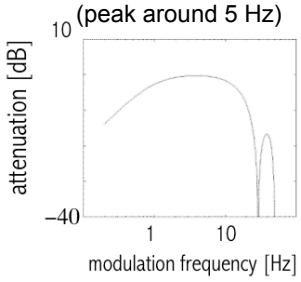
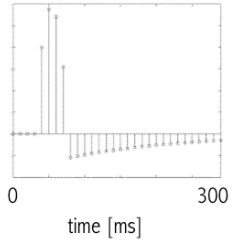
## Machine recognition of speech trained on data from New Jersey Labs

| tested in New Jersey | tested in Colorado |
|---|---|
| 2.8% error | 60.7% error |
| RASTA | RASTA |
| **2.2% error** | **2.9% error** |

Frequency response
(peak around 5 Hz)



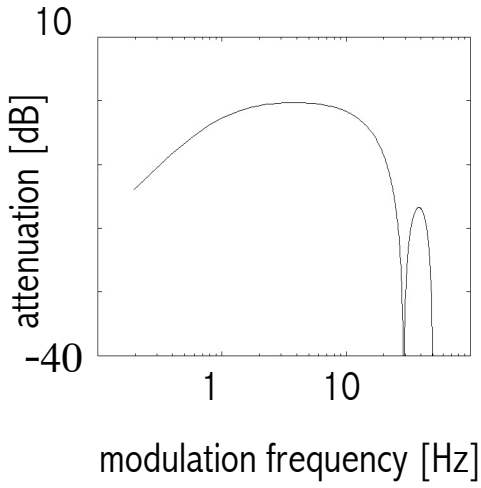

Impulse response
(effective length around 200 ms)




- sensitivity of hearing to modulation peaks at about 4 Hz
  - Riesz 1928, Zwicker 1952, …
- modulation transfer function of primary auditory cortex peaks at about 4 Hz
  - Schreiner via Greenberg, personal communication 1997
- modulation spectrum of speech peaks at about 4 Hz
  - Houtgast and Steeneken 1978
- intelligibility of speech significantly impaired when 4 Hz modulation frequency component attenuated
  - Drullman et al 1992, Arai et al 1996

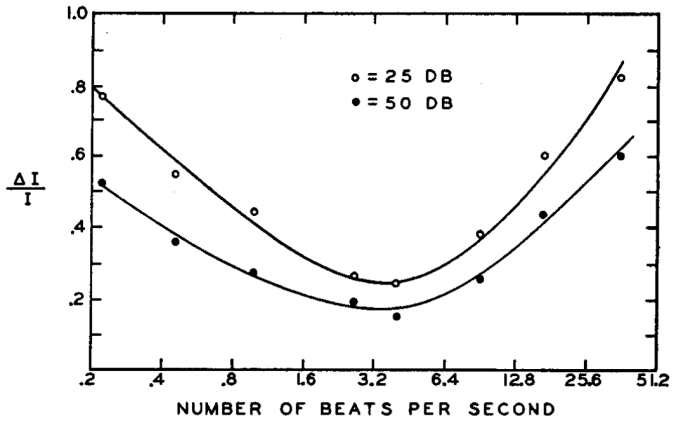
- frequency discrimination of short stimuli improves up to about 200 ms
- loudness of equal-energy stimuli grows up to about 200 ms
- minimum detectable silent interval indicates time constant of about 200 ms
- effect of forward masking lasts about 200 ms

# RASTA filter




# Perception of modulations

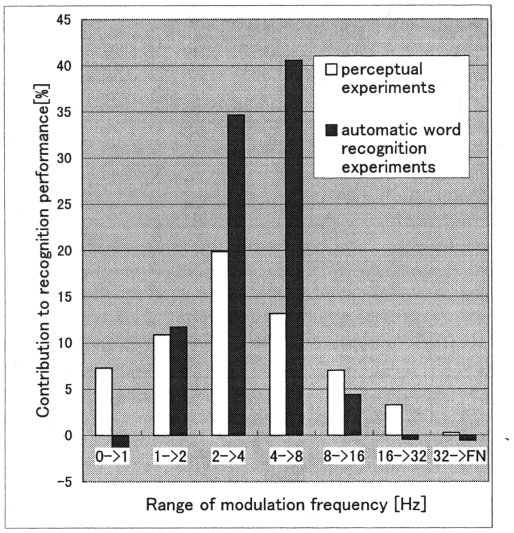
(Riesz 1923)



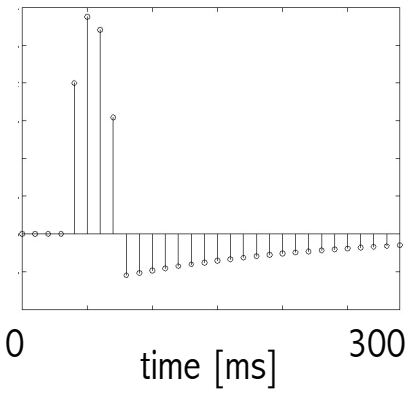
# 1-12 Hz Passband

- sensitivity of hearing to modulation peaks at about 4 Hz
  - Riesz 1928, Zwicker 1952, …
- modulation transfer function of primary auditory cortex peaks at about 4 Hz
  - Schreiner via Greenberg, personal communication 1997
- modulation spectrum of speech peaks at about 4 Hz
  - Houtgast and Steeneken 1978
- intelligibility of speech significantly impaired when 4 Hz modulation frequency component attenuated
  - Drullman et al 1992, Arai et al 1996

Relative importance of various components of modulation spectrum of speech for speech intelligibility and for ASR

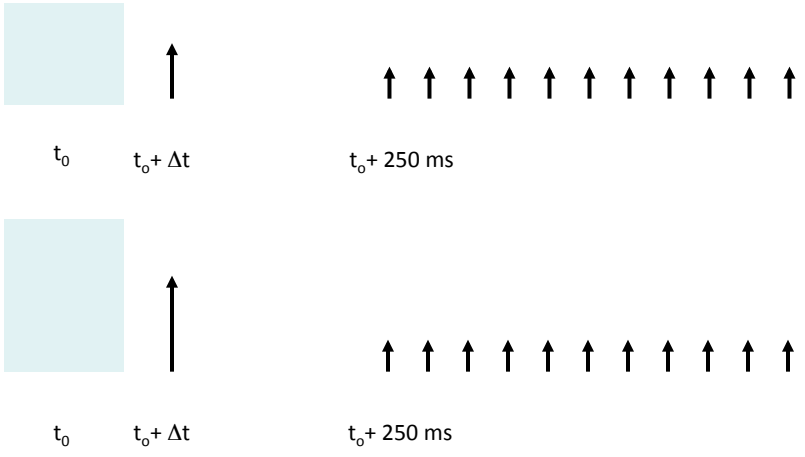


# RASTA filter




- average four neighboring frames
- subtract exponentially decaying past values ($\tau$=170 ms)

# Masking in Time



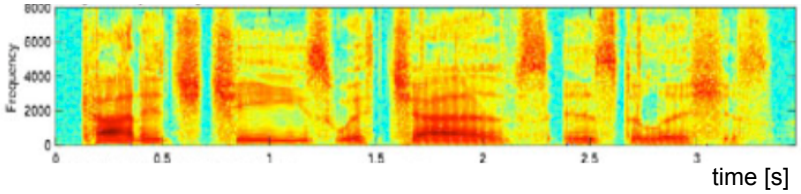

- suggests ~250 ms buffer (critical interval) in auditory system
  - **what happens outside the critical interval, does not affect detection of signal within the critical interval**

# ~ 200 ms length of impulse response

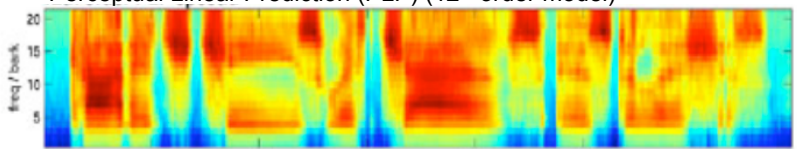
- frequency discrimination of short stimuli improves up to about 200 ms
- loudness of equal-energy stimuli grows up to about 200 ms
- minimum detectable silent interval indicates time constant of about 200 ms
- effect of forward masking lasts about 200 ms

**syllable-length buffer of human hearing ?**

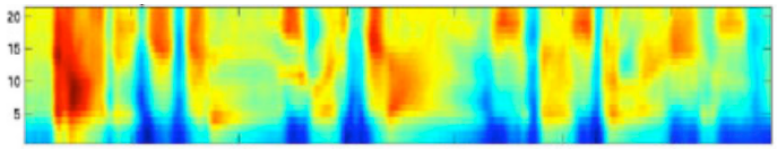
spectrogram (short-term Fourier spectrum)




Perceptual Linear Prediction (PLP) (12th order model)

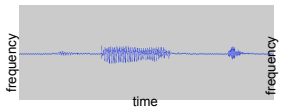


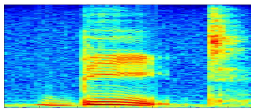

RASTA-PLP

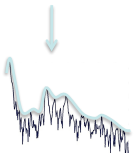


## Formant-Less Vowel

original speech

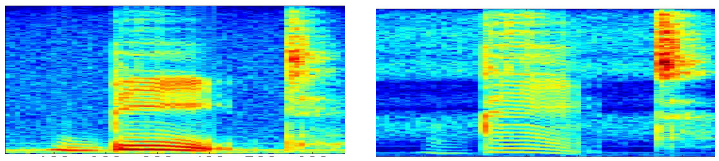




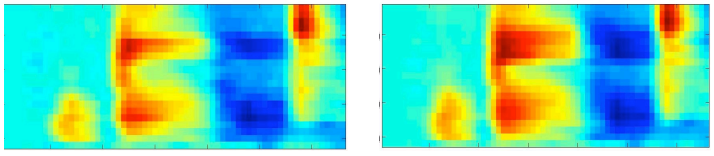
original speech filtered speech

spectrogram



frequency

spectrogram from RASTA



time

# Data Do Not Lie

Prof. Frederick Jelinek: “Airplanes don’t flap their wings”.

S. Lohr, New York Times, March 6, 2011

“Airplanes do not flap wings but have wings nevertheless,…..

Of course, we should try to incorporate the knowledge that we have **of hearing, speech production, etc.**, into our systems,....

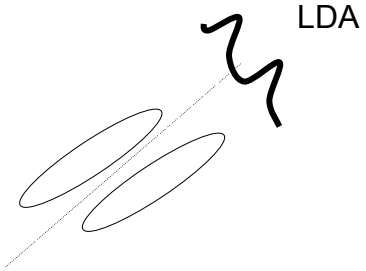
F. Jelinek, Five speculations (and a divertimento) on the themes of H. Bourlard, H. Hermansky, and N. Morgan, Speech Communication 18, 1996. 242–2



# Linear Discriminant Analysis (LDA)

Linear discriminants: eigenvectors of $S^{-1}_W S_B$

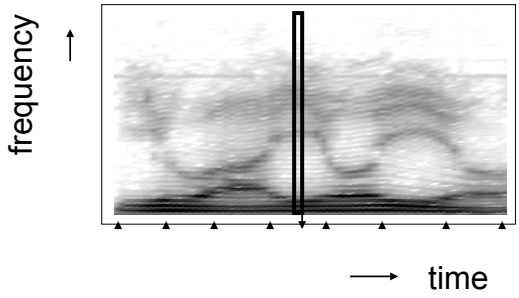
$S_W$ - within-class covariance matrix
$S_B$ - between class covariance matrix




- Needs labeled data
- Within-class distributions assumed Gaussian with equal $\sigma$ (take log of power spectrum)

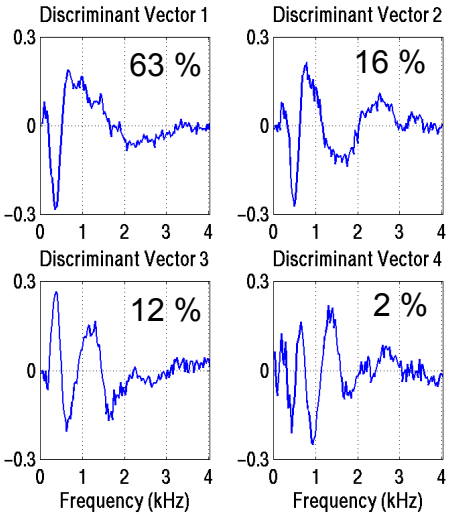
# Spectral Basis from LDA

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:

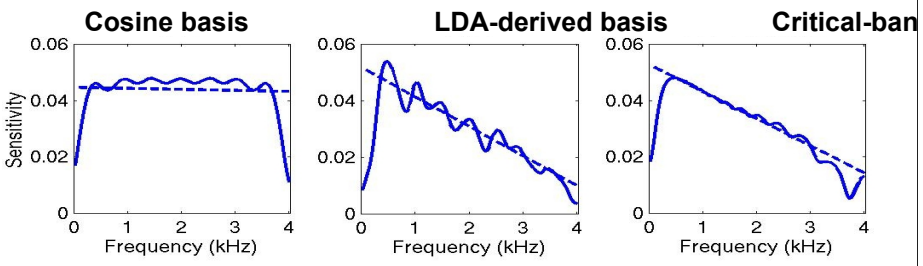
Critical bands of human hearing are broader at higher frequencies

Physiology:

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

## Sensitivity to Spectral Change

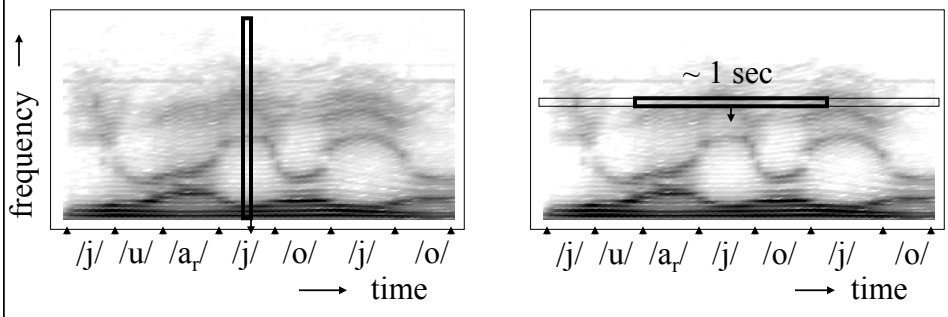
(Malayath 1999)



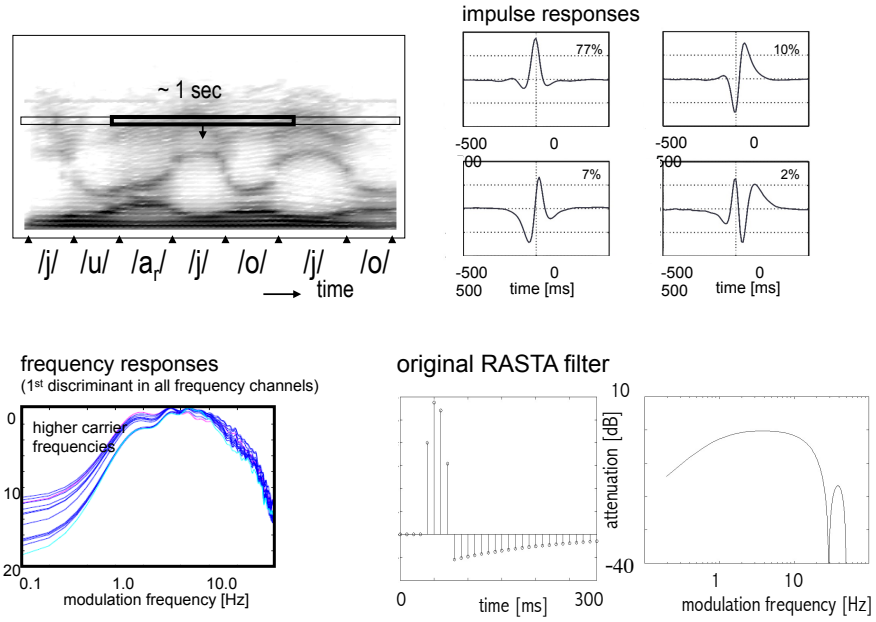

## Two ways of using LDA

LDA gives basis for projection of spectral space

LDA gives FIR filters for processing time trajectories of spectral energies



## RASTA Filters from Speech Data




**engineering**

***effect ( signal )***

***cognitive signal***

***effect ( signal )***

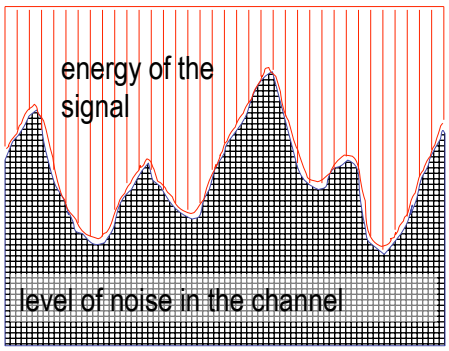
**perception**

good engineering could be consistent with biology

- physiology of sensory organs
- psychophysics of perception
- emulation of the knowledge in engineering

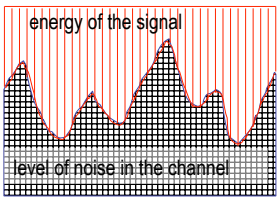
C. Shannon: Communication in Presence of Noise

- combination of channel and signal spectrum should be as flat (as random-like) as possible




frequency

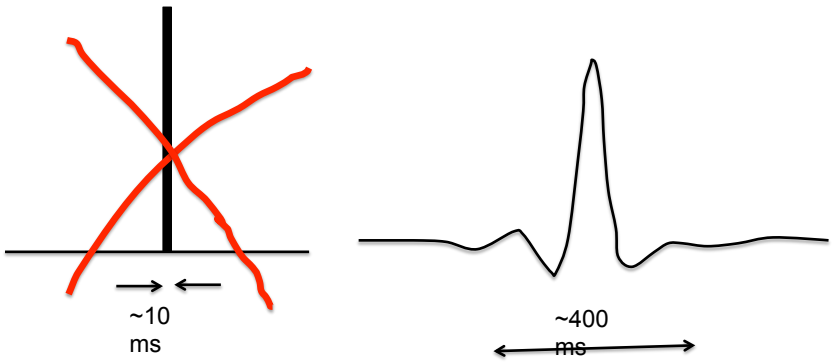
# Forces of Nature




resource space

if signal could be controlled (e.g. speech)

- put more signal where there is less noise
- sensory signal optimized for a given communication channel

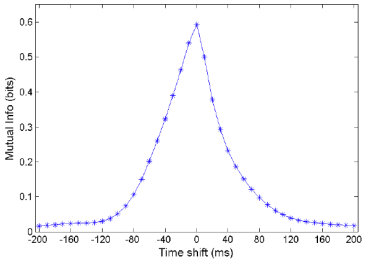



Evaluate spectra within a given speech sound **relative to** neighboring sounds

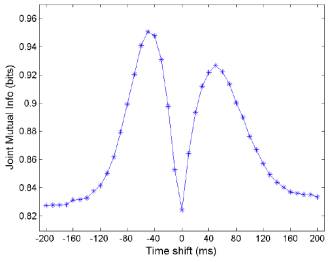
# Mutual Information Between Phoneme Labels and Measurement(s) in Time

H. Yang et al 2000, F. Li (unpublished)

first measurement




second measurement




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

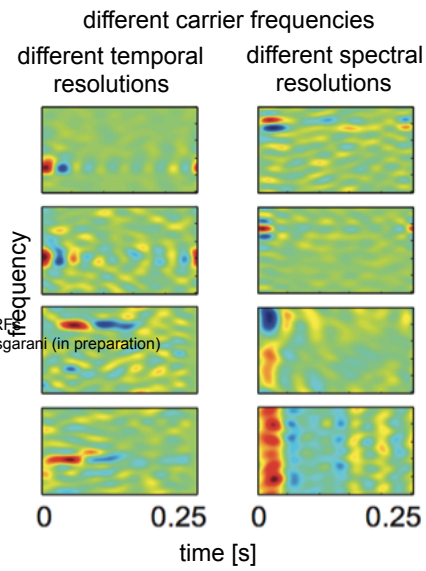
$$I(X;Y|Z) = \sum_{z \in Z} \sum_{y \in Y} \sum_{x \in X} p_{X,Y,Z}(x,y,z) \log \frac{p_Z(z) p_{X,Y,Z}(x,y,z)}{p_{X,Z}(x,z) p_{Y,Z}(y,z)}$$

## Auditory cortical receptive fields

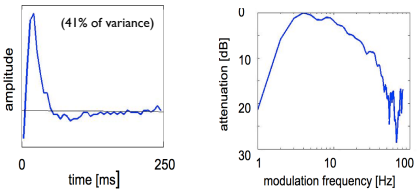
from N. Mesgarani

Time-frequency patterns that optimally excite a given cortical neuron

Most often frequency-localized and often rather long




1st principal component along temporal axis from about 300 STRFs
Nima Mesgarani (in preparation)




## Short Term Spectral Envelope?

Ear is frequency selective !

Simultaneous masking: Sound elements outside a critical band do not corrupt decoding of elements inside the band

Temporal masking: Sound elements outside a critical interval (about 250 ms) do not corrupt decoding of elements inside the interval

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

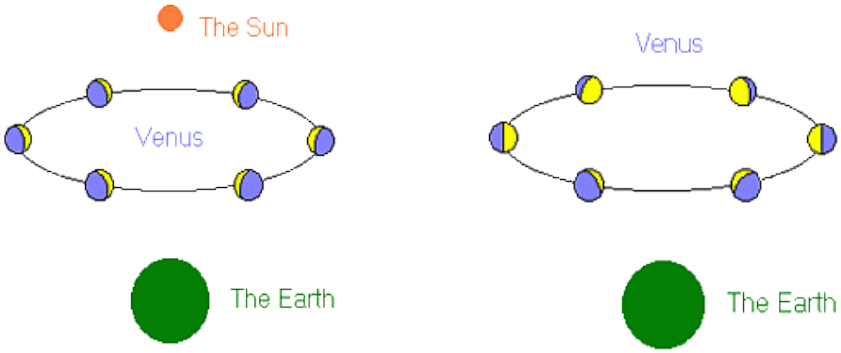
**Human listeners recognize speech in independent frequency bands**

Jont Allen's interpretation of earlier works of Fletcher et al at the 1993 Summer Workshop at Rutgers University

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

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

## Power of Experimental Results



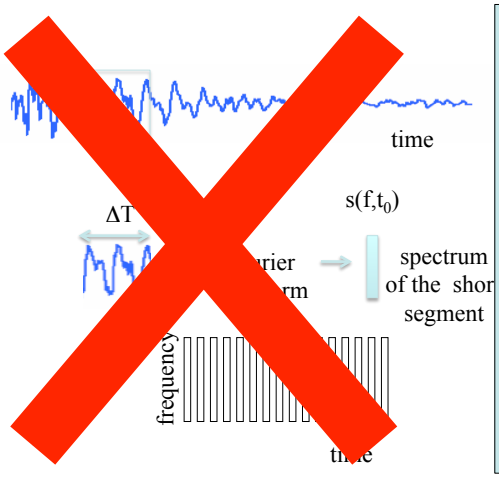

Ptolemy

Galileo

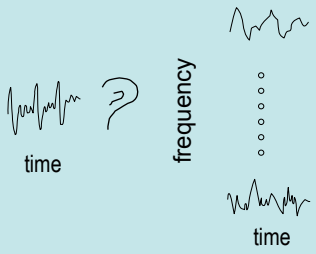
Ear is frequency selective

however, it is NOT to derive spectrum of the signal but to yield frequency-localized temporal patterns, which carry the information about underlying acoustic events.

# Away from Short-Term Spectrum




back to human hearing




109

## Frequency Domain Linear Prediction (FDLP)

FDLP

- means for all-pole estimation of Hilbert envelopes (instantaneous spectral energies) in individual frequency channels