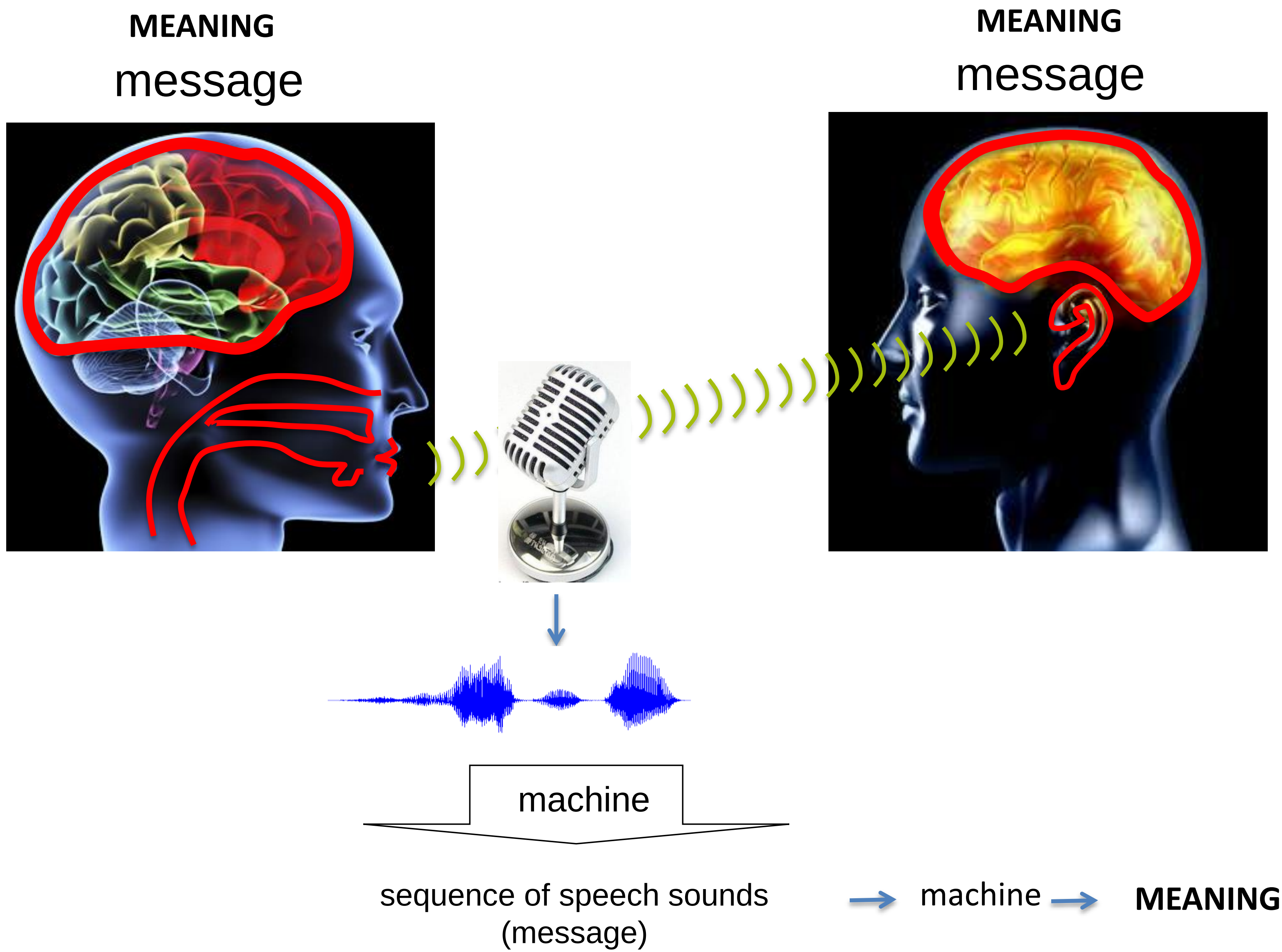
The Center For Language and Speech Processing
at the Johns Hopkins University

speech@fit BUT

human language technology
center of excellence

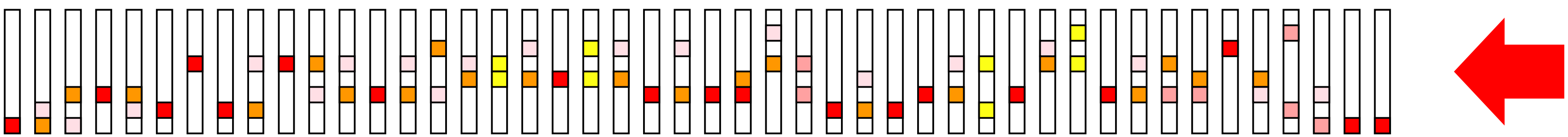
Deep, Long, and Wide Artificial Neural Networks in ASR

Hynek Hermansky


MEANING
message
MEANING
message
machine
sequence of speech sounds
(message)
machine
MEANING

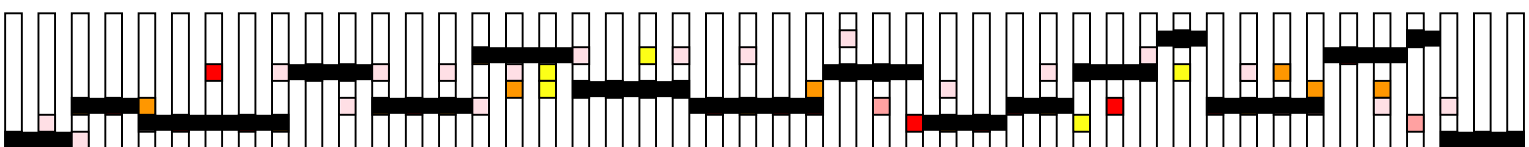
speech signal

estimate likelihoods $p(x|W_i)$, *where* W_i *are constituents of* W *(speech sounds)*

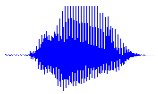


stochastic search

$\hat{W} = argmax_W\ p(x|W)\ P(W)$ ← language model and lexicon

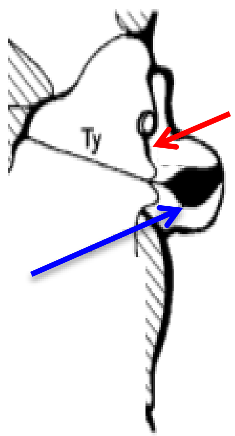


Multi-Layer Perceptron can emulate any nonlinear mapping

 → MLP → likelihood of speech sound

cicada

~ 10^6 neurons

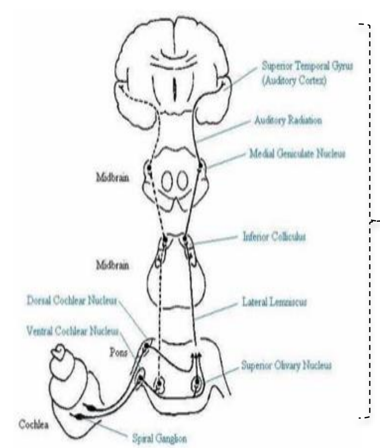


sensory organ

central
nervous
system

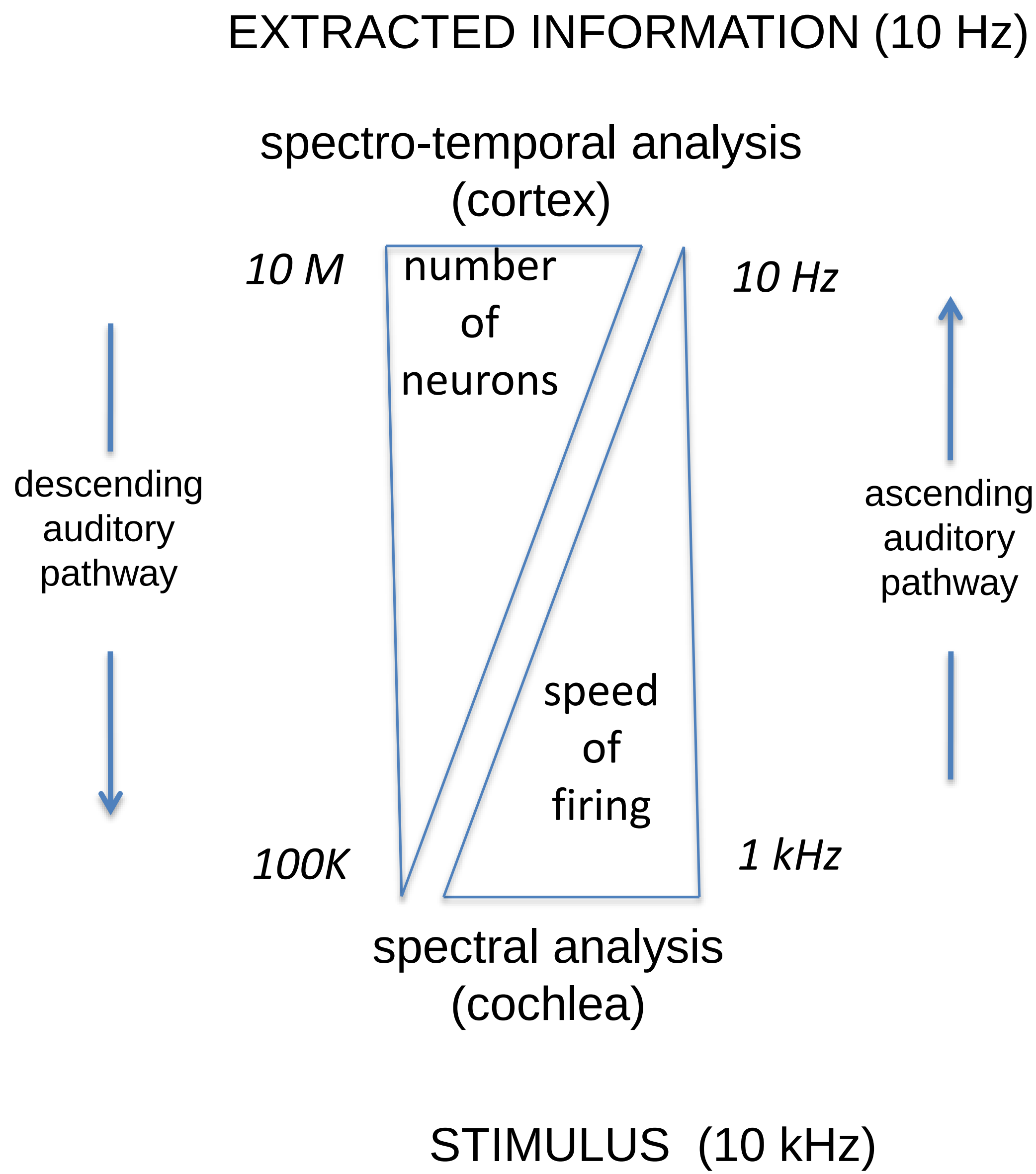
human

~ 10^{11} neurons

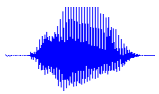


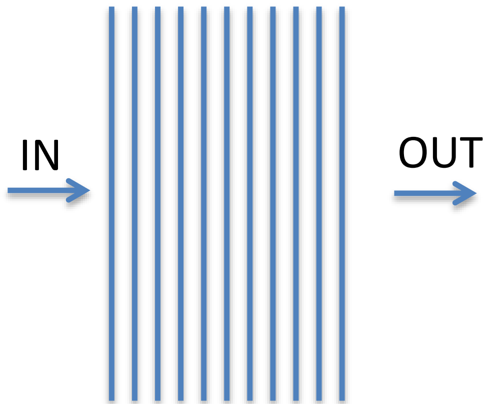
central
nervous
system

sensory organ


EXTRACTED INFORMATION (10 Hz)
spectro-temporal analysis
(cortex)
10 M
number of neurons
10 Hz
descending auditory pathway
ascending auditory pathway
speed of firing
100K
1 kHz
spectral analysis
(cochlea)
STIMULUS (10 kHz)

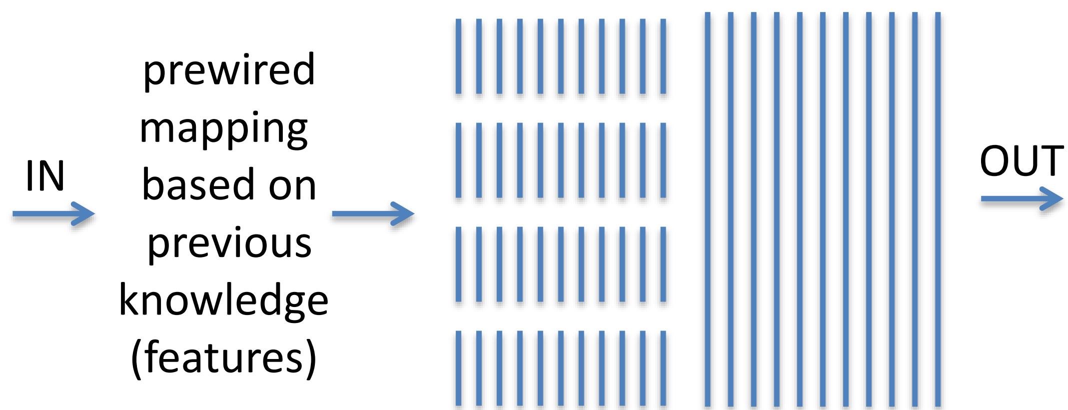
Multi-Layer Perceptron can emulate **any** nonlinear mapping ☺

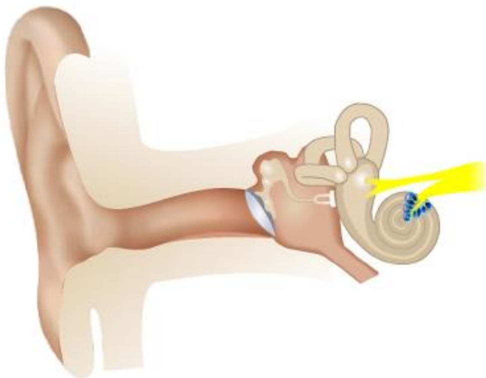
 → MLP → likelihood of speech sound



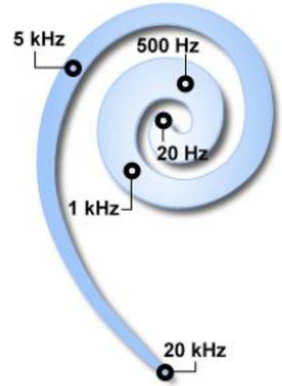
(given infinite size of the MLP and an infinite amount of training data ☹)_

Combinations of MLPs can emulate **useful** nonlinear mappings and perhaps can do it more efficiently

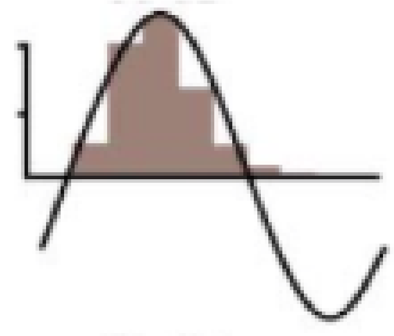




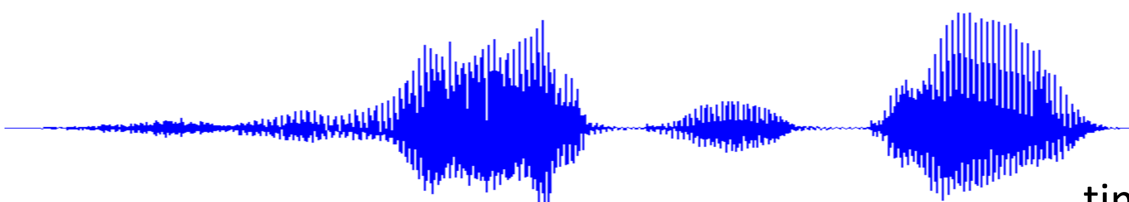
hearing periphery



cochlear frequency analysis



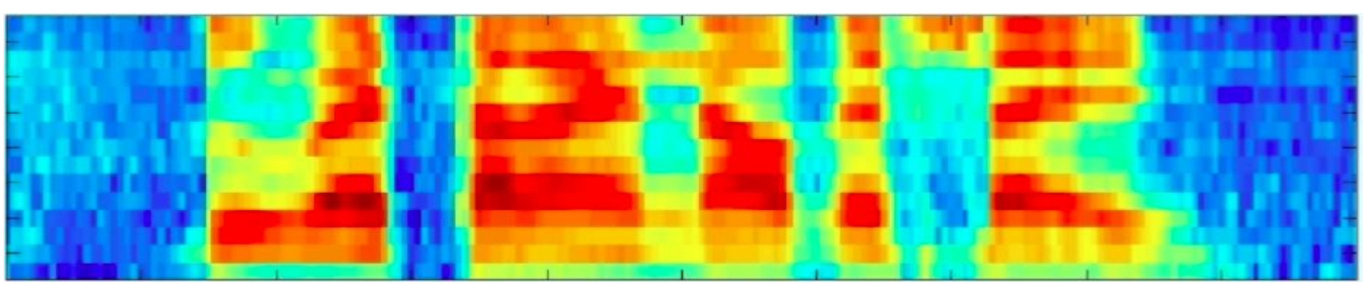
hair cell one-way rectification



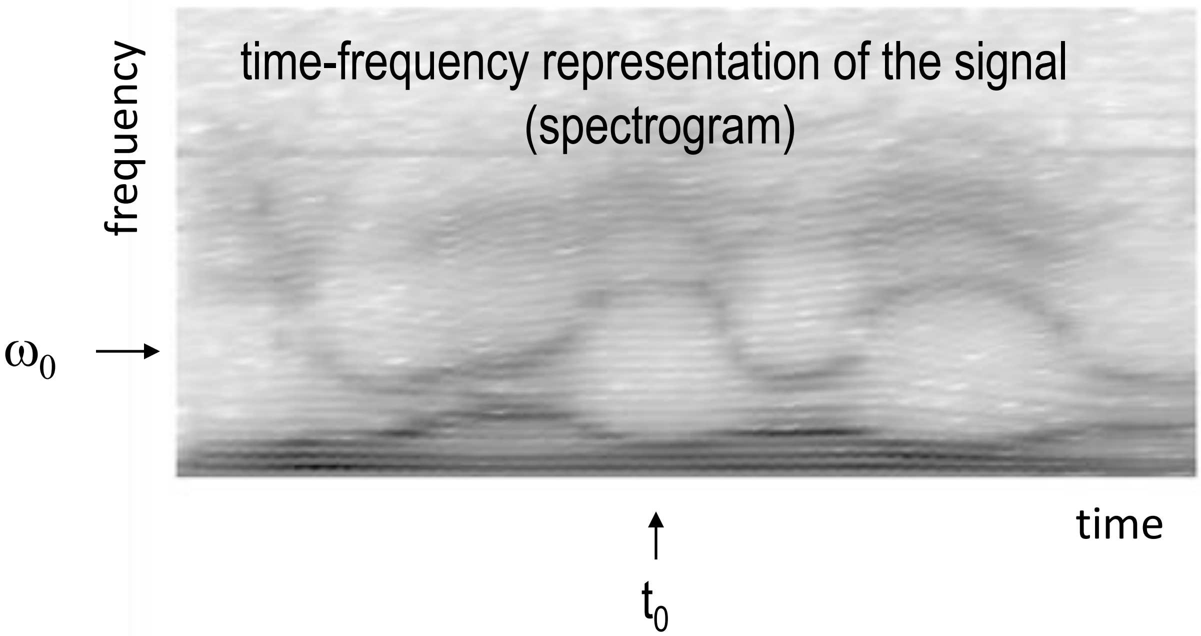
time

spectral analysis

frequency

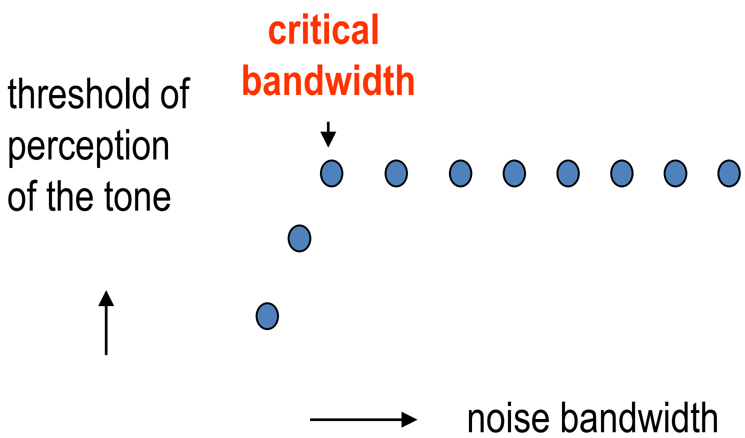


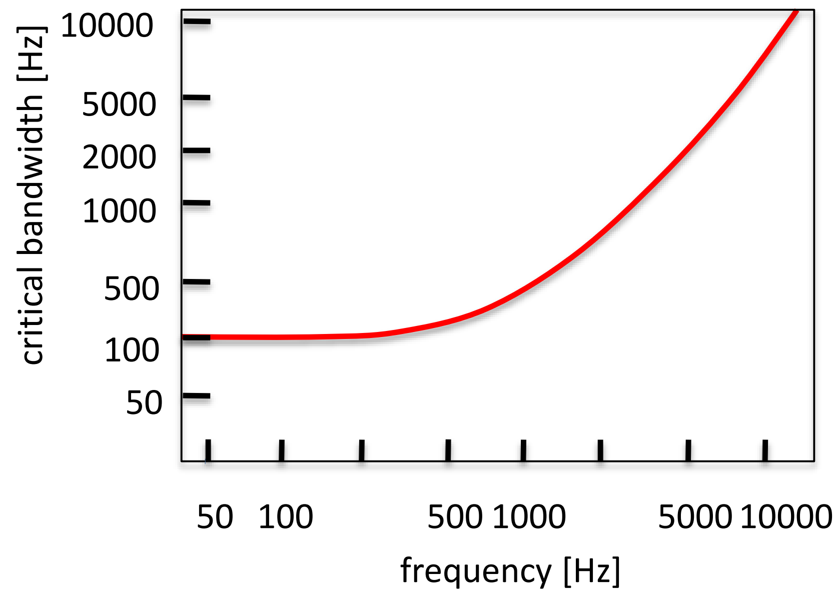
time



Which frequency bands ?

Simultaneous (frequency) masking





Better frequency resolution at lower frequencies

- also seen in
  - growth of loudness
  - perception of subthreshold stimuli

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

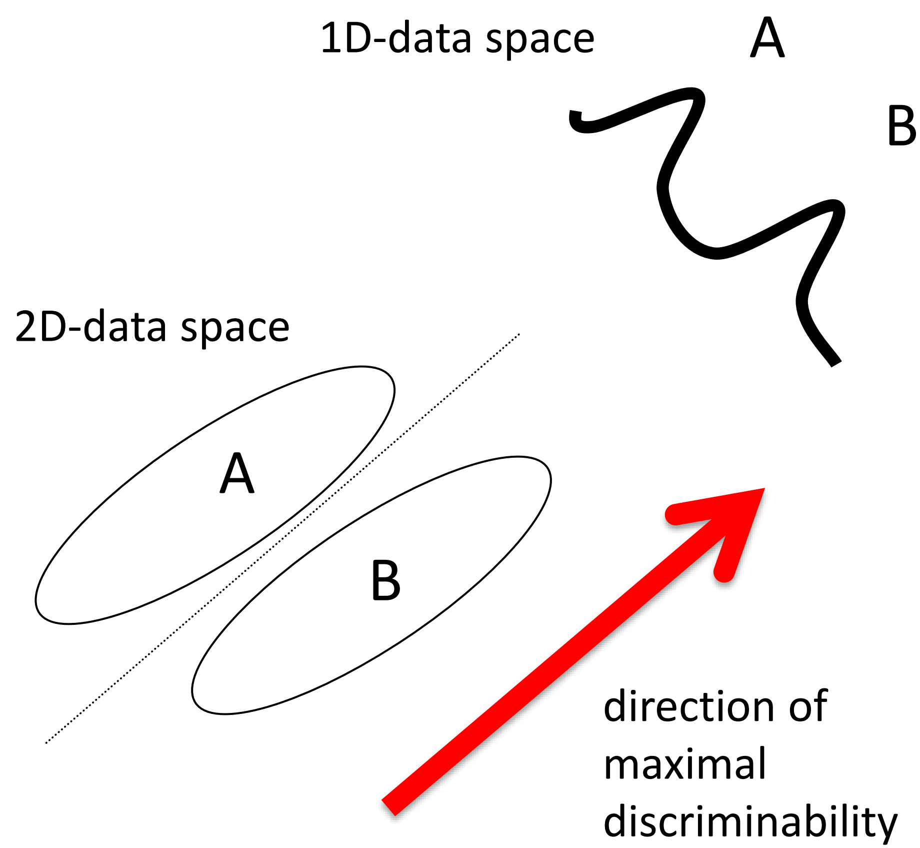
Linear Discriminant Analysis (LDA)

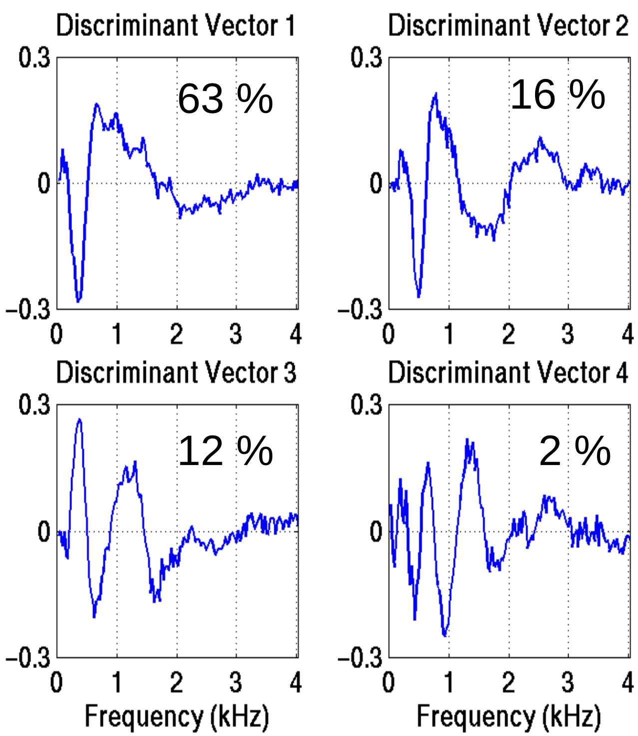
Linear discriminants: eigenvectors of $\Sigma^{-1}_{W}\Sigma_{B}$

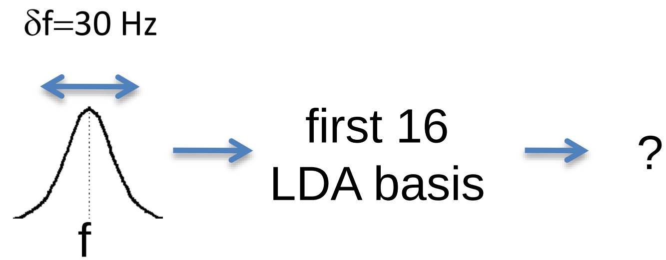
Σ_{W} - within-class covariance matrix

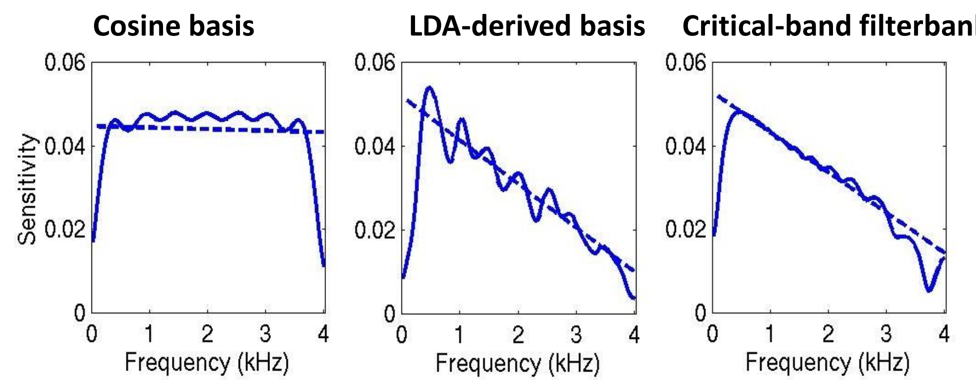
Σ_{B} - between class covariance matrix

- Needs labeled data









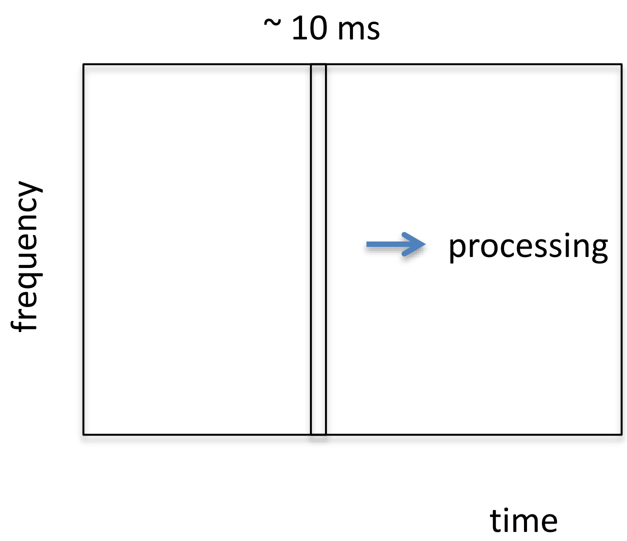
Malayath and Hermansky 2003, Valente and Hermansky 2006

Better frequency resolution at lower frequencies is desirable

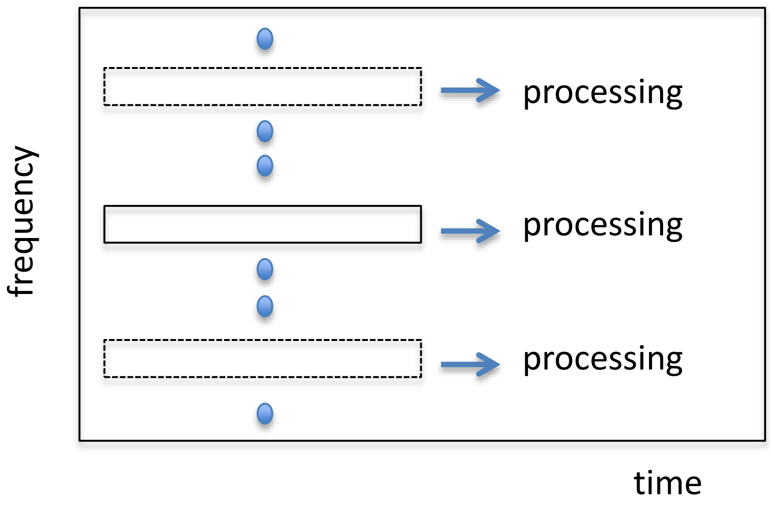
Better frequency resolution at lower frequencies

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

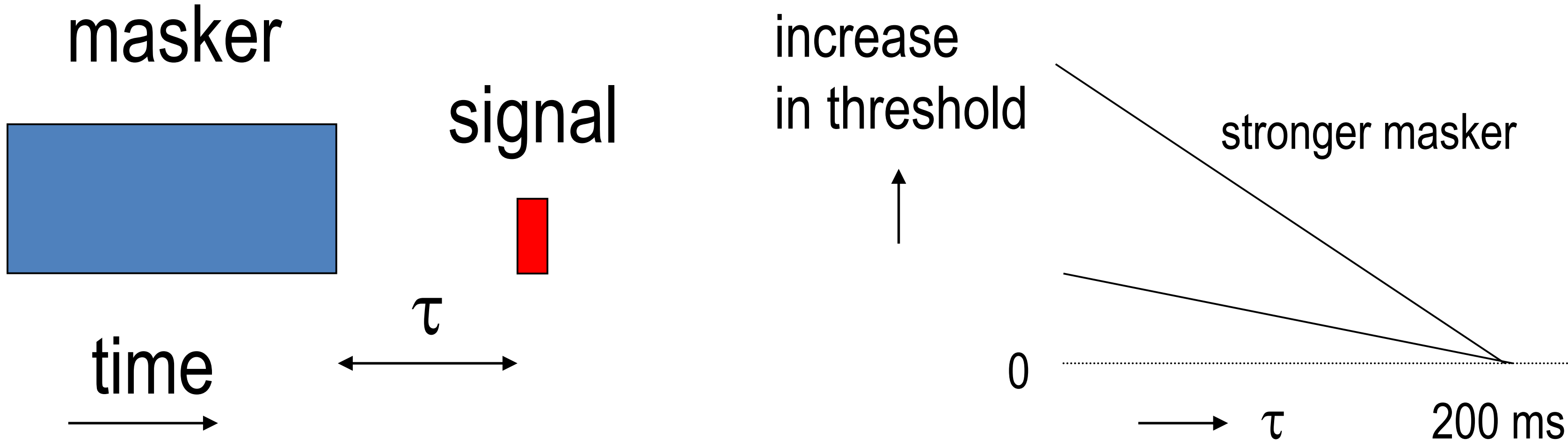
conventional



HOW LONG ?



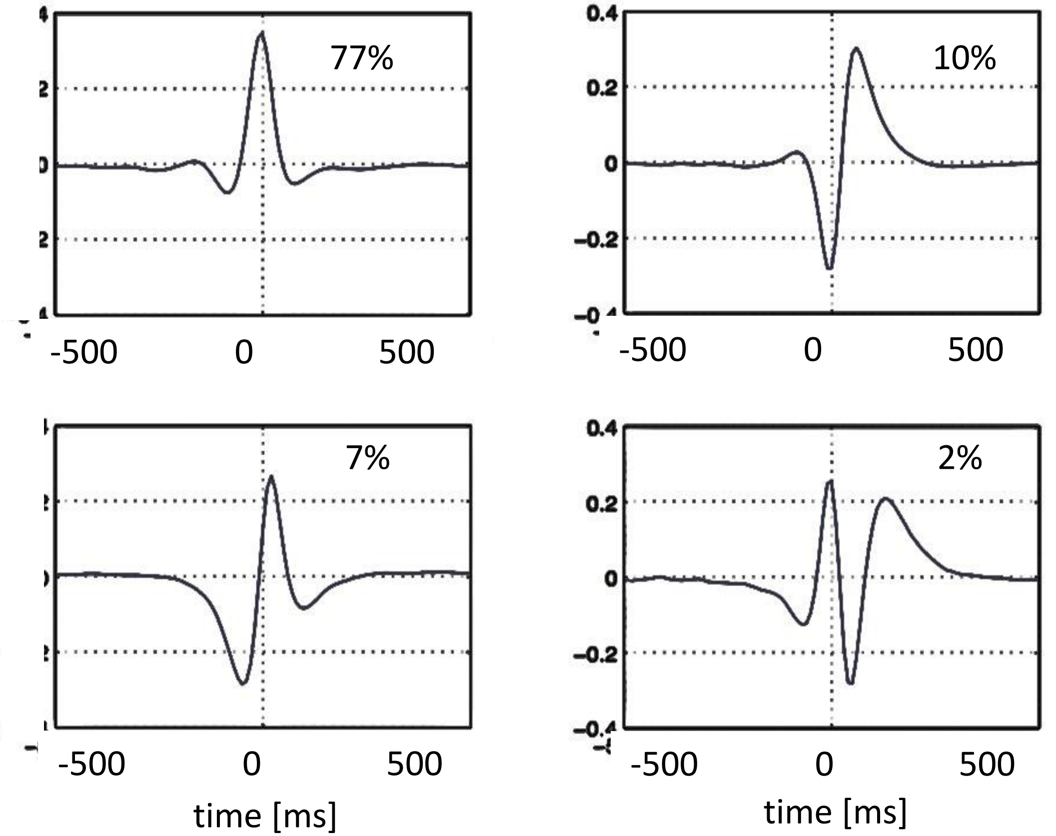
Masking in Time



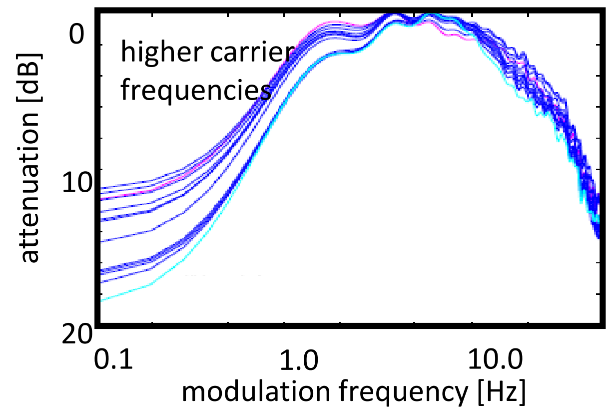
- suggests ~200 ms buffer in auditory system
  - also seen in perception of loudness, detection of short stimuli, gaps in tones, auditory afterimages, binaural release from masking, …..
- **Sound elements outside this buffer do not affect detection of signal within the buffer**

LDA on temporal trajectories of spectral energies

impulse responses

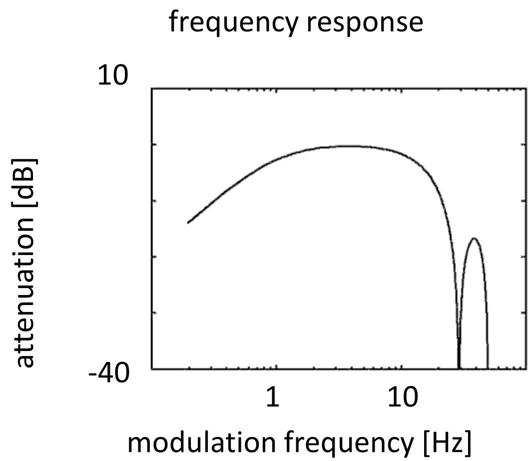


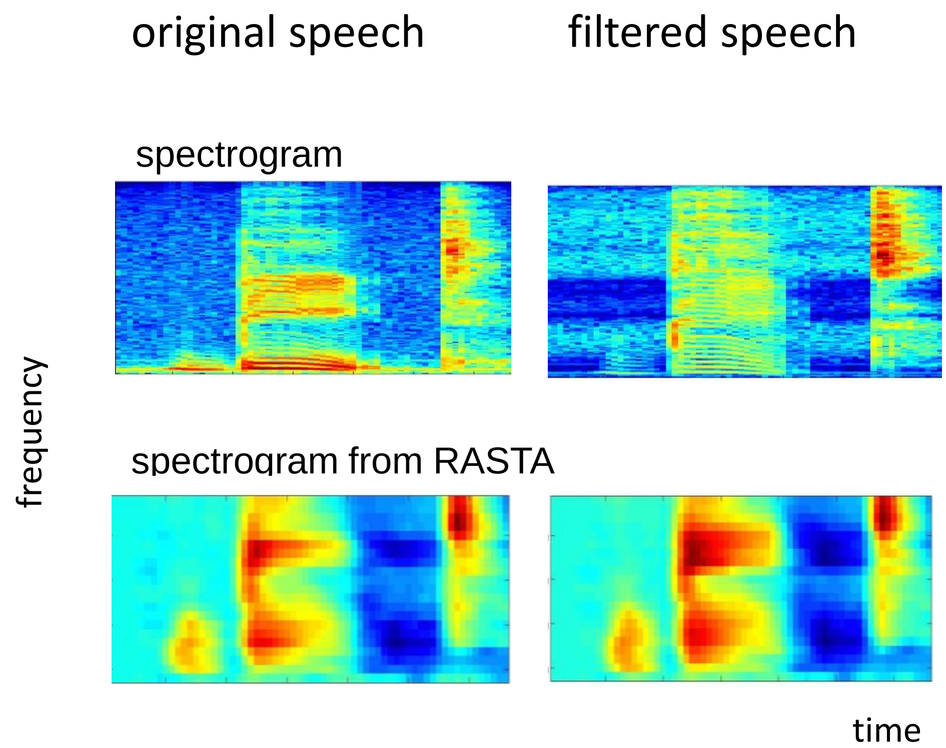
frequency responses
(1st discriminant in all frequency channels)



RASTA

Filter **each critical band** output by a band-pass filter





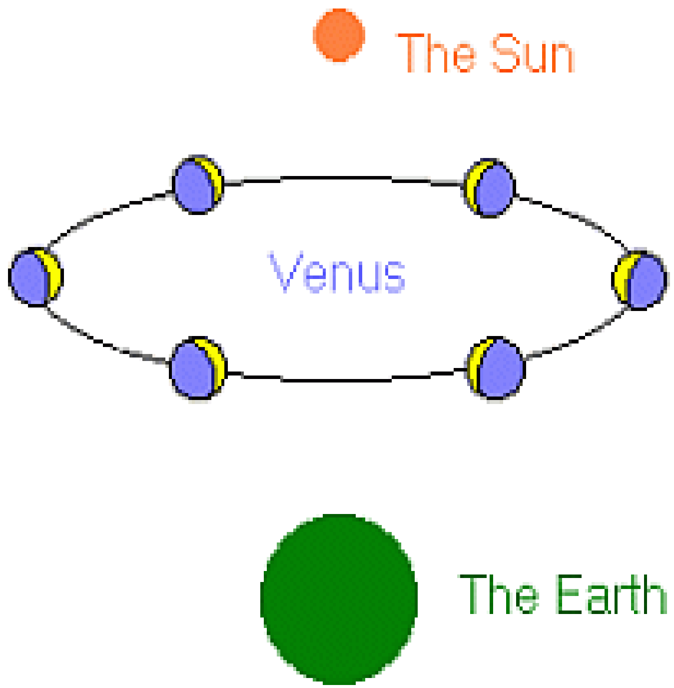
- pass modulations between 1-15 Hz

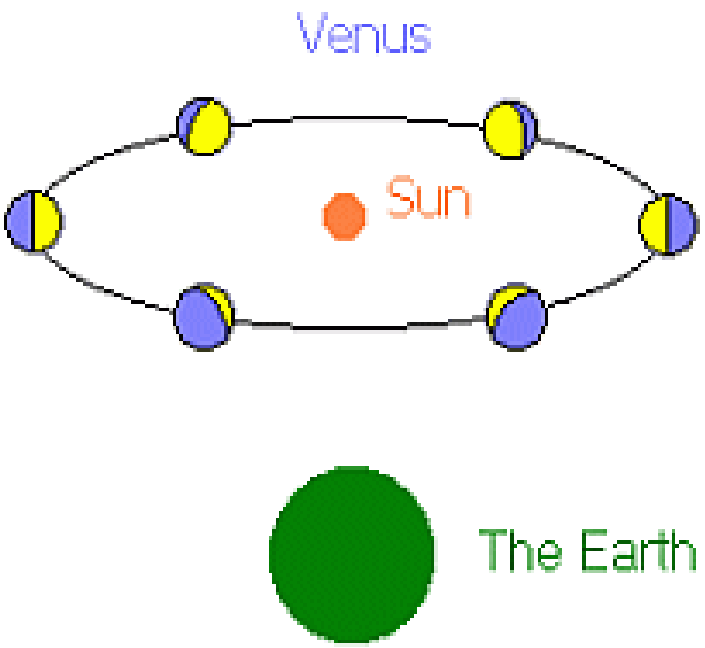
Environmental mismatch in training and in test

	matched	mismatched
conventional	2.8 % error	60.7% error
RASTA	**2.2 % error**	**2.9 % error**

Lesson From History

Galileo





Ptolemy

Ear is frequency selective

~~NOT in order to derive spectrum of the signal~~

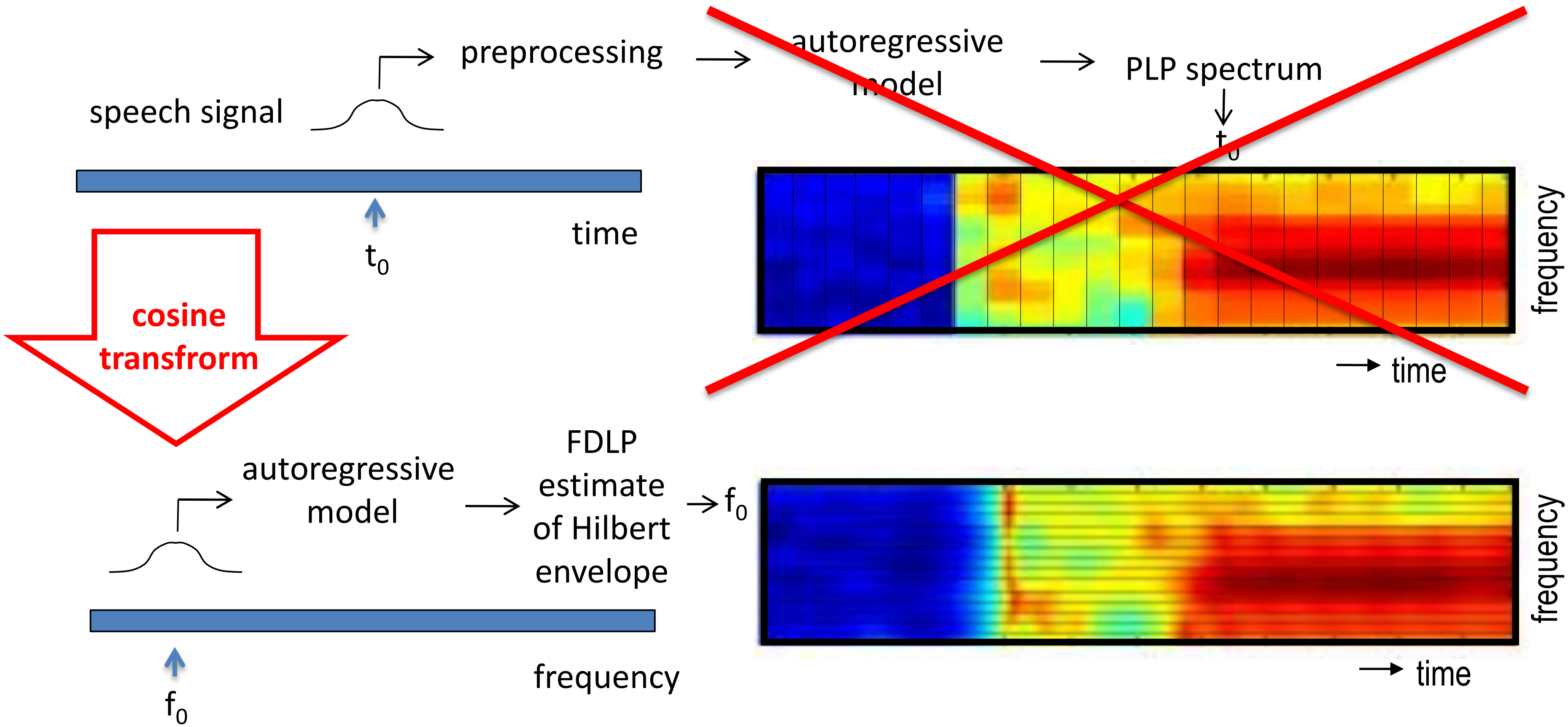
but

in order to yield frequency-localized temporal patterns.

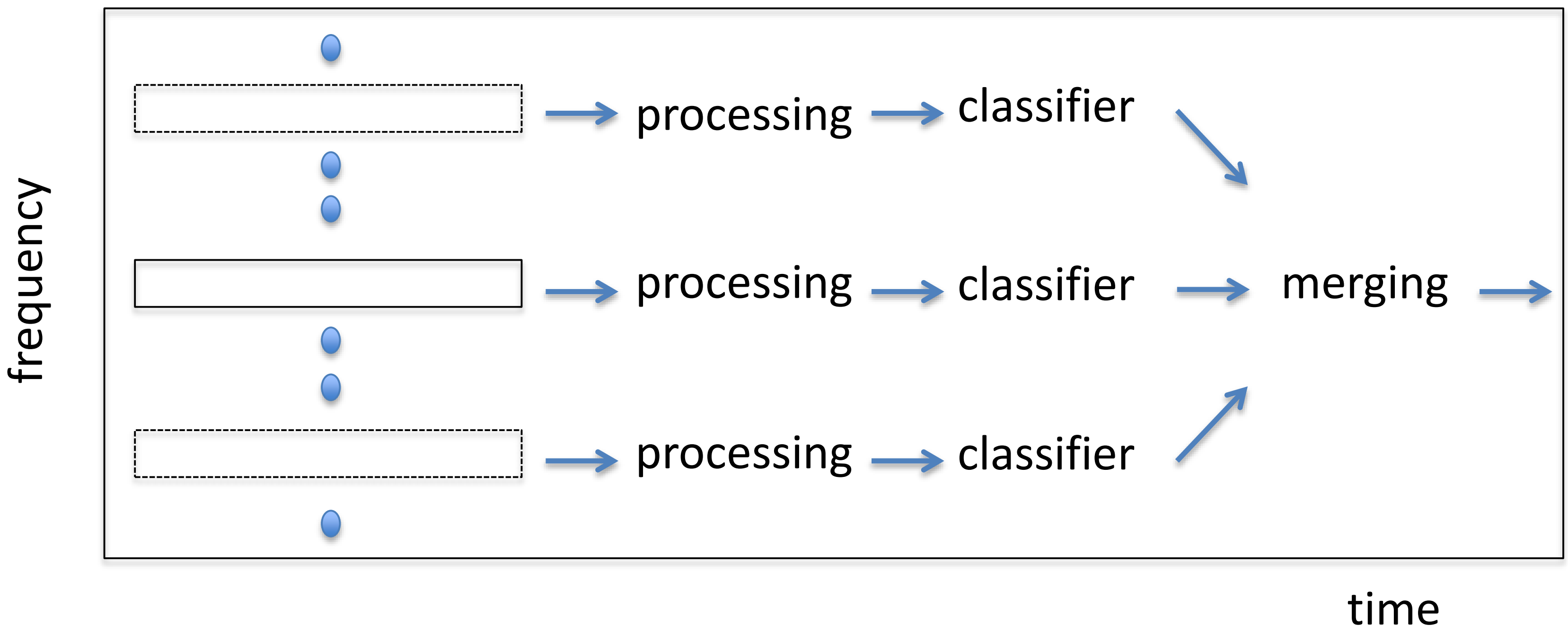
Frequency Domain Linear Prediction (FDLP)

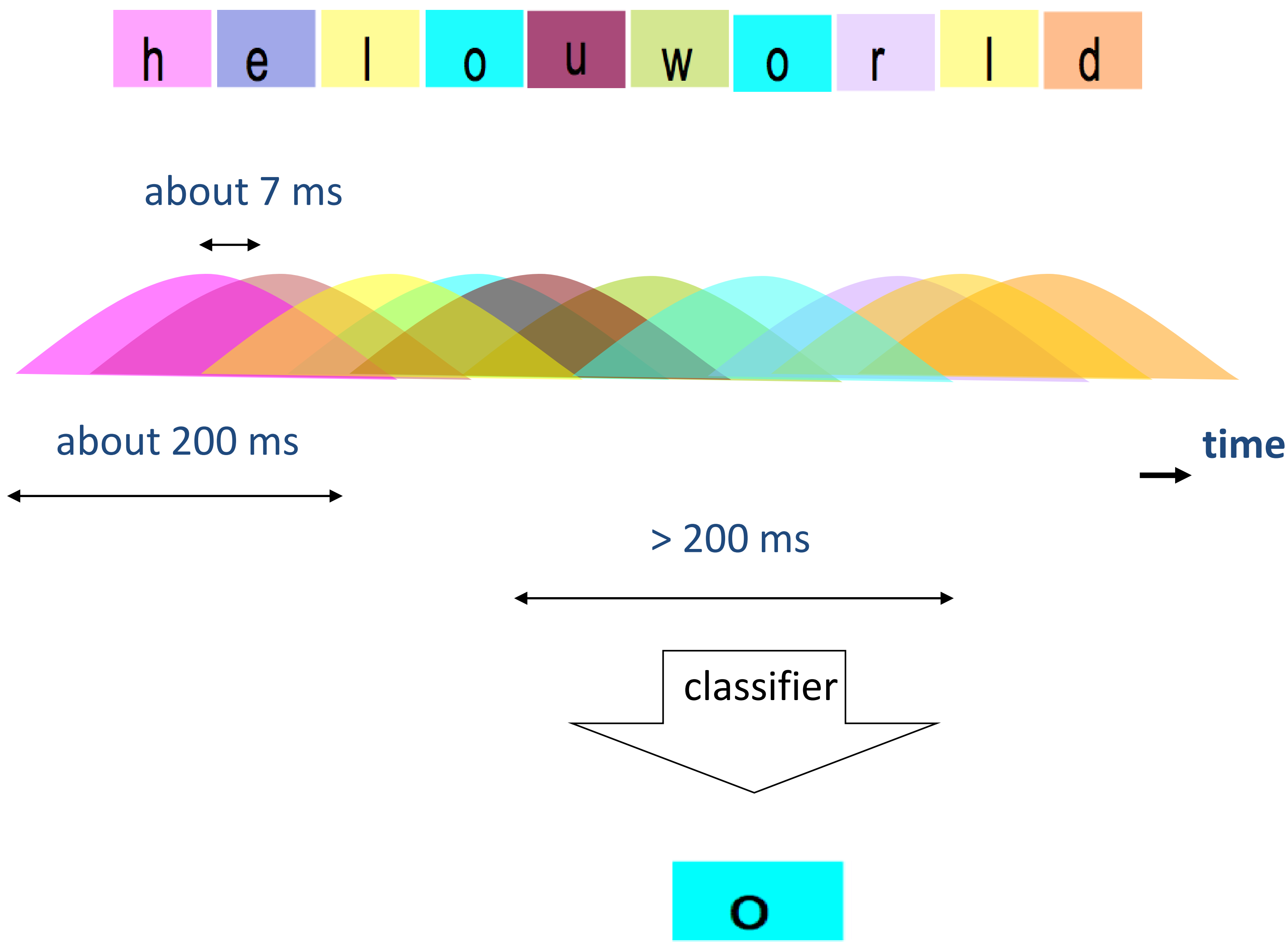
FDLP

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



200-400 ms

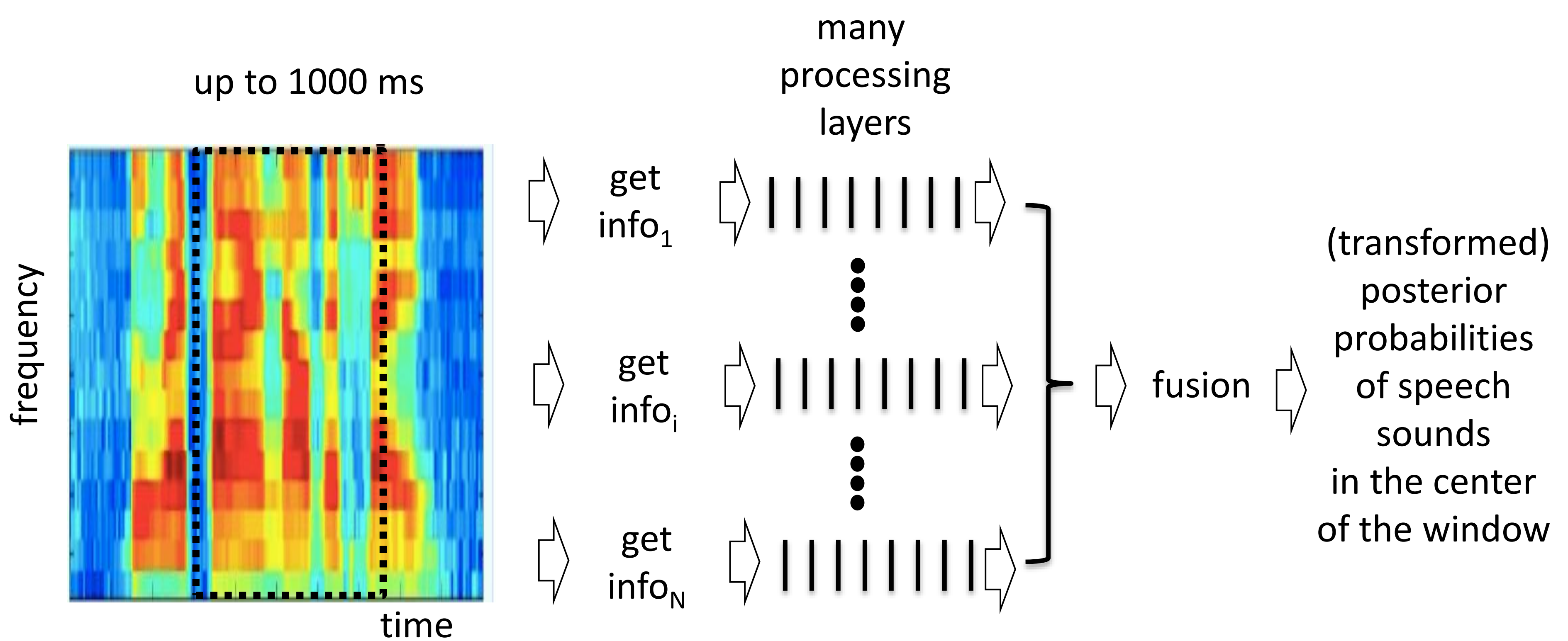

frequency
processing
classifier
processing
classifier
merging
processing
classifier
time



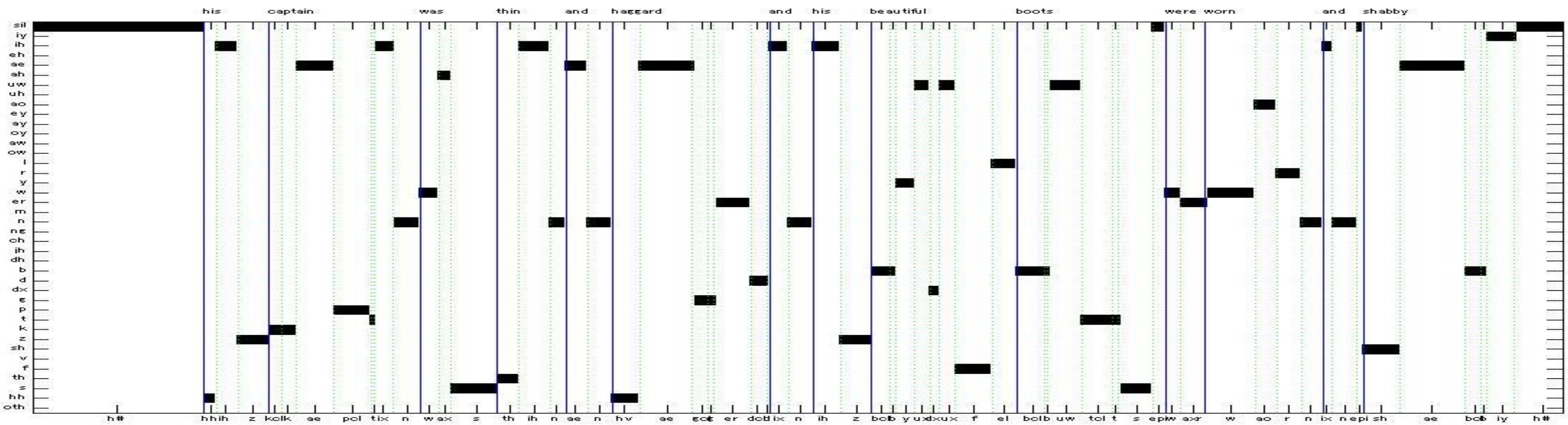
h
e
l
o
u
w
o
r
l
d
about 7 ms
about 200 ms
time
> 200 ms
classifier
o

Information in speech is coded hierarchically **(deep)**
in temporal dynamics **(long)**
and in many redundant dimensions **(wide)**

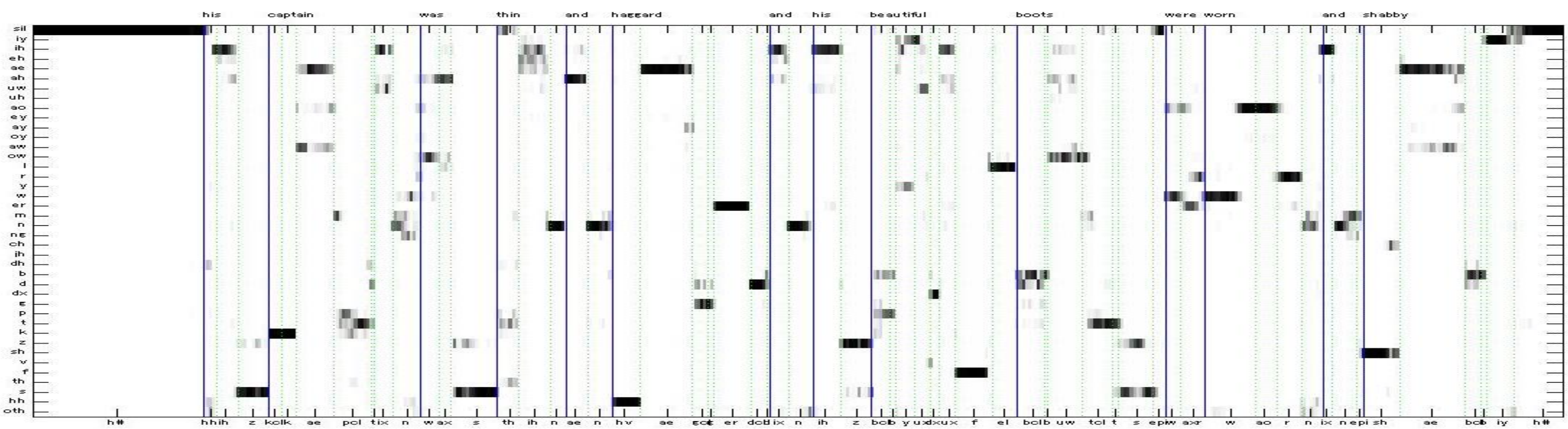
Deep, Long, and Wide Neural Nets



Labels

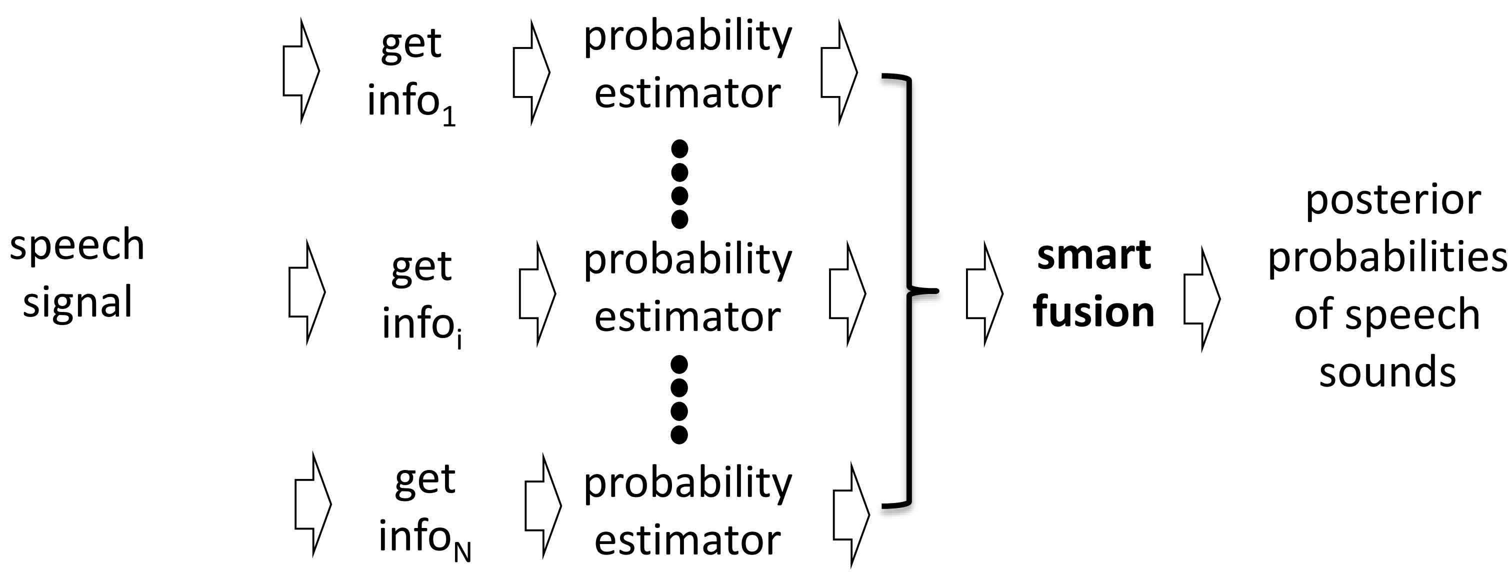


Long, wide and deep ANN estimates



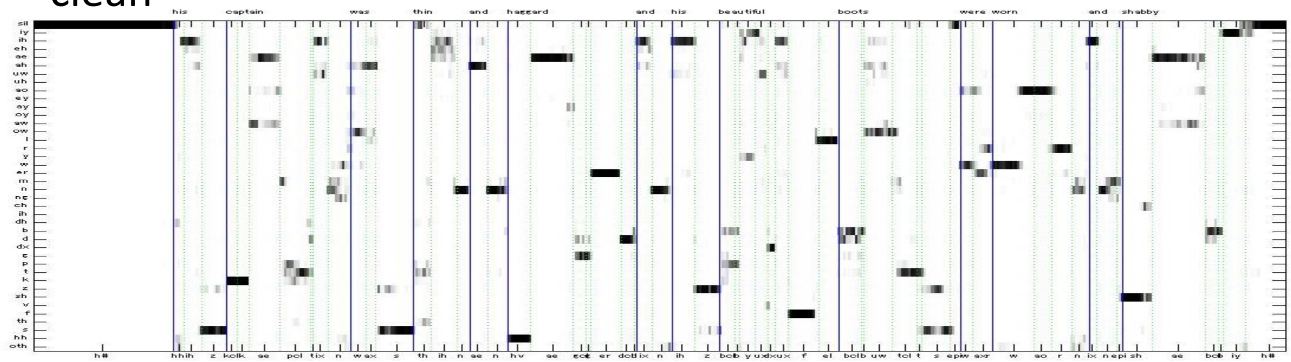
thanks Tetsuji Ogawa

Information in speech is coded in many redundant dimensions.
Not all dimensions get corrupted at the same time.

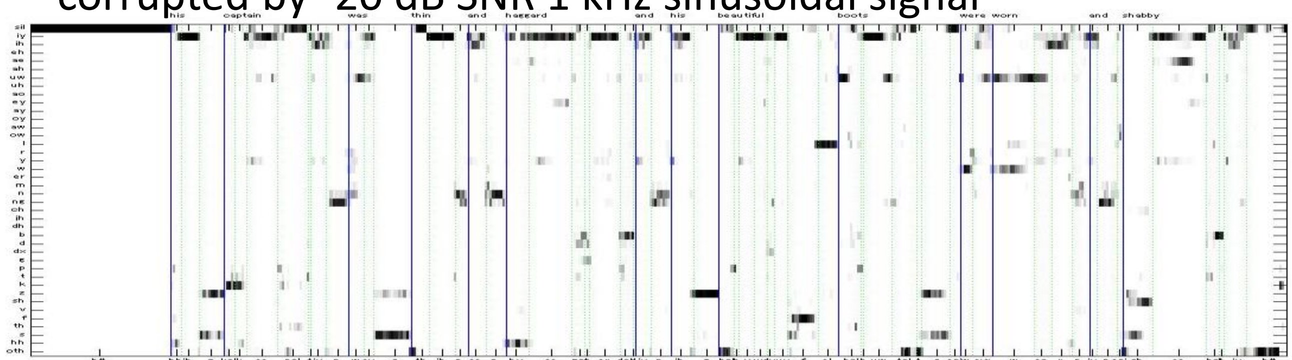


Smart fusion – alleviates unreliable processing streams

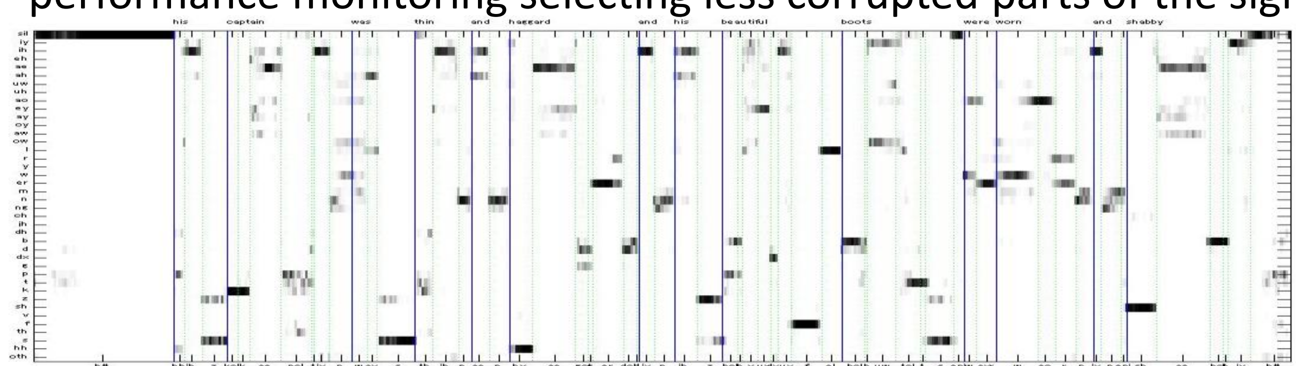
Probability estimator, which knows when it does not know clean



corrupted by -20 dB SNR 1 kHz sinusoidal signal



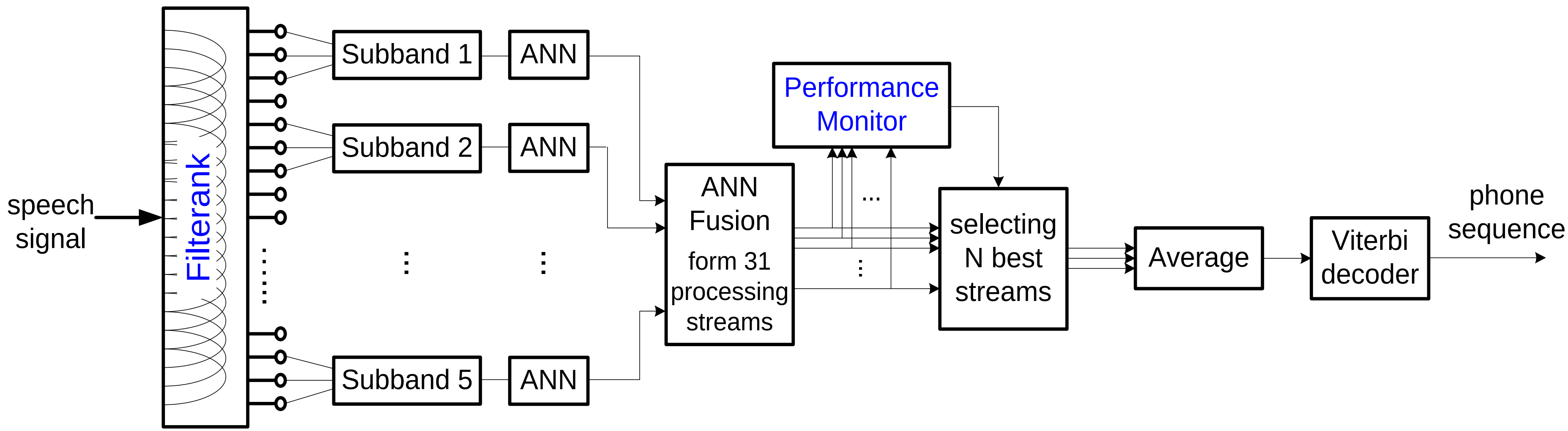
performance monitoring selecting less corrupted parts of the signal



thanks Tetsuji Ogawa

Multi-stream speech recognition

Variani, Li and Hermansky 2013

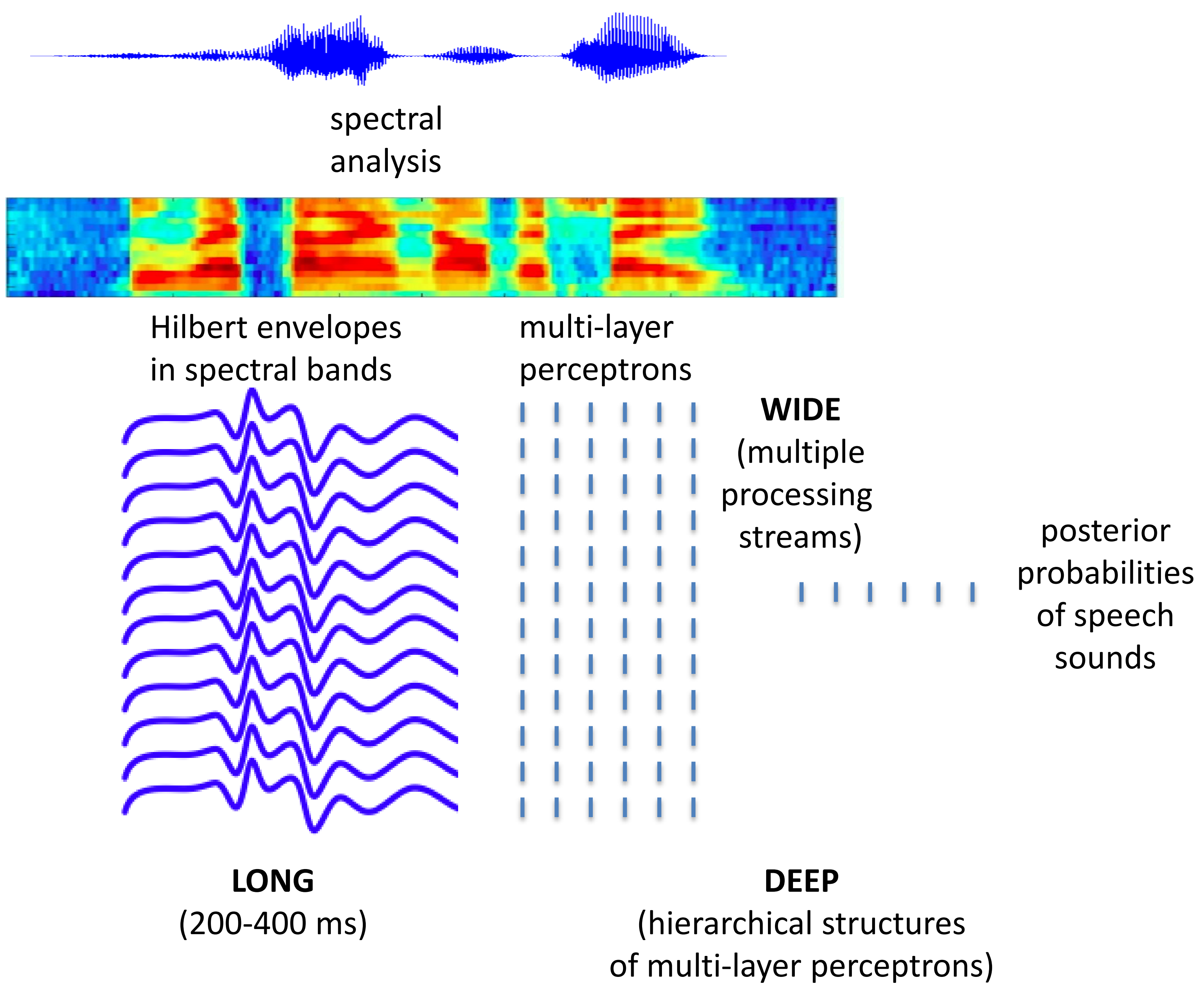


Phoneme recognition error rates

environment	conventional	proposed	best by hand
clean (matched training and test)	**31 %**	**29 %**	**24 %**
TIMIT with car noise at 0 dB SNR (training on clean)	**54 %**	**35 %**	**30 %**
RATS data (Channel E – matched training and test)	**70 %**	**54 %**	**49 %**

Conclusions

- Inputs to each local Deep Neural Net (DNN) should be frequency localized
- Data to each local DNNs should cover larger than 200-300 ms time spans
- Fusion from local DNNs should be done in a way that alleviates unreliable processing on local DNN levels


spectral analysis
Hilbert envelopes in spectral bands
multi-layer perceptrons
WIDE
(multiple processing streams)
posterior probabilities of speech sounds
LONG
(200-400 ms)
DEEP
(hierarchical structures of multi-layer perceptrons)