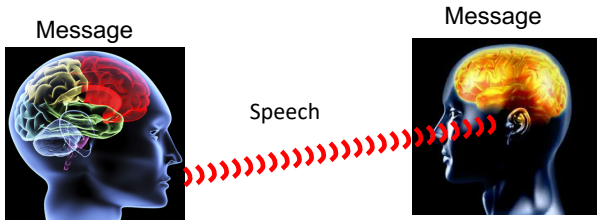
Human Speech

Hermansky Spring 2020

EN.520.680
Speech and Auditory Processing by Humans and Machines



Messages

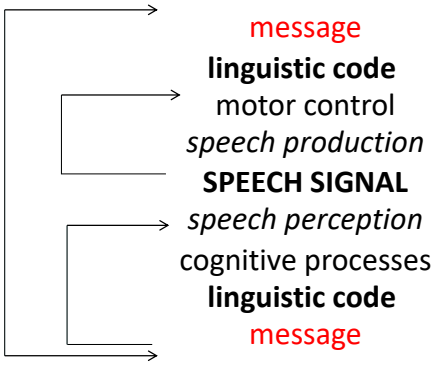
Problem

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

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.

Human Speech



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

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

p_i - probability of i-th symbol

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

Entropy : measure of information in the source

Entropy of the source

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

p_i - probability of i-th symbol

Property of the information source (alphabet)

Average amount of information per a symbol in the alphabet

26 letters in the English alphabet + one space = 27 symbols
entropy of the Enhlish alphabet when all symbols would be equally probable

H(s)= 1/27 $\log_2$(1/27)= 4.74 bit

how could English text look like if all letters were equally probable

xfoml rxklrjffjuj zlpwcfwkcyj ffjey

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%

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."

example of text generated when all letters are equally probable (zero order)

H(s)= 4.74 bit

xfoml rxklrjffjuj zlpwcfwkcyj ffjey

Respecting relative frequencies of letters
(first order)
H(s)= 4.279 bit

tocro hli rhwr nmielwis eu ll nbnes

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

H(s)= 2.77 bit

In no ist lat why cratict froure demonstures of the reptgain is

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	j	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 sounds 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

Consonants		
Phoneme	Graphemes	Examples
/p/	p pp	*pie pepper*
/b/	b bb	*bee ribbon*
/t/	t tt bt ght ed	*tie pretty doubt light stopped*
/d/	d dd de ld	*dig addroade should*
/k/	c k ck ch q cc que	*car scheme folk kite quiet occur cheque*
/g/	g gg gu gue gh	*garden eggg guess fatigue ghetto*
/s/	ps s ss c sc se ce	*psycho sit kiss lace scene worse peace*
/z/	z zz se s ze ss	*zoo jazz raise laser freeze scissors*
/ʃ/ (sh)	sh ti ch s ss c c+ion s+ion	*shop lotion charade sugar tissue ocean tension mention*
/ʒ/ (zh)	s g s+ion	*measure beige vision*
/f/	f ff ph lf fe	*food cliff photo half cafe*
/v/	v ve f	*vet glove of*

Vowels		
Phoneme	Graphemes	Examples
Short Vowel Sounds		
/æ/ (a)	a	*apple*
/ɛ/ (e)	e ea ai	*set head said*
/ɪ/ (i)	i y ui o u a	*sit lyric build women business pillage*
/ɒ/ (o)	o a ow au ough	*hot want knowledge station cough*
/ʌ/ (u)	u o-e o oo ou oe a	*hut done son flood country does pupil*
Long Vowel Sounds		
/eɪ/ (ae)	a a-e ai ay eigh ey	*lady lake sail hay neigh obey*
/i/ (ee)	ea ee ie ei y e ey i eo	*beach see piece receive many be alley key people*
/aɪ/ (ie)	i i-e ie igh uy ai ey y	*child divine tie sight buy aisle eye dry*
/əʊ/ (oa)	oe oa ough o-e ow o oa	*toe oat dough bone snow rogue loan*
/u/ (oo)	oo o-e oe ou o oo-e ew u-e wo ui u ue	*food move shoe you who mature never future two suit rude flue*

Consonants		
Phoneme	Graphemes	Examples
/θ/ (th)	th	*thorn*
/ð/ (th)	th the	*leather breathe*
/tʃ/ (ch)	ch tch c t+ure t+ion	*chain kitchen cello vulture question*
/dʒ/ (j)	g j ge dge gg d	*general jar sage ledge suggest*
/l/	l ll le il al el ul	*leaf pull isle soulful metal novel useful*
/r/	r rr wr rh	*row merry write rhino*
/m/	m mm mb me mn	*man mammal climb home autumn*
/n/	n nn kn ne pn gn en an	*nest runner knife done pneumatic align spoken human*
/ŋ/ (ng)	ng n	*sing sink*
/h/	h wh	*hello whole*
/w/	w wh u o	*well whale quiet once*
/y/	y u io	*yell using onion*

Long Vowel Sounds cont...		
/ɑ/ (ar)	ar a al au ear	*start sinatra path laugh heart*
/ɔ/ (or)	or a au oor ore oar aw ar ough augh	*for tall haunt floor tore board award warn bought naughty*
/ɜ/ (er)	er ir ur or ear our ere	*herb third turn word hearse journey were*
/ɪə/ (air)	air are ere eir ear ayor	*hair square there their pear mayor*
/eə/ (ear)	ear ere eer ier	*learn here deer tier*
Other Vowel Sounds		
/u/ (oo)	oo u ou	*look put should*
/aʊ/ (ou)	ou ow ough o-e	*mouse now bough towel*
/ɔɪ/ (oi)	oi oy	*join boy*
/ə/ (uh)	schwa vowel	

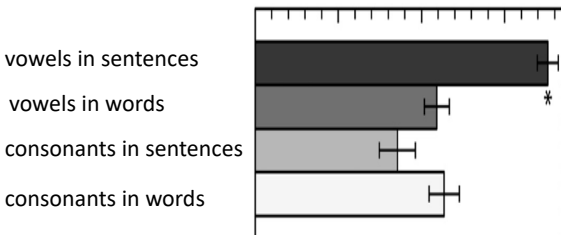
vowels – mouth open
consonants - mouth not so open

typical syllable

cvc
onset – nucleus – coda
cv
onset – nucleus

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

relative contribution



Forgety et al JASA 2012

BUT

The quick brown fox jumps over the lazy dog

Th qck brwn fx jmps vr th lzy dg

e ui o o y oe e a o

pronunciation dictionary

/prəˌnʌnsɪˈeɪʃ(ə)nˈdɪkʃən(ə)ri/

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%

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 ☺

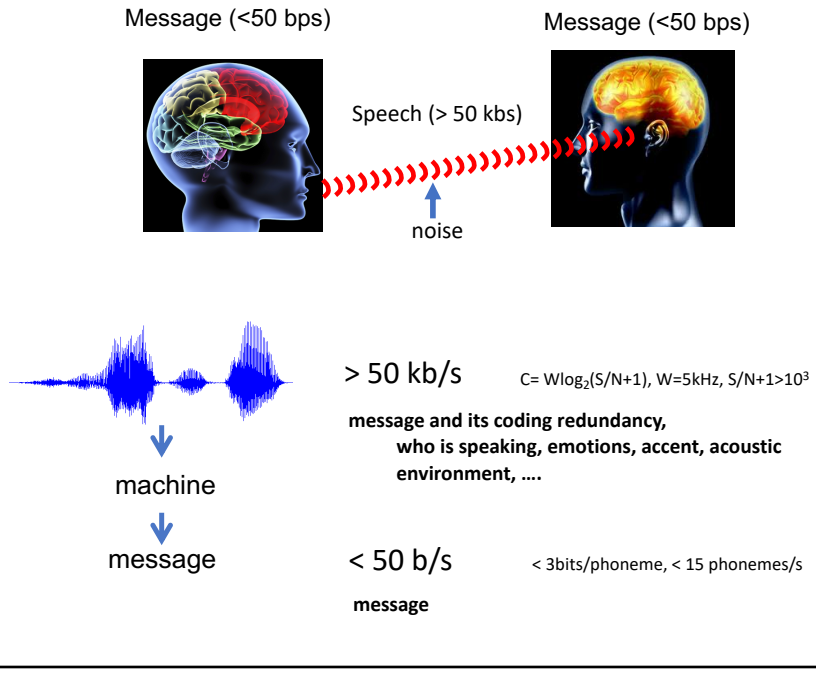
Predictability and unpredictability

- 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.

Variability

- Wanted variability:
  carries information about message, which we want to extract (signal)
- Unwanted variability:
  carries "other" information (**noise**)

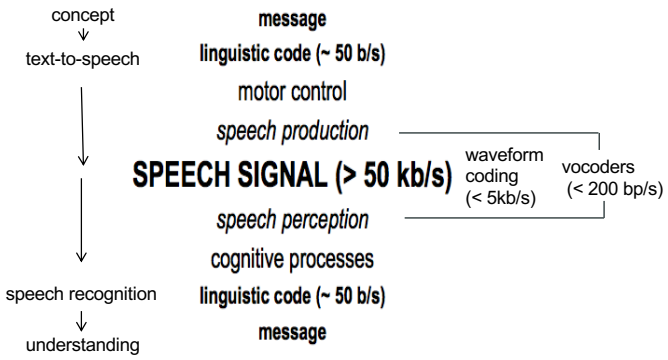


Noise: the good, the bad, and the ugly

- 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,…

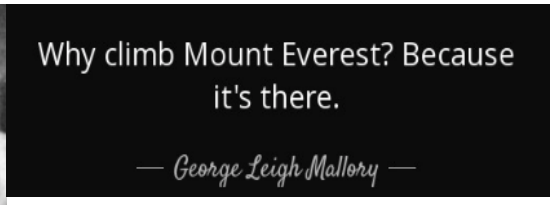
- We know this noise may come but its effect is not known
  - e.g., various environmental noises, reverberations, speaker peculiarities, language phonetics, accents, ….
    - multistyle training,…

- 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, …)



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



Most people think the famous climbing phrase "because it is there" was first uttered by Edmund Hillary when he and Tenzing Norgay conquered Mount Everest in 1953. Not so. Actually George Leigh Mallory, three decades earlier, said it as he prepared to scale the world's highest peak.

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

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

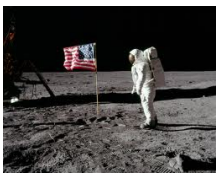
Speech recognition
Research field of “mad inventors or untrustworthy engineers”.

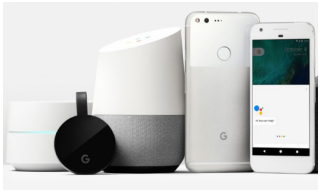
To succeed, machine needs intelligence and knowledge of language comparable to those of a native speaker.

- 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

John Pierce

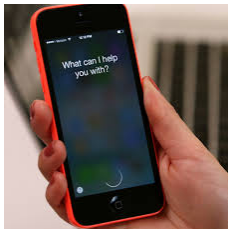
To succeed, machine needs intelligence and knowledge of language comparable to those of a native speaker.





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

Are We There Yet ?

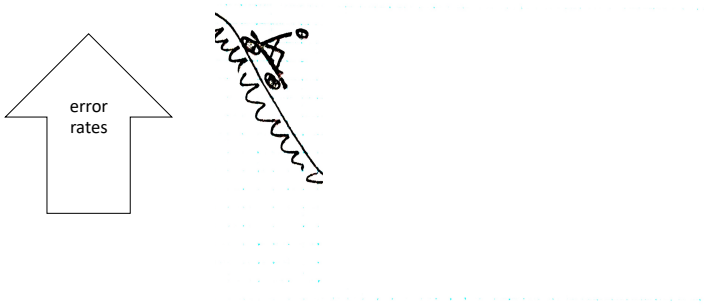


- Repetition, fillers, hesitations, interruptions, unfinished and non-grammatical sentences, new words, dialects, emotions, …
- Hands-free operation in noisy and reverberant environments,…

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 2020

EN.520.680
Speech and Auditory Processing by Humans and Machines