phonemes)

Rotokas language – East of New Guinea, 11 phonemes,
12 symbols, 1 symbol per sound

Taa language – Botswana (Africa), ~ 200 phonemes ,
20-22 symbols, up to 6 symbols per sound

English
~45 phonemes, 27 symbols,
~ 250 graphemes, up to 5 symbols per sound

40 speech sounds (phonemes) in American English
24 consonants
19 vowels and diphtongs

vowels – mouth open
consonants - mouth not so open

typical syllable

cvc
onset – nucleus – coda
cv
onset – nucleus

/l/,/r/,/w/,/y/ - semivowels
produced with open mouth
can stand as nucleus in syllable

Words

- ordered combinations of speech sounds
- represent objects, ideas, actions, relationships, qualities, e.t.c., **as agreed on by a particular society (language)**
- new words constantly invented and old words changing their meanings
- learned using interventions and rewards from other human beings
- particular word meanings often depend on context

Word sequences (sentences, phrases,..)

- Words organized into larger units (sentences, phrases,..) using rules of the language (syntax, grammar)

- Order also carries information
    - John beats Frank. Frank beats John.
    - I went home and had a dinner. I had a dinner and went home.

Relative frequencies of words in written English [%]

7.31	the	.58	not	.31	their	.20	time	.15	these
3.99	of	.58	at	.30	there	.20	up	.14	two
3.28	and	.57	this	.30	were	.20	do	.14	very
2.92	to	.54	are	.30	so	.20	out	.13	before
2.12	a	.52	we	.29	my	.19	can	.13	great
2.11	in	.51	his	.26	if	.19	than	.13	could
1.34	that	.50	but	.25	me	.18	only	.13	such
1.21	it	.47	they	.25	what	.18	she	.13	first
1.21	is	.46	all	.25	would	.17	made	.12	upon
1.15	I	.45	or	.24	who	.16	other	.12	every
1.03	for	.45	which	.23	when	.16	into	.12	how
.84	be	.44	will	.23	him	.16	men	.12	come
.83	was	.43	from	.22	them	.16	must	.12	us
.78	as	.41	had	.22	her	.16	people	.12	shall
.77	you	.39	has	.21	war	.16	said	.11	should
.72	with	.36	one	.21	your	.16	may	.11	then
.68	he	.33	our	.21	any	.15	man	.11	like
.64	on	.33	an	.21	more	.15	about	.11	well
.61	have	.32	been	.21	now	.15	over	.11	little
.60	by	.32	no	.20	its	.15	some	.11	say

In spoken language most frequency word is pronoun “I”

- Telephone conversations 5%
- Schizophrenics 8.4%

Predictability and unpredictability

- 100 % predictable message has no information value
  - When knowing exactly what will be said, no need to listen

- Speech is to large extent predictable since is follows rules
  - Grammar, use of words, word order, …

- The predictability allows for easier communication

To communicate effectively, the right balance between predictability and unpredictability need to be maintained.

How predictable is language? - Claude Shannon

1. Think about the English sentence
2. Ask people to think about the first letter in the sentence
3. When correct, tell them, mark it by "-" and ask for the second letter
4. When incorrect, tell them the correct one and ask for the second letter
5. Go on until the end of the sentence

```
(1) THE ROOM WAS NOT VERY LIGHT A SMALL OBLONG
(2) ----ROO------NOT-V-----I------SM----OBL----
(1) READING LAMP ON THE DESK SHED GLOW ON
(2) REA----------O------D----SHED-GLO--O--
(1) POLISHED WOOD BUT LESS ON THE SHABBY RED CARPET
(2) P-L-S-----O---BU--L-S--O------SH-----RE--C------
```

69% of letters guessed correctly

Both line (1) and (2) contain the same information

- The line (1) can be guessed from the info in the line (2) – by the identical twin ☺

Connection between written and spoken language

pronunciation dictionary

/prəˌnʌnsɪˈeɪʃ(ə)nˈdɪkʃən(ə)ri/

Variability

- Wanted variability:

  carries information about message, which we want to extract (signal)

- Unwanted variability:

  carries “other” information (**noise**)

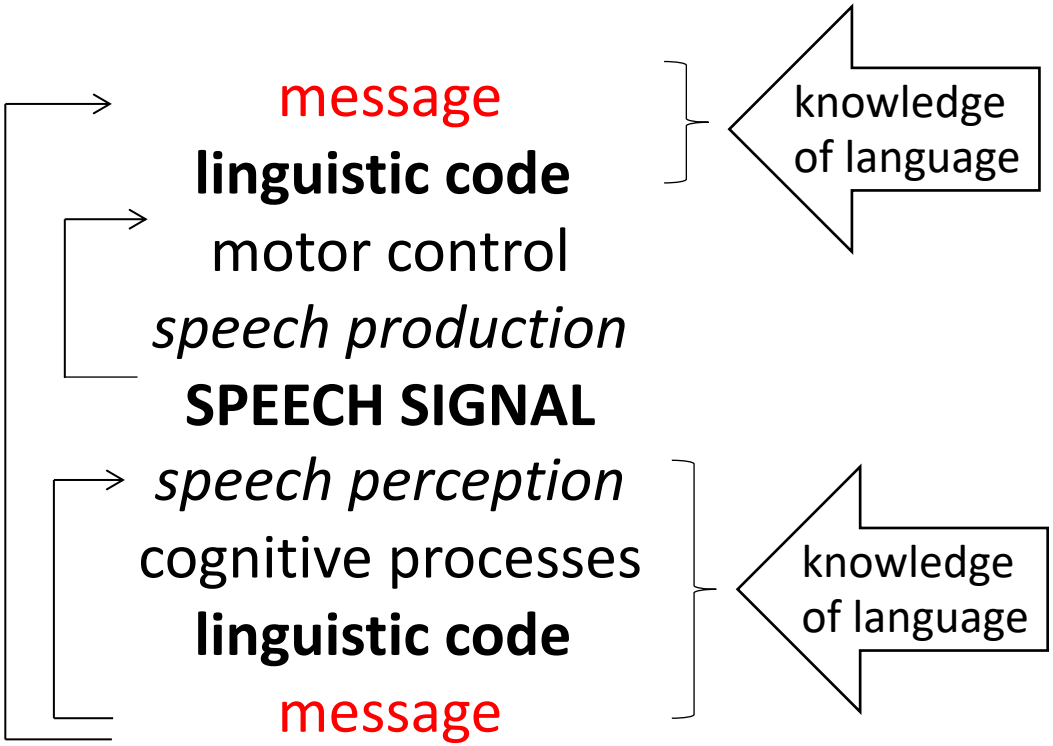
Noise: the good, the bad, and the ugly

- The effect of the noise is known
  - e.g., known additive noise, linear distortions, first order effects of speaker vocal tract anatomy,…
    - spectral subtraction, RASTA filtering, vocal tract normalization,...

- We know this noise may come but its effect is not known
  - e.g., various environmental noises, reverberations, speaker peculiarities, language phonetics, accents, ….
    - multistyle training,...

- A new unexpected and previously unseen noise is coming and we do not know its effect
  - e.g. noise with new spectral and temporal composition, another new speaker is speaking (cocktail party effect)
    - high-level cognitive processing (adaptation with performance monitoring, attention, …)

Human Speech

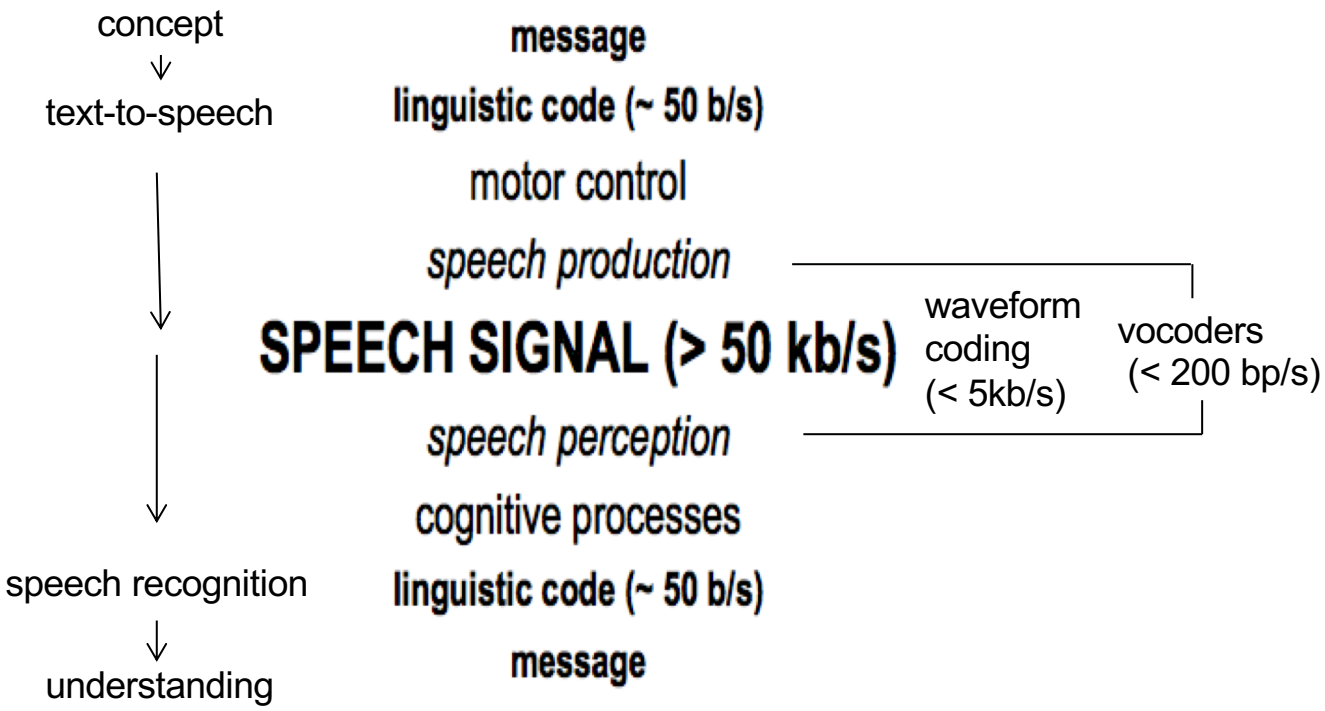


27 symbols of English alphabet
~15 symbols/s, each symbol ~ 3 bits/symbol
> 50 b/s

standard PCM coding
8 kHz sampling, 11 bit accuracy
88 kb/s

> 50 b/s

INFORMATION in speech signal: **message,** who is speaking, health, language, emotions, mood, social status, acoustic environment, etc,…


concept
text-to-speech
speech recognition
understanding
message
linguistic code (~ 50 b/s)
motor control
speech production
SPEECH SIGNAL (> 50 kb/s)
speech perception
cognitive processes
linguistic code (~ 50 b/s)
message
waveform coding (< 5kb/s)
vocoders (< 200 bp/s)

Where is this linguistic information in speech?

The filling of a very deeepe flaggon wth a constant streame of beere or water sounds y^{e} ^{v}owells in this order w, u, **ω**, o, a, e, I, y,

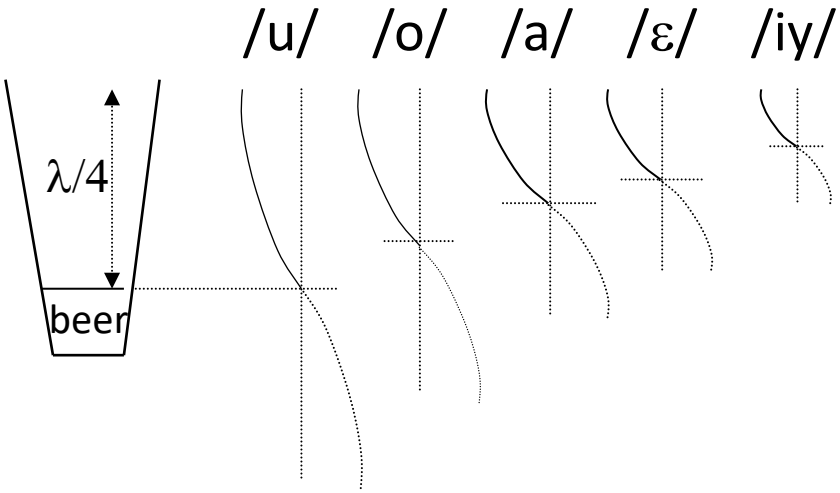
Isaac Newton

from his notes, probably 1659-1662
(17-20 years old)

Ralph W. V. Elliott: Isaac Newton as Phonetician, The Modern Language Review, Vol. 49, No. 1 (Jan., 1954), pp. 5-12

Is it in spectral components of the speech signal ?

Newton's Spectral Analysis of Speech



/u/ f1 = 300 Hz,

$\lambda/4$ = 25 cm

Loving ffreind

It is commonly reported y^t you are sick. Truely I am sorry for y^t. But I am much more sorry y^t you got yo^r sicknesse (for y^t they say too) by drinking too much.

Yo^r very loving freind

I. N.

from his notes, probably 1659-1662 (17-20 years old)

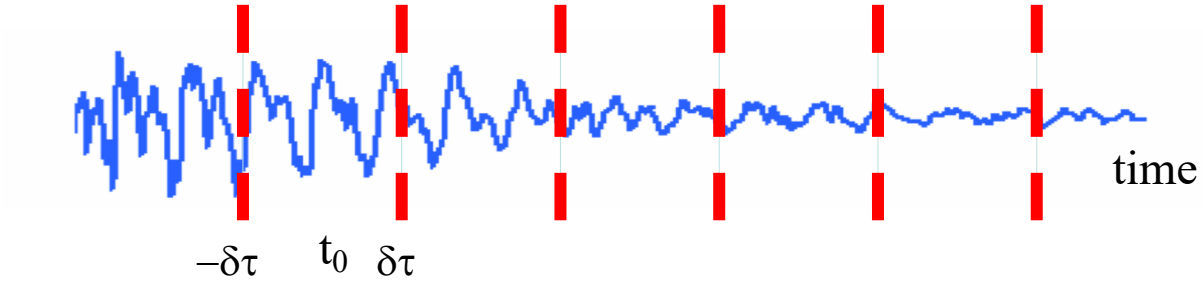
Why speech?

- Profit
  - searching large speech databases, transcription, voice control,…
- *voice will do to touch what touch did to keyboards*.
  - Mooly Eden, senior vice president Intel

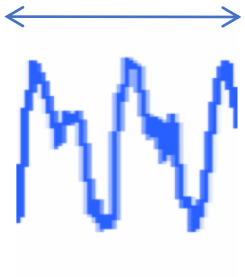
- Important spin-offs
  - Digital signal processing
  - Sequence classification (Hidden Markov Models)
    - financial predictions
    - human DNA matching
    - action recognition
  - Image processing techniques

Spoken language is one of the most amazing accomplishments of human race.

non-stationary speech signal s(t)

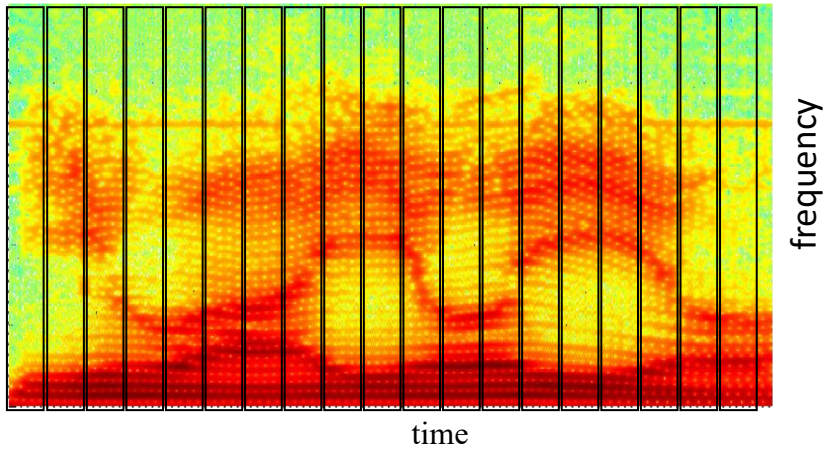


$\Delta t=2\delta\tau$

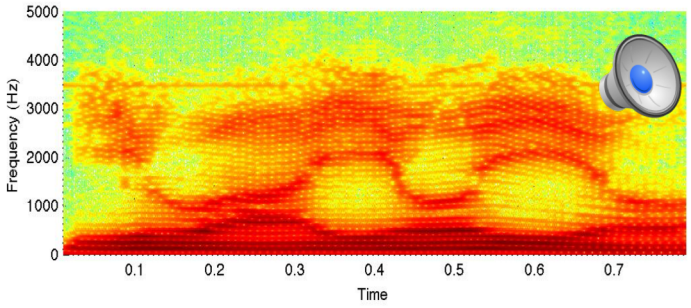


$\rightarrow$ **S(ω)**= $\mathcal{F}\{s(t_0\pm\delta\tau)\}$

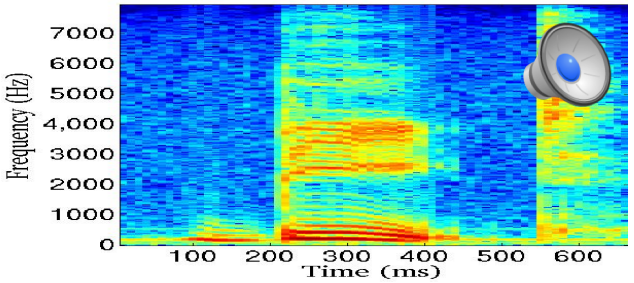
time-frequency representation of the signal
(spectrogram)



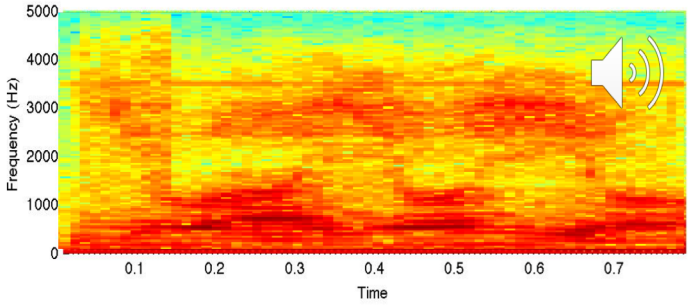
NORMAL



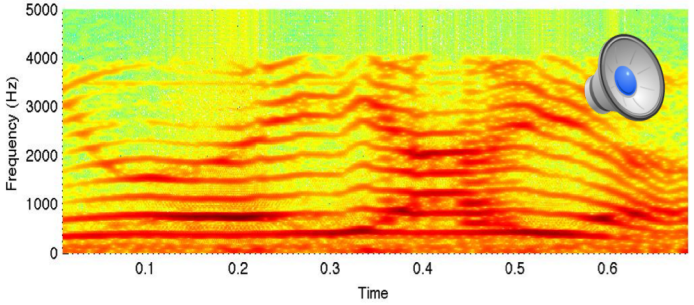
NORMAL



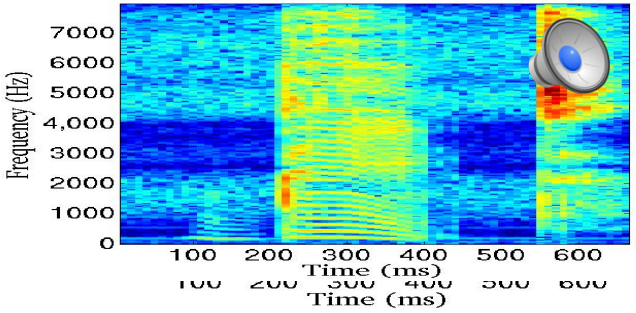
REVERBERATED



CHILD



SEVERELY FILTERED (FLAT SPECTRUM VOWEL)

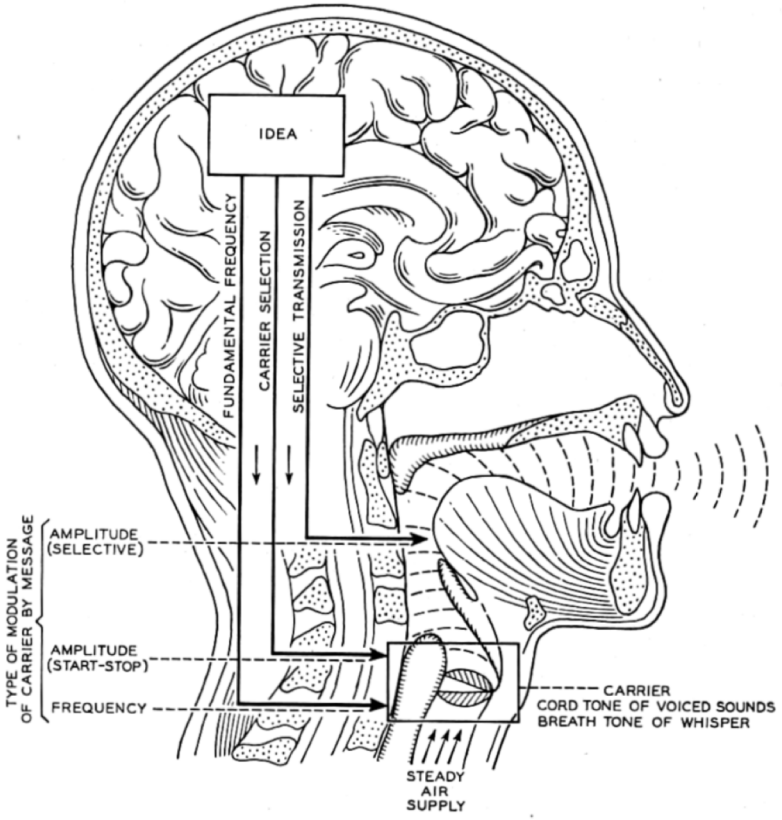


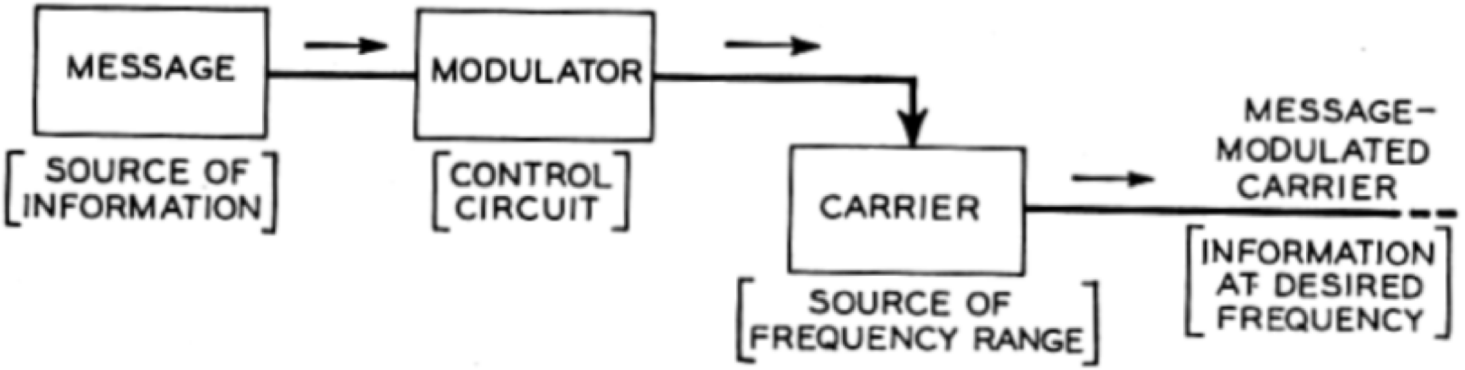
The Bell System Technical Journal

Vol. XIX *October, 1940* *No. 4*

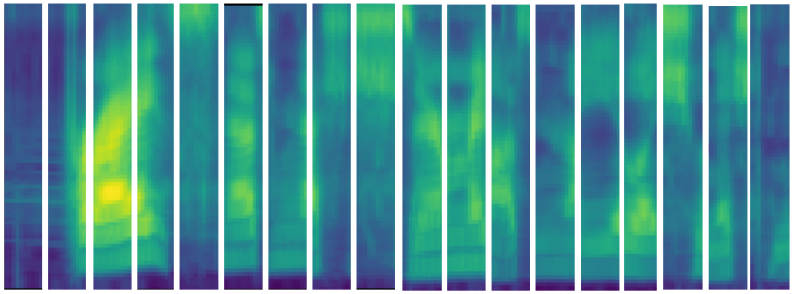
The Carrier Nature of Speech

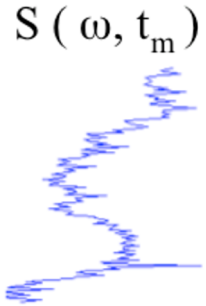
By HOMER DUDLEY





Short-term analysis Oppenheim 1970



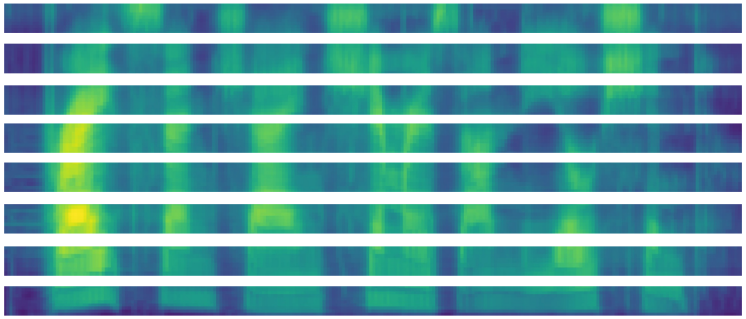


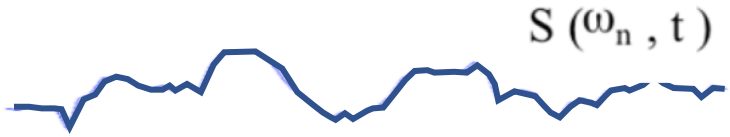
Short-time spectrum of speech?

It is fortunate that speech intelligibilty does resist erosion of frequency selectivity, for **normal environment plays havoc with speech spectrum.**

G.A. Miller Language and Communication, p. 96

Filter-bank analysis Dudley and Gruenz 1946

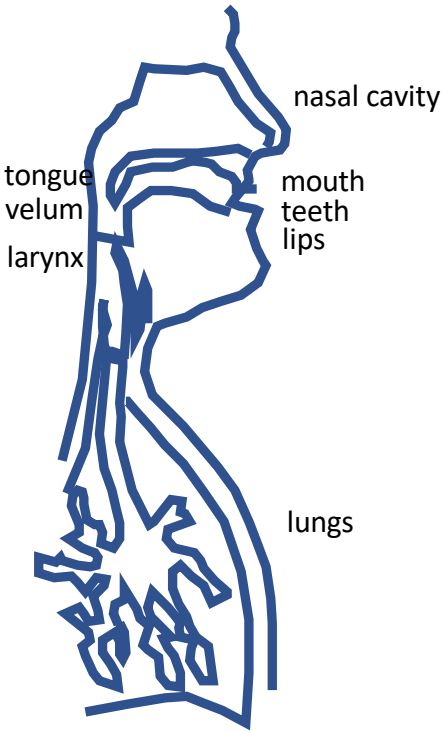




Temporal trajectories of frequency-localized spectral energies (spectral modulatons)?

Message is carried in **changes** in vocal tract shape, which modulate spectral components of speech

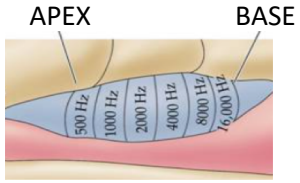
Dudley 1940



Redundant spread of information

- every change of the tract shape shows at all frequencies of speech spectrum
- tract shape changes do not happen very fast

brain



Medial geniculate body

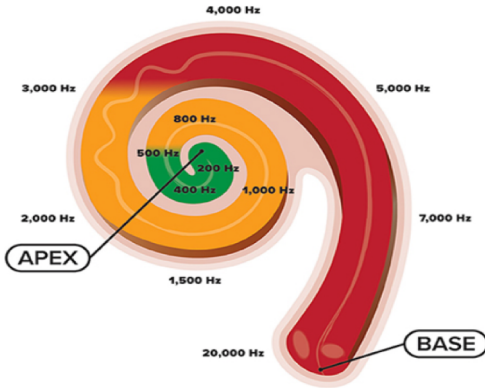
Inferior colliculus

Superior olive

Cochlear nucleus

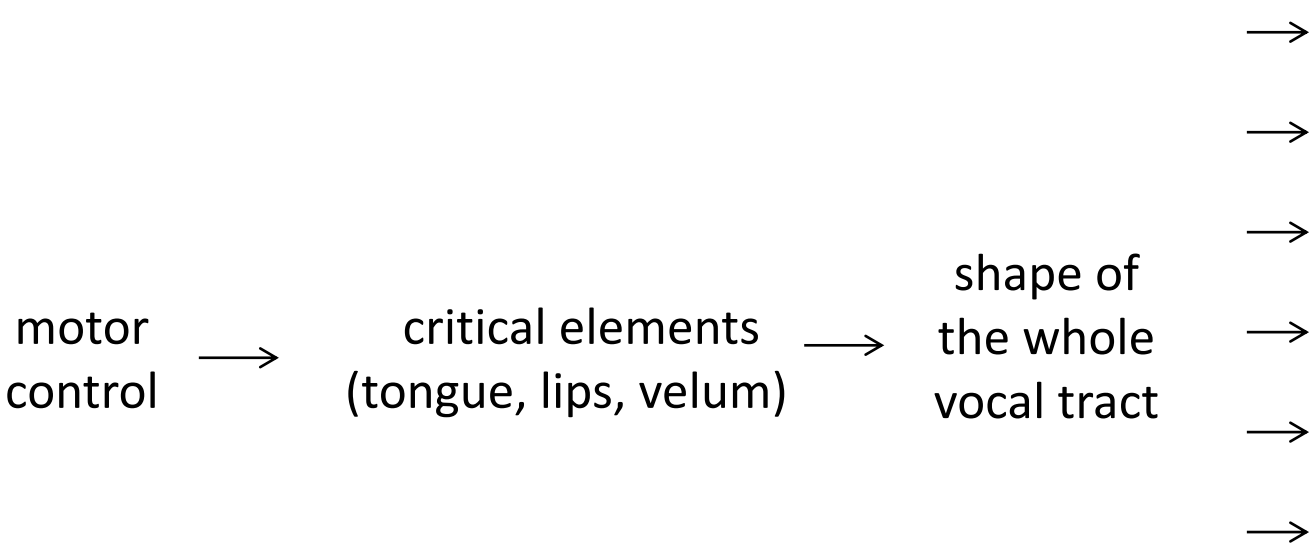
Auditory nerve

ear



- frequency selective (about 20 bands)
- sluggish (tenths of seconds)

INFORMATION ABOUT TRACT SHAPES DISTRIBUTED IN FREQUENCY



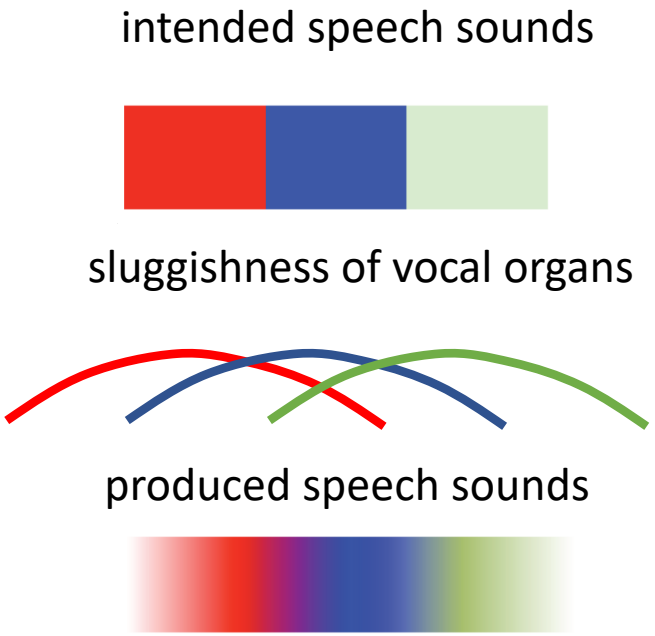
spectrum of speech signal

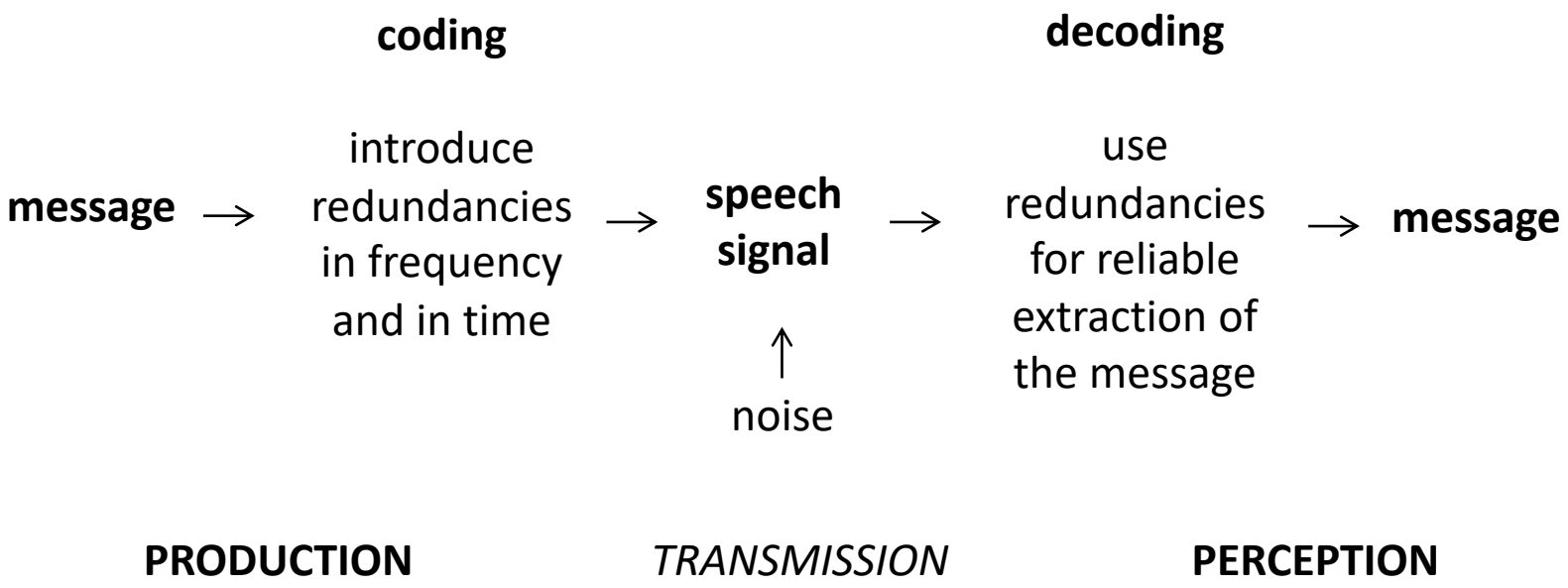
(redundant contributions of movements of critical elements in different frequency bands)

INFORMATION ABOUT TRACT SHAPES DISTRIBUTED IN TIME

from Sri Narajanan

movements of vocal organs are rather sluggish





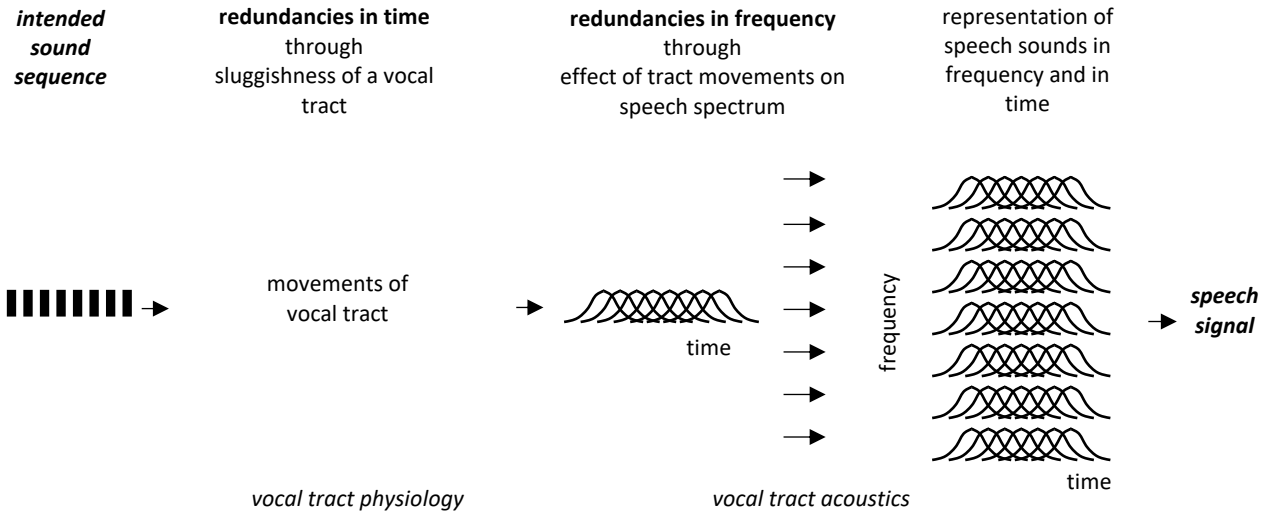
redundancy in frequency

production*:* tract acoustics distributes the information to all frequencies of the speech spectrum
perception: hearing selectivity allows for decoding the information in separate frequency bands

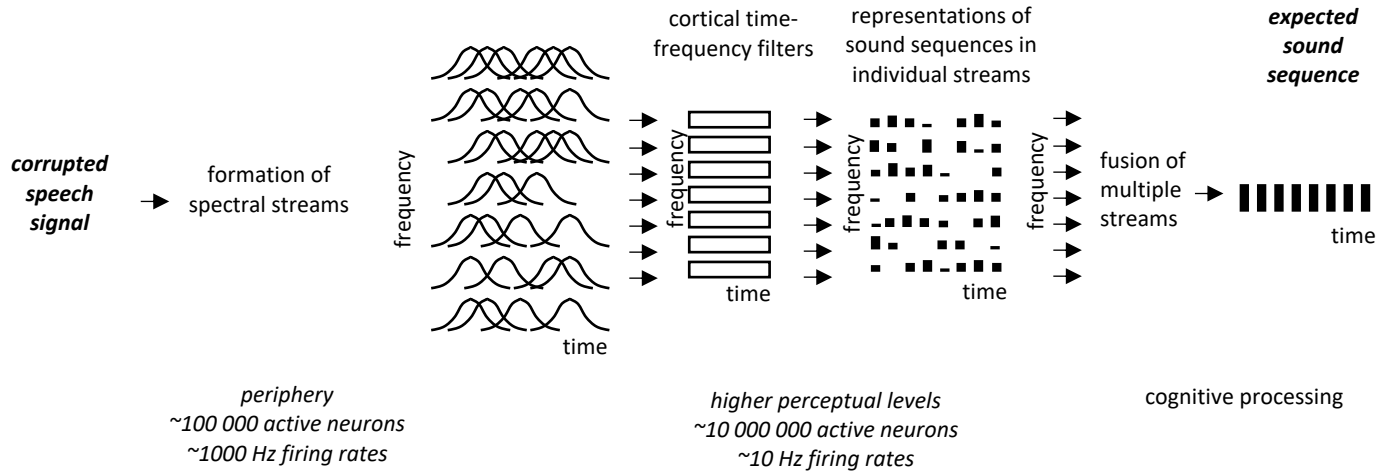
redundancy in time

production: tract sluggishness (coarticulation) distributes information about each speech sound in time
perception: temporal sluggishness of hearing collect the information distributed in time

PRODUCTION



PERCEPTION



Received 20 June 1969 9.10, 9.1

Whither Speech Recognition?

Letter to Editor
J.Acoust.Soc.Am.

J.R. PIERCE

Bell Telephone Laboratories, Inc., Murray Hill, New Jersey 07971

Speech recognition

Research field of “mad inventors or untrustworthy engineers”.

To succeed, machine needs intelligence and knowledge of language comparable to those of a native speaker.

- supervised the Bell Labs team which built the first transistor
- President’s Science Advisory Committee
- developed the concept of pulse code modulation
- designed and launched the first active communications satellite

Why to rock the boat ?

We have good thing going !

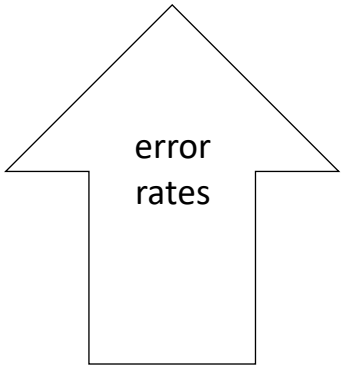
Are We There Yet ?

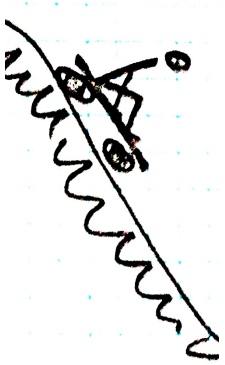
- Repetition, fillers, hesitations, interruptions, unfinished and non-grammatical sentences, new words, dialects, emotions, …
- Hands-free operation in noisy and reverberant environments,…

Alleviate need for large amounts of annotated training data

- Robustness to speech distortions, which do not seriously impact human speech communication
- Dealing with new unexpected lexical items
- Unsupervised learning/adaptation?

Why to rock the boat?
We have good thing going.





How to Get There ?

Fred Jelinek

Speech recognition
…a problem of maximum likelihood decoding
information and communication theory, machine learning, large data,….

Roman Jakobson

We speak, in order to be heard, in order to be understood
human communication, speech production, perception, neuroscience, cognitive science,..

Gordon Moore

The complexity for minimum component costs has increased at a rate of roughly a factor of two per year…

John Pierce

..devise a clear, simple, definitive experiments. So a science of speech can grow, certain step by certain step.

Signal processing, information theory, machine learning, …

&

neural information processing, psychophysics, physiology, cognitive science, phonetics and linguistics, ...

Engineering and Life Sciences together !

Hermansky Spring 2022

EN.520.680
Speech and Auditory Processing by Humans and Machines