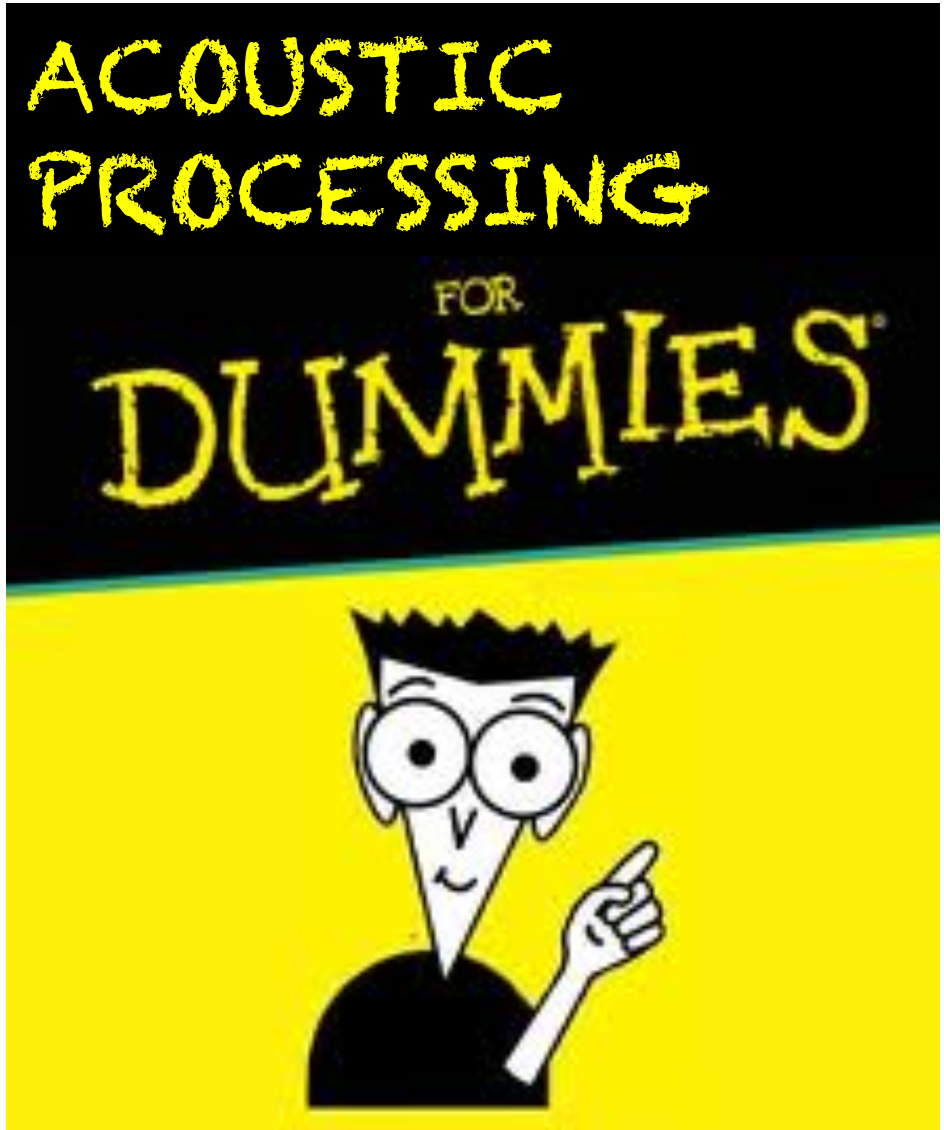

ACOUSTIC
PROCESSING
FOR
DUMMIES

Machine recognition of speech

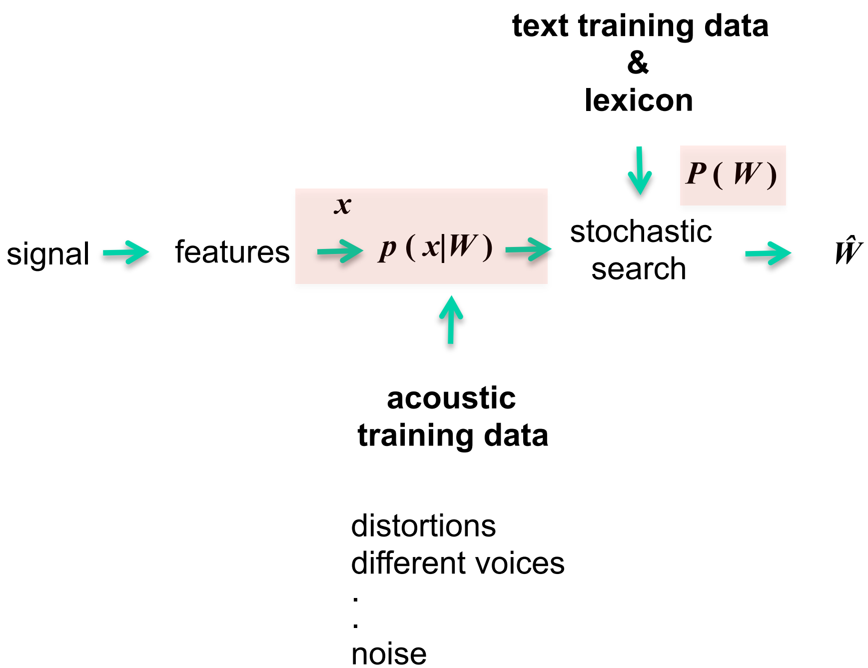
$$w = \arg\max_i (P(M(w_i)/x))$$

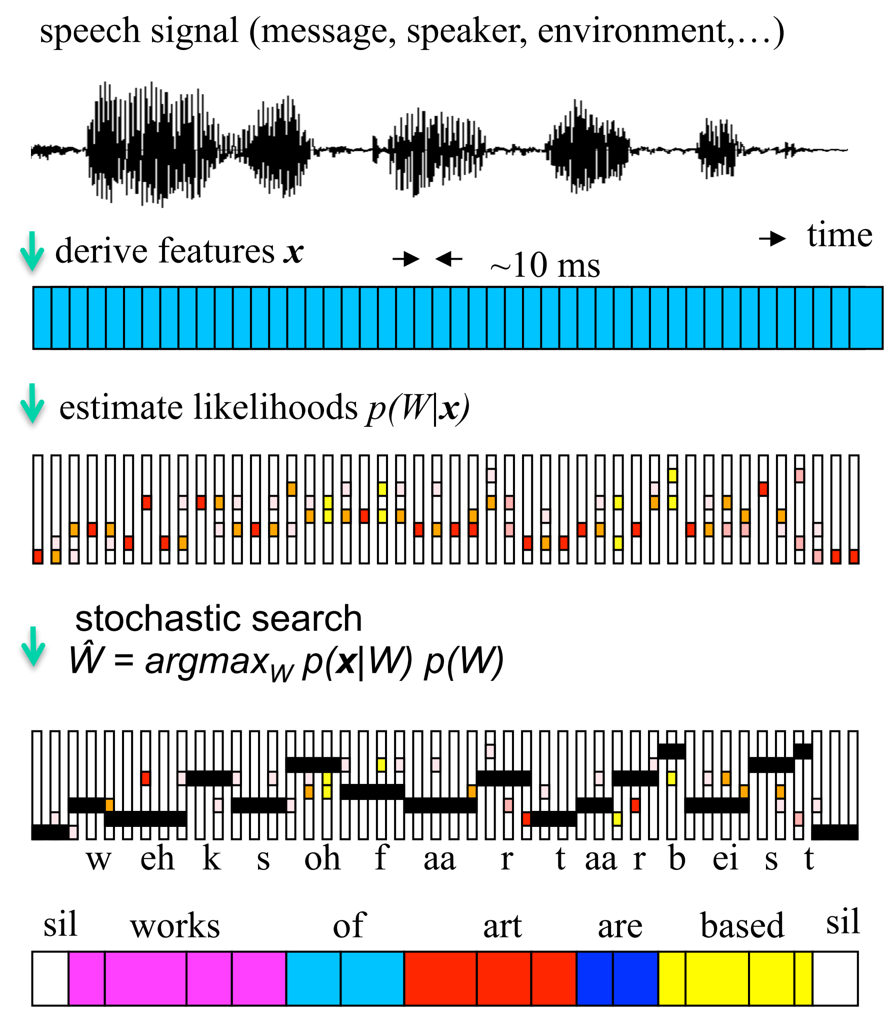
The "best" model found through Bayes rule

$$w \propto \arg\max_i (p(x \mid M(w_i)P(M(w_i))$$

1. How to find *w* ?
2. What is the model $M(w_i)$?
3. **What is the data *x*** ?

Machine Recognition of Speech





- Info lost in ***x***, is lost forever
- Info left must be dealt with later

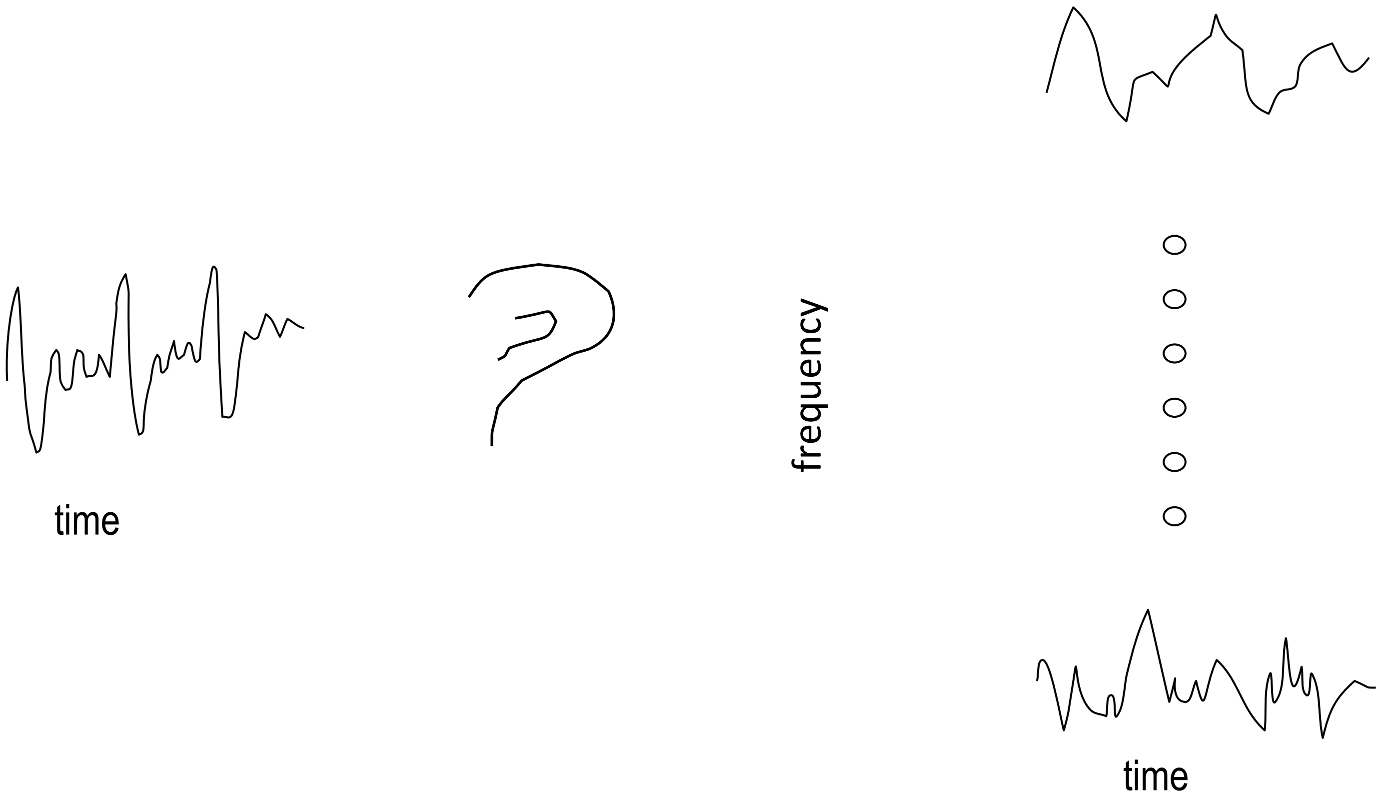
Data x ?

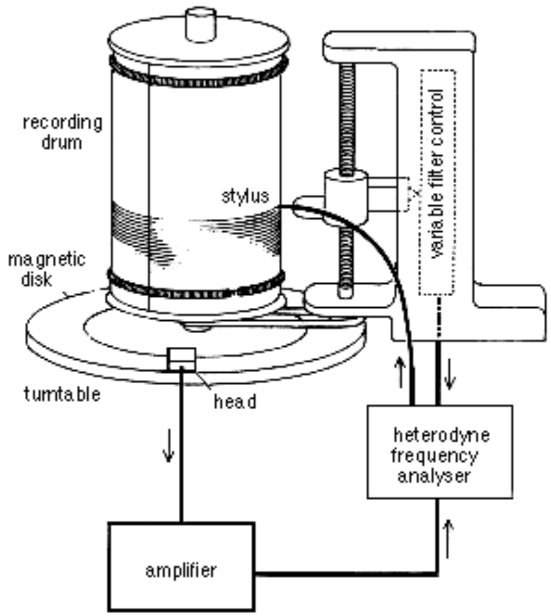
Speech signal ?

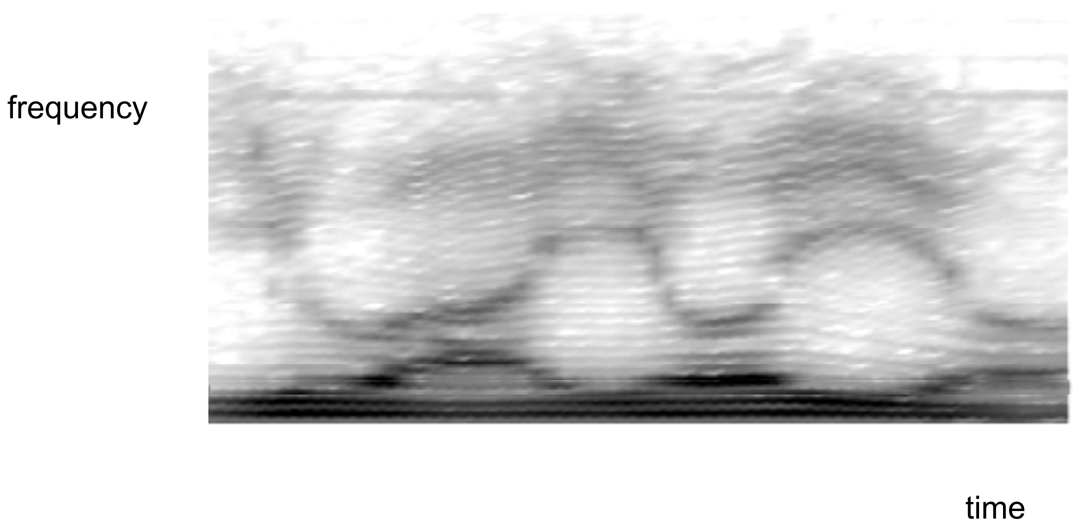
- Describes changes in acoustic pressure
  - original purpose is reconstruction of speech
  - rather high bit-rate

- additional processing is necessary to alleviate the irrelevant information

- besides information lost and retained, additional requirements on ***x*** may exist (Normal distributions, de-correlated,..)

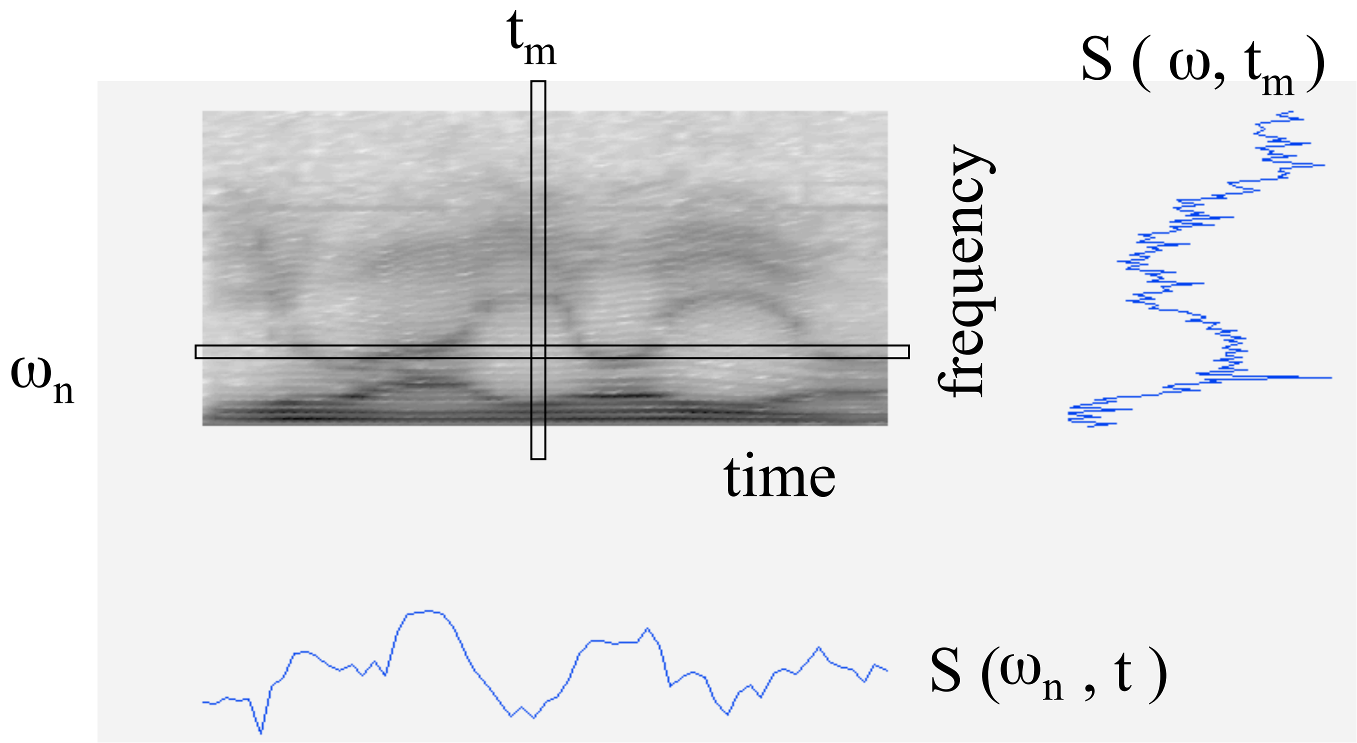

time
frequency
time

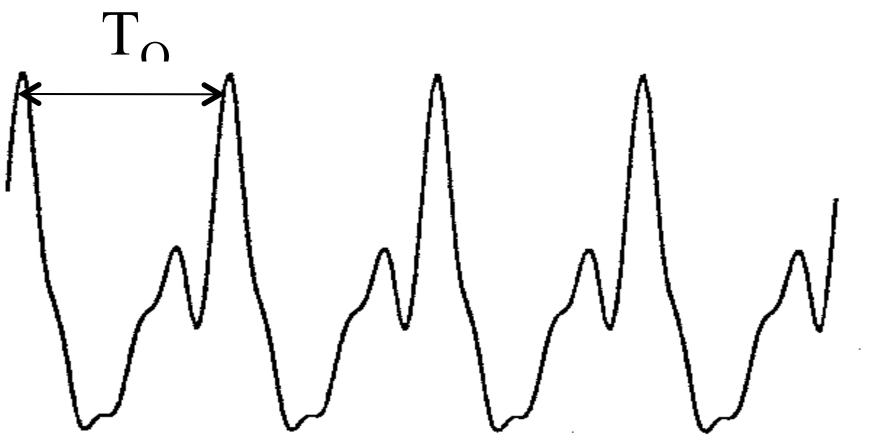

recording
drum
stylus
magnetic
disk
turntable
head
variable filter control
heterodyne
frequency
analyser
amplifier

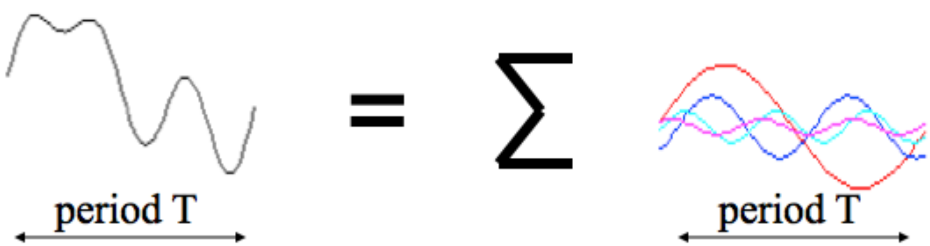

frequency
time

Short-term spectrum

Spectrogram







Joseph Fourier
(1768-1830)
Student of Lagrange
Adviser of e.g. Dirichlet or Navier

- One of Fourier ideas
  - Describe a periodic signal by an (infinite) sum of other well defined periodic signals (sines and cosines)

Orthogonality

$$\int_0^T \sin n\omega t \cdot \cos m\omega t dt = 0$$

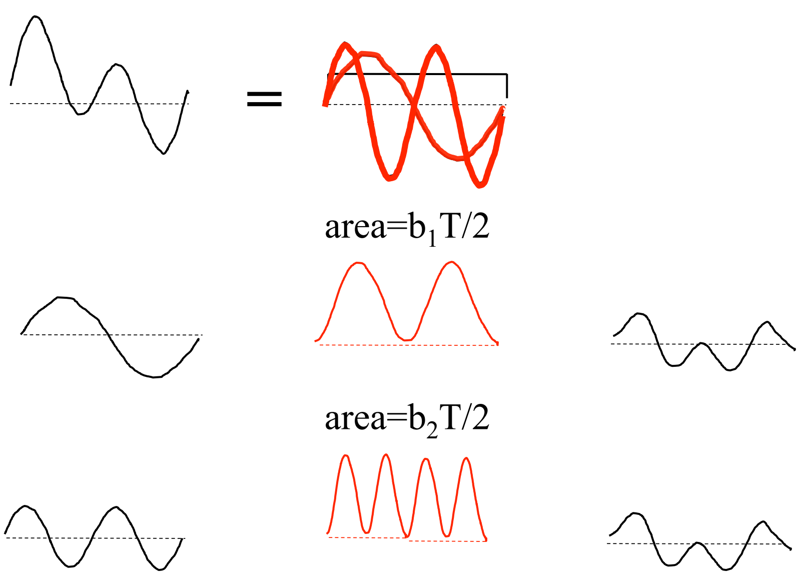
$$\int_0^T \cos n\omega t \cdot \cos m\omega t dt = 0 \text{ for } n \neq m \text{ and } \tfrac{T}{2} \text{ for } n = m$$

$$\int_0^T \sin n\omega t \cdot \sin m\omega t dt = 0 \text{ for } n \neq m \text{ and } \tfrac{T}{2} \text{ for } n = m$$

$$f(t) = DC + \sum_{i=1}^{\infty}\left[a_i \cos\left(\frac{2\pi i t}{T}\right) + b_1 \sin\left(\frac{2\pi i t}{T}\right)\right] = DC + a_1 \cos\left(\frac{2\pi t}{T}\right) + b_1 \sin\left(\frac{2\pi t}{T}\right) + a_2 \cos\left(\frac{4\pi t}{T}\right) + b_2 \sin\left(\frac{4\pi t}{T}\right) + a_3 \cos\left(\frac{6\pi t}{T}\right) + b_3 \sin\left(\frac{6\pi t}{T}\right) + \ldots\ldots\ldots$$

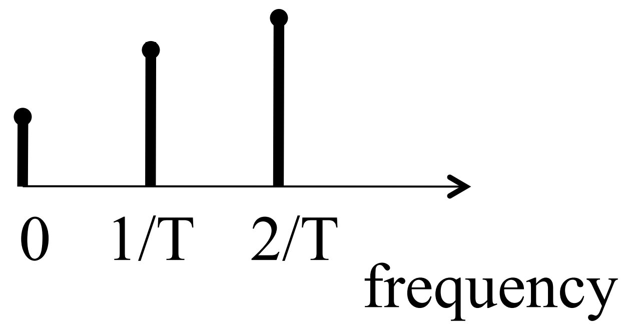
$$\int_0^T f(t)\sin\left(\frac{2\pi t}{T}\right)dt = \int_0^T \left\{DC \sin\left(\frac{2\pi t}{T}\right) + a_1 \cos\left(\frac{2\pi t}{T}\right)\sin\left(\frac{2\pi t}{T}\right) + b_1 \sin\left(\frac{2\pi t}{T}\right)\sin\left(\frac{2\pi t}{T}\right) + a_2 \cos\left(\frac{4\pi t}{T}\right)\sin\left(\frac{2\pi t}{T}\right) + b_2 \sin\left(\frac{4\pi t}{T}\right)\sin\left(\frac{2\pi t}{T}\right) + \ldots\ldots\ldots\right\}dt$$

$0 \qquad 0 \qquad b_1T/2 \qquad 0 \qquad 0 \ldots\ldots\ldots\ldots\ldots$

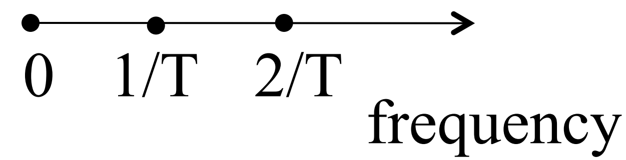


$$\int_0^T \sin^2\left(\frac{t}{T}\right) dt = \frac{T}{2}$$

Magnitude spectrum

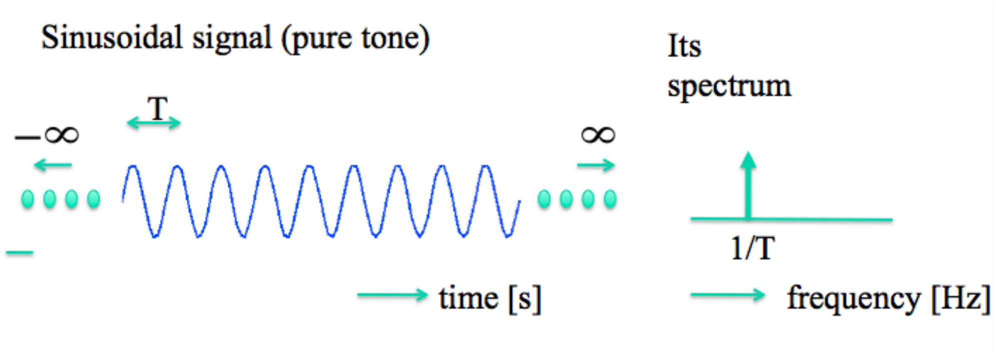


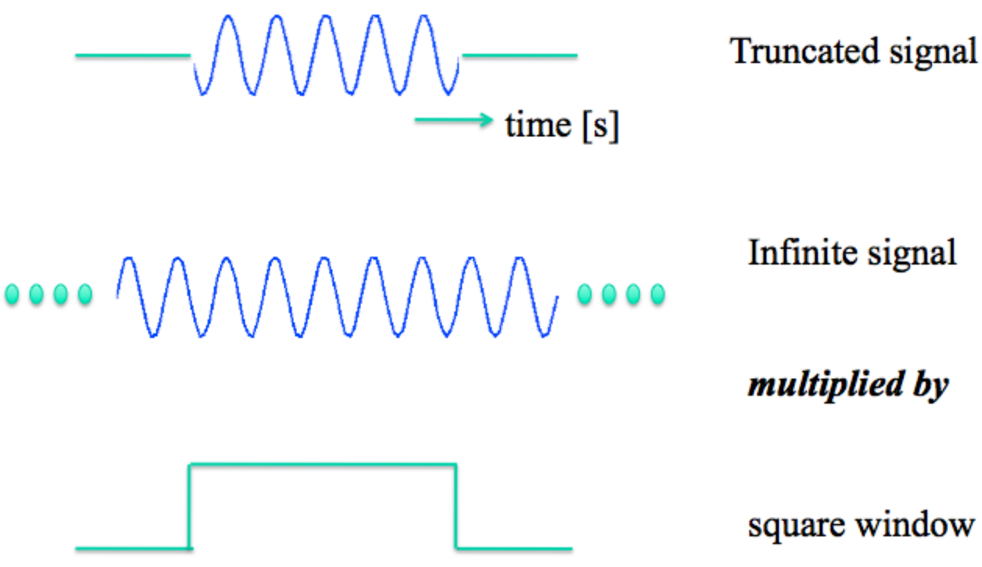
Phase spectrum

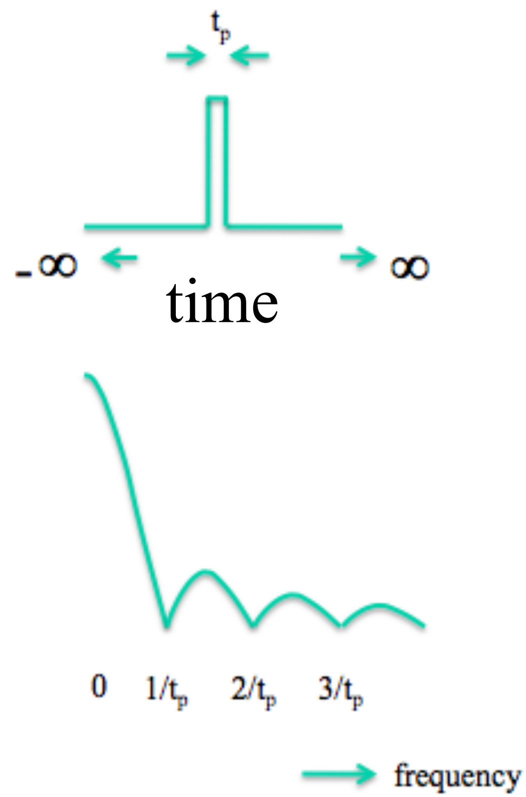


Spacing of spectral components is 1/T

Periodicity in one domain (here time) implies discrete representation in the dual domain (here frequency)



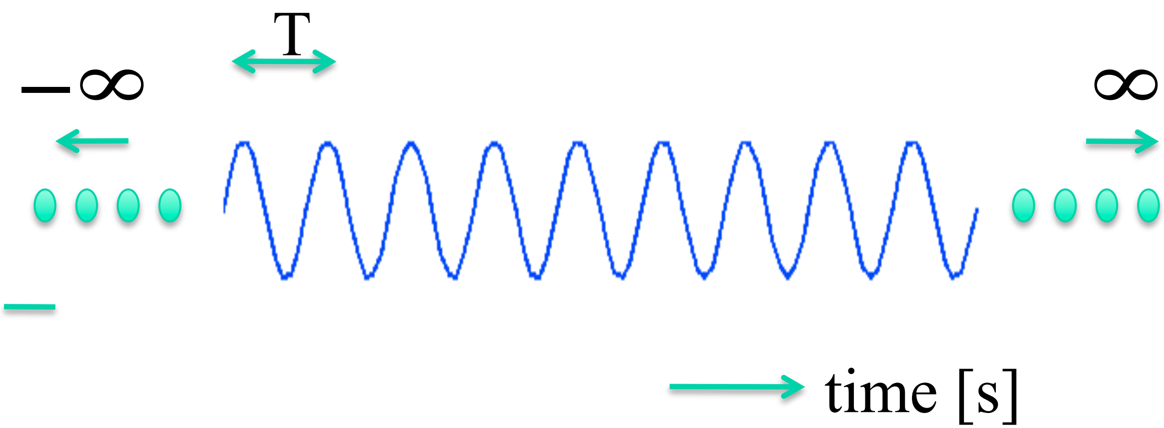


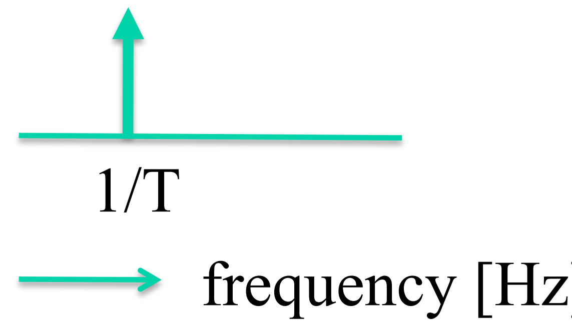


Multiplication in one (time) domain is convolution in the dual (frequency) domain

Sinusoidal signal (pure tone)

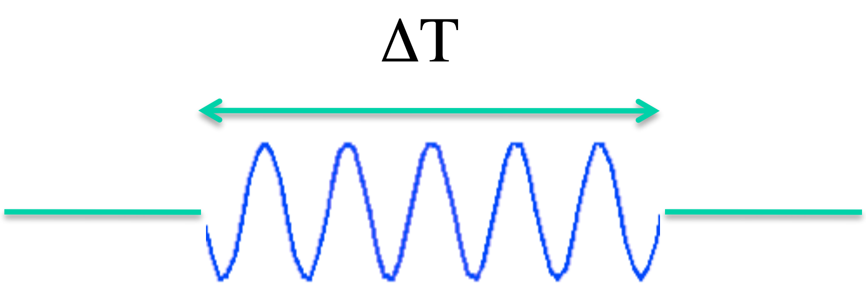
Its spectrum



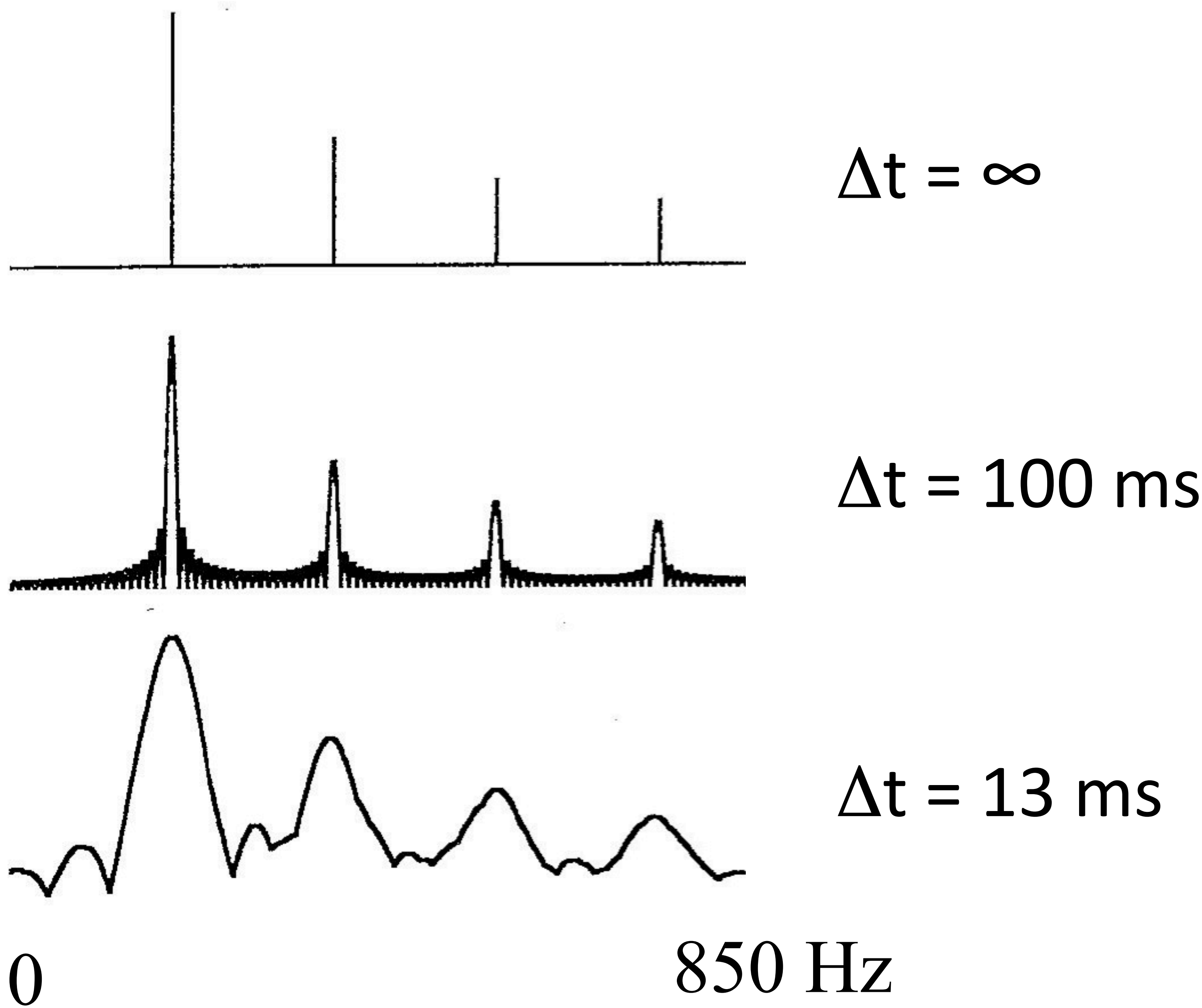


Truncated sinusoidal signal

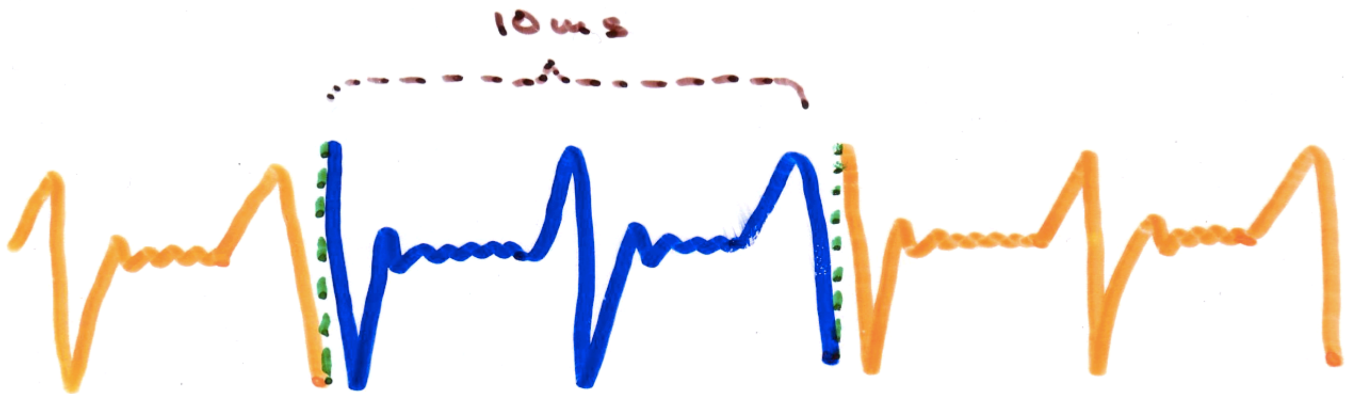
Its spectrum



?


Δt = ∞
Δt = 100 ms
Δt = 13 ms
0
850 Hz

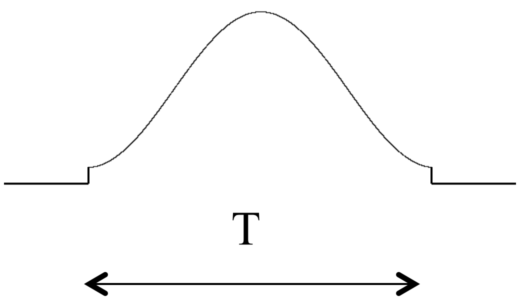
Non-stationary turns into periodic



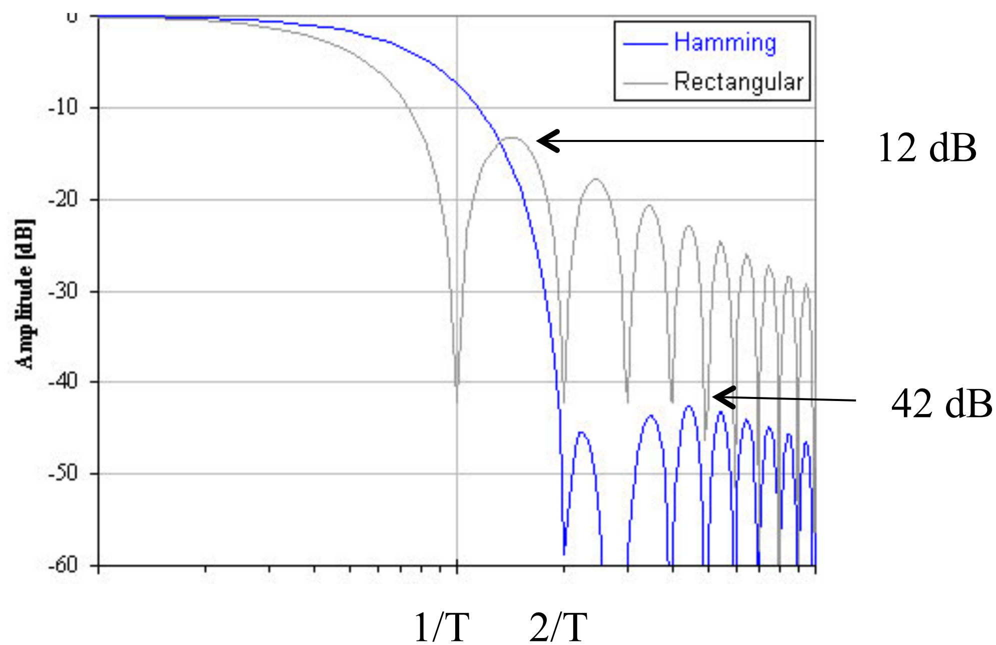
Time domain

square

Hamming

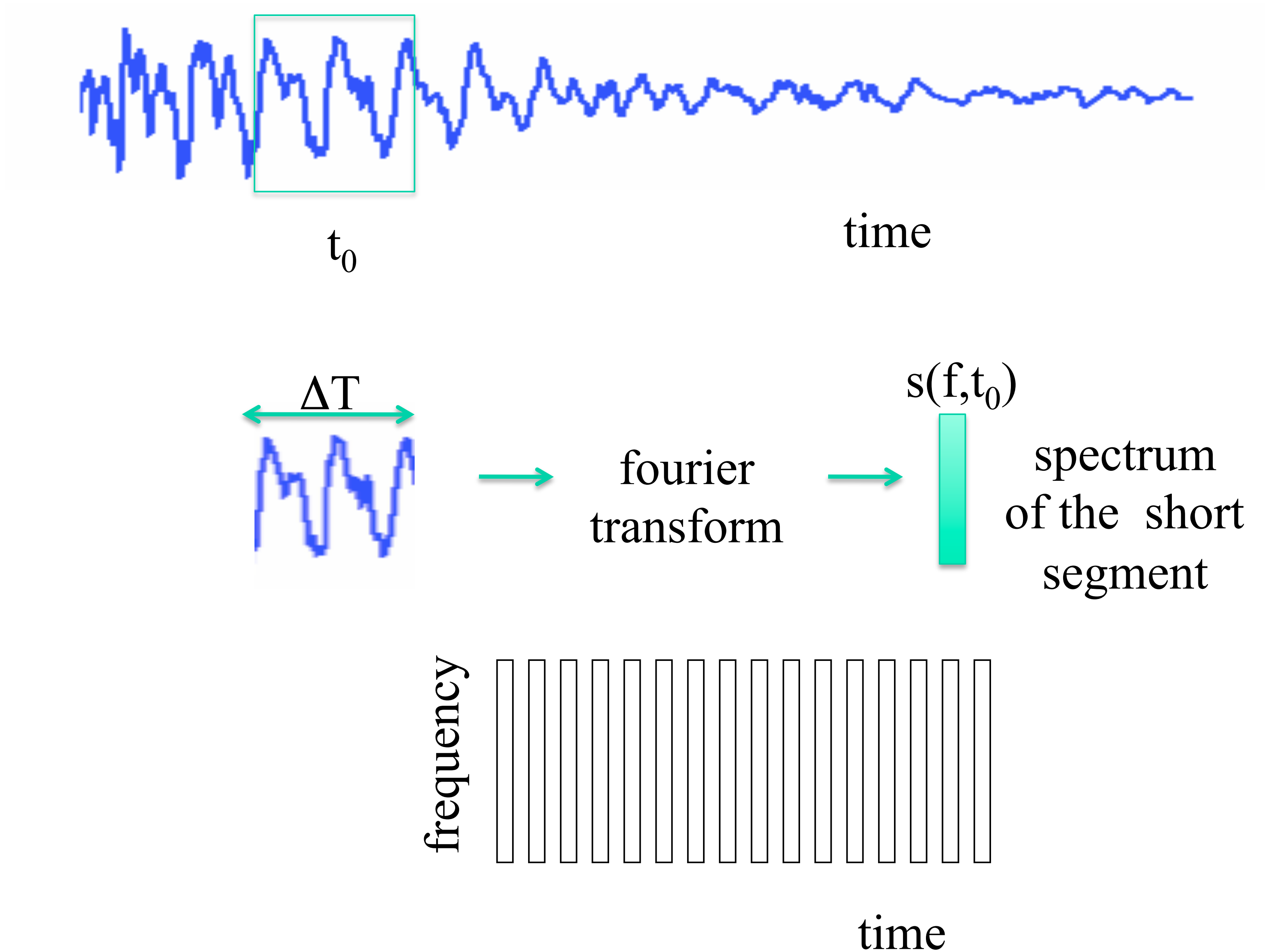


Frequency domain



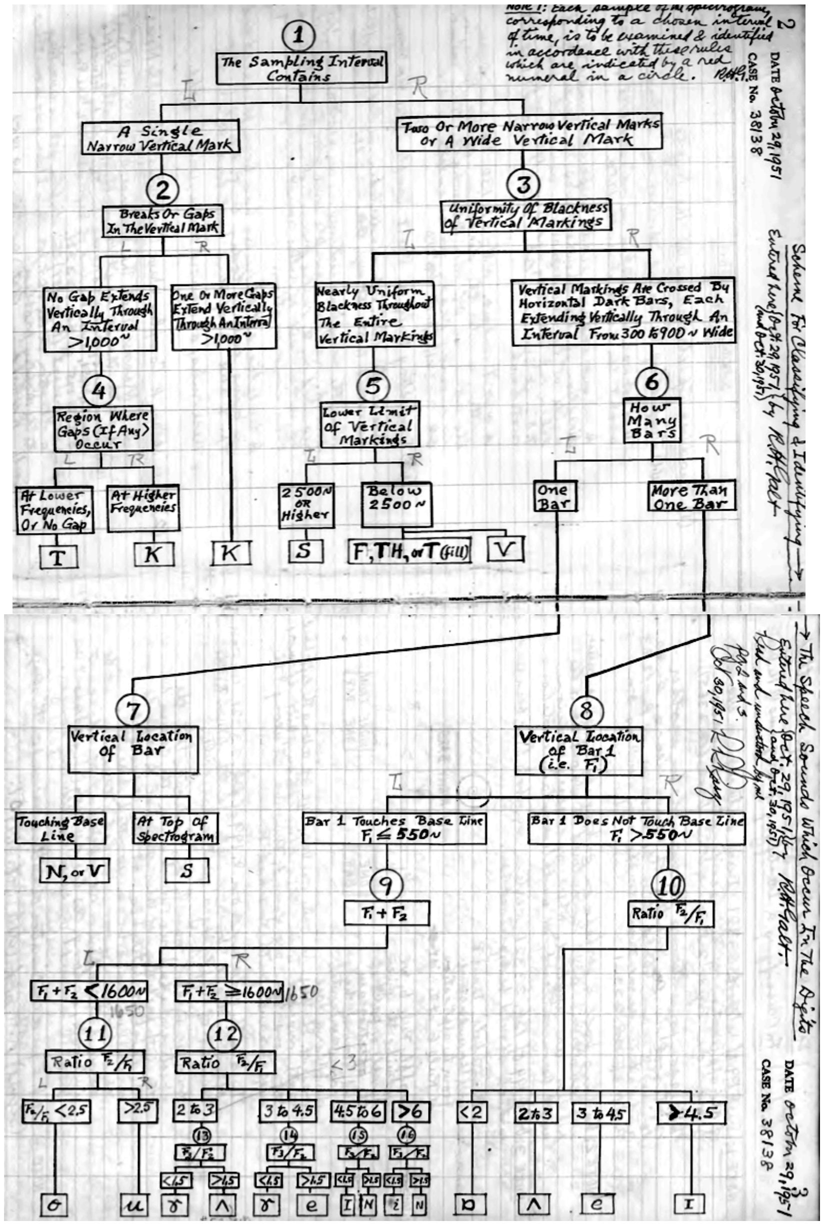
Multiplication
with signal

Convolution
with signal spectrum

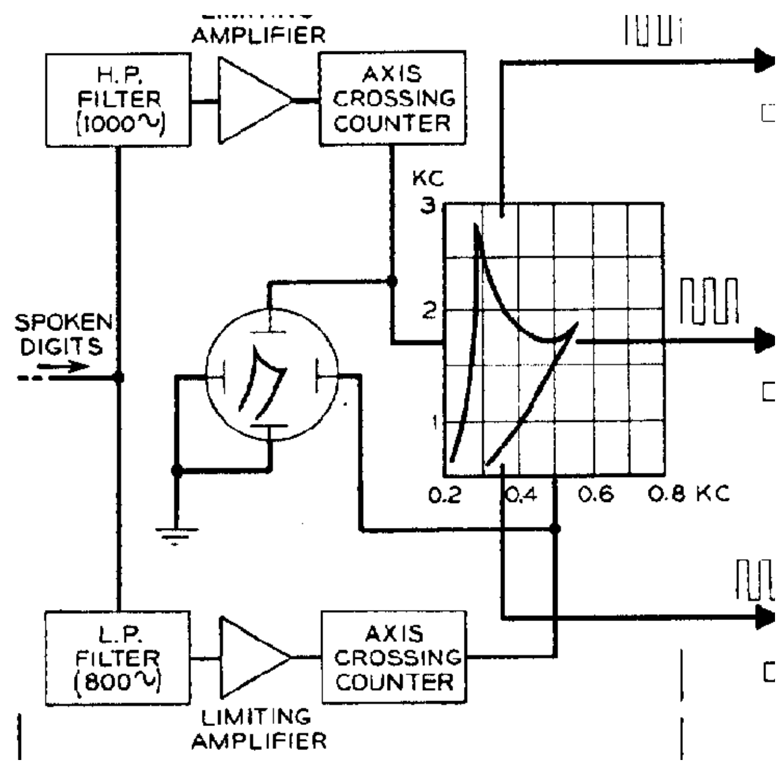


t_0
time
ΔT
fourier
transform
$s(f,t_0)$
spectrum
of the short
segment
frequency
time

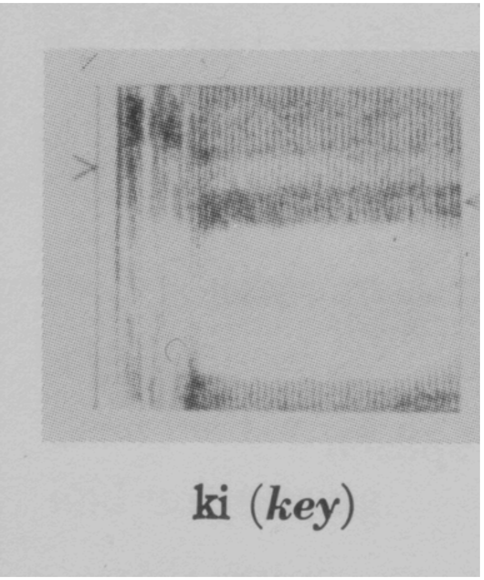
Concept of the first “real” automatic speech recognizer(R.H. Galt 1951)



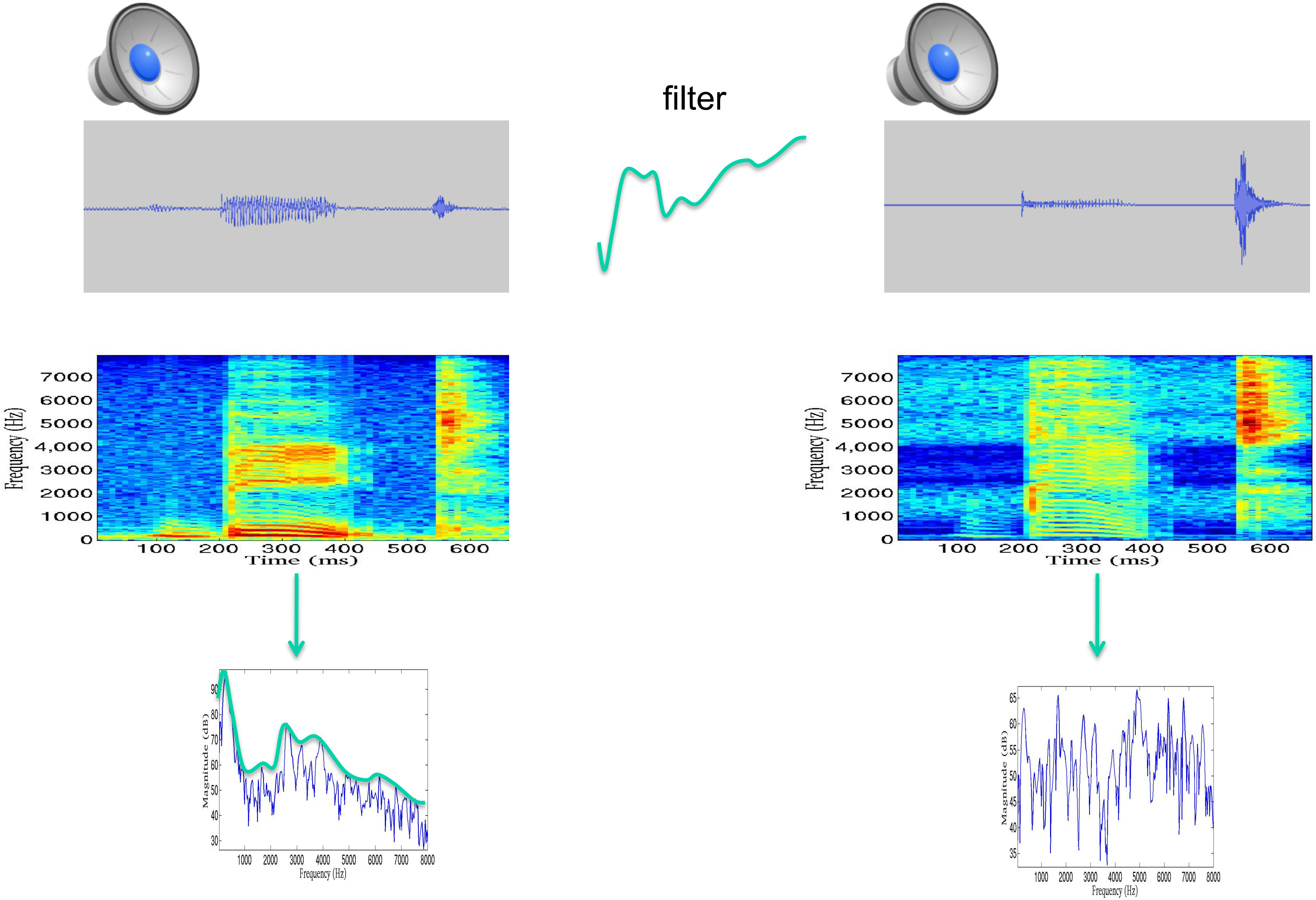
First “real” recognizer ever build (Davis, Biddulph, Balashek 1952)



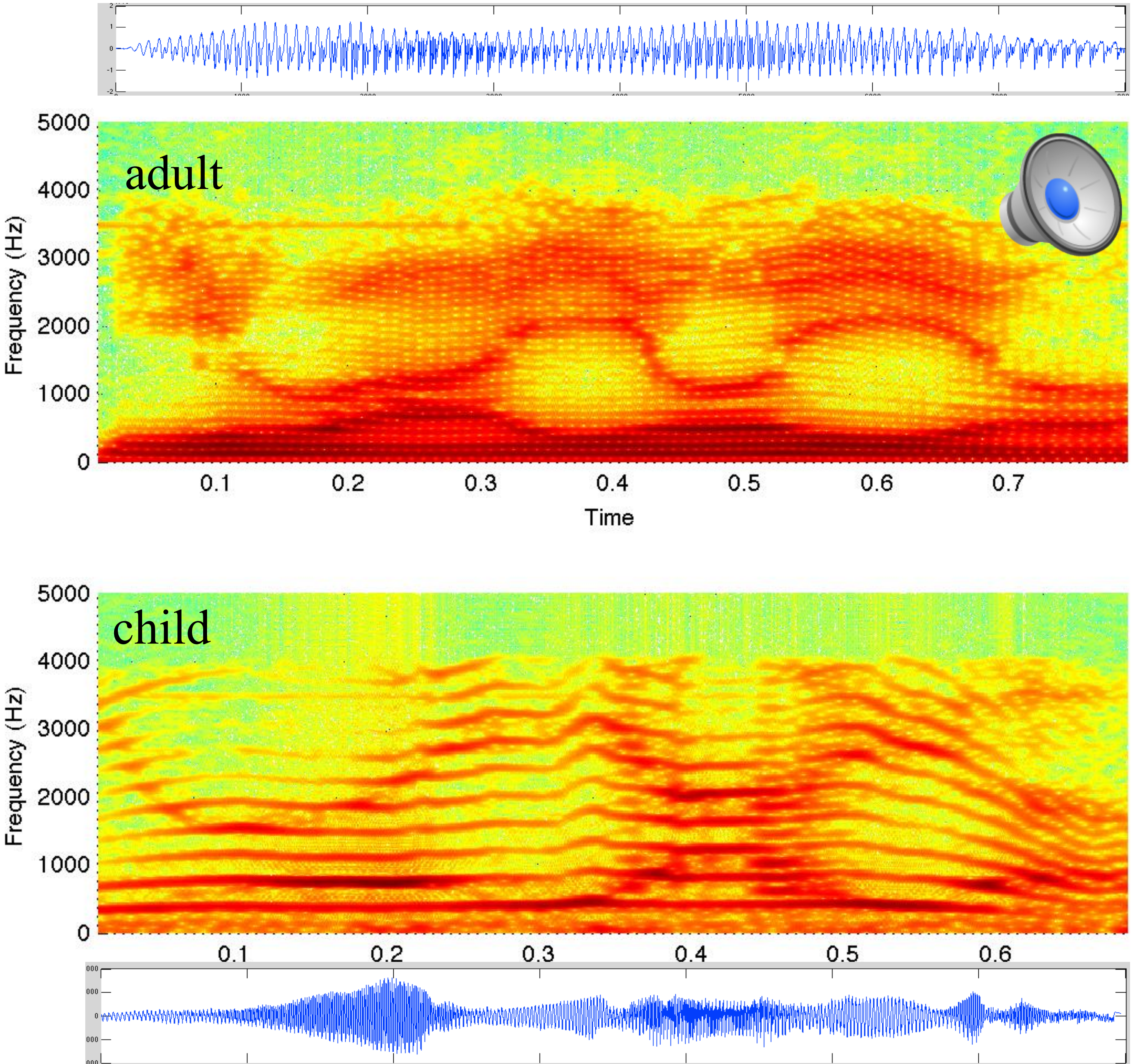
/k/



Potter, Kopp, and Green, Visible Speech 1947



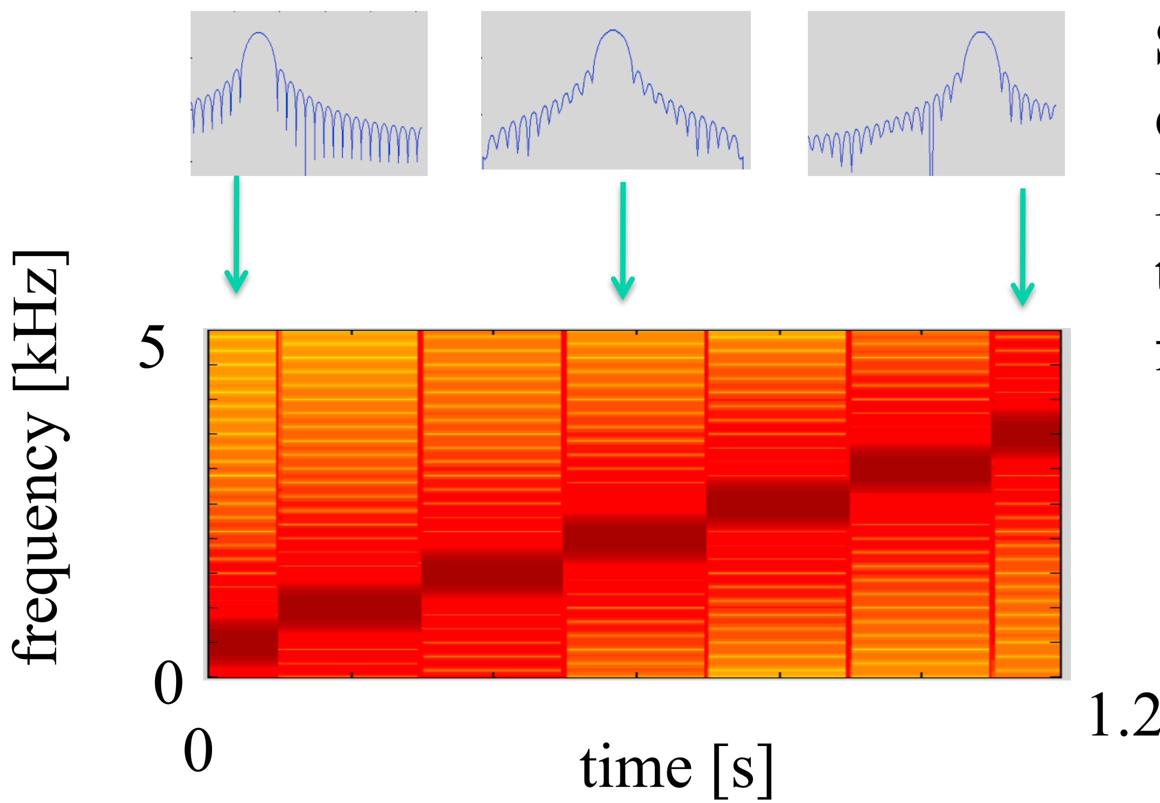
filter
Frequency (Hz)
7000
6000
5000
4,000
3000
2000
1000
0
100
200
300
400
500
600
Time (ms)
Magnitude (dB)
Frequency (Hz)


adult
Frequency (Hz)
5000
4000
3000
2000
1000
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Time
child
Frequency (Hz)
5000
4000
3000
2000
1000
0
0.1
0.2
0.3
0.4
0.5
0.6

$$S_n(e^{j\omega}) = \sum_{m=-\infty}^{\infty} s(m) \cdot w(n-m) e^{-jm\omega}$$

Fourier transform of the signal *s(m)* multiplied by the window *w(n-m)*

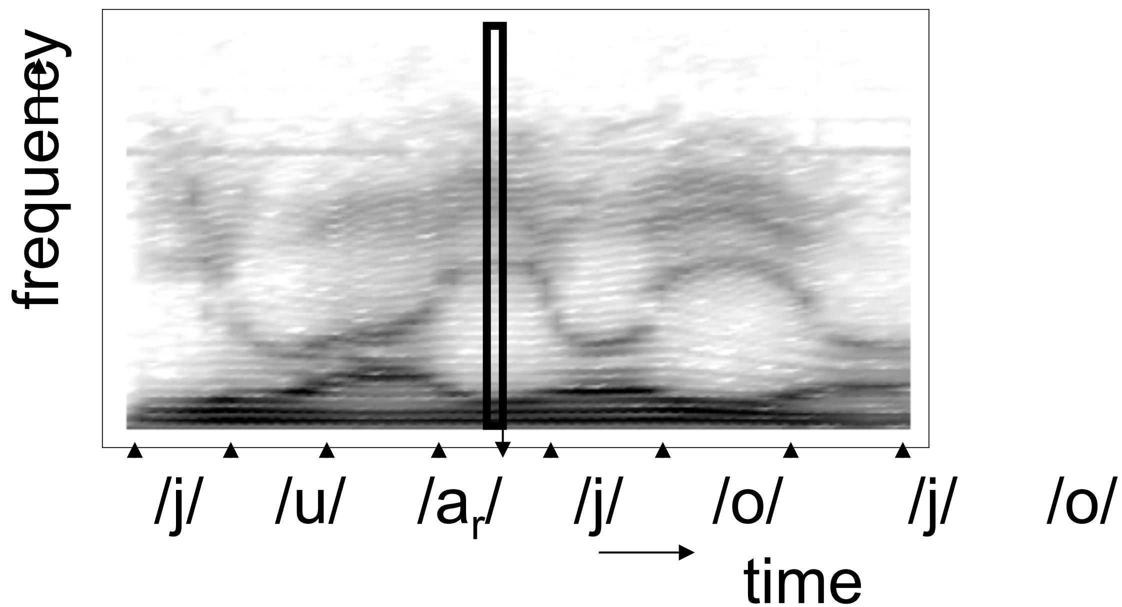
Spectrum is the line spectrum of the signal convolved with the spectrum of the window



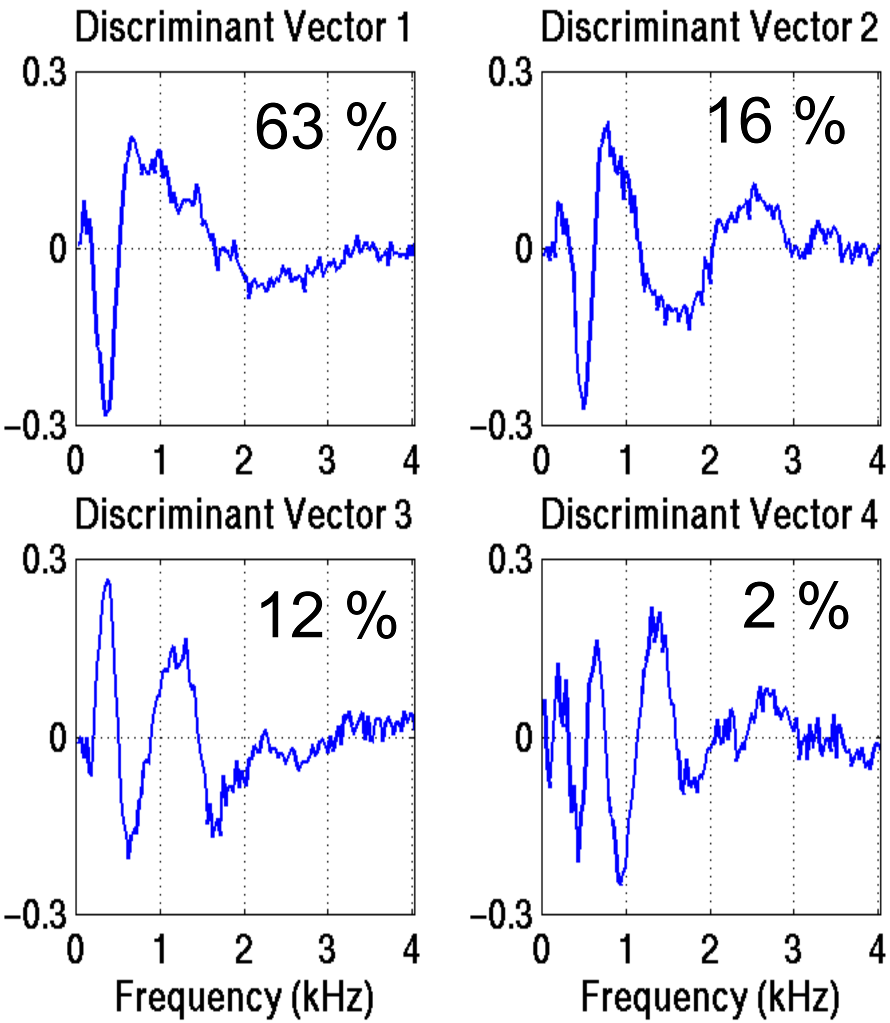
Spectral resolution of the short-term Fourier analysis is the same at all frequencies.

Spectral Basis from LDA

LDA gives basis for projection of spectral space



LDA vectors from Fourier Spectrum
(OGI 3 hour stories hand-labeled database)



- Spectral resolution of LDA-derived spectral basis is higher at low frequencies

Psychophysics:

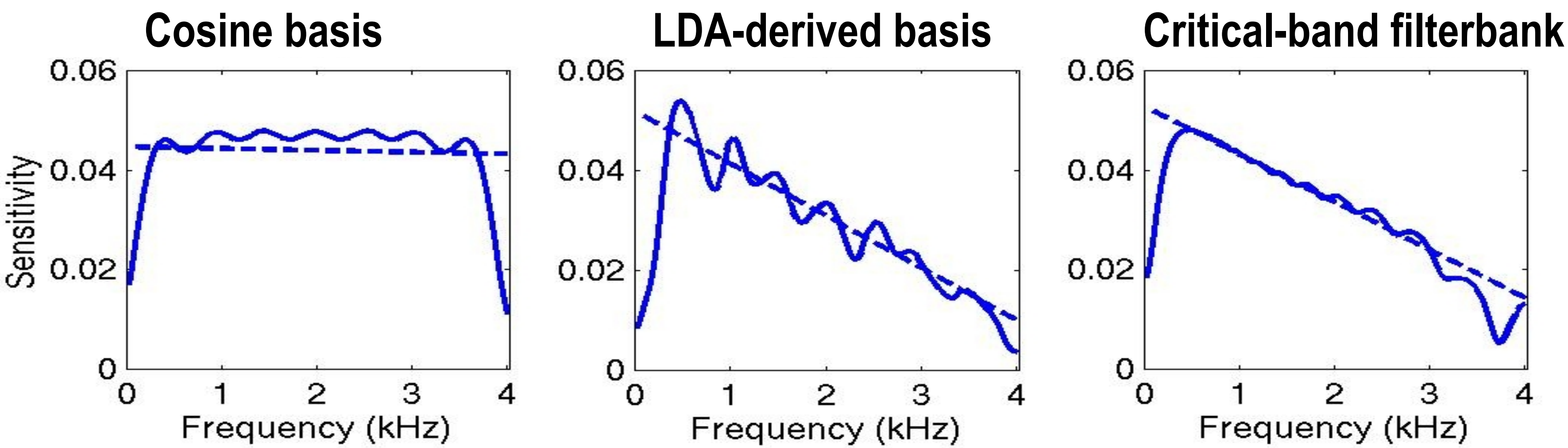
Critical bands of human hearing are broader at higher frequencies

Physiology:

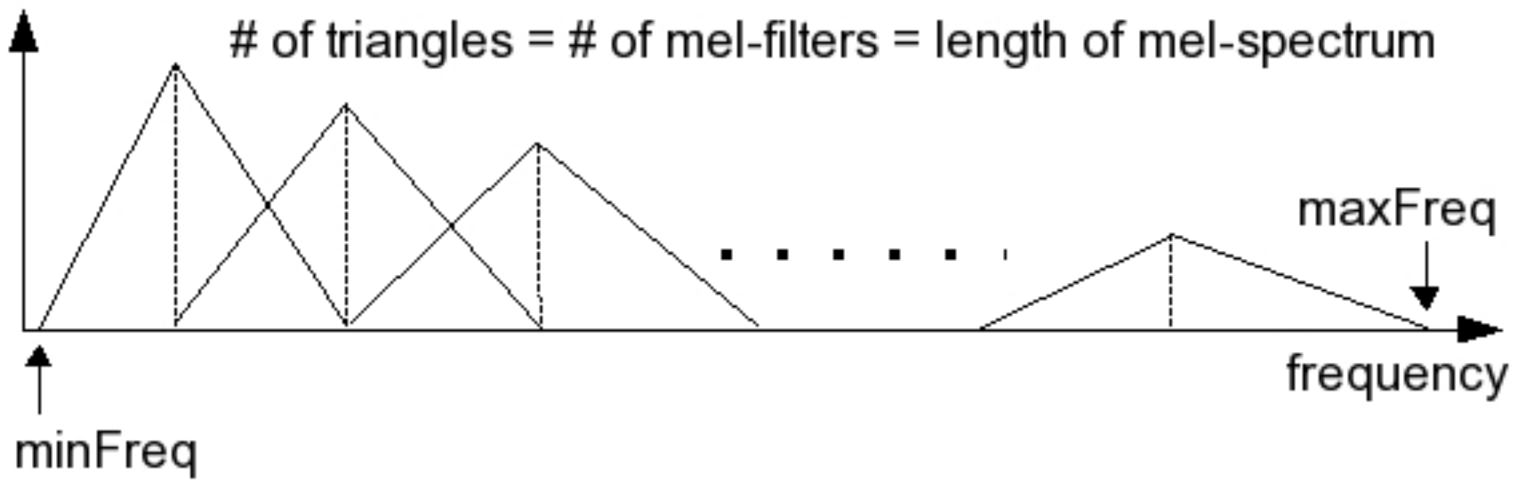
Position of maximum of traveling wave on basilar membrane is proportional to logarithm of frequency

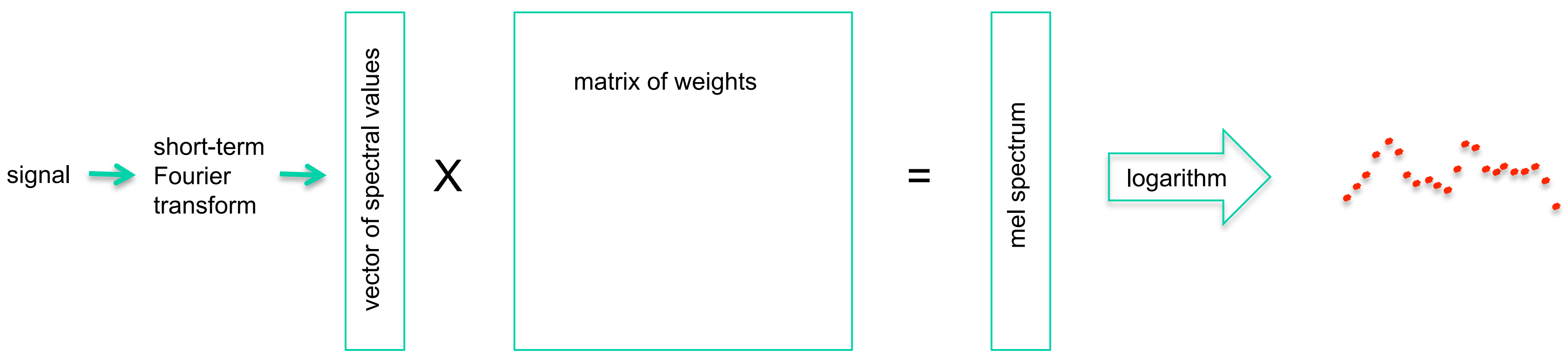
Sensitivity to Spectral Change

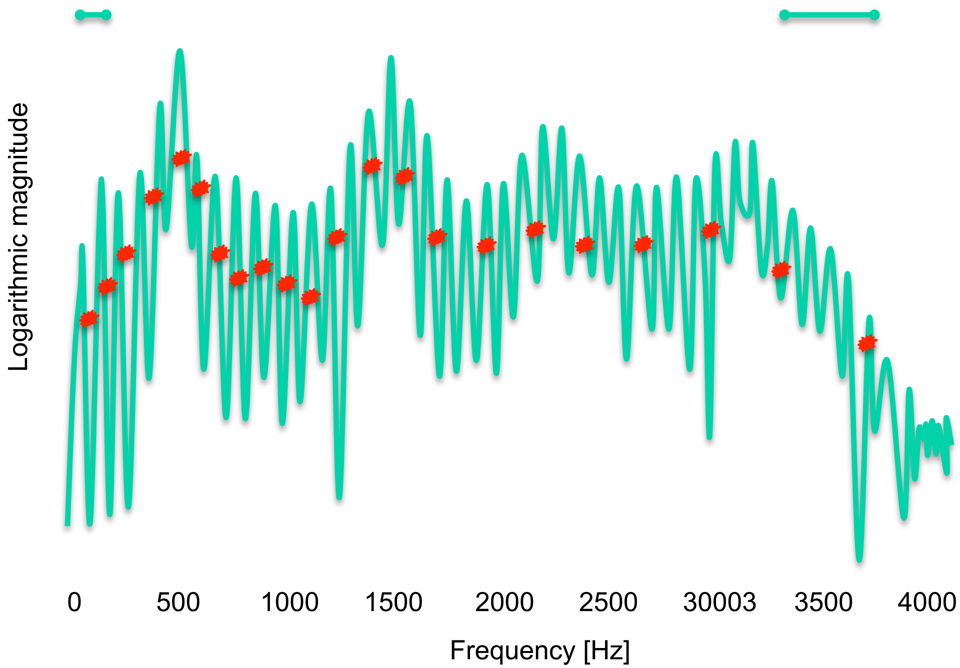
(Malayath 1999)



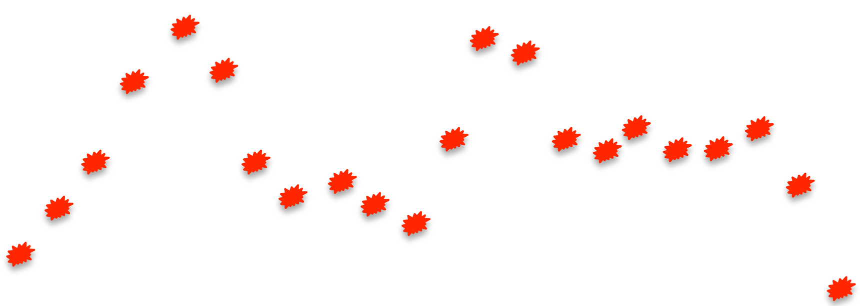
Spectral weights





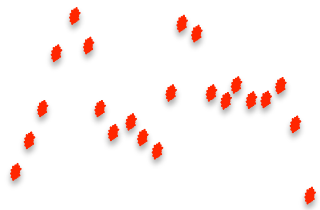


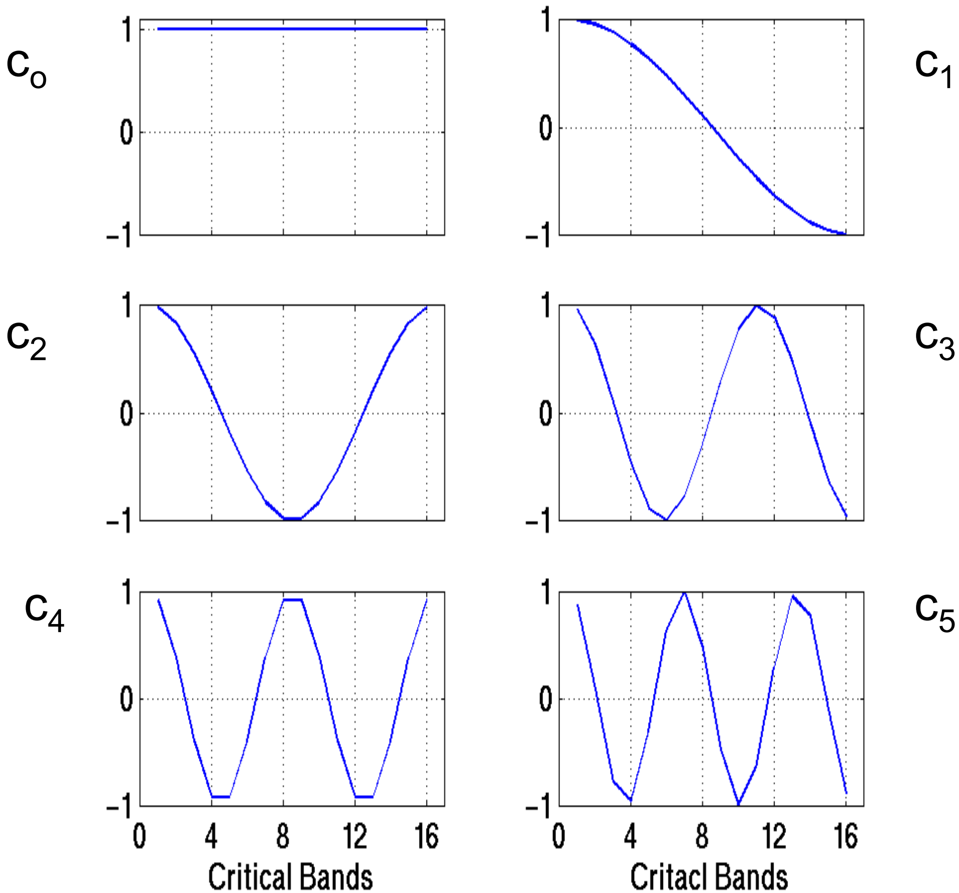
- Spectral resolution decreases with frequency.
- Temporal resolution stays the same (given by the length of the analysis window in computing spectrum)



cosine transform

Mel cepstrum

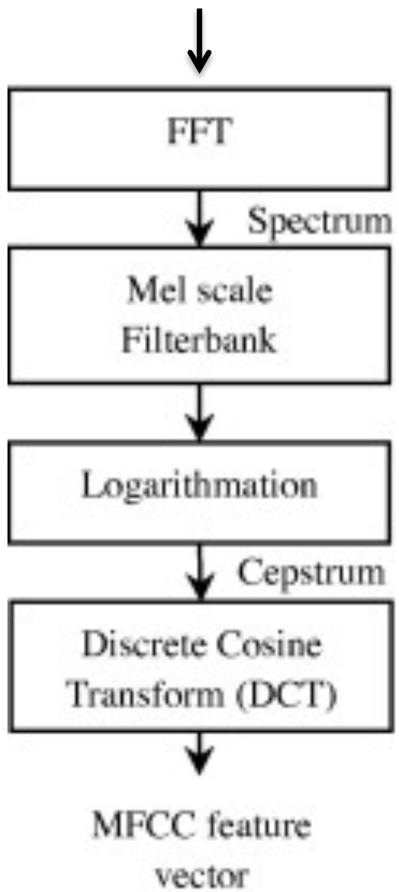




c_o
c_1
c_2
c_3
c_4
c_5
1
0
-1
0
4
8
12
16
Critical Bands
Critacl Bands

Mel cepstrum

Segment of signal (~ 20 ms – windowed)



Short-term spectrum

- Frequency selectivity of hearing

Project on spectral weights

- Non-equal spectral resolution of hearing

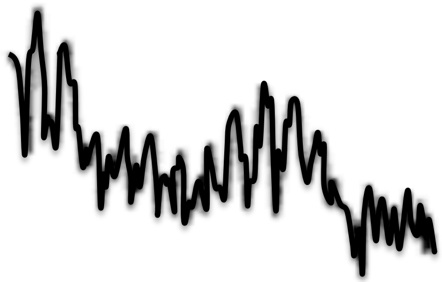
Take logarithm

- make distribution more Normal

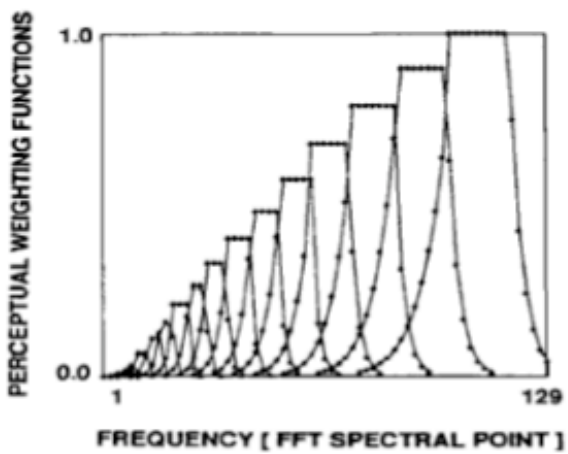
Cosine transform

- de-correlate

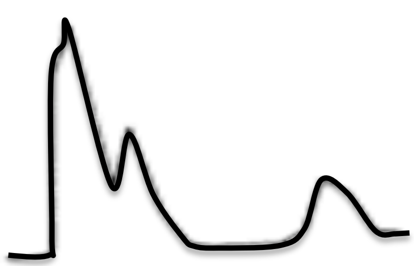
Perceptual Linear Prediction



spectrum

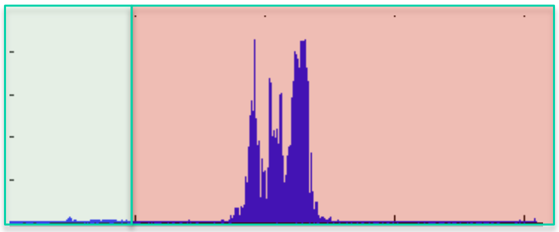


summation windows

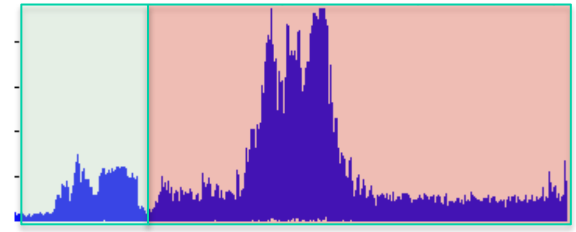


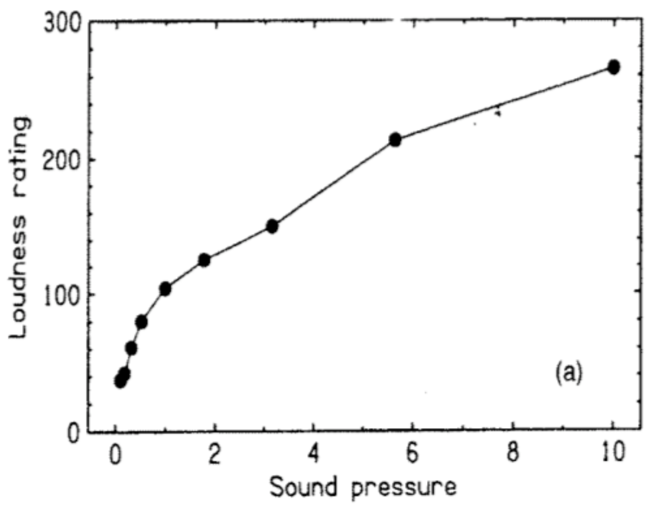
spectrum with auditory-like resolution

intensity ≈ signal ^{2} [w/m^{2}]

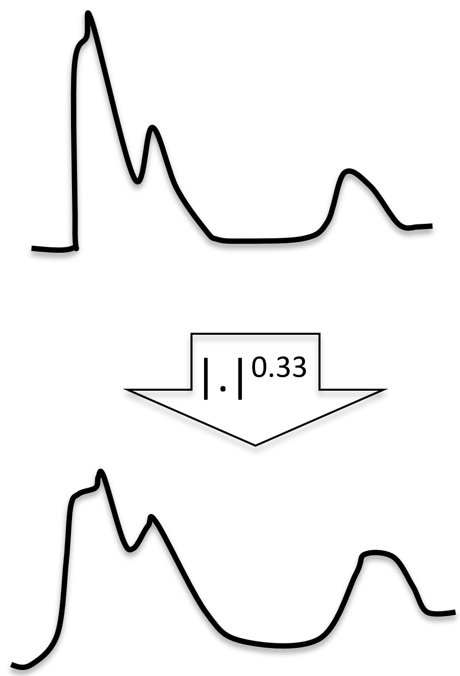


loudness [Sones]





$$\text{loudness} = \text{intensity}^{\,\mathbf{0.33}}$$



intensity
(power spectrum)

loudness

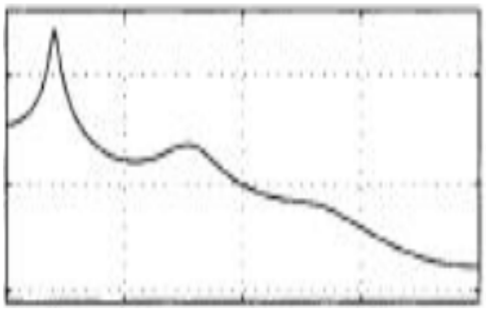
Not all spectral details are important

a) compute Fourier transform of the auditory spectrum and truncate it (cepstrum)

b) approximate the auditory spectrum by an autoregressive model

6th order AR model

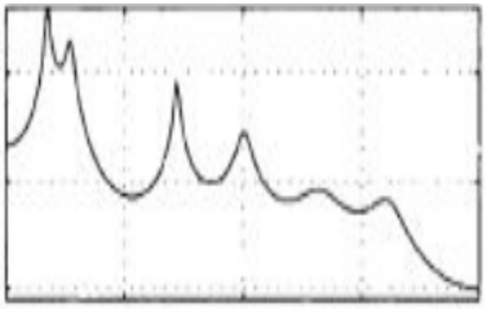
power (loudness)



frequency (tonality)

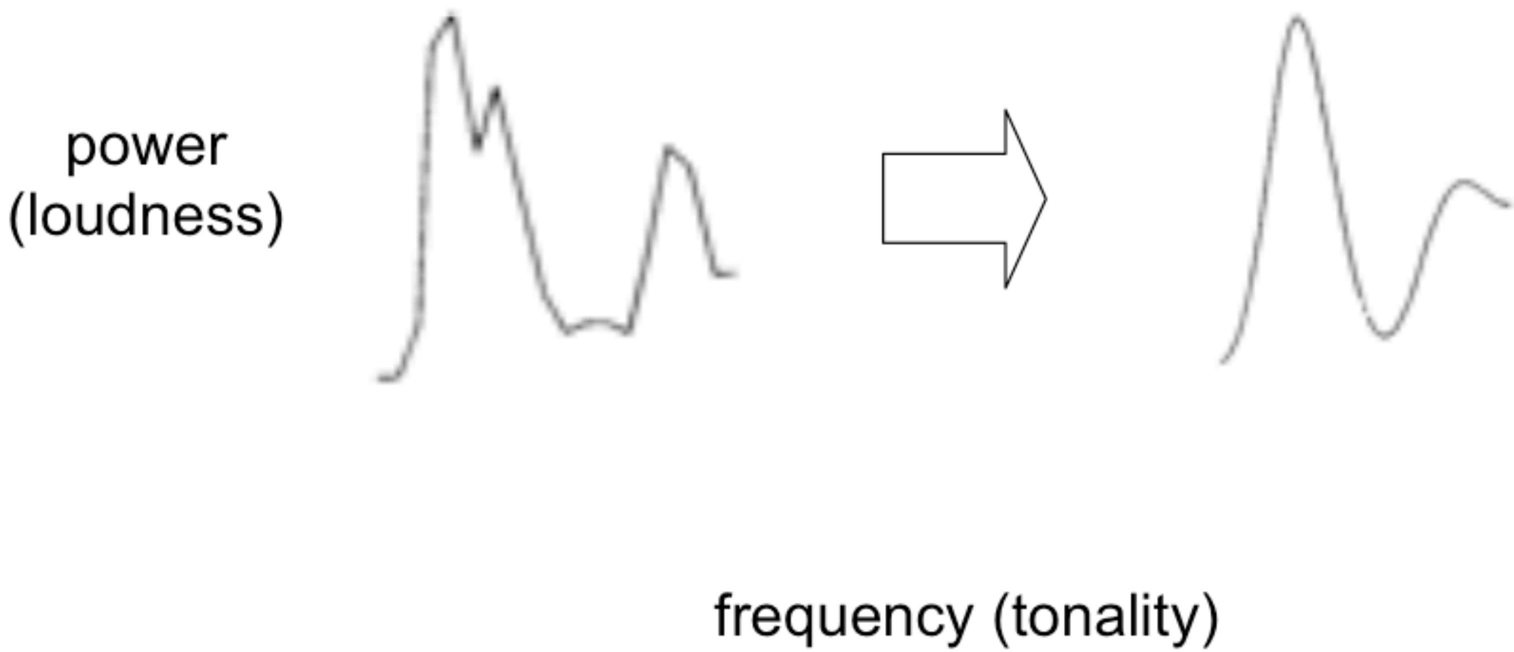
14th order AR model

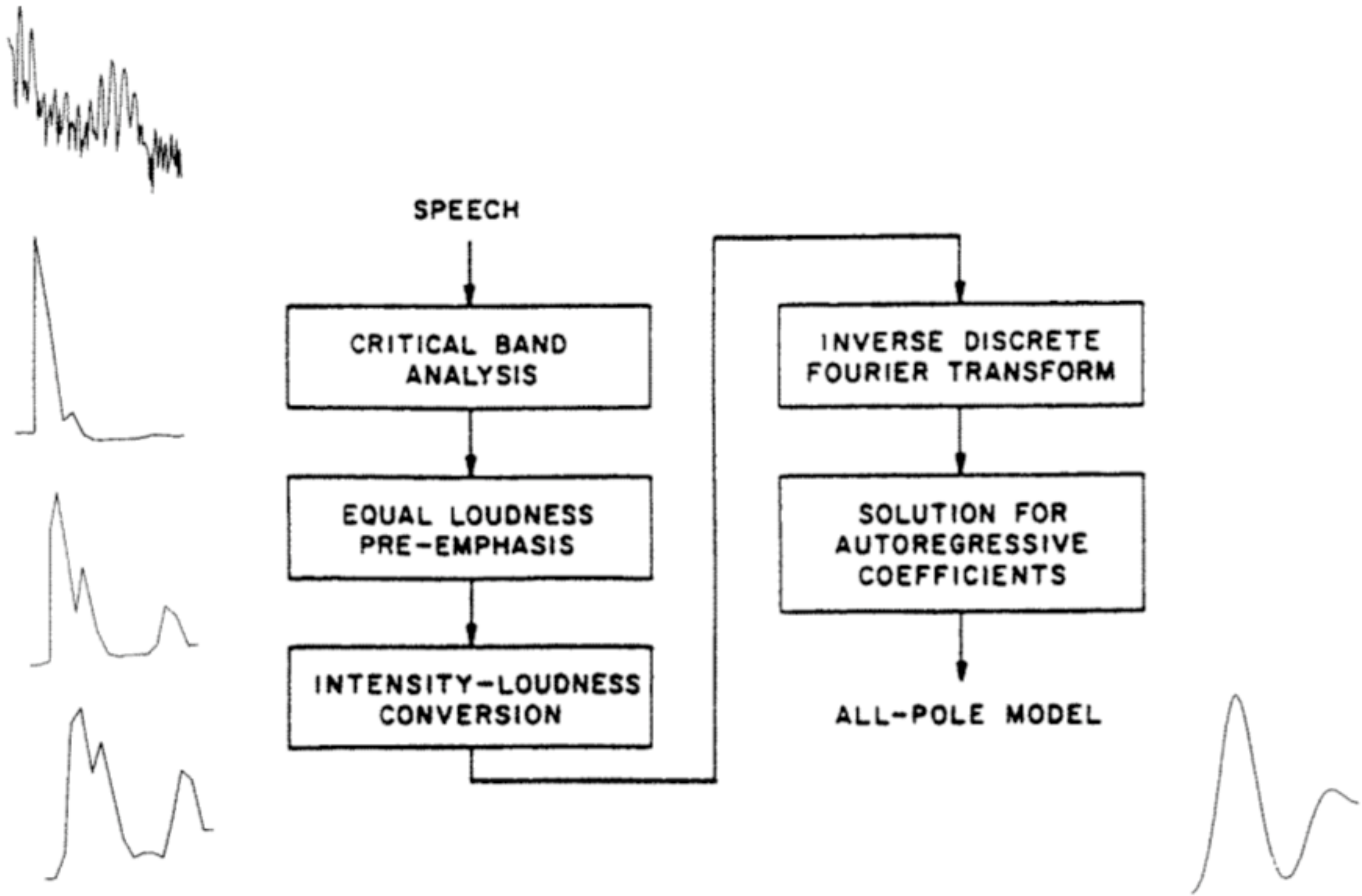
power (loudness)



frequency (tonality)

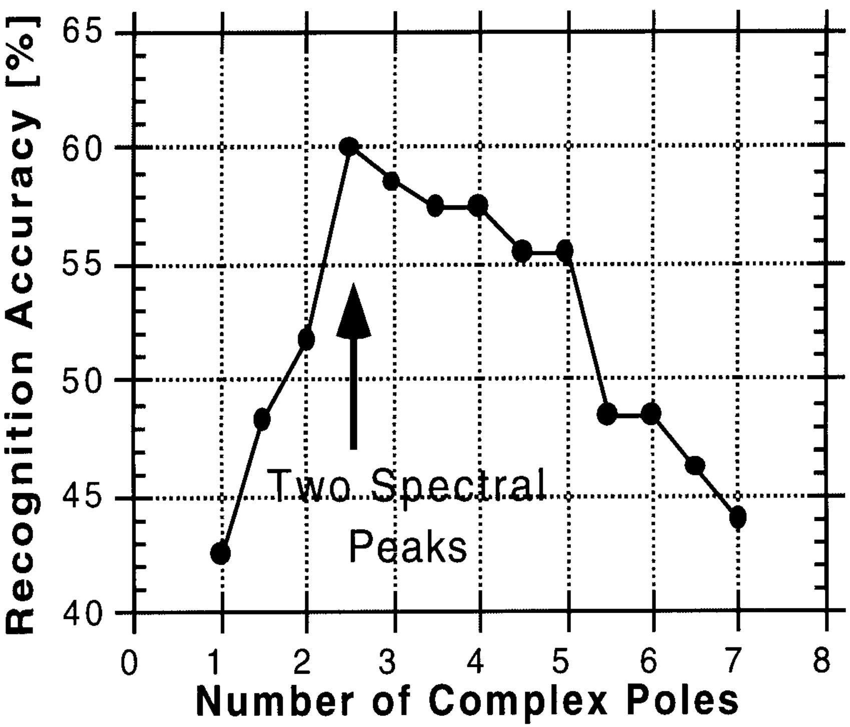
Perceptual Linear Prediction (PLP) Autoregressive fit to the auditory-like spectrum





SPEECH
CRITICAL BAND ANALYSIS
EQUAL LOUDNESS PRE-EMPHASIS
INTENSITY-LOUDNESS CONVERSION
INVERSE DISCRETE FOURIER TRANSFORM
SOLUTION FOR AUTOREGRESSIVE COEFFICIENTS
ALL-POLE MODEL

Optimal Amount of Spectral Smoothing (order of PLP autoregressive model)



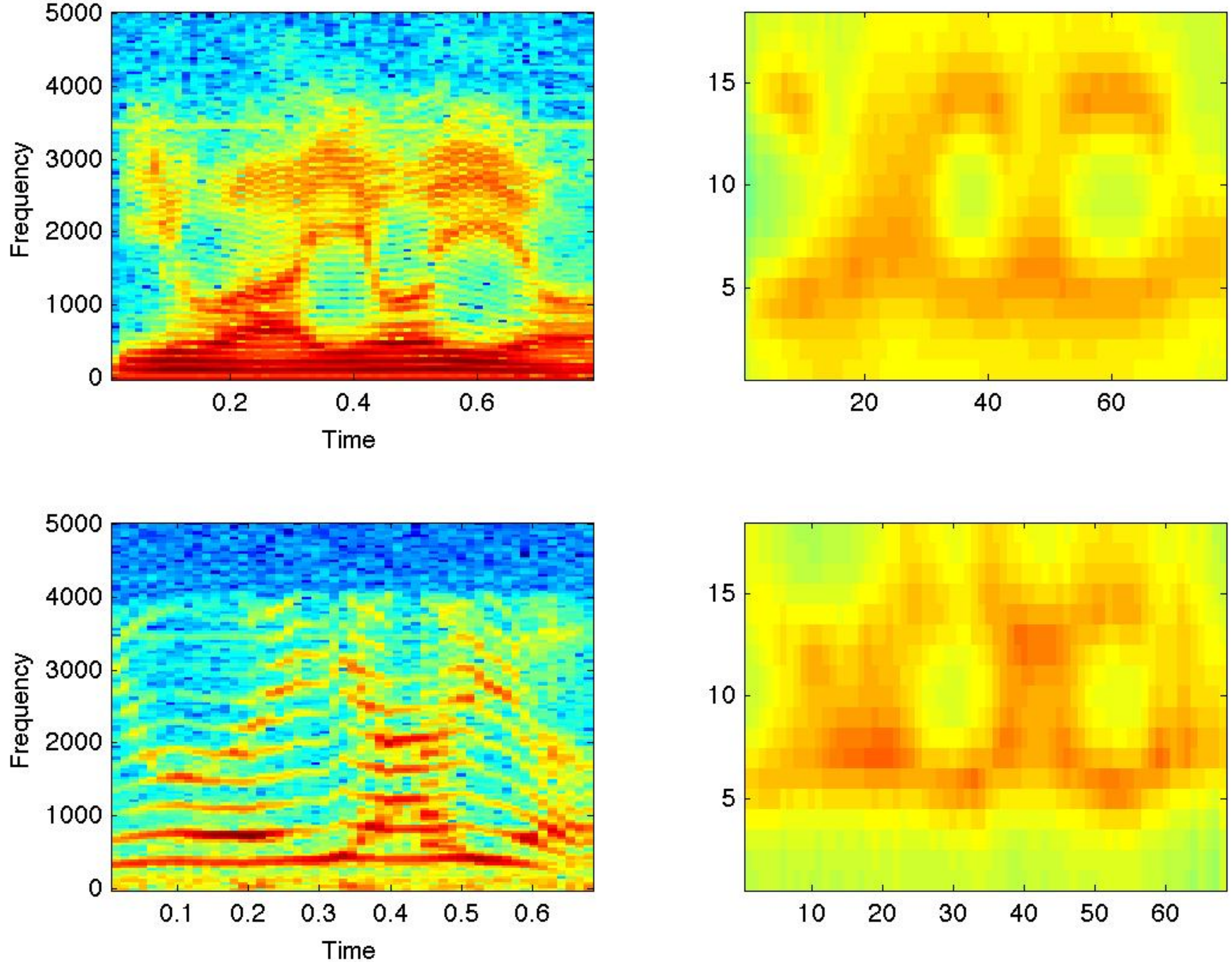
- cross-speaker ASR (trained on one speaker and tested on another)
- all speaker-dependent information harmful

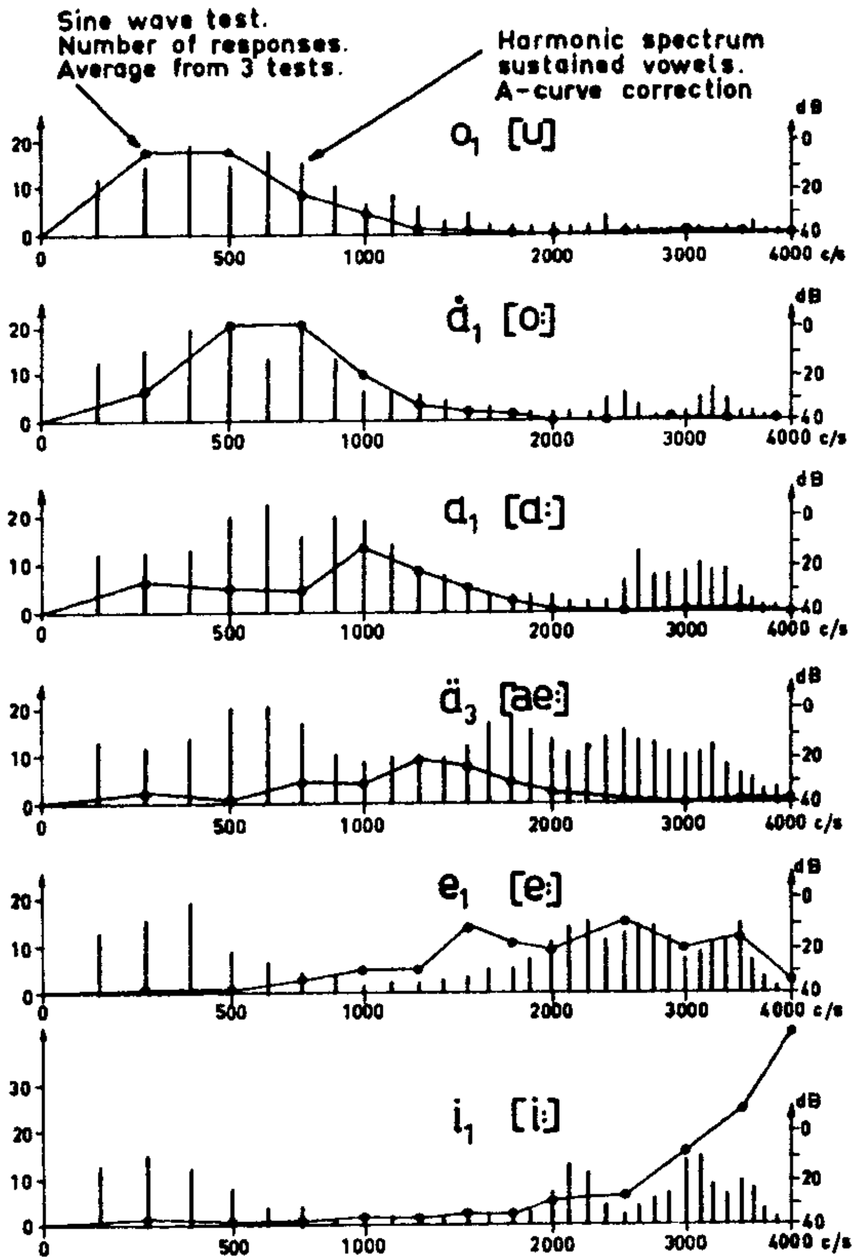
short-term spectrum

5th order PLP spectrum

adult male

4 year old child

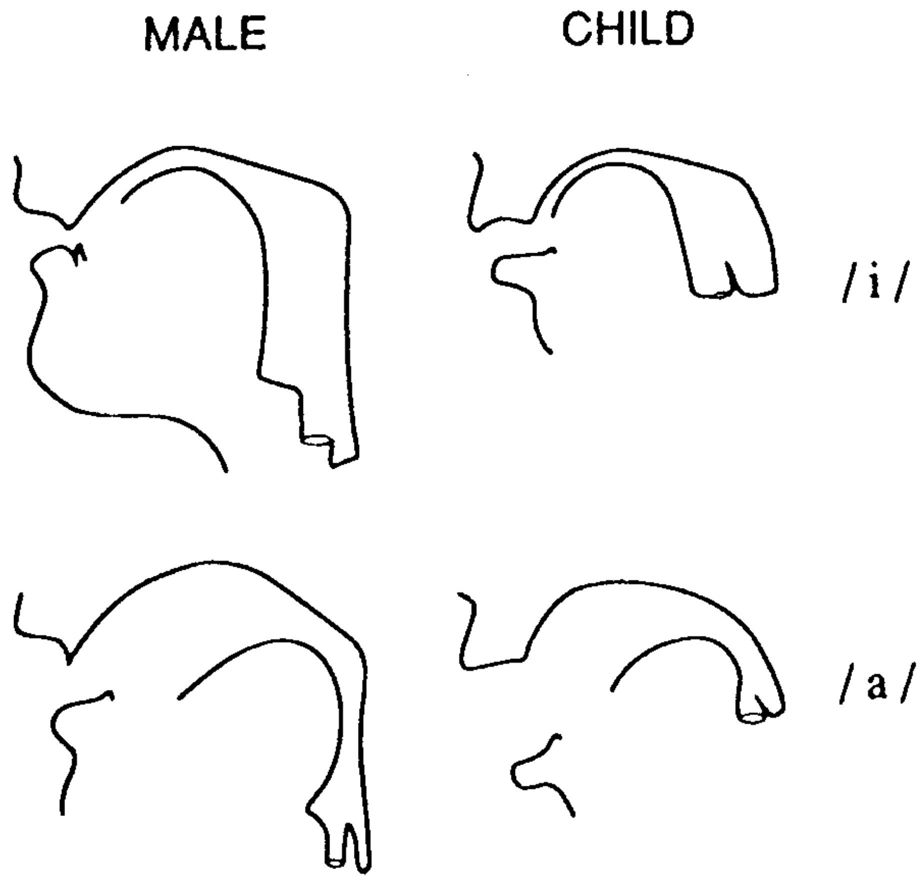




- affiliate vowel with sine wave tone (forced judgment)
- peak of histograms would correspond to resonance frequency of uncoupled front cavity in production of a given vowel
  - Fant 1947
- Perceptual F2'
  - position of second peak in two-peak simulation of vowels

X-rays of Male and Child Vocal Tract in Production of Vowels

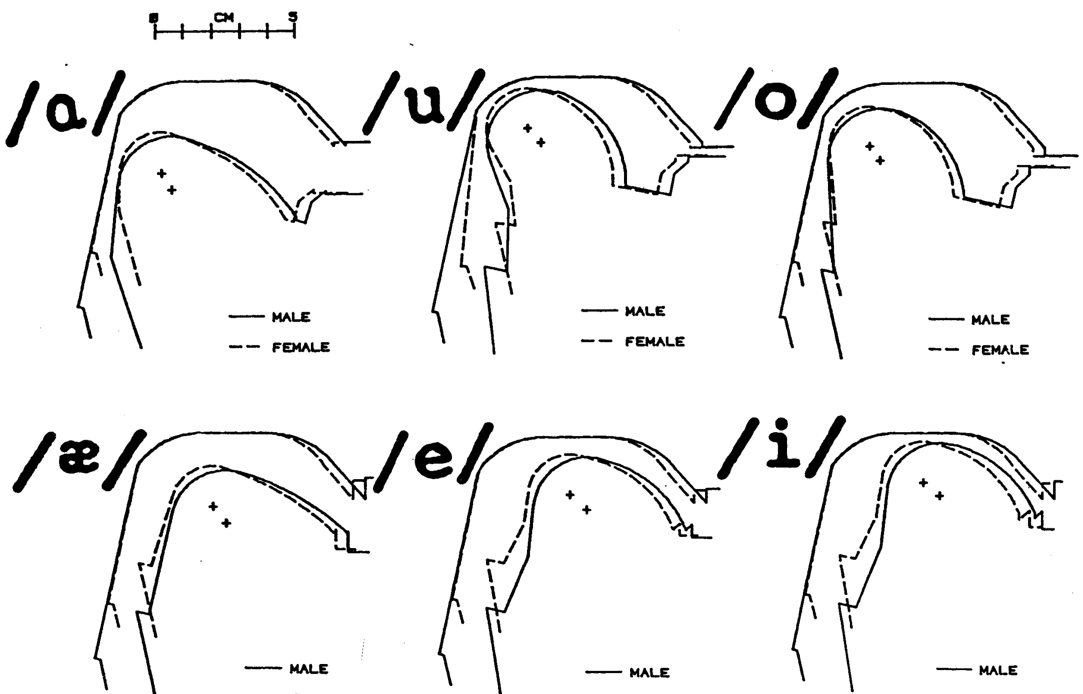
- In production of vowels, the front part of the vocal tract appears to be less speaker dependent than its back part
  - Hermansky and Broad 1990



Female vocal tract from male

Ursula Goldstein, MIT PhD. Thesis 1980

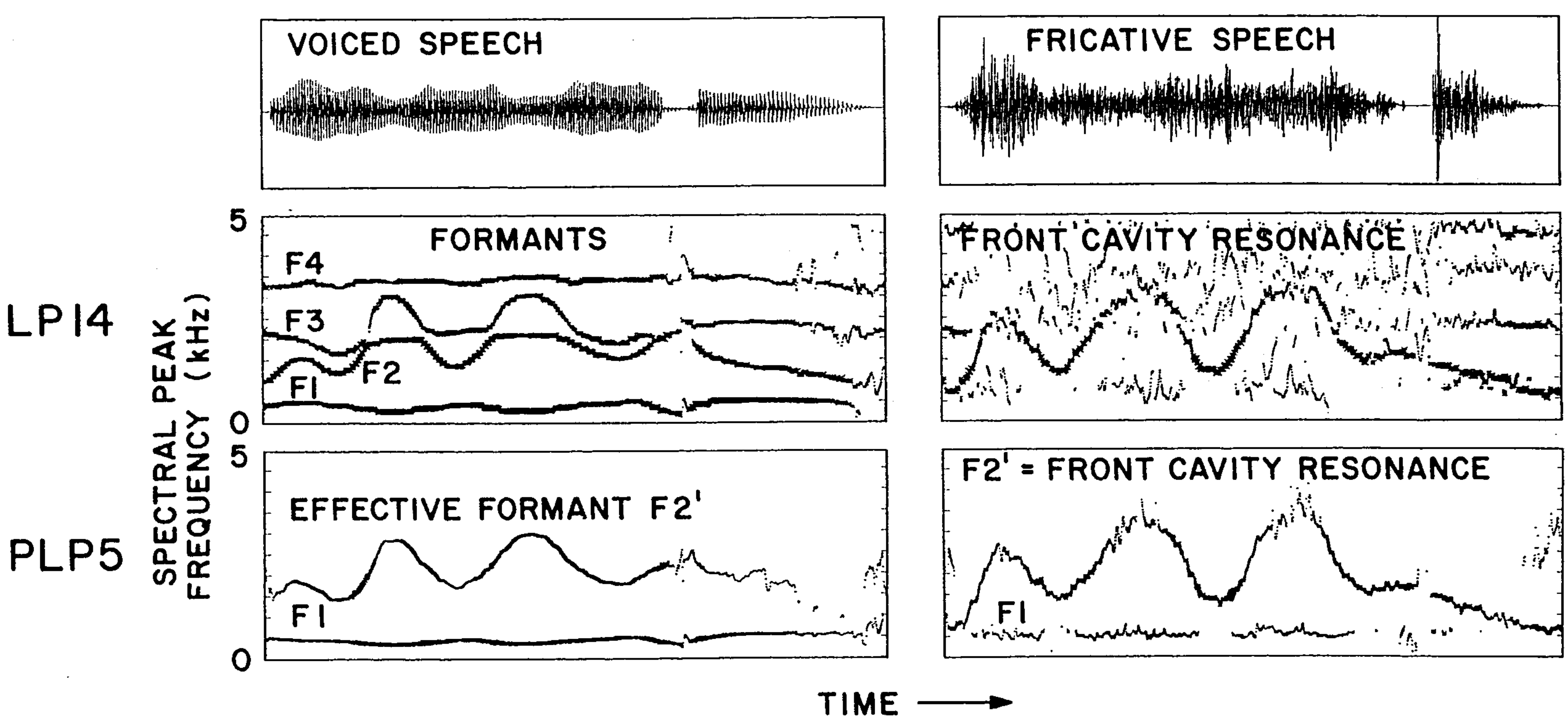
- Start with male vocal tract x-ray
- implement male-female anatomical differences
- change “resting dimensions” to “female”



Front Cavity - F2' Hypothesis

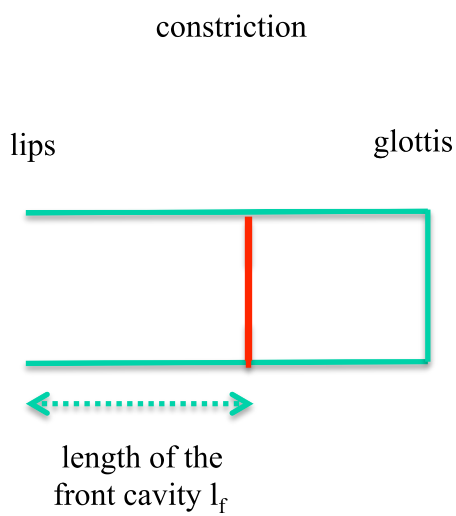
- F2' correlates with resonance frequency of decoupled front cavity of vocal tract in production of vowels
  - Fant 1960
- Front part of the vocal tract
  - grows relatively little during lifetime
  - is easy to manipulate without special training
  - for many consonants, the front part dominance is well accepted

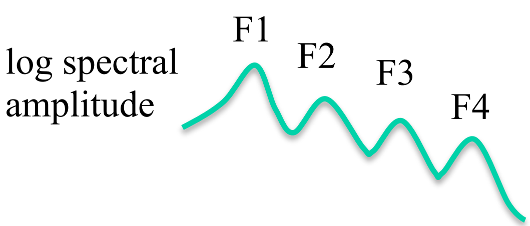
Voiced and fricative speech



PLP-estimated F2' and Front Cavity Resonance Frequency

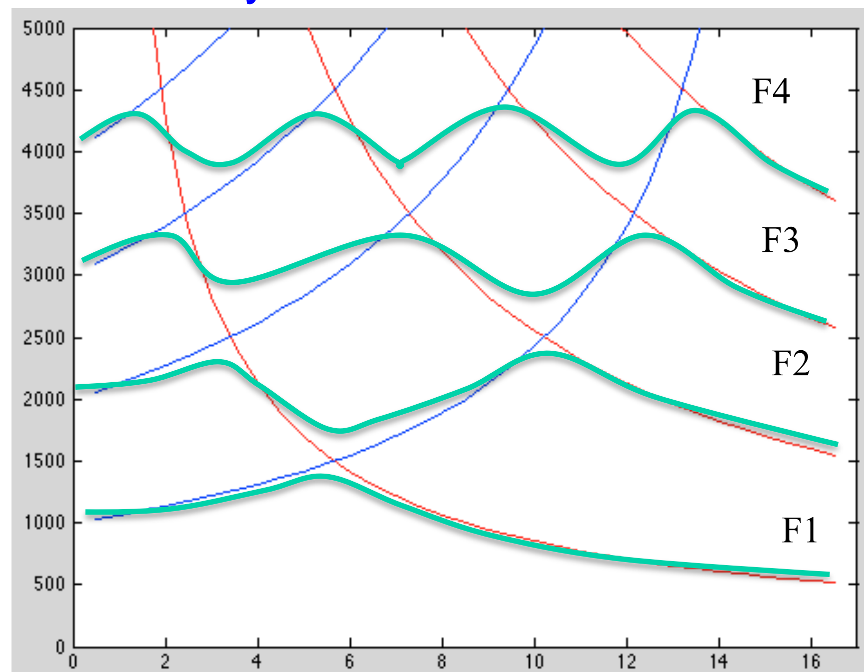
- Articulatory Synthesis
  - formants known
  - resonance frequency of decoupled front cavity can be computed
  - synthetic speech is available for analysis by PLP (F2' can be estimated)





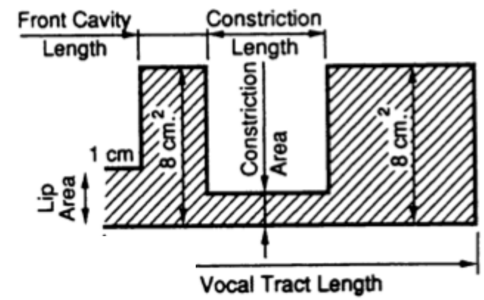
"quarter wave resonator"
F1 = 500 Hz, F2=1500 Hz, f3=2500 Hz,…
if the length = 17 cm and c = m/s

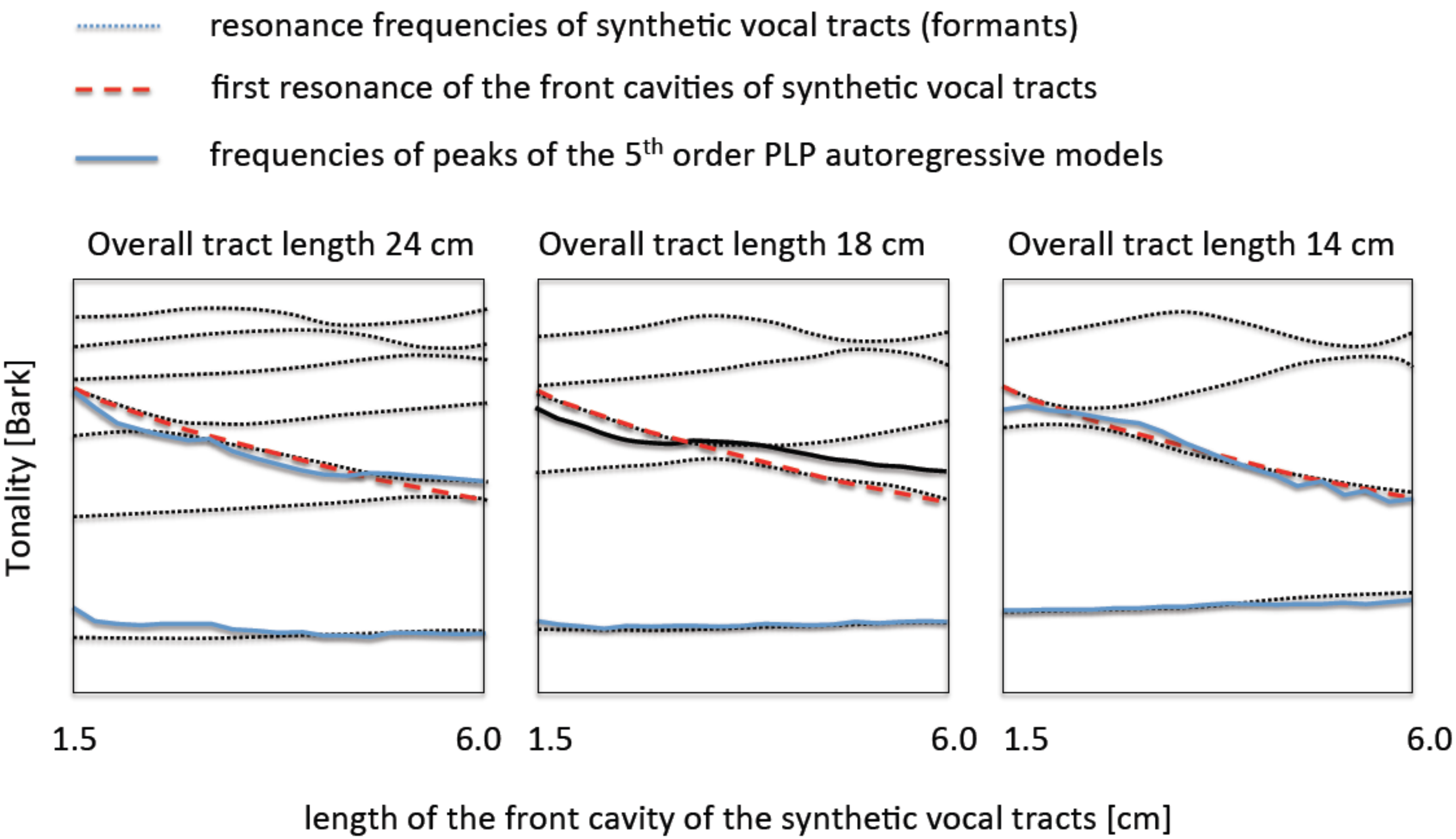
front cavity resonance modes
back cavity resonance modes



length of the
front cavity l_f

Front Cavity Resonance Experiment Using Articulatory Synthesis



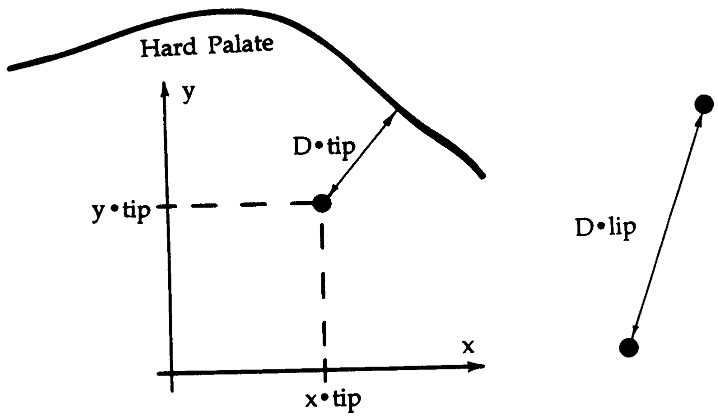


Result of Experiment with Synthetic Vowels

- correlations on about 11 000 synthetic front vowels
  - (back vowels for which PLP formed only one peak were excluded)
  - tract length varied between 14 and 24 cm

	tract length	front cavity resonance
Second peak of PLP model	**-0.18**	**0.9**
formants (averaged)	-0.71	0.22

X-ray Microbeam Experiment (Broad and Hermansky 1989)



$$L = k1 - \alpha \mathbf{x}$$

(a) $$\mathbf{x} = x_{tip} \cos\theta + y_{tip} \sin\theta$$

$$\Phi = k2 + b1 \ln D_{tip} + b2 \ln D_{lip}$$

(b) $$\boxed{\frac{1}{F2'} = \frac{4L}{c} \frac{2}{2 + \Phi}}$$

(c) PARAMETERS:

k1, k2, α, θ, b1, b2

- Shape approximated by cosine with period of 2**L** and amplitude Φ
- Resonance frequency given by **L** and $\mathbf{\Phi}$ (Schroeder, Mermelstein)

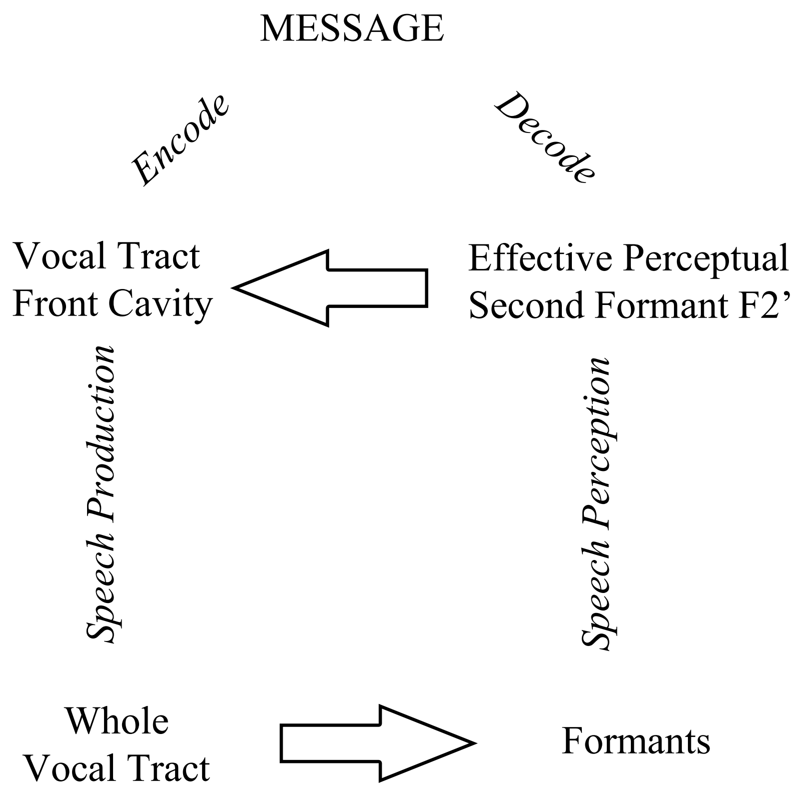
Results of X-Ray Microbeam Experiment

- two male speakers
  - “where were you a year” three times each
- front cavity resonance from articulations
- PLP-estimated F2’ from acoustic data

CORRELATION BETWEEN RESONANCE FREQUENCY
OF FRONT CAVITY AND PLP-DERIVED F2’

speaker 1	correlation 0.95
speaker 2	correlation 0.96

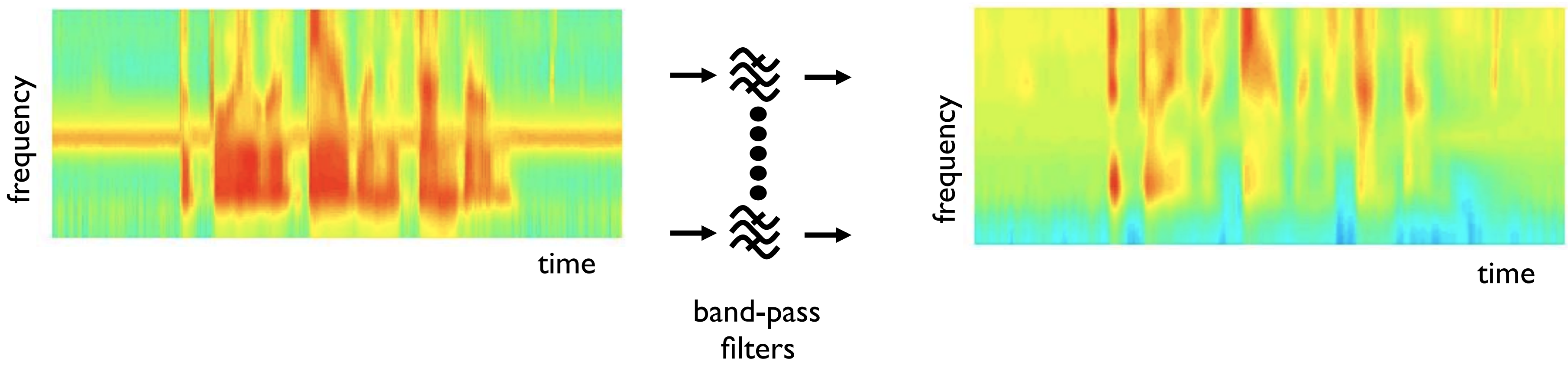
Front Cavity - F2' Hypothesis

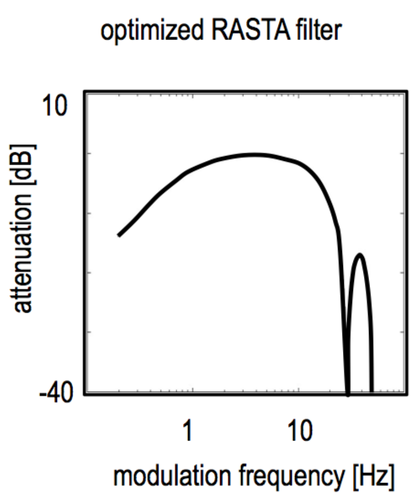


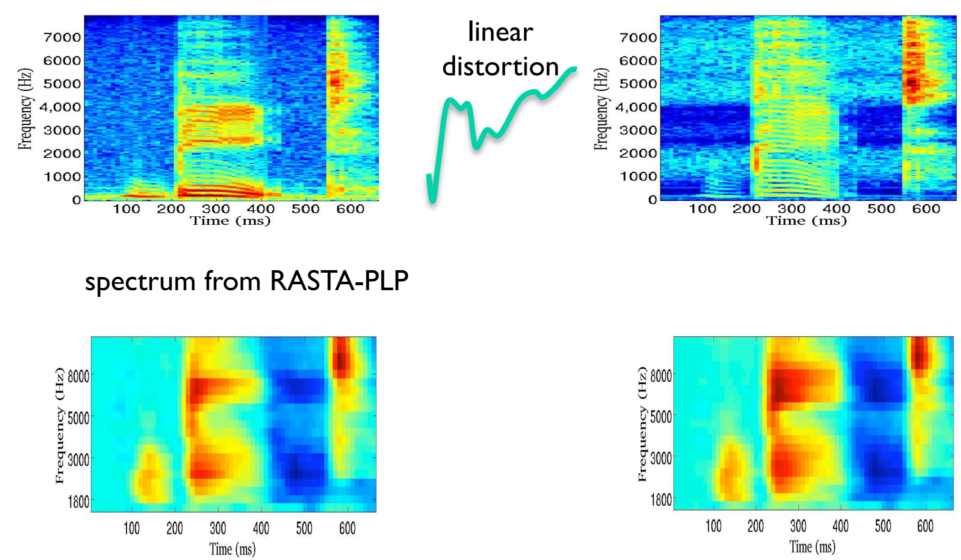
- **Our limited experimental data do not contradict the hypothesis**

RASTA processing

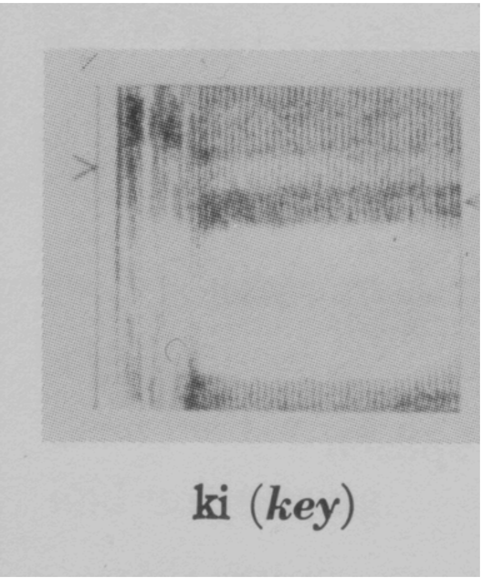
Hermansky and Morgan 1990







/k/



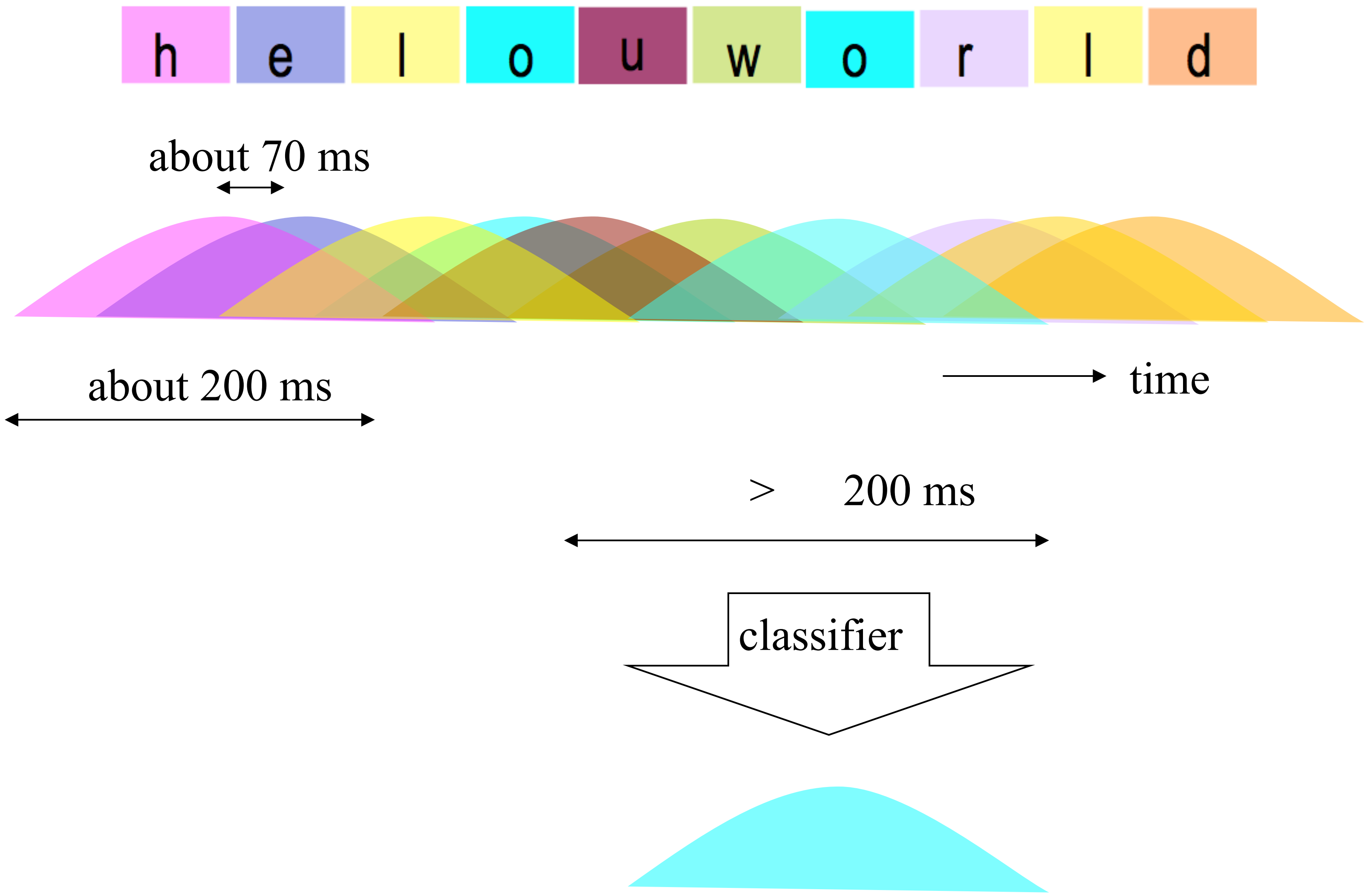
Potter, Kopp, and Green, Visible Speech 1947

need to know the following vowel before identifying the consonant ?

recognize whole syllables ?

recognize phonemes but use information from syllable-length segments of the signal !

- V. A. Kozhevnikov and L. A. Chistovich, Speech: Articulation and perception. 1967


h
e
l
o
u
w
o
r
l
d
about 70 ms
about 200 ms
time
> 200 ms
classifier

spectro temporal features

serial

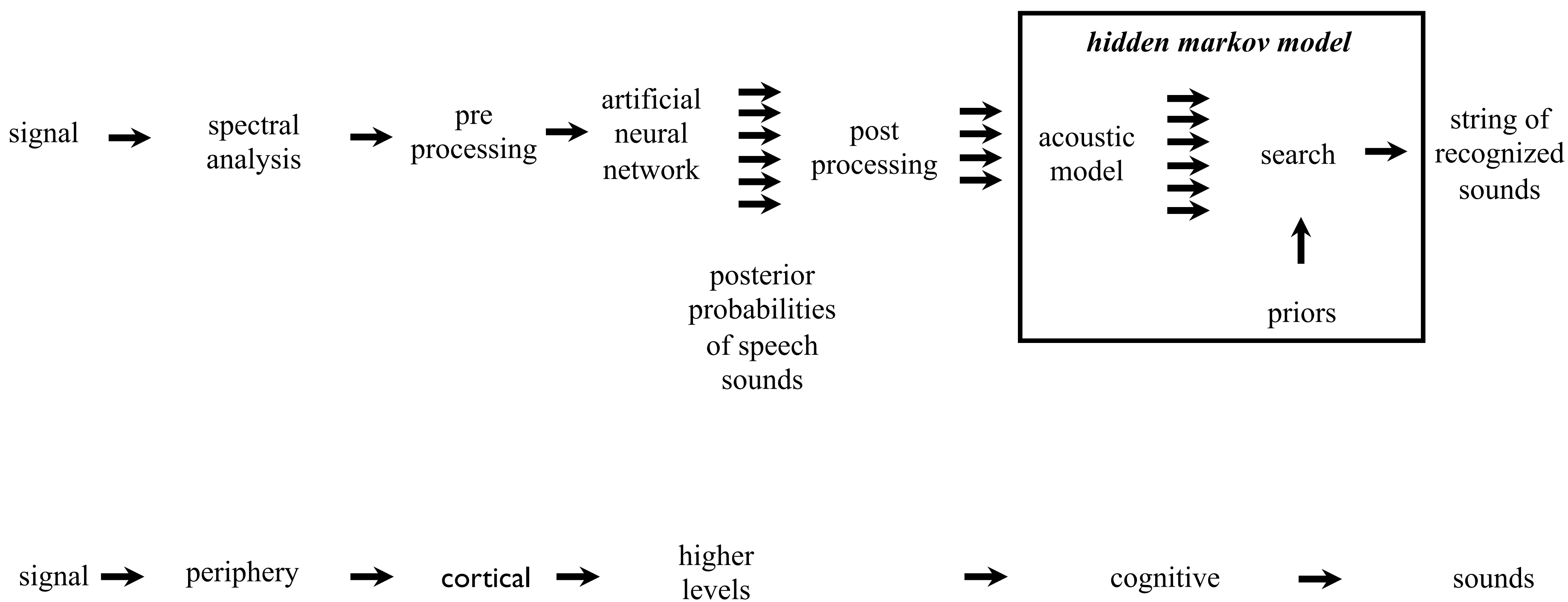
TANDEM

FDPLP

MRASTA

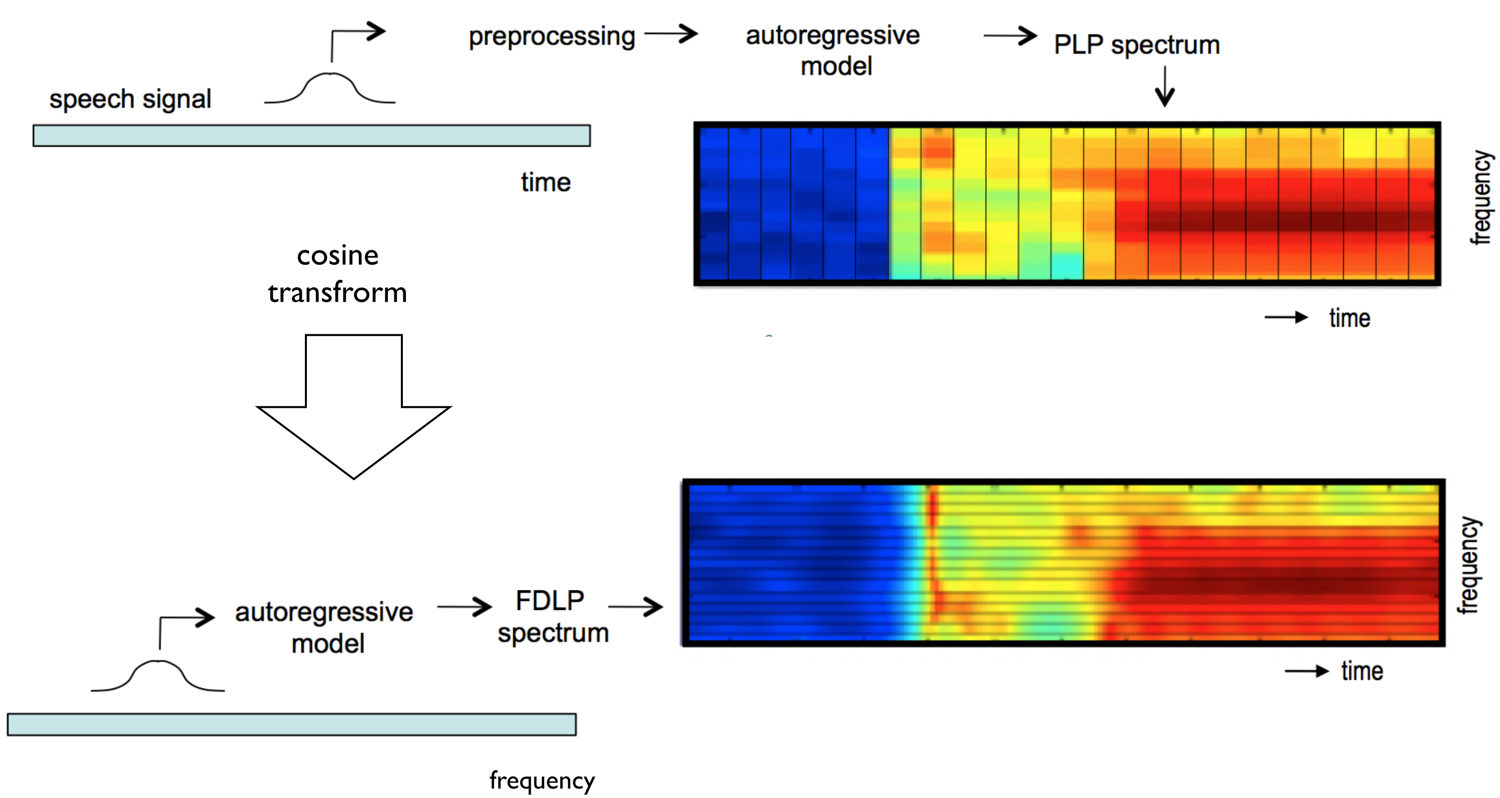
parallel

HATS (bottleneck)



Frequency Domain Perceptual Linear Predicton (FDPLP)

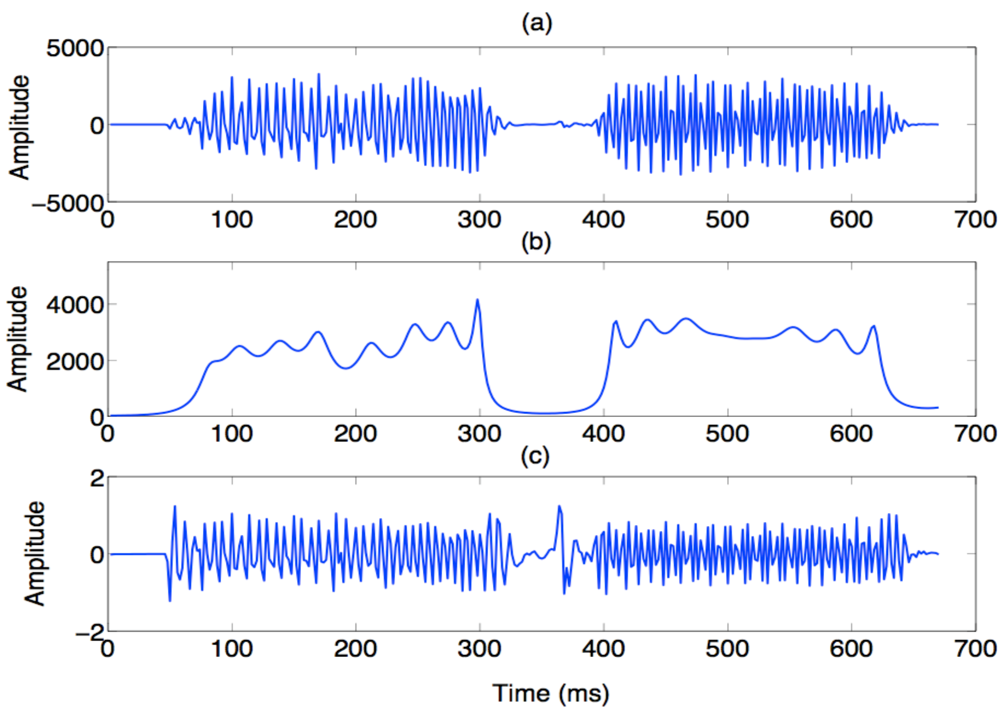
-with Marios Athineos, Dan Ellis, Sriram Ganapathy and Samuel Thomas



Decomposition into AM and FM components.

Straightforward alleviation of effects of linear distortions and , reverberations .

FDLP decomposition of the signal



signal

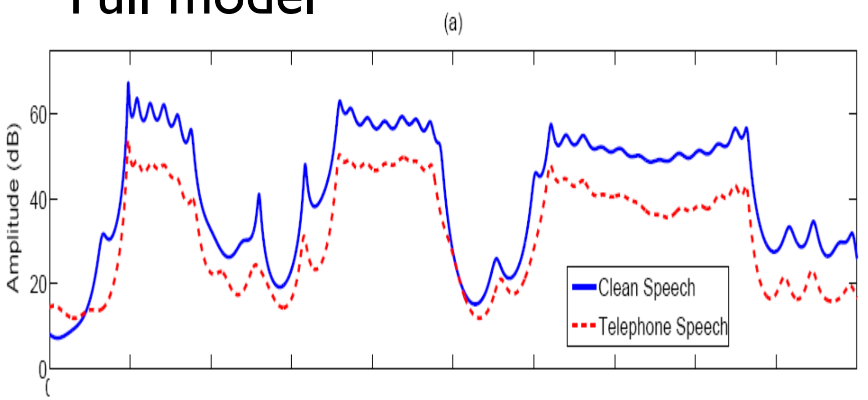
AM component (temporal envelope)

FM component (carrier)

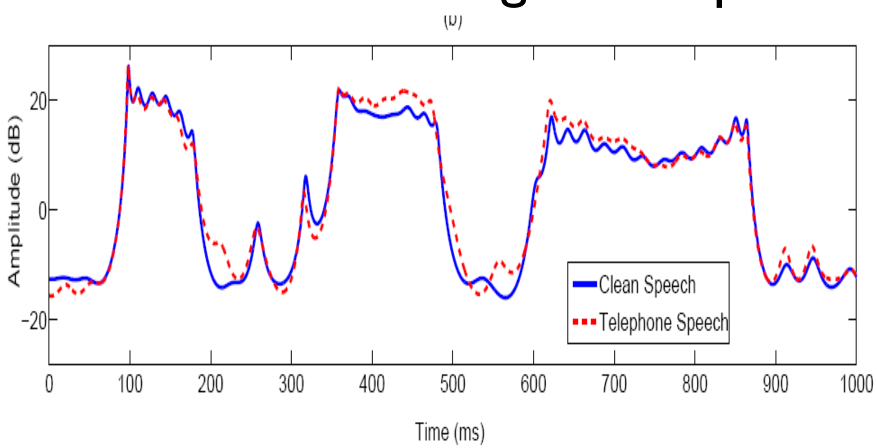
Varying communication channels
(convolution with a short impulse response of a channel)

Convolution turns into addition in log spectral domain

Full model



Model without its gain component



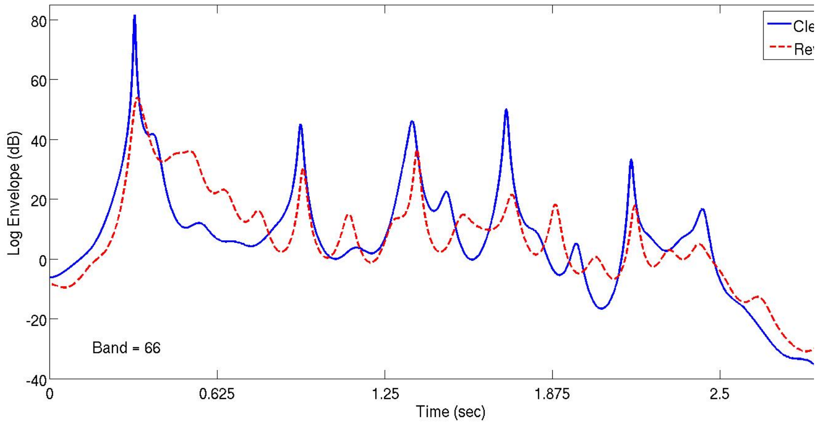
Ignoring FDPLP model gain makes the representation invariant to linear distortions introduced by the communication channel.

Reverberant speech
(convolution with a long impulse response of the room)

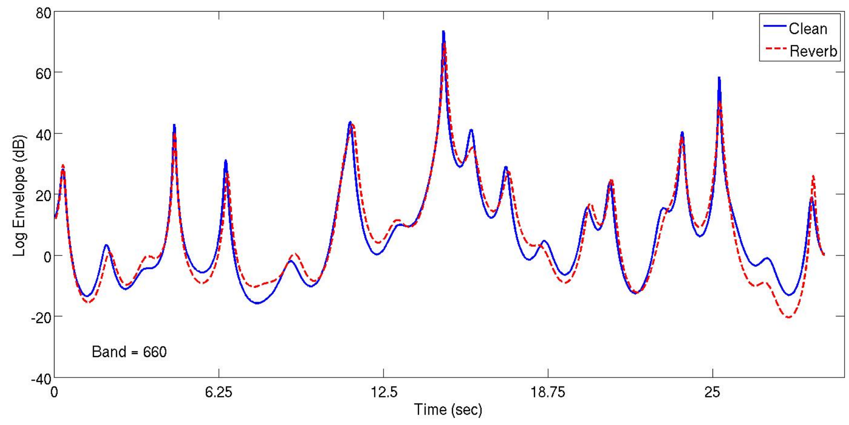
Convolution turns into addition in log spectral domain, **as long as the most of the room impulse response fits into the analysis window!**

Ignoring FDLP model gain makes the representation invariant to revebs.

3 s window

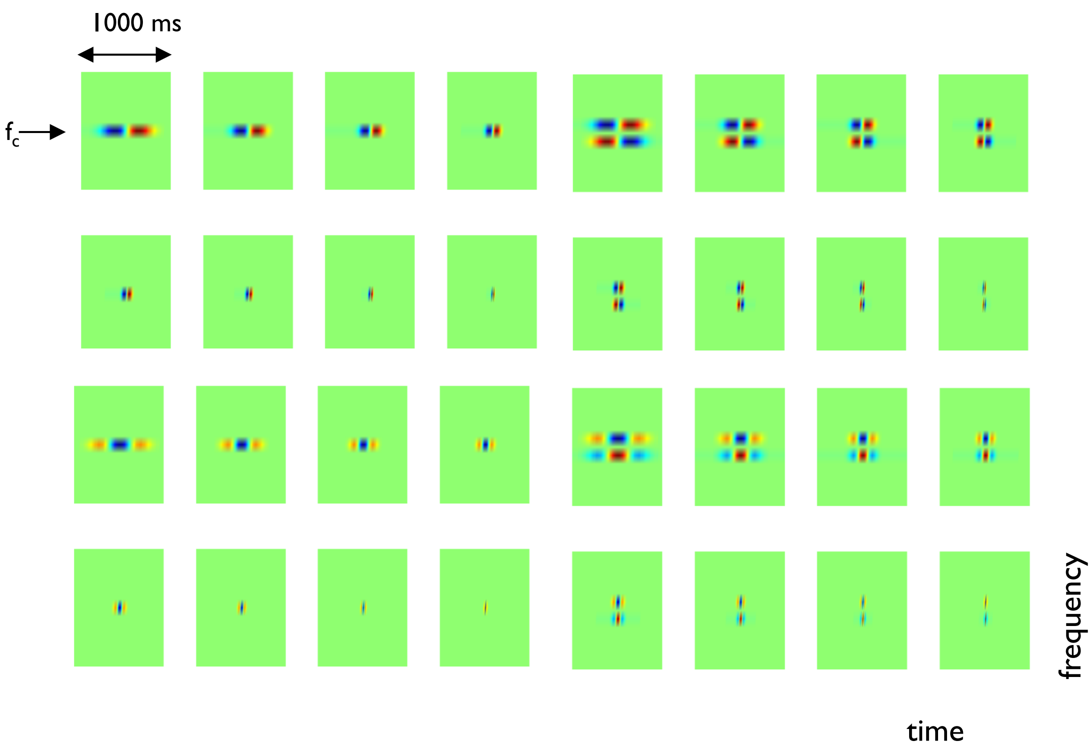


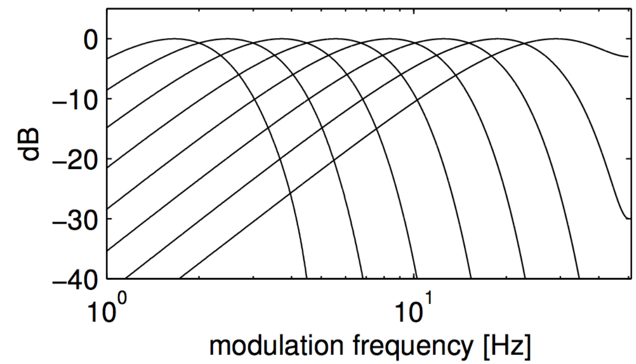
30 s window



MRASTA

Hermansky and Fousek 2005





32 features at each of 14 frequencies

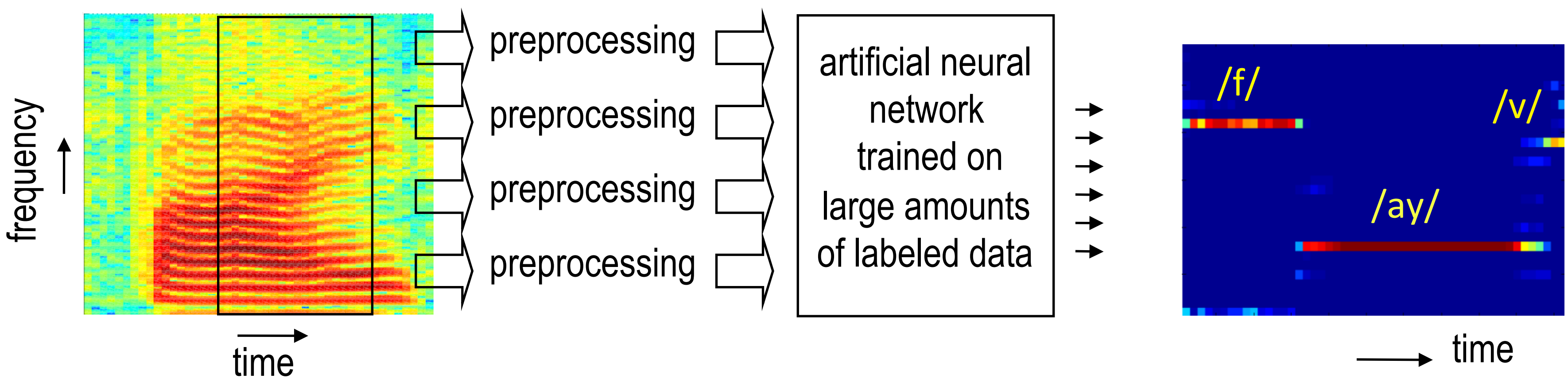
448 dimensional vector of features every 10 ms

multi-resolution band-pass filtering of modulation spectrum

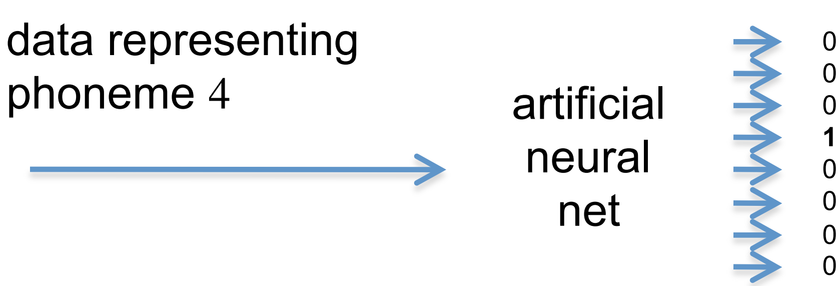
Optimal (lowest dimensionality) features are posterior probabilities of classes

Spectrogram

Posteriogram



Training of the artificial neural net

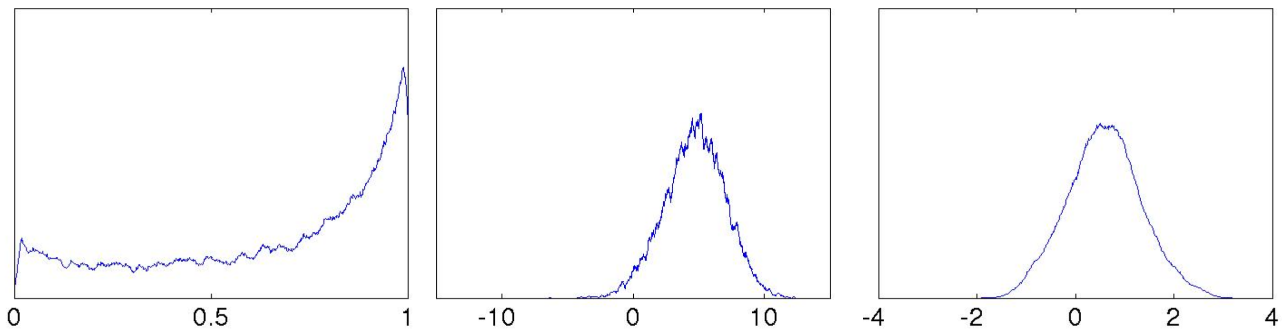


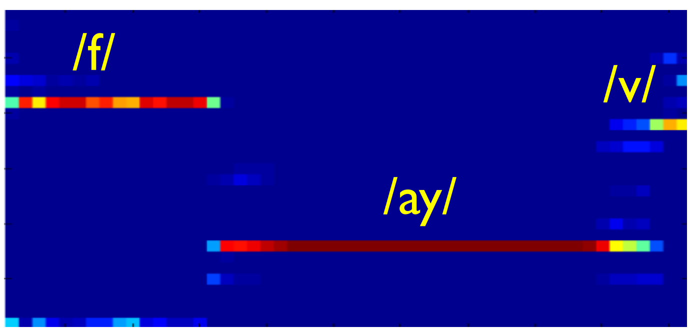
TANDEM Features for HMM/GMM System

Hermansky, Ellis, and Sharma 2000

good attributes for state-of-the-art ASR systems
should be Normally distributed and uncorrelated

histogram of one feature



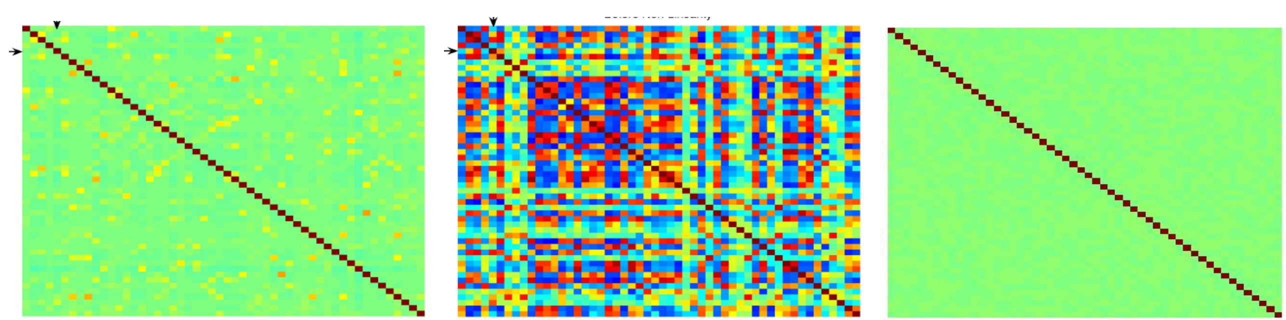


pre-softmax outputs

principal component projection

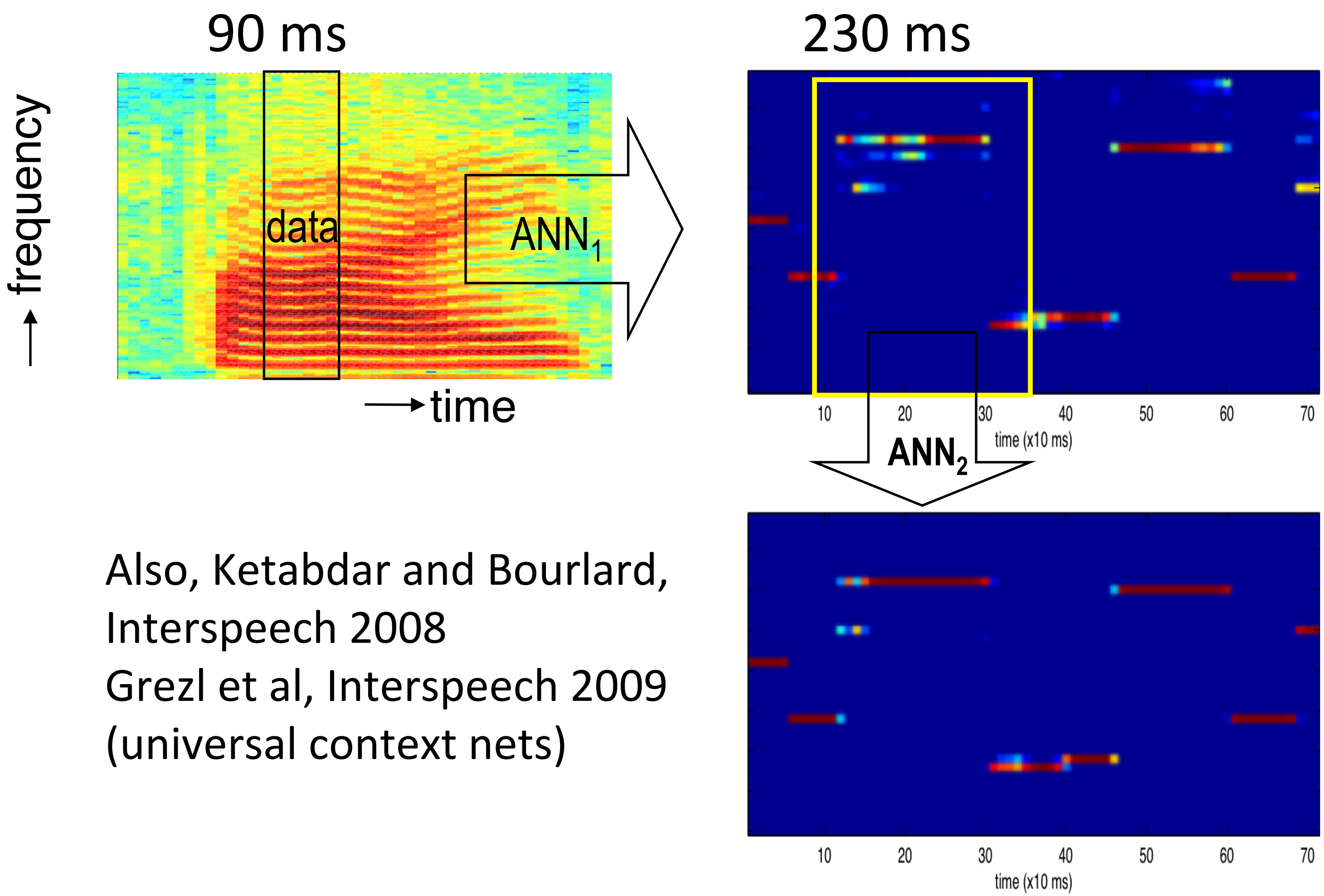
to HMM

correlation matrix of features



Serial hierarchical estimation

(Pinto et al, Interspeech 2008)

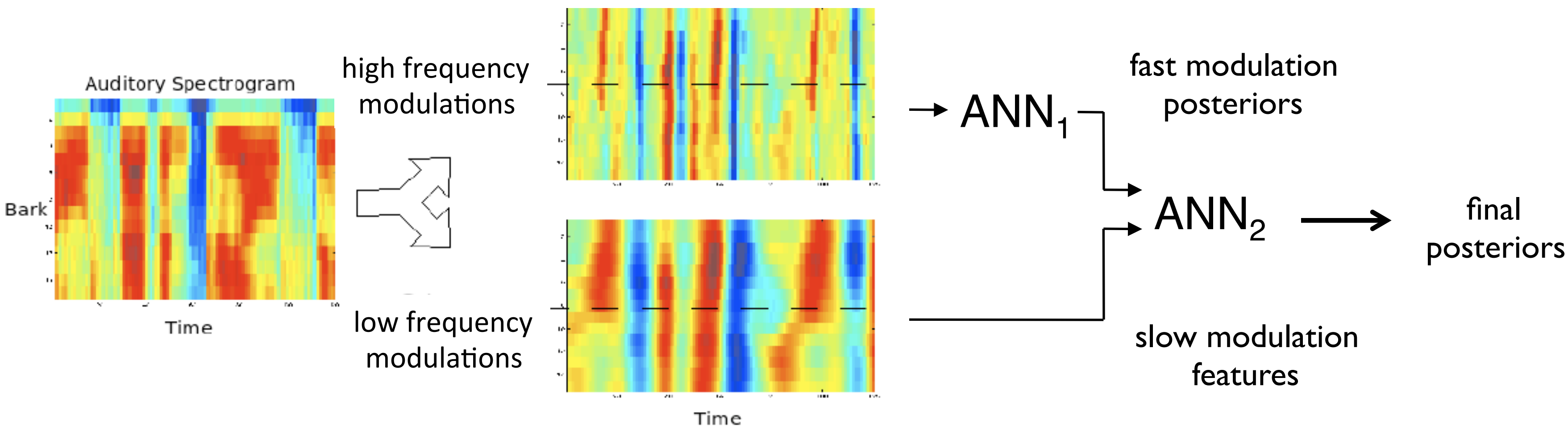


Also, Ketabdar and Bourlard, Interspeech 2008
Grezl et al, Interspeech 2009 (universal context nets)

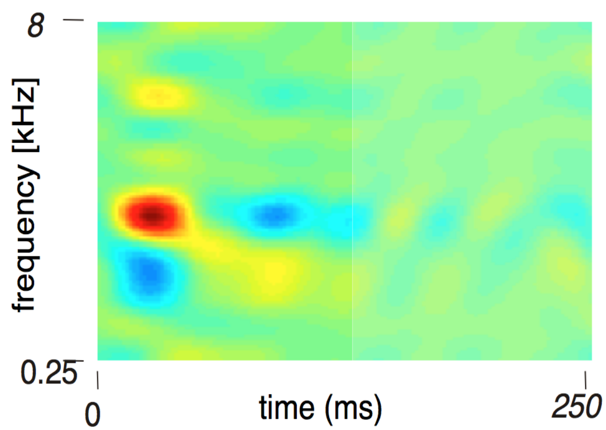
Parallel hierarchical estimation

Valente and Hermansky, ICASSP 08, Interspeech 2008

- one-stage processing on coarse (slow modulations) representation
- two-stage processing of finer (faster modulations) representation



Auditory cortical spectro-temporal receptive fields (STRFs)

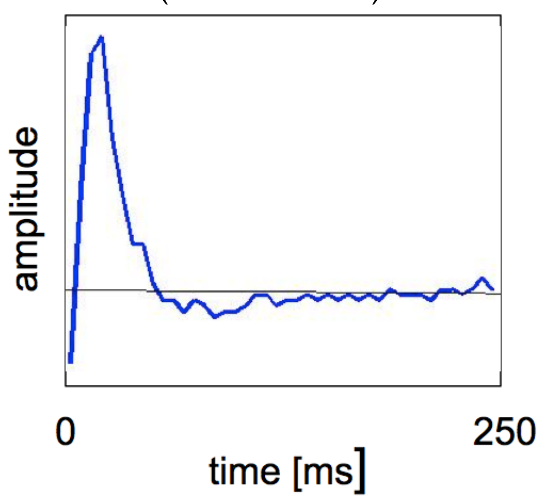


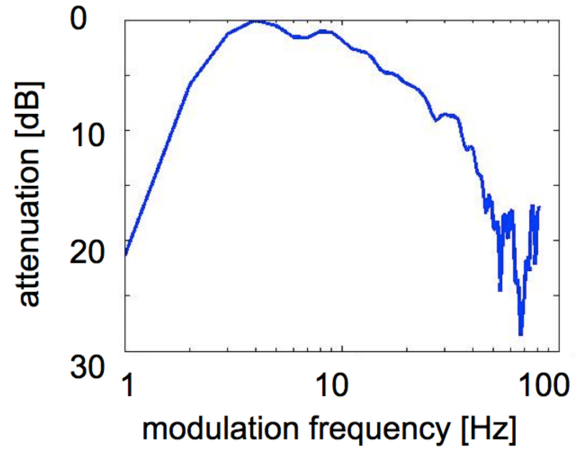
from Shihab Shamma

indicate "optimal" stimulus that excites a given cortical neuron

compute principal components along temporal axis of about 300 STRFs

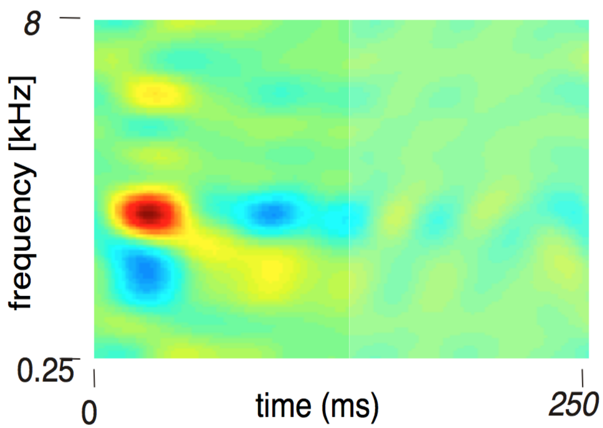
Nima Mesgarani (in preparation)





[higher components similar but shifted in time]()

Auditory cortical spectro-temporal receptive fields (STRFs)



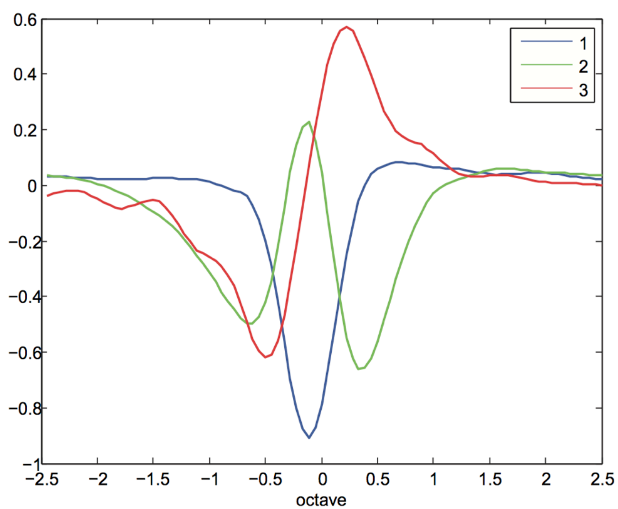
from Shihab Shamma

indicate “optimal” stimulus that excites a given cortical neuron

align maxima of STRFs in frequency and compute principal components along frequency axis of about 300 STRFs

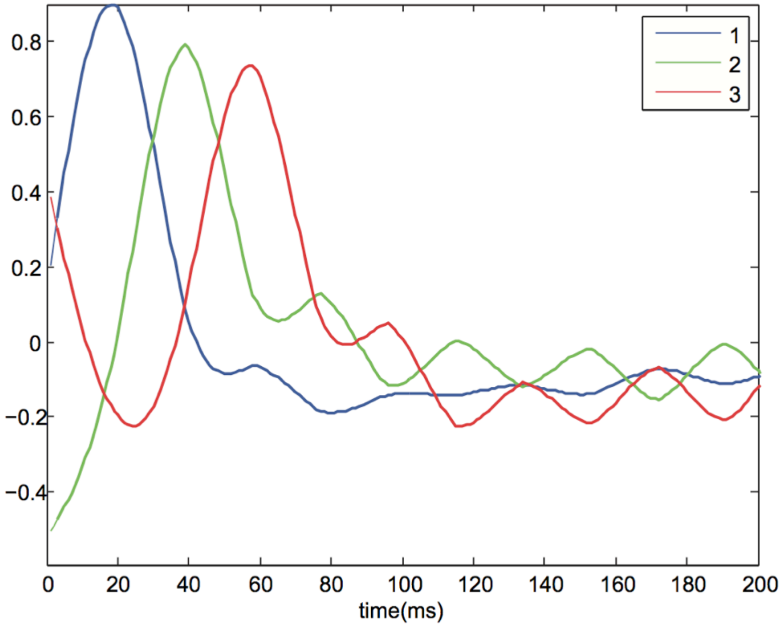
Nima Mesgarani (in preparation)

Principal components of spectral axis

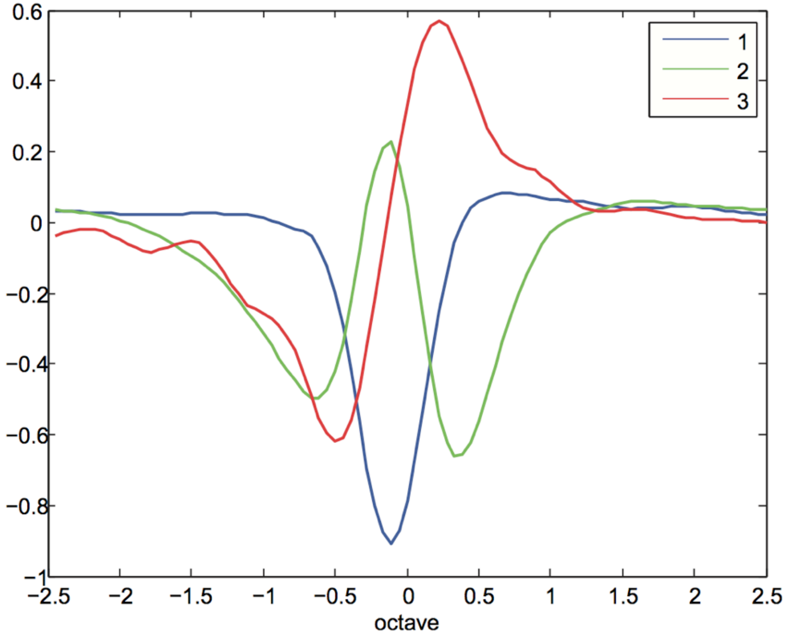


- principal components of about 300 STRFs
  - Nima Mesgarani (in preparation)

Principal components of temporal axis

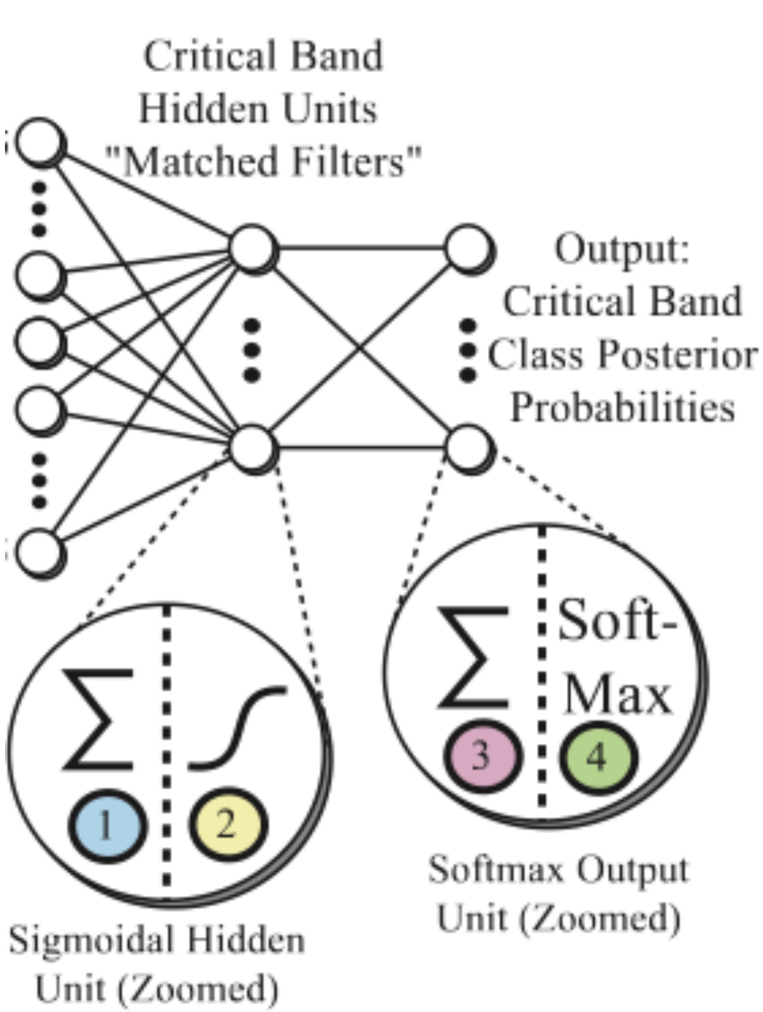


Principal components of spectral axis



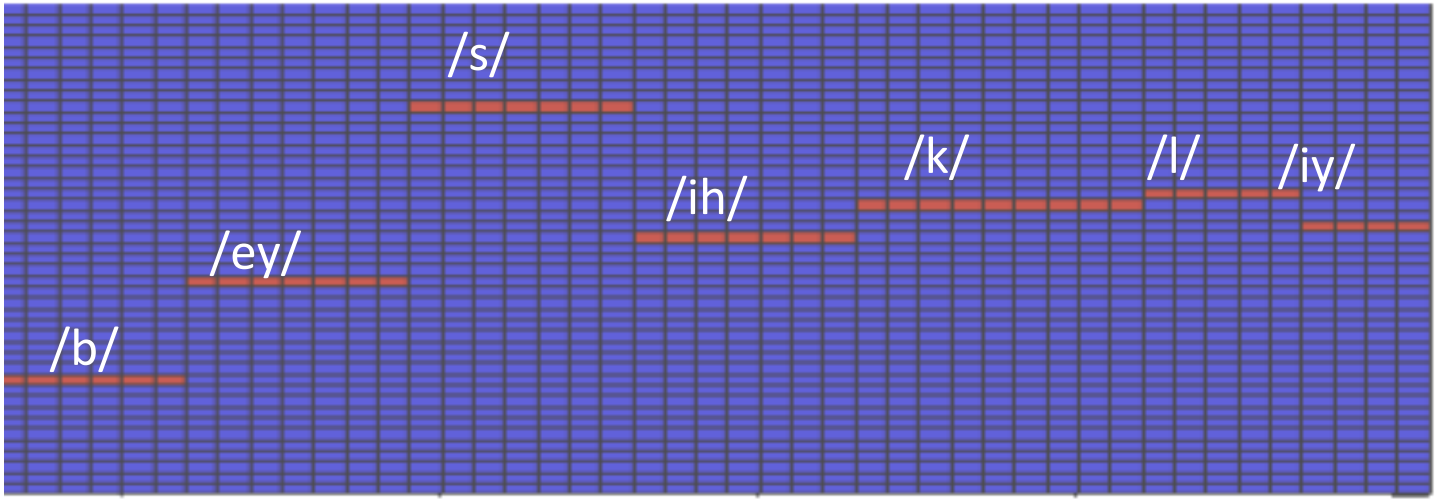
HATS (bootleneck features)

Chen et al 2005, Grezl et al 2007

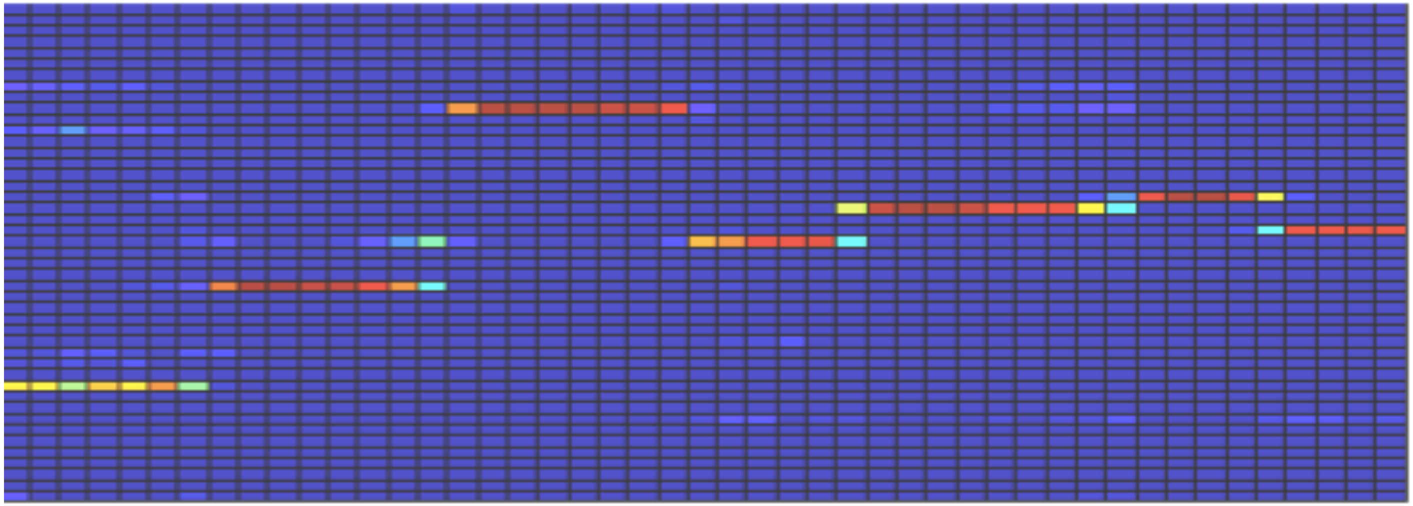


4 - posteriors
3 - for TANDEM
2 - HATS

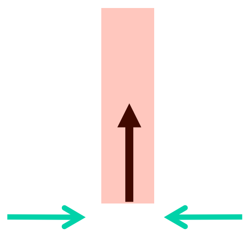
truth

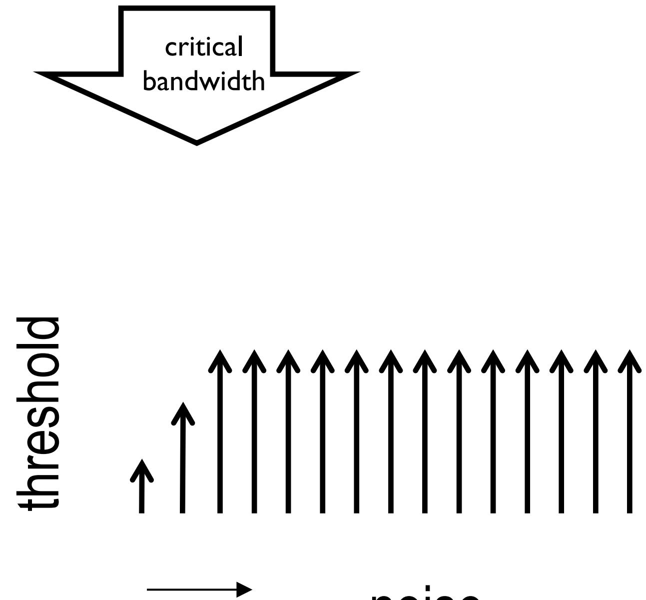


estimate



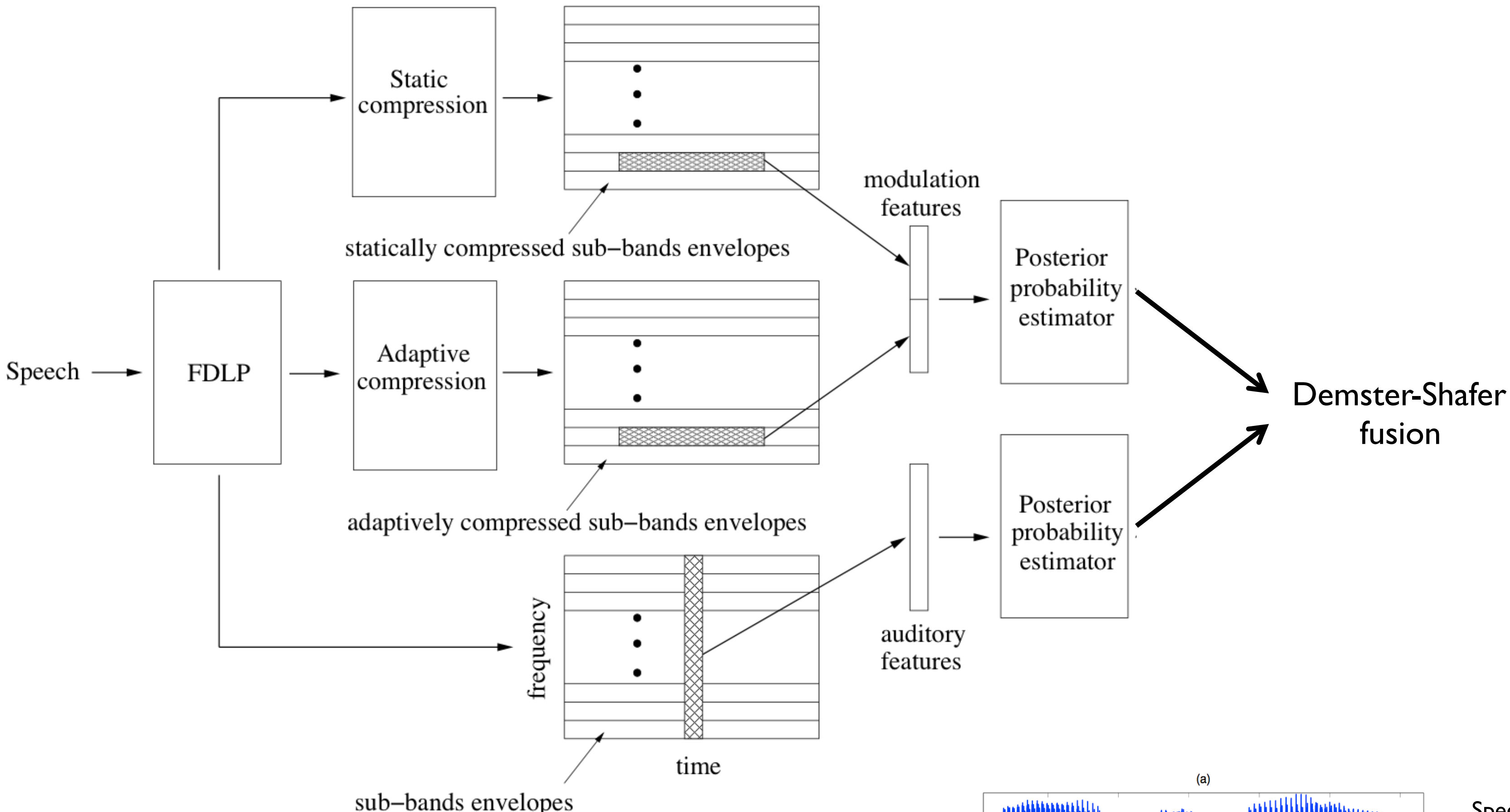
simultaneous masking

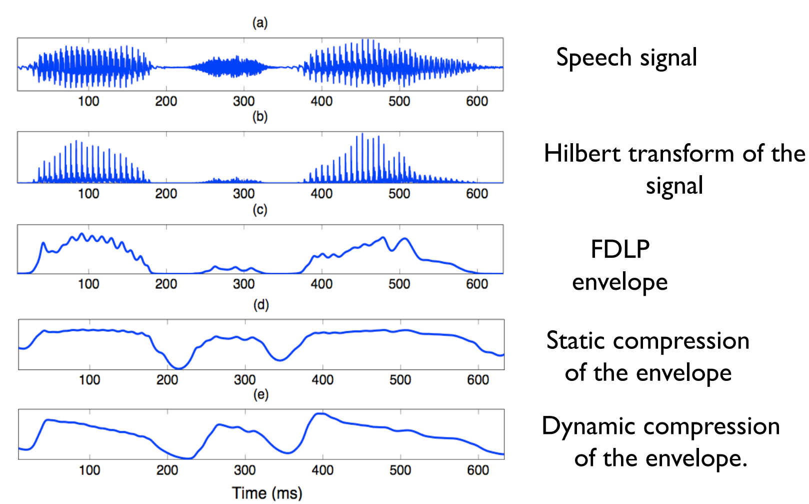




spectral components inside the (critical) band interact differently with components inside the band than they do with components, which are outside the band

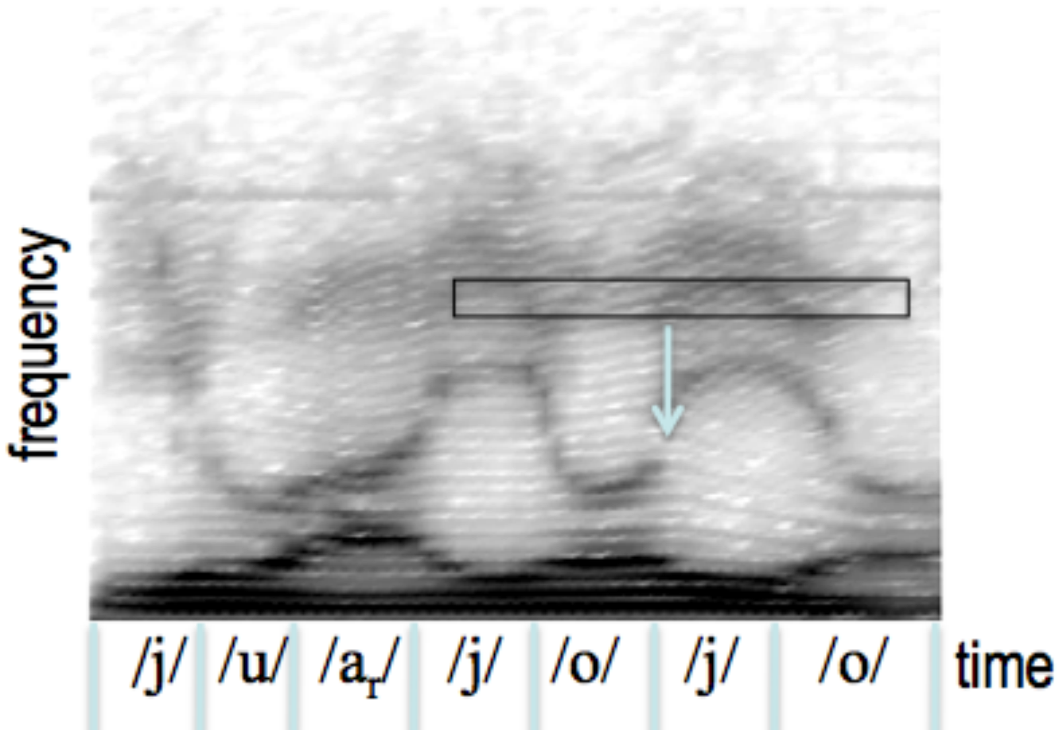
hearing periphery does spectral analysis to allow for separation of corrupted signal elements at higher levels of auditory processing


Speech
FDLP
Static compression
Adaptive compression
statically compressed sub–bands envelopes
adaptively compressed sub–bands envelopes
sub–bands envelopes
frequency
time
modulation features
auditory features
Posterior probability estimator
Posterior probability estimator
Demster-Shafer fusion


(a)
(b)
(c)
(d)
(e)
100
200
300
400
500
600
Time (ms)
Speech signal
Hilbert transform of the signal
FDLP envelope
Static compression of the envelope
Dynamic compression of the envelope.

Data-guided FIR RASTA filters

van Vuuren and Hermansky 1997, Valente and Hermansky 2006

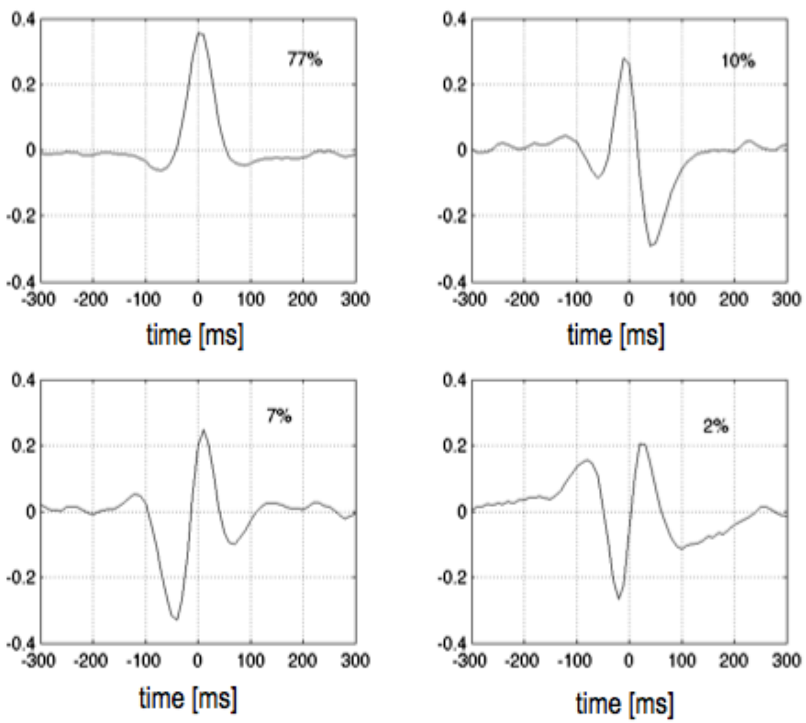


Data-guided (LDA-based) FIR RASTA filters

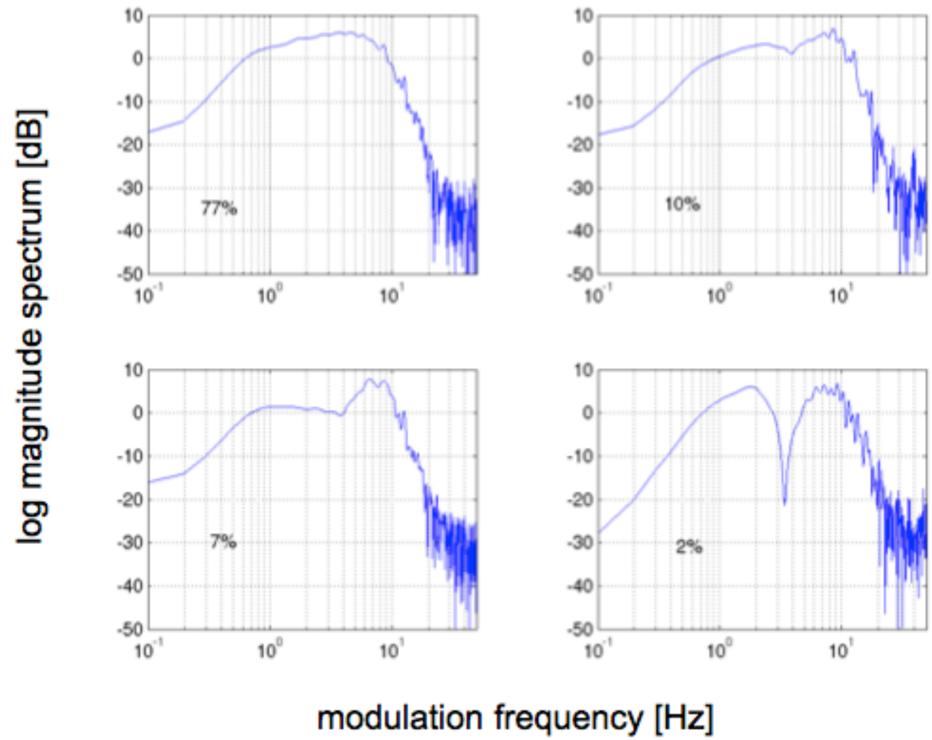
van Vuuren and Hermansky 1997, Valente and Hermansky 2006

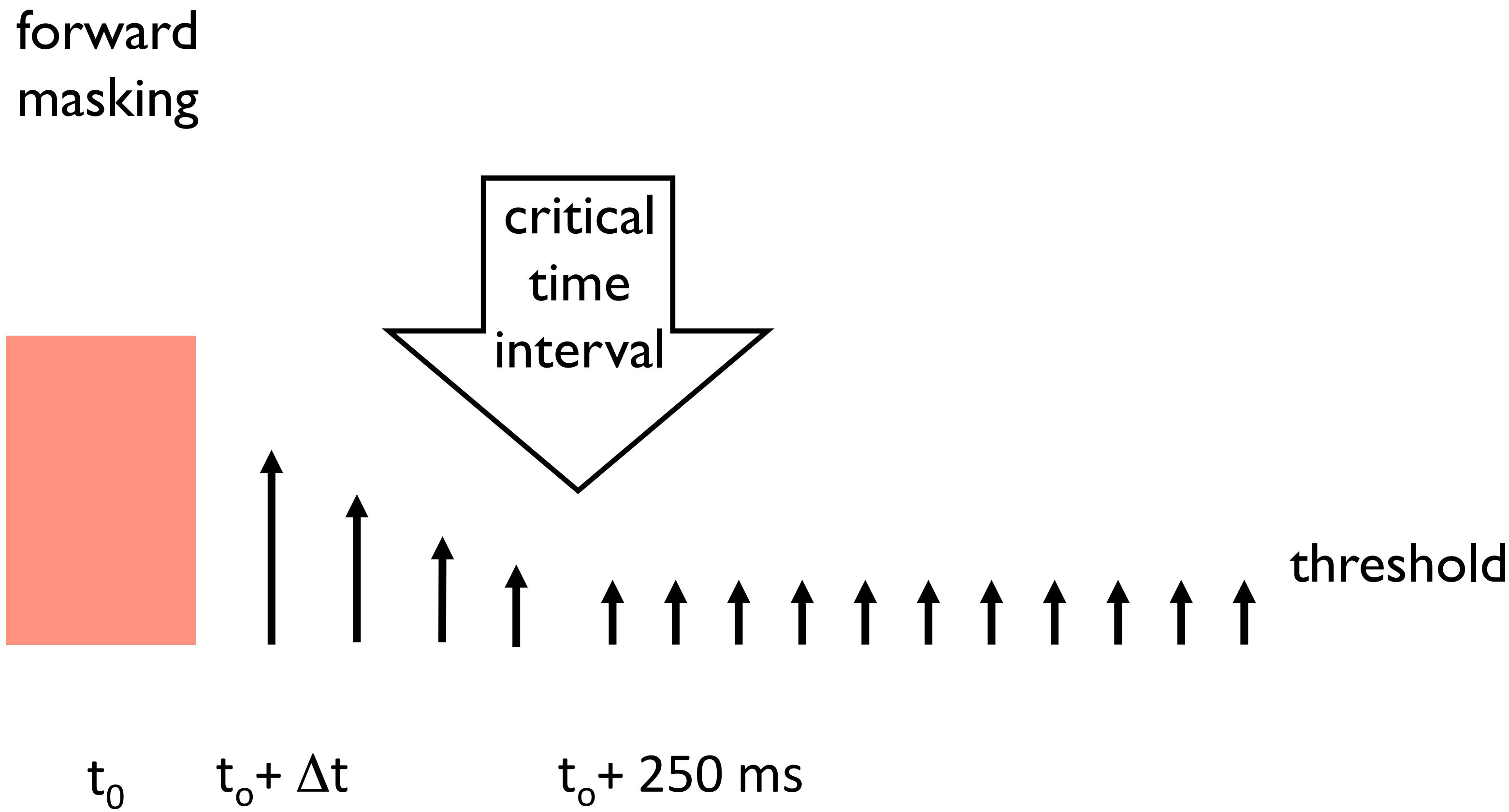
first 4 temporal linear discriminats

impulse responses



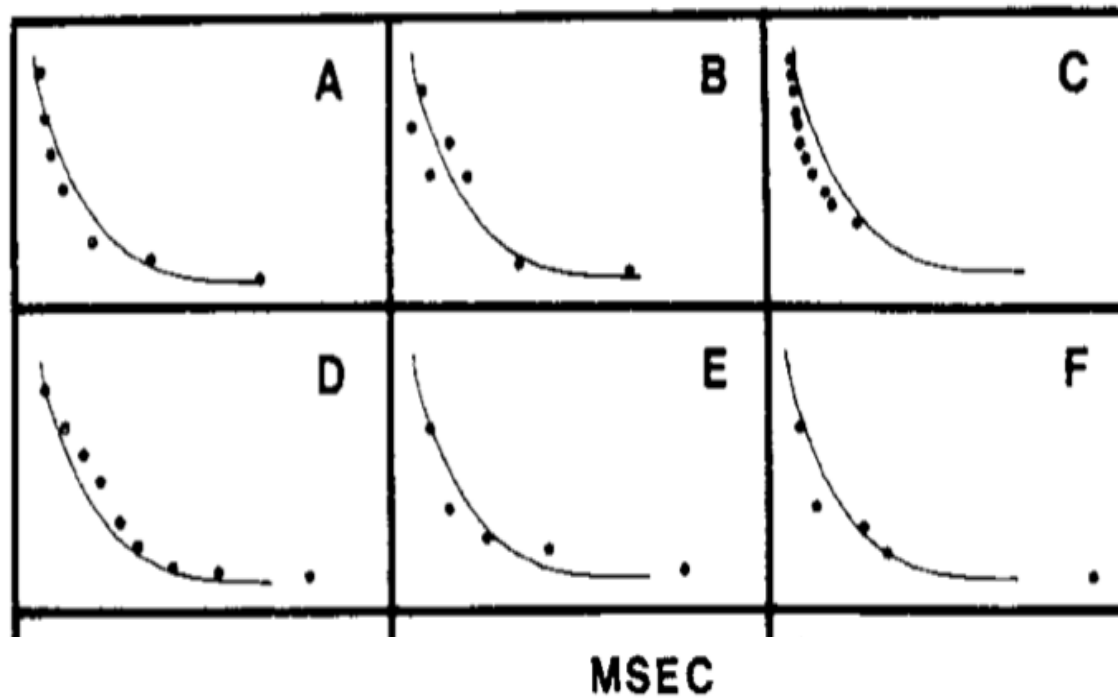
frequency responses





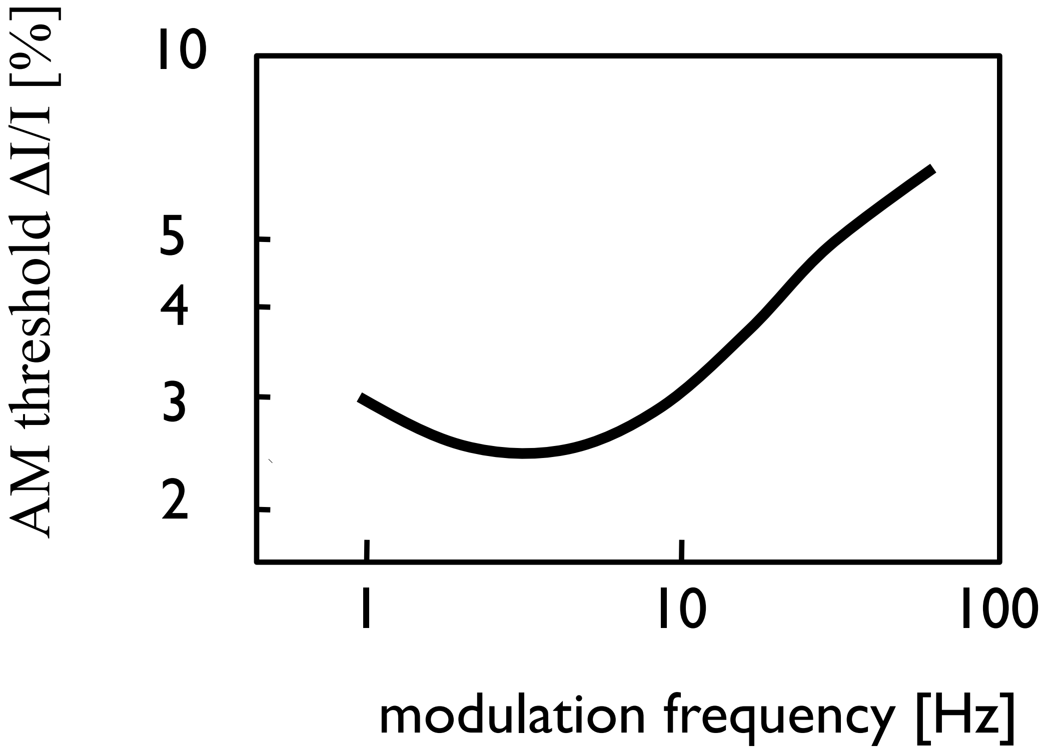
forward
masking
critical
time
interval
threshold
t0
to+ Δt
to+ 250 ms

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



N. Cowan, On Short and Long Auditory Stores, Psychological Bulletin 1984

Riezs 1928



Kozhevnikov and Chistovich
(Speech: Articulation and Perception, 1965)

- reaction times for identifying consonants and vowels in CV syllables

- consonant always identified before a vowel

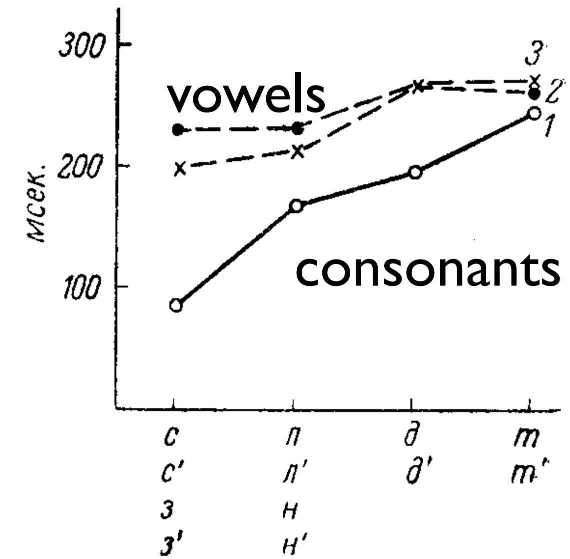


Рис. 6.6. Задержки буквенной записи согласных в зависимости от их качества (*1*) и задержки буквенной записи гласных в зависимости от качества предшествующих согласных (*2* — твердый согласный, *3* — мягкий согласный).

To recognize phoneme one needs to collect information distributed over the whole syllable