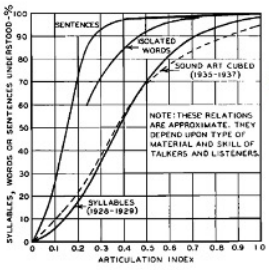
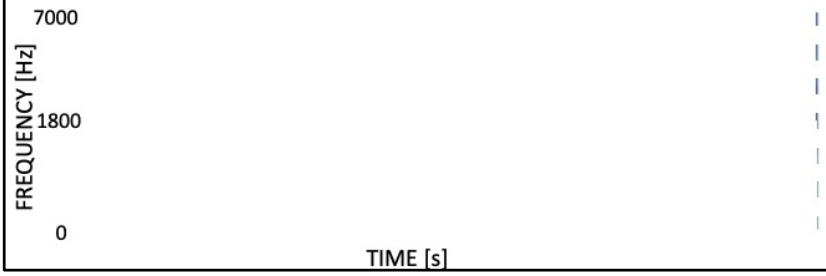
Articulatory Bands

French and Steinberg 1949

250-375-505-654-795-995-1130-1315-1515-1720-1930-2140-2355-2600-2900-3255-3680-4200-4860-5720-7000 Hz

- 20 frequency bands in speech spectral region
- each band with SNR > 30 dBcontributes equally to human speech recognition
- bands with SNR < 0 dB do not contriute at all
- any 10 bands sufficient for 70% correct recognition of nonsense syllables, better than 95% correct recognition of meaningful sentences [Fletcher and Steinberg 1929]





This turned out to be true not only for two bands, but also for more bands (up to 20 bands). Thus, the multichannel model predicts that the total error will be given by

$$e = \prod_{i=1}^{K} e_i$$

- SERIAL PROCESSING
  - phonemes in a nonsense syllable are decoded independently of each other S=c.v.c (**probabilities of correct recognition multiply**)
- PARALLEL PROCESSING
  - errors in phonetic judgment in nonsense syllables in individual sub-bands are independent (**probabilities of errors multiply**)

Articulation Index

Bell Labs 1921-1950
P(message) weak - no words, only speech sounds with their prior probabilities

Recognition of nonsense CV and CVC syllables under filtering and noise
Nonsense CVC syllables: mav-pok-seng-*run*-kiz-***veb***……

66 syllables. 34 correctly recognized
34/66=0.515

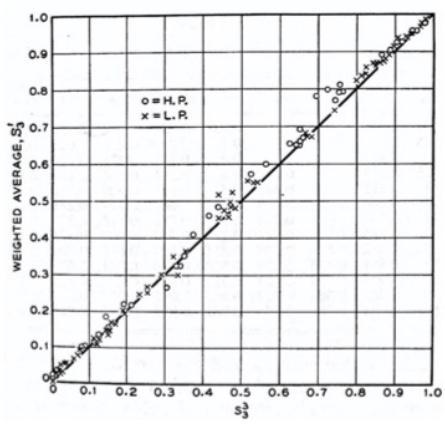
198 phonemes, 157 correctly recognized
s = 157/198=0.793 chance to correctly recognize phoneme

0.909 correct vowels, 0.735 correct consonants
cvc = 0.753 x 0.909 x 0.753 = 0.491

When phonemes recognized independently, probability of correctly recognizing 3 phoneme syllable would be

$S=s^3 = 0.793^3 = 0.499$

0.499 close to 0.515
Phonemes recognized independently of each other !



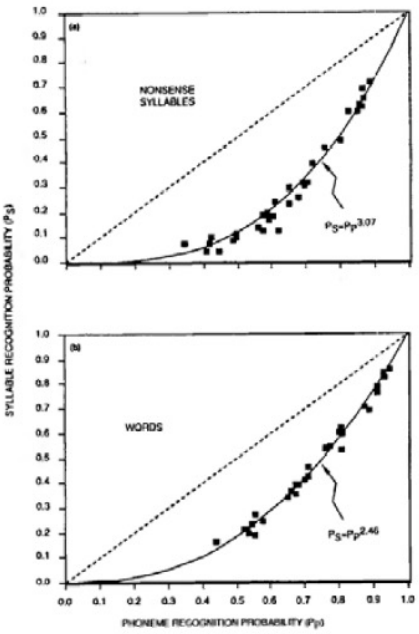
This result replicated later by others

Boothroyd, A., & Nittrouer, S. (1988). Mathematical treatment of context effects in phoneme and word recognition. *The Journal of the Acoustical Society of America*, *84*(1), 101-114

• Probability of correct identification of CVC $S(\alpha)$ for various values of distortion α is given by a **product** of probabilities of identification of individual phonemes in the CVC

$$S(\alpha) = cvc$$

That implies the individual phonemes in the syllable are decoded **independently** of each other (i.e. the coarticulation is perceptually compensated for!)



to recognize nonsense syllable, all three phonemes need to be recognized independently (product of three probabilities of phonemes)

$p_s=p_p^{3.07}$

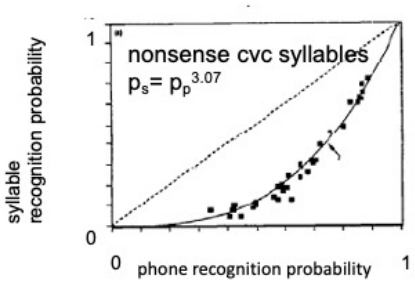
prior experience, expectations, language model
to recognize meaningful syllable, in average only about 2.5 phoneme need to be recognized

$p_s=p_p^{2.46}$

Boothroyd, A., & Nittrouer, S. (1988). Mathematical treatment of context effects in phoneme and word recognition. *The Journal of the Acoustical Society of America*, *84*(1), 101-114

human recognition (serial model)
(Boothroyd and Nittrouer 1988)

- **whole** (e.g. CVC syllable) consists of **parts** (C and V phonemes)
- to recognize nonsense syllable, one needs to recognize all phonemes
- $p_S = p_p^N$ N – number of phonemes in the word

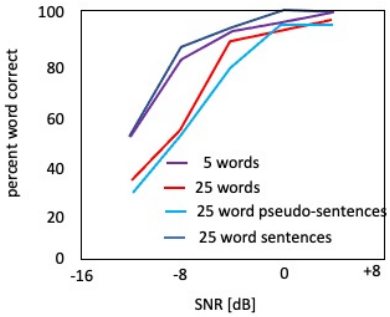


- to recognize meaningful word from the closed set of words, one often does not have to recognize all phonemes
- $p_W = p_p^M$ M < N

- Probability given by a product of probabilities from individual processes is necessary condition for the stochastic independence of` the processes involved.
- In human recognition of nonsense syllables, phonemes are decoded independently of each other, i.e. the coarticulation is perceptually compensated for !

Effect of number of words to be recognized

Human recognition of monosyllabic words from closed vocabulary in noise (Miller 1962)



Things has no wet Don

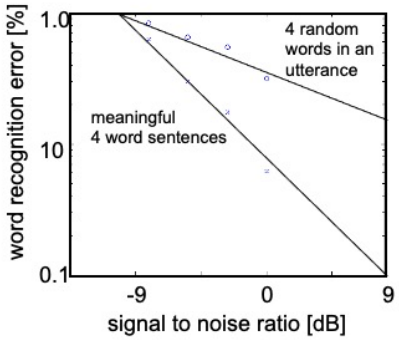
Don has no wet things

It is more difficult to recognize words from larger vocabularies

It is easier to recognize words in real sentences (expectations, language model,…)

How do Human Listeners Recognize Words in Context?

Boothroyd and Nittrouer 1988



Confirmation of Miller 1962

- 20 sentences
- 40 pseudo-sentences

5 adult listeners

$wer_{context} = wer_{no\ context}^{k}$

$k > 1$ ($k \cong 2.7$)

$wer_{context} = wer_{no\ context} \cdot \mathbf{wer}_{\mathbf{context\ channel}}^{k-1}$

*The assumption is made that the effect of contexts is quantitatively equivalent to **adding statistically independent channels of sensory data** to those already available from the speech units themselves.*

Boothroyd and Nittrouer 1988

PARALLEL

Fletcher (Allen)

$$P(\varepsilon) = \prod P(\varepsilon_i)$$

Boothroyd (Allen)

$$P(\varepsilon_{\text{incontext}}) = P(\varepsilon_{\text{no context}})P(\varepsilon_{\text{context channel}})$$

Low error probability in any channel makes the final error low

SERIAL

Bayes rule (current ASR)

$$\hat{W} = \text{argmax}_W\{p(X|W)P(W)\}$$

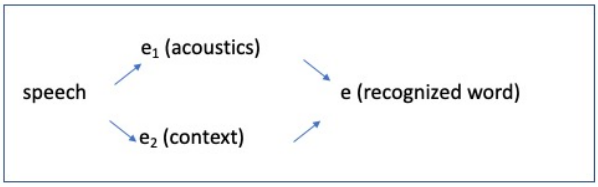
Both the likelihood *p(X|W)* and the prior probability *P(W)* needs to be high

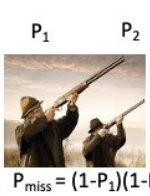
back

Probability summation (independent decision) model
(e – probability of error, p = probability of correct response)

$e=e_1e_2 = (1-p_1)\ (1-p_2)= 1-p_1 - p_2 + p_1p_2$
$p = p_1 + p_2 - p_1p_2$

applies to two independent information channels, where either channel can yield the correct answer





$P_{miss} = (1-P_1)(1-P_2)$

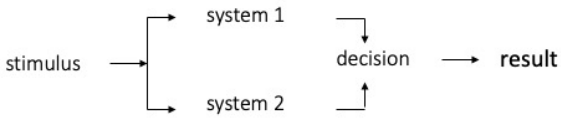
the correct answer must be recognized as being correct
for the model to hold

Serial systems (AND)

stimulus ⟶system 1 ⟶ system 2 ⟶result

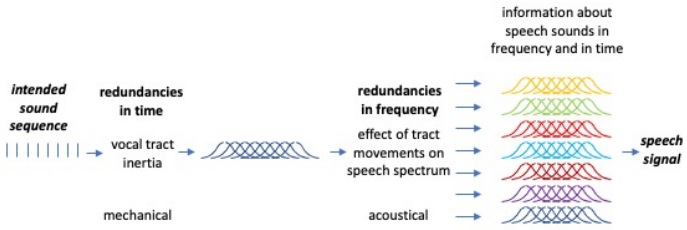
- Both systems need to be correct for the result to be correct

Parallel systems (OR)

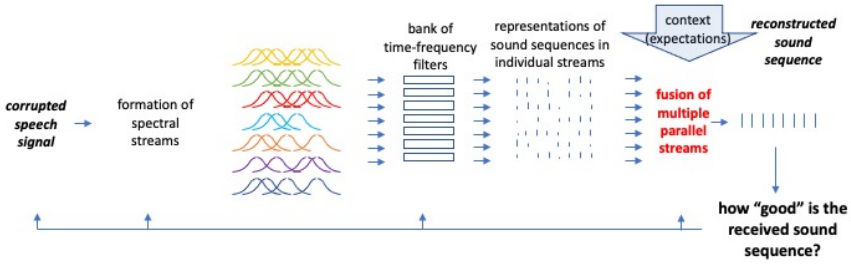


- Only one system needs to be correct for the result to be correct, but **decision box needs to recognize that the answer is correct**

Redundanies introduced in speech production



Proposed system for multistream speech recogntion



Parallel combination of bottom-up multiple acoustic streams and top-down context (languaege) information

Watt Regulator

- negative **feedback** to adjust for optimal speed of the steam engine

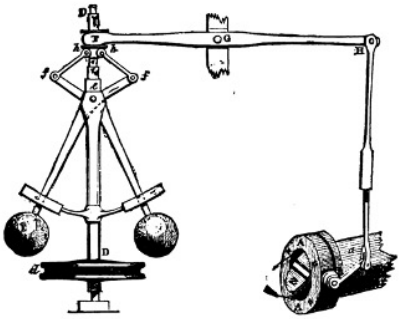
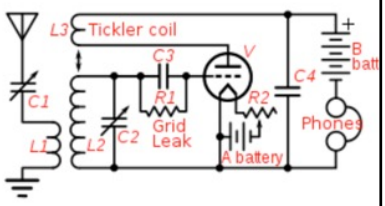
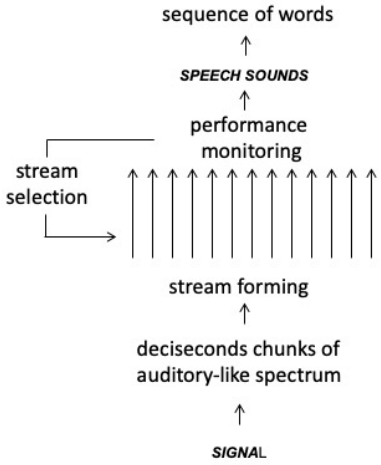


FIG. 4.—*Governor and Throttle-Valve.*

Armstrong circuit

- positive feedback to sharpen frequency response of a resonant circuit





streams

- enhancing different parts of speech spectrum
- enhancing different spectral and temporal modulations
- different sensors (micriphones)
- different modalities
- different strengths of prior assumptions (language model)

performance monitoring

- **estimate quality of information without knowing what is the information**

Competence Monitoring

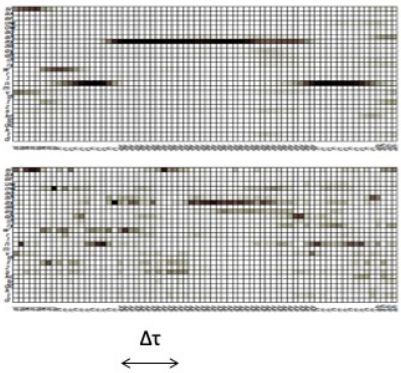
1) knowing how the classifier output should look like

   some prior knowledge about the character of the output is available

2) learing how the outut should look like

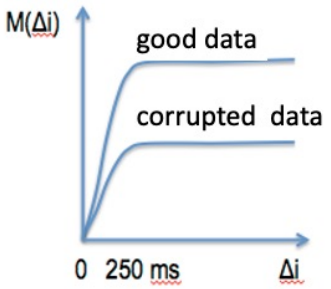
   - classifier works the best when it is applied to its training data
     - derive output model (statistics) while the trained classifier is used with its training data
     - how different is the classier output when used with new test data ?
       - fit the output on new data to the model (compare statistics) from the training data

Representations (posteriors) from different parts of the signal should be different



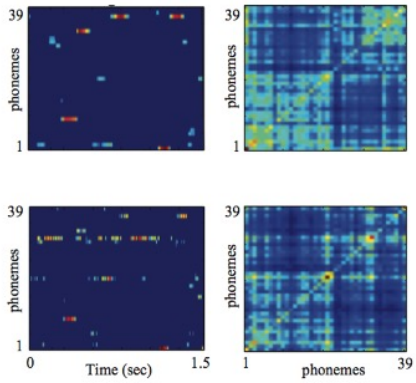
$$M(\Delta\tau) = \frac{\sum_{i=0}^{N-\Delta\tau} D(\boldsymbol{p}_i, \boldsymbol{p}_{i+\Delta\tau})}{N - \Delta\tau}$$

Δi – time delay
D(.) – symmetric Kl divergence



How “similar” are the estimator outputs on its training data and in the test?

Mesgarani et al, JASA Acoustic Letters 2011



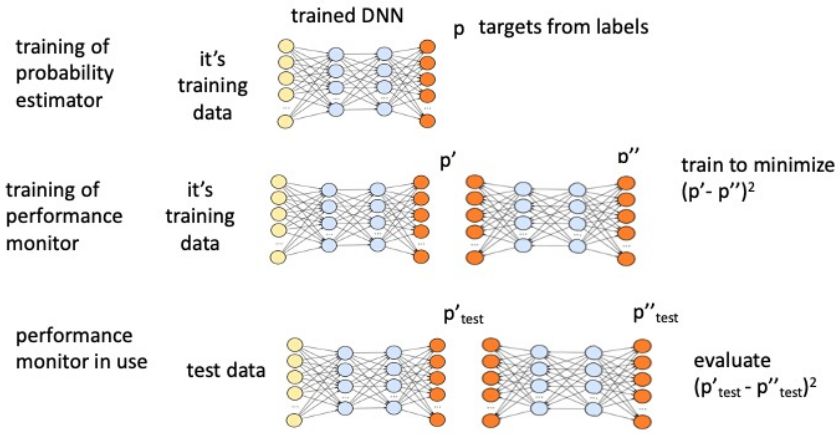
$$AC = \frac{1}{N}\sum_{i=1}^{N}\mathbf{P}(i)\mathbf{P}(i)^{T},$$

where $\mathbf{P}(i)$ – posterior probability vector at time i,

N - length of the data to be described

Compare matrices derived on training data of the DNN and in the test.

DNN auto-encoder, trained on output of the estimator when applied to its training data



Knowing that the result is no good

- warn when the results is unreliable
- additional training when necessary
  - reinforcement learning
- select appropriate processing among the available ones (on-line adaptation)
  - e.g., select among many processing streams in the multi-stream paradign
- semi-superwised training (which machine labeled data to use?)

Received 20 June 1969

9.10, 9.1

Whither Speech Recognition?

Letter to Editor
J.Acoust.Soc.Am.

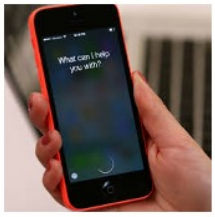
J.R. PIERCE

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

Implement.... ***intelligence and knowledge of language comparable to those of a native speaker !***

.... should people continue work towards speech recognition by machine ? Perhaps it is for people in the field to decide.

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,…

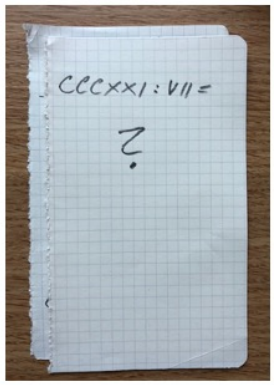
Alleviate need for large amounts of annotated training data

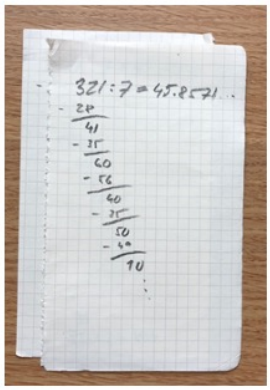
- Robustness to speech distortions, which do not seriously impact human speech communication
- Unsupervised learning/adaptation?
- Dealing with new unexpected lexical items

Why is speech recogntion hard?

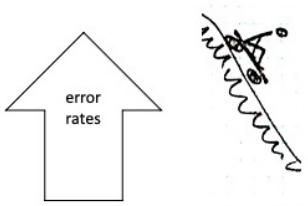
languages are different, speaker knowledge is different, people are different, acoustsic environments are different, sounds are coarticulated,....

Why is long division hard?

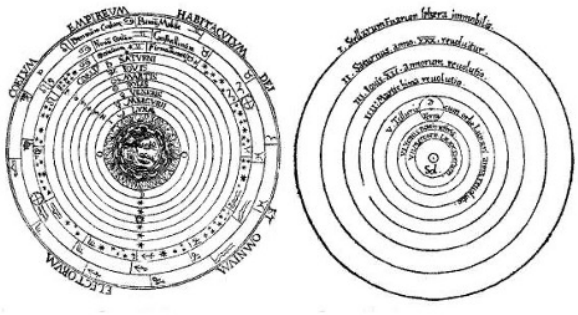




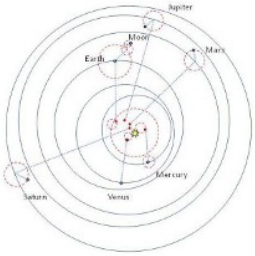
Why to rock the boat?
We have good thing going.

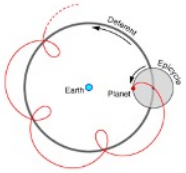


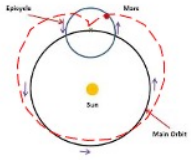
Geocentric (Ptolemy) Heliocentric (Copernicus)



Geocentric (Ptolemy) Heliocentric (Copernicus)

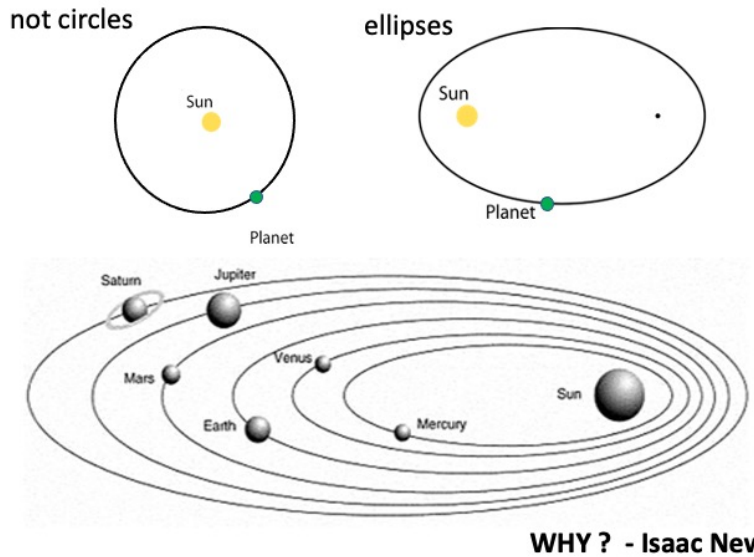




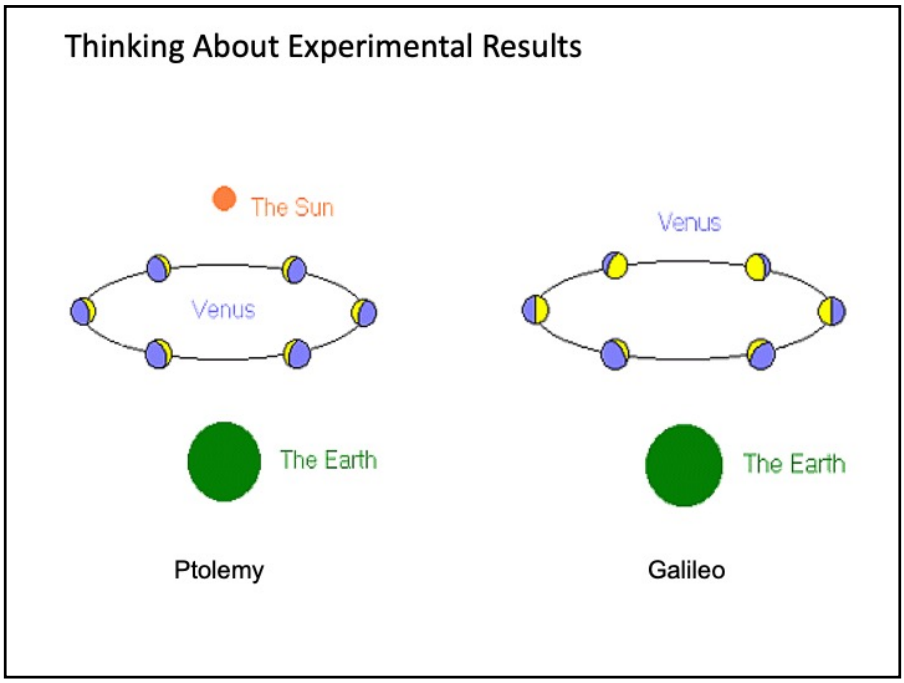


40 epicycles 48 epicycles !!

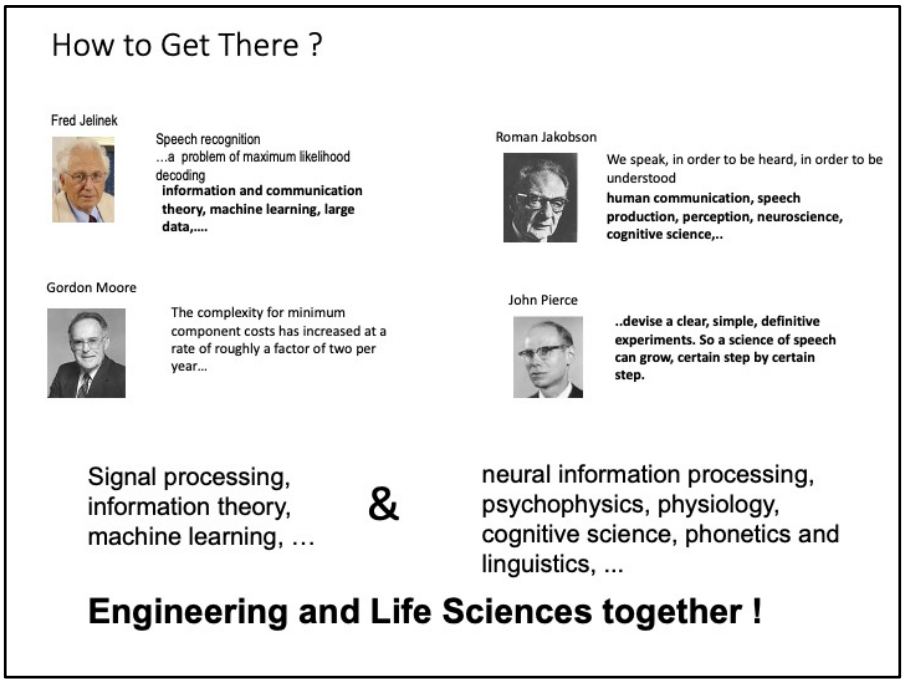
Kepler



WHY ? - Isaac Newton



It is better to se large part of an animal to decide what it is.
Spectrograph - meaning of spectral slices?
(Also linear frequency scale because of heterodyne)
Coarticulation "problem"?
Simultaneous masking - remote spectral components do not influence detection in critical band.
Spectral envelope must be wrong!
Felt like Galileo. Had data which do not fit the concept!



Speech processing – dominated by data (Fred-stochastic approaches).
Strive for more permanent knowledge (oldfashioned speech science – can be derived from data).
Get help from powerful tools (GPUs)
Remember John Pierce – step-by-step incremental knowledge acquisition (hypothesis does not always be – things will get better)

Why not to strive for emulation of knowledge of a native speaker?
Impossible may become possible.