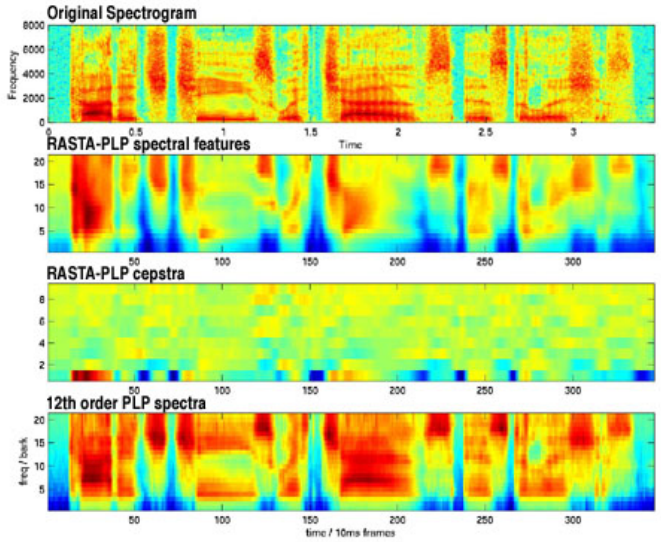
Curso "Reconocimiento automático del habla"

sinc(*i*)

Técnicas clásicas para análisis de voz



Hugo Leonardo Rufiner

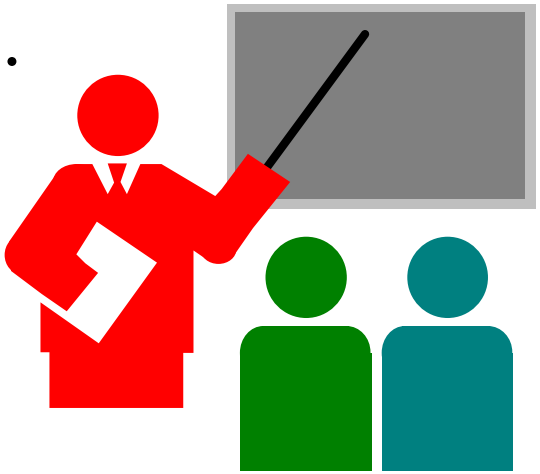
Doctorado en Ingeniería

CONICET

FICH UNL

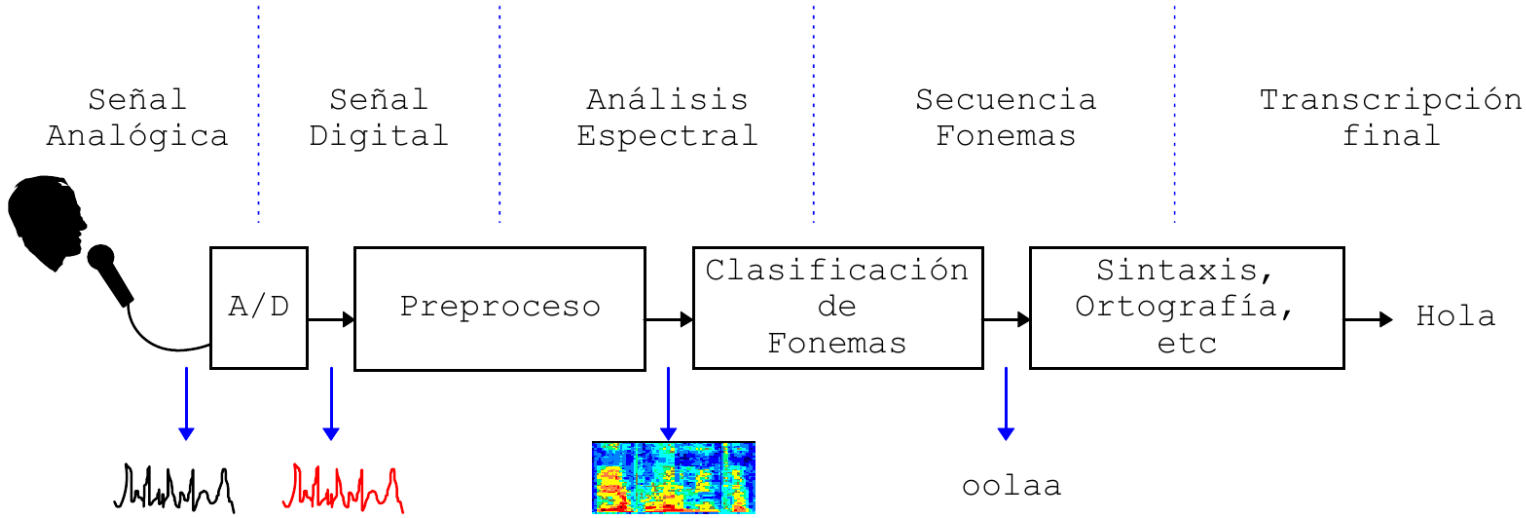
Temas a Tratar

- Introducción a las técnicas clásicas.
- Análisis de energía por bandas (FB).
- Modelos de predicción lineal (LPC).
- Coeficientes cepstrales en escala de mel (MFCC).
  - Deltas (velocidad y aceleración).
- Métodos de cuantización vectorial.
- Otros:
  - PLP, RASTA
  - Modelos auditivos.



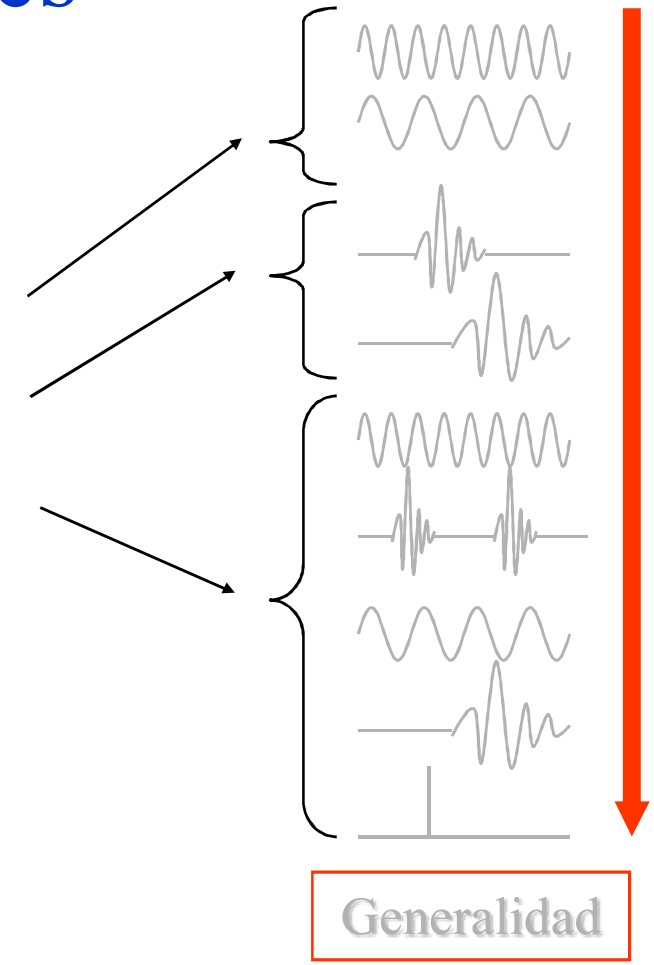
Objetivos

- Presentar los fundamentos de las técnicas clásicas para análisis de la señal de voz.
- Analizar la robustez de las mismas y comprender sus alcances y limitaciones.
- Poder aplicarlas en la implementación de un sistema completo de RAH.



Análisis de señales

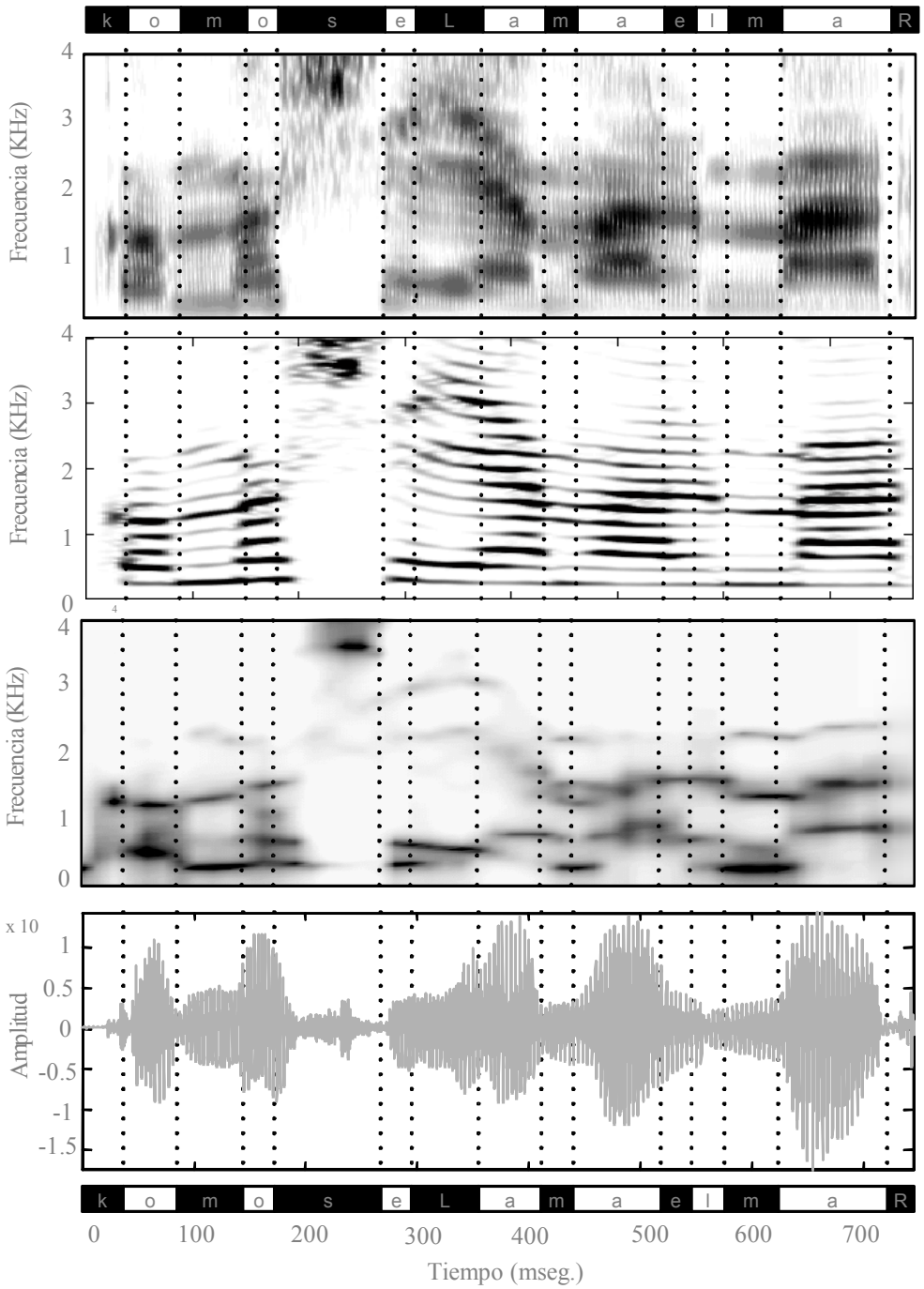
- Lineal invariante en el tiempo (*Fourier*).
- Lineal no estacionario (*STFT*, *onditas*).
- No lineal y/o no estacionario:
  - *Distribuciones tiempo-frecuencia*.
  - *Representaciones basadas en diccionarios.*
    - *Representaciones ralas y/o independientes.*

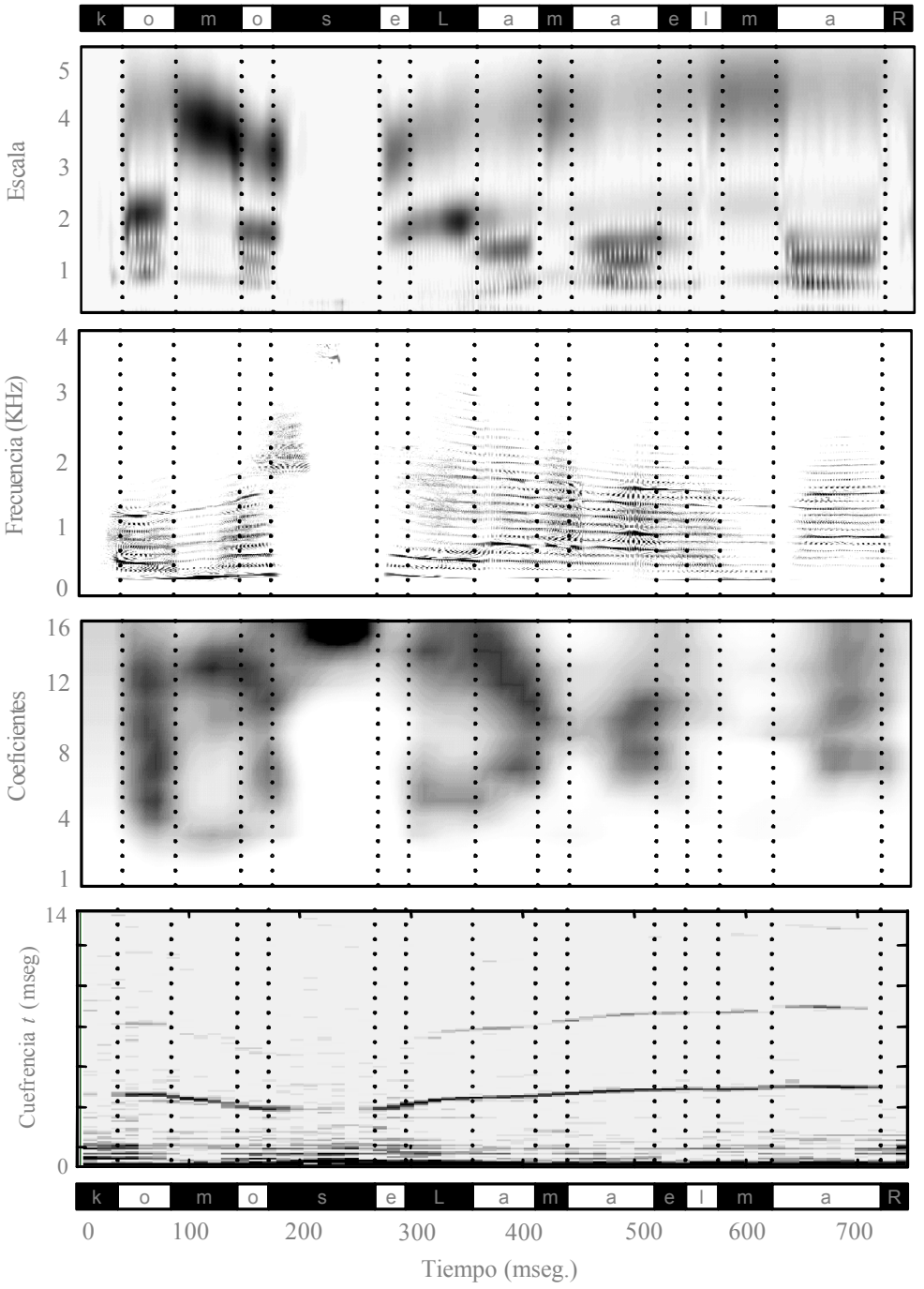


- Otros:
  - Específicos señal habla
    - *Cepstra*, *LPC*, PLP, RASTA, *modelos oído*

Estado del arte

    - F0 y formantes…
- ¿Robustez?

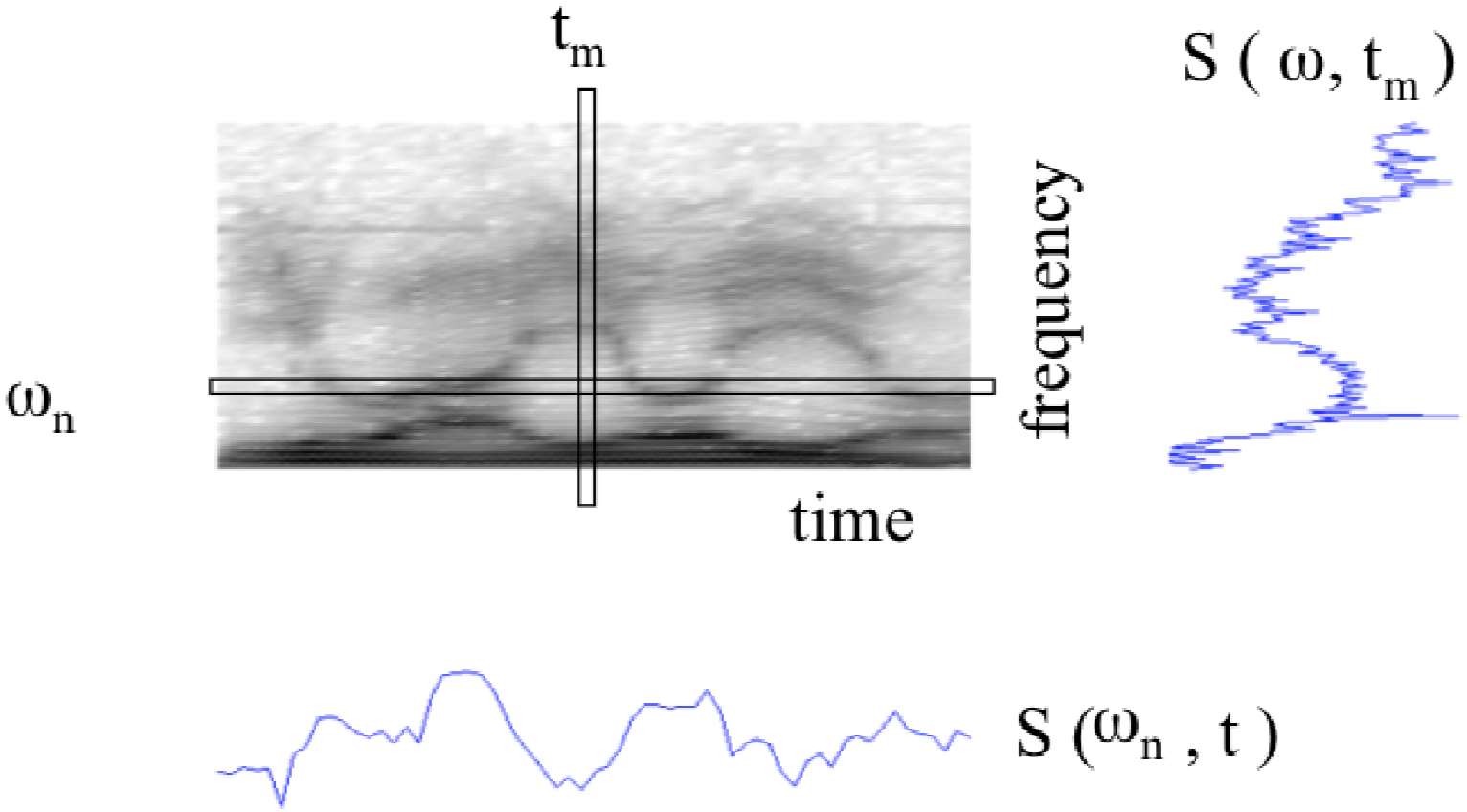

k o m o s e L a m a e l m a R
Frecuencia (KHz)
Frecuencia (KHz)
Frecuencia (KHz)
x 10
Amplitud
0 100 200 300 400 500 600 700
Tiempo (mseg.)


k o m o s e L a m a e l m a R
Escala
Frecuencia (KHz)
Coeficientes
Cuefrencia t (mseg)
0 100 200 300 400 500 600 700
Tiempo (mseg.)

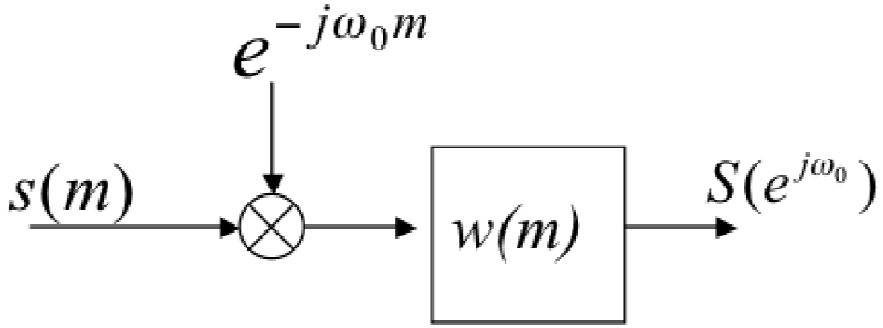
Predicción Lineal Perceptual (PLP)

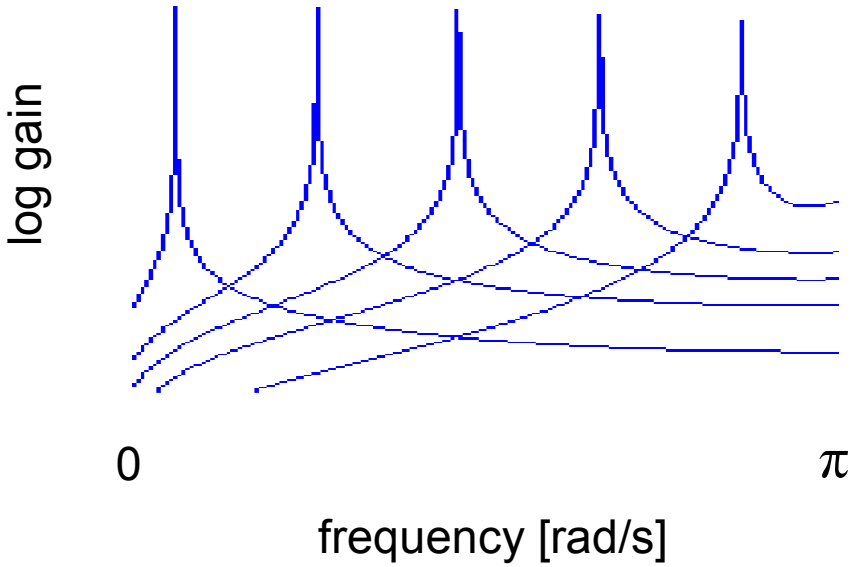
H. Hermansky (1990)

Spectrogram – 2D representation of sound



Short-term Fourier analysis

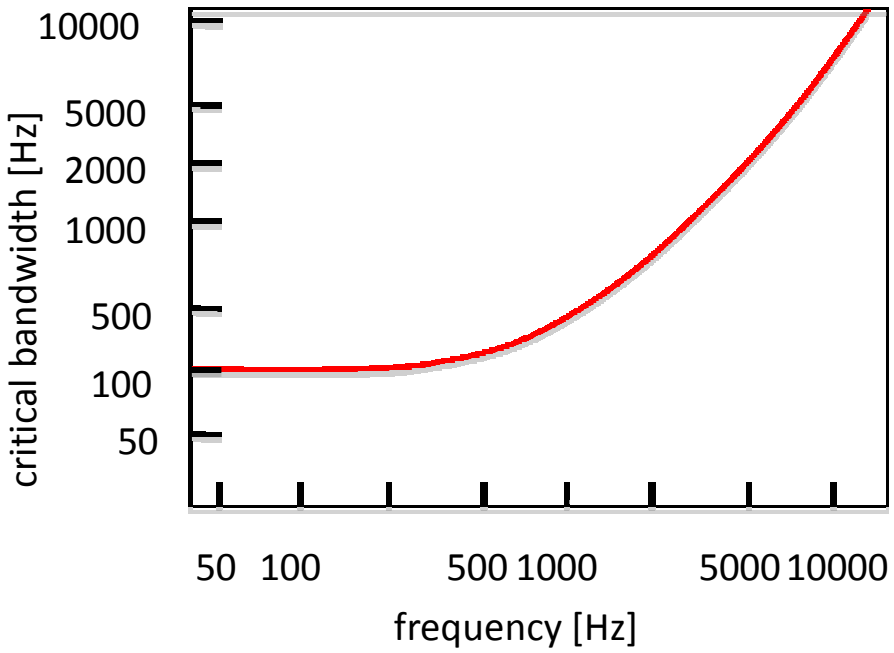




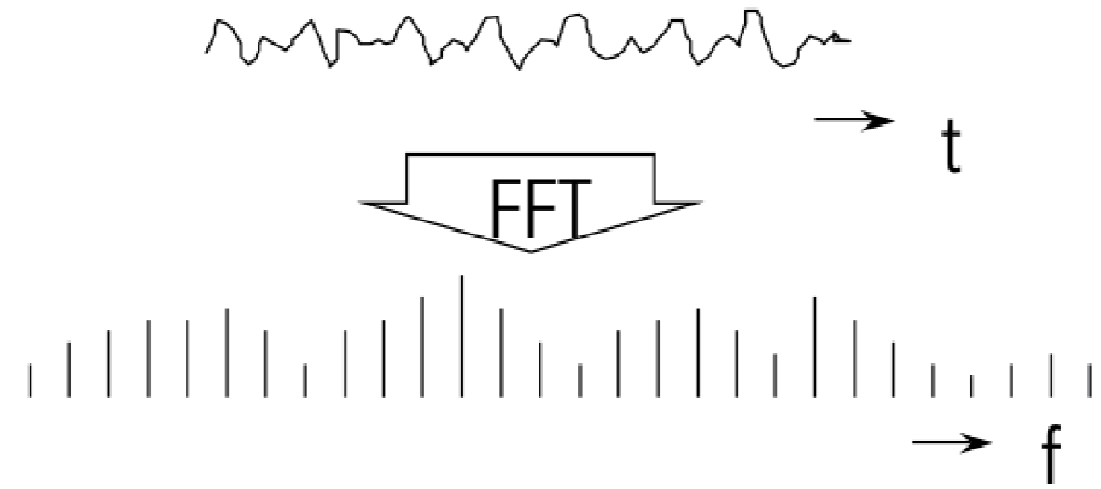
- Short-term Fourier analysis
  - filterbank with filter impulse responses given by the shape of the analysis window
  - equal bandwidth analysis (all filters with identical bandwidths)

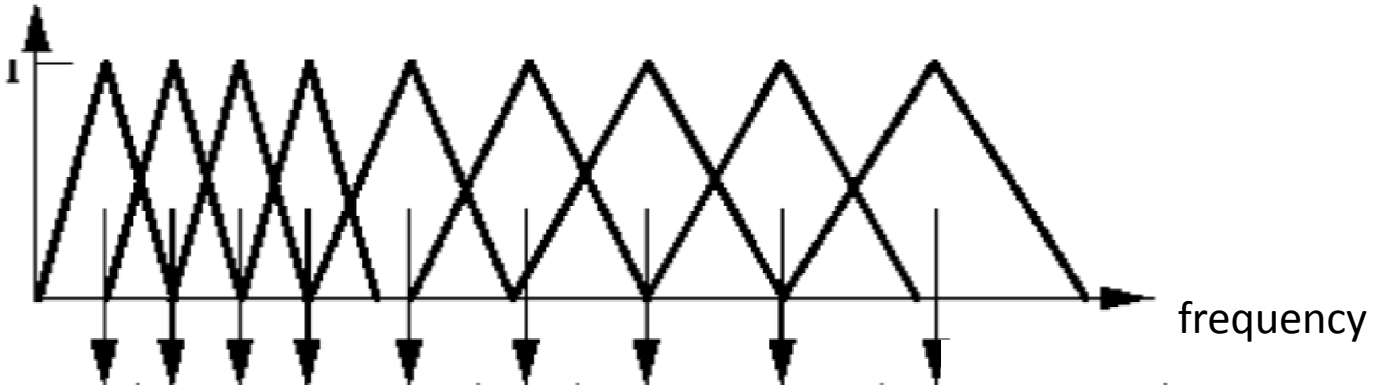
Spectral resolution of hearing

spectral resolution of hearing decreases with frequency
(critical bands of hearing, perception of pitch,…)



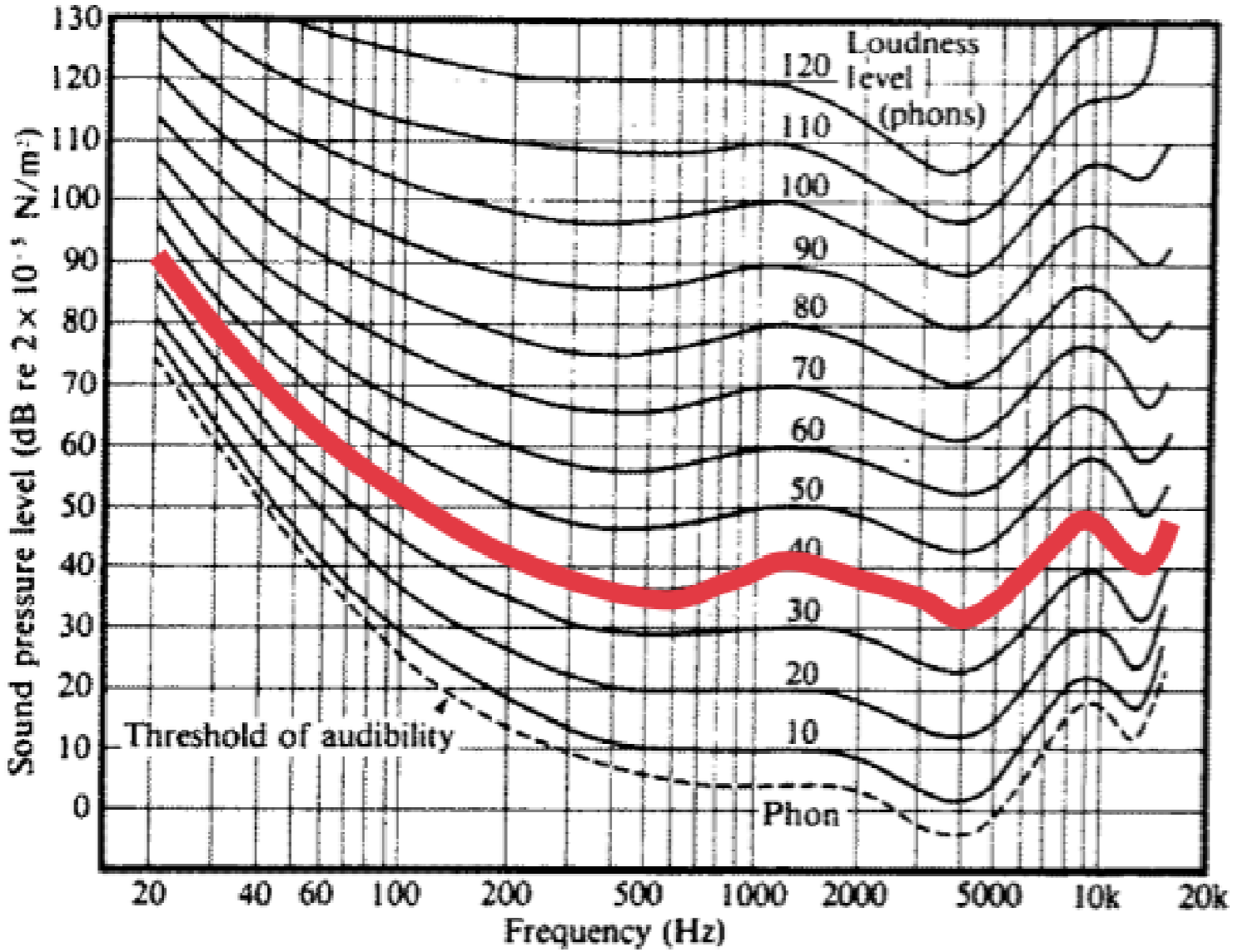
"Poor man's" way of emulating critical-band filter





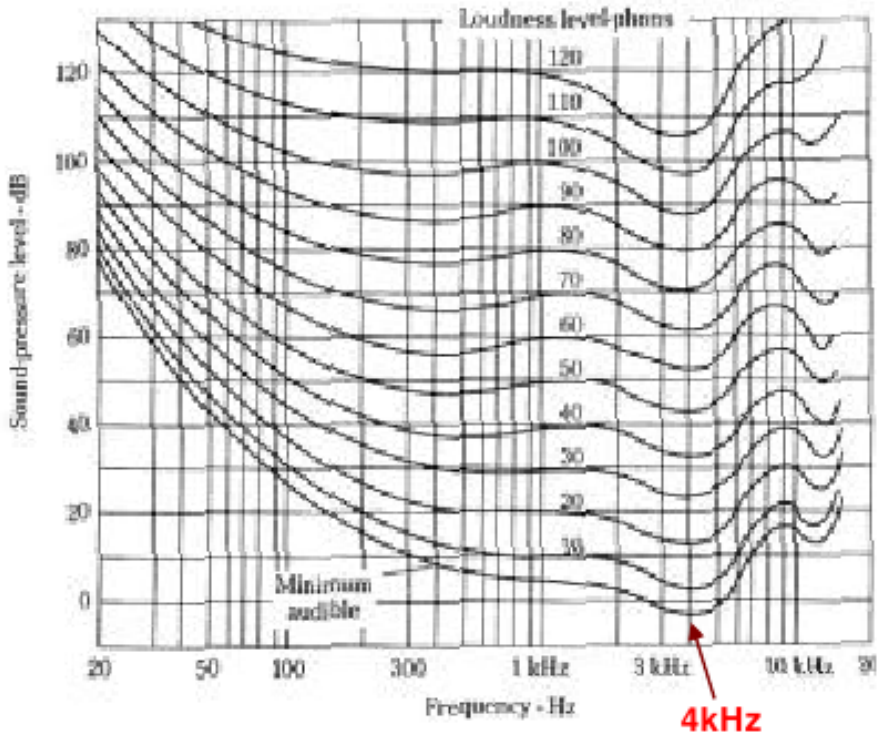
energies in "critical bands"

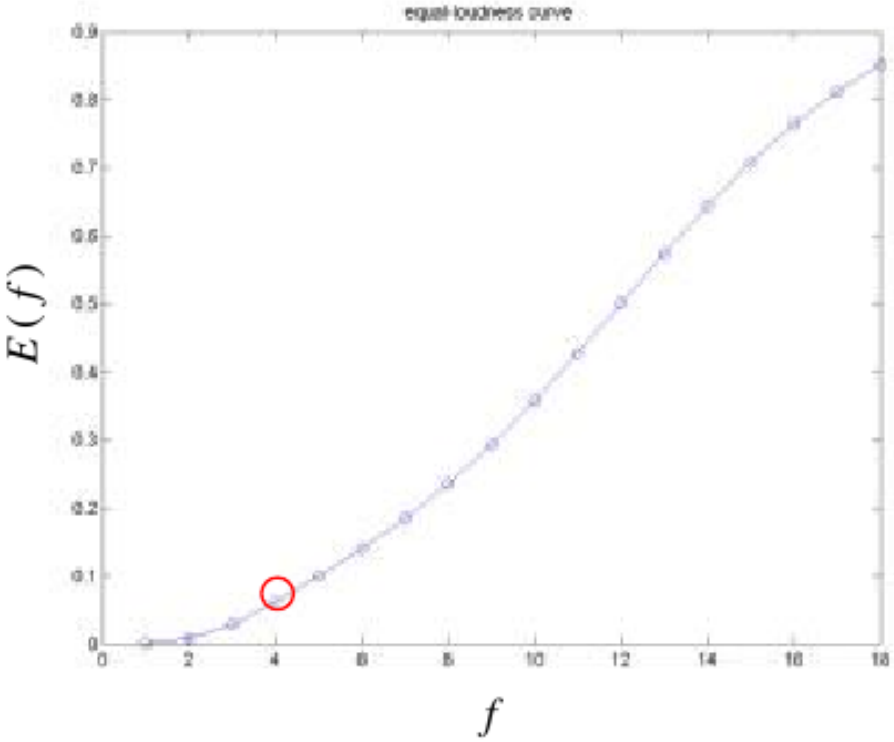
Sensitivity of hearing depends on frequency

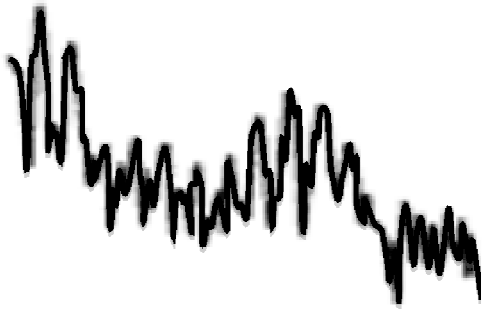


Equal loudness preemphasis:

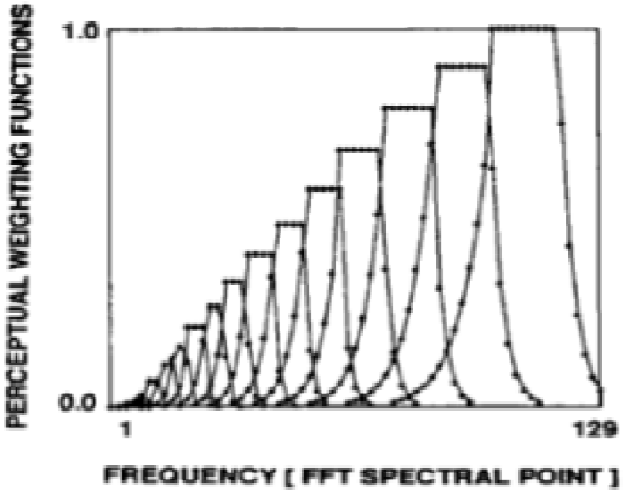
$$E(f) = \frac{(f^2 + 1.44e6) f^4}{(f^2 + 1.6e5)^2 (f^2 + 9.61e6)}$$



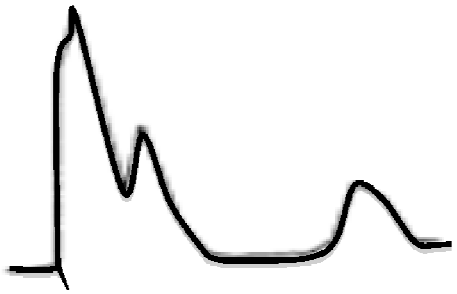




spectrum



summation windows

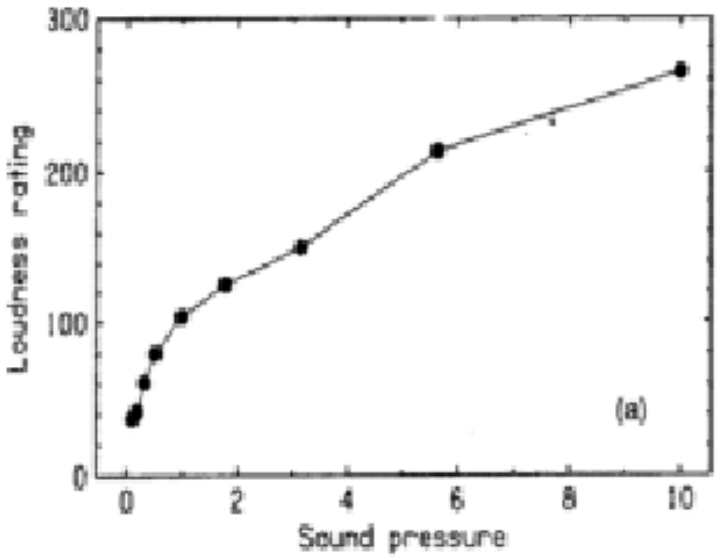


spectrum with auditory-like resolution

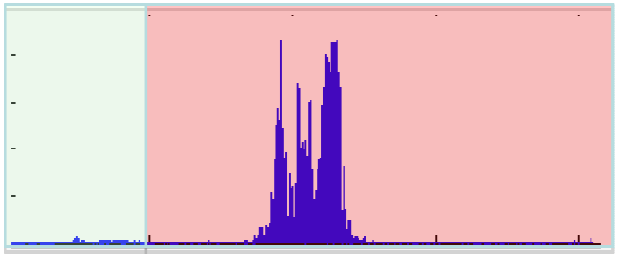
Intensity-loudness power law

Perceived loudness, $L(w)$, is approximately the cube root of the intensity, $I(w)$

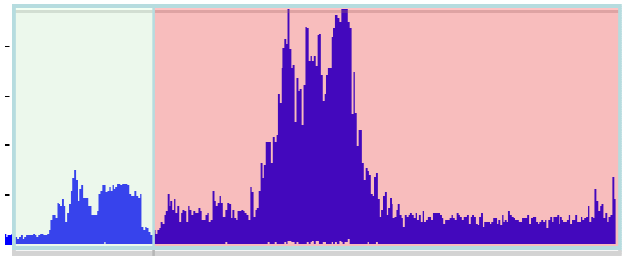
$$L(w) \cong I(w)^{\frac{1}{3}}$$

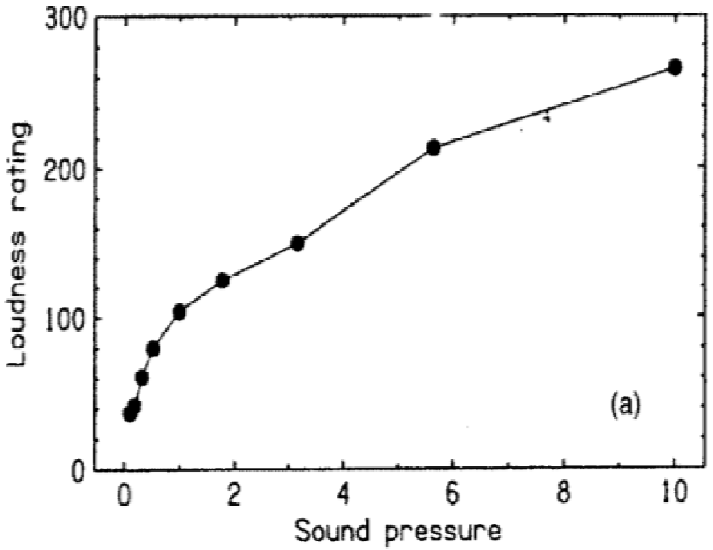


intensity ≈ signal ^{2} [w/m^2]

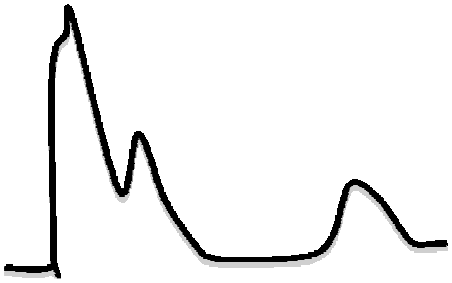


loudness [Sones]



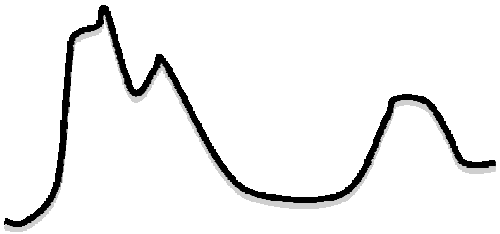


loudness = intensity $^{0.33}$



intensity
(power spectrum)

$|.|^{0.33}$

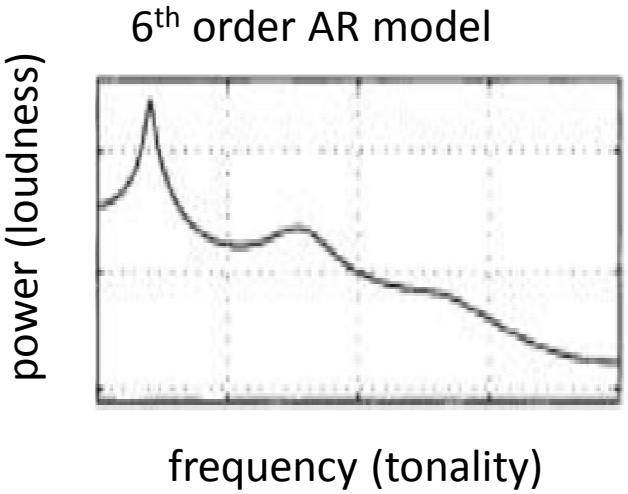


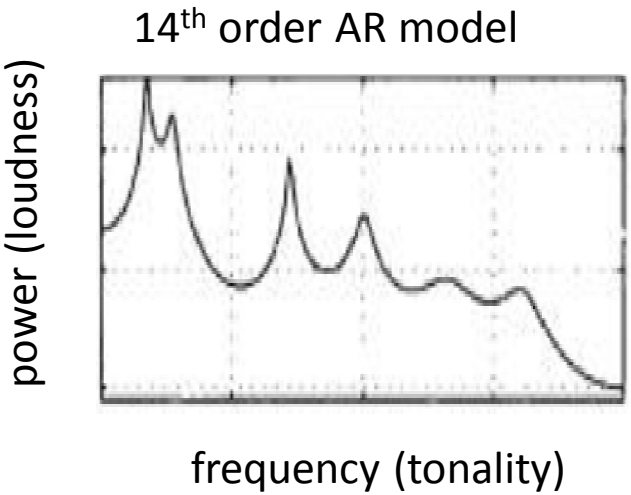
loudness

Not all spectral details are important

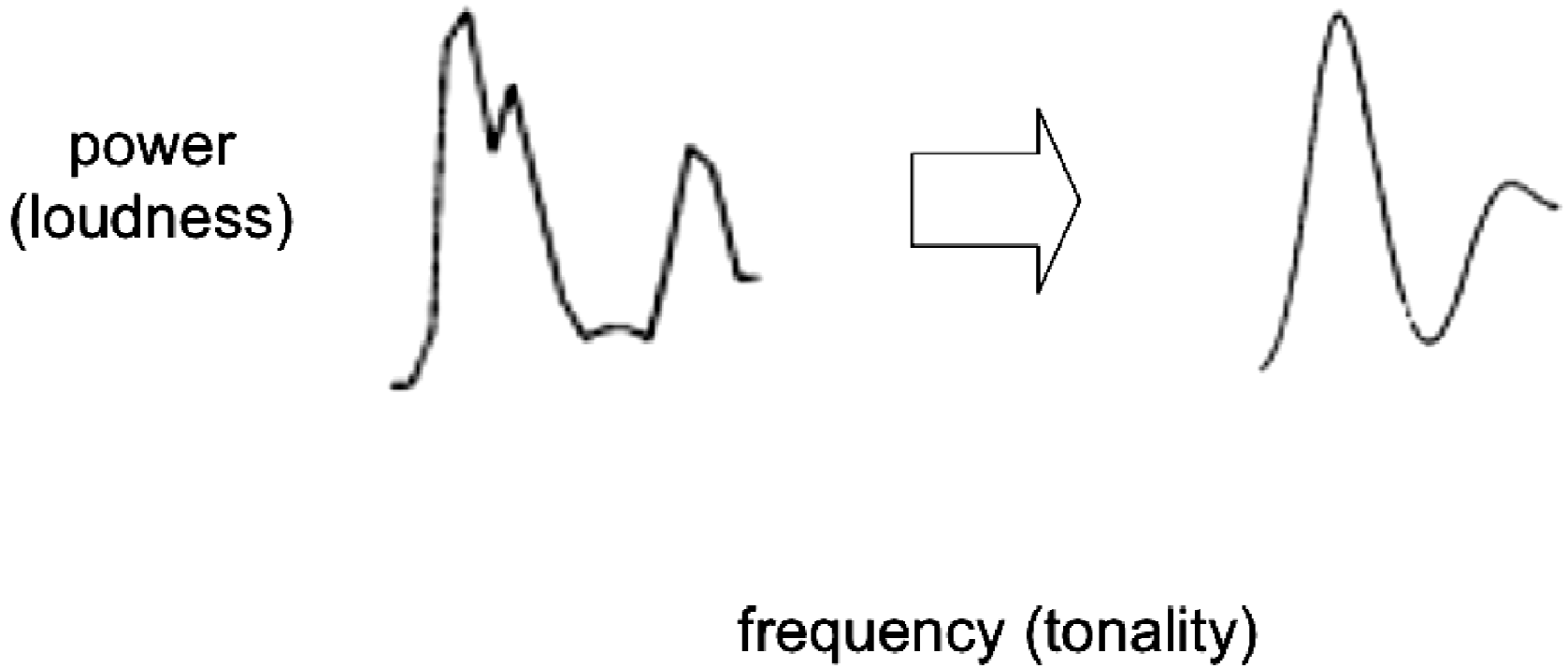
a) compute Fourier transform of the logarithmic auditory spectrum and truncate it (Mel cepstrum)

b) approximate the auditory spectrum by an autoregressive model (Perceptual Linear Prediction – PLP)





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



Perceptual Linear Prediction

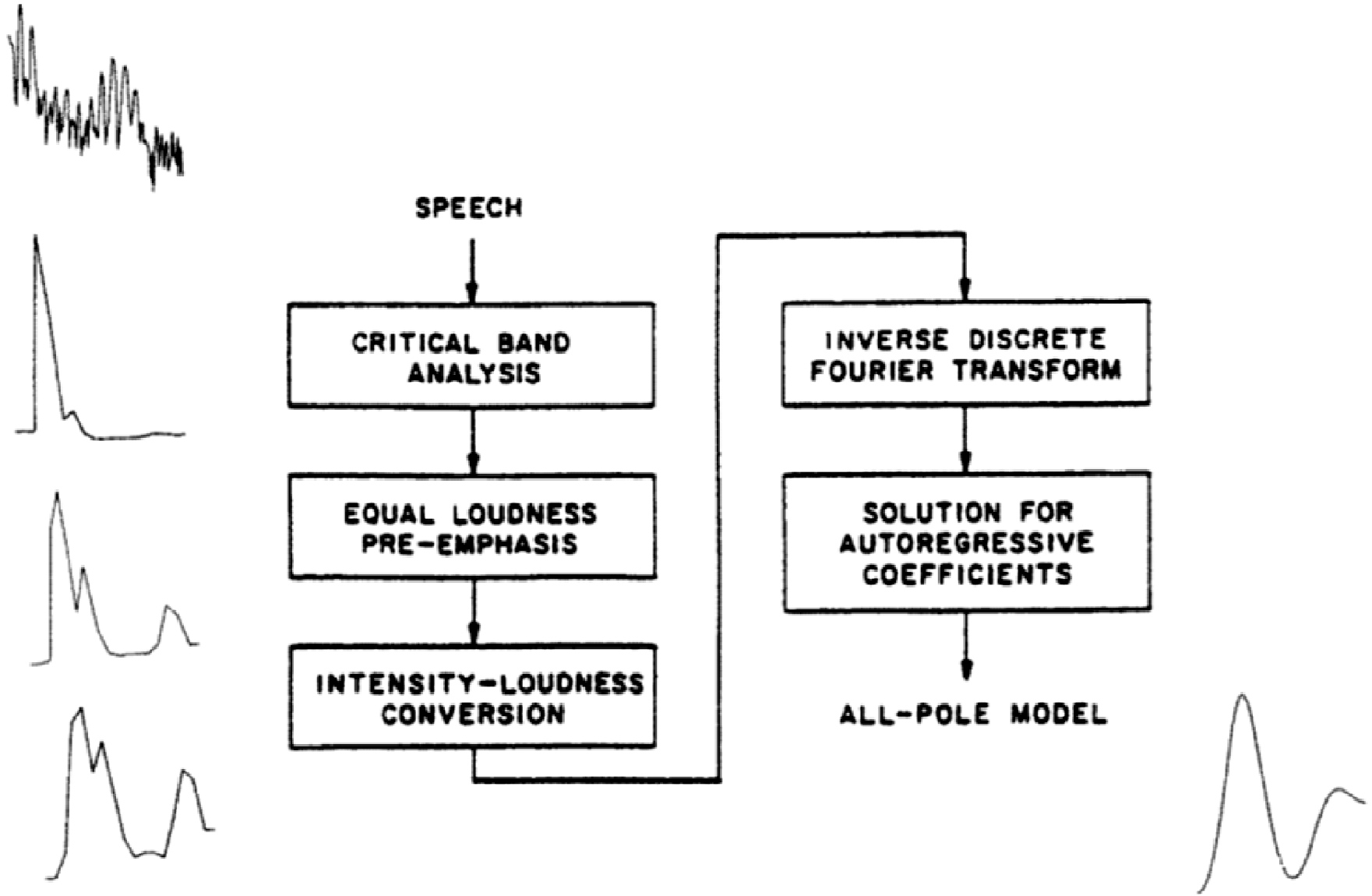
Limited spectral resolution

formant clusters as may be interpreted by auditory perception

Perceptual Linear Prediction (PLP)

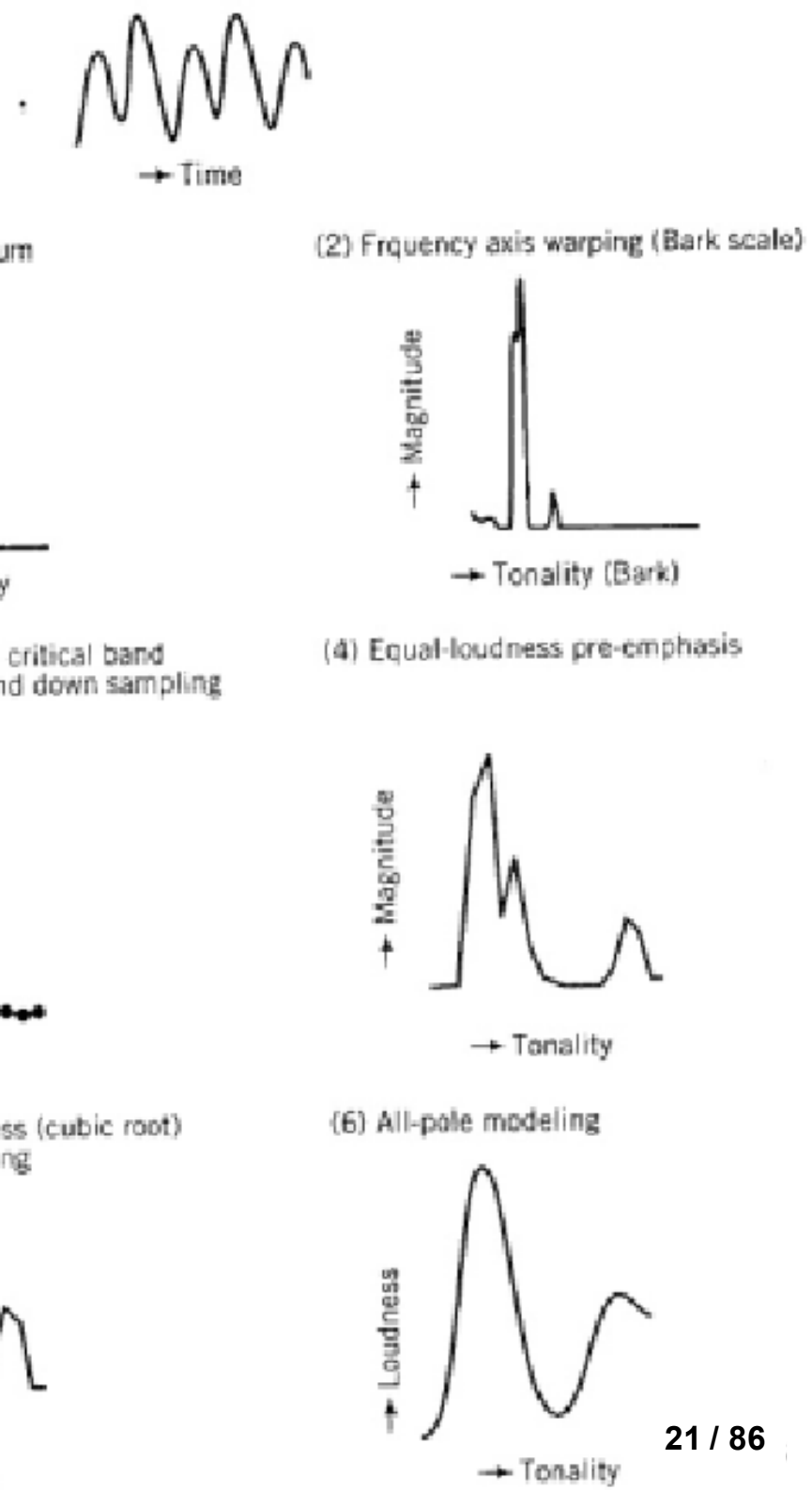
critical-band (Bark) spectral analysis
loudness domain (cubic root of intensity)
equal loudness curve (at 40 dB)
autoregressive spectral fit (fits well at peaks)

PLP calculation

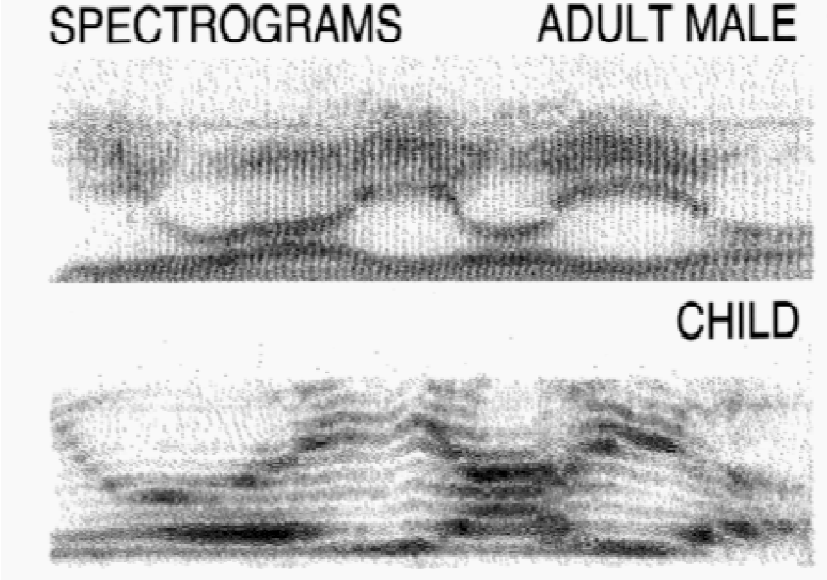


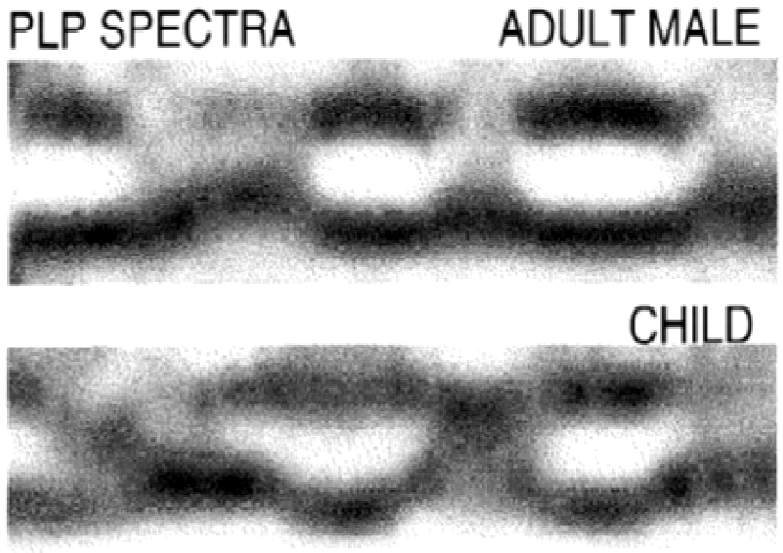
PLP calculation

- Succesive domain transformations
  ….

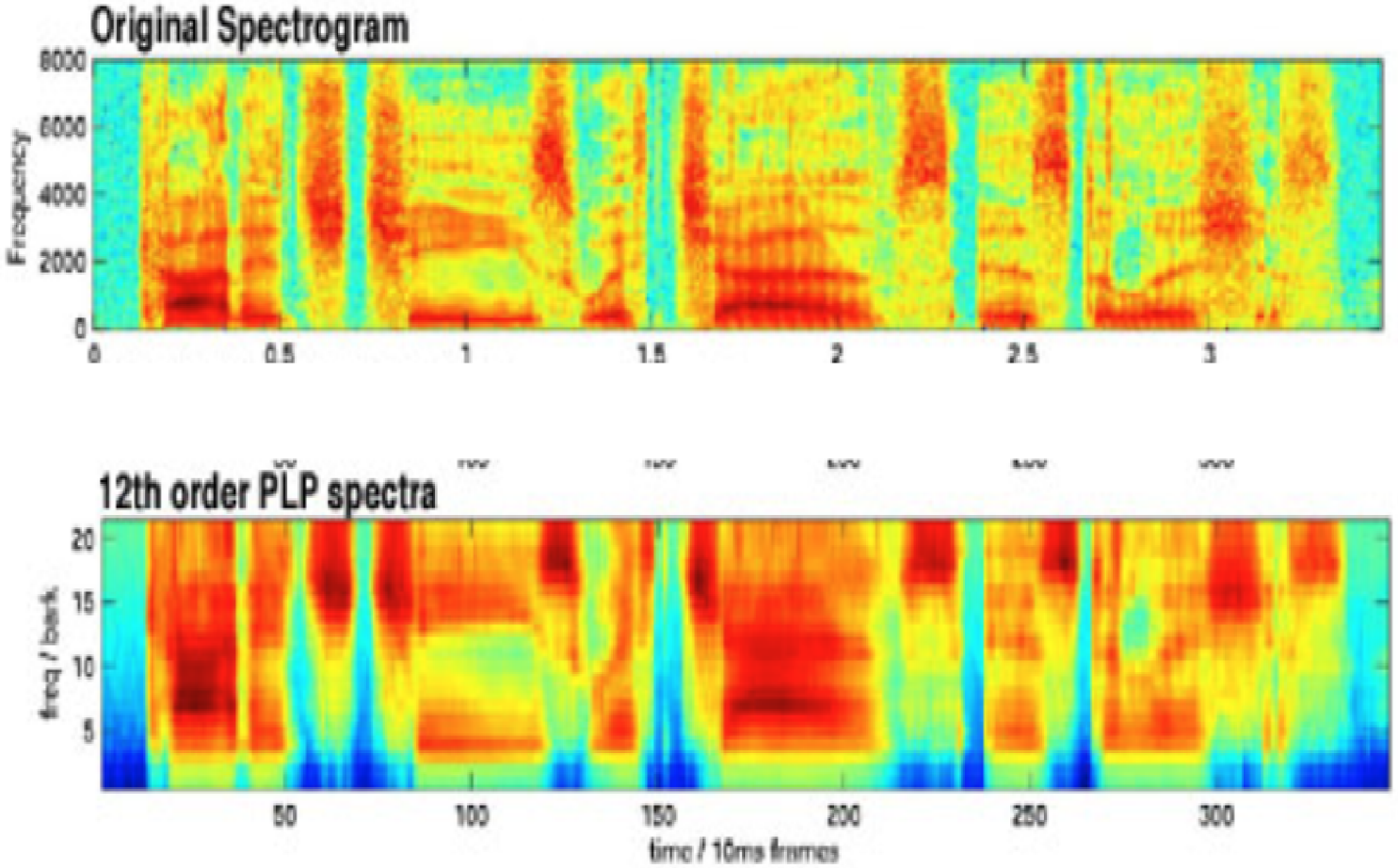


Spectral Comparison



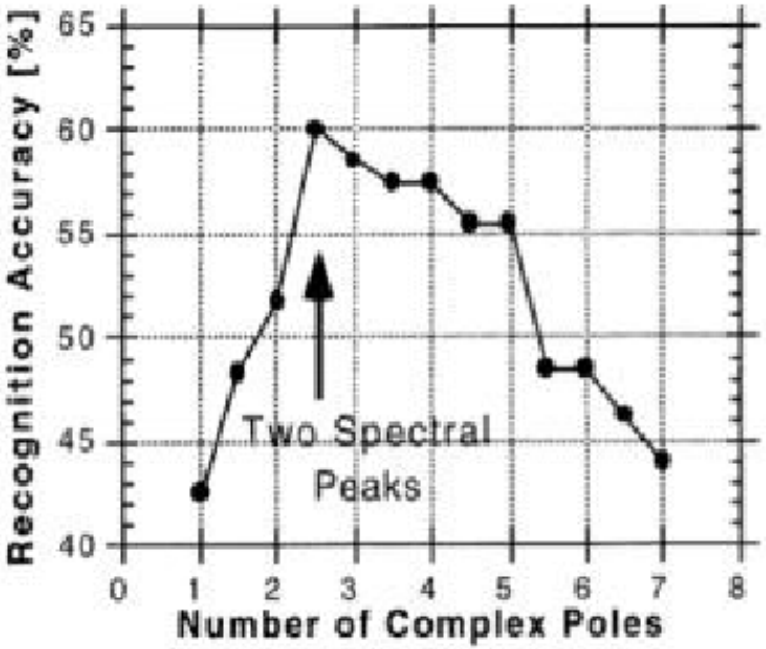


Spectral Comparison



Current state-of-the-art speech recognizers typically use high model order PLP

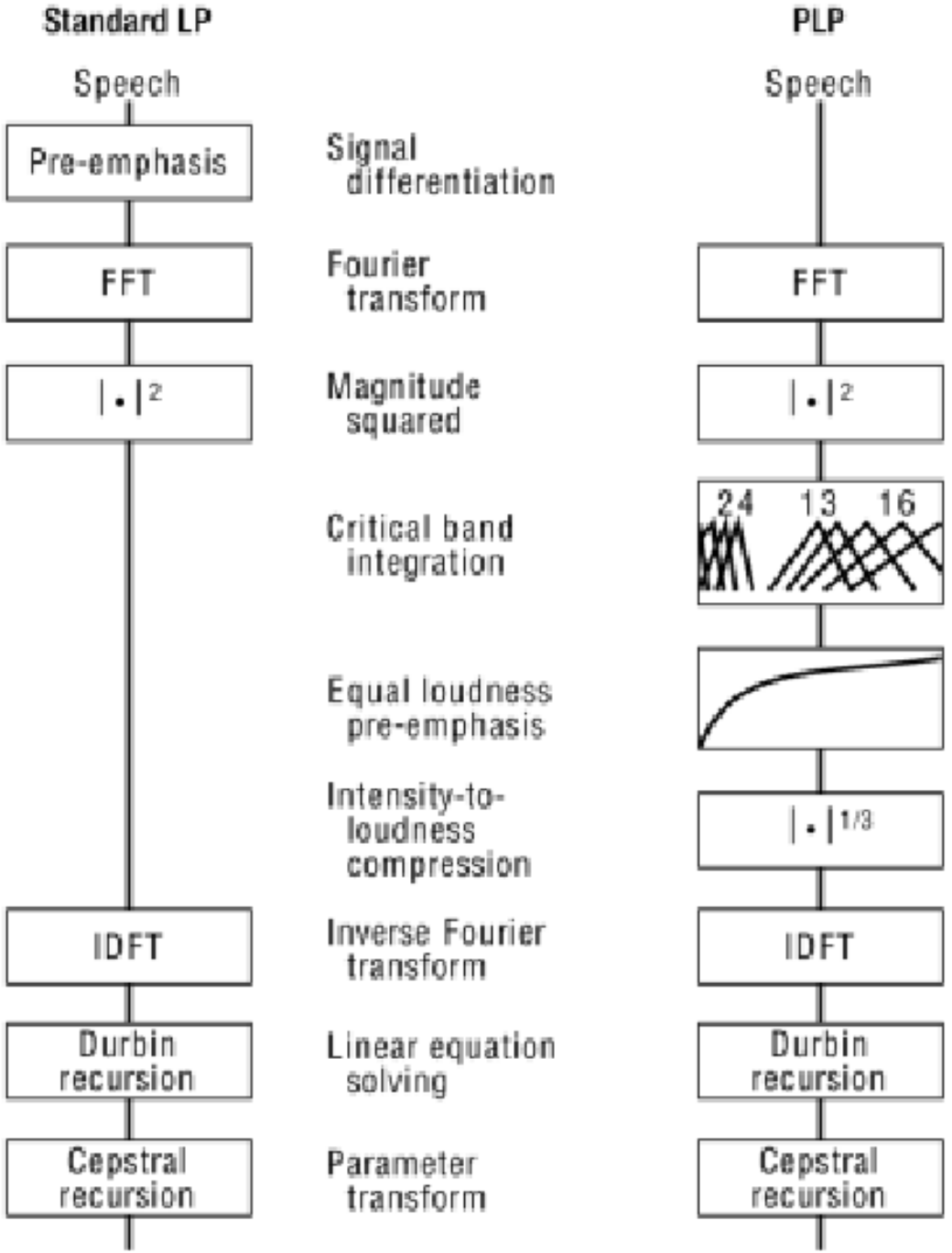
Optimal Amount of Spectral Smoothing (order of PLP autoregressive model)



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

LP and PLP based Cepstral comparison

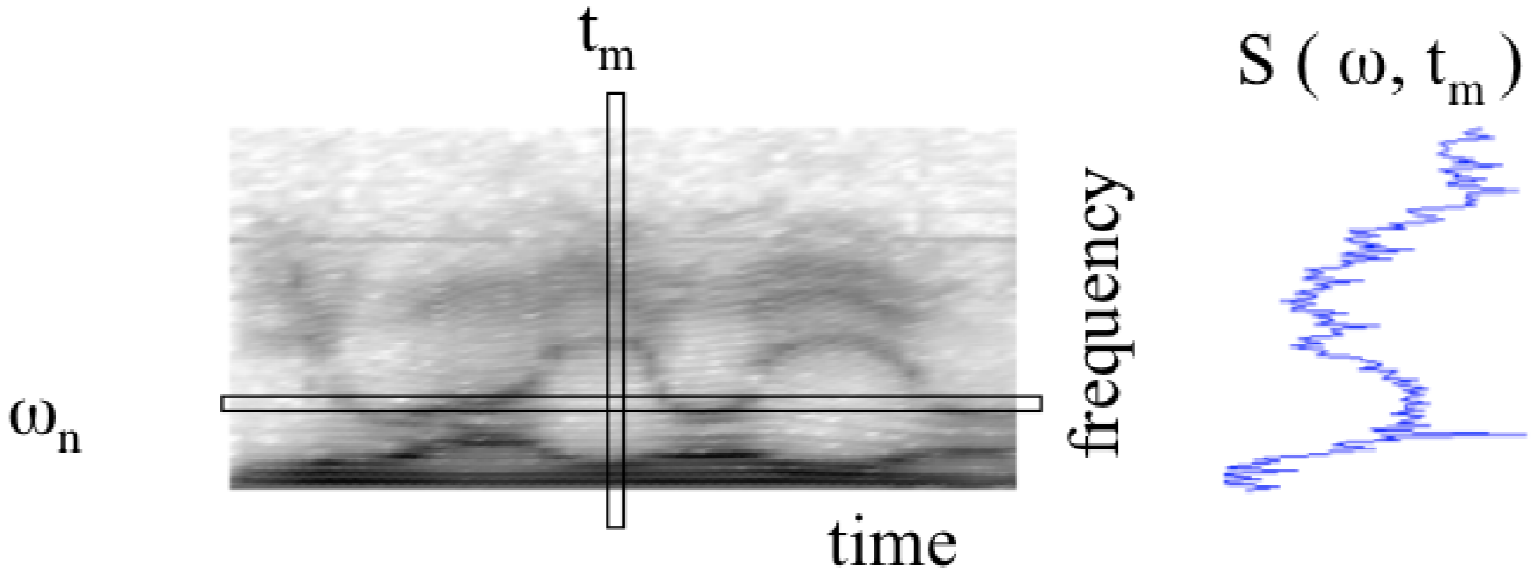
- Cepstral coefficients can be recursively calculated starting from PLP, in similar way as in the LP case.



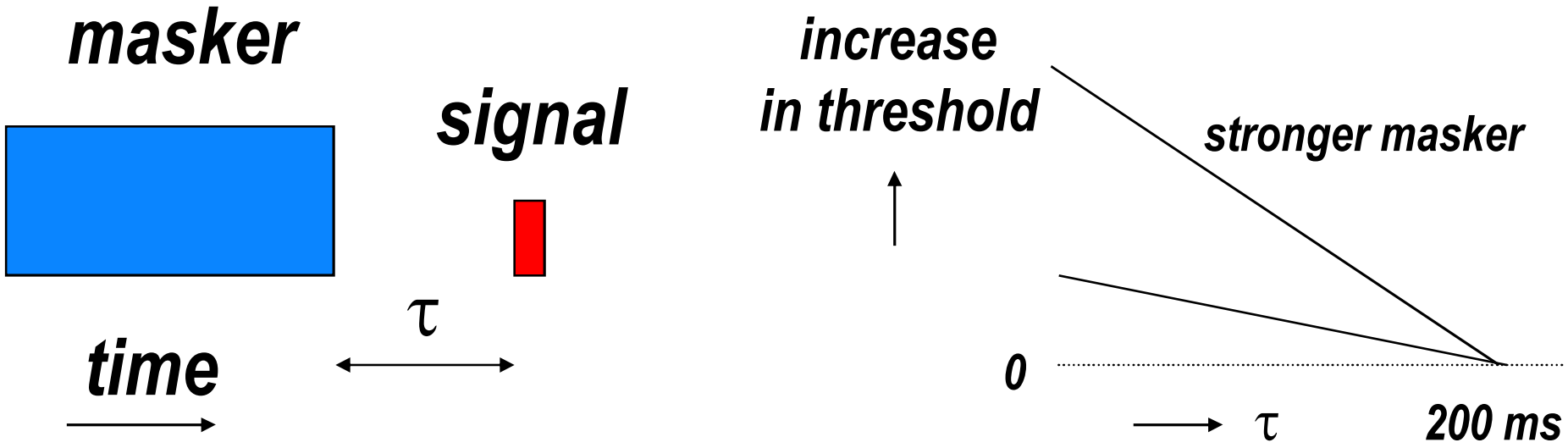
RelAtive SpecTrAl (RASTA)

Hermansky (1994)

It’s about time
(to talk about TIME)



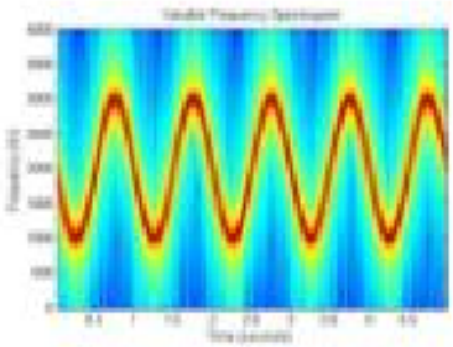
Masking in Time

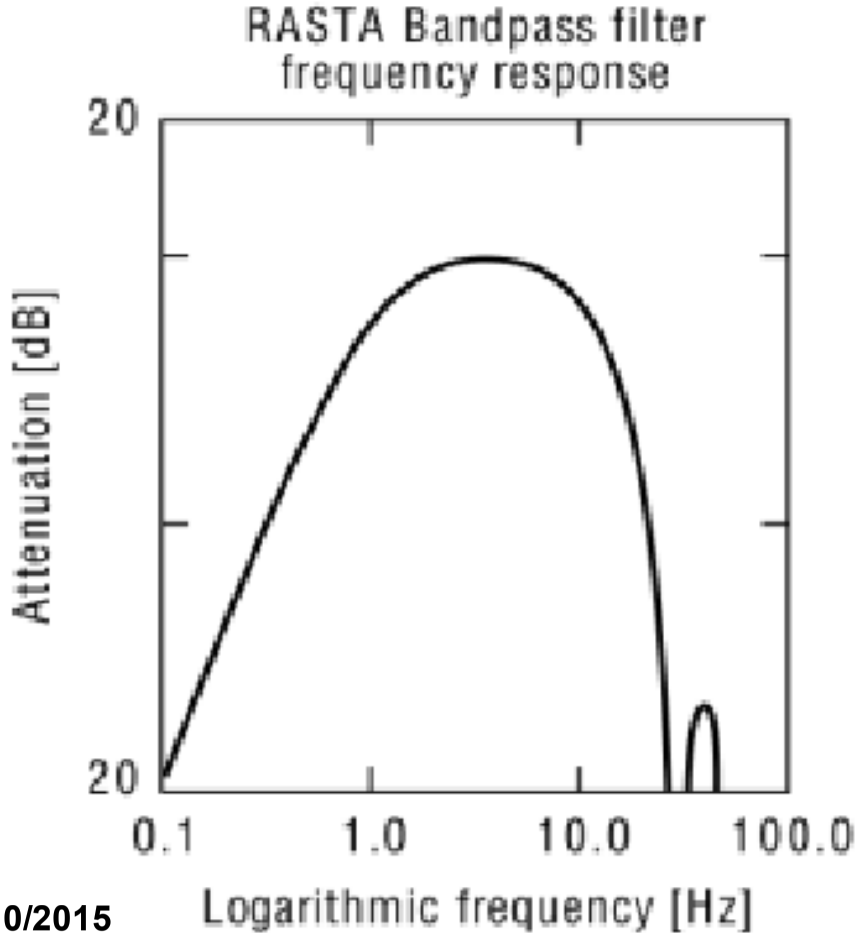


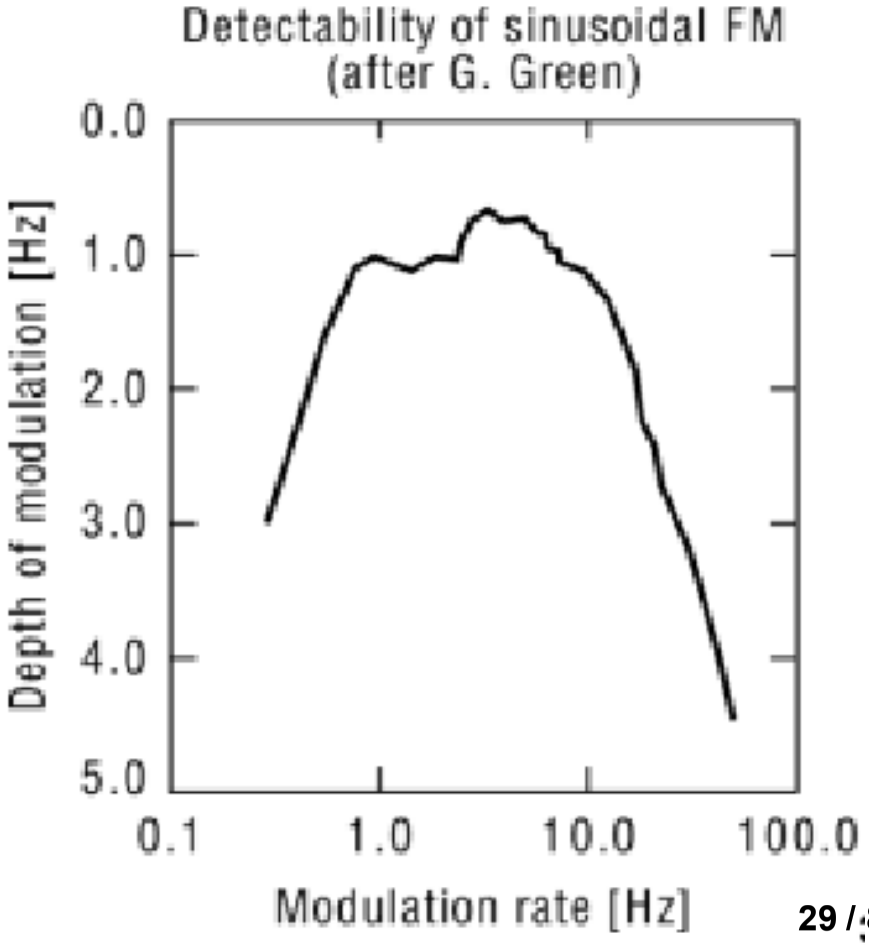
- suggests ~200 ms buffer (critical interval) in auditory system

Rate of Spectrum Change Sensibility

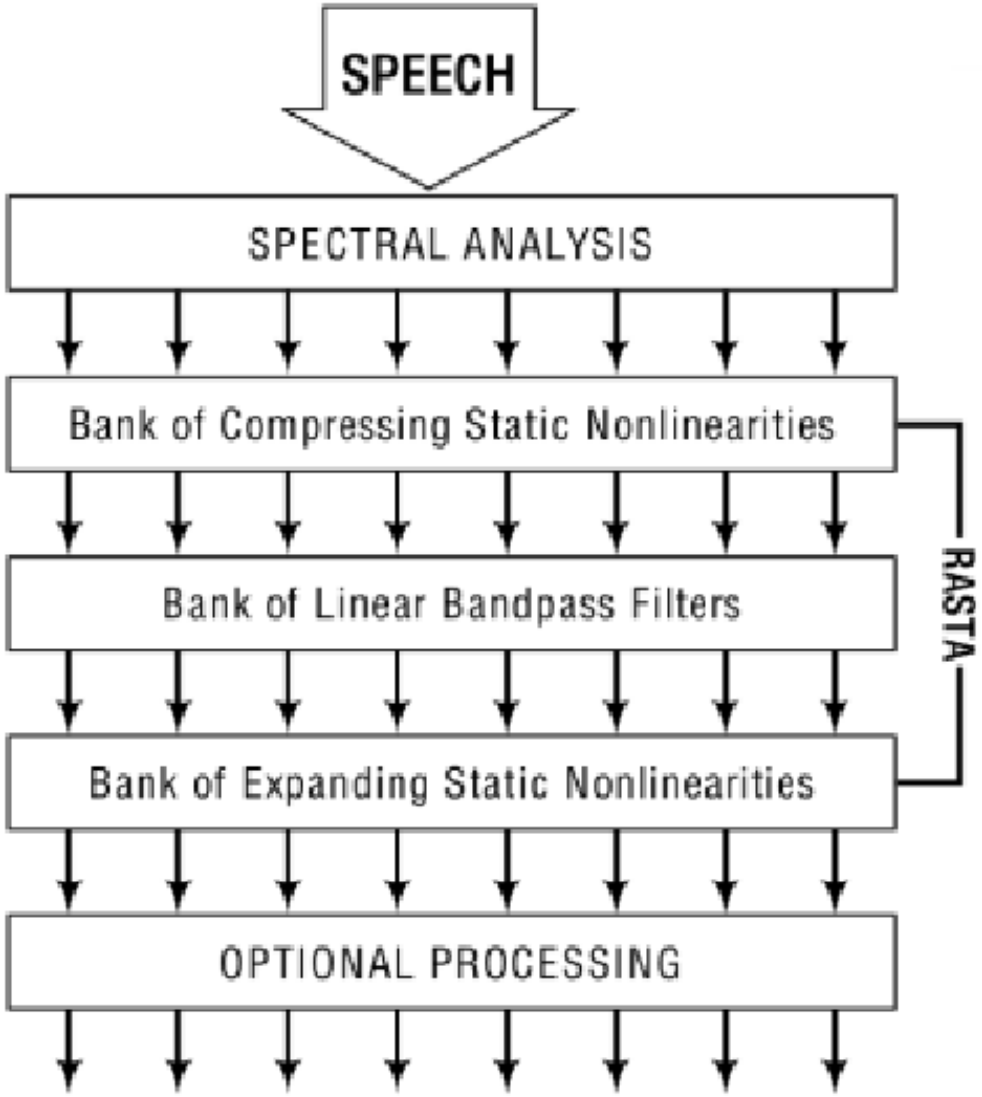
$$x(t) = A_c \cos[\omega_c t + \phi_c + A_m \sin(\omega_m t + \phi_m)],$$







RASTA



time trajectories of the spectrum
in critical bands of hearing

filter with time constant > 200 ms
(temporal buffer > 200 ms)

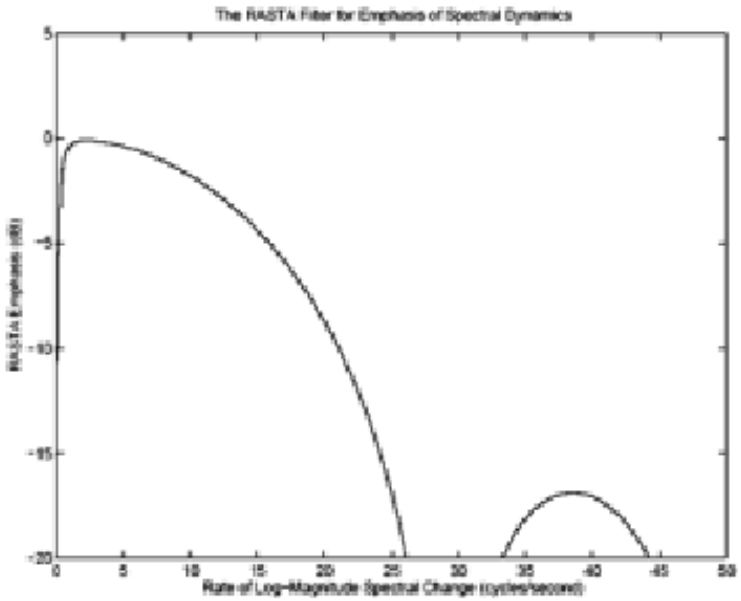
Hermansky and Morgan,
IEEE ASSP Transactions, 1994

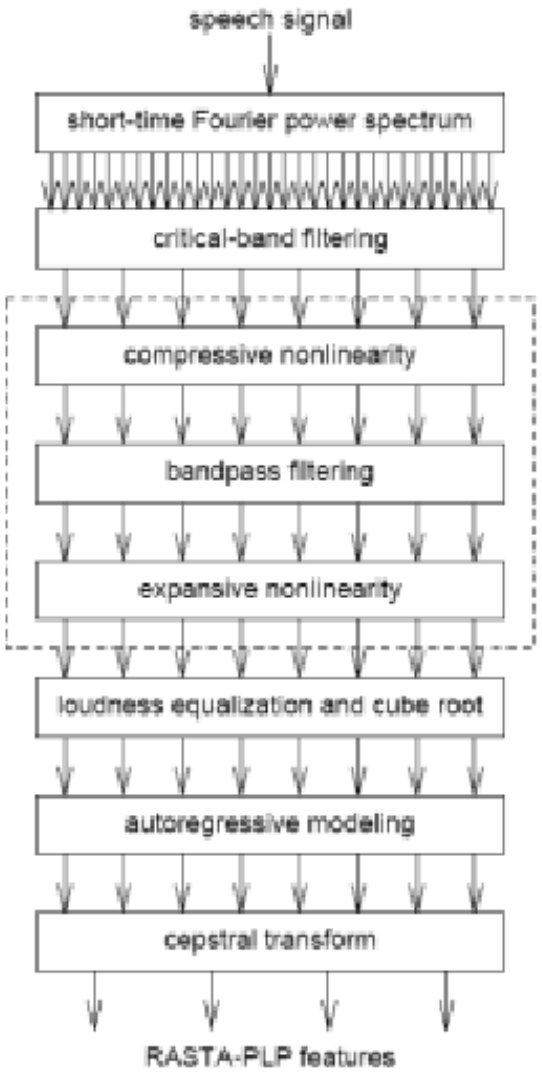
RASTA-PLP

- Modulation filtering

RASTA filter

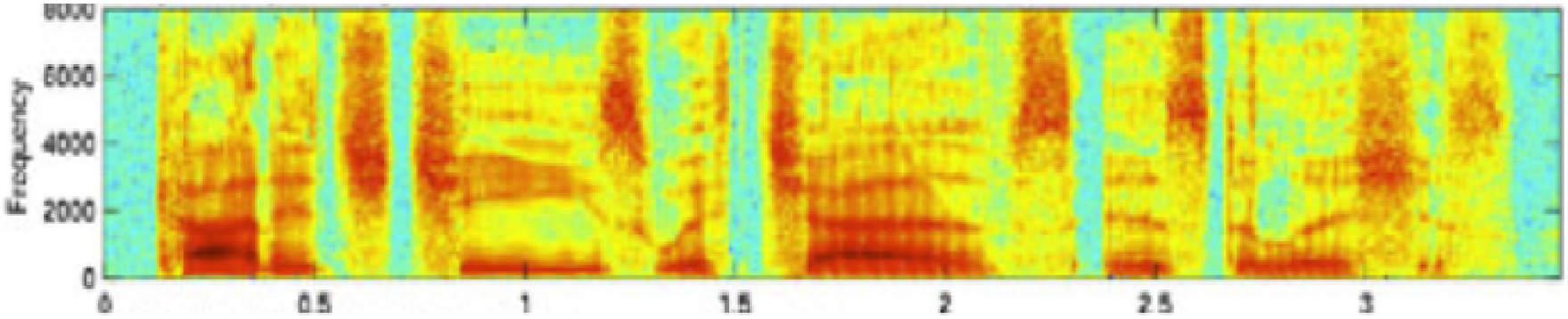
$$H(z) = 0.1 \times \frac{2z^2 + z - z - 2z^{-2}}{1 - 0.94z^{-1}}$$



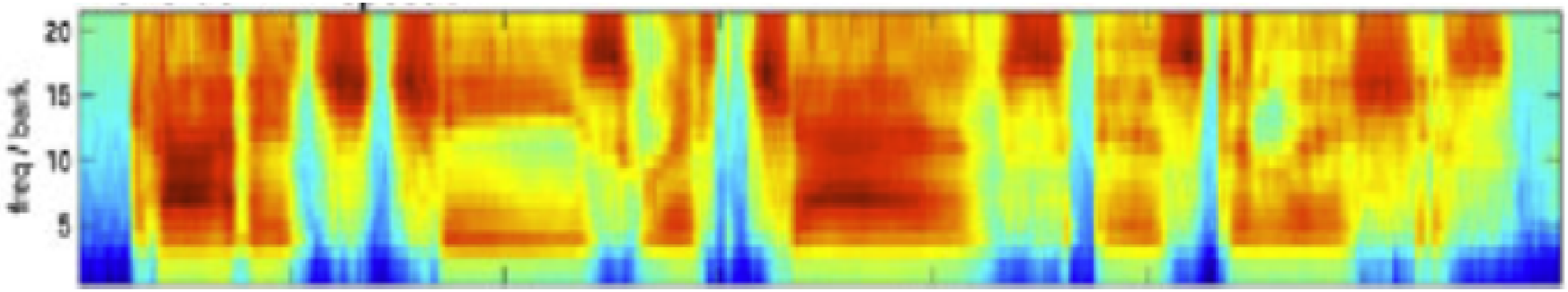


Perceptually Inspired Signal-processing Strategies for Robust Speech Recognition in Reverberant Environments, 1998

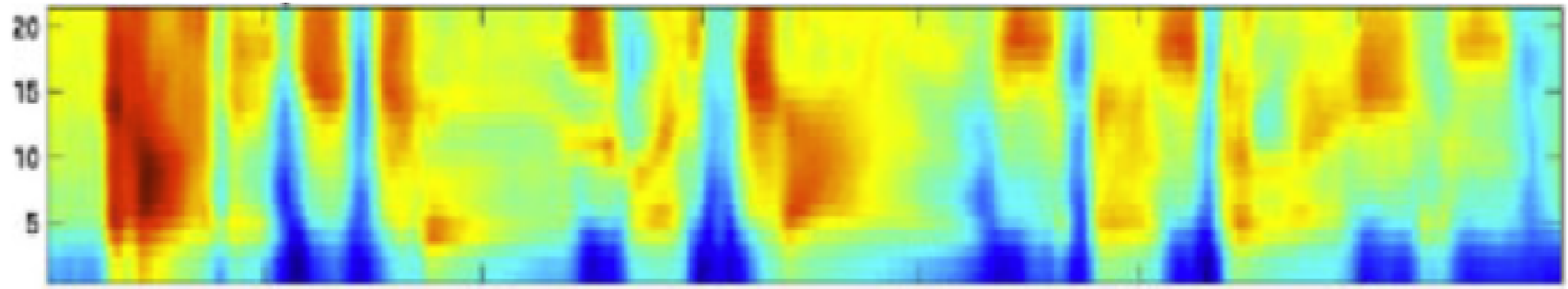
spectrogram (short-term Fourier spectrum)

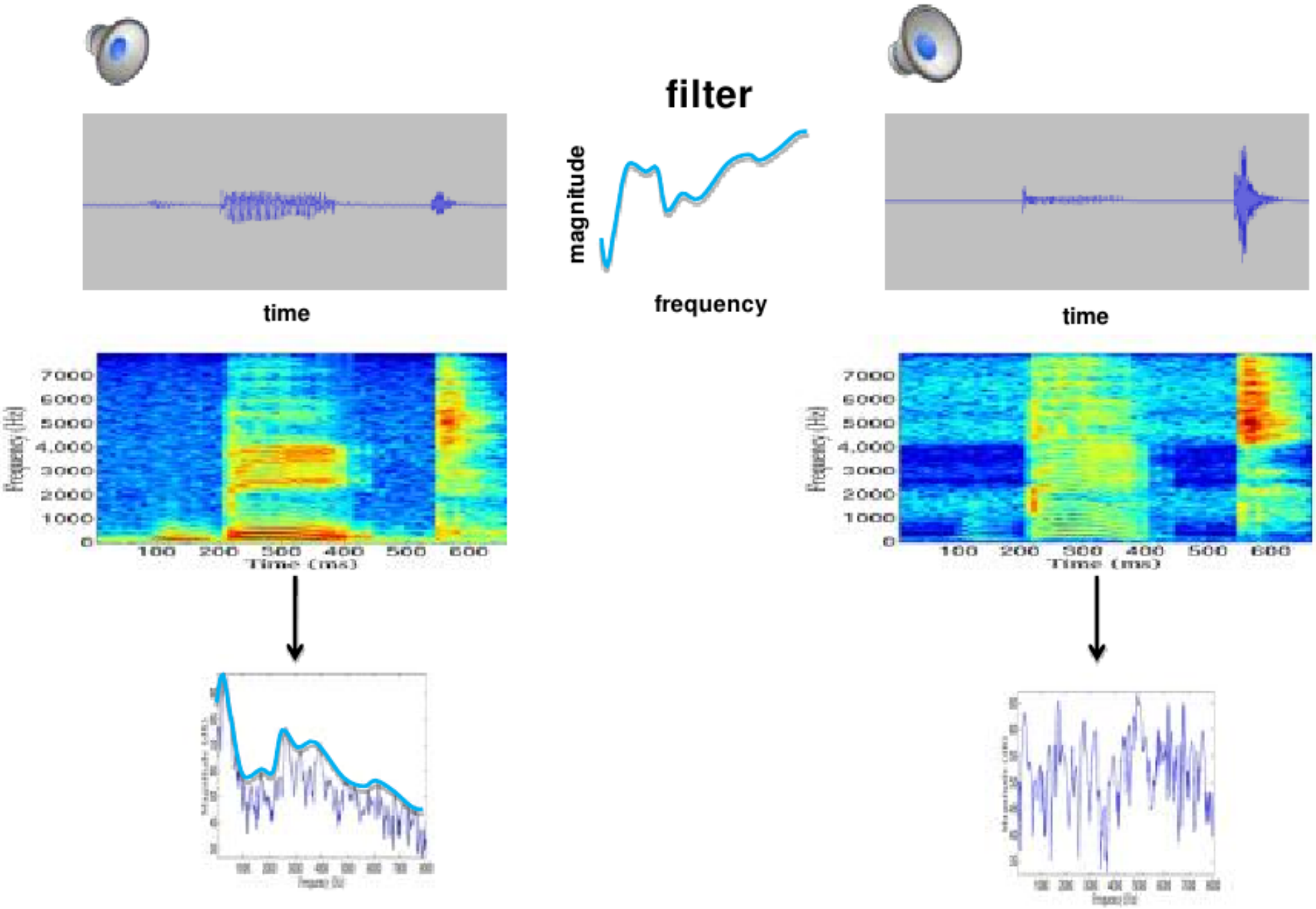


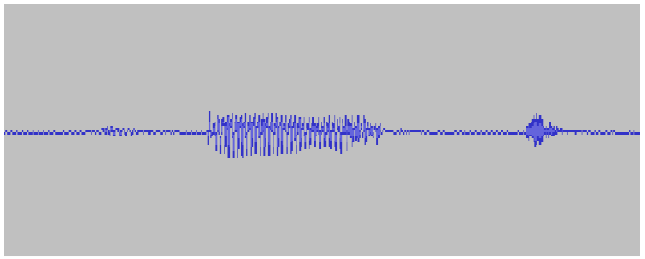
Perceptual Linear Prediction (PLP) (12th order model)



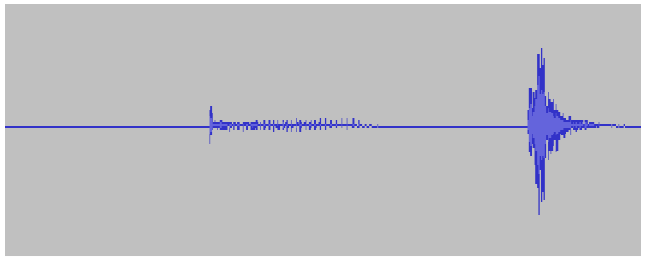
RASTA-PLP



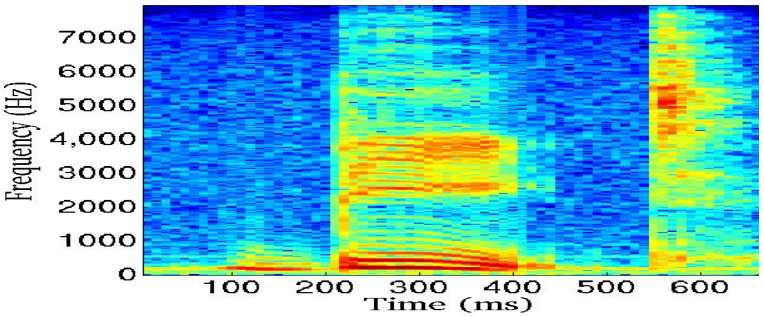

filter
magnitude
frequency
time
time
Frequency (Hz)
Time (ms)
Frequency (Hz)
Time (ms)

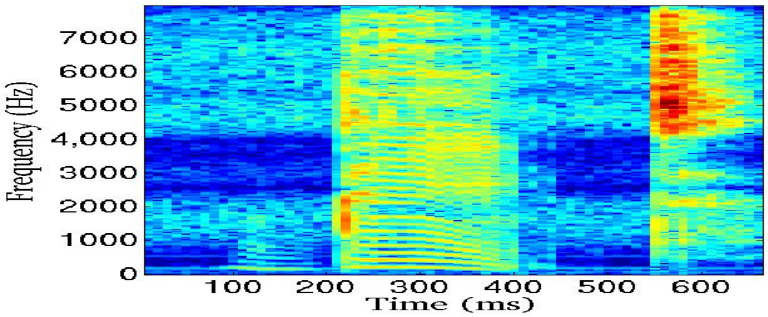


filter

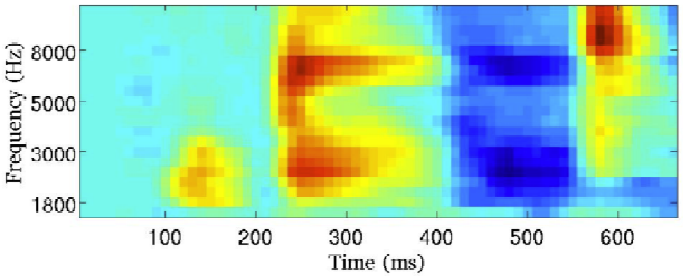


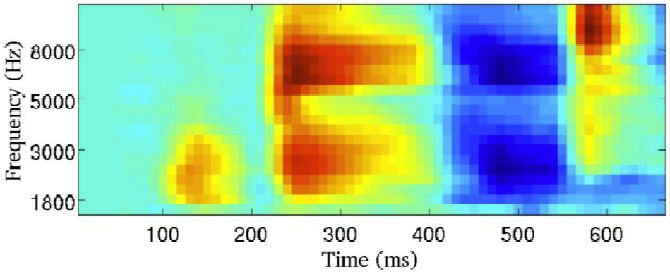
spectrogram





spectrum from RASTA-PLP

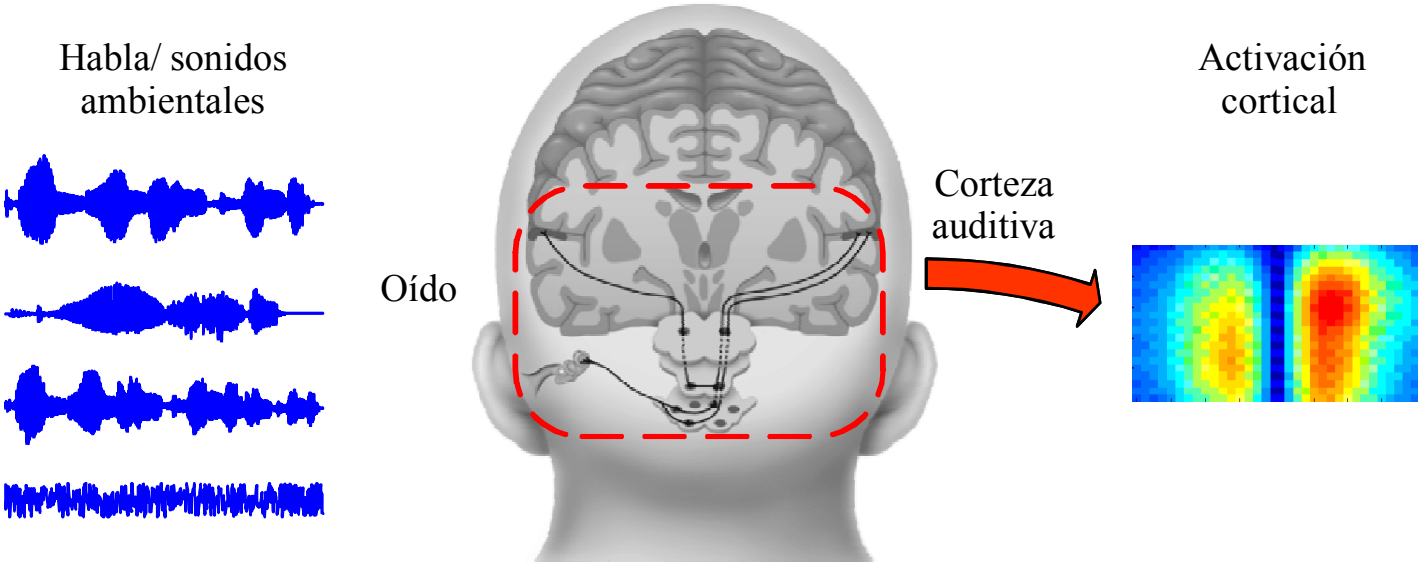




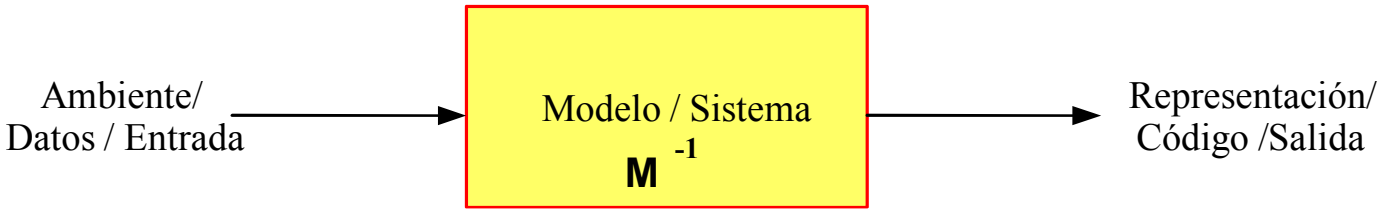
Modelos Auditivos

Shamma, 1985

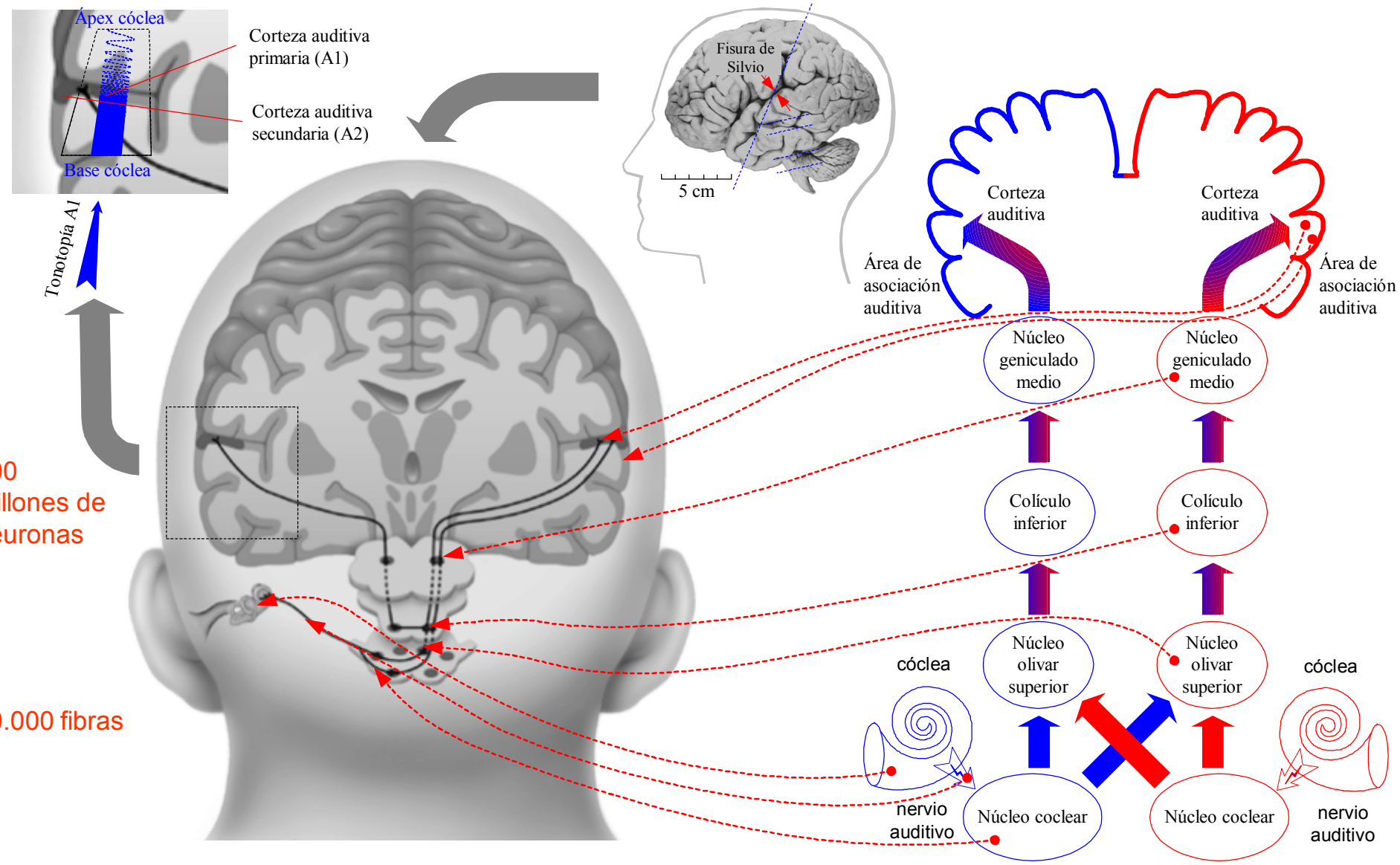
Modelos auditivos



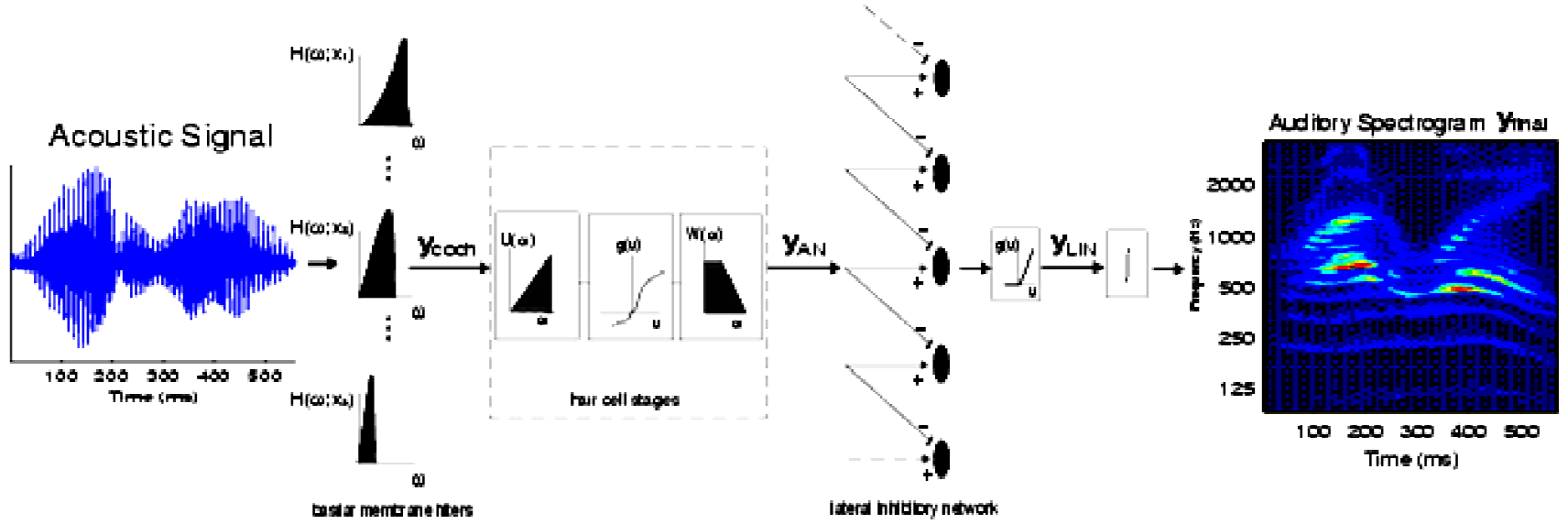
$$\mathbf{x} = \mathbf{M}(\Phi, \mathbf{a}) + \varepsilon$$



Vía auditiva



Etapa temprana
Formulación matemática

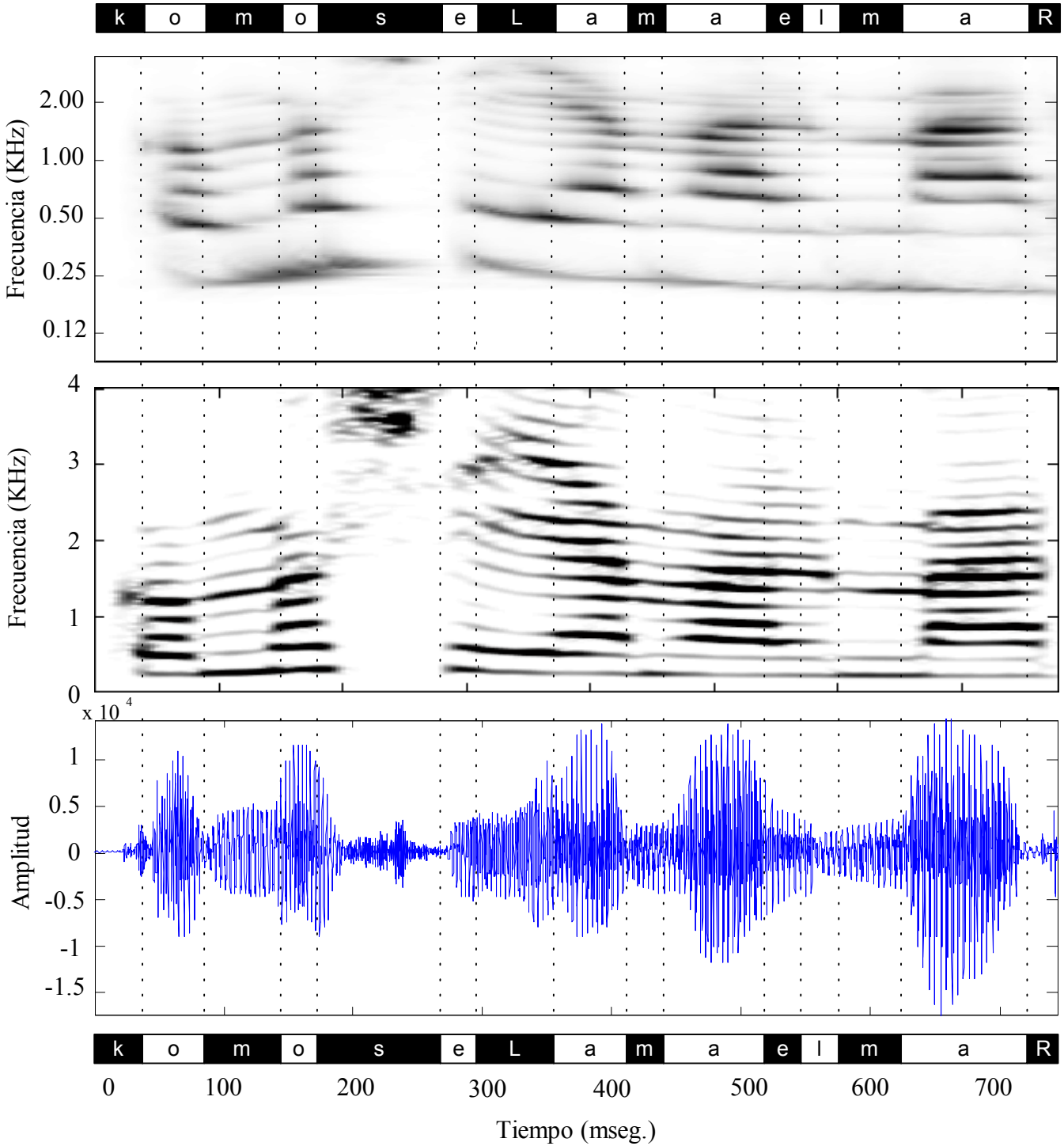


$$y_{coch}(t,x) = s(t) *_t h(t;x)$$

$$y_{AN}(t,x) = g(\partial_t y_{coch}(t,x)) *_t w(t)$$

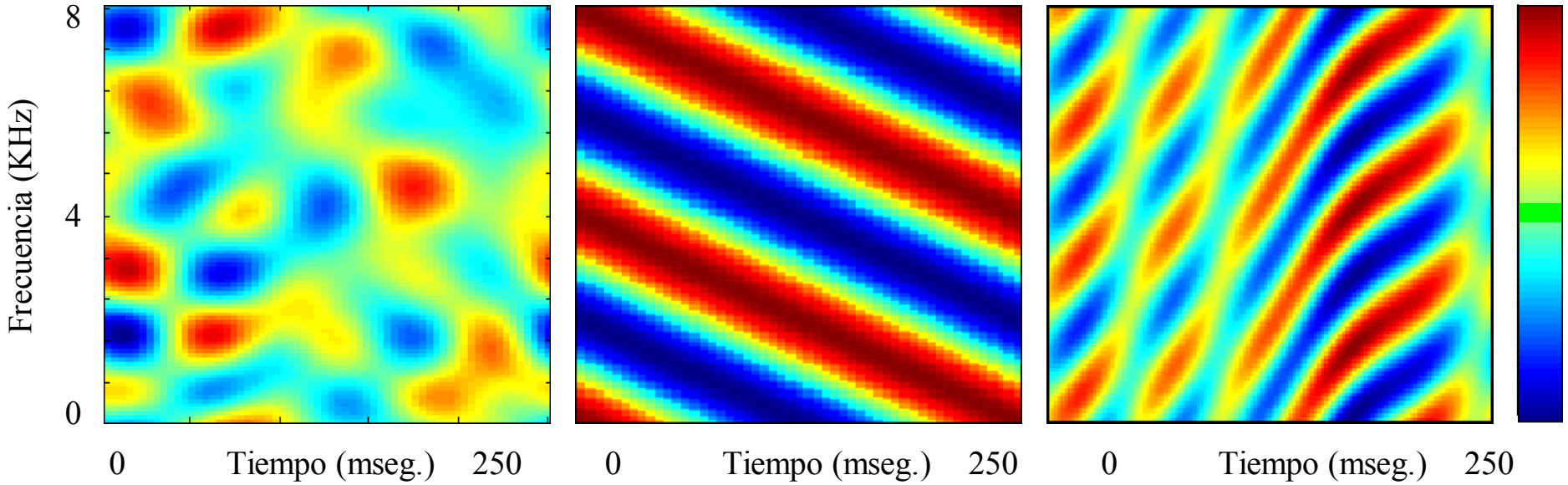
$$y_{LIN}(t,x) = \max(\partial_x y_{AN}(t,x), 0)$$

$$y_{final}(t,x) = y_{LIN}(t,x) *_t \mu(t;\tau)$$



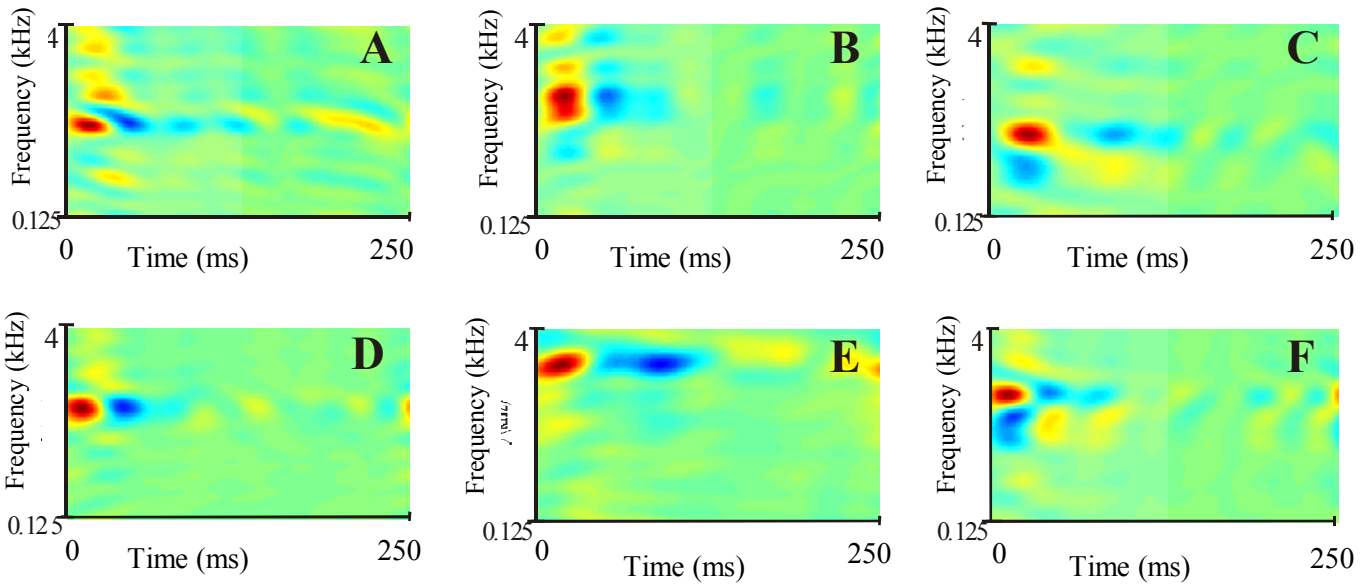
Representación auditiva temprana o espectrograma auditivo (Shamma)

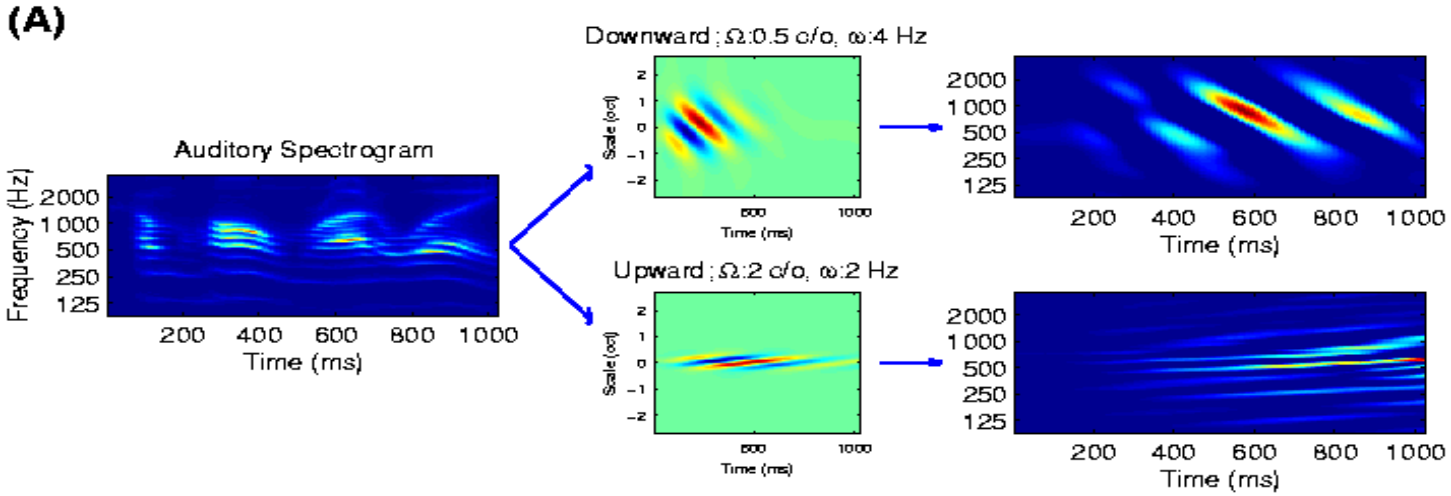
Estímulos complejos



Ruido modulado espectro-temporalmente, ondas móviles y combinaciones.

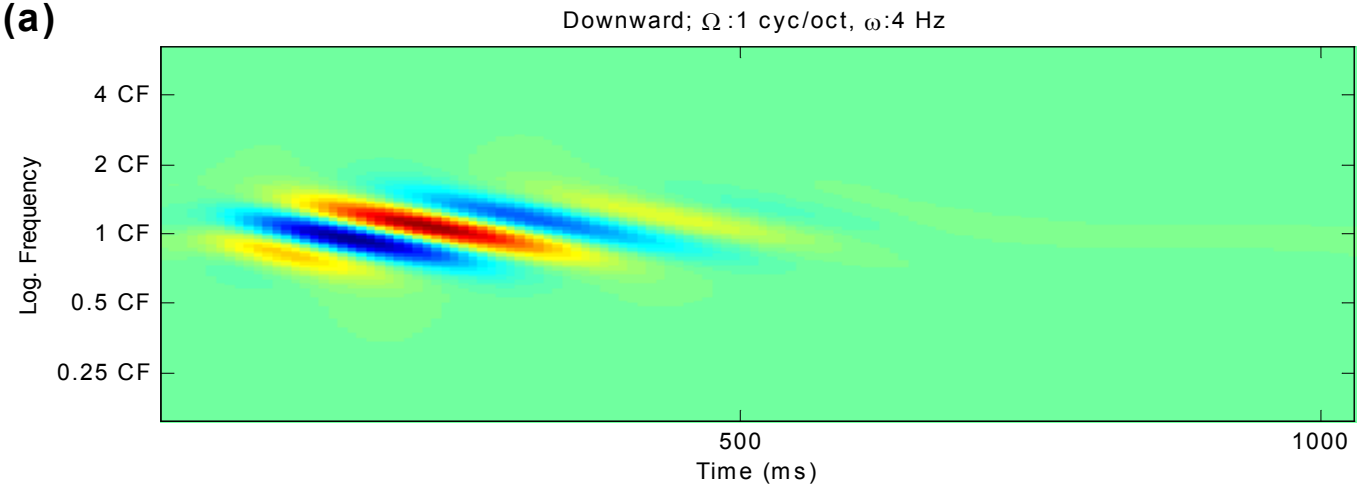
Etapa cortical: STRFs





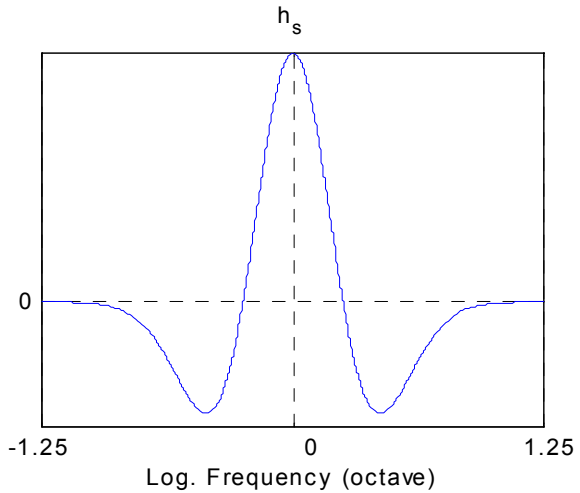
Etapa cortical: implementation

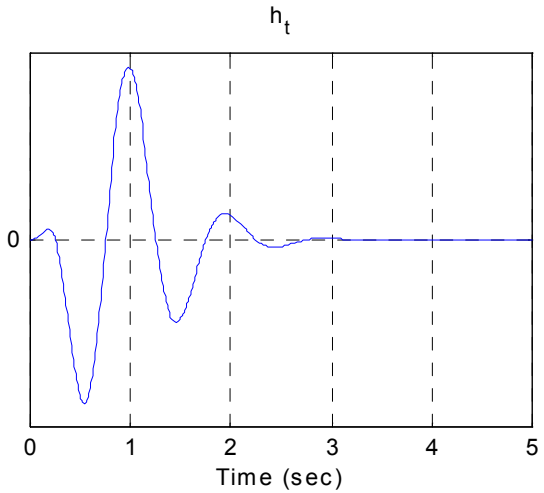
(a)



$$\text{STRF} \equiv h_{IRT}(t) \cdot h_{IRS}(x)$$

(b)





$$h_s(x) = (1 - x^2)e^{-\frac{x^2}{2}}$$

$$h_t(t) = t^3 e^{-4t} \cos(2\pi t)$$

$$h_s(x;\Omega) = \Omega h_s(\Omega x)$$

$$h_t(t;\omega) = \omega h_t(\omega t)$$

Etapa cortical: Formulación matemática

where $$\mathrm{STRF} \equiv h_{IRT}(t) \cdot h_{IRS}(x)$$

$$h_{IRS}(x;\Omega,\phi) = h_s(x;\Omega)\cos\phi + \hat{h}_s(x;\Omega)\sin\phi$$

$$h_{IRT}(t;\omega,\theta) = h_t(t;\omega)\cos\theta + \hat{h}_t(t;\omega)\sin\theta$$

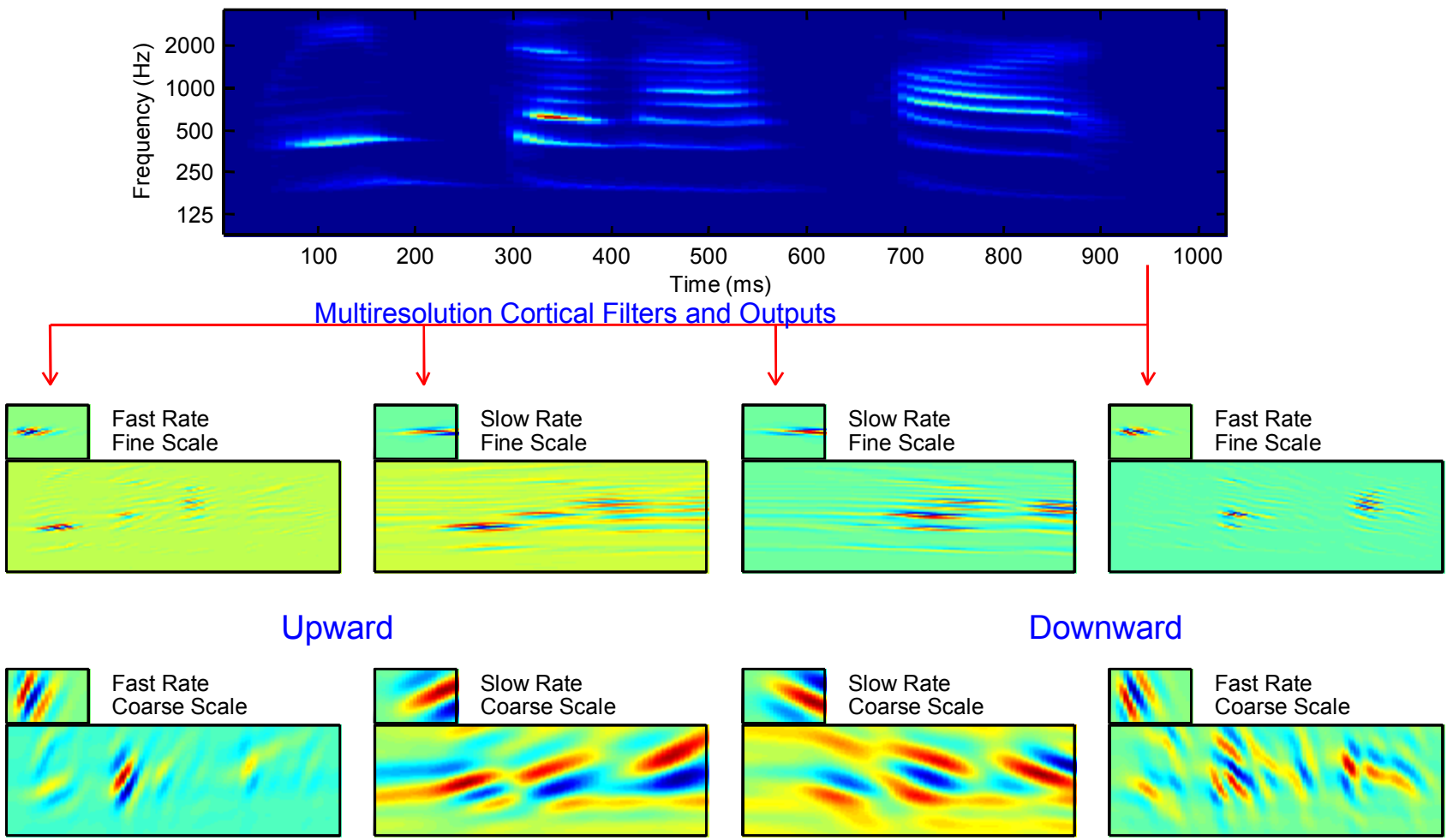
then the spectrotemporal cortical response:

$$r_c(t,x;\omega_c,\Omega_c,\theta_c,\phi_c) = y(t,x) *_{tx} [h_{IRT}(t;\omega_c,\theta_c) \cdot h_{IRS}(x;\Omega_c,\phi_c)]$$

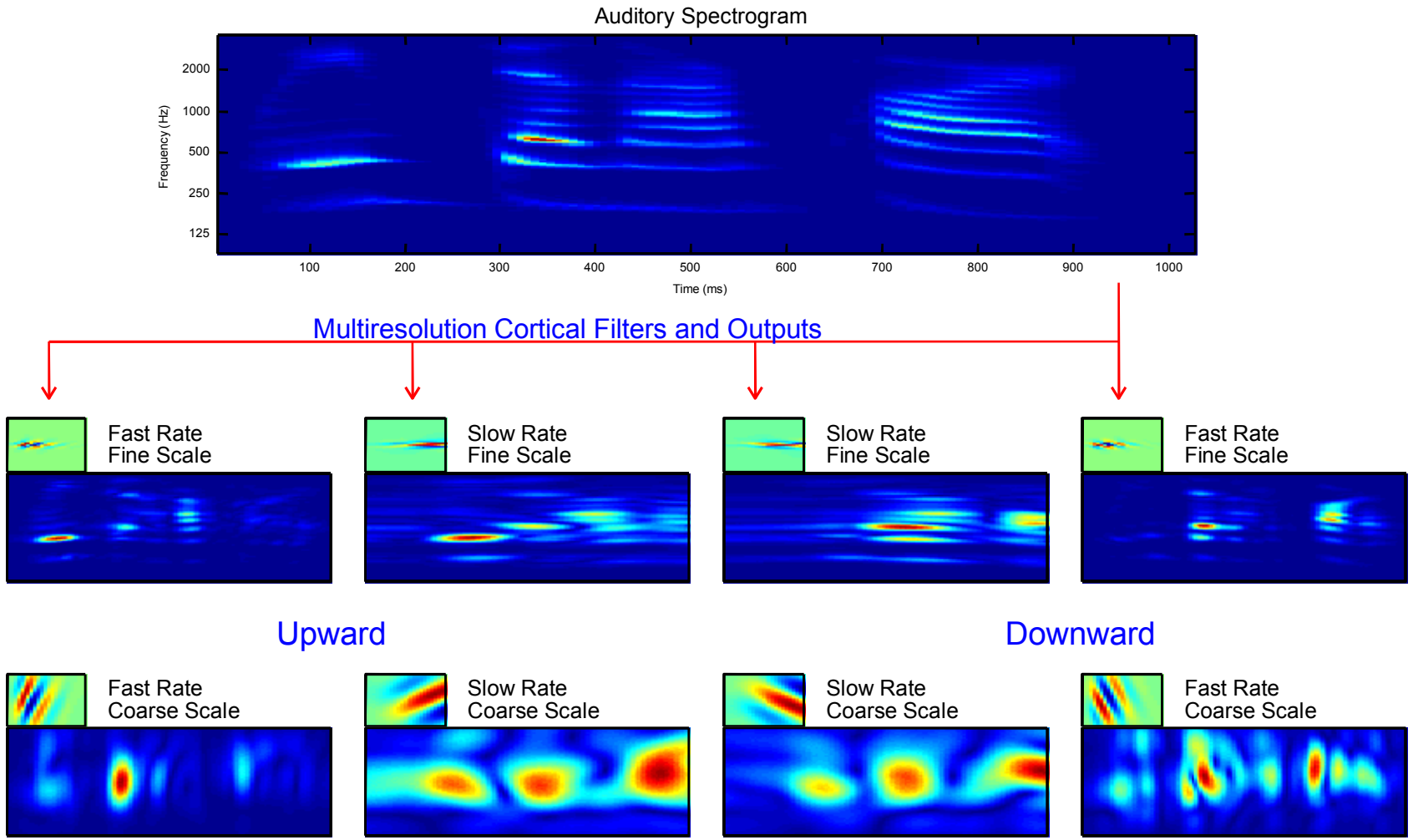
$$= y(t,x) \quad *_{tx} \quad [h_t h_s \cos\theta_c \cos\phi_c + h_t \hat{h}_s \cos\theta_c \sin\phi_c$$

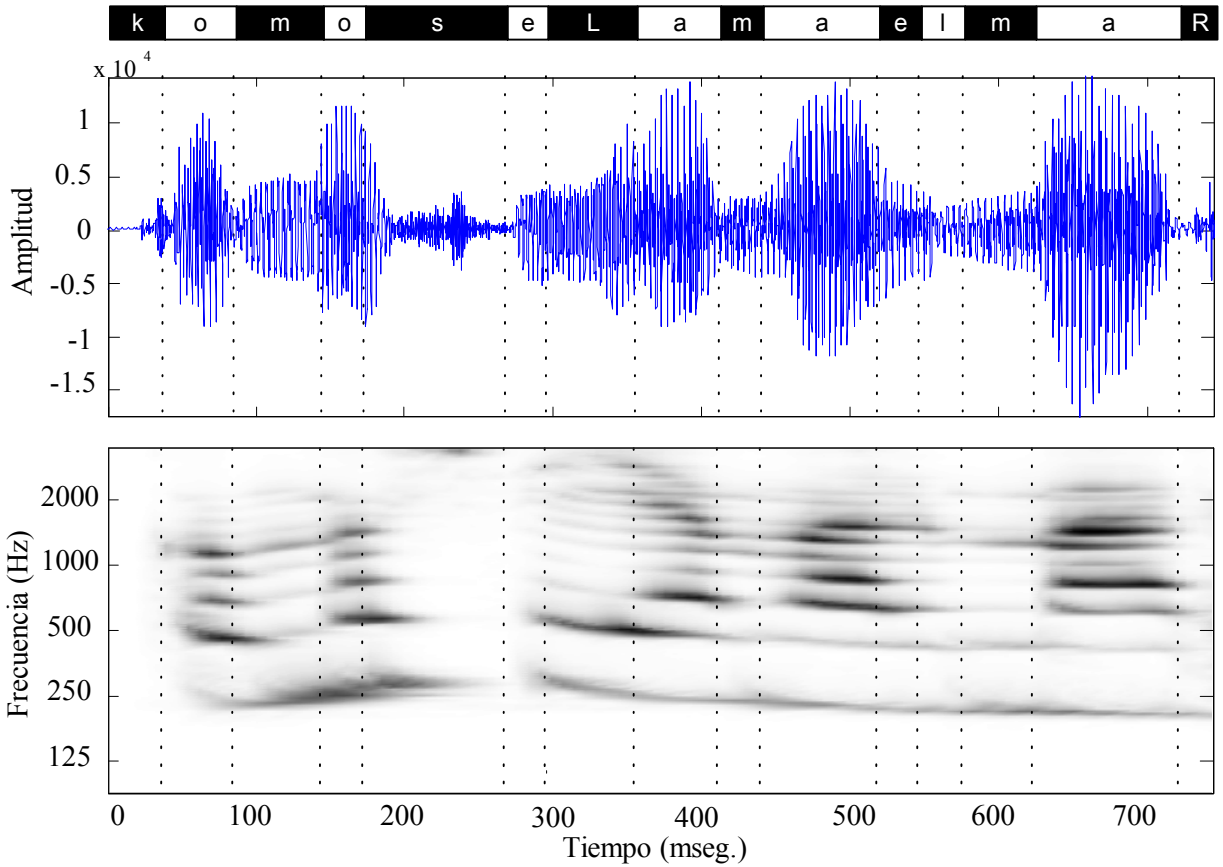
$$+ \hat{h}_t h_s \sin\theta_c \cos\phi_c + \hat{h}_t \hat{h}_s \sin\theta_c \sin\phi_c]$$

Representación cortical



Representación cortical (magnitud)





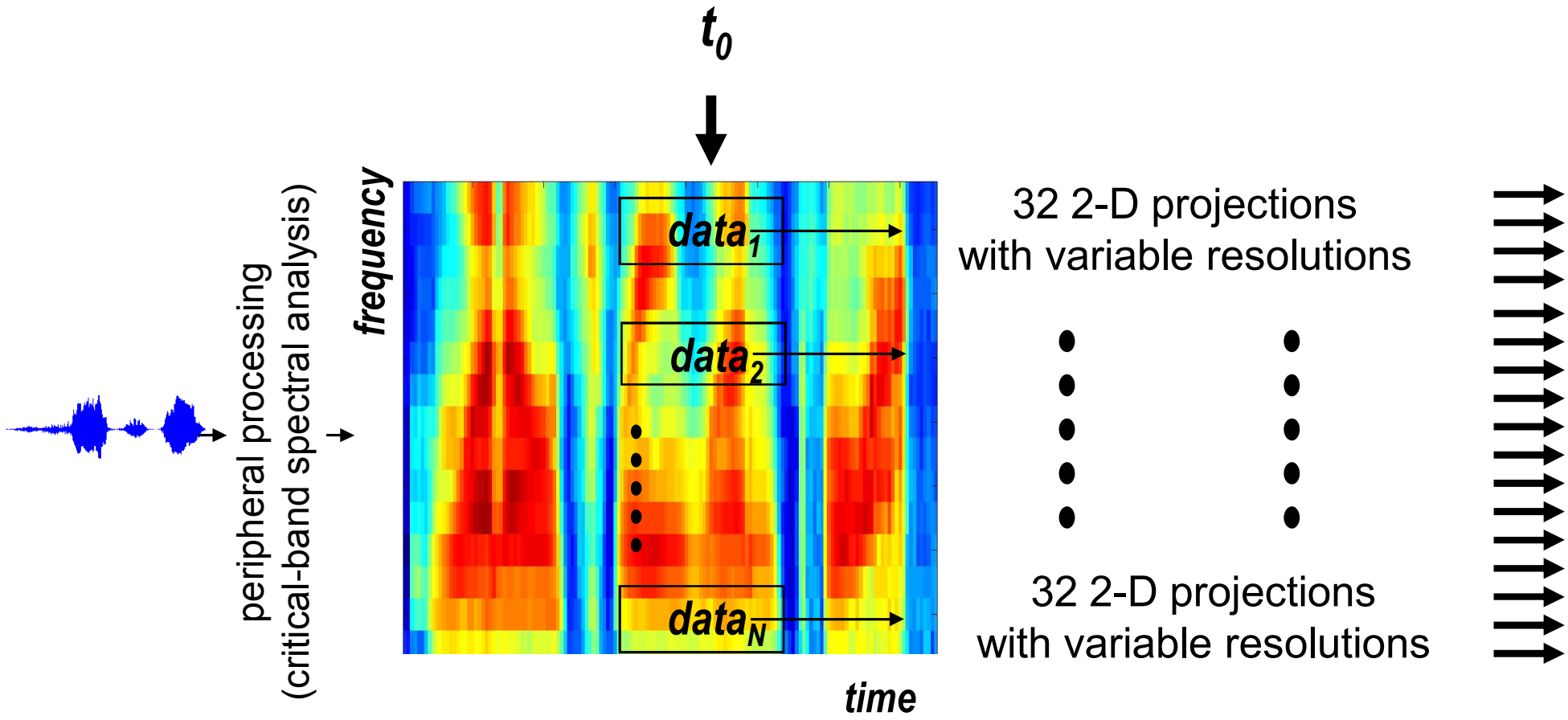
Representación Cortical (Shamma)

RASTA Multiresolución

Hermansky et al., 2005

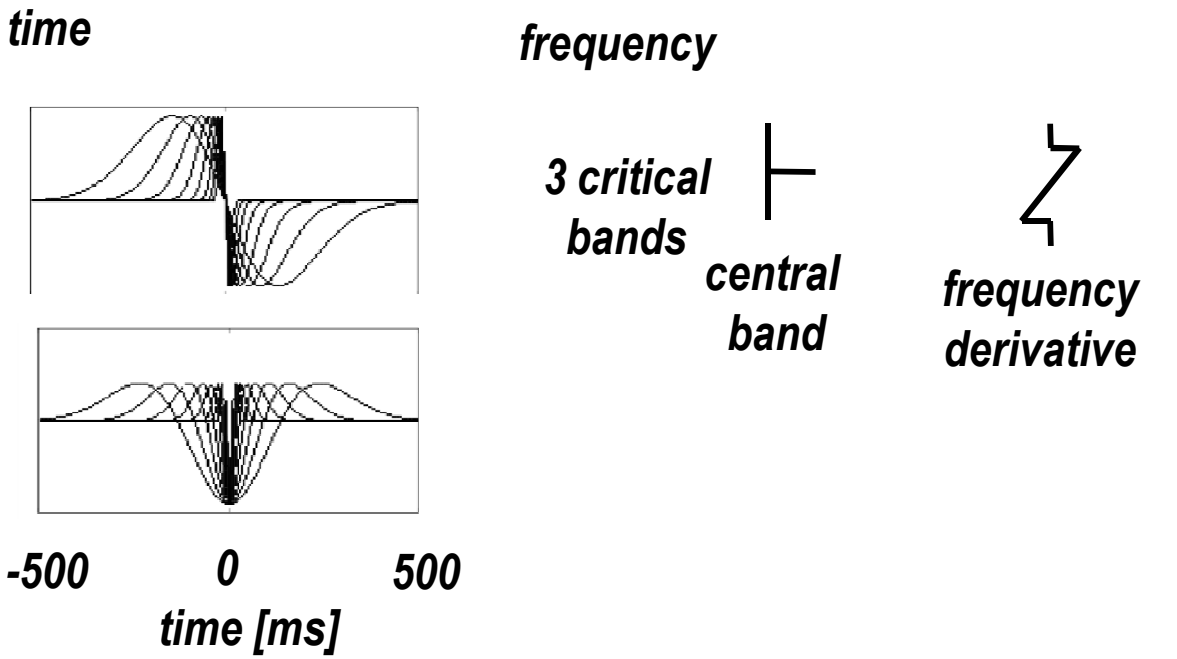
Emulation of cortical processing (MRASTA)

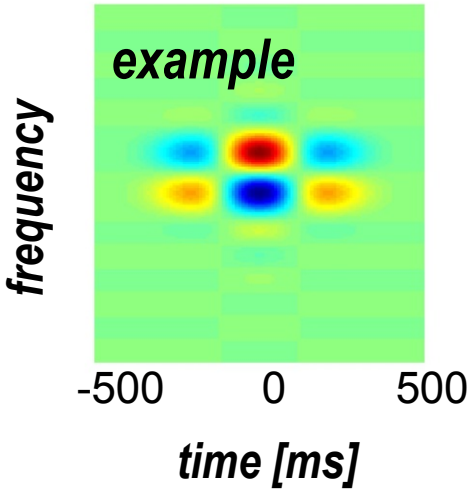
16 x 14 bands = 448 projections

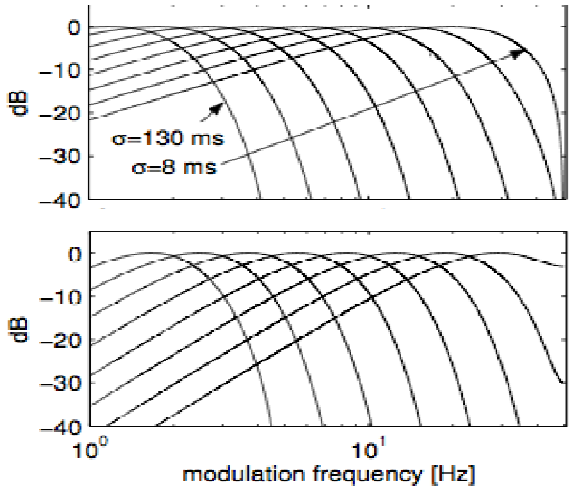


Multi-resolution RASTA (MRASTA) (Interspeech 05)

Spectro-temporal basis formed by outer products of





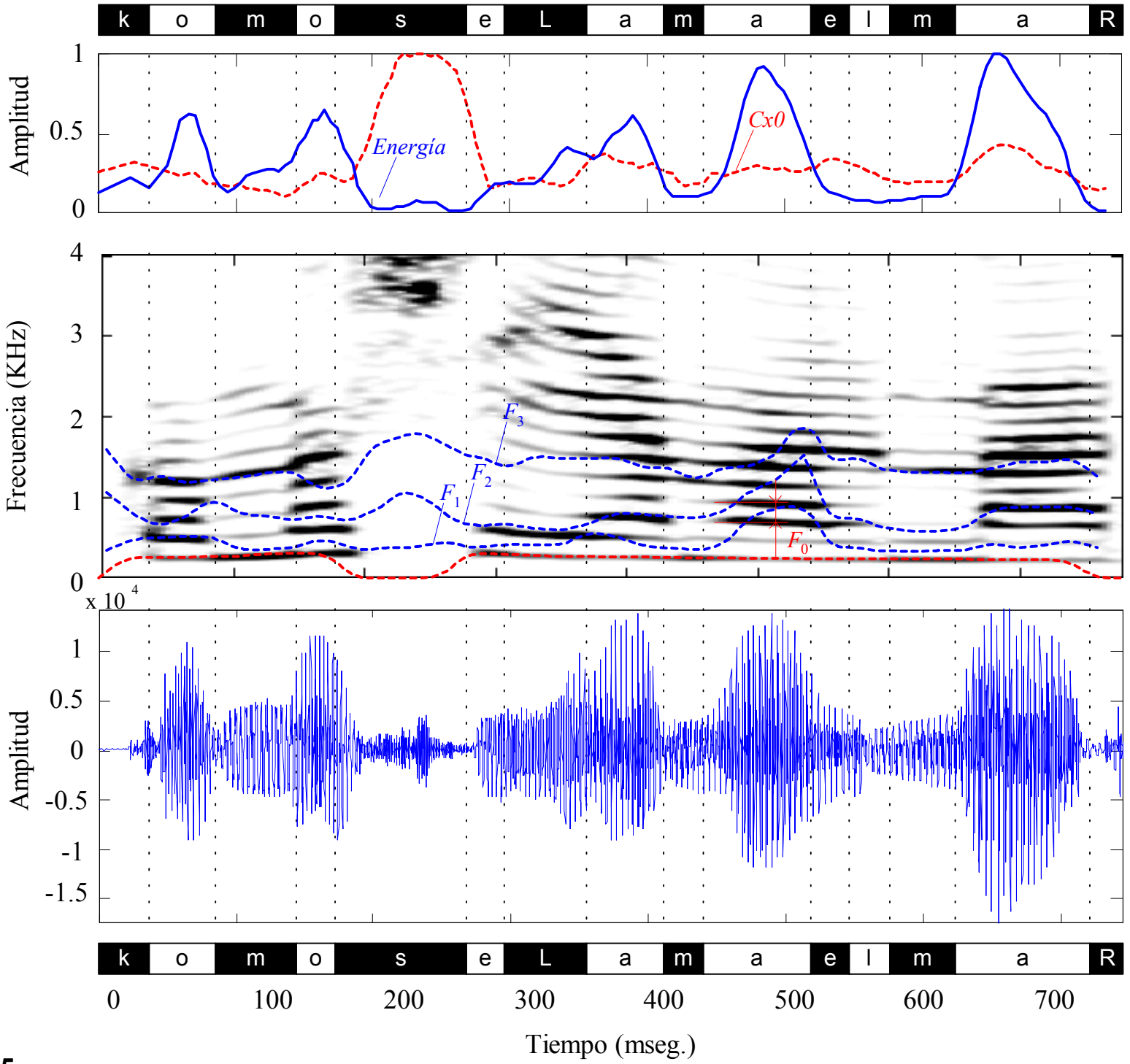


Bank of 2-D (time-frequency) filters
(band-pass in time, high-pass in frequency)

1.RASTA-like: alleviates stationary components
2.multi-resolution in time

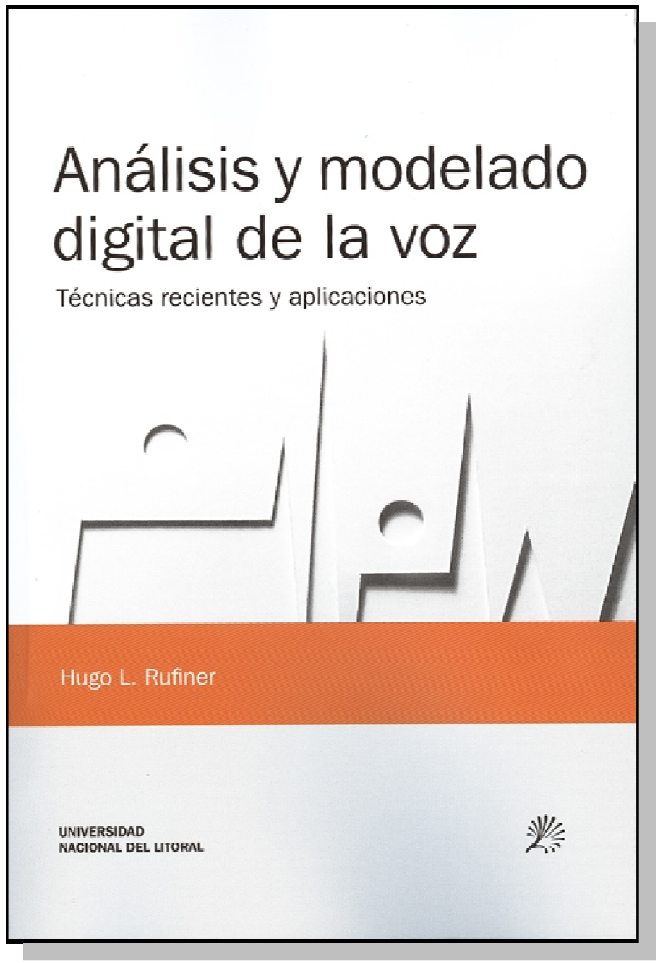
Métodos para extracción de F0 y formantes

Basados en análisis cepstral y correlación


k o m o s e L a m a e l m a R
Amplitud
1
0.5
0
Energía
Cx0
Frecuencia (KHz)
4
3
2
1
0
F_3
F_2
F_1
F_0
x 10^4
Amplitud
1
0.5
0
-0.5
-1
-1.5
k o m o s e L a m a e l m a R
0
100
200
300
400
500
600
700
Tiempo (mseg.)

Bibliografía

- H.L. Rufiner, "Análisis y modelado digital de la voz: Técnicas recientes y aplicaciones", Editorial UNL, 2009. (Capítulo 3).



Bibliografía

- J. Deller, J. Proakis, J. Hansen, "Discrete Time Processing of Speech Signals". Macmillan Publishing, NewYork, 1993.
- J. W. Piccone, "Signal Modeling Techniques in Speech Recognition", Proceedings of the IEEE, Vol. 81, Nº9, pp. 1215-1247, 1993.
- H. Fletcher, "Speech and Hearing in Communication", Van Nostrand, New York, NY, 1953.
- H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", J. Acoust. Soc. Am., vol. 87, no. 4, pp. 1738-1752, Apr. 1990.
- H. Hermansky and N. Morgan, "RASTA processing of speech", IEEE Trans. on Speech and Audio Proc., vol. 2, no. 4, pp. 578-589, Oct. 1994.
- Shamma S., "Speech Processing in the Auditory System. Part I: The Representation of Speech Sounds in the Responses of the Auditory-Nerve,", J. Acoust. Soc. Am. 78(5), 1612-1621, 1985.
- Shamma S., "Speech Processing in the Auditory System. Part II: Lateral Inhibition and the Central Processing of Speech Evoked Activity in the Auditory-Nerve," J. Acoust. Soc. Am. 78(5), 1622-1632, 1985.
- Yang X., Wang K., and Shamma S., "Auditory Representations of Acoustic Signals", IEEE Trans. Info. Theory, 38, 824-839, 1992.

Neural Systems Laboratory

Institute for System Research Department of Electrical Engineering University of Maryland

Home
Research
People
Publications
Downloads
Opportunities
Links
Contact

Faculty

Shihab Shamma, Director

Ph.D., Stanford University, 1980

Research Interests: Dr. Shamma's research deals with the question of how the acoustic signal is represented at various levels of the mammalian auditory system. The research spans a wide range of disciplines and techniques, ranging from theoretical models of auditory processing the early and central auditory stages, to neurophysiological investigations of the auditory cortex, to psychoacoustical experiments of human perception of acoustic spectral profiles. These studies complement each other in that theoretical models are directly based on experimental data, and in turn the models motivates the experimental paradigms and analysis.

Jonathan Fritz, Research Scientist

Ph.D., Brown University, 1995

Research Interest: The two broad topics of Dr. Fritz's current research are:

(1) task-related adaptive plasticity in auditory processing with a current focus of interest on behavioral physiology studies of ferret (and monkey) primary and secondary auditory cortices, and top-down influence of prefrontal cortex.

(2) neurobiology of auditory perception and memory, including psychophysical studies in the ferret at NSL, perceptual and behavioral lesion studies in the monkey at NIH with Mort Mishkin, PET and fMRI imaging studies of auditory processing in collaboration with Al Braun at NIH.

Neural Systems Laboratory

Institute for System Research Department of Electrical Engineering University of Maryland

Home
Research
People
Publications
Downloads
Opportunities
Links
Contact

NSL Matlab Toolbox

NSL Matlab Toolbox
Unzip the file in your Matlab toolbox directory and add it to the Matlab Path

NSL Toolbox Graphical User Interface
Unzip the file in NSL Matlab toolbox directory. Then run nslgui.m

A power point presentation about the toolbox
This PowerPoint presentation is prepared by Taishih Chi

Complete documentation of the toolbox
This documentation is a small portion of Powen Ru's PhD thesis

HRTF Data

HRTS Database

Readme file

Associated Publication: ICAD 2003

Welcome!

The Center for Auditory and Acoustic Research (CAAR) is a consortium of researchers from six universities working in partnership with Department of Defense laboratories and industry. CAAR is funded by the Office of Naval Research through a 1997 Department of Defense Multidisciplinary University Research Initiative. Learn more about us here.

Research **People** **Publications**

Consortium partners **Mission statement** **Conferences and talks**

Research People Publications
Consortium Mission Conferences
partners statement and talks

How to contact us

E-mail
shantanu@isr.umd.edu

Phone (301) 405-6596
Fax (301) 314-9920

Mailing address
Center for Auditory and Acoustic Research
Institute for Systems Research
University of Maryland
College Park, MD 20742

PLP and RASTA (and MFCC, and inversion) in Matlab using melfcc.m and invmelfcc.m

---

Introduction

One of the first decisions in any pattern recognition system is the choice of what features to use: How exactly to represent the basic signal that is to be classified, in order to make the classification algorithm's job easiest.

Speech recognition is a typical example. Through more than 30 years of recognizer research, many different feature representations of the speech signal have been suggested and tried. The most popular feature representation currently used is the Mel-frequency Cepstral Coefficients or MFCC.

Another popular speech feature representation is known as RASTA-PLP, an acronym for Relative Spectral Transform - Perceptual Linear Prediction. PLP was originally proposed by Hynek Hermansky as a way of warping spectra to minimize the differences between speakers while preserving the important speech information [Herm90]. RASTA is a separate technique that applies a band-pass filter to the energy in each frequency subband in order to smooth over short-term noise variations and to remove any constant offset resulting from static spectral coloration in the speech channel e.g. from a telephone line [HermM94].

RASTA-PLP is implemented in a number of programs, such as the 'rasta' program, and its enhanced version 'feacalc', which are distributed for Unix as part of the SPRACHcore package. In order to understand the algorithm, however, it's useful to have a simple implementation in Matlab. By using Matlab's primitives for FFT calculation, Levinson-Durbin recursion etc., the Matlab code can be made quite small and transparent.

Mike Shire started this implementation in 1997 while he was a graduate student in Morgan's group at ICSI. I have recently revised and extended his implementation to allow both spectral and cepstral outputs, and to allow independent selection of RASTA and/or PLP processing.

This implementation offers only a few control parameters, namely a switch to select or disable rasta filtering, and an option to set the order of PLP modeling (which disables PLP modeling when set to zero). Other important options, such as the basic window and hop sizes, can easily be altered by editing the relevant routines, if desired.

MFCCs

Since Mel-frequency Cepstral Coefficients, the other really popular speech feature, involve almost the same processing steps, I decided to make an implementation for them as well, using the same blocks as far as possible. See below.

Inverting Cepstra to Audio

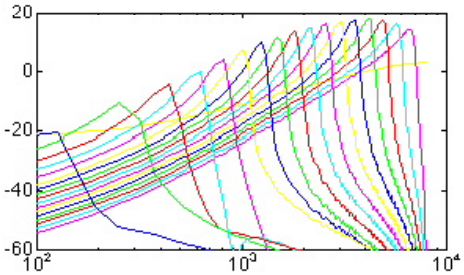
Sometimes it's interesting to `listen' to what it is that the cepstral representations are really capturing. You can do this, crudely, by recovering the short-time magnitude spectrum implied by the cepstral coefficients, then imposing it on white noise. The routine `invmelfcc` below does this

Auditory Toolbox

Version 2

Malcolm Slaney

Technical Report #1998-010
Interval Research Corporation
malcolm@interval.com



What is the Auditory Toolbox?

This report describes a collection of tools that implement several popular auditory models for a numerical programming environment called MATLAB. This toolbox will be useful to researchers that are interested in how the auditory periphery works and want to compare and test their theories. This toolbox will also be useful to speech and auditory engineers who want to see how the human auditory system represents sounds.

This version of the toolbox fixes several bugs, especially in the Gammatone and MFCC implementations, and adds several new functions. This

FIN