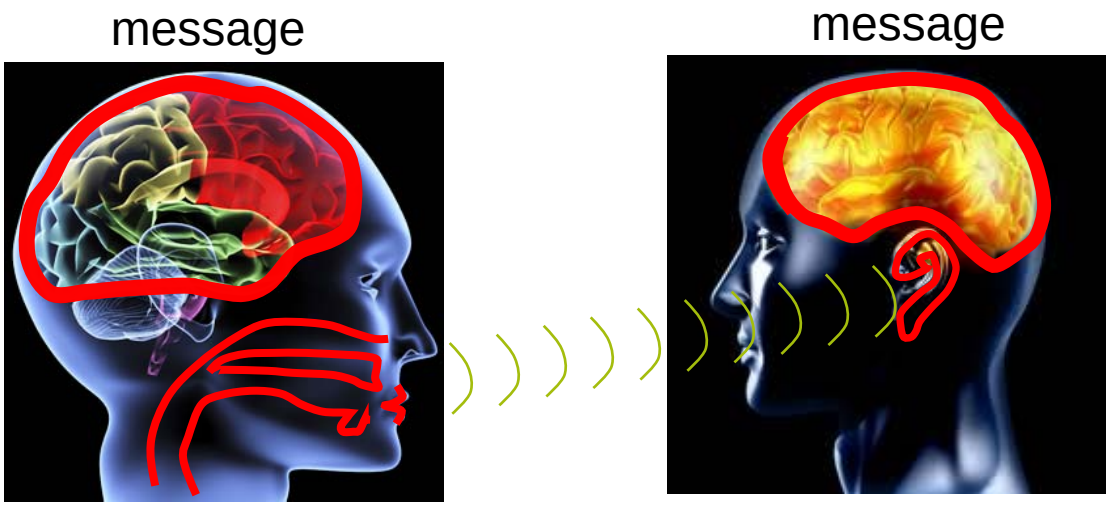
The Center For Language and Speech Processing
at the Johns Hopkins University

Human hearing and speech technology

Hynek Hermansky



Spoken language

- Limited number of **speech sounds** (phonemes) forming larger groups (words, phrases,..)
- Information about the speech sounds is coded in the signal
- Hearing must decode this code

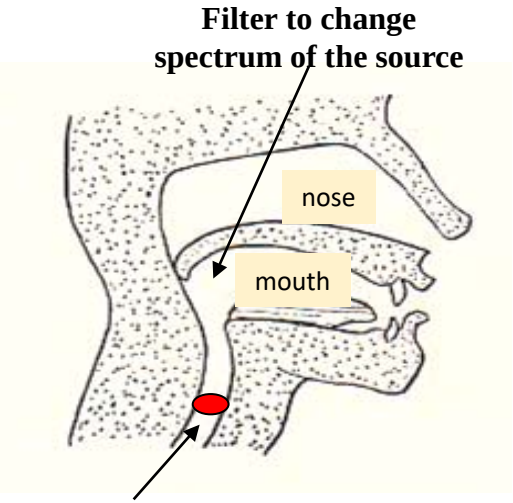
We speak, in order to be heard, and need to be heard in order to be understood

(Roman Jacobson 1979)

SPEECH

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

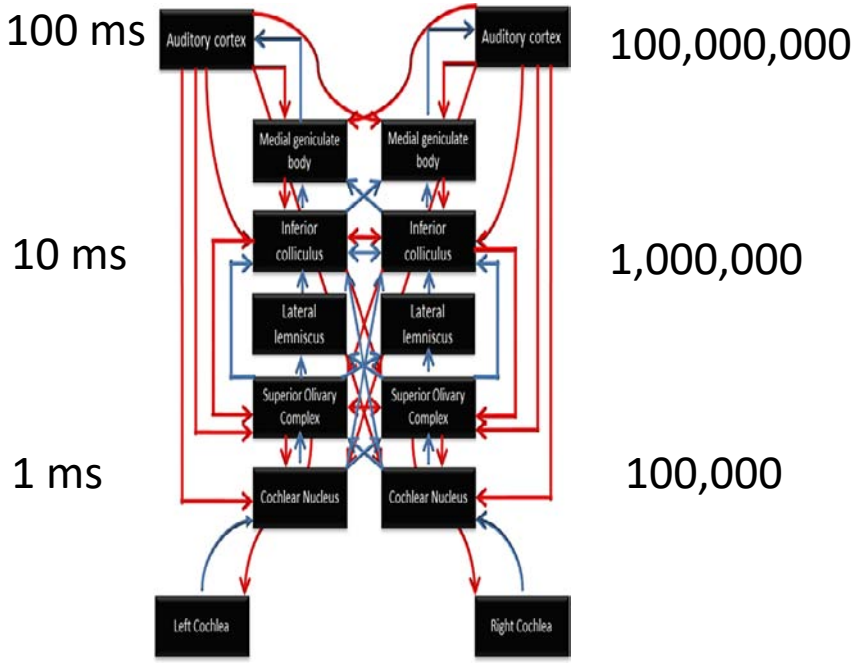
Homer Dudley 1940



HEARING

inter-spike interval

number of spiking neurons



Speech & Hearing

We hear to survive

- It has long been assumed that sensory neurons are adapted, through both evolutionary and developmental processes, to the statistical properties of the signals to which they are exposed.
    - Simoncelli and Olshausen 2001

……but 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 Communication

message

linguistic code

motor control

speech production

SPEECH SIGNAL

speech perception

cognitive processes

linguistic code

message

Information in speech signal

$C = W \log_2 [(S+N)/N)]$,
W-signal bandwidth,
S-power of signal, N-power of noise

W – about 8 000 Hz
(S+N)/N - about 10^3
$\log_2 1000$ – about 10

about 80 kb/s

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

p_i - probability of i-th symbol

41 phonemes in English
$H=\log_2 41 = 5.4$ bit/phoneme
about 15 phonemes/s

about 80 b/s

Human Speech Communication

message

linguistic code (< 80 b/s)

motor control

speech production

SPEECH SIGNAL (> 80 kb/s)

speech perception

cognitive processes

linguistic code (< 80 b/s)

message

Speech Engineering

concept
(translation)

↓

text-to-speech

↓

↓

speech recognition

↓

understanding
(translation)

message

linguistic code (< 80 b/s)

motor control

speech production

SPEECH SIGNAL (> 80 kb/s)

speech perception

cognitive processes

linguistic code (< 80 b/s)

message

waveform coding (< 5kb/s)

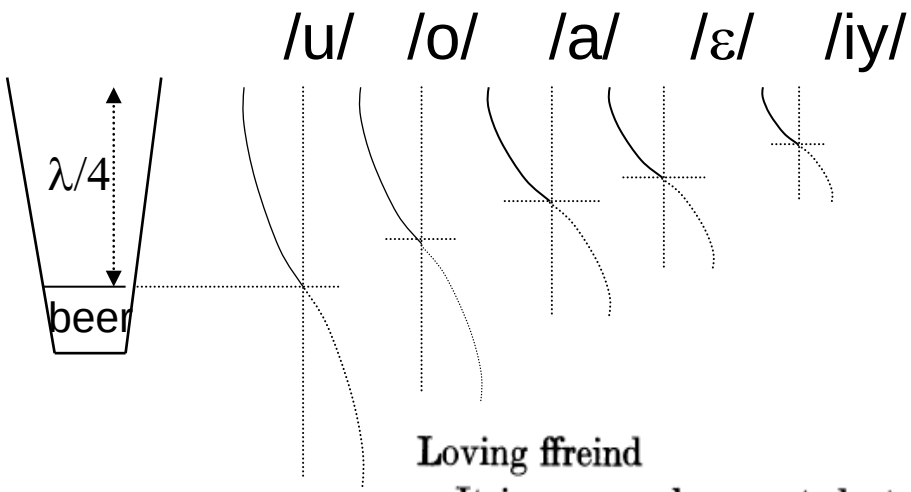
vocoders (< 200 bp/s)

How does hearing get the information about speech sounds ?

Isaac Newton

The filling of a very deepe flaggon wth a constant streame of beere or water sounds y^{e} vowells in this order w, u, ω, o, a, e, i, y.

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



/u/ f1 = 300 Hz,

λ/4 = 25 CM

/iy/ f2’ = 3000 Hz

λ/4= 2.5 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.

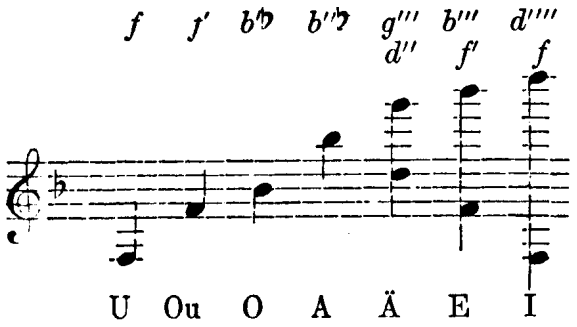
Speech experiment of Newton (around 1640)

Hearing can extract information about frequency composition of sounds

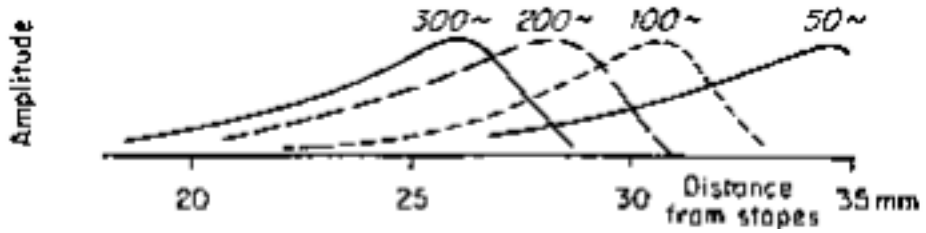
Ohm 1843: Ear separates acoustic signal into a series of sinusoidal signals

von Helmholtz 1863 Theory of Hearing

pitch a human voice is
determined by the places
where the membrane vibrates,

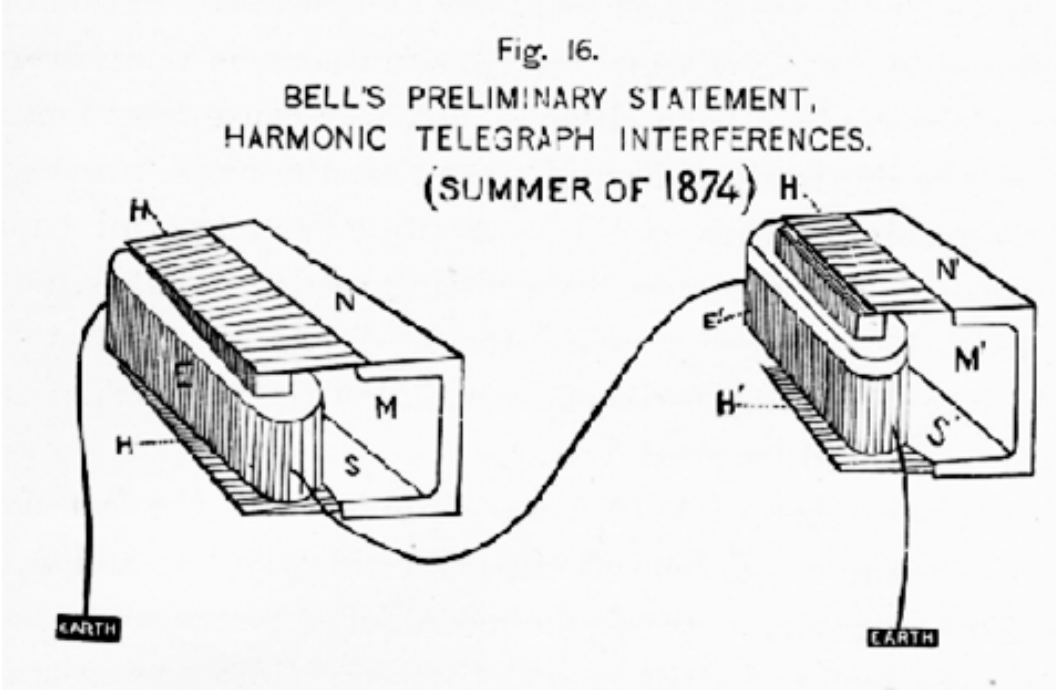


von Bekesy 1960



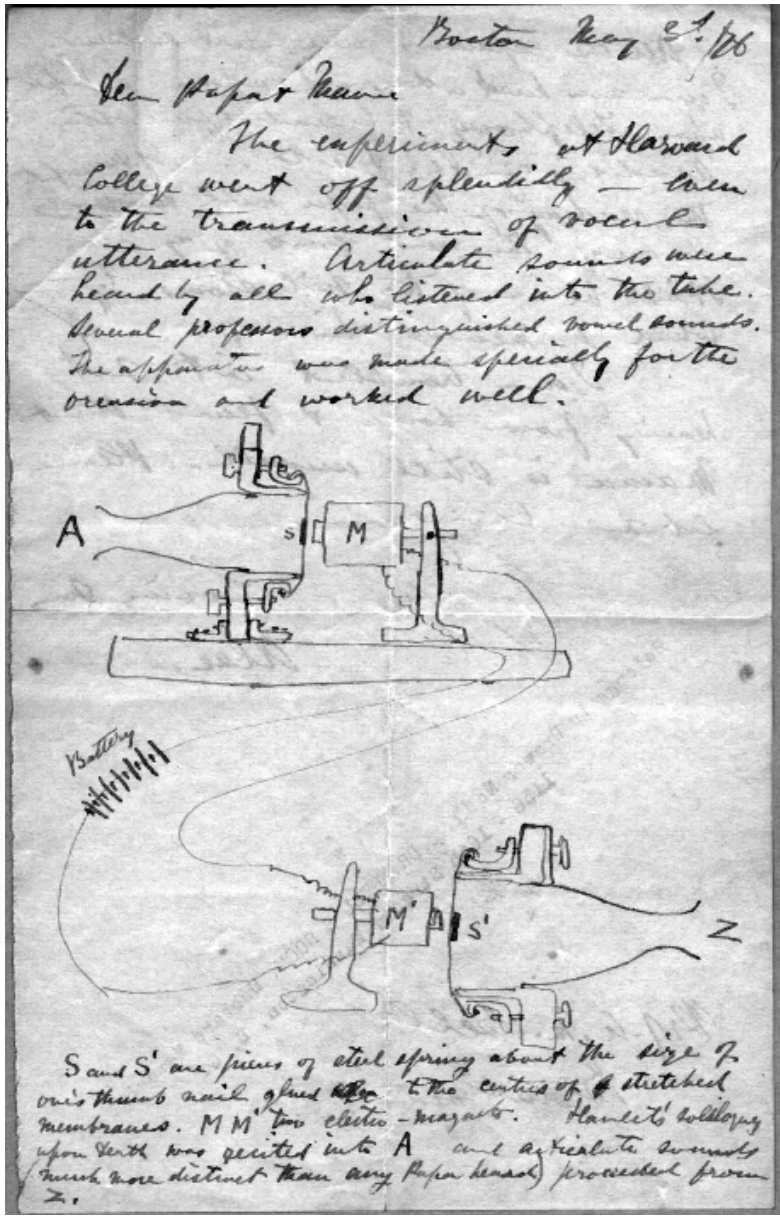
Alexander Graham Bell

emulate spectral analysis by hearing

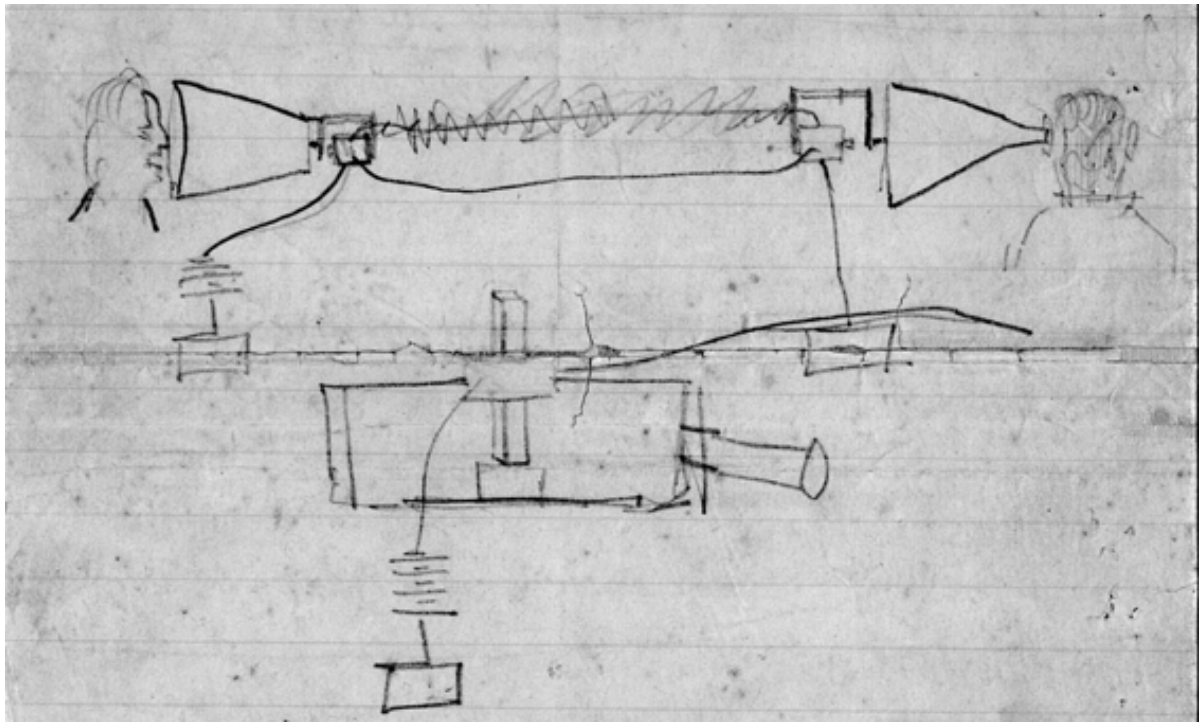


Had proven that different tones would vary the strength of an electric current in a wire.

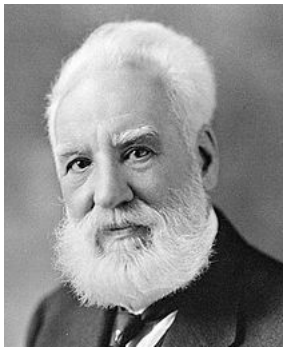
Until one of the reeds got stuck and reacted to all frequencies…



Back to speech signal



Alexander Graham Bell



Theodore Newton Vail

How much of information can be stripped from the speech signal (and still charge for the service) ?

Western Electric
Bell Telephone Laboratories

Harvey Fletcher and his colleagues
(1916-1949)

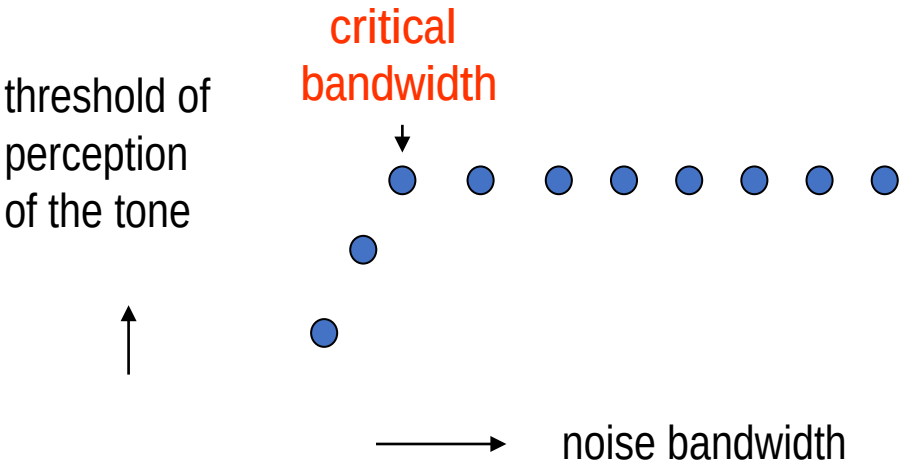
Arnold, Biddulph, Crandall, Dunn, French, Fry, Galt, Gardner, Graham, Hartley, Kingsbury, Koenig, Kranz, Lane, MacKenzie, Munson, Riesz, Sacia, Shower, Sivian, Steinberg, Stewart, Wegel, and Wente.

Speech intelligibility

Threshold of hearing, iso-loudness curves, power law of loudness, perception of modulations, pitch perception, critical bands of hearing, articulatory bands, …

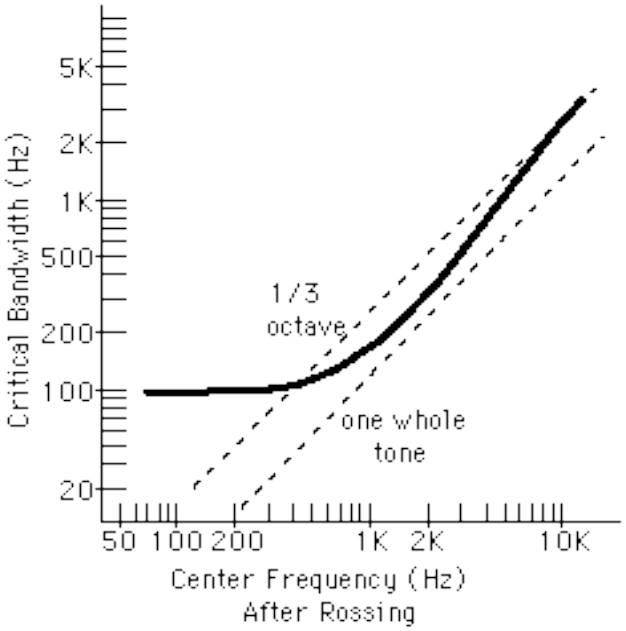
Simultaneous Masking (critical bandwidth)

(Fletcher 1934)



critical bandwidth increases with frequency

- non-equal resolution of hearing (Bark scale, mel scale,..)



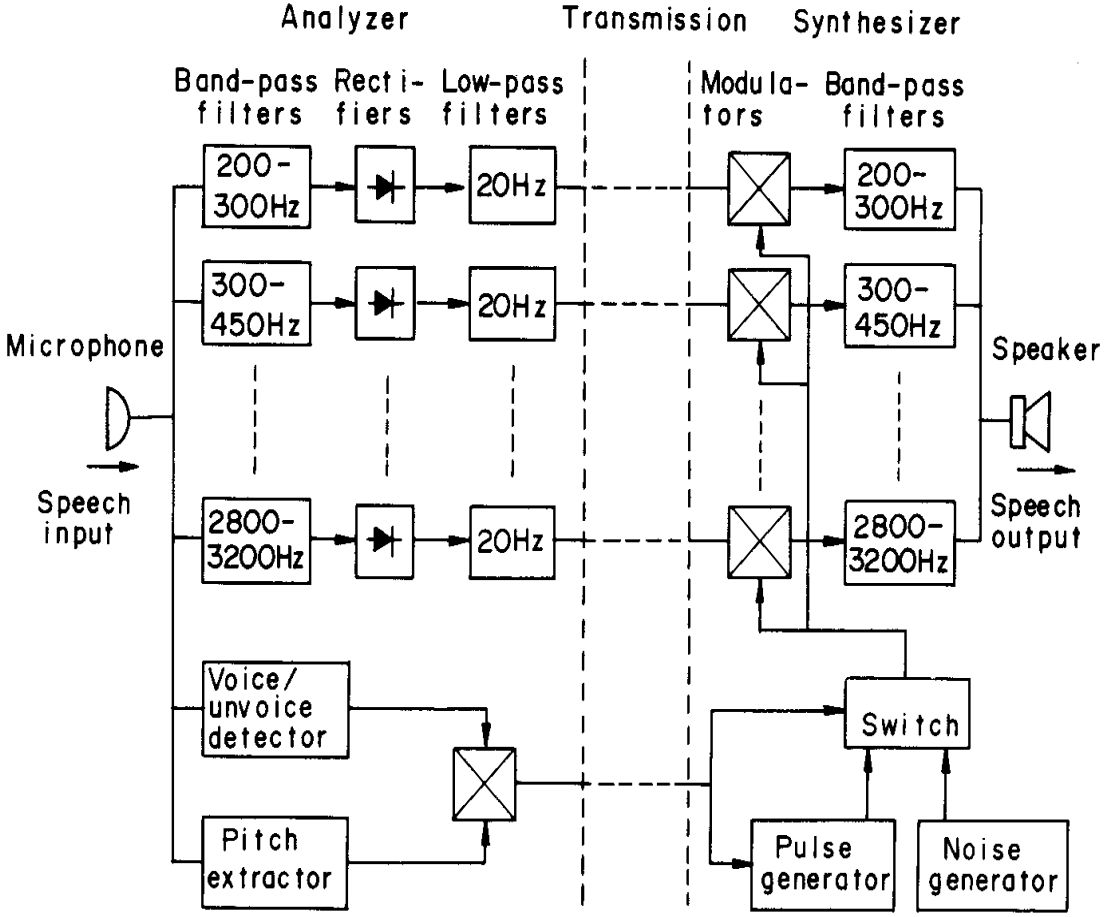
Articulatory Bands

French and Steinberg 1949

- 20 frequency streams in speech spectral region
- each stream with SNR> 30dB contributes about equally to human speech recognition
- streams with SNRs < 0dB do not contribute
- any 10 streams sufficient for 70% correct recognition of nonsense syllables, and better than 95% correct recognition of meaningful sentences [Fletcher and Steinberg 1929]

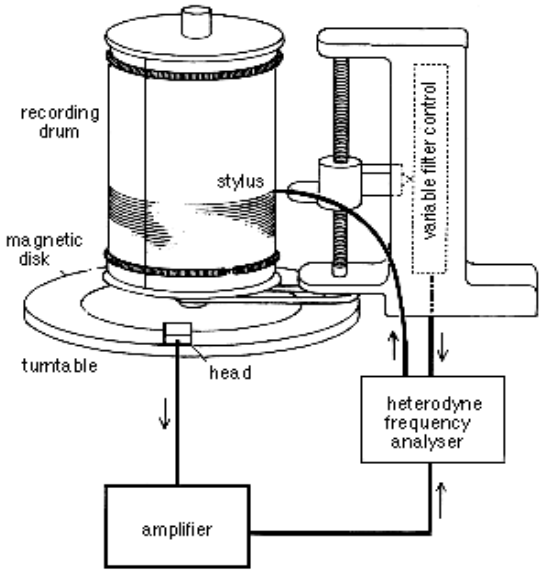
250-375-505-654-795-995-1130-1315-1515-1720-1930-2140-2355-2600-2900-3255-3680-4200-4860-5720-7000 Hz

VOCODER (Homer Dudley 1939)



- Predictability (production)
  - speech waveform changes "slowly" (inertia of air mass in vocal tract cavities)
  - spectral envelope changes slowly
    - 20 Hz low-pass
  - voiced speech is periodic
    - pulse generator for excitation
- Hearing properties (perception)
  - spectral resolution of hearing
    - wider band-pass filters at higher frequencies

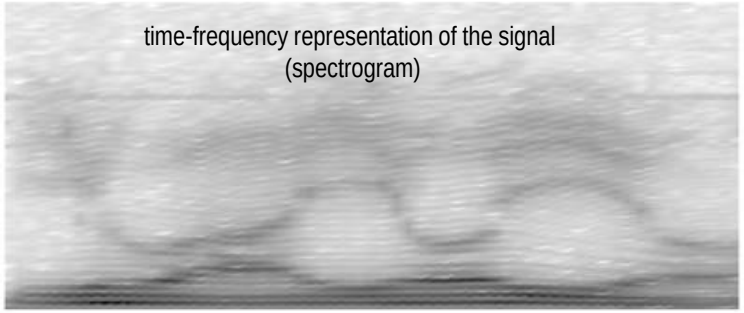
Spectrograph ™



developed by Bell Labs engineers
to visualize sounds

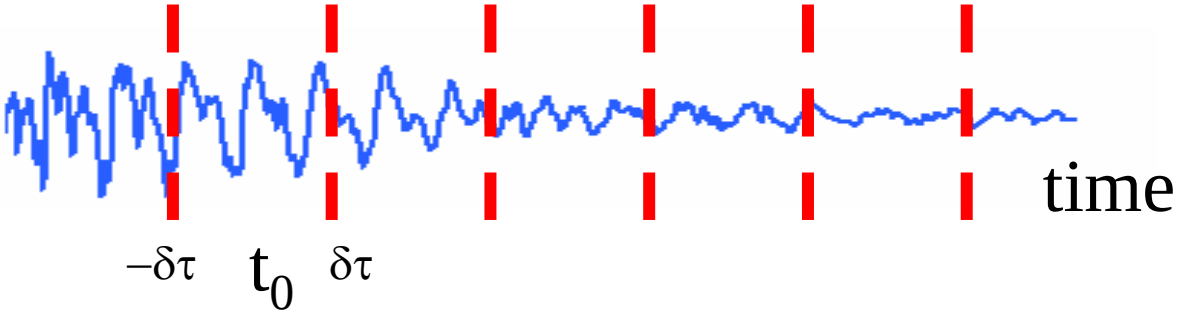
heterodyne frequency analyzer
(constant-bandwidth frequency resolution)

/j/ /u/ /a$_r$/ /j/ /o/ /j/ /o/

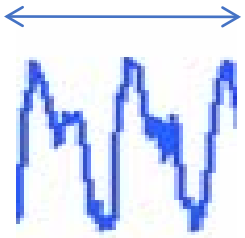


original Spectrograph
temporal evolution of spectral energy $P(\omega_0,t)$

non-stationary speech signal s(t)

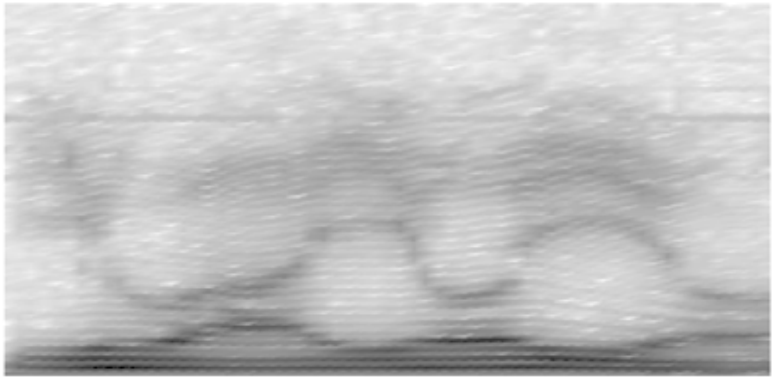


$\Delta t = 2\delta\tau$



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

time-frequency representation of the signal
(spectrogram)

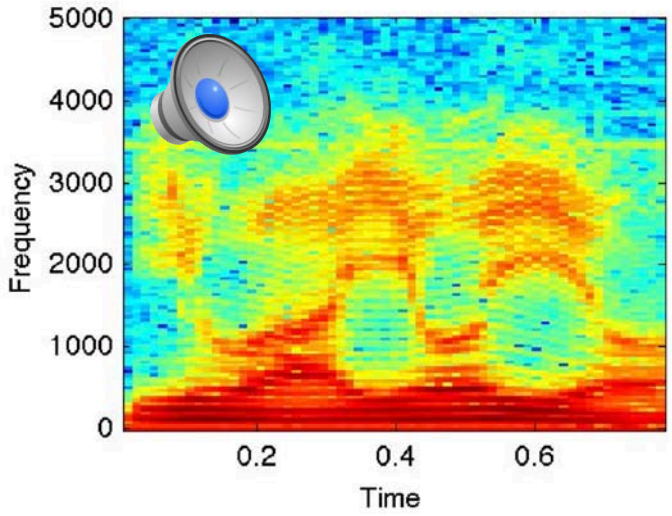


time

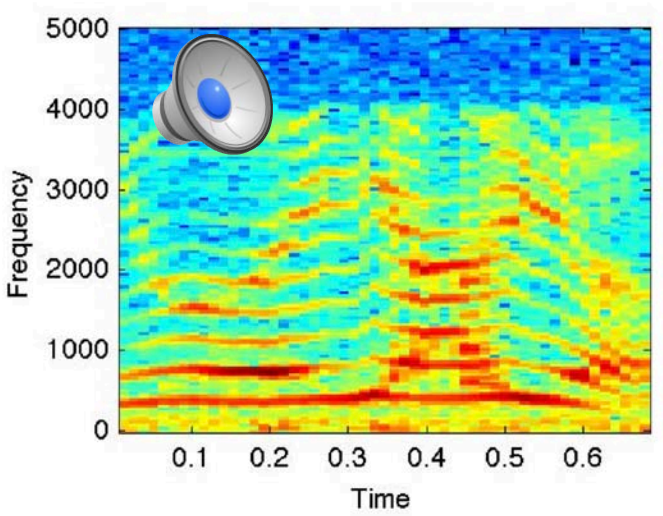
Constant-bandwidth frequency resolution (given by the analysis window)

Different human species

adult male

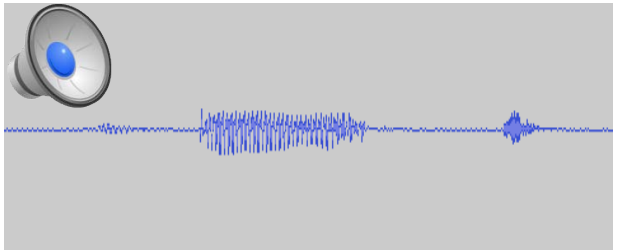


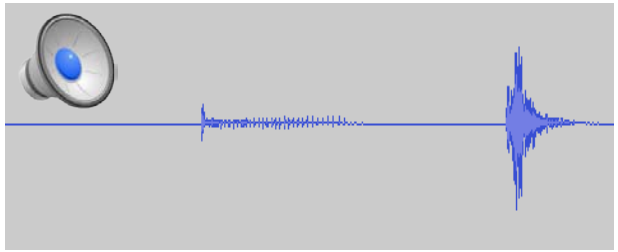
4 year old child

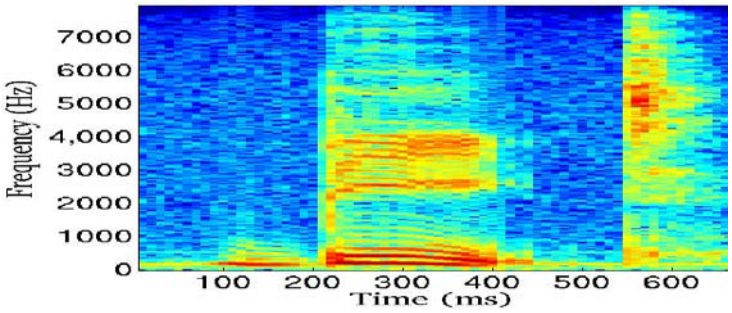


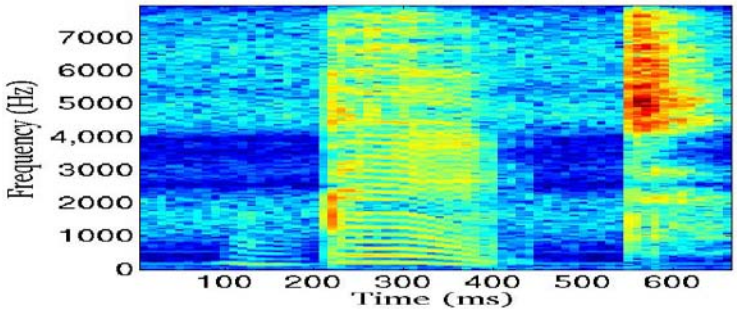
Linear distortions (frequency filtering)

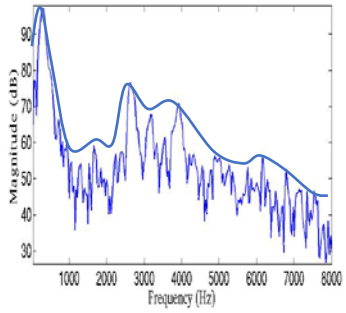
filter

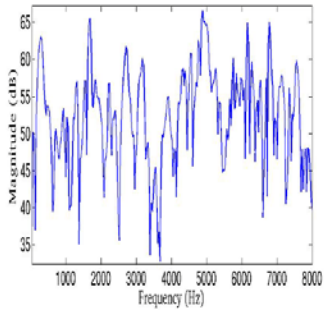




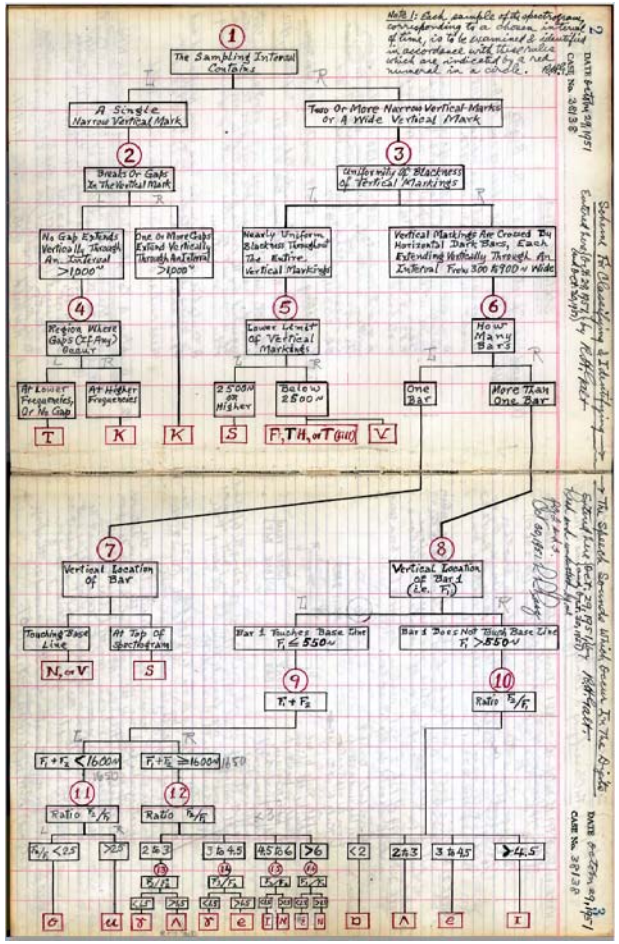




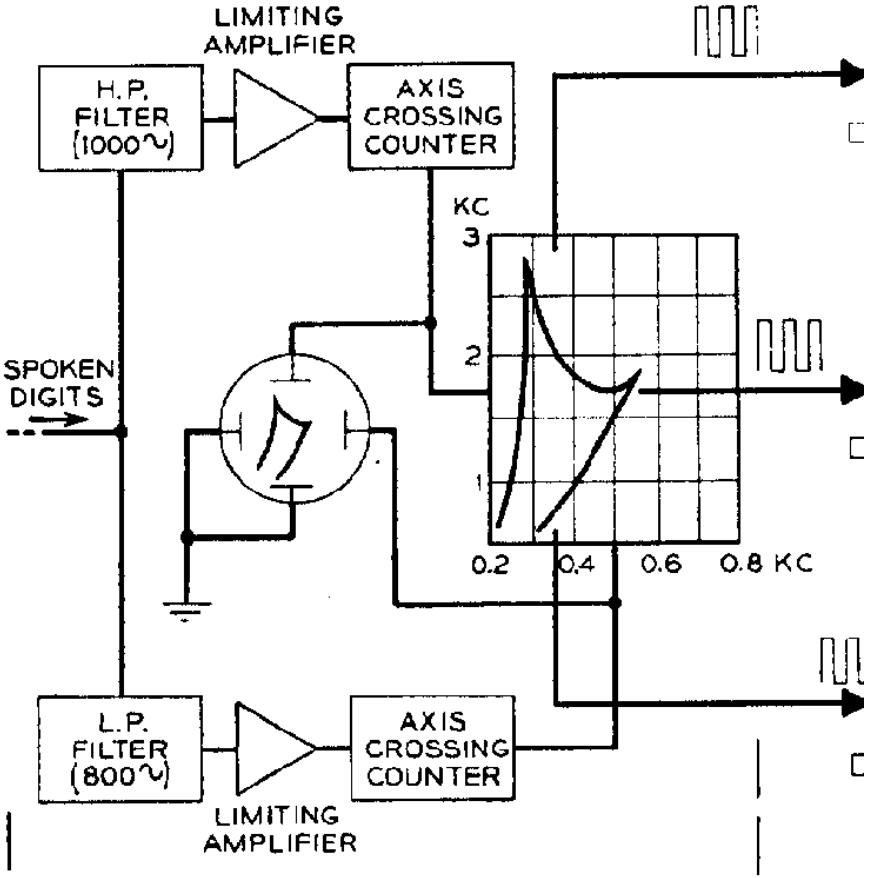




Concept of the first automatic speech recognizer (R.H. Galt 1951)



Winning solution
(Davis, Biddulph, Balashek 1952)



Stochastic machine recognition of speech

$$w = \underset{i}{\operatorname{argmax}}(P(M_i \mid x))$$

How to find w ?

What is the form of the model M ?

What is the data **x** ?

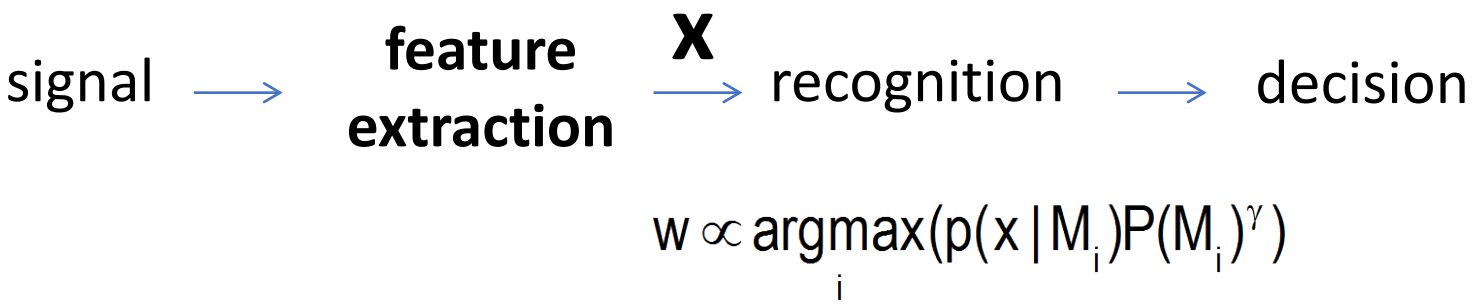
x - speech signal ?

- Changes in acoustic pressure
  - do carry the relevant information
  - always also carry some irrelevant information
    - original purpose is reconstruction of speech

signal

contains wanted and unwanted variability (information)

may be in a form that is not suitable for the recognition stage

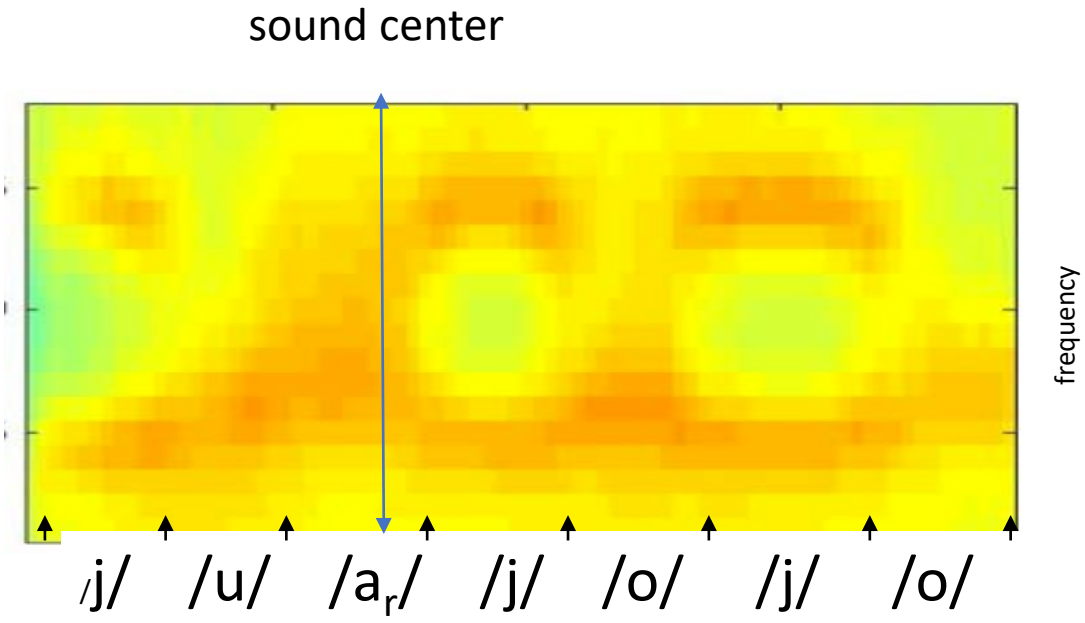


features

what is lost is lost forever

what is kept may cause problems later

Where is information about message in speech?



Mutual information between a points *X* in a time-frequency representation of speech (spectrogram) and a phoneme labels *Y*

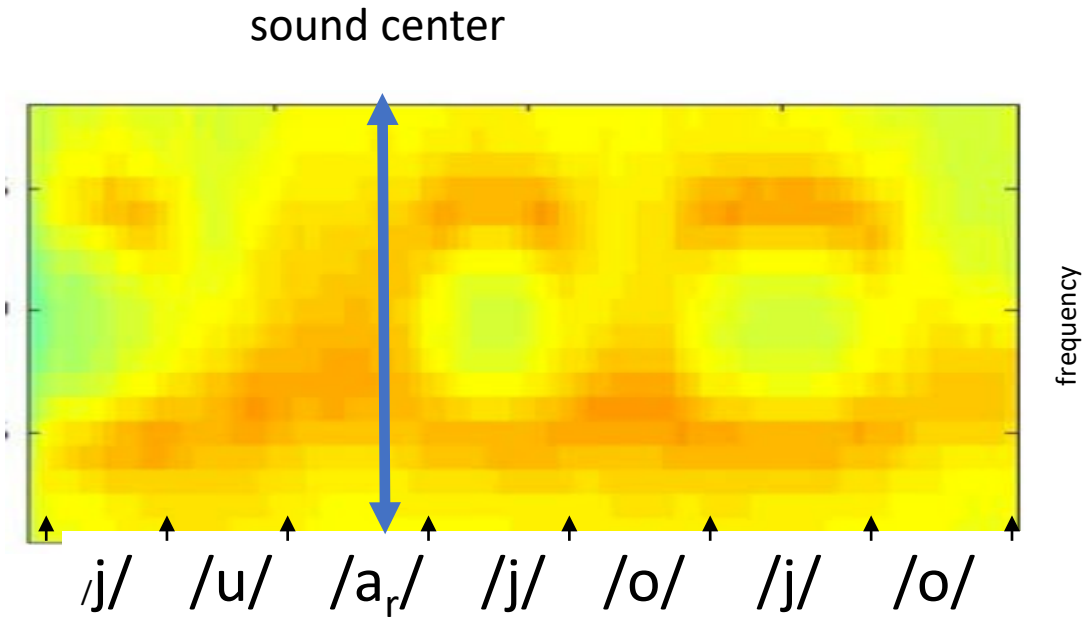
$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

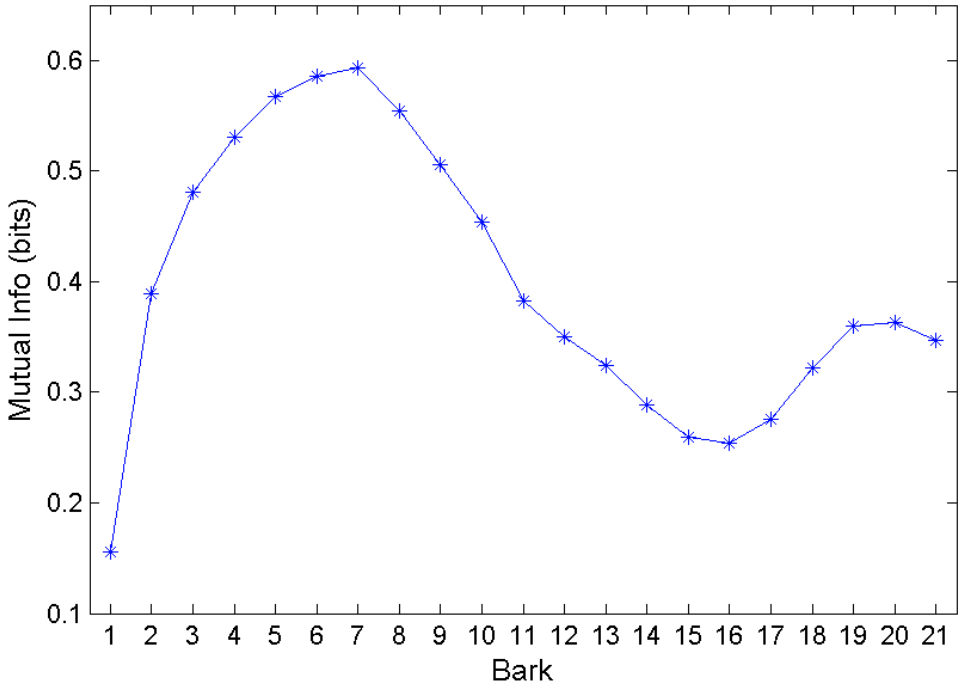
p(x)
Histogram of spectral values at a given frequency band from the whole database
(about 30-40 bins)

p(y)
Histogram of label indices from the whole database
(number of bins=number of phonemes)

p(x,y)
Histogram of instances when the given phoneme and the given quantized spectral value coincide)
number of bins = (30-40) x number of phonemes

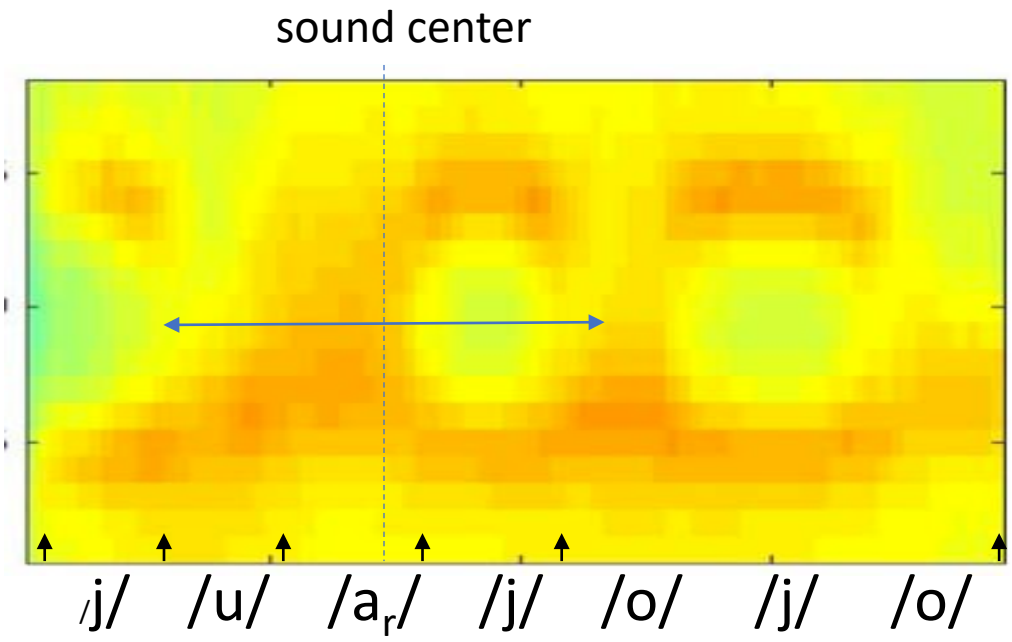
Yang et al, Speech Communication 2000

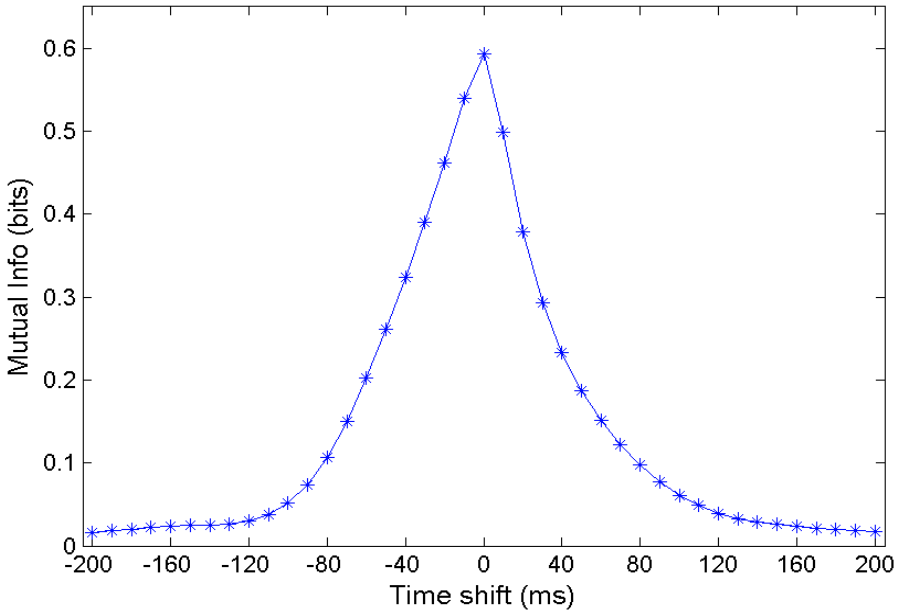




in frequency:
info is spread at all frequencies

Thanks Feipeng Li (now Apple)
for the figure

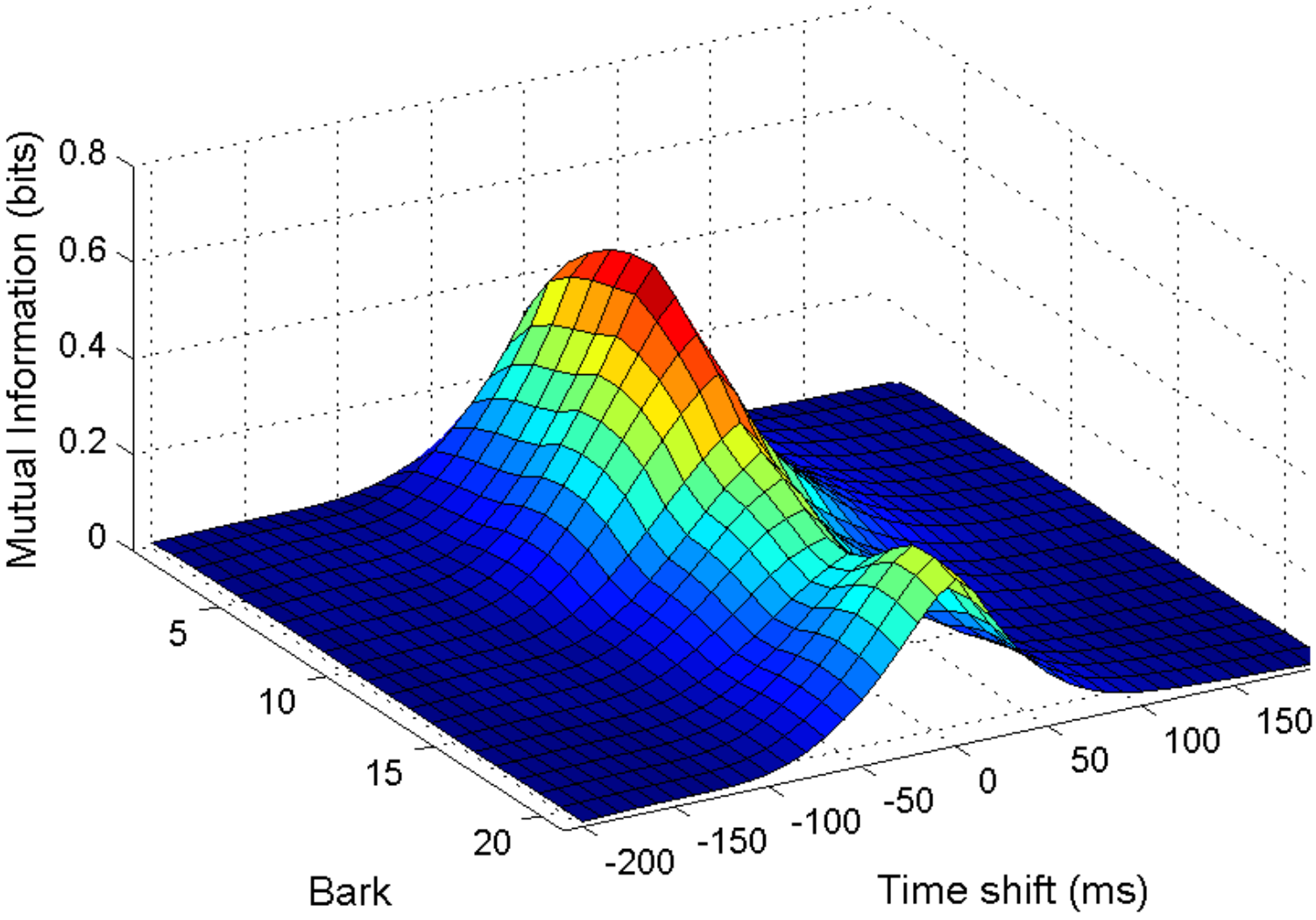




in time:
info spread over about 200 ms

Thanks Feipeng Li (now Apple)
for the figure

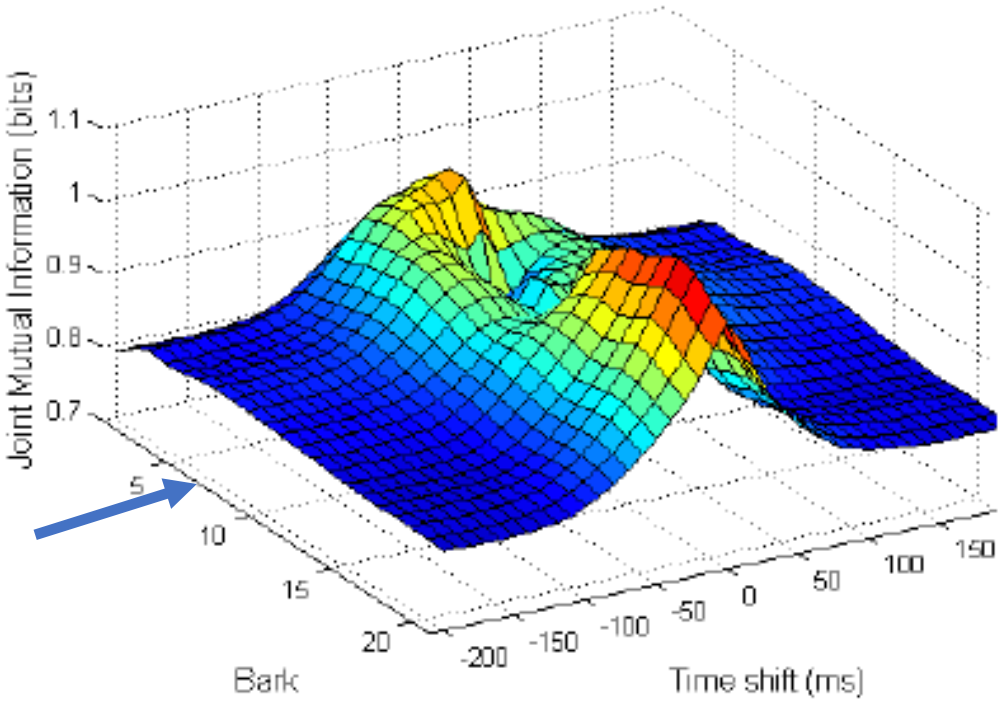
info is spread at all frequencies and over roughly 200 ms

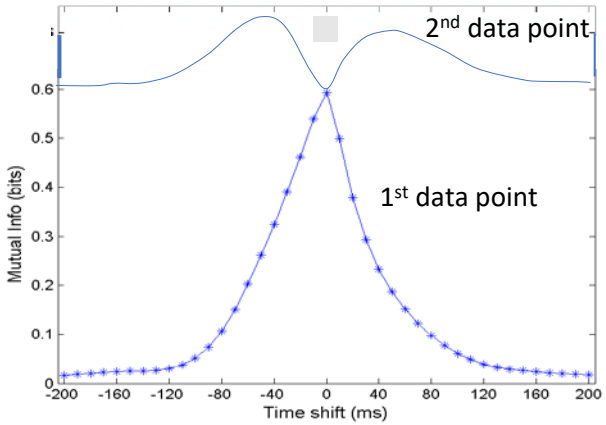


Thanks Feipeng Li (now Apple) for the figure

Additional point in time frequency plane

$$
\begin{aligned}
I(X;Y|Z) &= \mathrm{E}_Z(I(X;Y|Z)) \\
&= \sum_{z\in Z} p_Z(z) \sum_{y\in Y} \sum_{x\in X} p_{X,Y|Z}(x,y|z) \\
&\quad \log\left(\frac{p_{X,Y|Z}(x,y|z)}{p_{X|Z}(x|z)p_{Y|Z}(y|z)}\right)
\end{aligned}
$$





Mutual information when using a second spectral points delayed in time or in frequency (the first one is at 7 Bark and in the phoneme center).

Thanks Feipeng Li (now Apple) for the figure

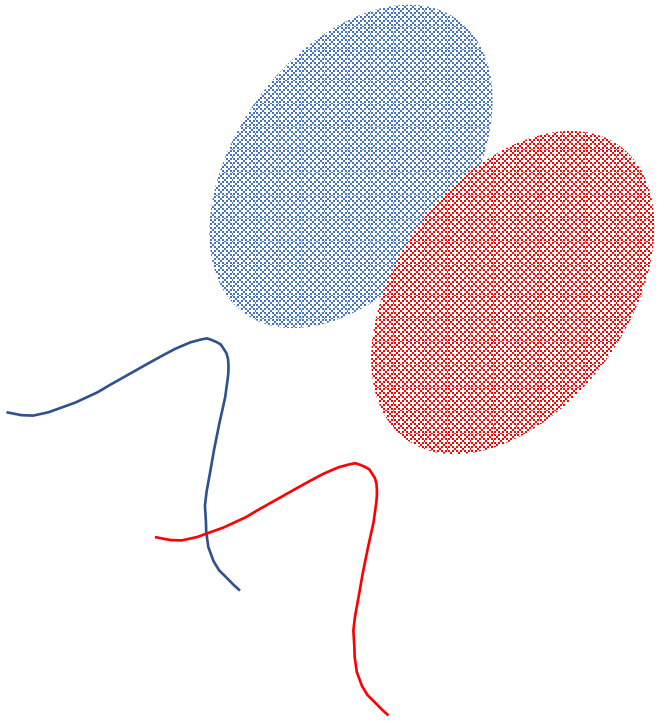
Where in the time frequency plane is the information about speech sounds ?

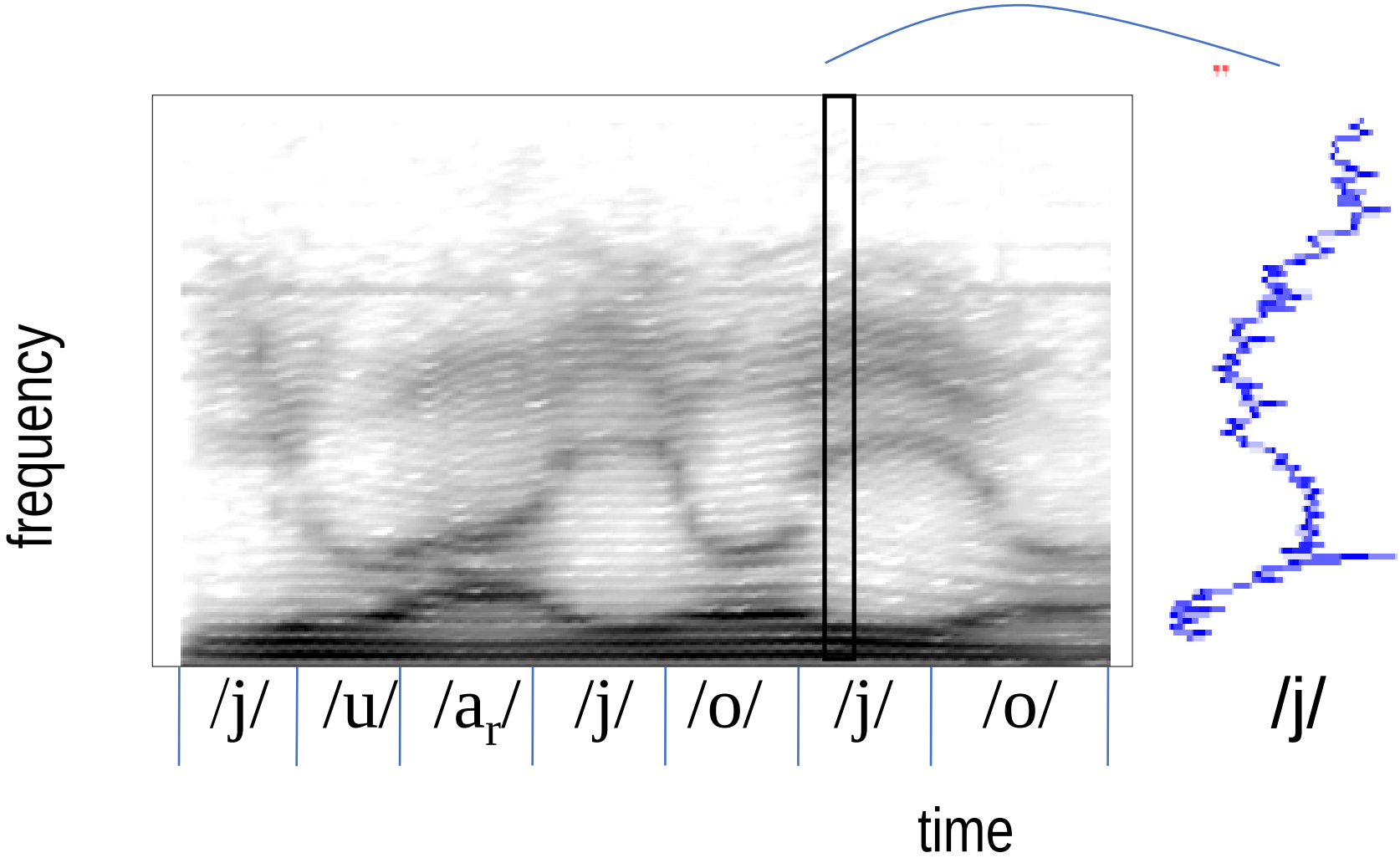
- At all frequencies of speech spectrum
  - most changes of the vocal tract shape are reflected at all frequencies of the spectrum
- Over about 200 ms
  - changes in tract shape obey dynamics of rhythmic movements of human body

Now : How to get the information?

Linear discriminant analysis (Ronald A. Fisher 1936)

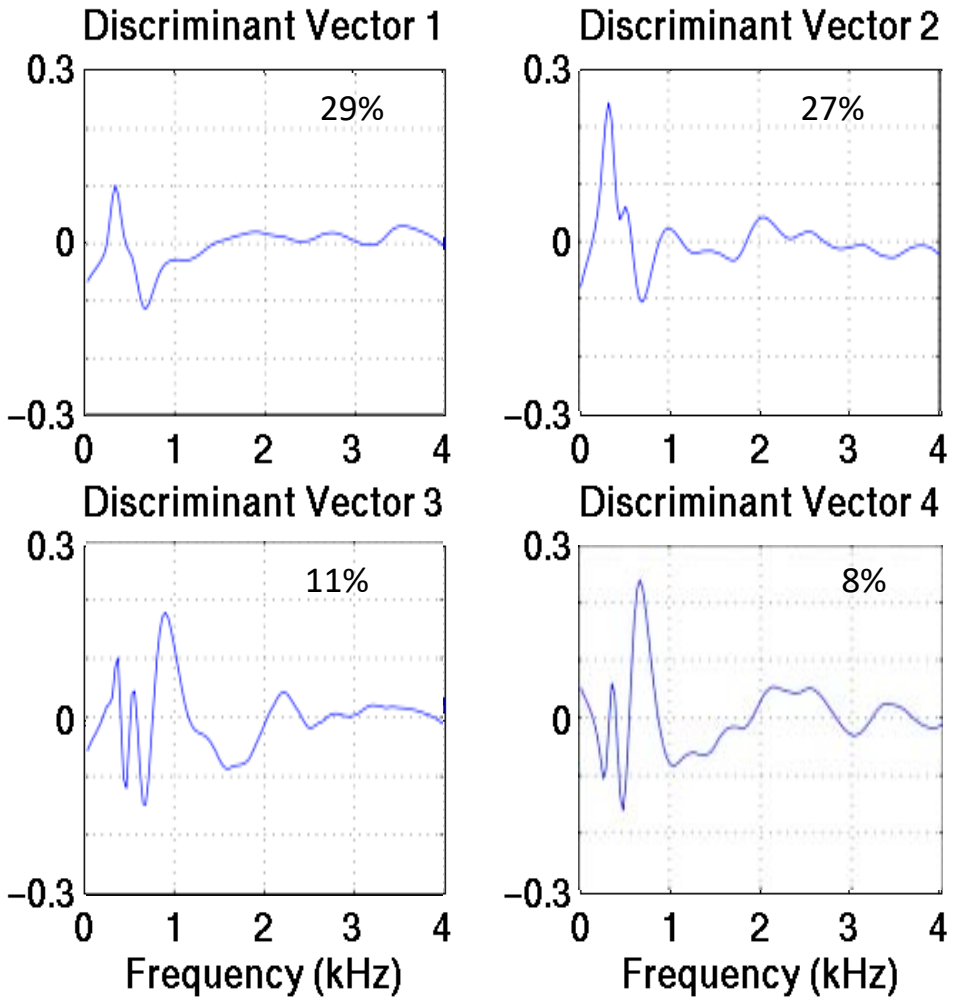
- find projections of data, which preserves most of the discriminability
- data vectors need to be labeled by classes
- yields matrix of discriminant vectors, ordered by their discrimination power
- discriminants are linear and therefore can be easily interpreted




frequency
/j/
/u/
/ar/
/j/
/o/
/j/
/o/
/j/
time

LDA-derived spectral bases

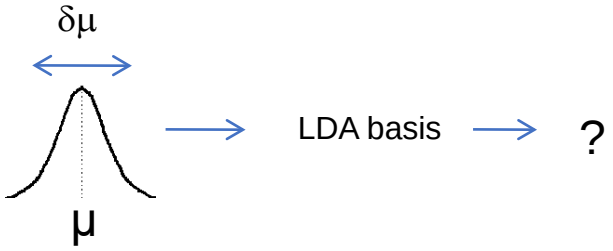
(30 hours of continuous telephone speech database – automatic labeling)



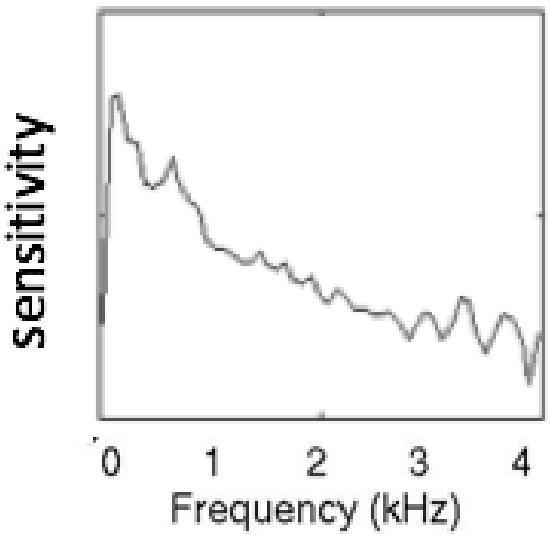
Malayath and Hermansky 1998, Valente and Hermansky 2006

Perturbation analysis

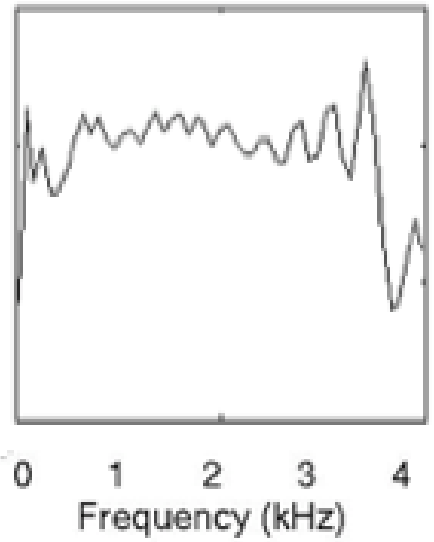
project Gaussian shape on 24 LDA-derived spectral basis and perturb its mean at different frequencies



Perturbation constant on linear frequency scale



Perturbation constant on perceptual critical-band Bark frequency scale



Optimizing spectral processing for discrimination among speech sounds yields human-like spectral resolution.

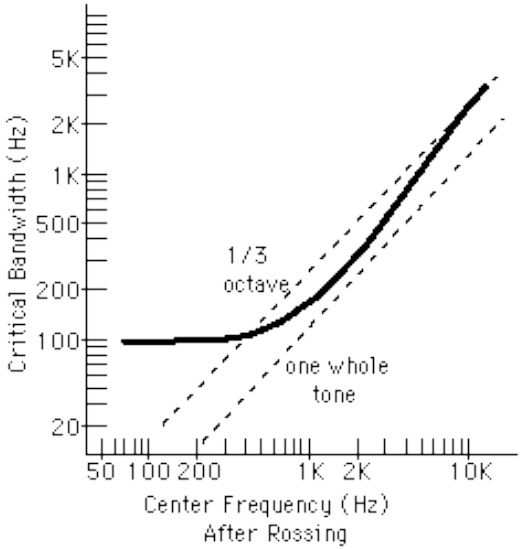
For the past 40 years, automatic speech recognition is using such spectral resolution.

Critical Bands

Fletcher 1934

critical bandwidth increases with frequency

- non-equal resolution of hearing (Bark scale, mel scale,..)



Articulatory Bands

French and Steinberg 1949

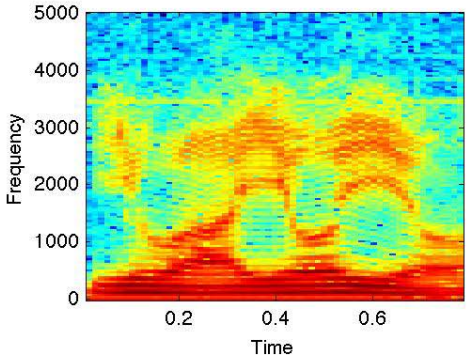
- any 10 streams sufficient for 70% correct recognition of nonsense syllables, and better than 95% correct recognition of meaningful sentences [Fletcher and Steinberg 1929]

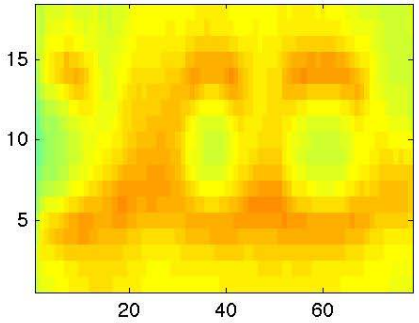
250-375-505-654-795-995-1130-1315-1515-1720-1930-2140-2355-2600-2900-3255-3680-4200-4860-5720-7000 Hz

spectrogram

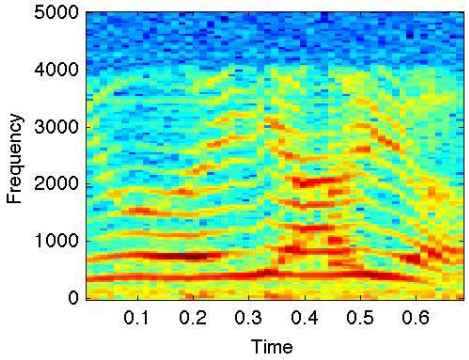
critical-band spectrogram

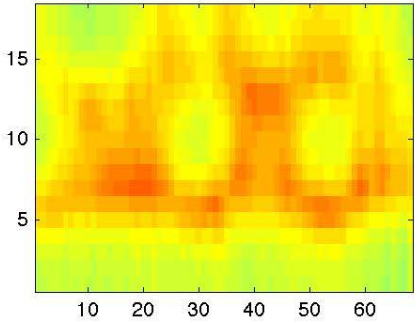
adult male

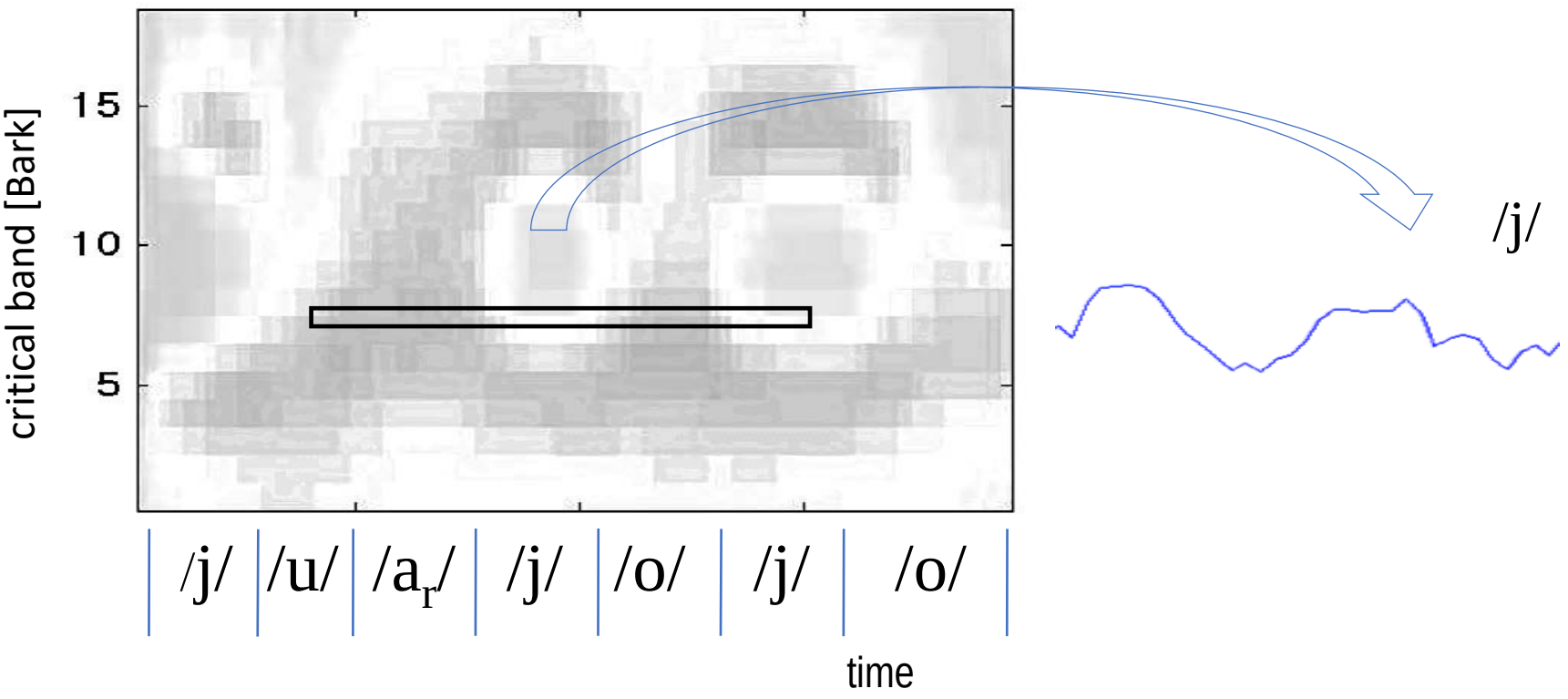




4 year old child





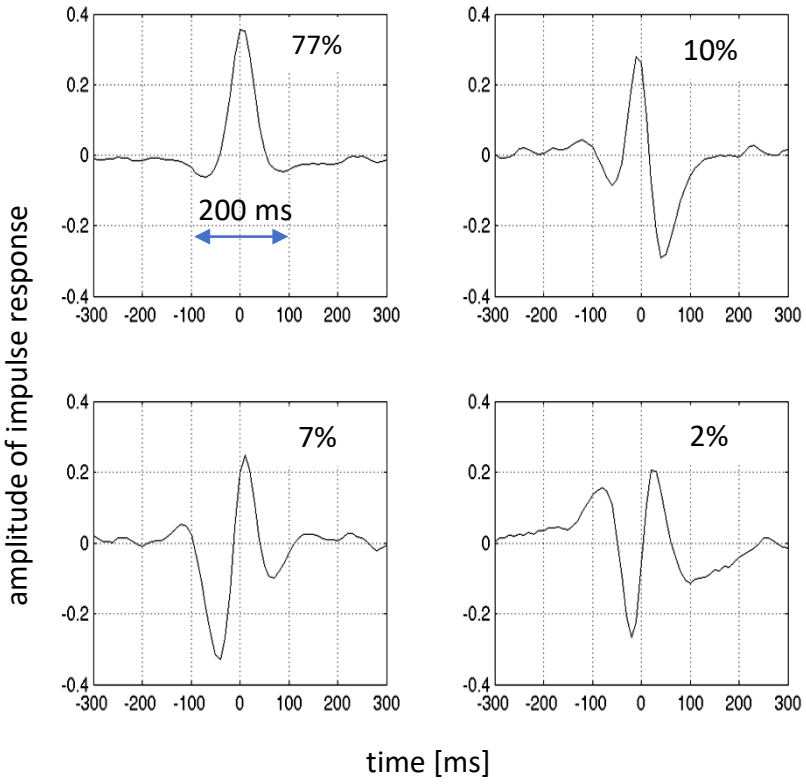

critical band [Bark]
15
10
5
/j/
/u/
/a_r/
/j/
/o/
/j/
/o/
time
/j/

LDA-derived FIR filters

(30 hours of continuous telephone speech database – automatic labeling)

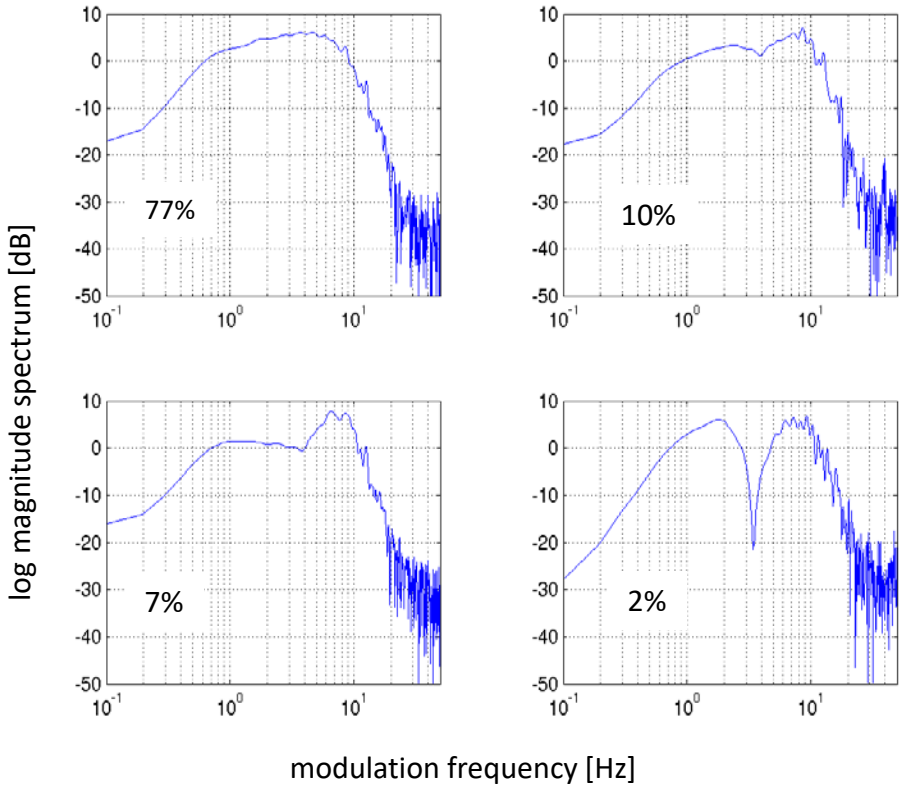
impulse responses

active parts of impulse responses > 200 ms



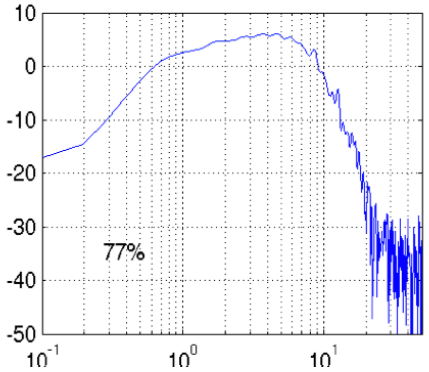
frequency responses

band-pass roughly 1-10 Hz

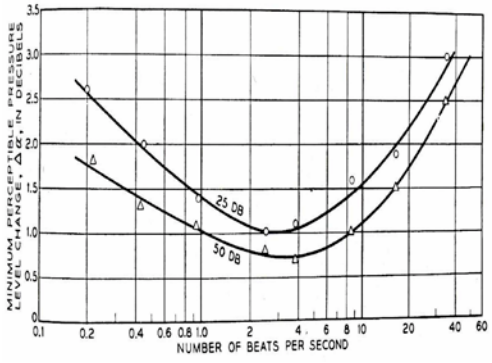


van Vuuren and Hermansky 1997, Valente and Hermansky 2006

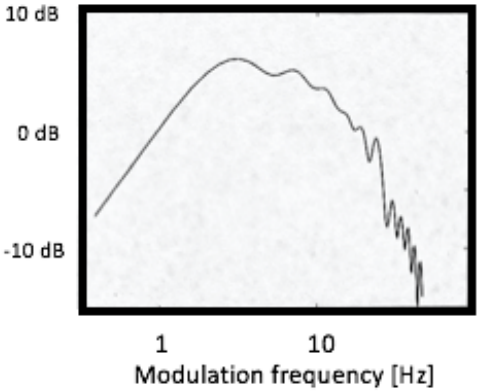
Frequency response of the 1st temporal discriminant



Sensitivity of human hearing to modulations (Riesz 1928)



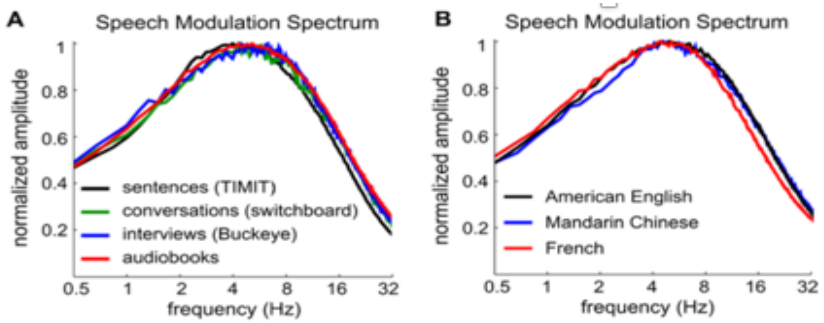
Frequency response of the 1st temporal principal component of about 3000 cortical spectro-temporal receptive fields (ferret)



Mahesan, Mesgarani, Hermansky (in preparation)

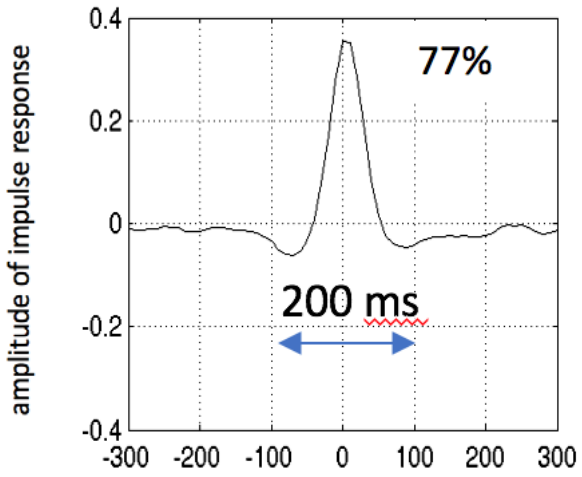
Optimizing temporal processing for discrimination among speech sounds yields filters, which are consistent with temporal properties of mammalian hearing.

Modulation spectra of speech

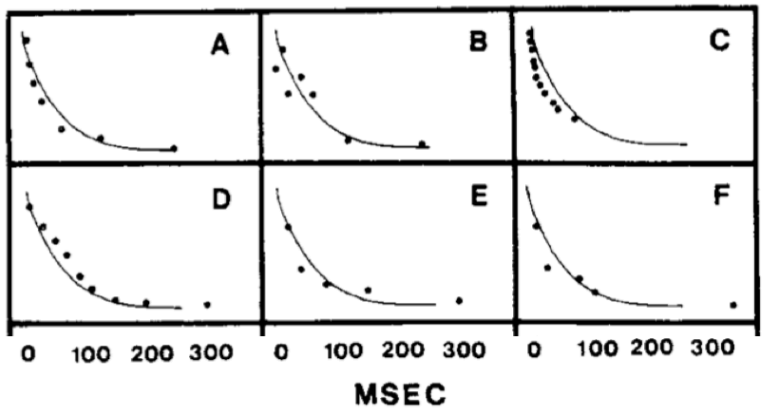


Ding, Patel and Poeppel 2015

Impulse response of the 1st temporal discriminant



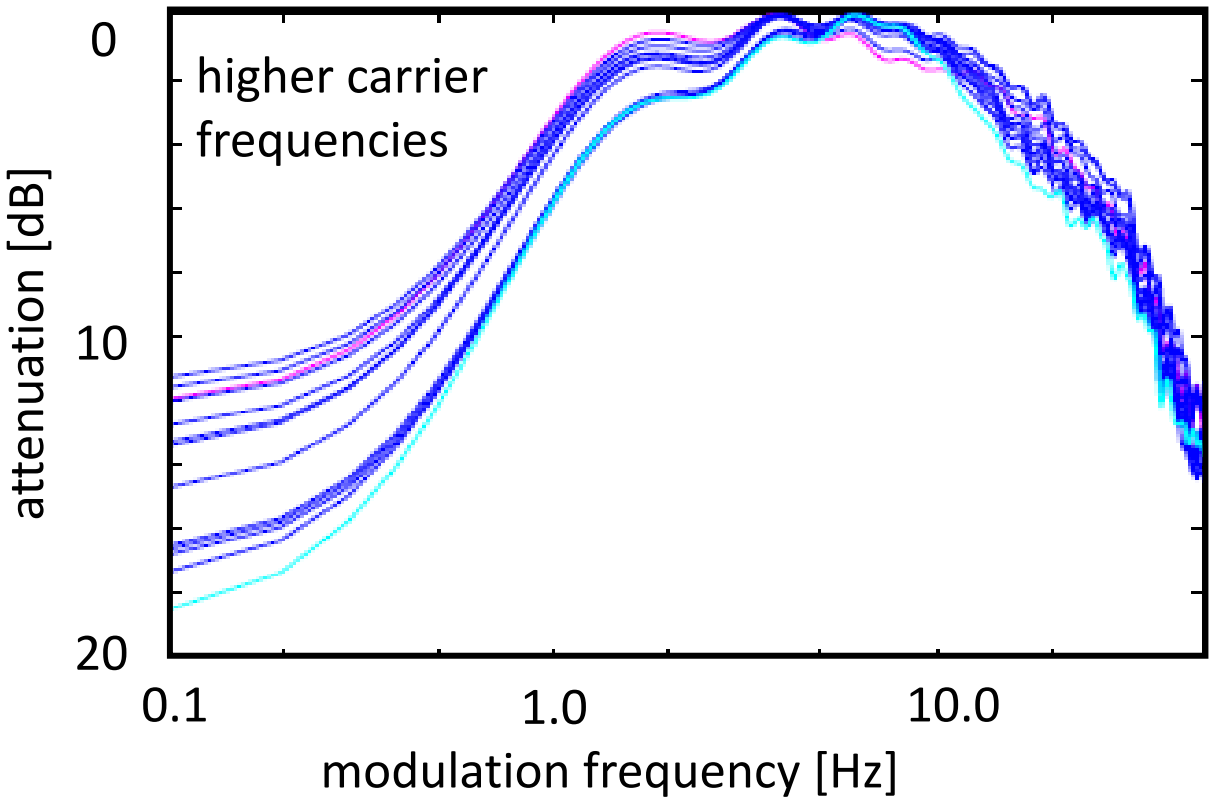
200 ms in auditory perception – Covan 1984



- A-forward masking
- B-backward masking
- C-gap detection
- D-overestimation of short burst duration
- E-loudness decrement
- F-JND in frequency

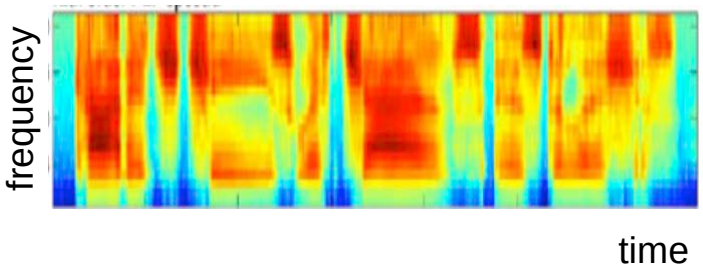
frequency responses

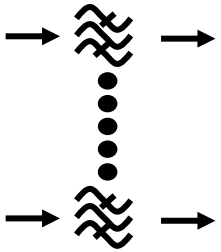
(1st discriminant in all frequency channels)



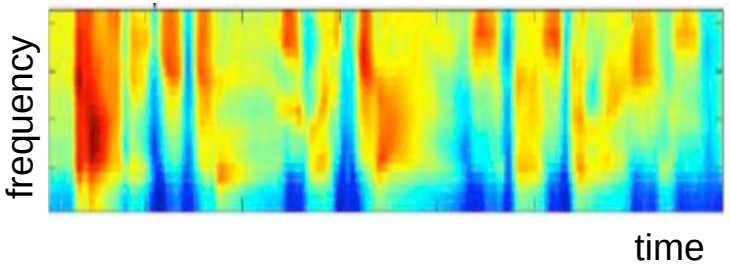
RASTA processing

Hermansky and Morgan 1990

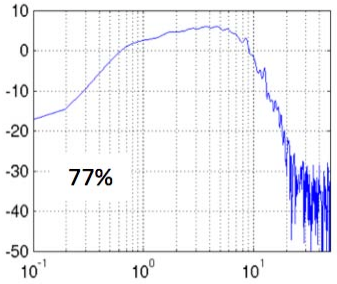




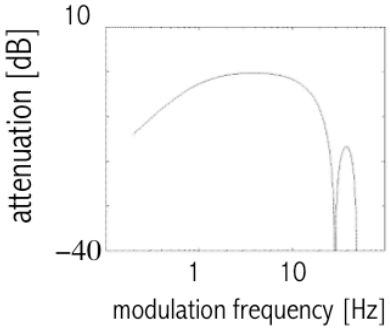
band-pass filters

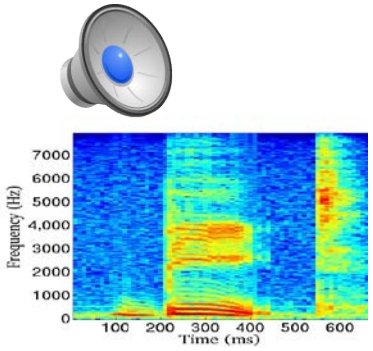


1st LDA discriminant

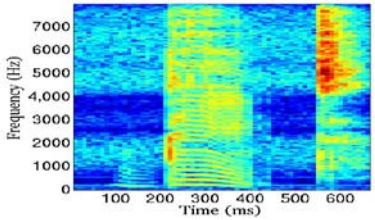


IIR RASTA implementation

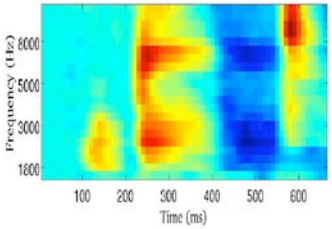


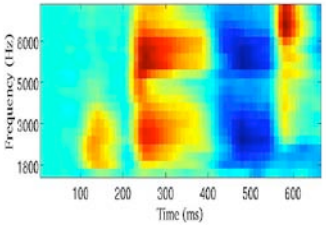


linear distortion

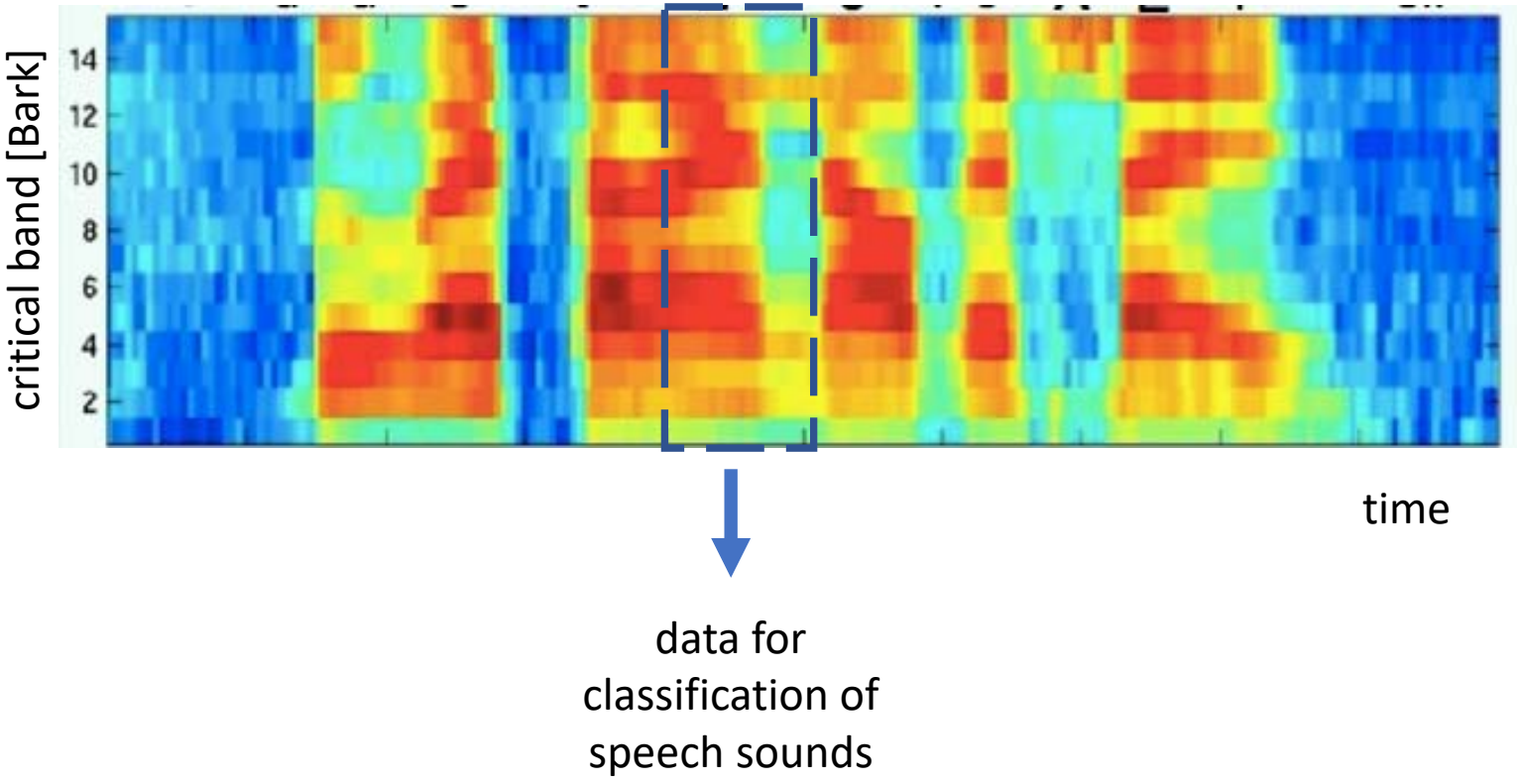


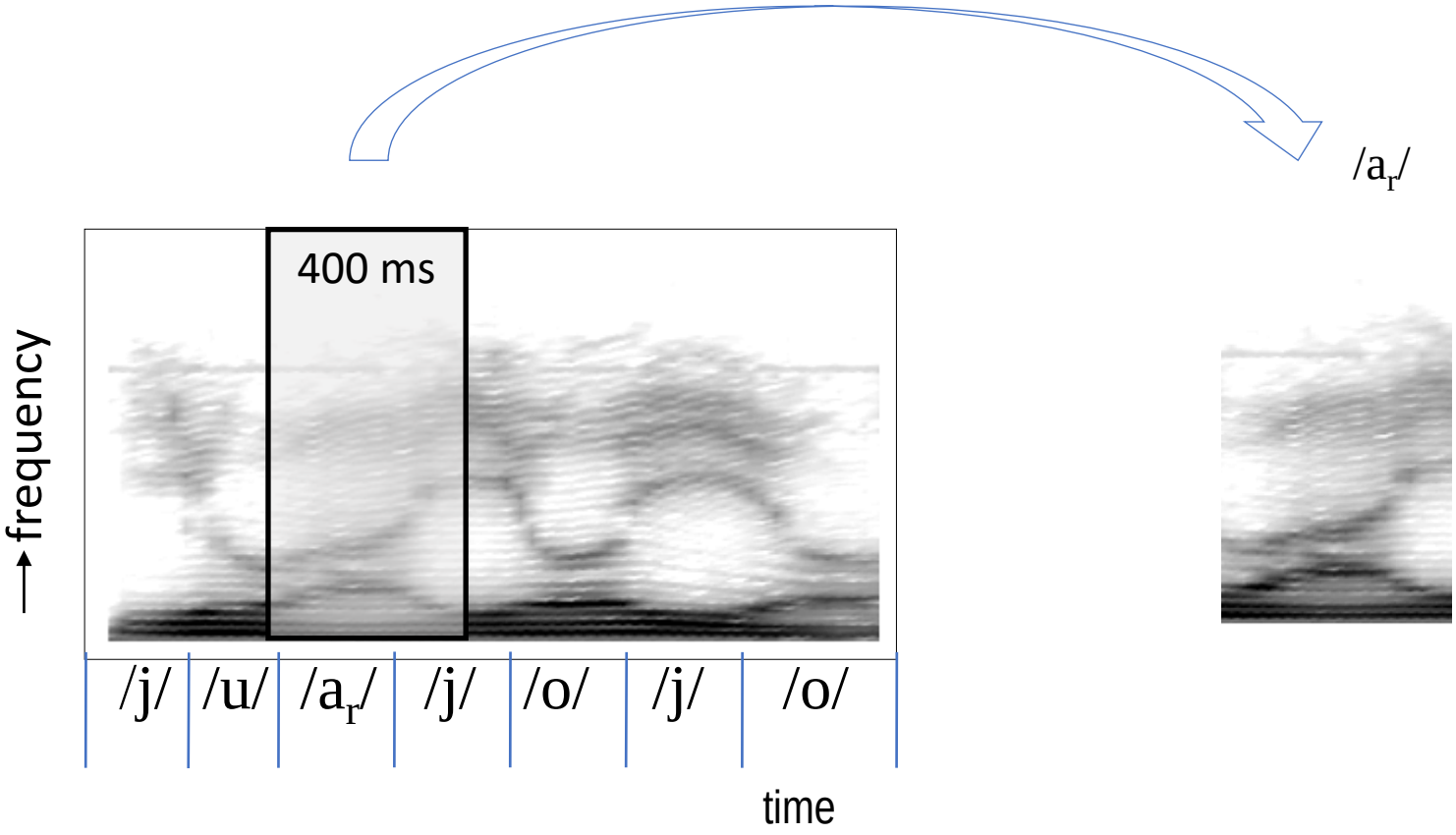
spectrum from RASTA-PLP



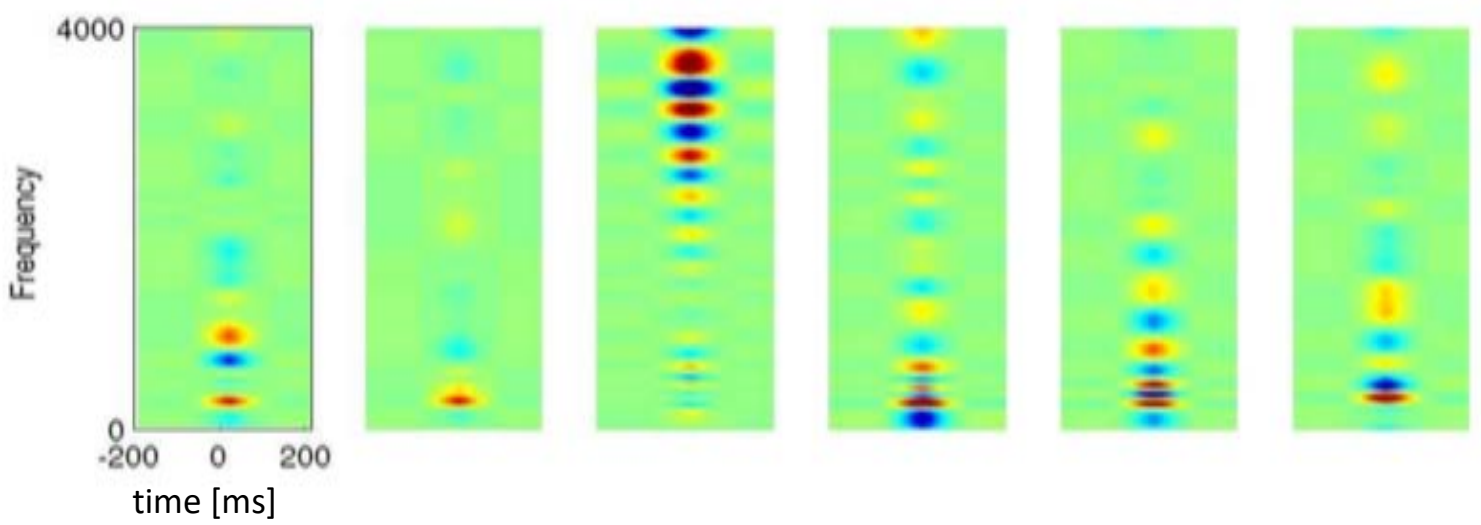


Optimizing for classification of speech sounds suggest critical-band-like spectral resolution and processing within at least 200 ms temporal intervals




/a_r/
400 ms
frequency
/j/ /u/ /a_r/ /j/ /o/ /j/ /o/
time

2D time-frequency discriminants

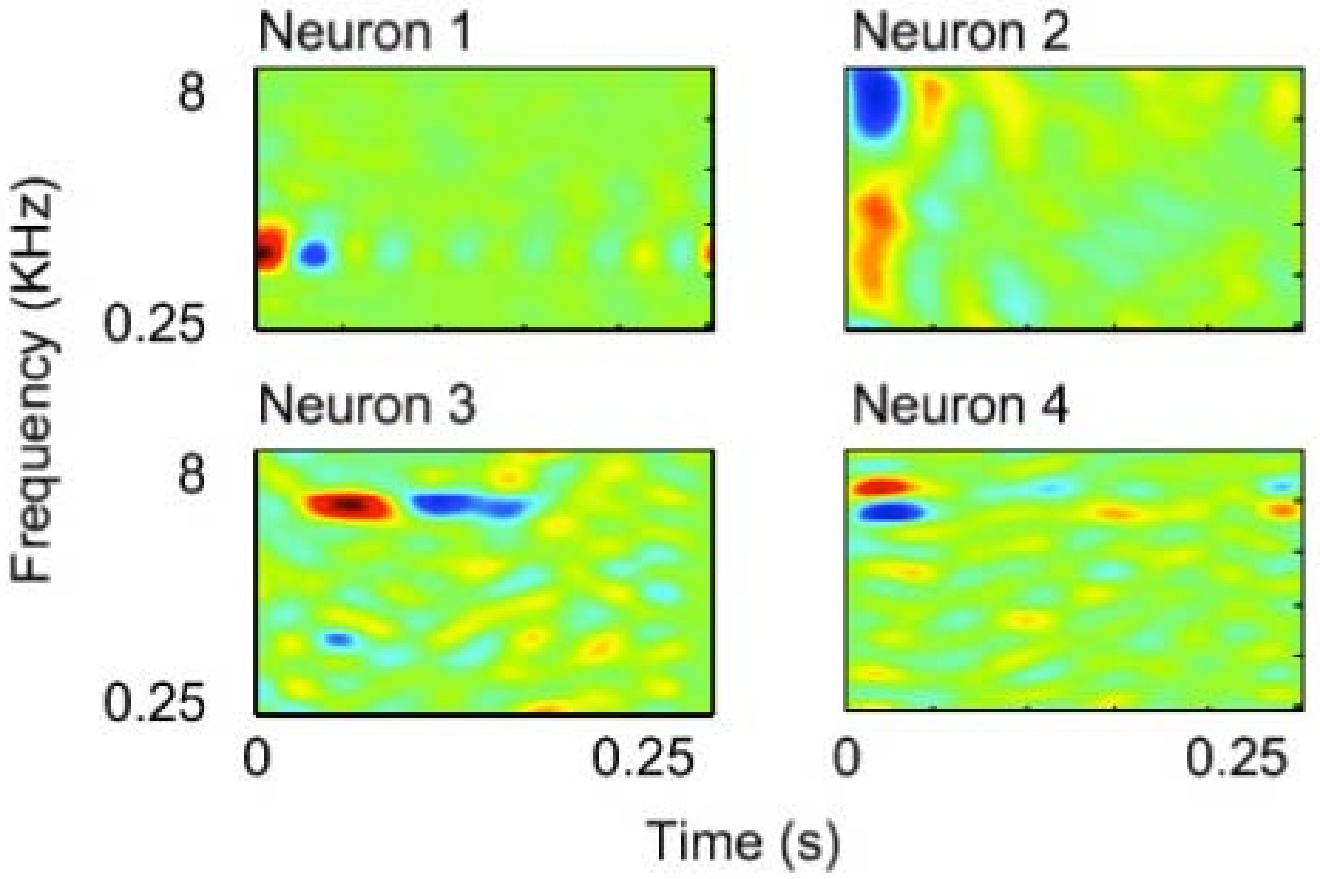


Valente and Hermansky 2006

Many 2D discriminants are frequency-selective, emphasizing particular parts of speech spectrum.

Cortical spectro-temporal receptive fields (STRFs)

[Mesgarani et al Interspeech 2010]



Environment (survival)

Hearing (communication)

Evolution of hearing

200 000 000 years

Evolution of speech

200 000 years

Pristerodon

Homo sapiens

We hear to survive

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

Simoncelli and Olshausen 2001

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 respect properties of human hearing, so properties of hearing is emerging in efficient speech technology.