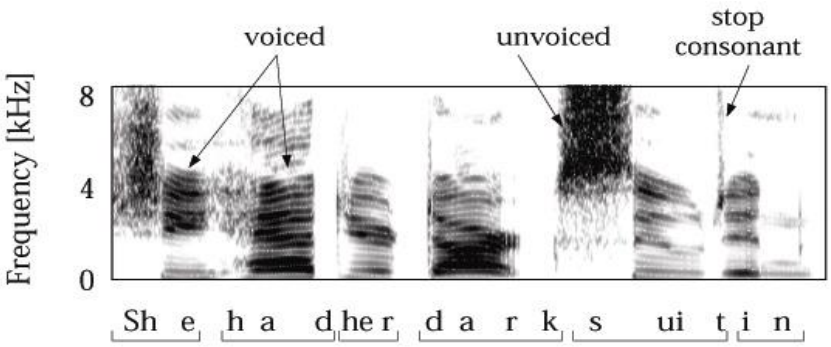
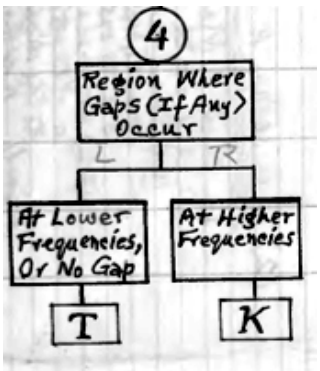
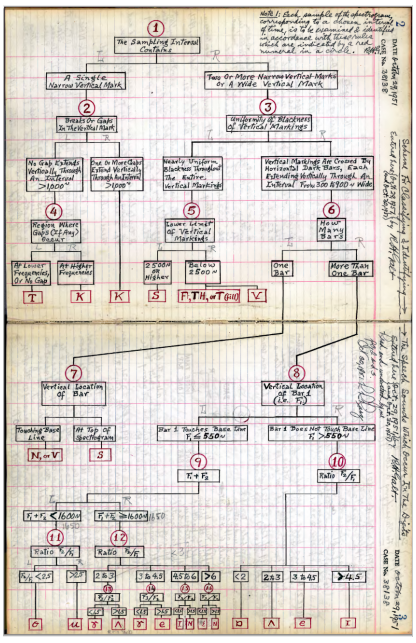
speech spectrogram



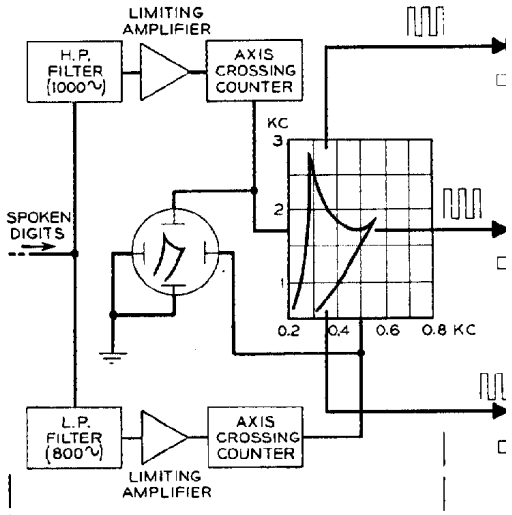
Concept of the first "real" automatic speech recognizer (R.H. Galt 1951)





First “real” recognizer ever build

(Davis, Biddulph, Balashek 1952) Automatic Speech Recognition of Spoken Digits, J. Acoust. Soc. Am. 24(6) pp.637 - 642



Note 1 (4/23/52): I called attention to the fact that in my tests, one voice was that of Balachek. The number of errors was either 2 or 1 the success score thus being either 97.5 % or 98.7 %. [The uncertainty was because of the use of the 11th box in one test.] Thus my rules were almost or equally as good, with Balachek's voice, as the rules governing E. Peterson's machine. However, my rules were almost as effective for a considerable range of voices; whereas for several men's voices, the machine gave scores perhaps as low as 65 %.

100

DATE 38138

CASE No. 4/29/52

Conference 4/29/52 with Karlin

Karlin reported conclusions reached in his recent conference with Potter.

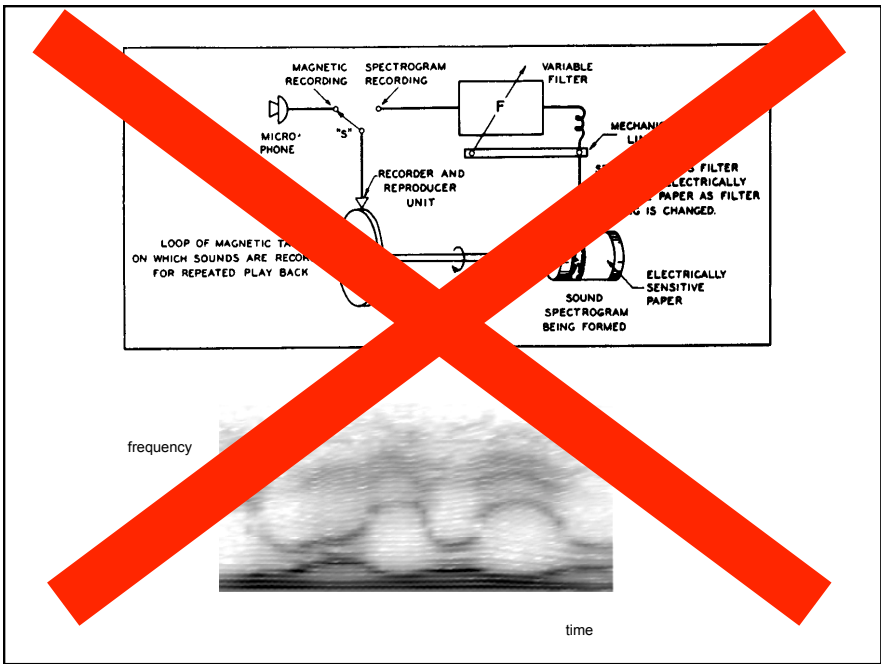
1. The V.O.T. job will not concern me any longer. This job as an exercise in simulation, will be one of the jobs in Kerstas hands.

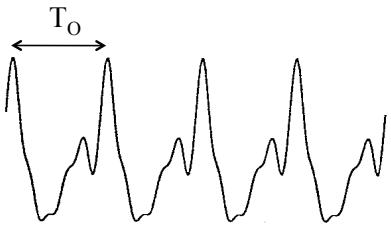
44

45

Computers

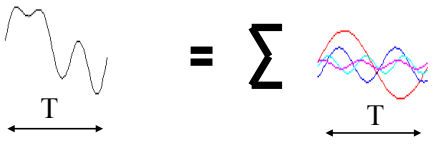
46



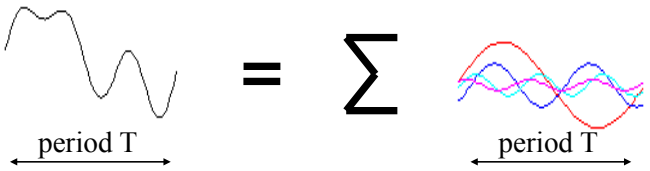


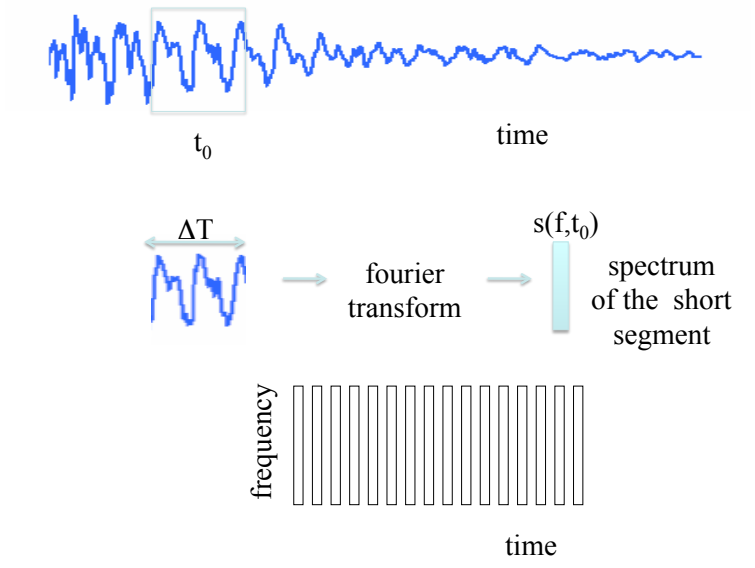
- One of Fourier ideas
  - Describe a periodic signal by an (infinite) sum of other well defined periodic signals

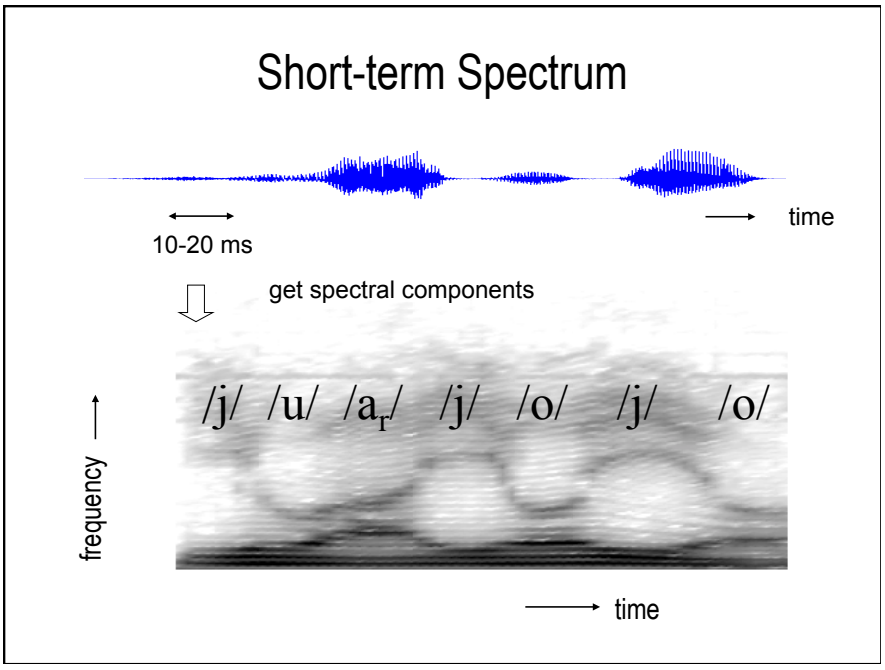
Joseph Fourier
(1768-1830)
Student of Lagrange
Adviser of e.g. Dirichlet or Navier

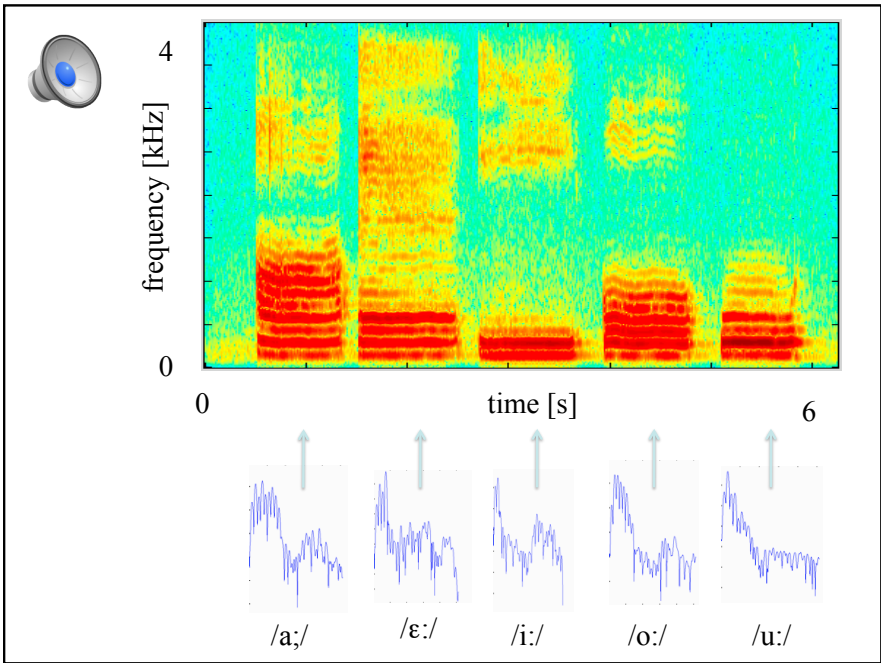


Fourier's Idea



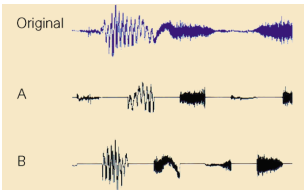


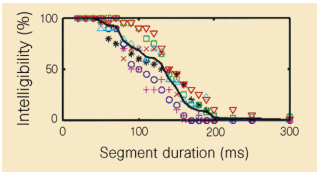

Short-term Spectrum
time
10-20 ms
get spectral components
/j/ /u/ /ar/ /j/ /o/ /j/ /o/
frequency
time


4
frequency [kHz]
0
0
time [s]
6
/a;/
/ɛ:/
/i:/
/o:/
/u:/

Ignore short-term phase: Locally reversed speech

200 ms
150 ms
100 ms
80 ms
60 ms
40 ms
20 ms
original





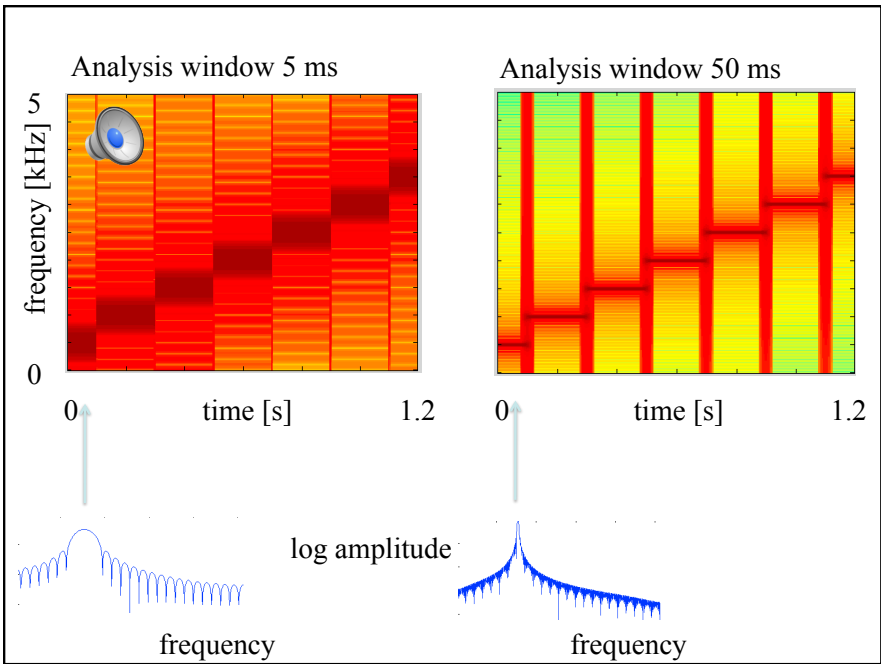
Nature **398**, 760 (29 April 1999)
Cognitive restoration of reversed speech
Kourosh Saberi and David R. Perrott

Uncertainty Principle

You cannot simultaneously know the exact frequency composition and the exact temporal locality of an acoustic event

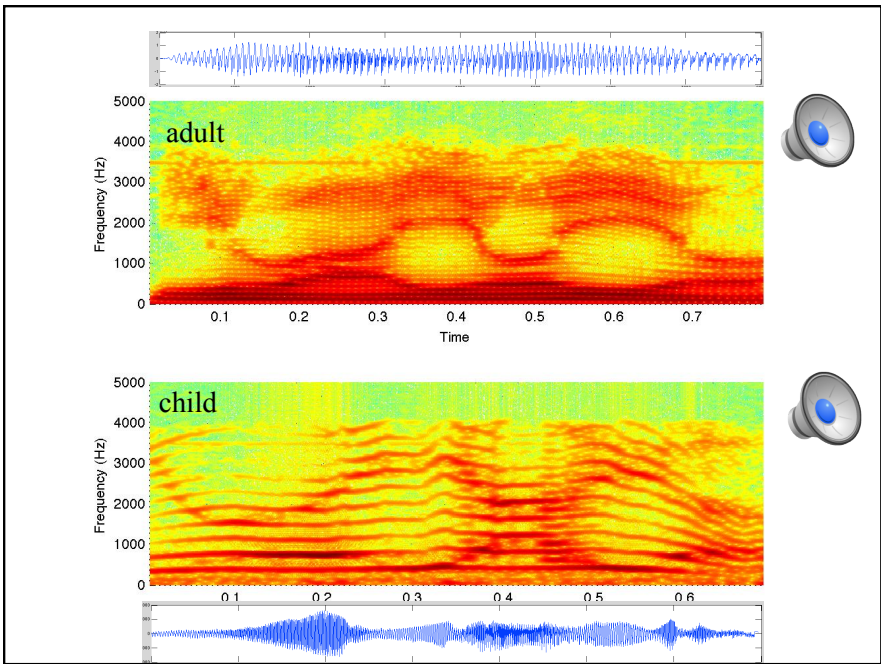
Dennis Gabor 1947

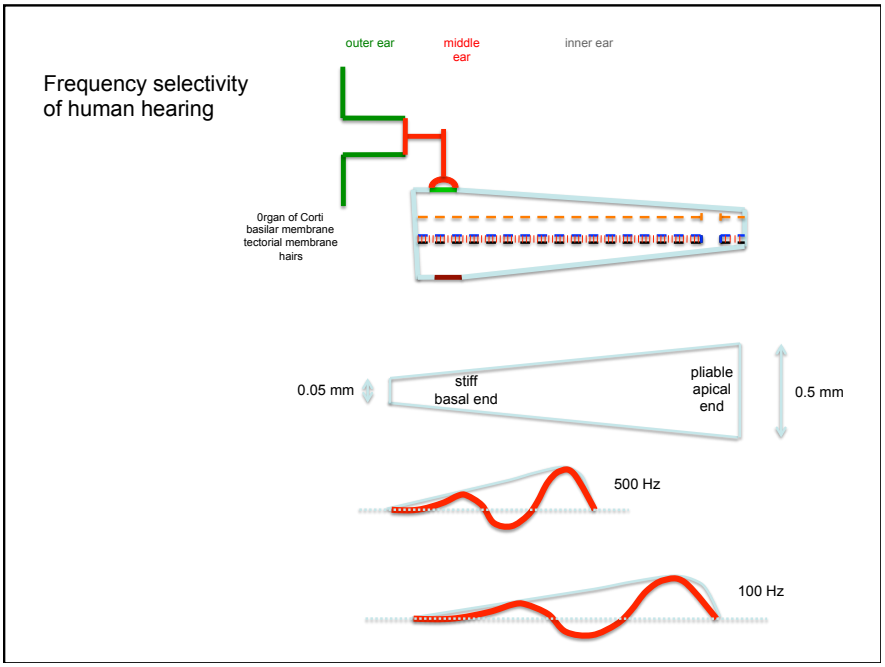
Higher spectral resolution needs longer analysis windows and hence yields lower temporal resolution



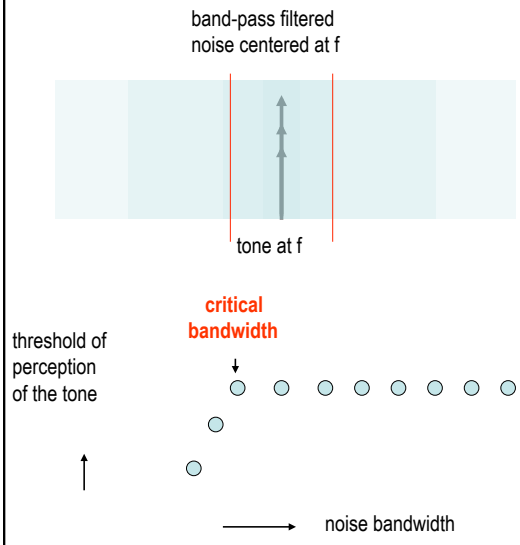
Different Speakers

56


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


Frequency selectivity
of human hearing
outer ear
middle
ear
inner ear
0rgan of Corti
basilar membrane
tectorial membrane
hairs
0.05 mm
stiff
basal end
pliable
apical
end
0.5 mm
500 Hz
100 Hz

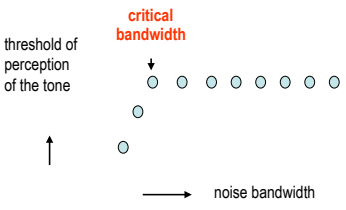
Simultaneous Masking (Fletcher 1940)

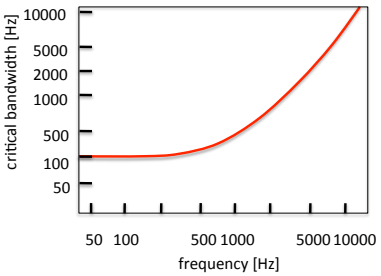


what happens outside the critical band does not affect decoding of the sound in the critical band

Spectral resolution of hearing

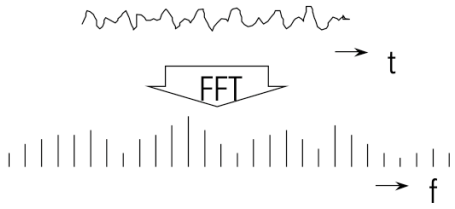
spectral resolution of hearing decreases with frequency
(critical bands of hearing, perception of pitch,…)



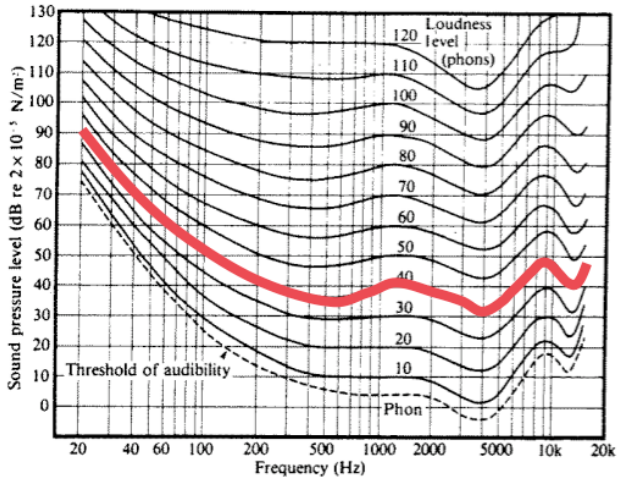


what happens outside the critical band does not affect decoding of the sound in the critical band

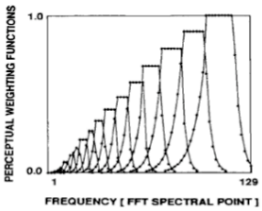
"Poor man's"
way of emulating
critical-band
filter



Sensitivity of hearing depends on frequency



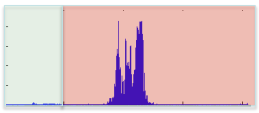
spectrum



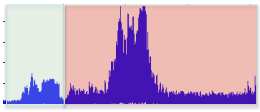
summation windows

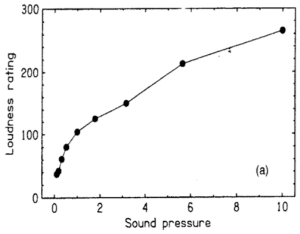
spectrum with auditory-like resolution

intensity ≈ signal ^{2} [w/m^2]



loudness [Sones]





loudness = intensity $^{0.33}$

intensity
(power spectrum)

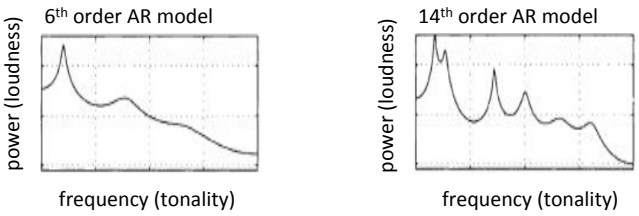
$|.|^{0.33}$

loudness

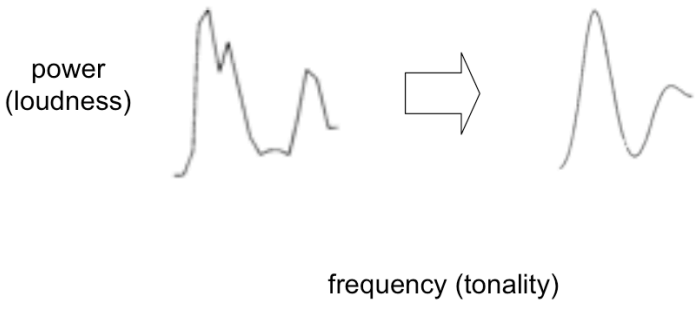
Not all spectral details are important

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

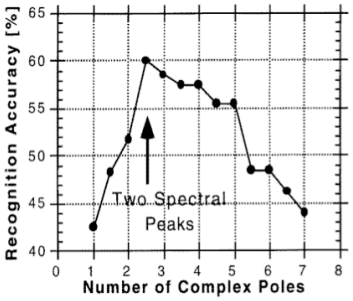
b) approximate the auditory spectrum by an autoregressive model



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



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

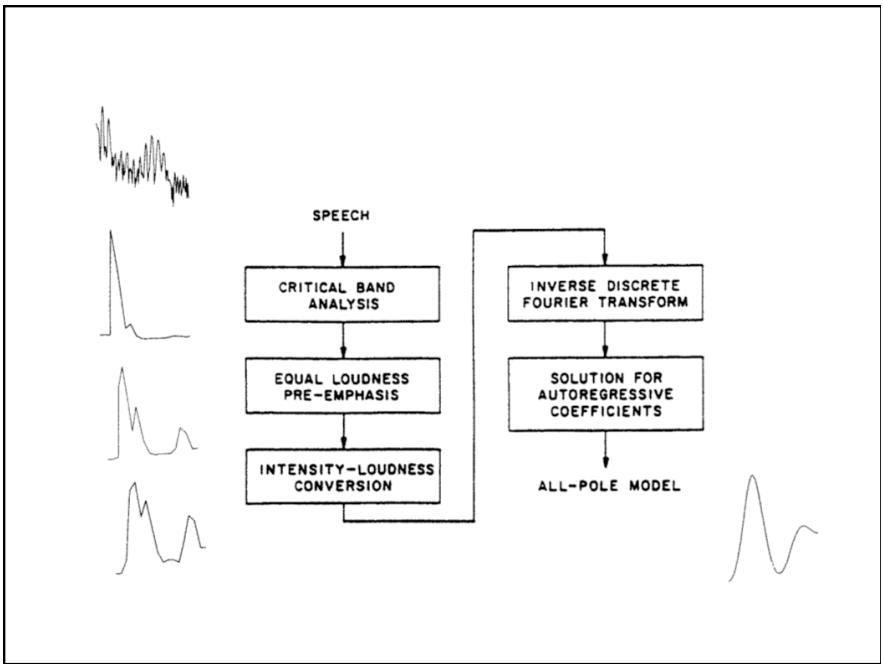
Perceptual Linear Prediction

Limited spectral resolution

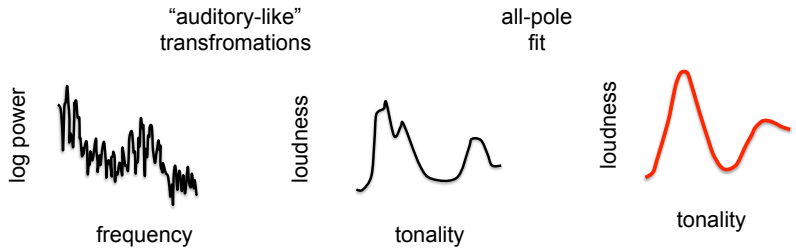
formant clusters as may be interpreted by auditory perception

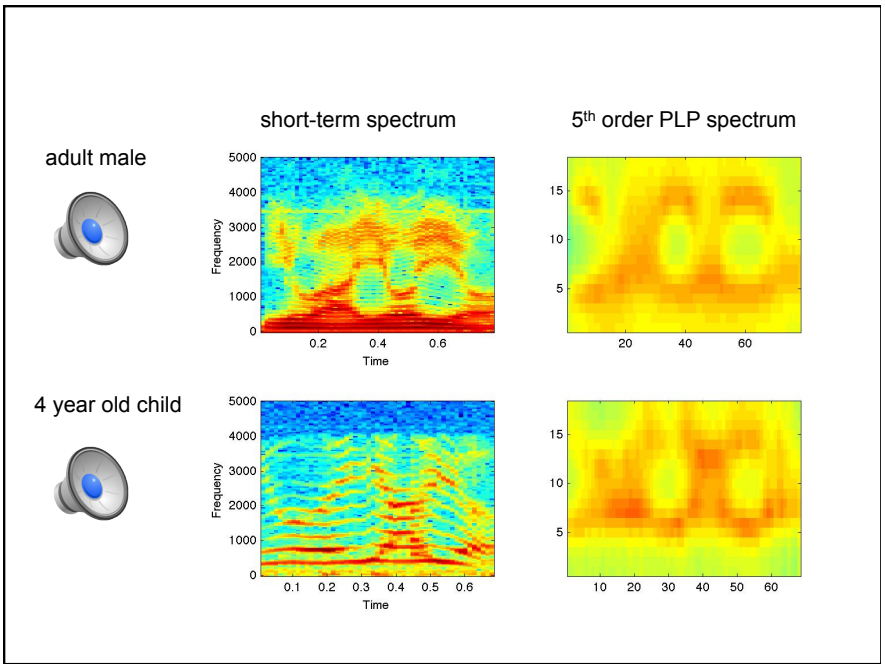
Perceptual Linear Prediction (PLP)

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



Perceptual Linear Prediction





H6. Effect of the spectral model order in automatic speech recognition.
Kazuhiro Tsuga and Hynek Hermansky (Speech Technology Laboratory, 3888 State Street, Santa Barbara, CA 93105)

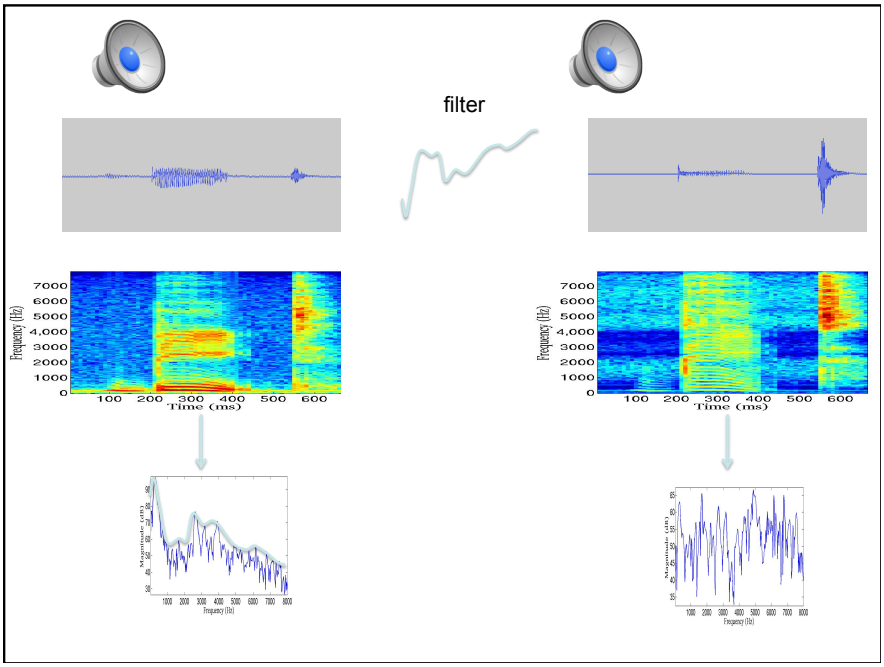
It has been observed that in speaker-independent multi-template digit recognition, the 5th-order perceptually based LP (PLP) analysis method yields about 40% lower error rates than does the standard 14th-order LP.

J. Acoust. Soc. Am. Volume 80, Issue S1, pp. S18-S18 (1986)

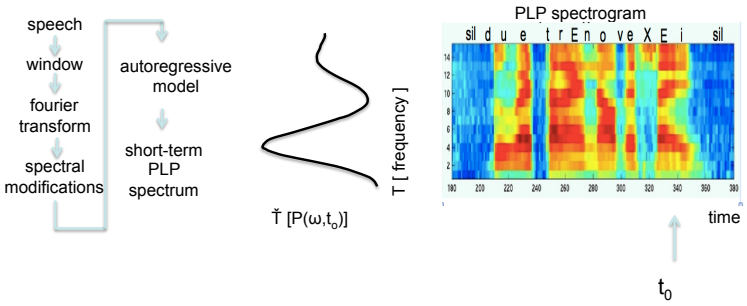
RE: TWO INVENTIONS BY DR. H. HERMANSKY ET AL
WE HAVE CAREFULLY STUDIED WHETHER TO FILE U.S. PATENT APPLICATIONS FOR THESE INVENTIONS. AS MR. NYUJI TELEPHONED YOU, OUR CONCLUSION IS:
(1) ''PERCEPTUALLY BASED LINEAR PREDICTIVE ANALYSIS OF SPEECH''. WE CANNOT SEE ANY PRESENT OR FUTURE PRODUCTS TO WHICH THIS INVENTION IS PRESUMED TO BE APPLIED. SO, THIS INVENTION DOES NOT HAVE ENOUGH PRACTICAL VALUE TO BE APPLIED FOR A U.S. PATENT.

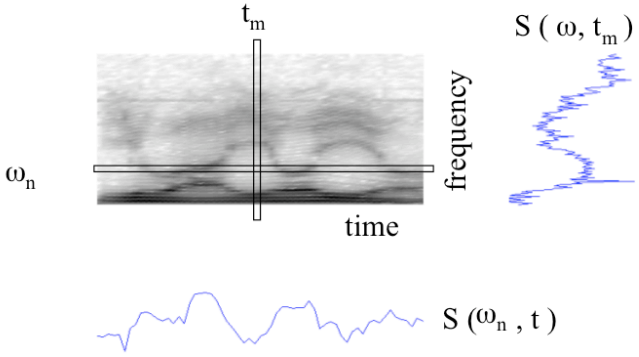
Different Communications Channels

73



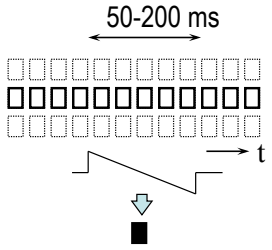
Short-Term Features

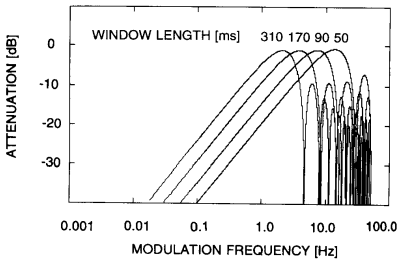




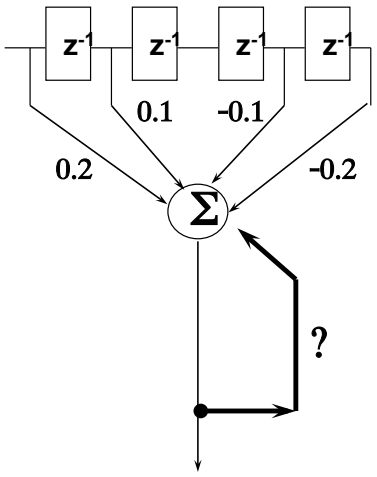
Linear distortions: DC bias on $S(\omega_n,t)$
different for each career frequency ω_n

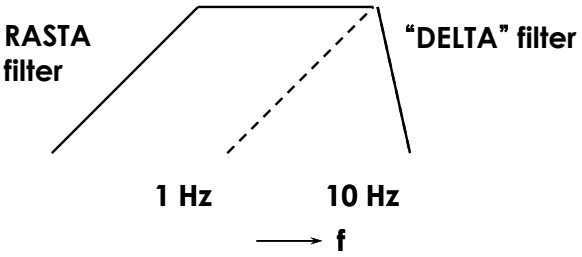
Delta Features





- linear combination of several short-term features
- equivalent to FIR filtering of feature trajectory
- selective band-pass with 6 dB/oct slope





Optimizing RASTA Filter