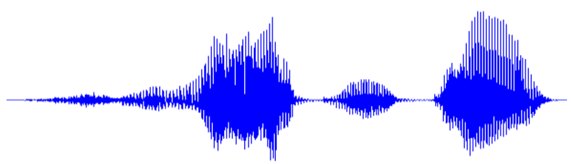
The Center For Language and Speech Processing
at the Johns Hopkins University

Auditory Perception in Speech Technology (Dealing with Unwanted Information)

Hynek Hermansky
Johns Hopkins University

Machine Recognition of Speech



↓

signal processing

↓

pattern classification

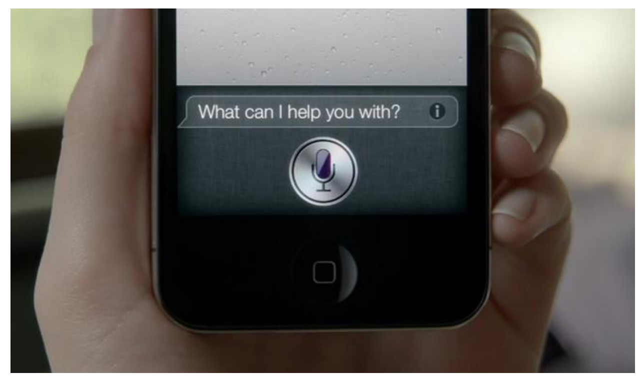
↓

decoder

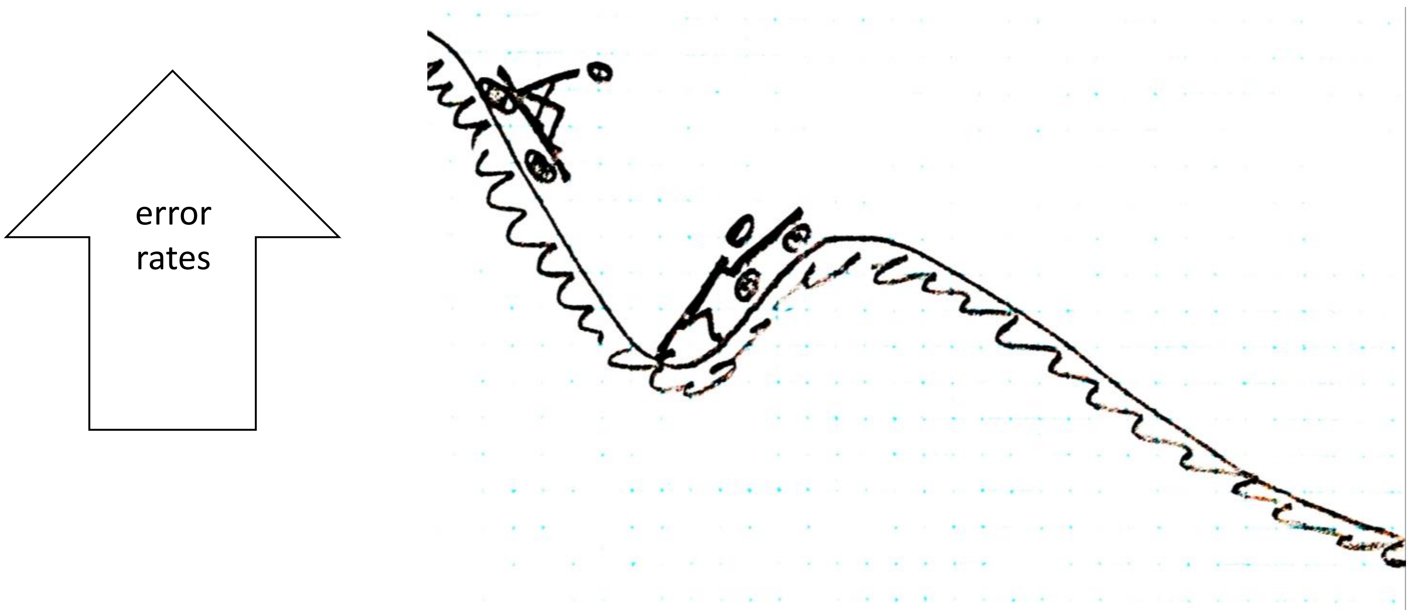
↓

message

Signal processing, information theory, machine learning, …



Why to rock the boat?
We have good thing going.



Repetition, fillers, hesitations, interruptions, unfinished and non-gramatical sentences, new words, dialects, emotions, …

Current DARPA and IARPA programs, research agenda of the JHU CoE HLT, industrial efforts (Google, Microsoft, IBM, Amazon,…)

Signal processing, information theory, machine learning, …

&

neural information processing, psychophysics, physiology, cognitive science, phonetics and linguistics, ...

Engineering and Life Sciences together !

… or at least engineering inspired by life sciences

1-2-3-6-7-49

first child
my mother's 2nd marriage
my father's 3rd marriage
was born of 6th of July

$6 \times 7 = 49$

Auditory perception

object

How to survive in this hostile world?

perceived signal

What is the message (is there a danger or opportunity ?

How to survive in this world?

"Eat vegetables, they are good for you"

"Eat vegetables, they are good for you"

Why machine recognition of speech?

Why did I climbed Mt. Everest?
Because it is there !
-Sir Edmund Hilary

Spoken language is one of the most amazing accomplishments of human race.

Addressing generic problems with human-like information processing (vision, e.t.c.)

access to information

- voice interactions with machines
- extracting information from speech data !

Job security - it will not be fully solved within your lifetime ☺

Speech

- Produced to be perceived
  - We speak in order to be heard in order to be understood

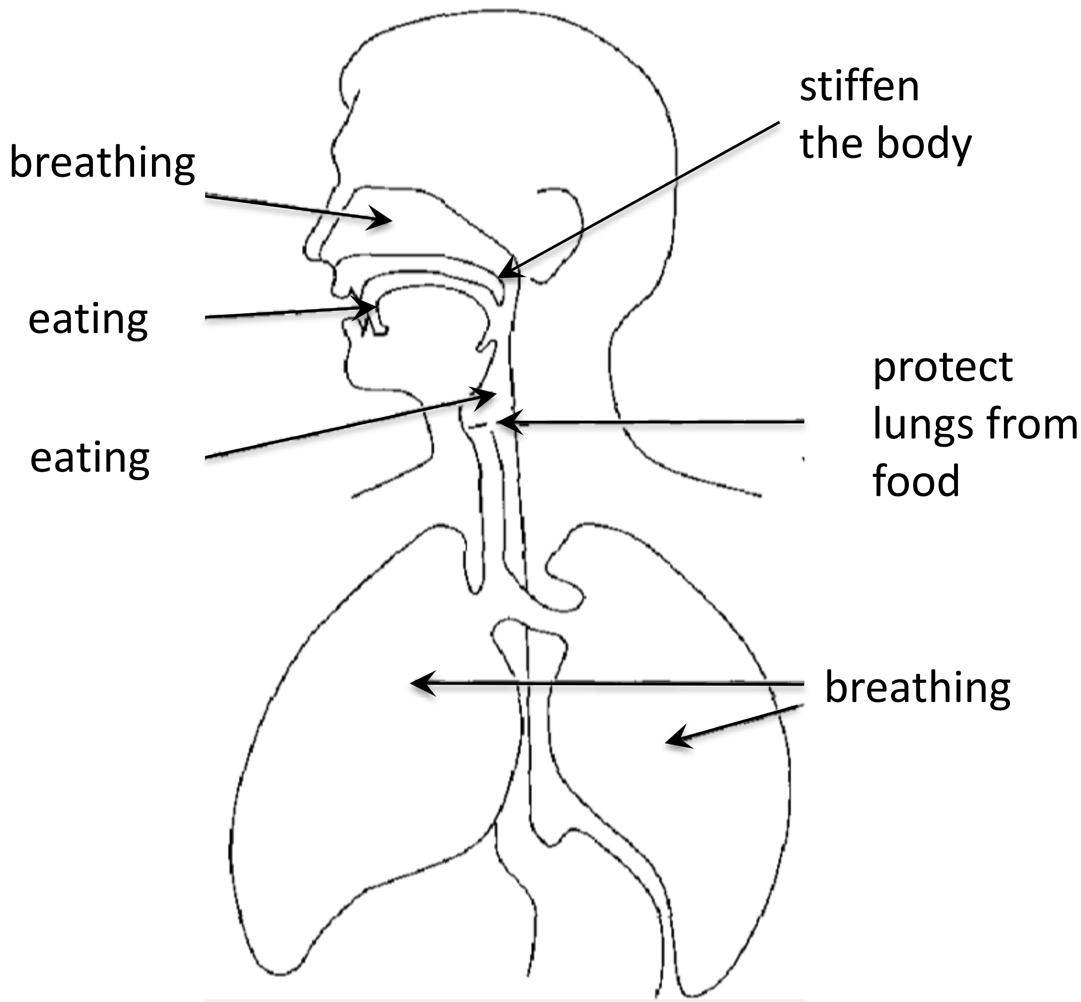
    Roman Jakobson

- Evolved over millennia to reflect properties of human hearing

Organs of speech production

Life sustaining functions:

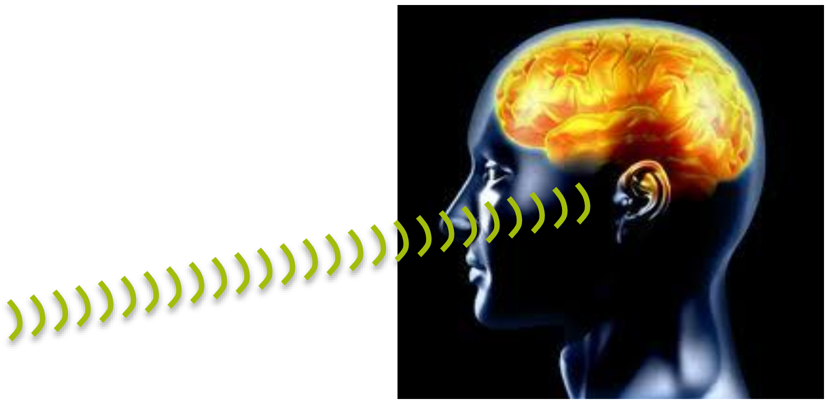
- eating
- breathing
- (and speaking)



How to survive in this world?

“Eat vegetables, they are good for you”

“Eat vegetables, they are good for you”



What is the message?

cognitive aspects

- common code (language), context, prior experience, ...

reliable signal carrying the message

Information in speech signal

$C = W \log_2 [(S+N)/N)]$,
W-signal bandwidth,
S-power of signal, N-power of noise

W – about 8 000 Hz *C about 80 kb/s*
(S+N)/N - about 10^3
log_2 *1000 – about 10*

standard PCM coding
8 kHz sampling, 11 bit
accuracy **= 88 kb/s**

$$H(s) = -\sum_{i=1}^{n} p_i \cdot \log(p_i)$$

p_i - probability of i - th symbol

41 phonemes in English
H=$\log_2 41$ = 5.4 bit/phoneme
about 15 phonemes/s
15 x 5.4 **= 80 bps**

150,000 words – about 18 bits,
300 words/min – 90 bits/s

considering relative frequencies of phonemes and phonotactic rules, the information in each phoneme decreases to about 1.5 bit/phoneme

15-25 bps ! (of course no other info but the phoneme sequences)

environmental noise

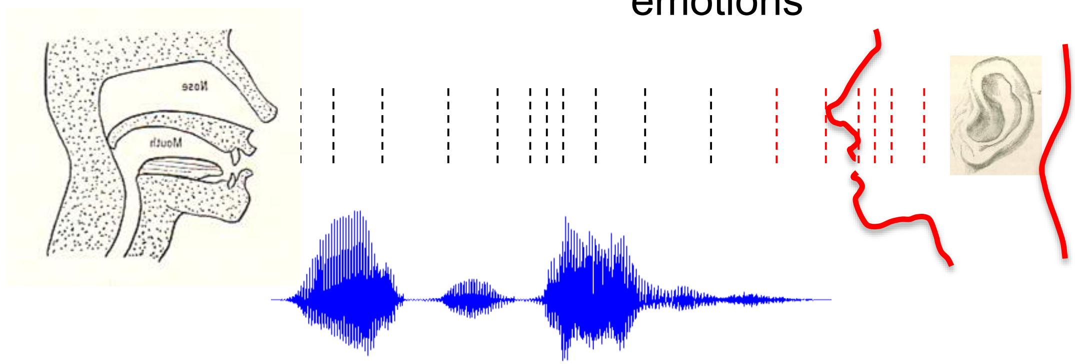
message

health
language
emotions

information

message

who is speaking
mood
social status



signal = message (wanted information)
noise = everything else (unwanted information)

*Get **signal** which carries desired information and ignores **noise***

The problem is NOT how to use all information but how to quickly IGNORE most of the information

frequency [Hz]

125 250 500 1000 2000 4000 8000

level [dB]

0 leaves

birds

20 water running

clock

40

speech

60 baby car noise

dog

80 piano telephone ringing

motorcycle

100 chain saw

120 rock band

Selectivity of perception

- different frequency spans
- different sound intensities
- different spectral and temporal dynamics
- different locations in physical space

e.t.c.

other

- selective attention (Mesgarani, Chang,..)
- e.t.c.

Human Speech Communication

message

linguistic code (~ 50 b/s)

motor control

speech production

SPEECH SIGNAL (> 50 kb/s)

speech perception

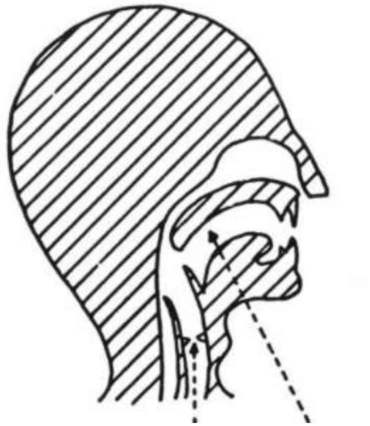
cognitive processes

linguistic code (~ 50 b/s)

message

Producing speech

We speak in order to be heard in order to be understood
Roman Jakobson



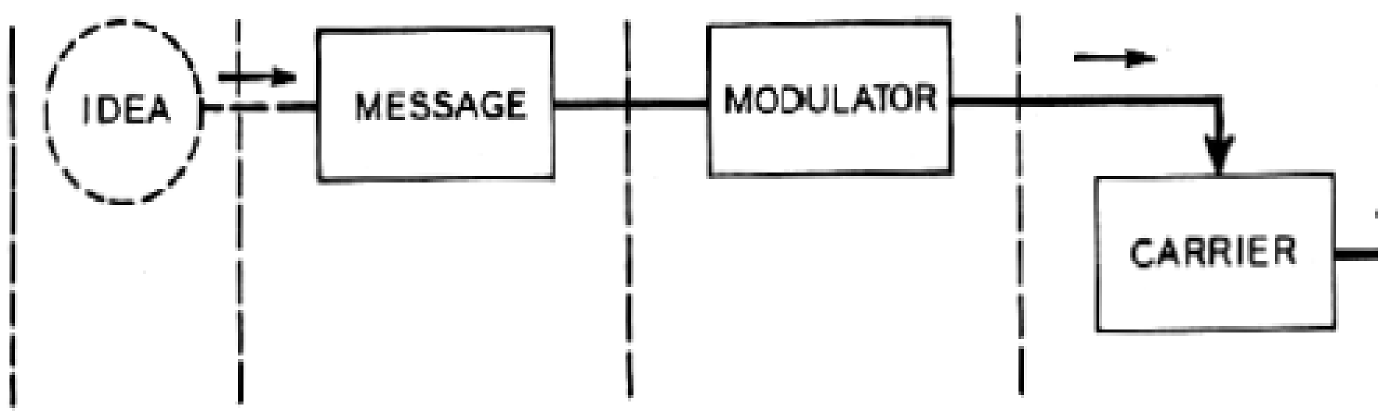
carrier message

H. Dudley 'The **carrier nature of speech** ', Bell System Technical Journal, vol. 19 (1940)

Inaudible **message** in slow motions of vocal tract is made audible by **modulating** the audible carrier

-Dudley 1940

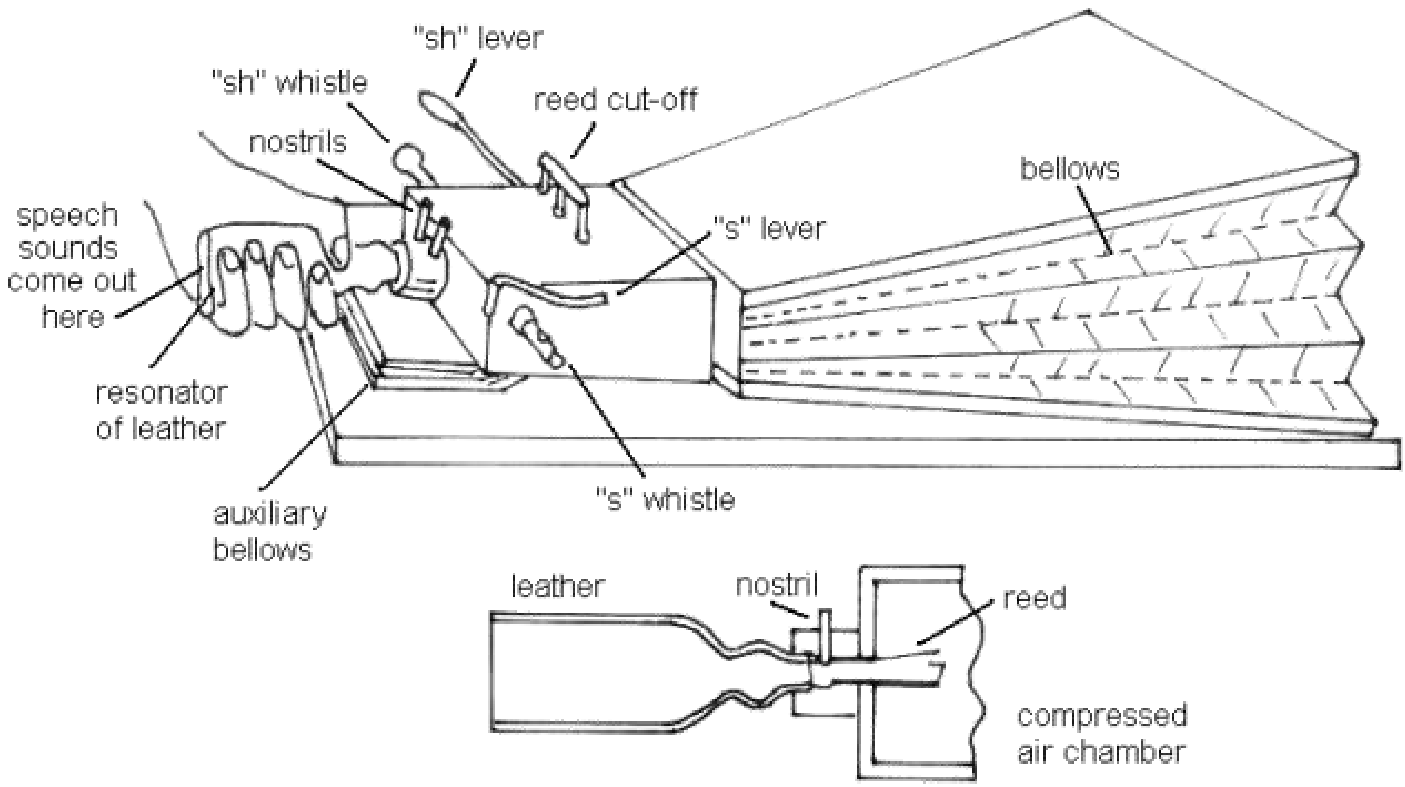
Source: Dudley, Homer. "The carrier nature of speech." Bell System Technical Journal 19, no. 4 (1940): 495-515.



speech

Producing speech

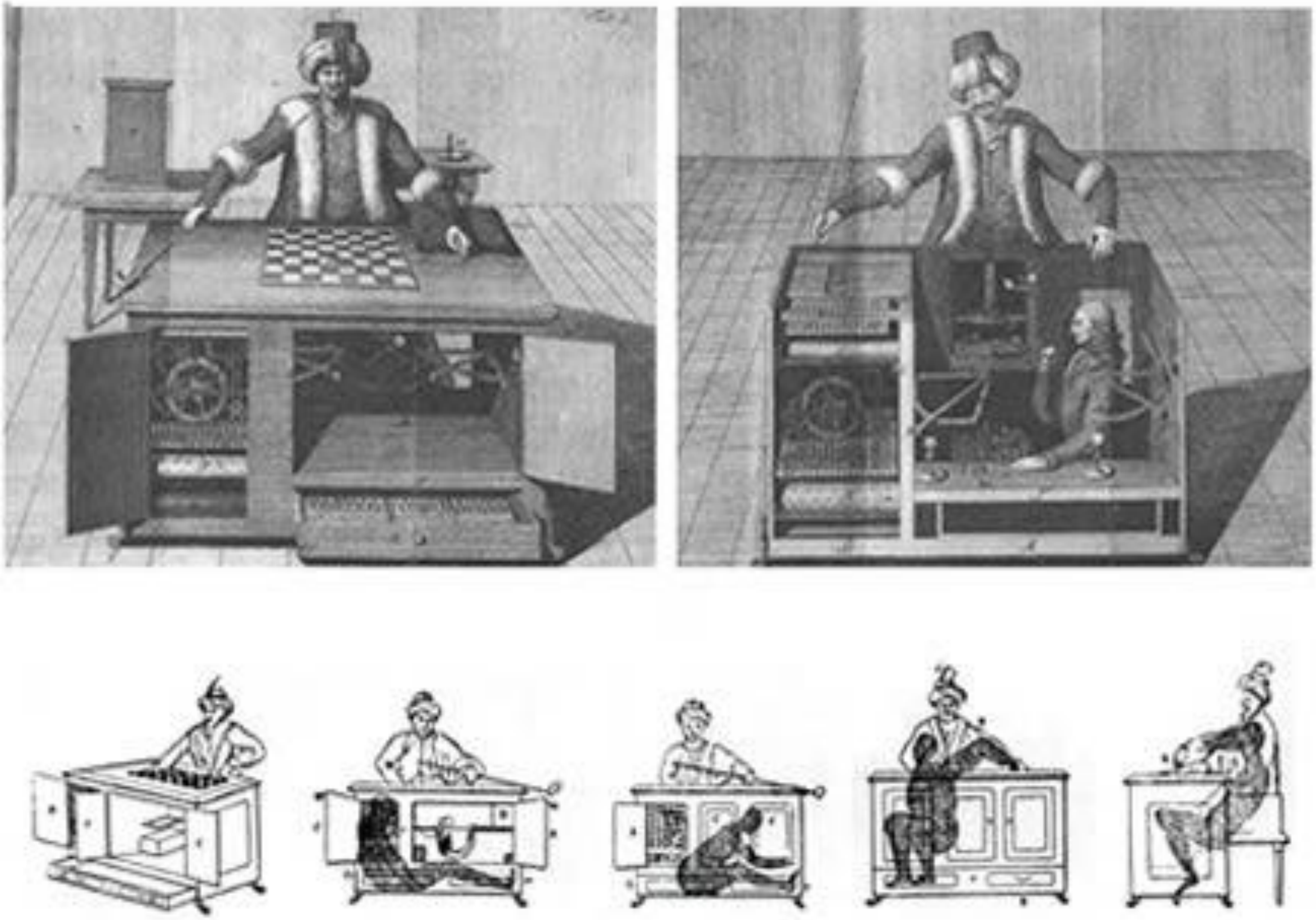
Johann Wolfgang Ritter **von Kempelen** de Pázmánd



This image of Wheatstone's construction of von Kempelen's speaking machine is in the public domain.

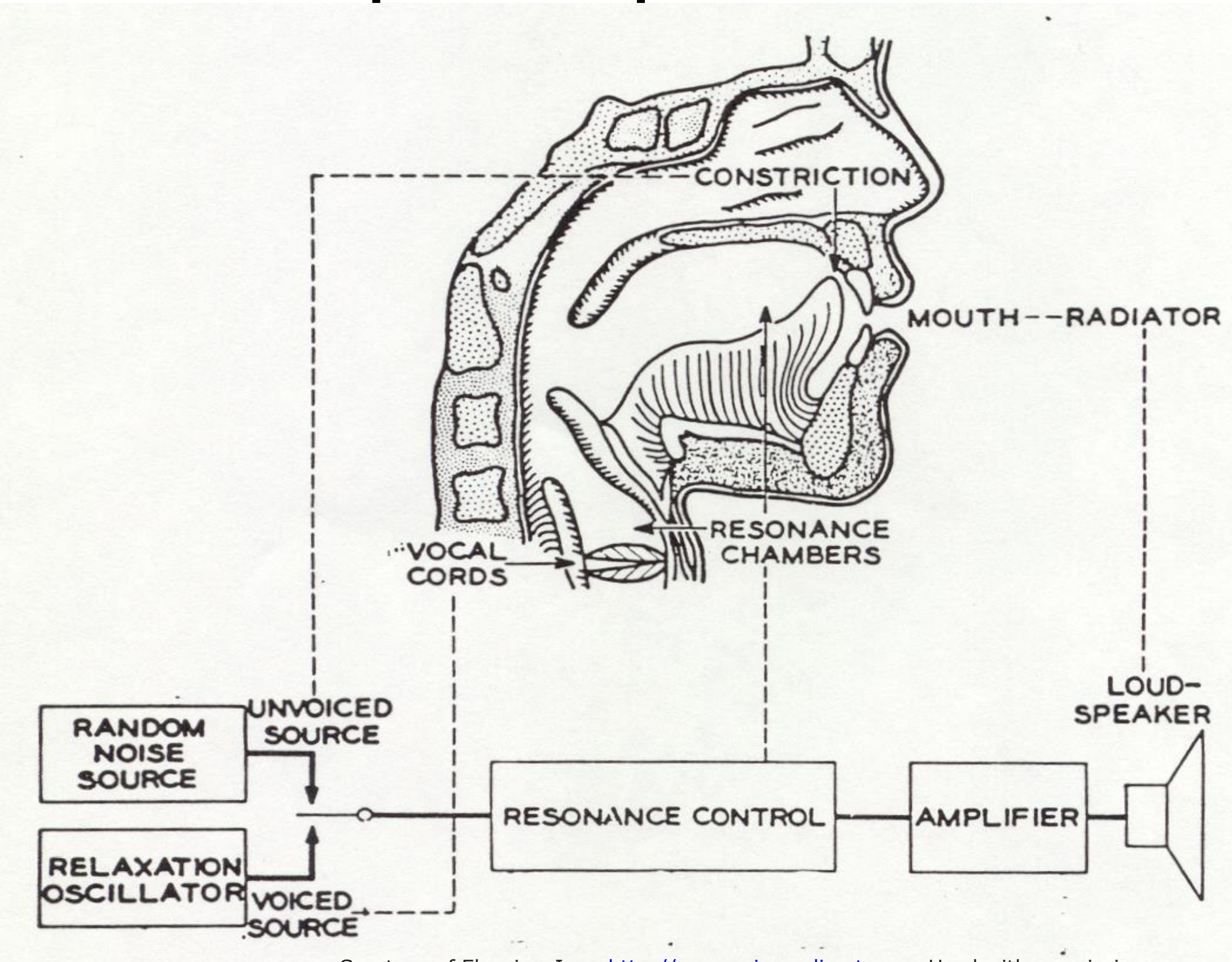
Mechanical Turk

Johann Wolfgang Ritter **von Kempelen** de Pázmánd



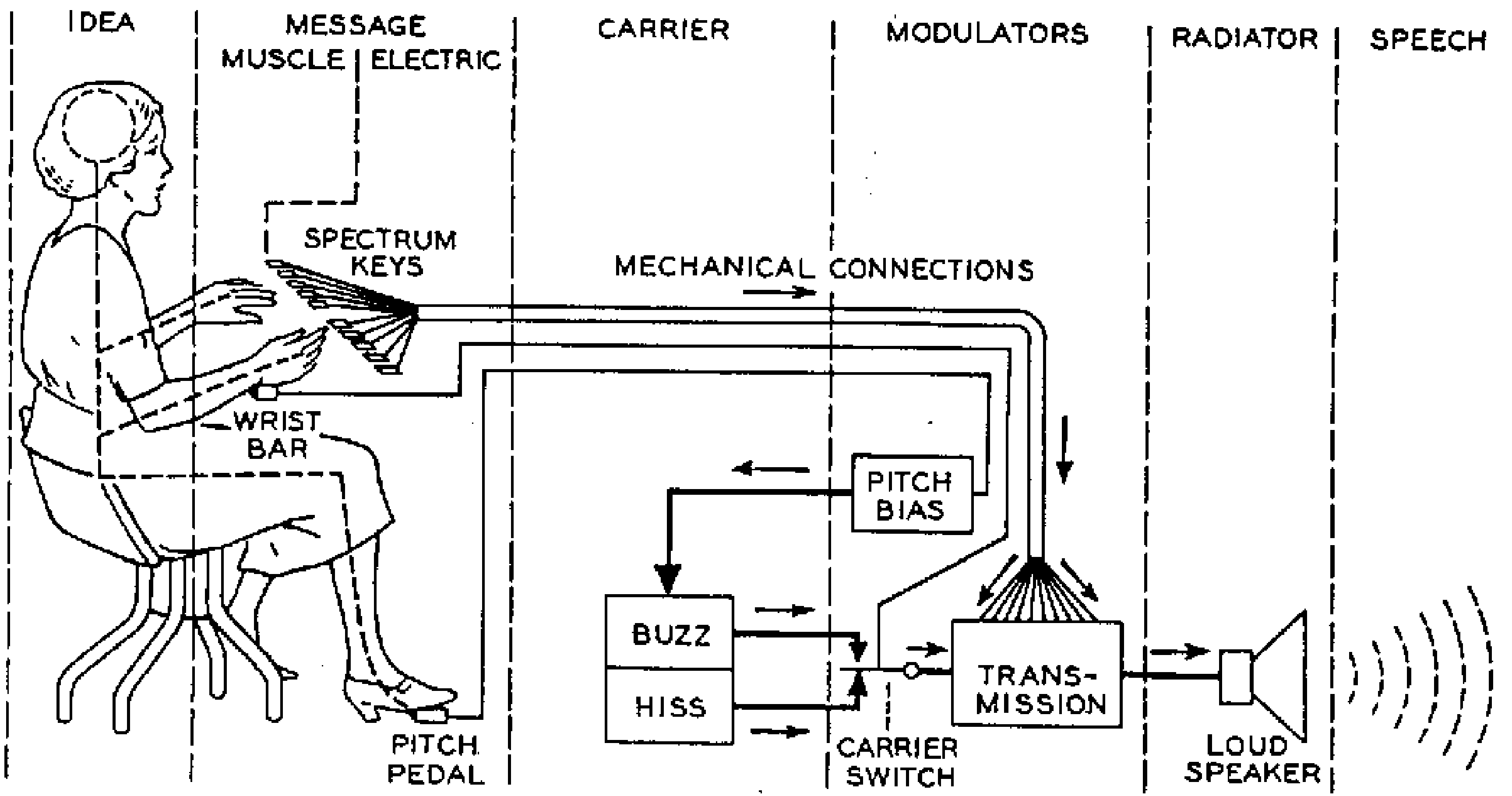
This image of the automaton chess player of von Kempelen is in the public domain.

Speech production



Source: Dudley, Riesz, and Watkins, "A synthetic speaker," Journal of the Franklin Institute. 227, 739 (1939).

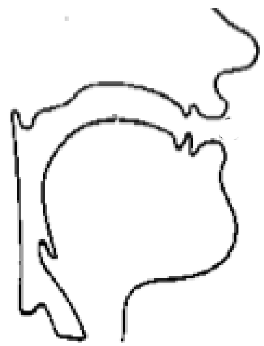
VODER
(Homer Dudley 1939)

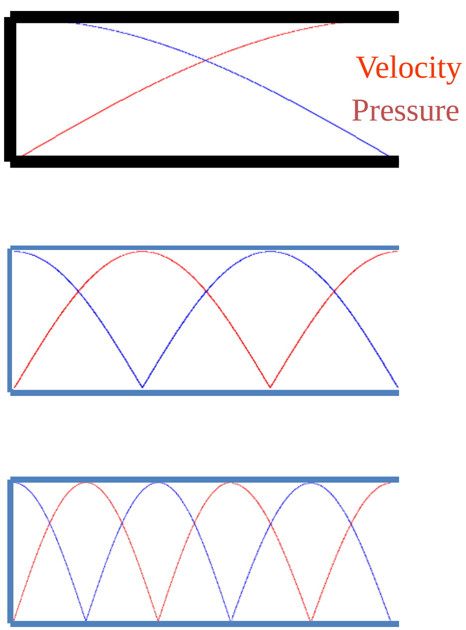


Source: Dudley, Homer. "The carrier nature of speech." Bell System Technical Journal 19, no. 4 (1940): 495-515.

- Constraining the tube af the point of its maximum velocity of the mode is the most efficient way to lower the mode frequency

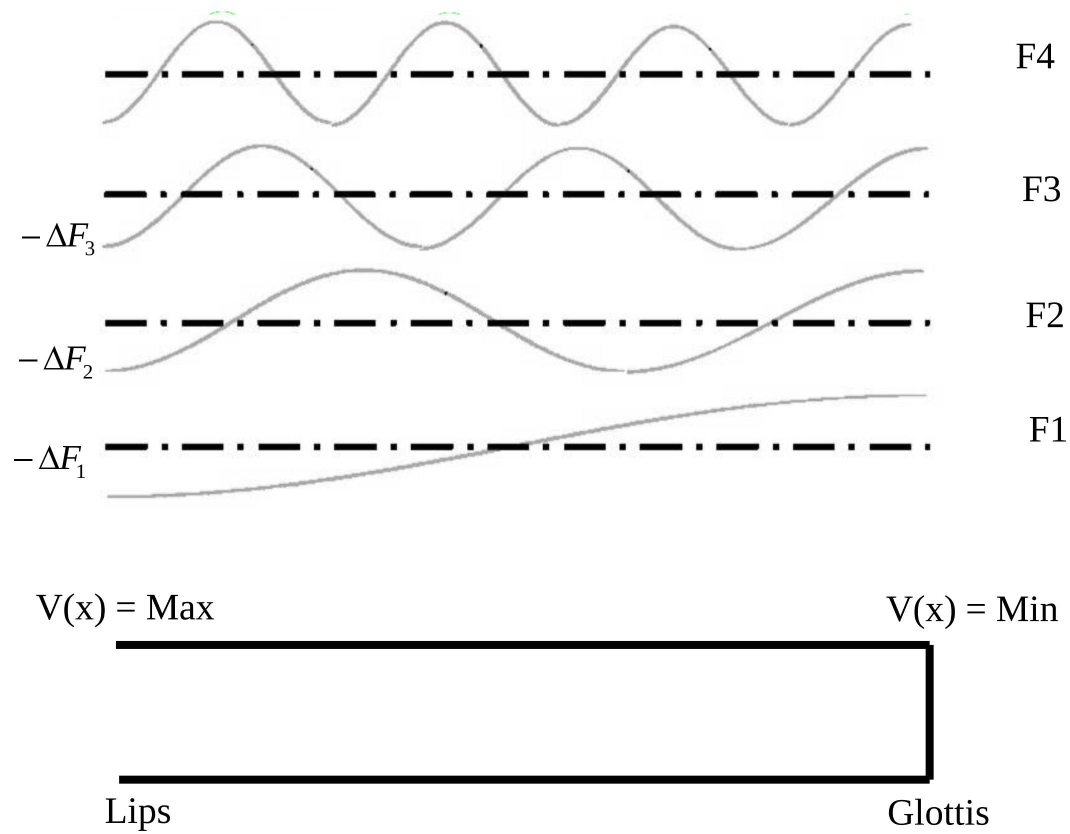
- Constraining it at the point of its maximum pressure lower the mode frequency

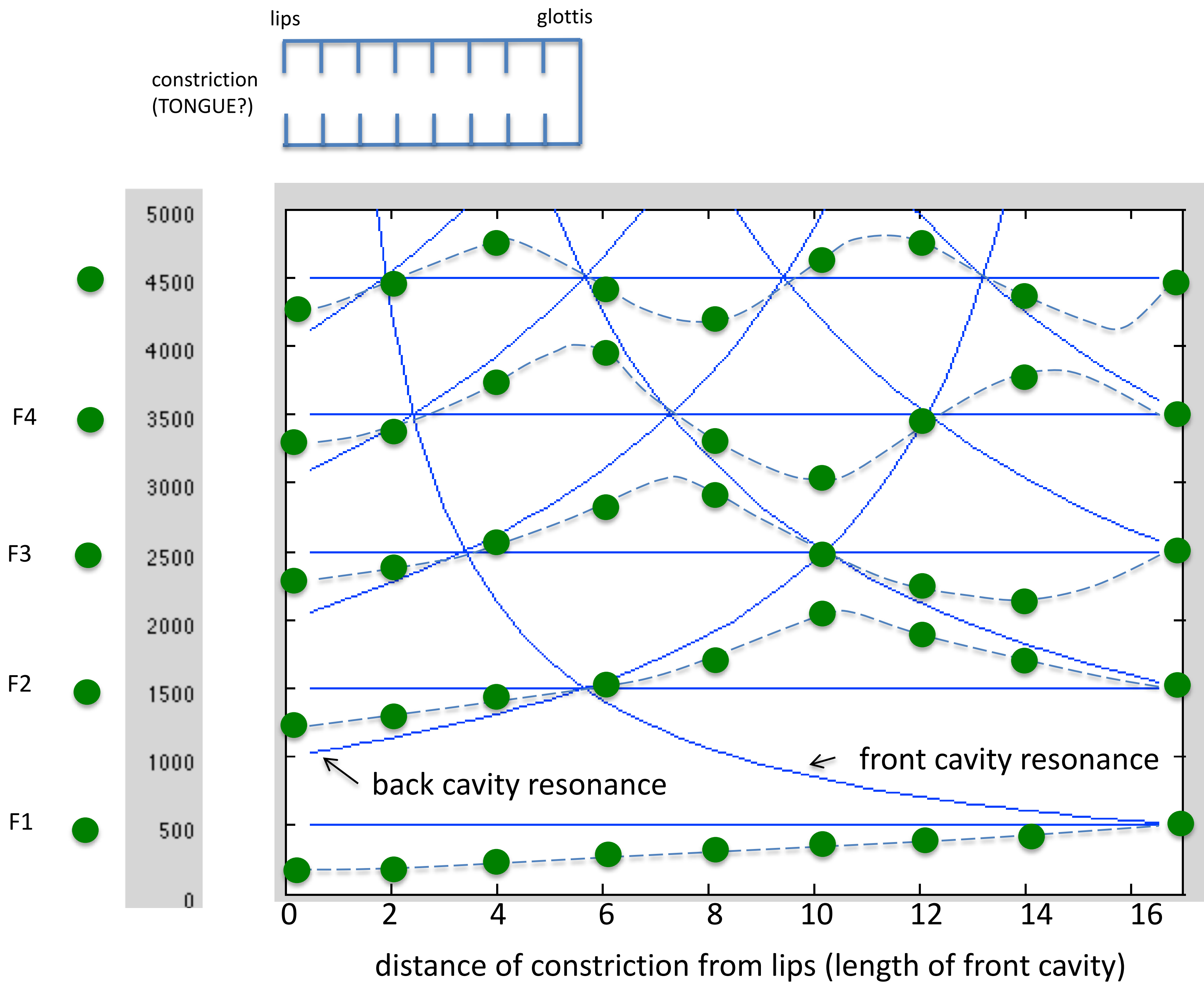


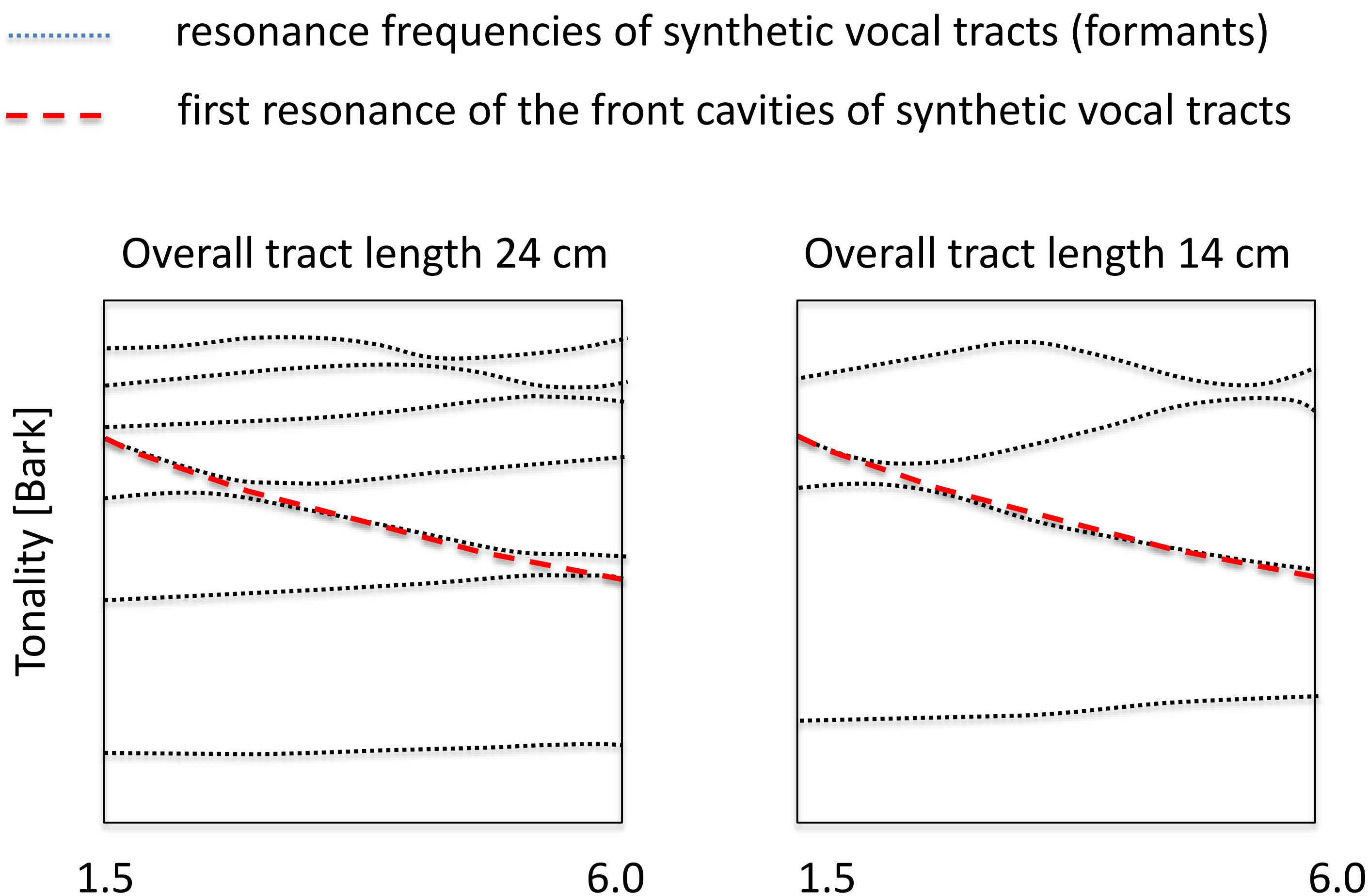


$$F_i = (2i-1)\frac{c}{4l}$$

$$i = 1, 2, 3, 4, 5, \ldots$$




lips
glottis
constriction
(TONGUE?)
F4
F3
F2
F1
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
front cavity resonance
back cavity resonance
0
2
4
6
8
10
12
14
16
distance of constriction from lips (length of front cavity)



length of the front cavity of the synthetic vocal tracts [cm]

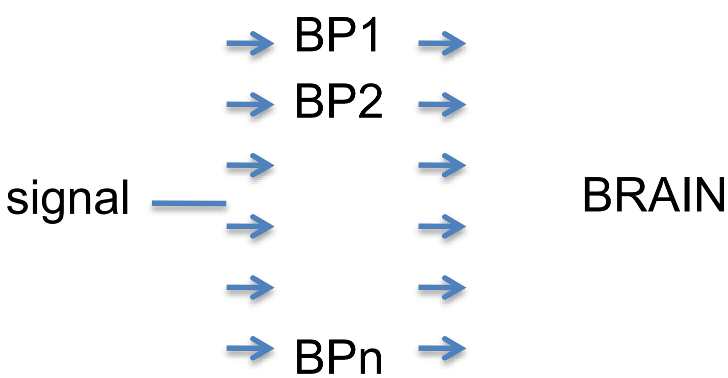
adopted from Hermansky and Broad ICASSP 1990

Hearing

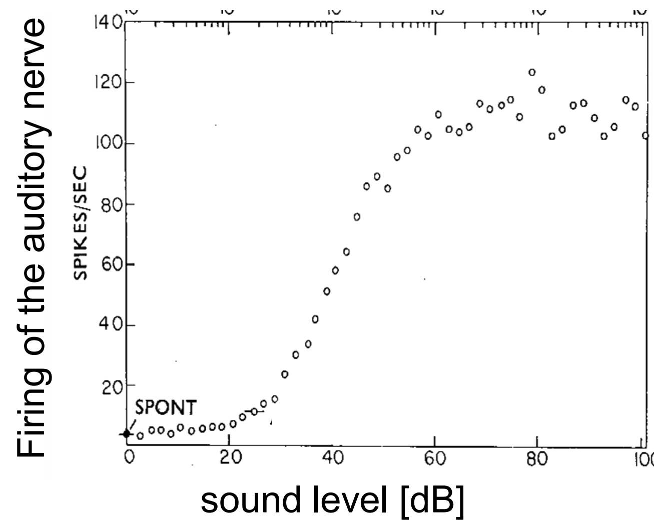
We speak **in order to be heard** in order to be understood
Roman Jakobson

Place theory of peripheral auditory processing

bank of cochlear band-pass filters



firing rate depends on sound intensity



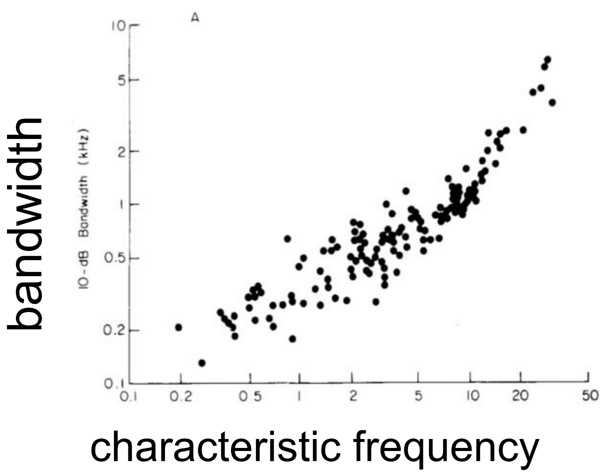


Figure of auditory processing from inner hair cells to auditory cortex removed due to copyright restrictions. Please see the video.

Brain wetware

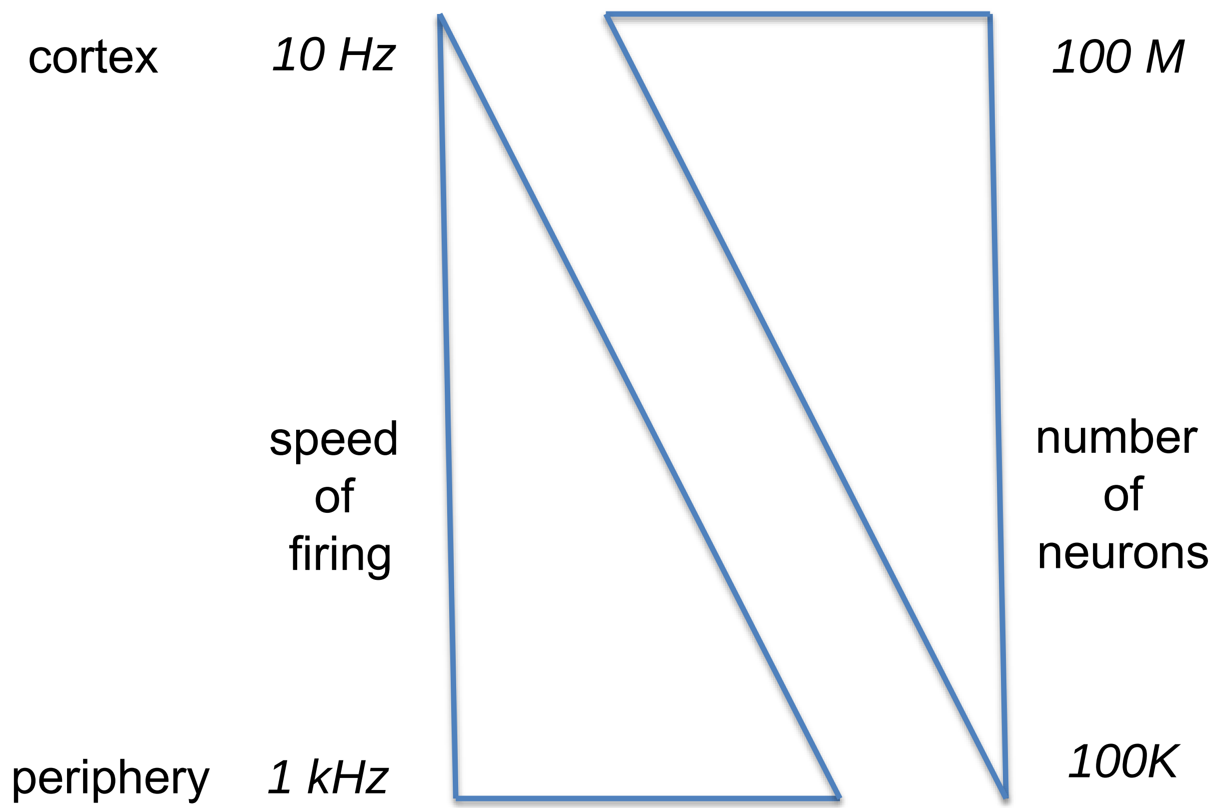
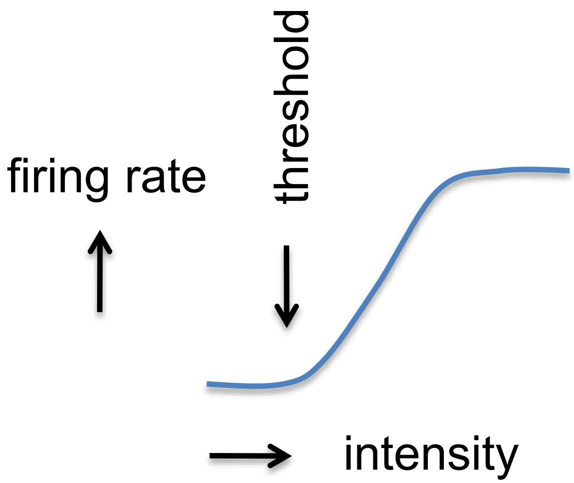


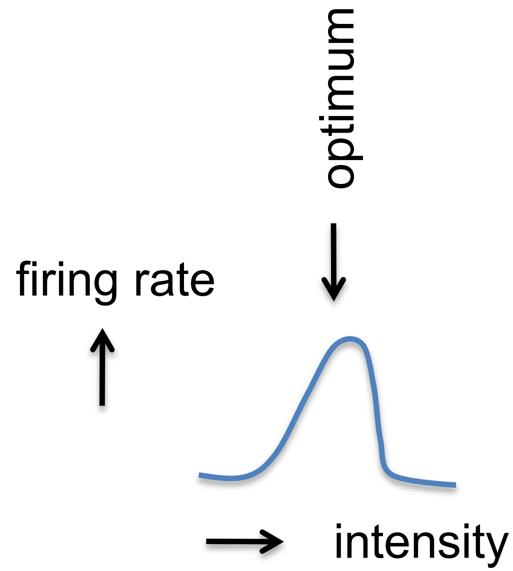
Figure of auditory processing from inner hair cells to auditory cortex removed due to copyright restrictions. Please see the video.

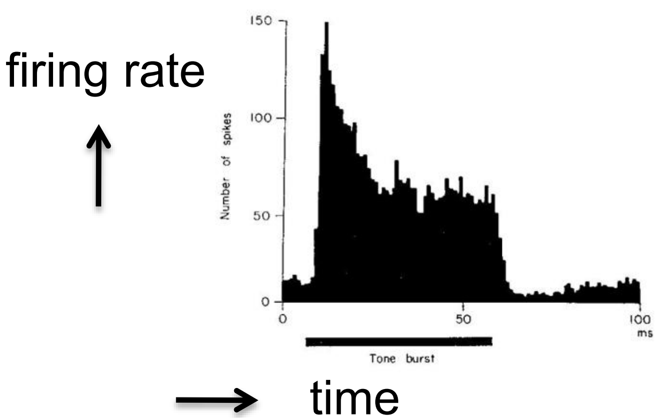
Figure of auditory processing from auditory cortex to hair cells removed due to copyright restrictions. Please see the video.

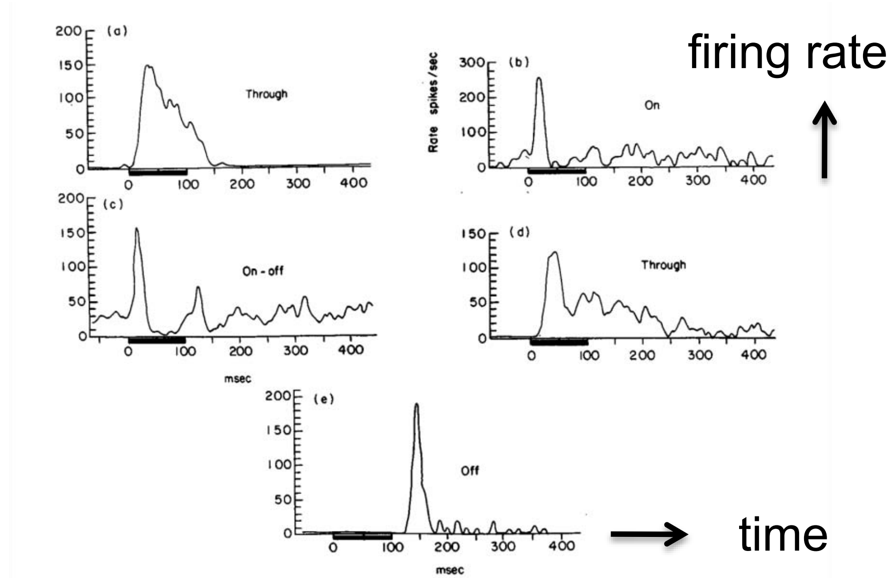
lover levels
(auditory nerve)

higher levels
(cortex)









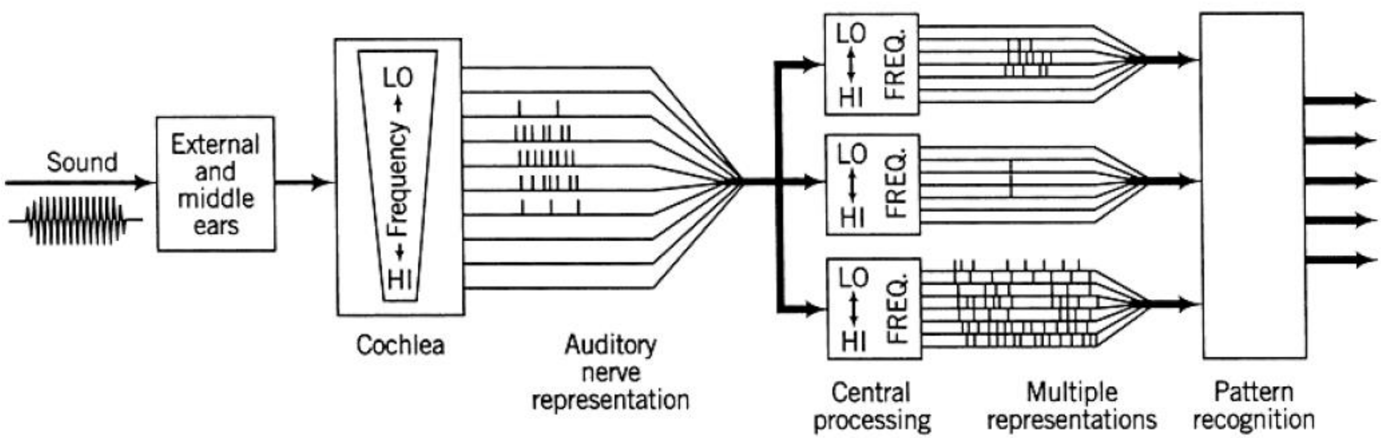
Auditory cortical spectro-temporal receptive fields

Obtained through a kind of "spike triggered averaging" (dynamic ripples as inputs)

Figure removed due to copyright restrictions. Please see the video.

Many different STRFs

Courtesy of S. Shamma UMD lab



Sachs et al 1988

Psychophysics

stimulus →

perception
(whole system)

response →

- What is the response of the whole organism to a stimulus?
- Present the stimulus and ask

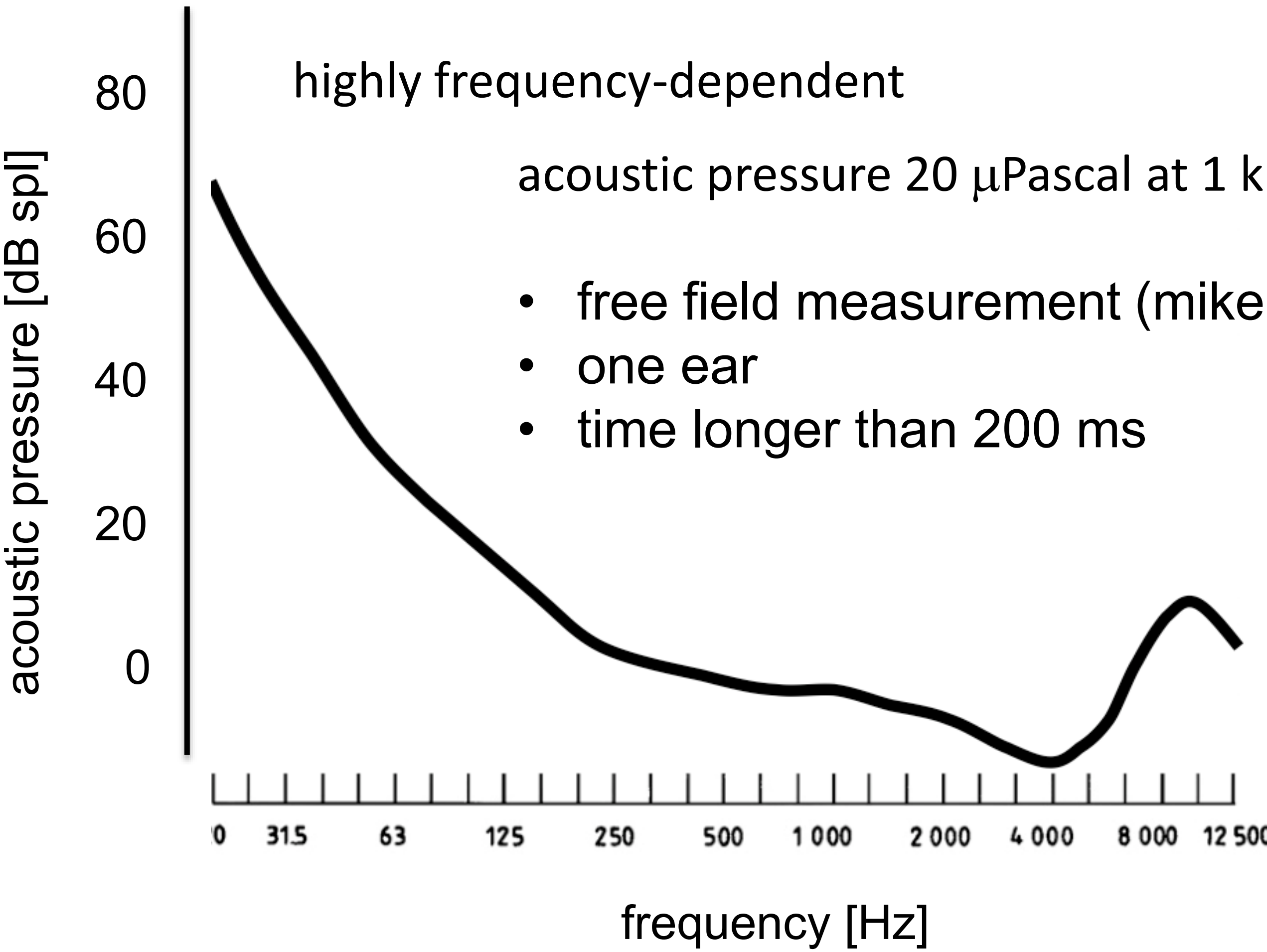
Threshold of hearing

can you hear it?

highly frequency-dependent

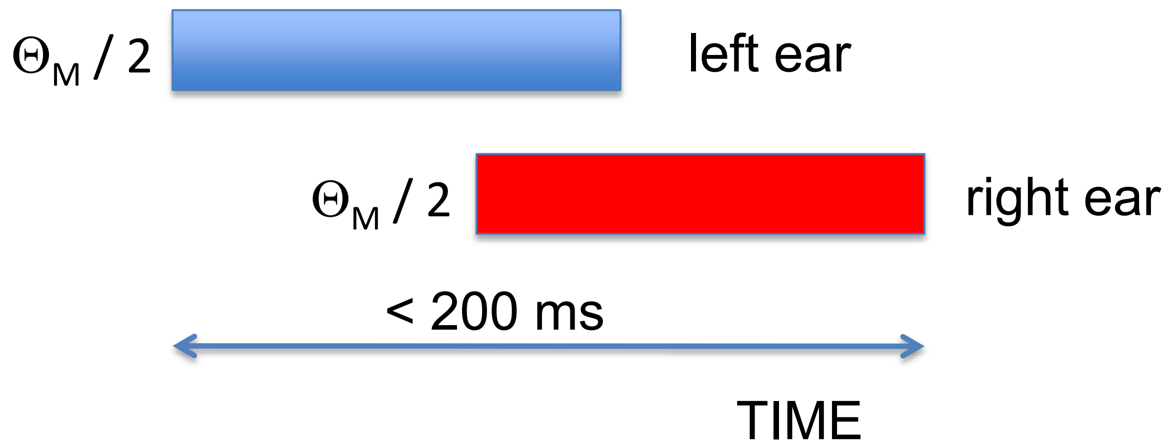
acoustic pressure 20 μPascal at 1 kHz is 0 dB spl

- free field measurement (mike in place of the ear)
- one ear
- time longer than 200 ms



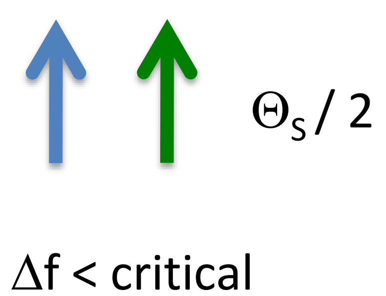
when signals are applied in in both ears, threshold for each is $\Theta_B = \Theta_M / 2$
(signals integrate)

the tones do not have to occur simultaneously as long as they are within 200 ms

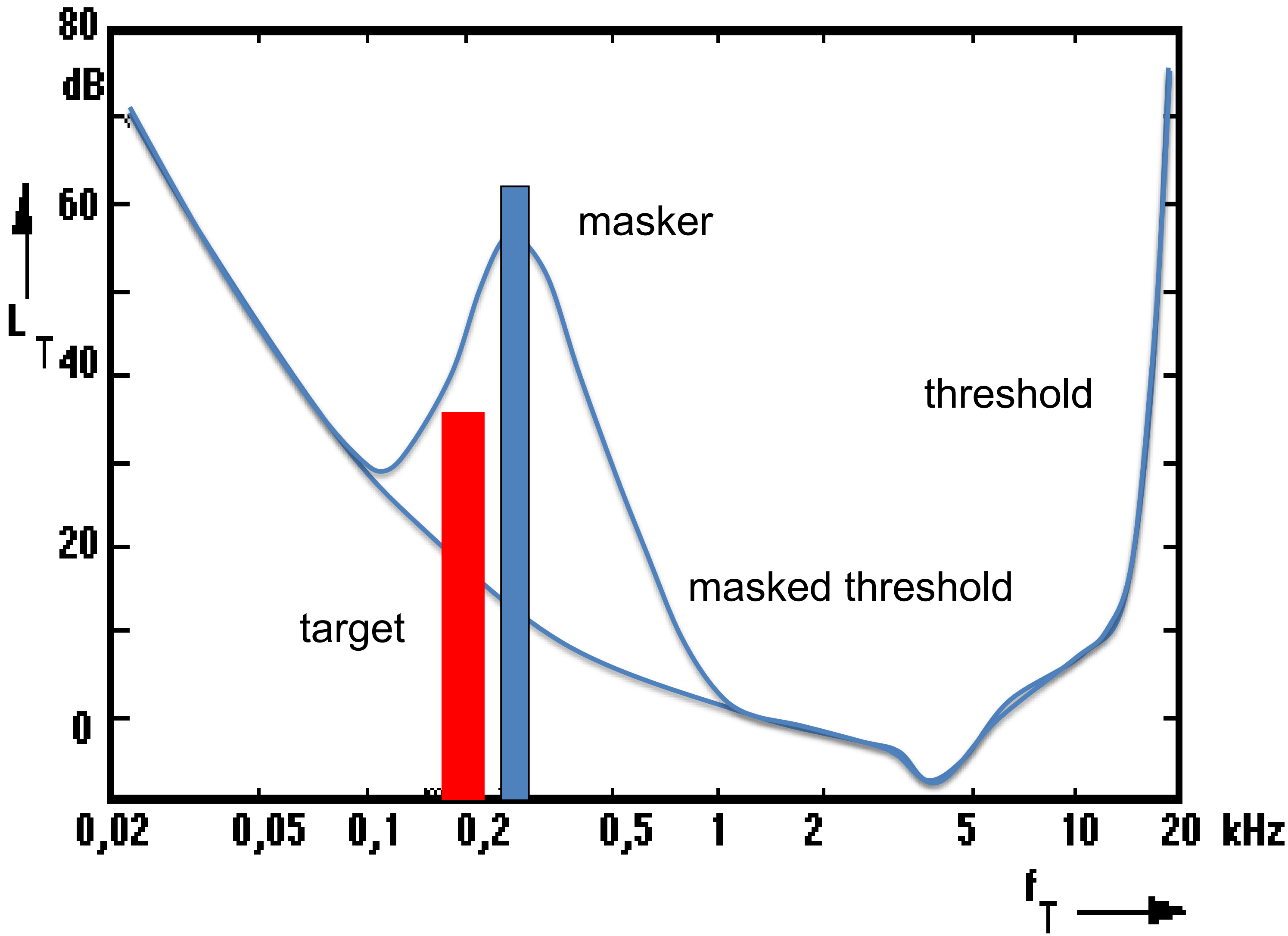


when two tones in one ear, the threshold $\Theta_D = \Theta_S / 2$,

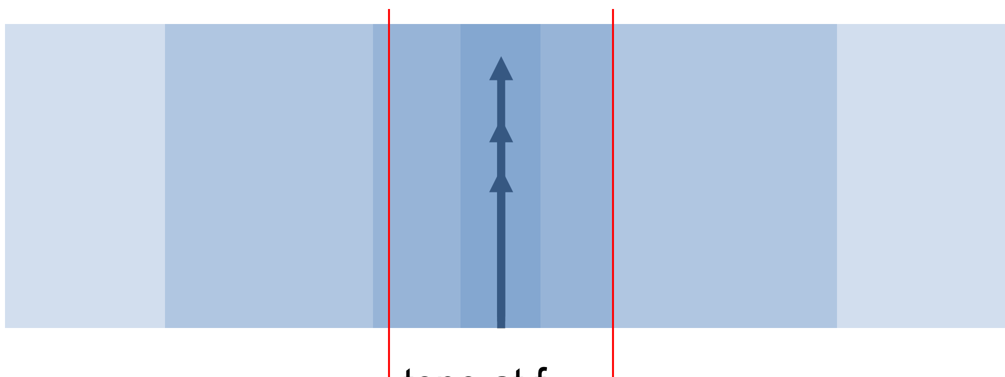
as long as the signals are "close" in frequency (within "critical band")



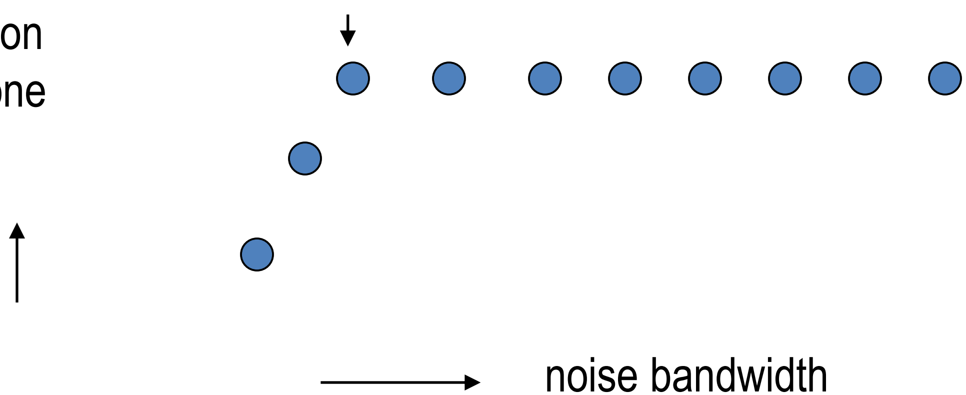
Simultaneous masking



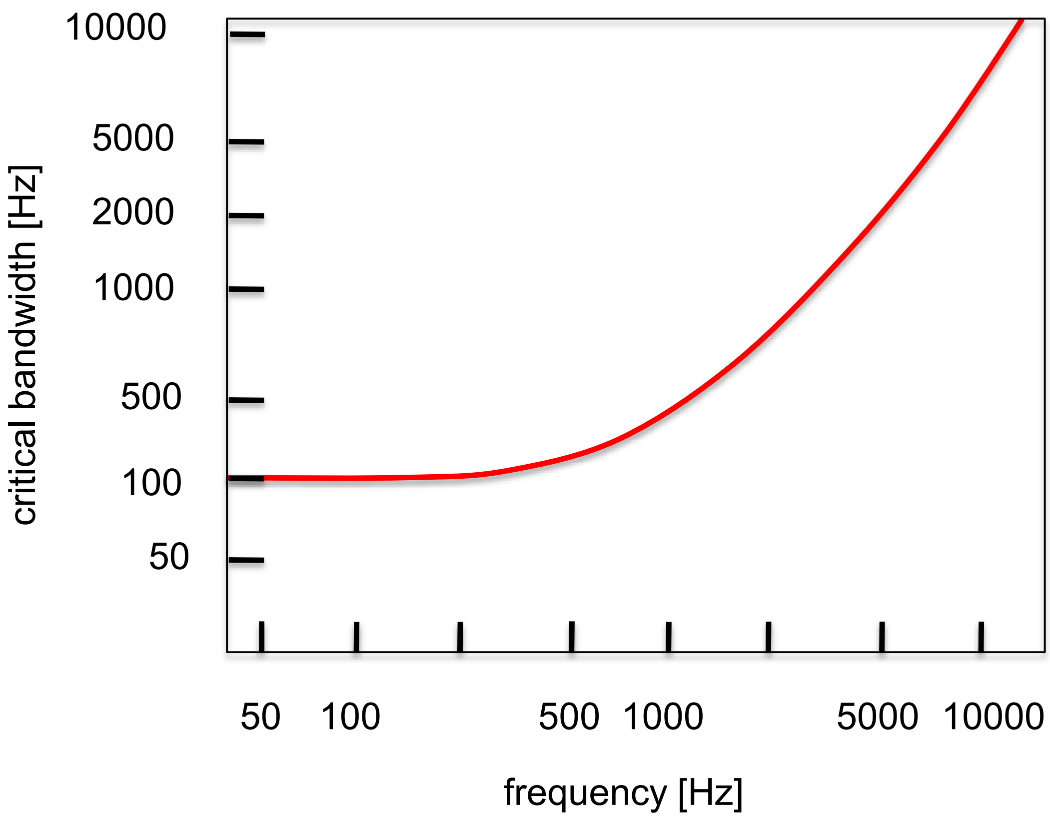
Simultaneous Masking (Fletcher 1940)



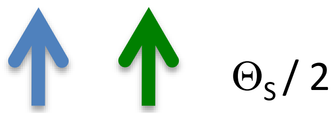
what happens outside the critical band does not affect decoding of the sound in the critical band



"critical bandwidth" again



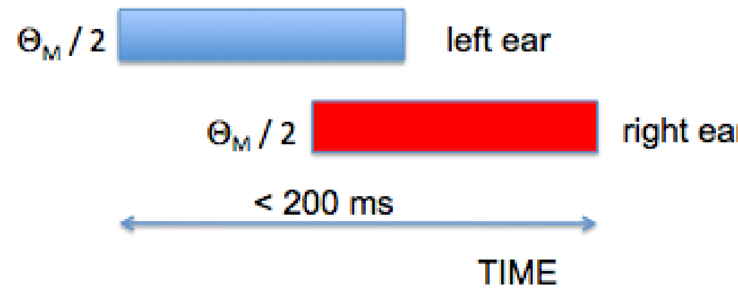
remember ?

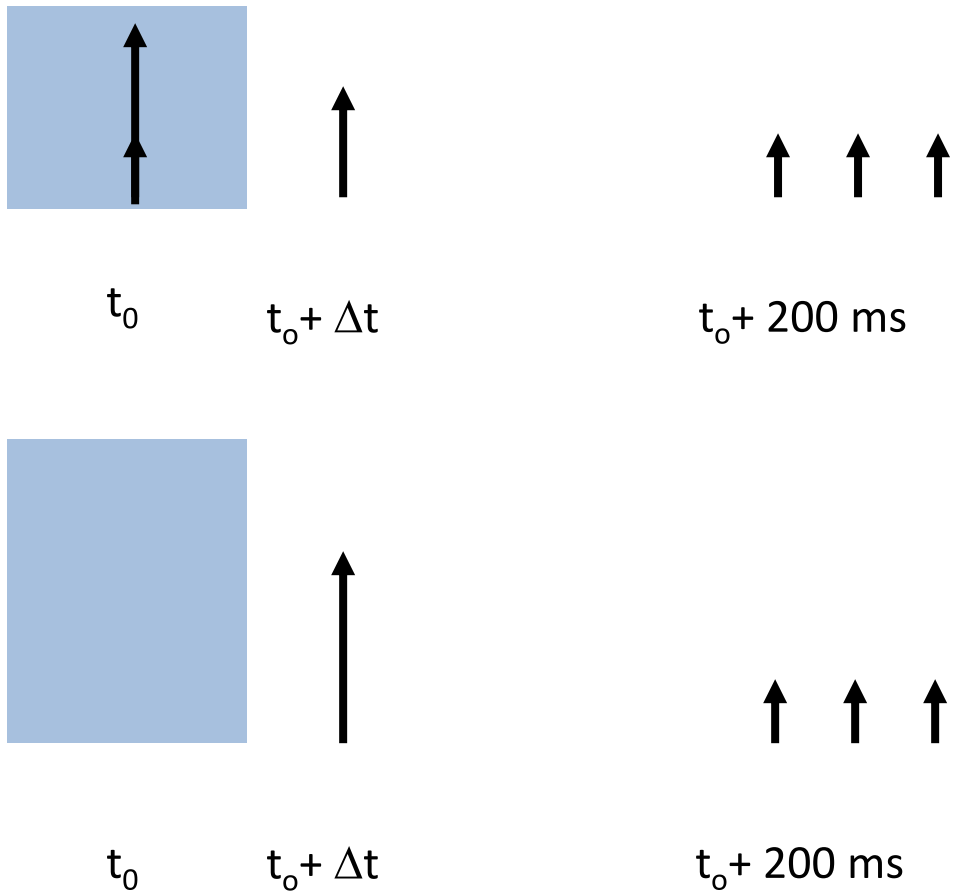


$\Delta f <$ critical

Masking in Time

remember ?





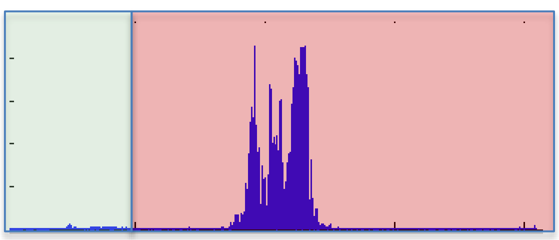
- **what happens outside the critical interval, does not affect detection of signal within the critical interval**

Loudness

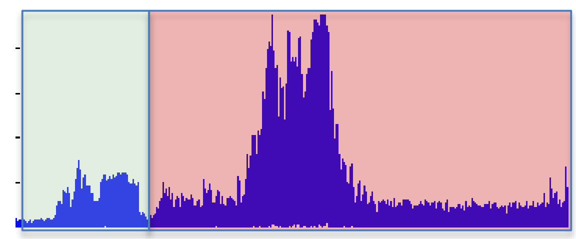
how much louder is one sound comparing to another?

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

intensity ≈ signal ^{2} [w/m^2]



loudness [Sones]

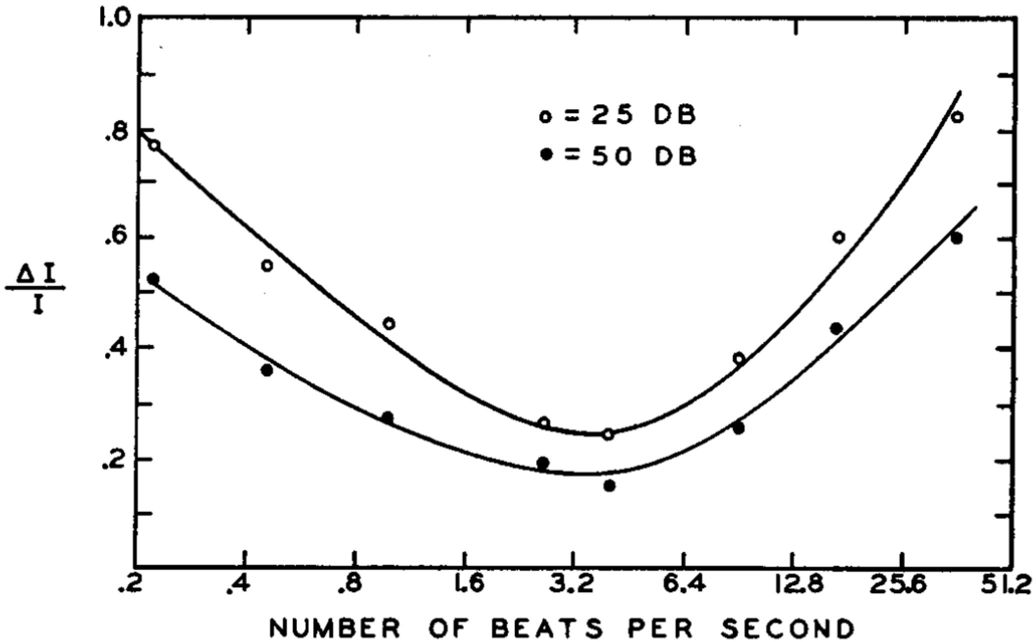


for stimuli longer than 200 ms

Equal loudness curves

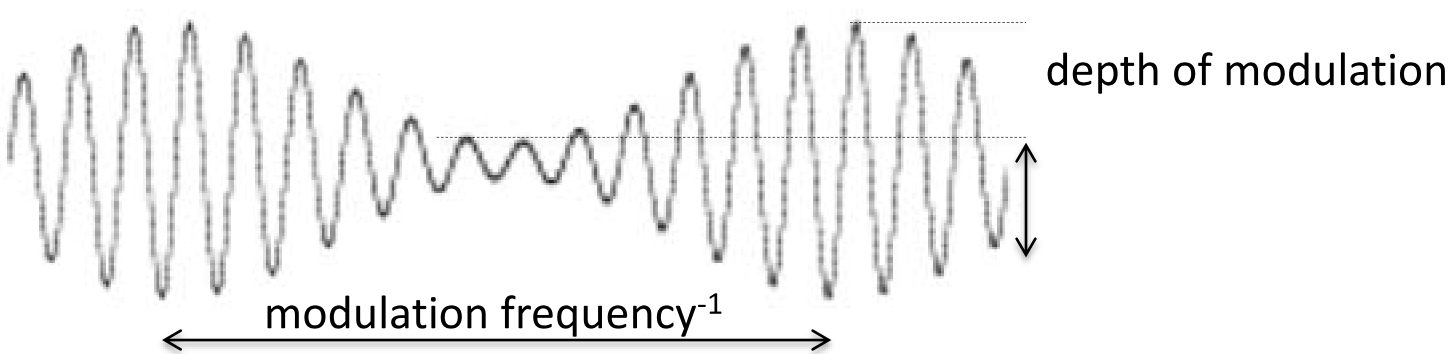
Figure of equal loudness curves removed due to copyright restrictions. Please see the video.

Perception of modulations
(Riesz 1923)



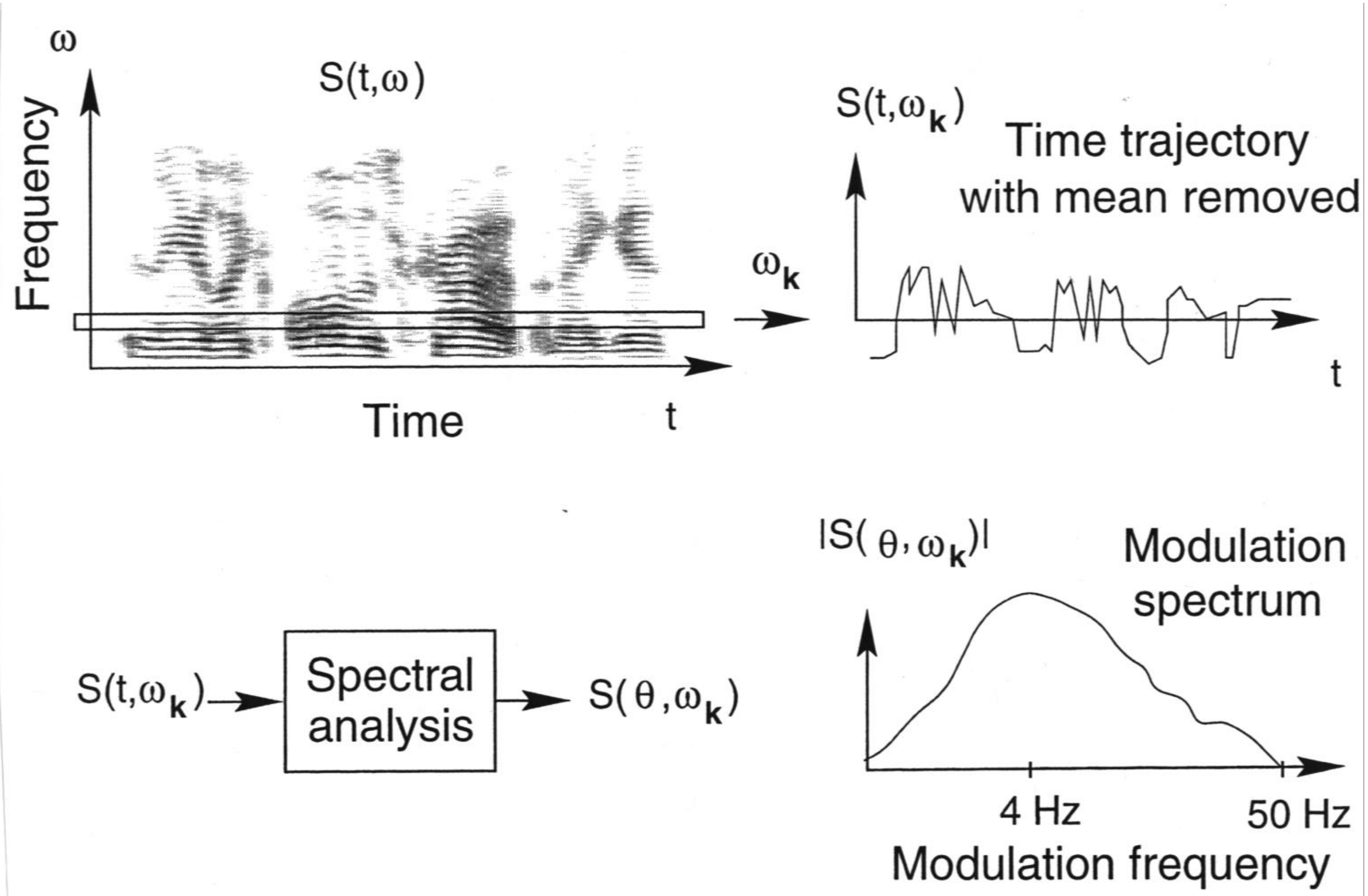
change the depth of modulation and modulation frequency

is the signal modulated or not?



Source: Riesz, R. R. "Differential intensity sensitivity of the ear for pure tones." Physical Review 31, no. 5 (1928): 867.

Modulation spectrum of speech



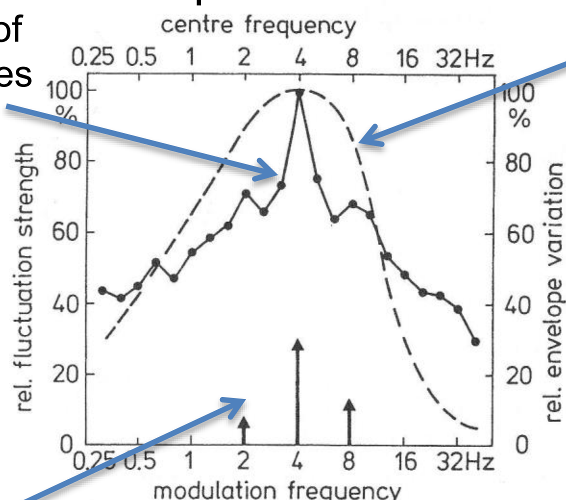
Rhythm

Perception of rhythm: tap on a Morse-code key to the rhythm of the sound

modulation frequency of music pieces

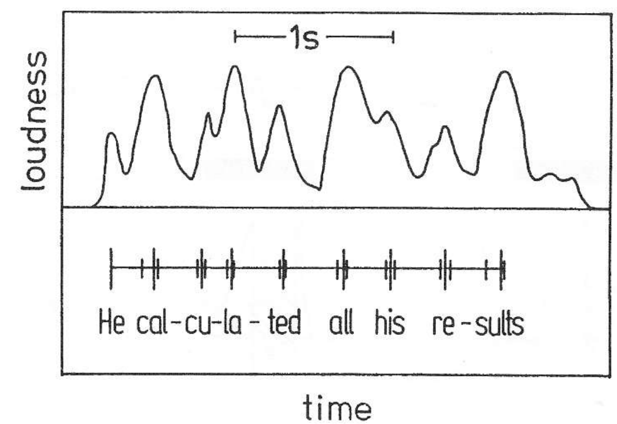
60 pieces of music

sensitivity of hearing to modulations



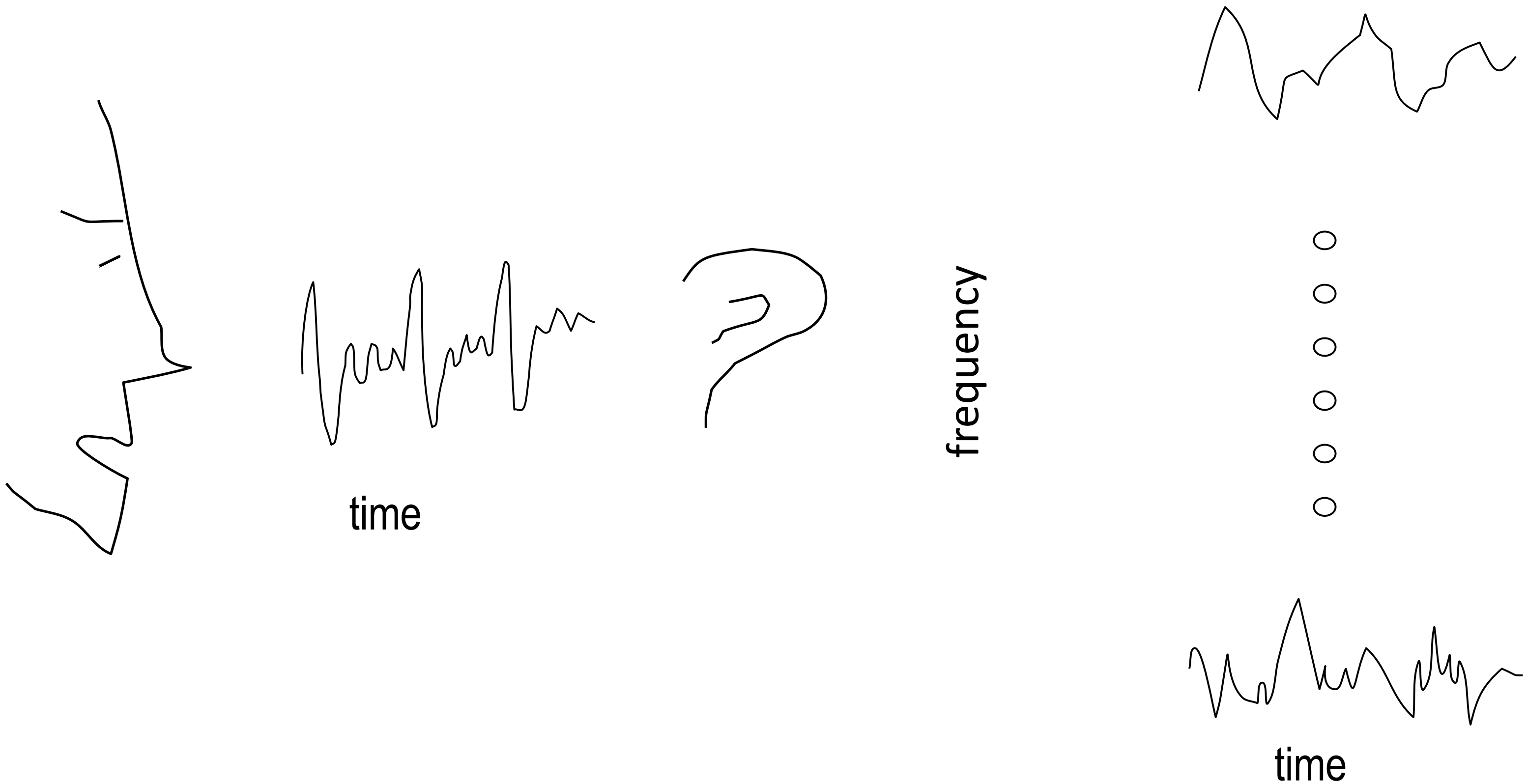
histogram of tapping frequencies

Speech sentence

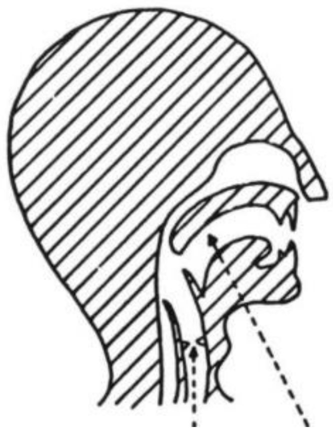


In average 4 taps per s

Where is the information ?



H. Dudley 'The **carrier nature of speech** ', Bell System Technical Journal, vol. 19 (1940)

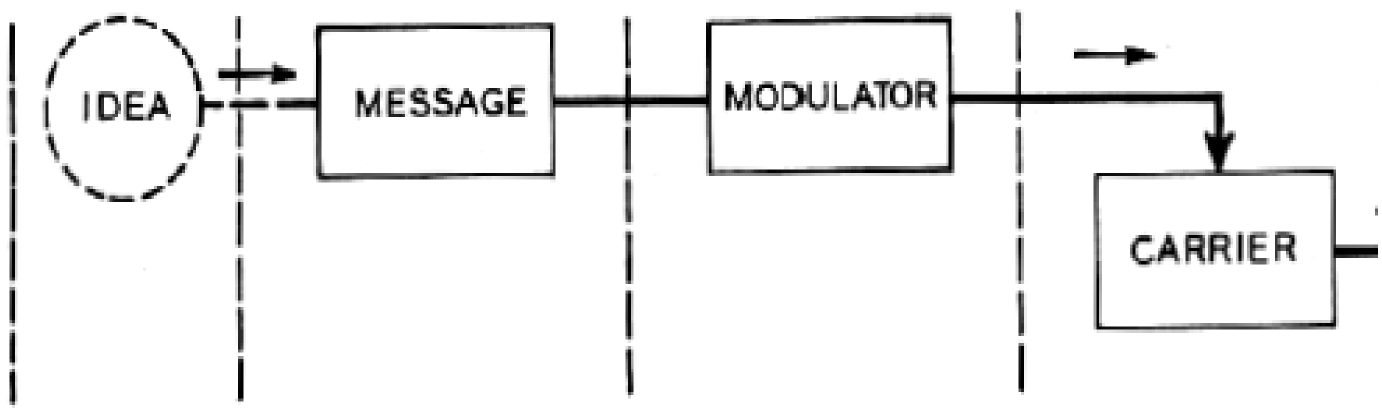


carrier message

Source: Dudley, Homer. "The carrier nature of speech." Bell System Technical Journal 19, no. 4 (1940): 495-515.

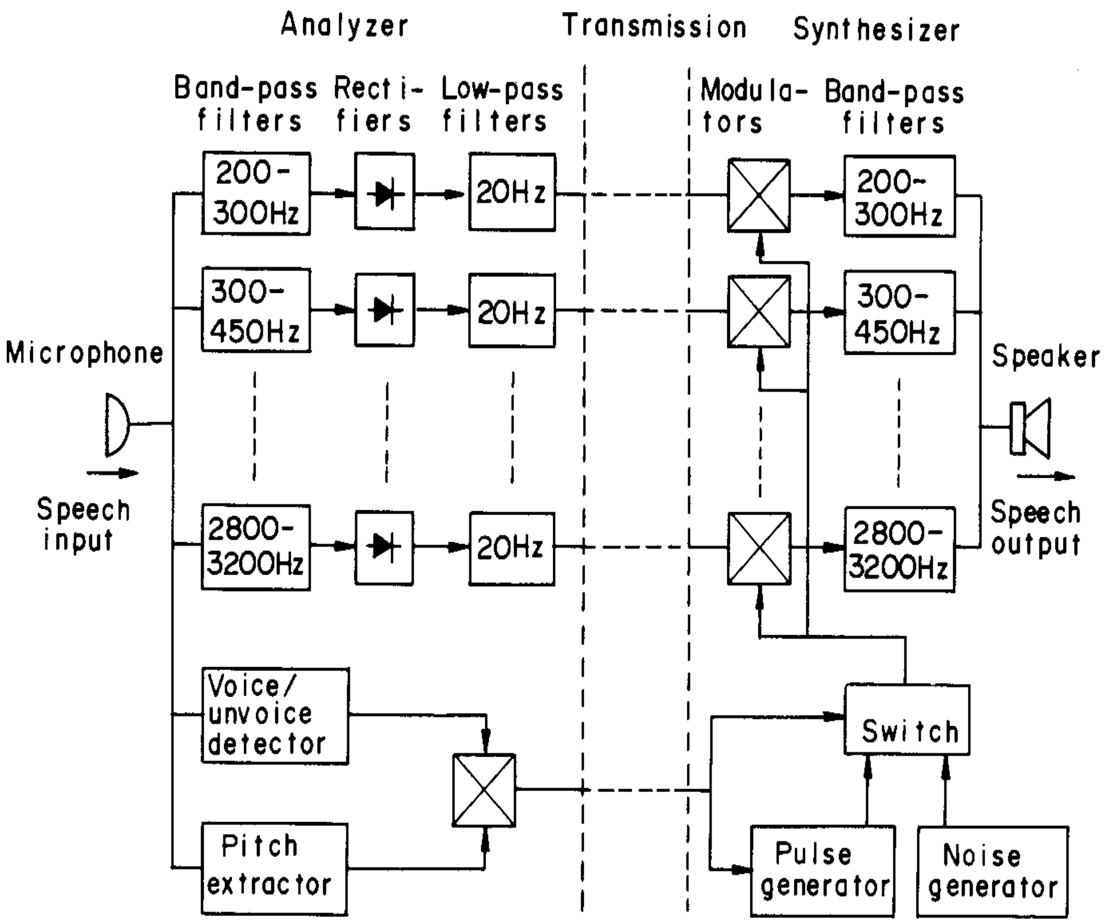
Inaudible **message** in slow motions of vocal tract is made audible by **modulating** the audible carrier

-Dudley 1940



speech

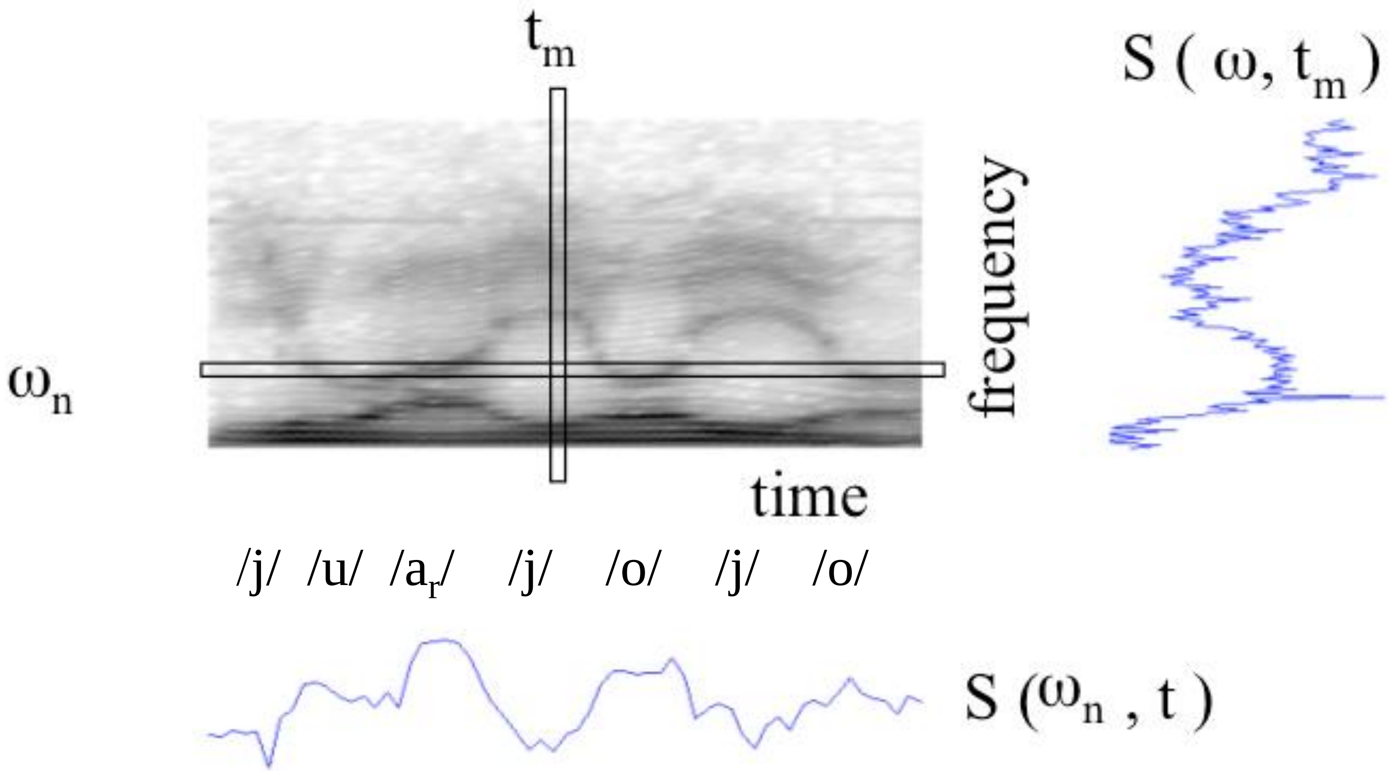
VOCODER
(H. Dudley, U.S. patent US2194298 A 1939)



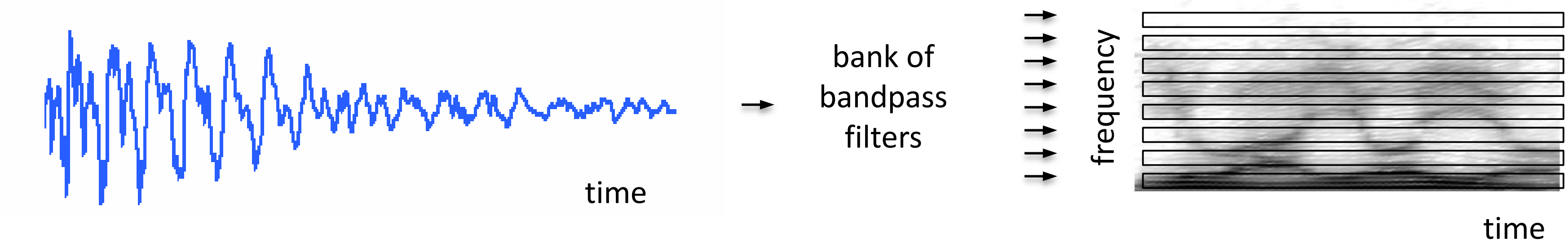
- Predictability (production)
  - speech waveform changes "slowly" (inertia of air mass in vocal tract cavities)
  - spectral envelope changes slowly
    - 20 Hz low-pass
  - voiced speech is periodic
    - pulse generator for excitation
- Hearing properties (perception)
  - spectral resolution of hearing
    - wider band-pass filters at higher frequencies

Figure removed due to copyright restrictions. Please see the video.
Source: Dudley, Homer, and Otto O. Gruenz Jr. "Visible speech translators with external phosphors." The Journal of the Acoustical Society of America 18, no. 1 (1946): 62-73.

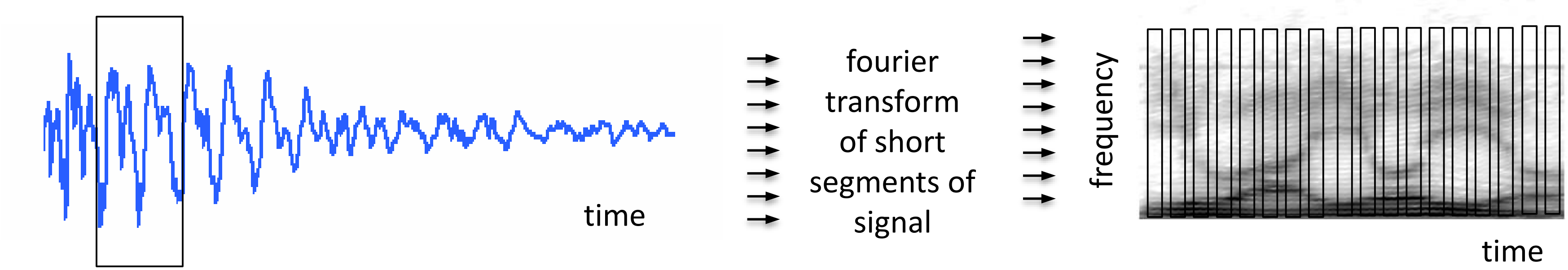
SPECTROGRAM

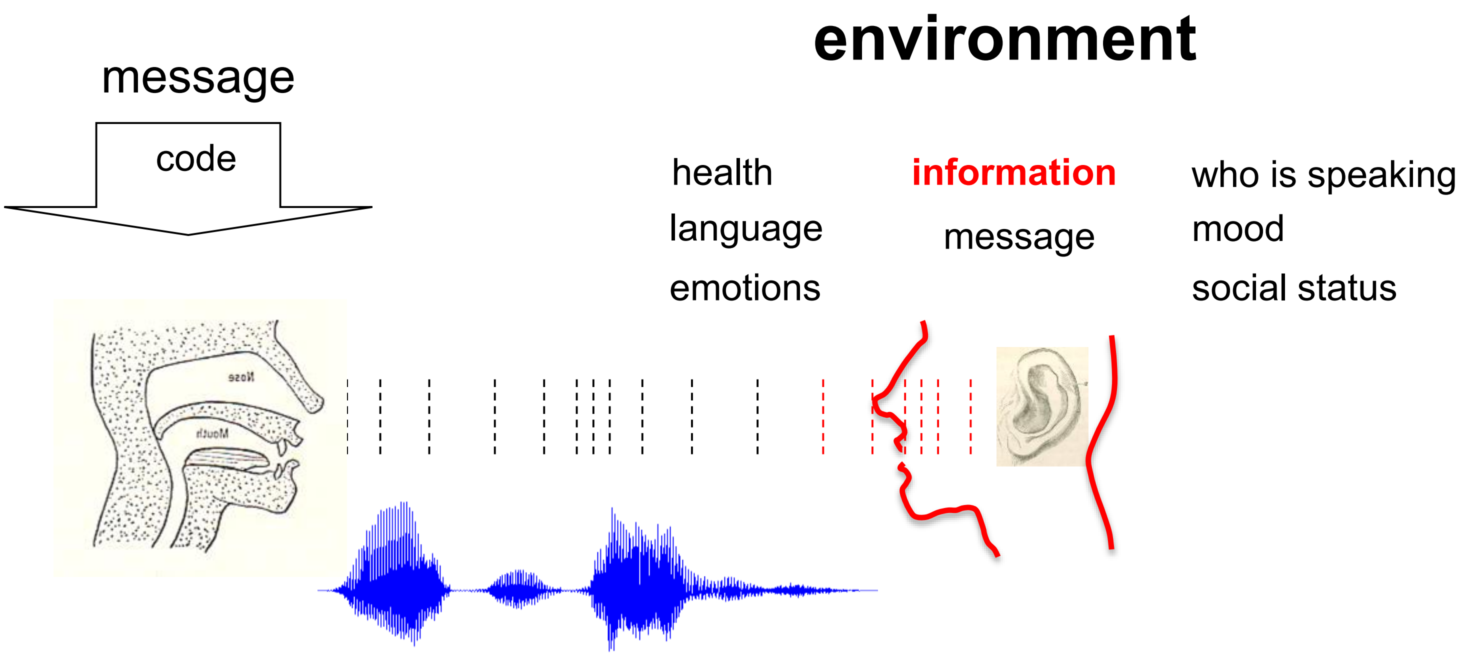


spectrogram through band-pass filtering



spectrogram through short-time fourier transform





Machine recognition of speech:
Transcribe the code which carries the message

Speech

- Produced to be perceived
  - We speak in order to be heard in order to be understood

    Roman Jakobson

- Evolved over millennia to reflect properties of human hearing

- Machine recognition of speech is a powerful way to support perceptual theory.

Better understanding of human perception through studying successful engineering solutions?

Listening for the message in speech is not the only task that human auditory perception must accomplish. Knowing what to emulate and what not when recognizing the message in speech is important. We suggest that one way to proceed is to focus on successful and well accepted ASR solutions and compare their properties with what we know about the perception of signals, and of speech in particular. Often, the engineering solution turns out to be a reflection of particular characteristics of hearing.

Hynek Hermansky, Jordan R. Cohen, and Richard M. Stern. "Perceptual properties of current speech recognition technology." *Proceedings of the IEEE* 101.9 (2013): 1968-1985.

RECOGNITION

message

health

language

noises *emotions*

who is speaking

mood

social status

e.g., text describing the message

SPEECH SIGNAL
high bit rate → RECOGNIZER → INFORMATION
low bit rate

↑

KNOWLEDGE

Knowledge

- From textbooks, teachers, intuitions, beliefs, …
  – hardwired, so no need to learn it over and over again

    but

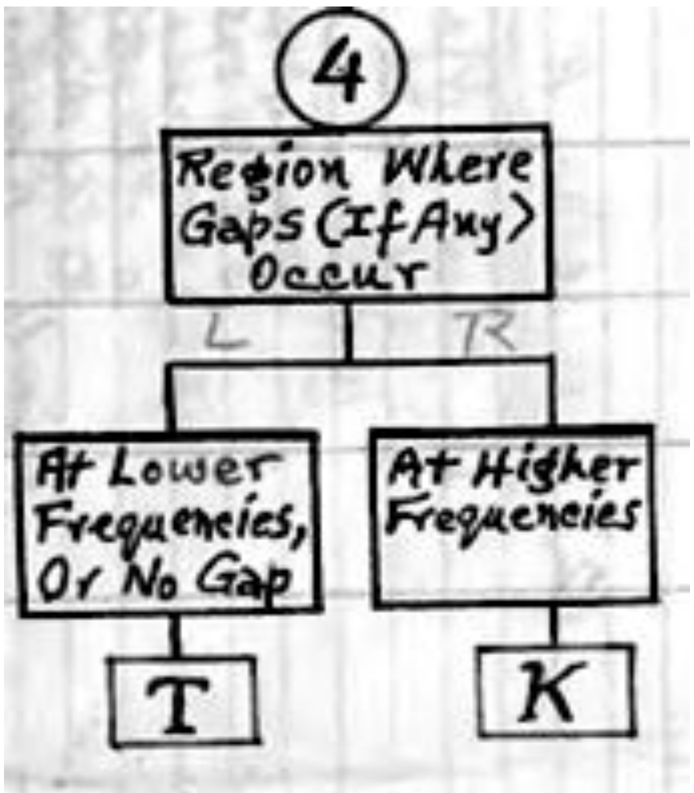
  – incomplete, irrelevant, can be wrong

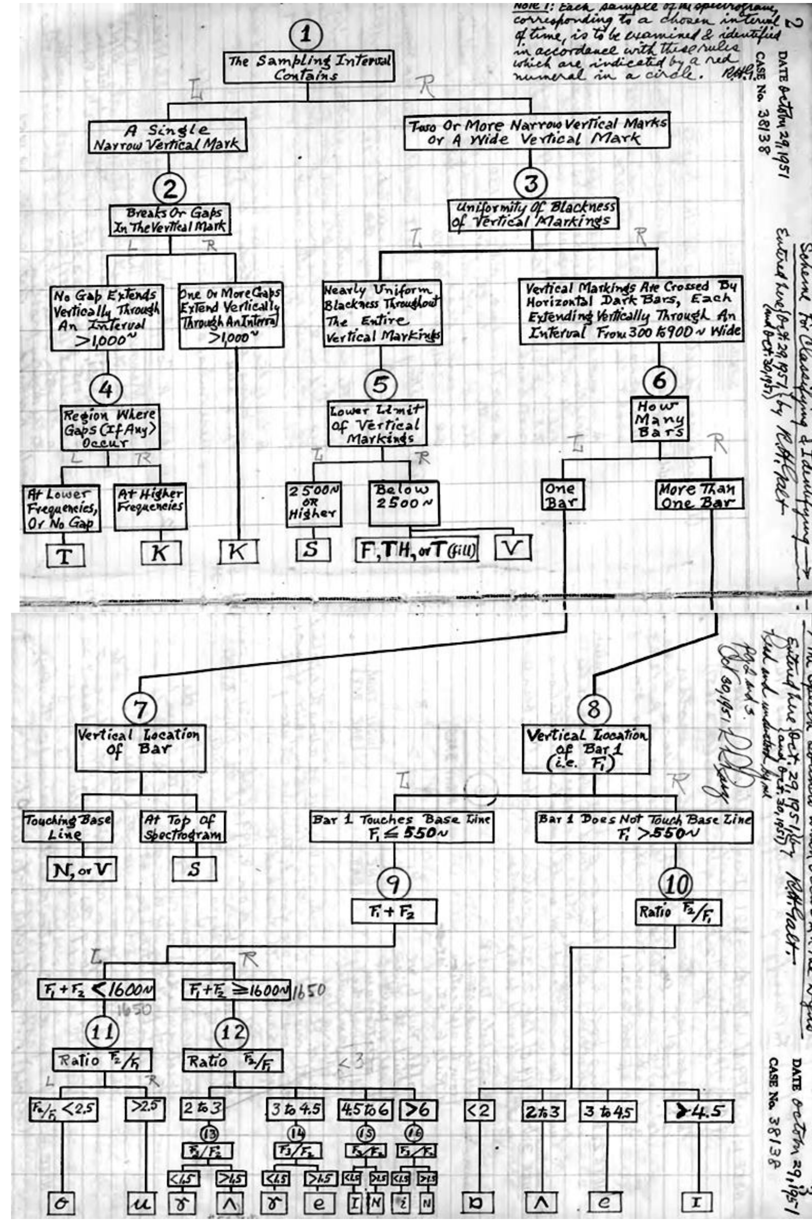
- Directly from data
  – relevant and unbiased

    but

  – large amounts of (transcribed) data may be required
  – how to get **architecture** of a machine from data ?

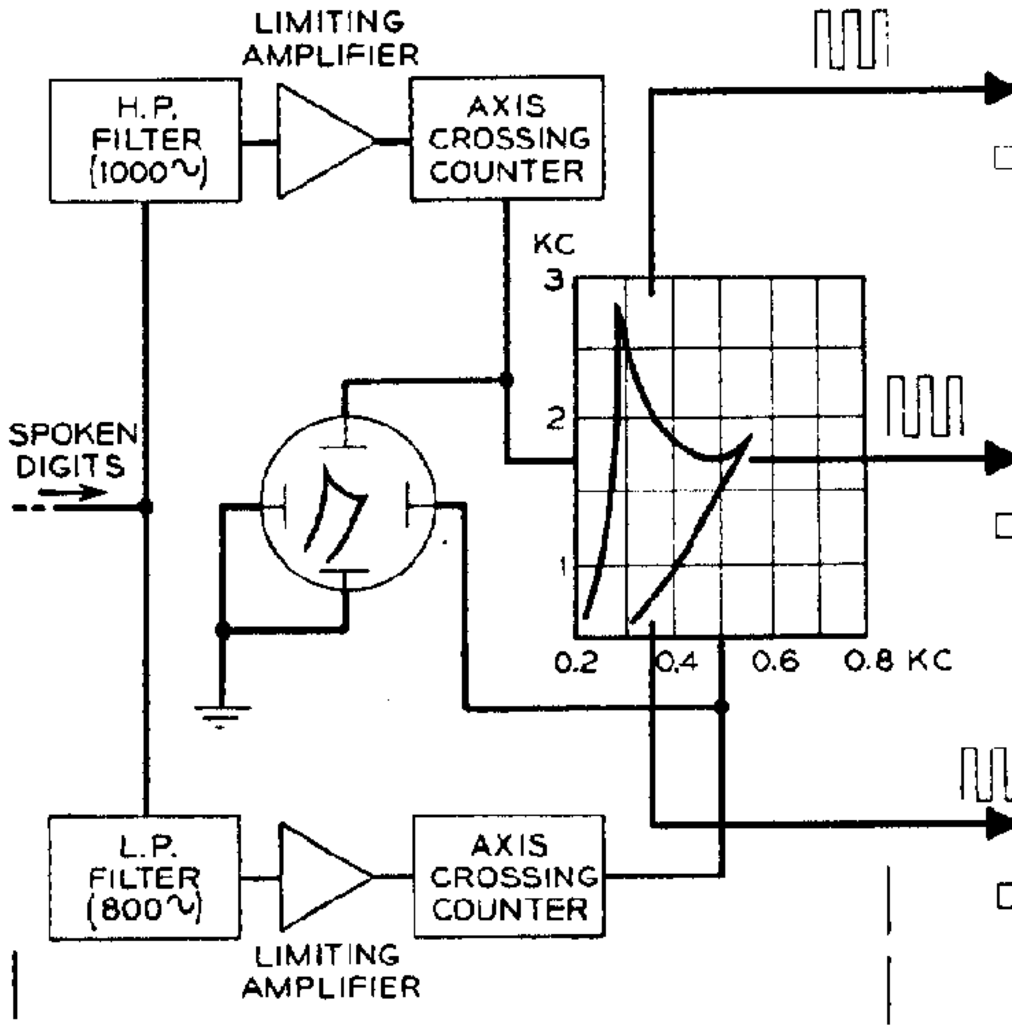
Concept of the first "real" automatic speech recognizer (R.H. Galt 1951)





First “real” recognizer ever build

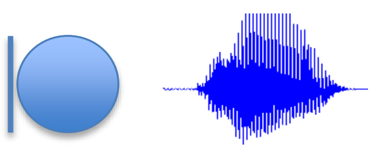
(Davis, Biddulph, Balashek 1952) Automatic Speech Recognition of Spoken Digits, J. Acoust. Soc. Am. 24(6) pp.637 - 642



Source: Davis, K. H., R. Biddulph, and Stephen Balashek. "Automatic recognition of spoken digits." The Journal of the Acoustical Society of America 24, no. 6 (1952): 637-642.

speech recognition in 21st century?

training data containing
ALL
sources of anticipated
harmful variability (noises)

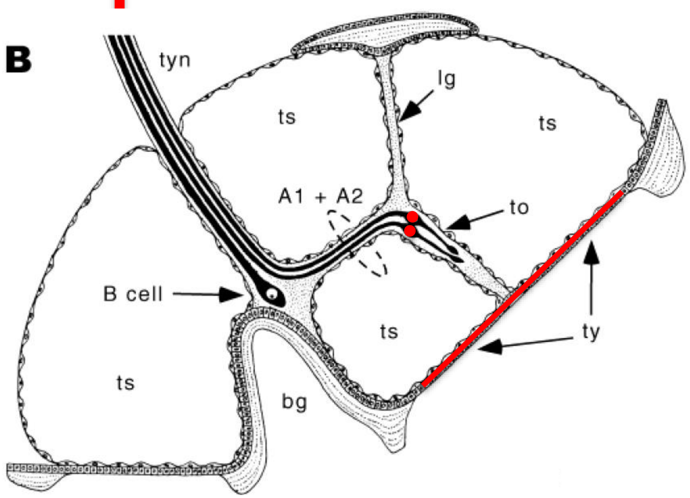


→ single flat
deep
neural net → speech message

needs to hear
a hungry bat
and to avoid it

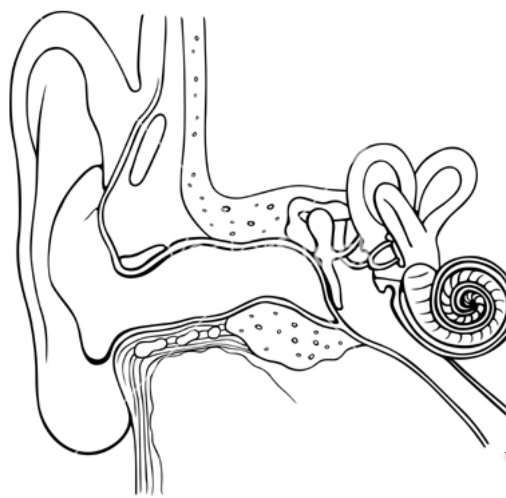
needs to
understand
speech

brain



tuned to 25-50 kHz

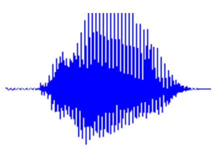
brain
medial geniculate body
inferior colliculus
lateral lemniscus
superior olivary complex
cochlear nucleus



bank of parallel
bandpass filters

training data containing
ALL
sources of anticipated
harmful variability (noises)

highly structured
deep neural net
convolutive pre-processing,
recurrent structures,
long-short-term memory,
hierarchical subsampling
(connectionist temporal classification),
e.t.c.



speech
message

A reasonable compromise ?

training data containing noise with unknown effects

speech signal → features (representation) → structured machine → wanted information

noise with known effects

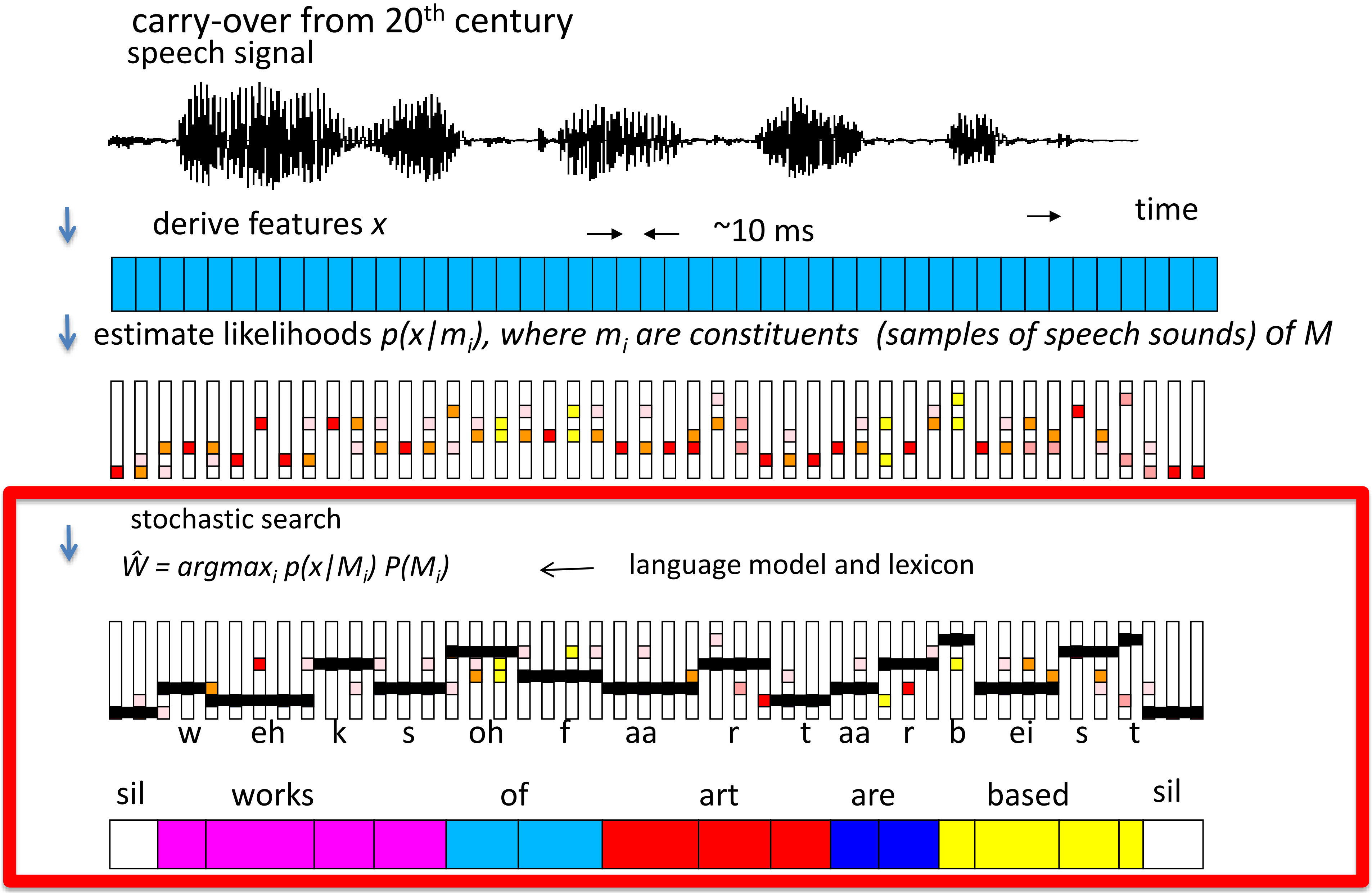
….. we suggest that the fundamental challenges in neural modeling are about representation rather than learning per se

Stuart Geman, Elie Bienenstock, and René Doursat. "Neural networks and the bias/variance dilemma." *Neural computation* 4.1 (1992): 1-58.

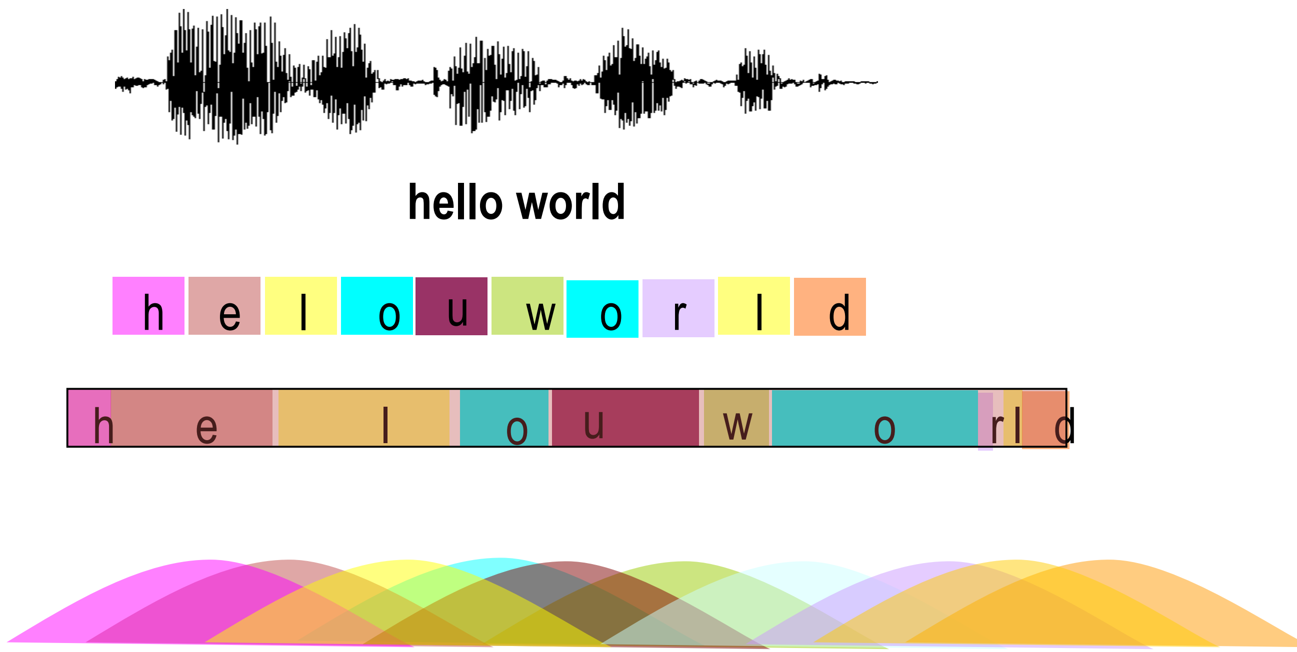
Features (representations)

- **wanted information, which is lost in this stage, is lost for recognition forever**
- **unwanted information (noise), which is kept needs to be dealt with in later stages**

Features can be also designed using development data !



Source: Hermansky, Hynek, Jordan R. Cohen, and Richard M. Stern. "Perceptual properties of current speech recognition technology." Proceedings of the IEEE 101, no. 9 (2013): 1968-1985; DOI: 10.1109/JPROC.2013.2252316.



Source: Hermansky, Hynek, Jordan R. Cohen, and Richard M. Stern. "Perceptual properties of current speech recognition technology." Proceedings of the IEEE 101, no. 9 (2013): 1968-1985; DOI: 10.1109/JPROC.2013.2252316.

coarticulation+ talker idiosyncrasies + environmental variability = **a big mess**

Two dominant sources of variability in speech

1. different people sound different, communication environment different,… (feature variability)
2. people say the same thing with different speeds (temporal variability)

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

through (modified) Bayes rule

$$w \propto \arg\max_i (p(x \mid M(w_i)) P(M(w_i))^{\gamma})$$

Model parameters from training data

How to find unknown utterance w ?

Form of the model $M(w_i)$?

What is the data x ?

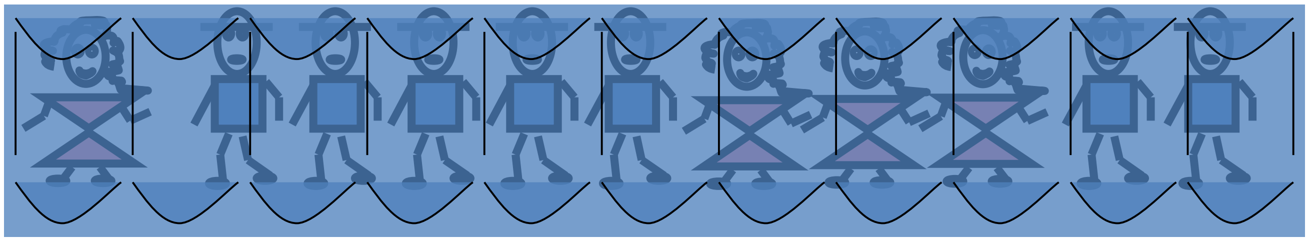
“Doubly stochastic” process (Hidden Markov Model)

Speech as a sequence of hidden states (speech sounds) - recover the sequence

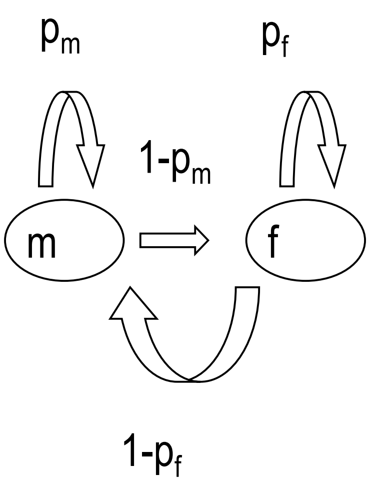
1. never know for sure which data will be generated from a given state
2. never know for sure in which state we are in

f_0= 195 125 140 120 185 130 145 190 245 155 130 Hz

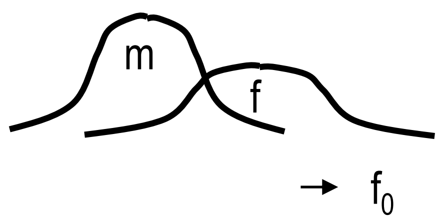
hi hi hi hi hi hi hi hi hi hi hi



know



P(sound|gender)



These parameters are typically learned from training data.

p_{1m} - probability of the first group being male group

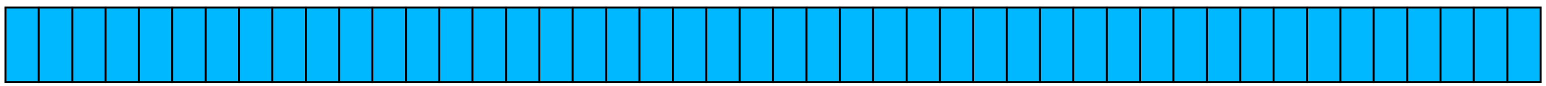
p_n – probability of group having n subgroups

Want to know

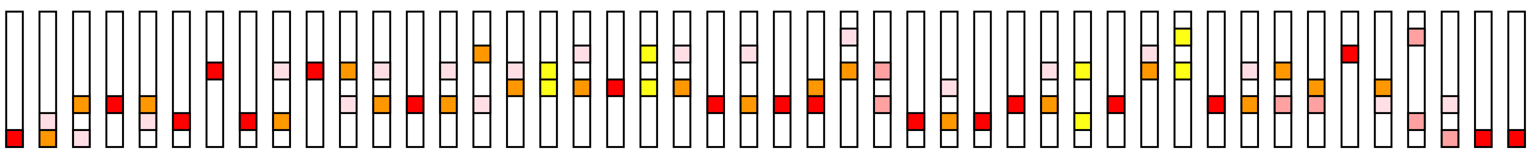
where are the boys (or girls) ?

speech signal

derive features x → ← ~10 ms → time



estimate likelihoods $p(x|m_i)$, *where* m_i *are constituents (samples of speech sounds) of M*



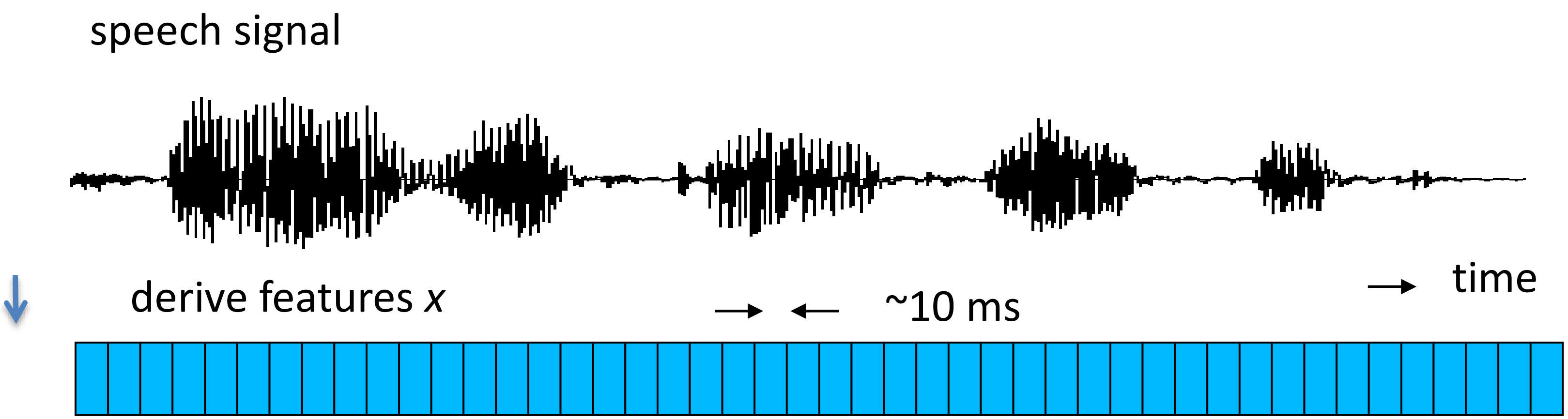
connectionist temporal classifier

| w | eh | k | s | oh | f | aa | r | t | aa | r | b | ei | s | t |

works of art are based

A.Graves et al, Proc. ICML 2006

Features (representations)



Source: Hermansky, Hynek, Jordan R. Cohen, and Richard M. Stern. "Perceptual properties of current speech recognition technology." Proceedings of the IEEE 101, no. 9 (2013): 1968-1985; DOI: 10.1109/JPROC.2013.2252316.

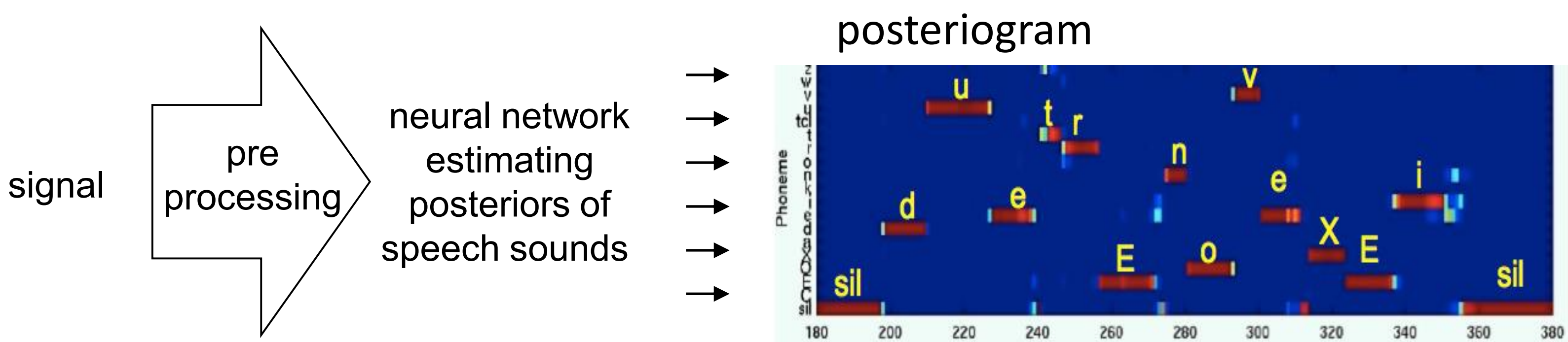
Features (representations)

- **wanted information, which is lost in this stage, is lost for recognition forever**
- **unwanted information (noise), which is kept needs to be dealt with in later stages**

1. One of important tasks of perception is to focus on relevant information (eliminating the irrelevant)
2. Feature extraction may benefit from emulations of relevant properties of hearing
3. Features can be also designed using development data (current trend)
    - what emerges, is very likely relevant to speech perecption

Artificial Neural Nets

Most efficient (smallest) set of features are posterior probabilities of classes



Classes – speech sounds:

- context independent phonemes
- context dependent phonemes
- parts of context dependent phonemes

a) Convert (divide by training priors) posterior probabilities to likelihoods for Viterbi search for the best word sequence

Bourlard and Morgan, NIPS 1990

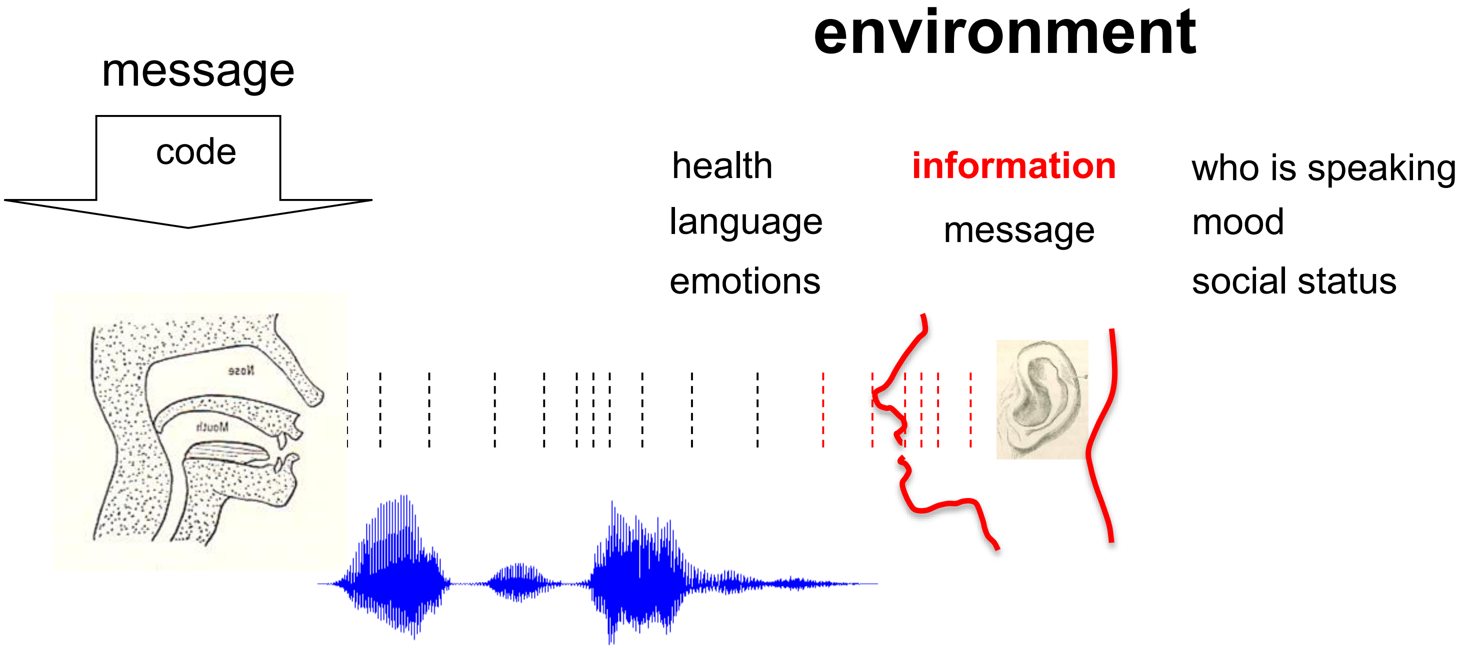
b) bottleneck (TANDEM)

output from inside of neural network estimating posteriors of speech sounds → optional additional processing (PCA, LDA) → features for HMM/GMM

Fontaine, Ris and Boite, Eurospeech 1997

Hermansky, Ellis and Sharma, ICASSP 2000

Grezl, Karafiat, Kontar, Cernocky, ICASSP 2007



signal = message (wanted information)

noise = everything else (unwanted information)

Not all noises are created equal

- expected and effects are partially understood

  e.g. linear distortions

- expected but effects are not well understood

  e.g. various environmental noises

- **unexpected**

  e.g. unexpected distortions - the real problem

training data containing noise with unknown effects

↓

speech signal → features → machine → wanted information

↓

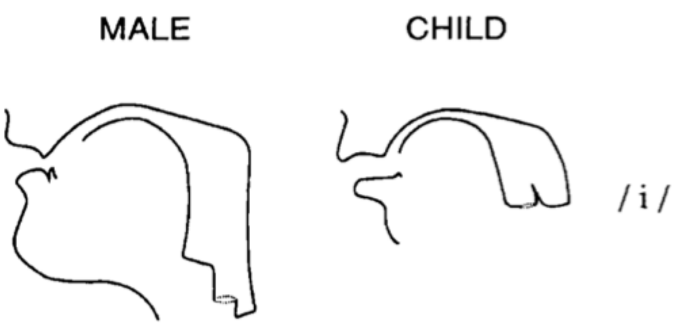
noise with known effects

Noise with known effects

Harmful information about speaker (speaker variability)

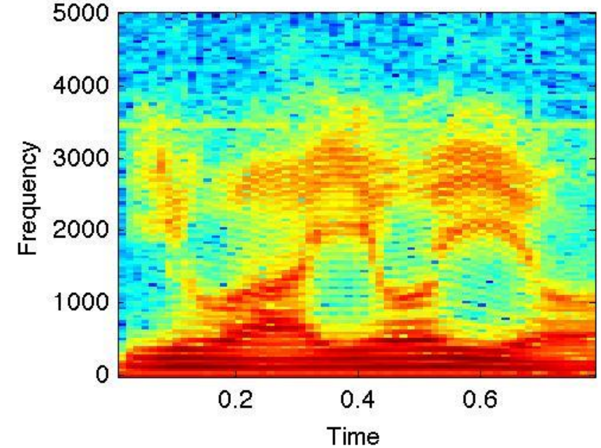
the same message

- different vocal tracts
- different speech signals

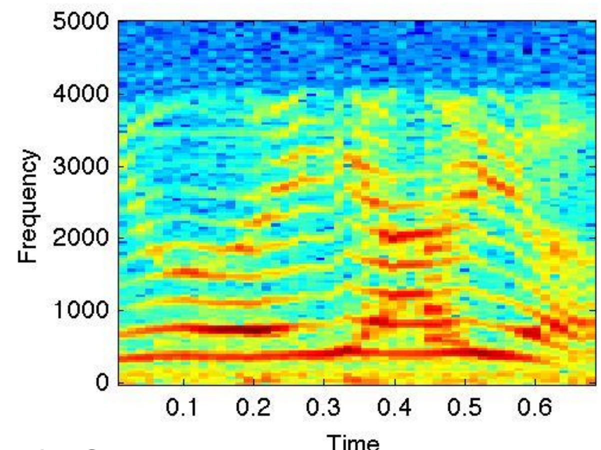


adult male

short-term spectrum



4 year old child



Source: Hermansky, Hynek, Jordan R. Cohen, and Richard M. Stern. "Perceptual properties of current speech recognition technology." Proceedings of the IEEE 101, no. 9 (2013): 1968-1985; DOI: 10.1109/JPROC.2013.2252316.

Perceptual Linear Prediction

Limited spectral resolution

formant clusters as may be interpreted by auditory perception

Perceptual Linear Prediction (PLP)

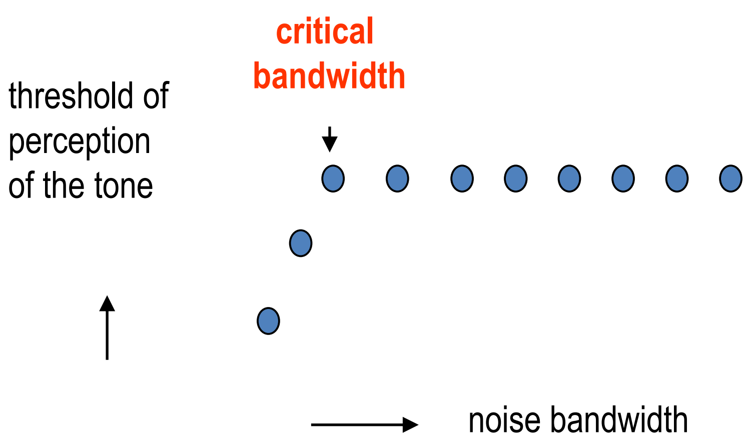
critical-band (Bark) spectral analysis
loudness domain (cubic root of intensity)
equal loudness curve (at 40 dB)
autoregressive spectral fit (fits well at peaks)

Equal loudness curves

Figure of equal loudness curves removed due to copyright restrictions. Please see the video.

Spectral resolution of hearing

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



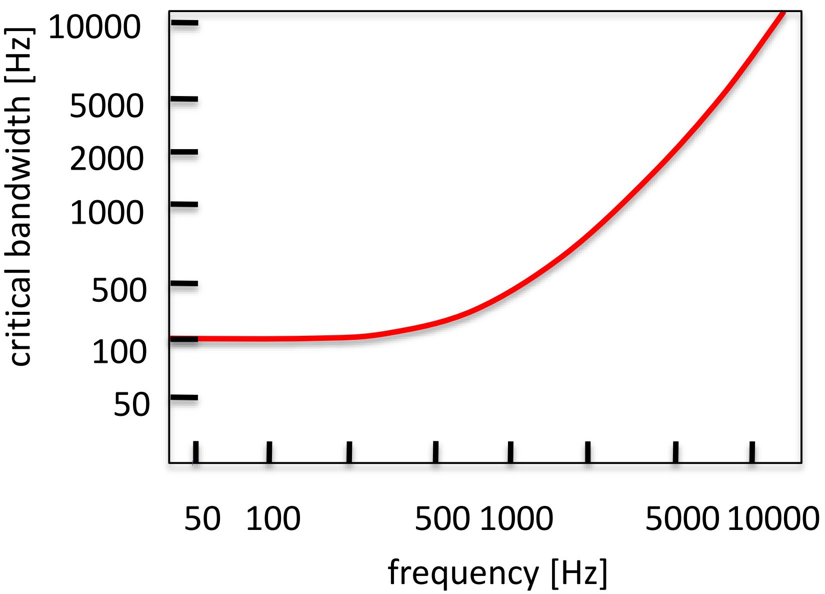
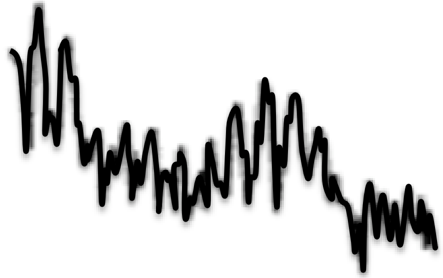
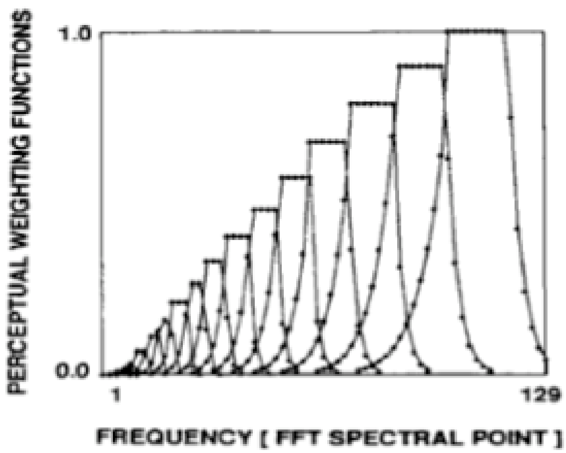


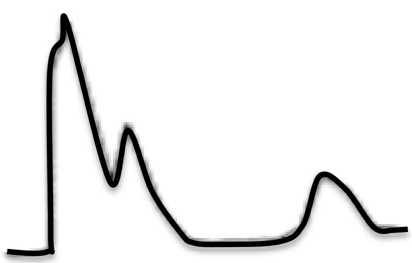
Figure removed due to copyright restrictions. Please see the video.



spectrum



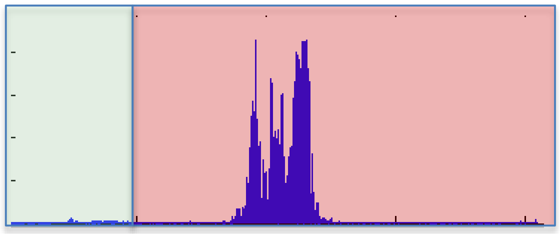
summation windows



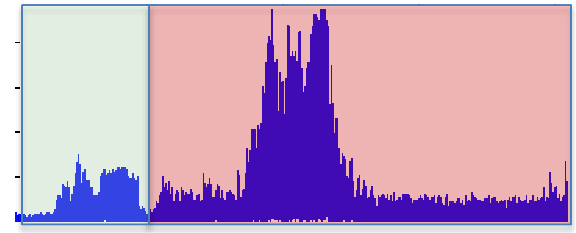
spectrum with auditory-like resolution

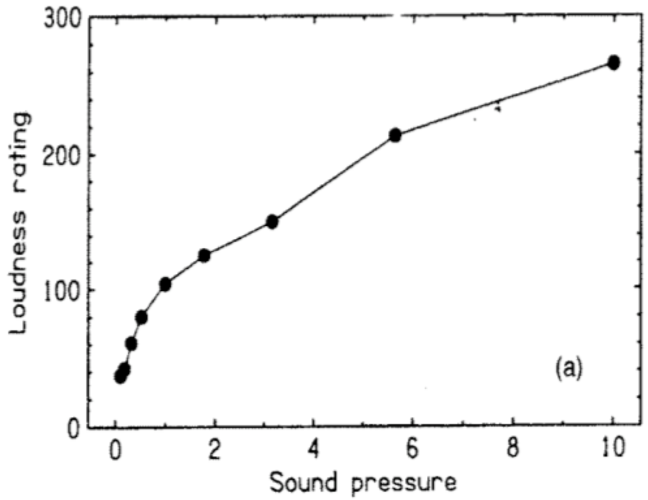
Source: Hermansky, Hynek. "Perceptual linear predictive (PLP) analysis of speech." The Journal of the Acoustical Society of America 87, no. 4 (1990): 1738-1752.

intensity ≈ signal ^{2} [w/m^2]

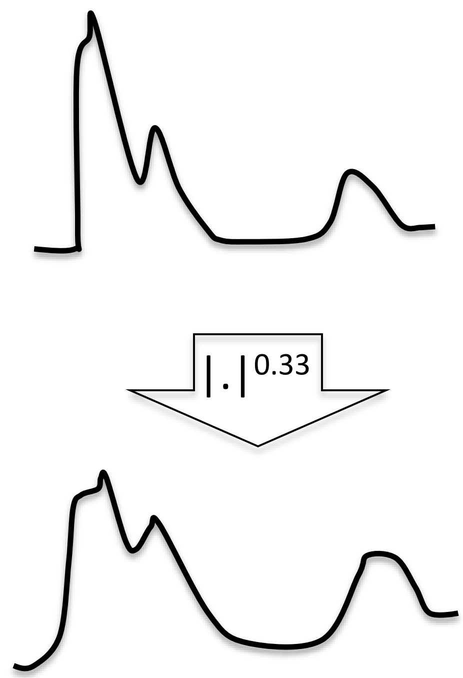


loudness [Sones]





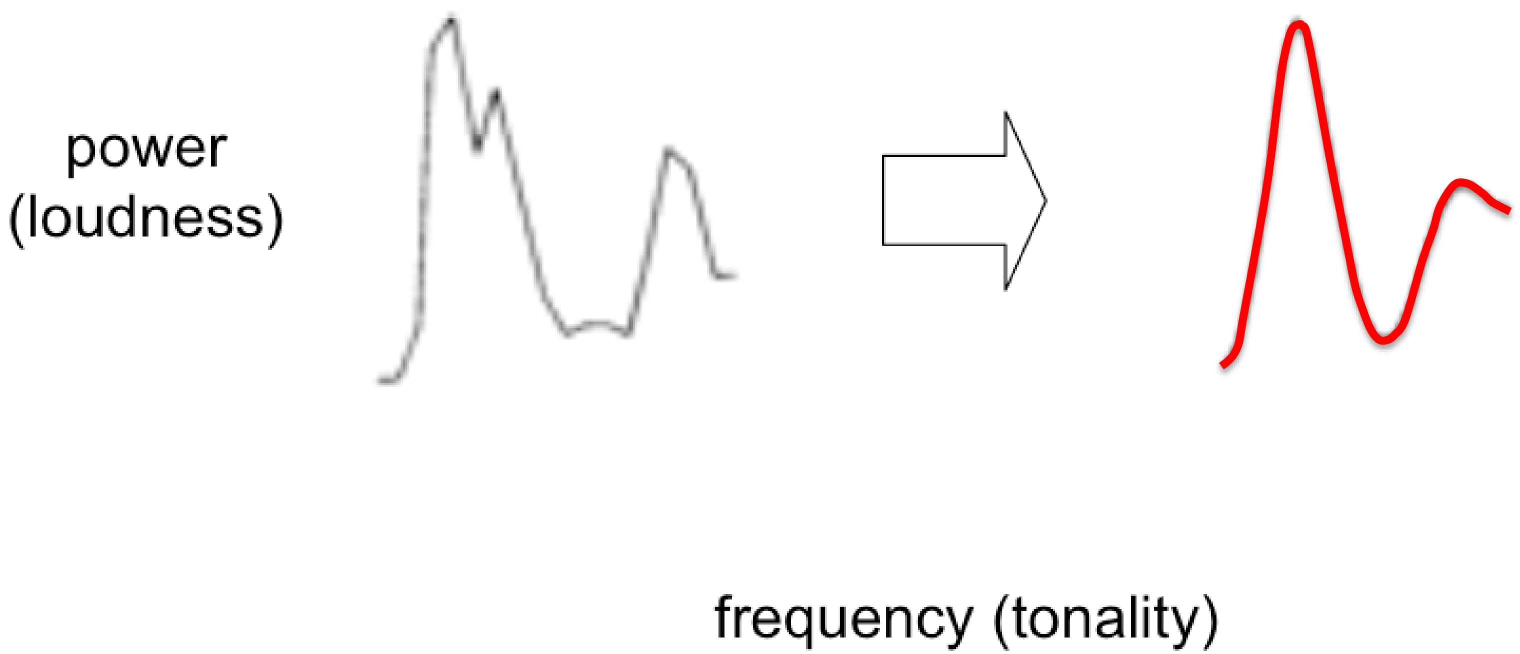
$$\text{loudness} = \text{intensity}^{\mathbf{0.33}}$$



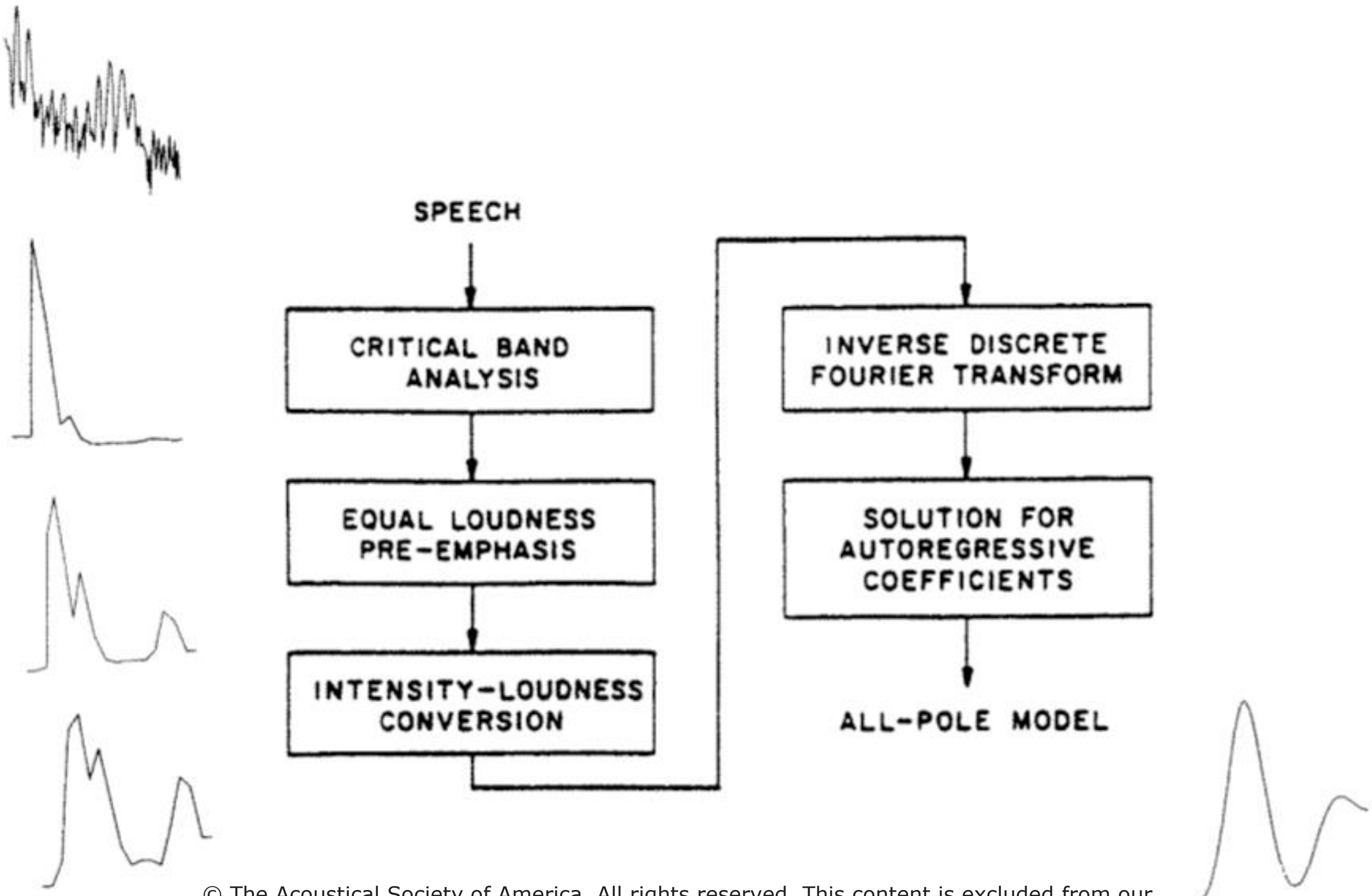
intensity
(power spectrum)

loudness

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

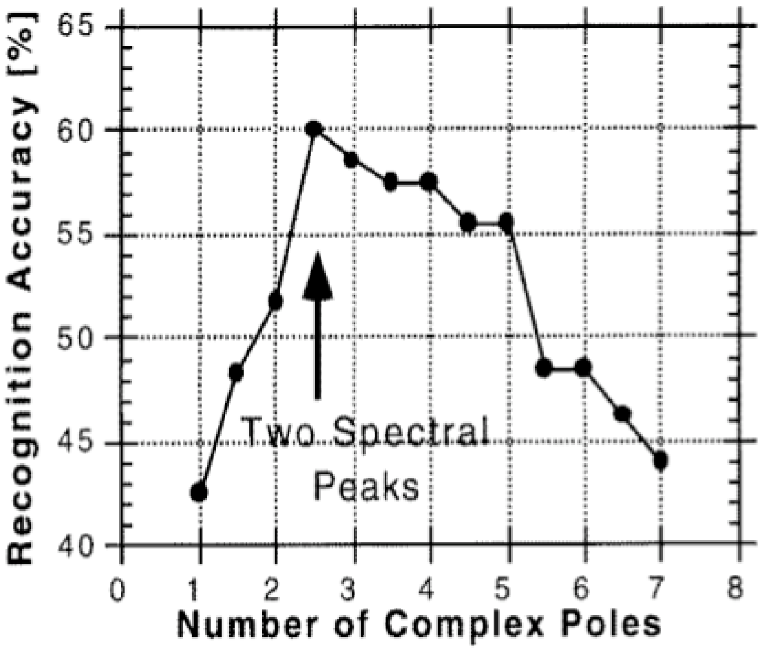


Perceptual Linear Prediction



Source: Hermansky, Hynek. "Perceptual linear predictive (PLP) analysis of speech." The Journal of the Acoustical Society of America 87, no. 4 (1990): 1738-1752.

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



Source: Hermansky, Hynek. "Should recognizers have ears?" Speech communication 25, no. 1 (1998): 3-27.

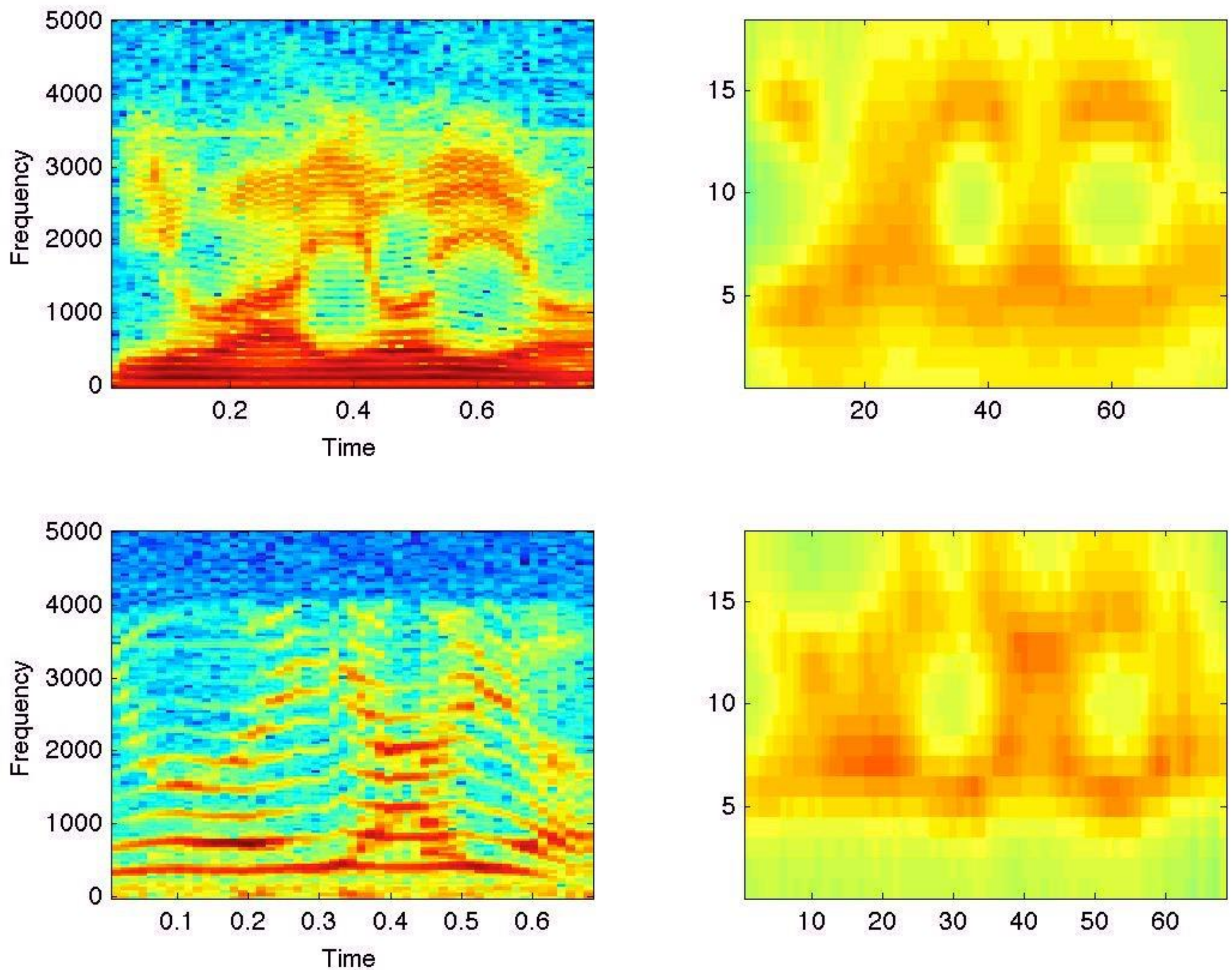
- cross-speaker ASR (trained on one speaker and tested on another)
- all speaker-dependent information harmful

short-term spectrum

5th order PLP spectrum

adult male

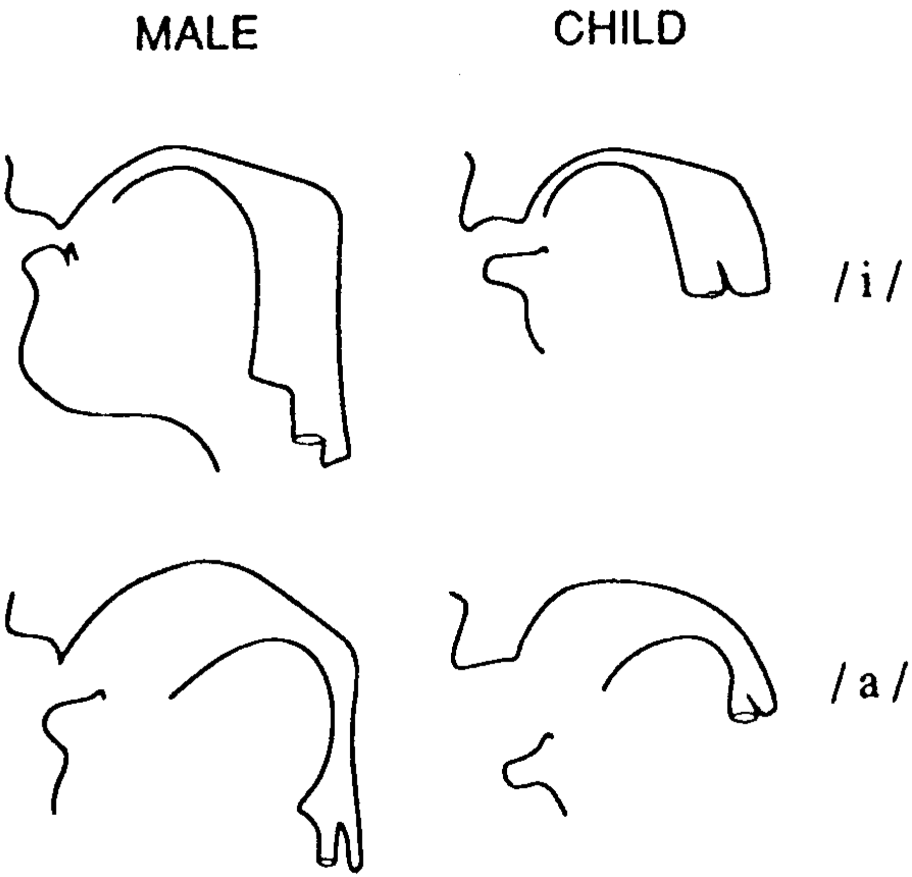
4 year old child



Source: Hermansky, Hynek, Jordan R. Cohen, and Richard M. Stern. "Perceptual properties of current speech recognition technology." Proceedings of the IEEE 101, no. 9 (2013): 1968-1985; DOI: 10.1109/JPROC.2013.2252316.

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 ICASSP 1990



Source: Hermansky, Hynek, Jordan R. Cohen, and Richard M. Stern. "Perceptual properties of current speech recognition technology." Proceedings of the IEEE 101, no. 9 (2013): 1968-1985; DOI: 10.1109/JPROC.2013.2252316.

Figure removed due to copyright restrictions. Please see the video.
Source: Hermansky, Hynek, and D. J. Broad. "The effective second formant F2' and the vocal tract front-cavity." In Acoustics, Speech, and Signal Processing, 1989. ICASSP-89.,1989 International Conference on, pp. 480-483. IEEE, 1989.

Hermansky and Broad ICASSP 1990, Hermansky JASA 1990,
Hermansky, Cohen, Stern, Proc. IEEE 2013

Listening for Shape of Front Cavity of Vocal Tract ?

message

encode

decode

shape of front cavity

perceptual second formant F2’

speaker-dependent vocal tract

speech perception

speech spectrum

Hermansky and Broad ICASSP 1990, Hermansky JASA 1990

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,…..

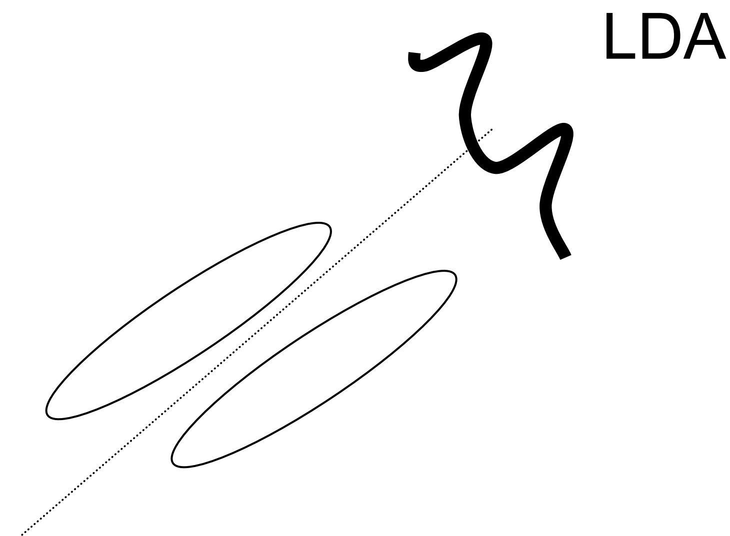
Of course, we should try to incorporate the knowledge that we have **of hearing, speech production, etc.**, into our systems,....but we need to estimate the parameter values from the data. There is no other way

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)

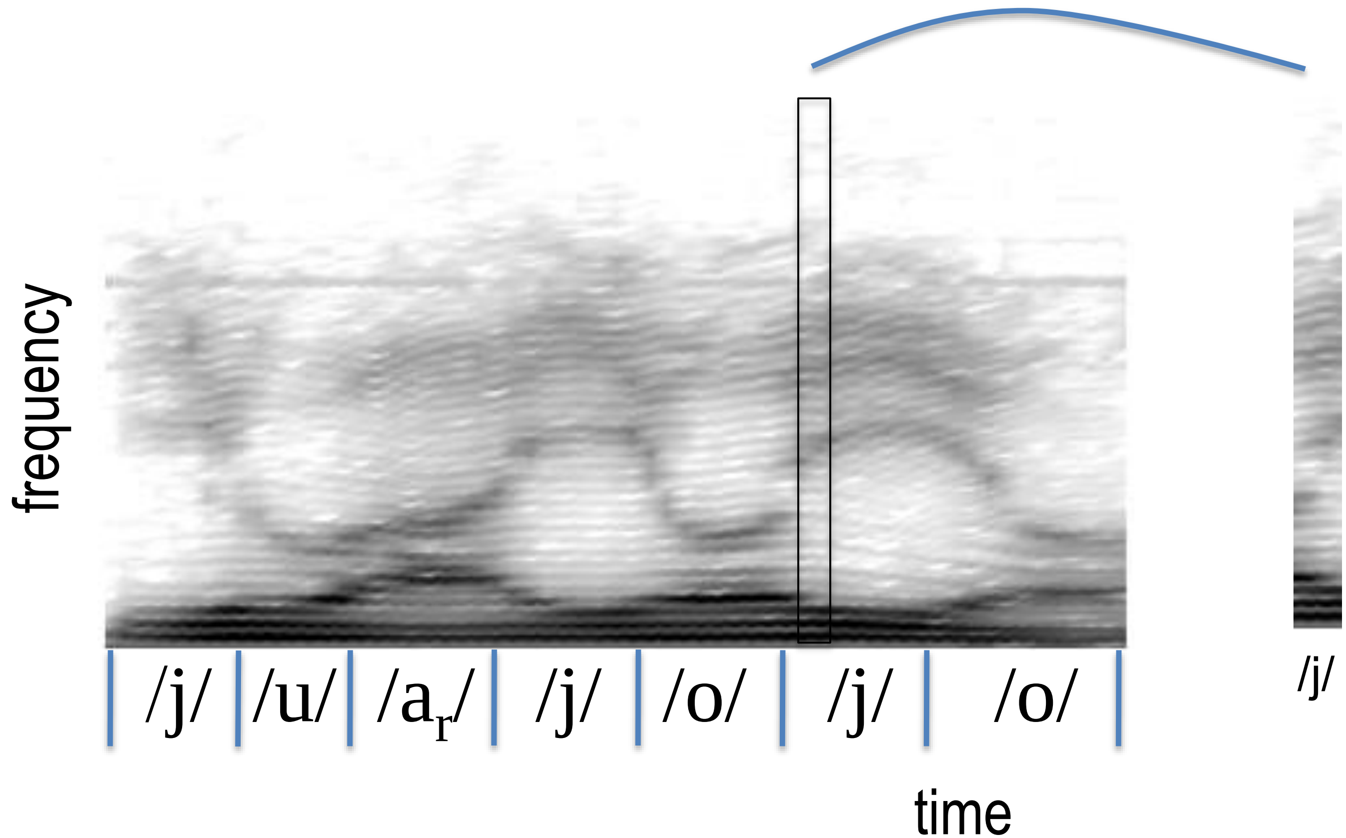
Linear discriminants: eigenvectors of $S^{-1}_W S_B$

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

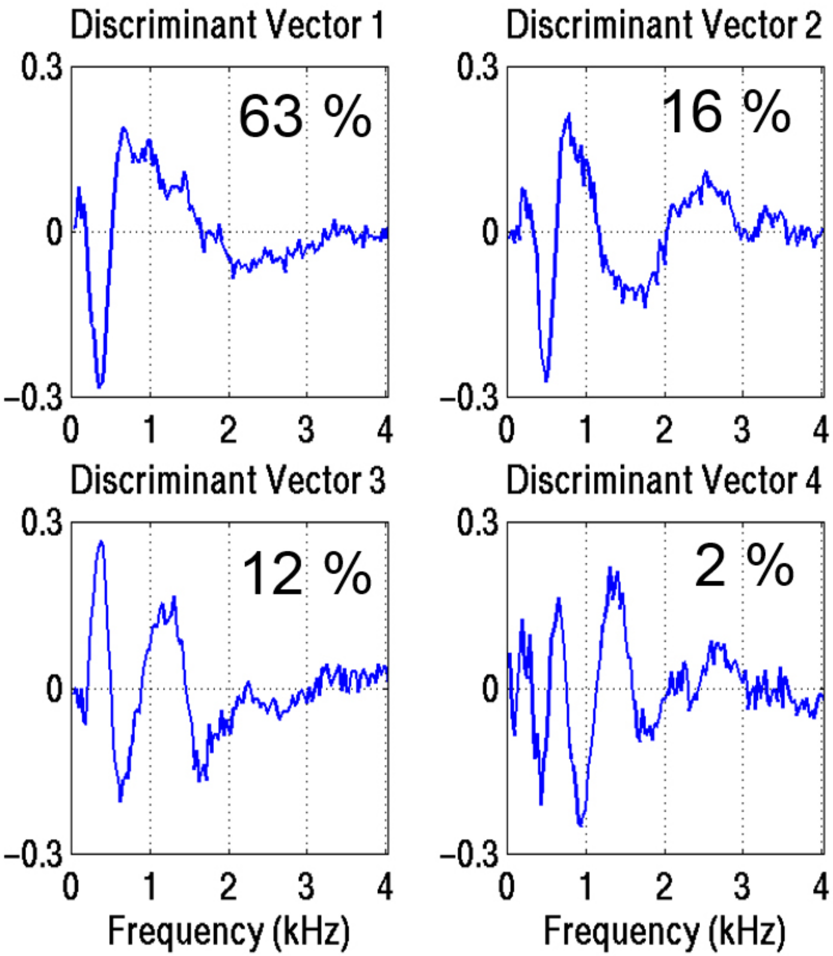


- Needs labeled data
- Within-class distributions assumed Gaussian with equal σ (take log of power spectrum)

Linear discriminant analysis (LDA) on short term spectral vectors



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

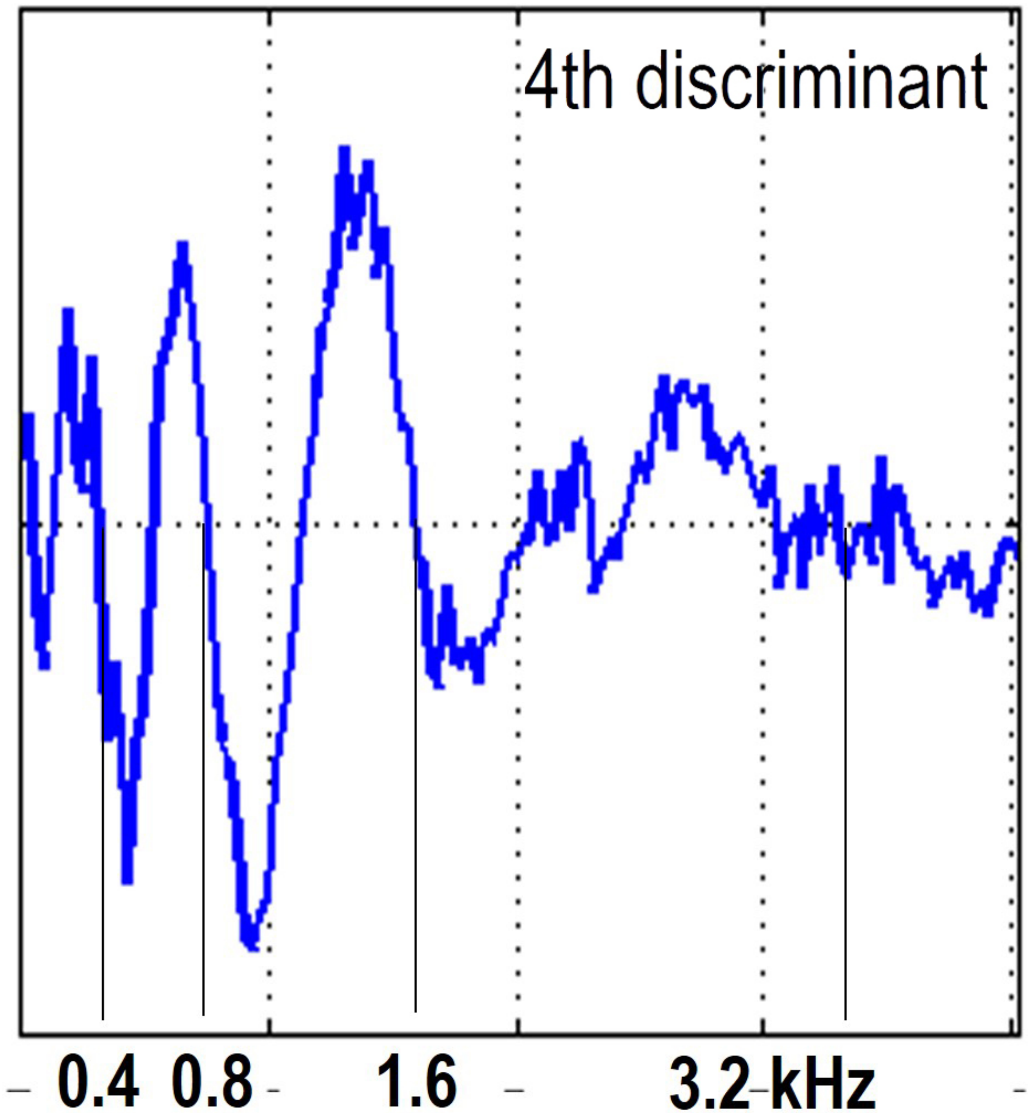
Source: Hermansky, Hynek. "Data-guided processing of speech." In Workshop on Spoken Language Processing. 2003.

4 discriminants
(92 % of variance)

<u>1 period / octave</u>

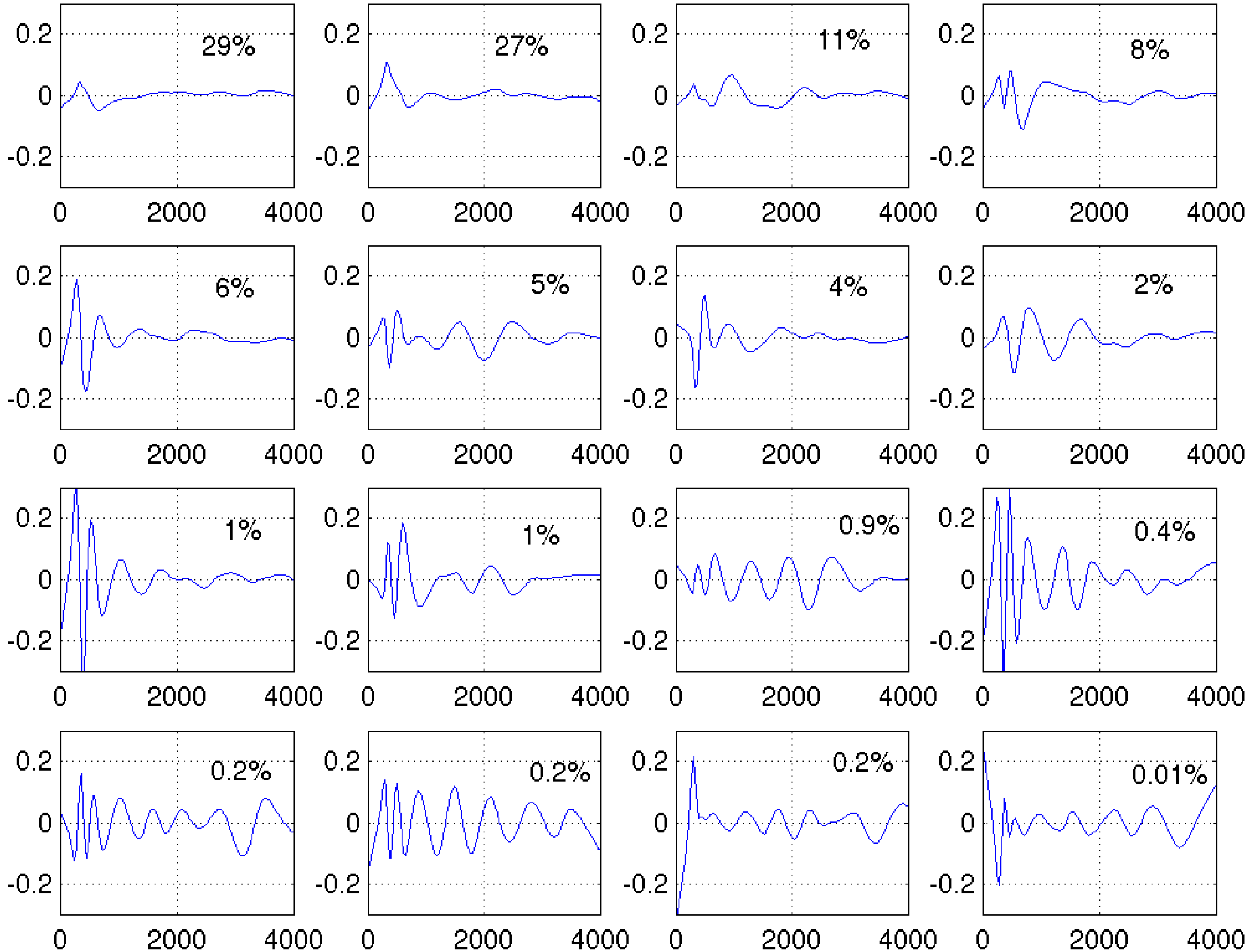
- resolution decreases with frequency

- resolution coarser than critical bands

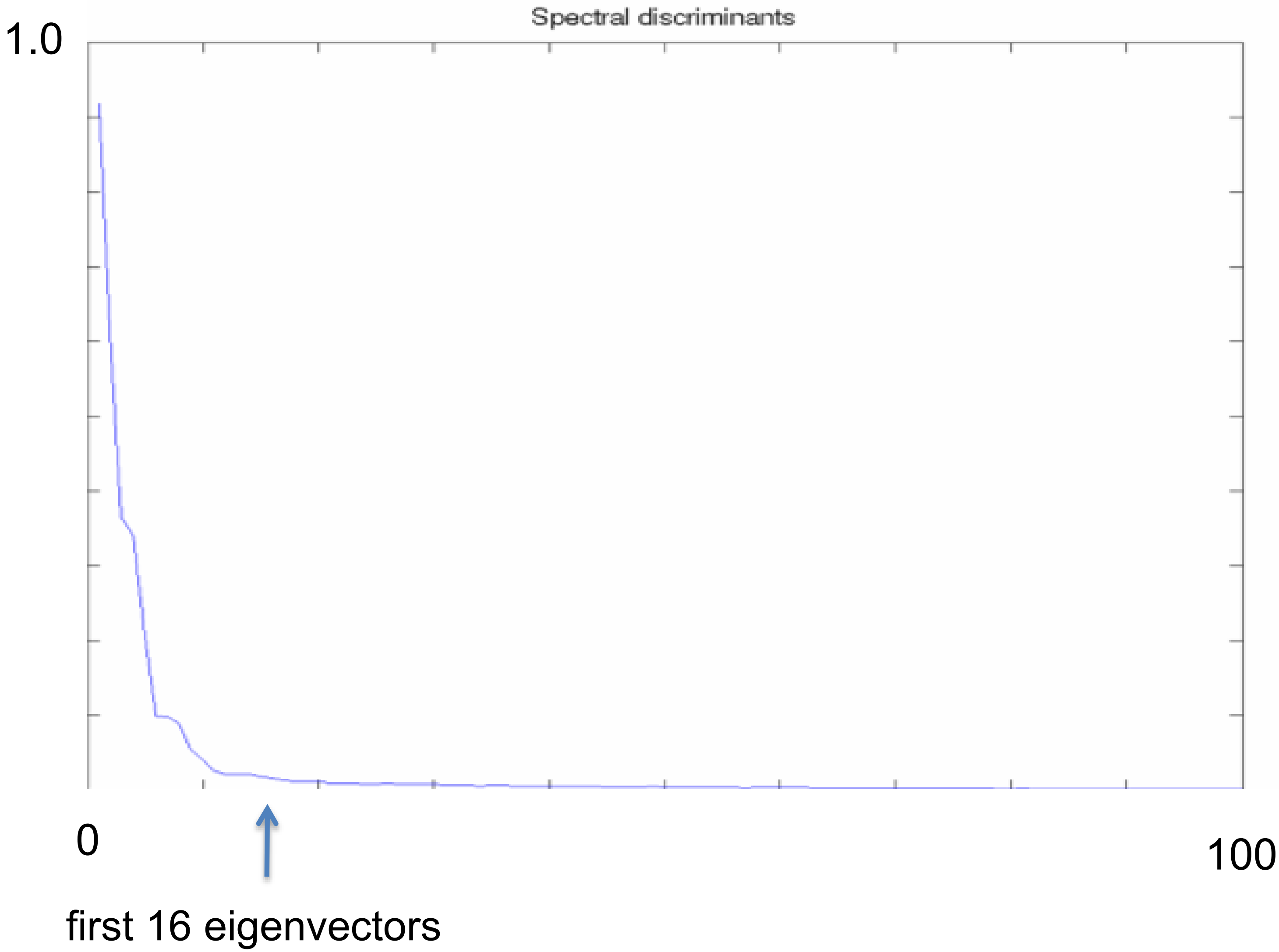


Source: Hermansky, Hynek. "Data-guided processing of speech." In Workshop on Spoken Language Processing. 2003.

30 hours of Resource Management and Switchboard labeled speech data (courtesy of SRI)

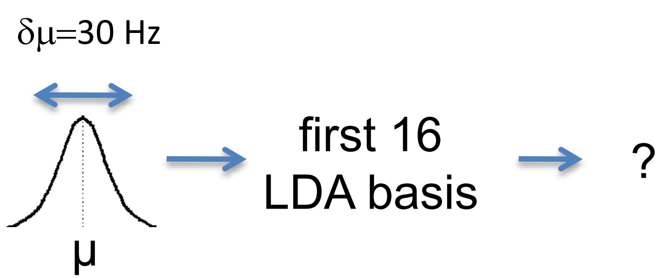


Eigenvalues of the discriminant matrix

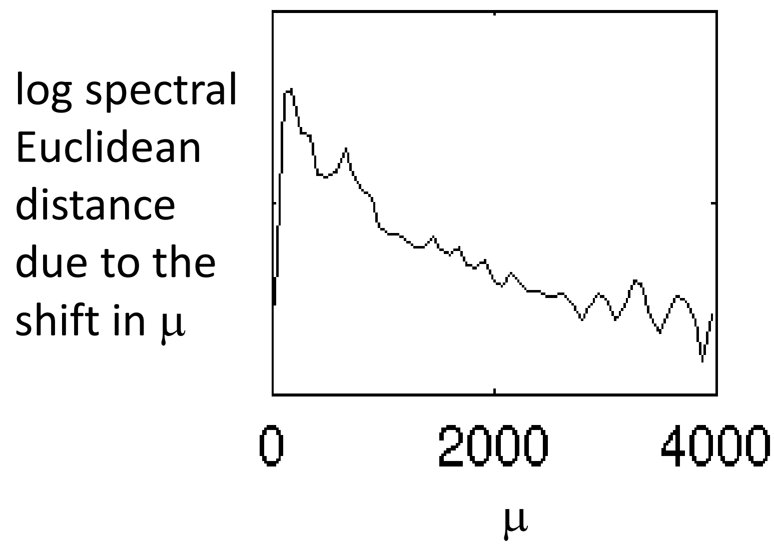


Spectral sensitivity of projections

- Perturbation analysis
  - project Gaussian shape on the first 16 spectral basis and evaluate the effect of the shift in μ by 30 Hz as the function of μ



Shift in μ constant on the Hz scale

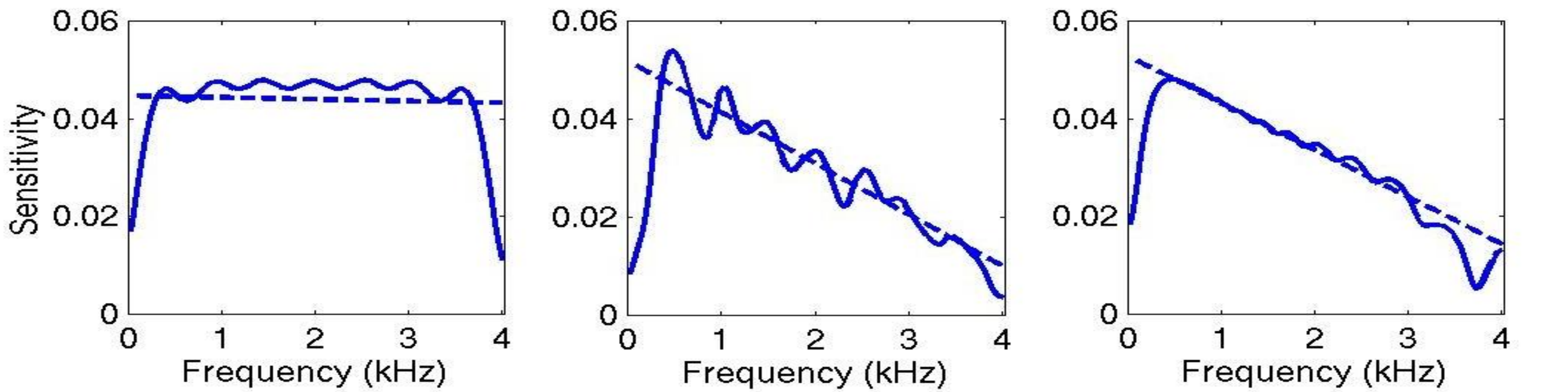


Decreasing spectral sensitivity with increasing frequency
- consistent with spectral resolution of hearing

Sensitivity to Spectral Change

(Malayath 1999)

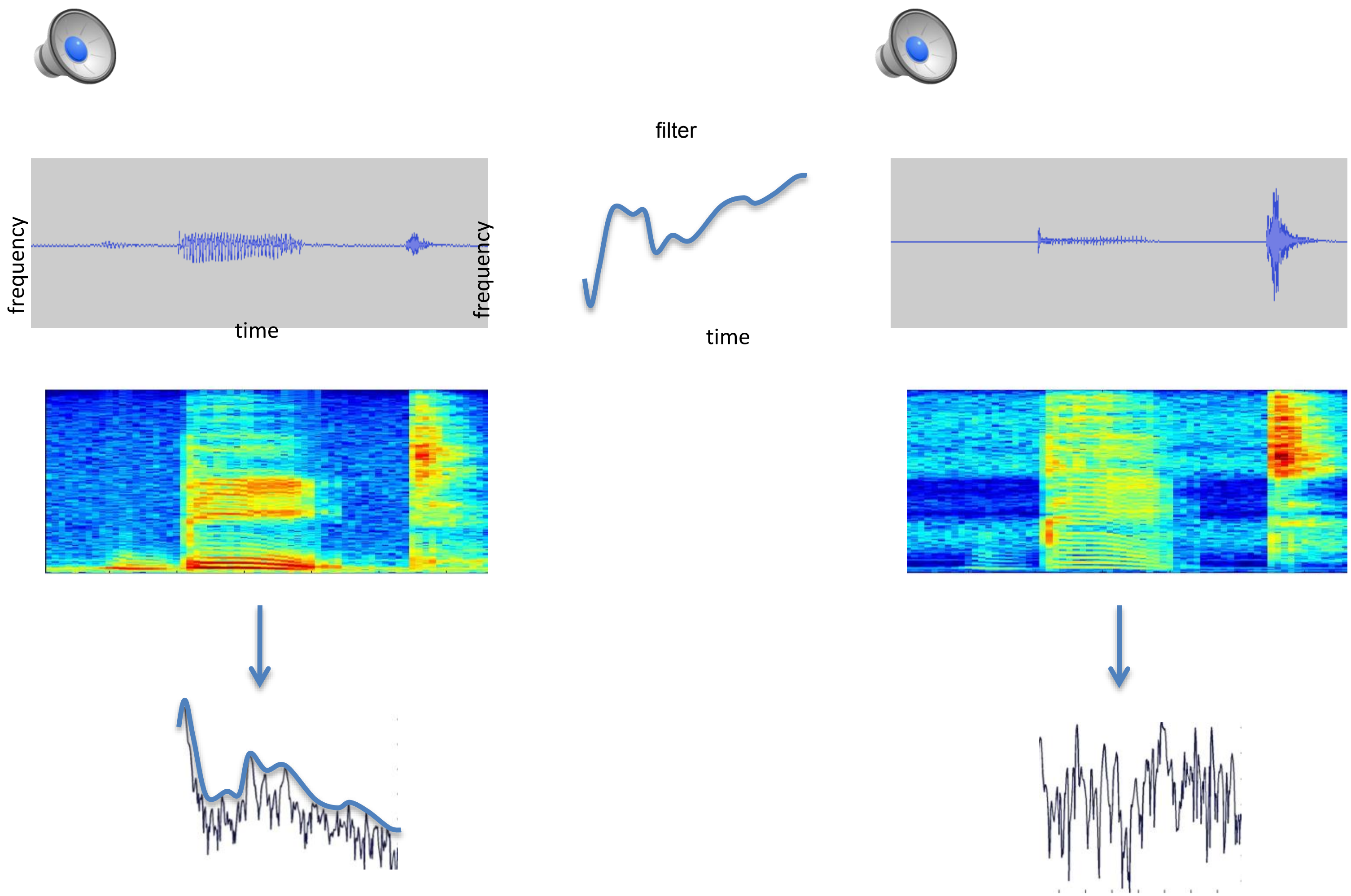
Cosine basis **LDA-derived basis** **Critical-band filterbank**



Linear distortions (filtering)

original speech

filtered speech



Source: Hermansky, Hynek, Jordan R. Cohen, and Richard M. Stern. "Perceptual properties of current speech recognition technology." Proceedings of the IEEE 101, no. 9 (2013): 1968-1985; DOI: 10.1109/JPROC.2013.2252316.

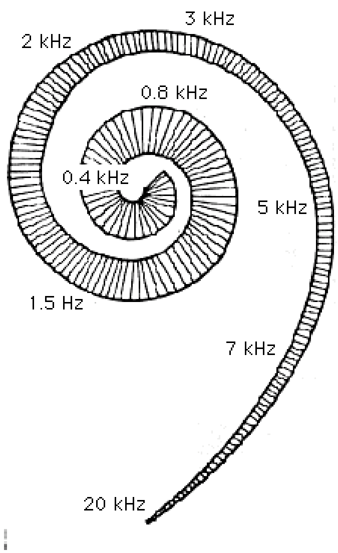
Effect of fixed linear distortions

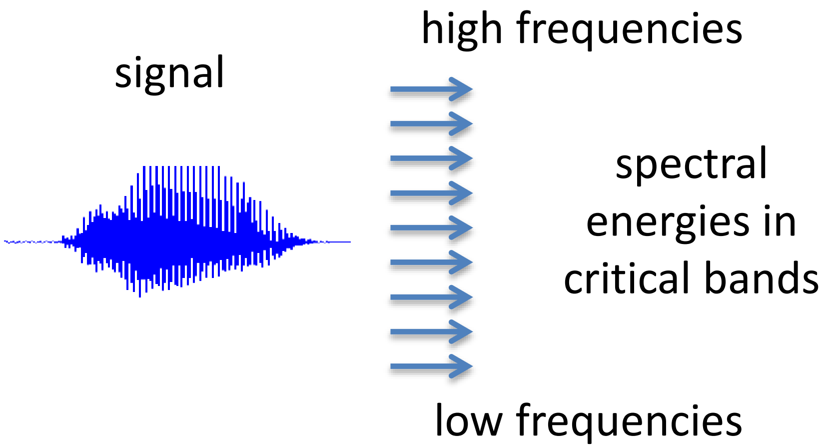
$$x(t) = s(t) * e(t)$$

$$\log[FT\{s(t) * e(t)\}] = \log S(\omega) + \log E(\omega)$$

- Convolution of speech with impulse response of the distorting filter
- Results in **different** additive constant **at different frequencies** in logarithmic spectral domain

Spectral analysis in ear





Ear is frequency selective in order to yield **frequency-localized temporal patterns** for processing by higher processing levels in hearing.

Exploiting spectral selectivity in engineering

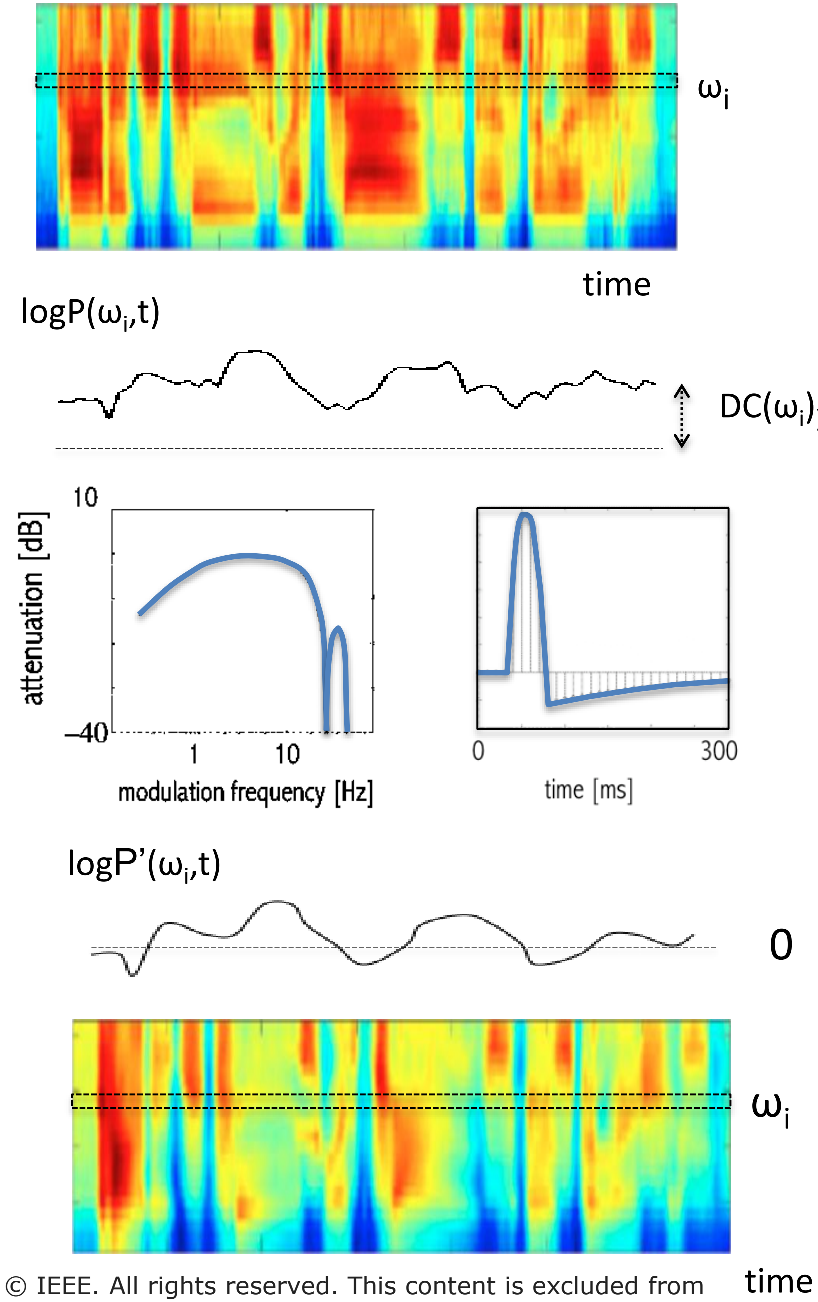
1. Separate speech into different frequency channels

1. Do independent processing in each frequency channel

RASTA processing

inspired by Marr 1974
"lightness" = luminance with slowly varying components removed

Figure removed due to copyright restrictions. Please see the video.
Source: Hermansky, Hynek, and Nelson Morgan."RASTA processing of speech." IEEE transactions on speech and audio processing 2, no. 4 (1994): 578-589.

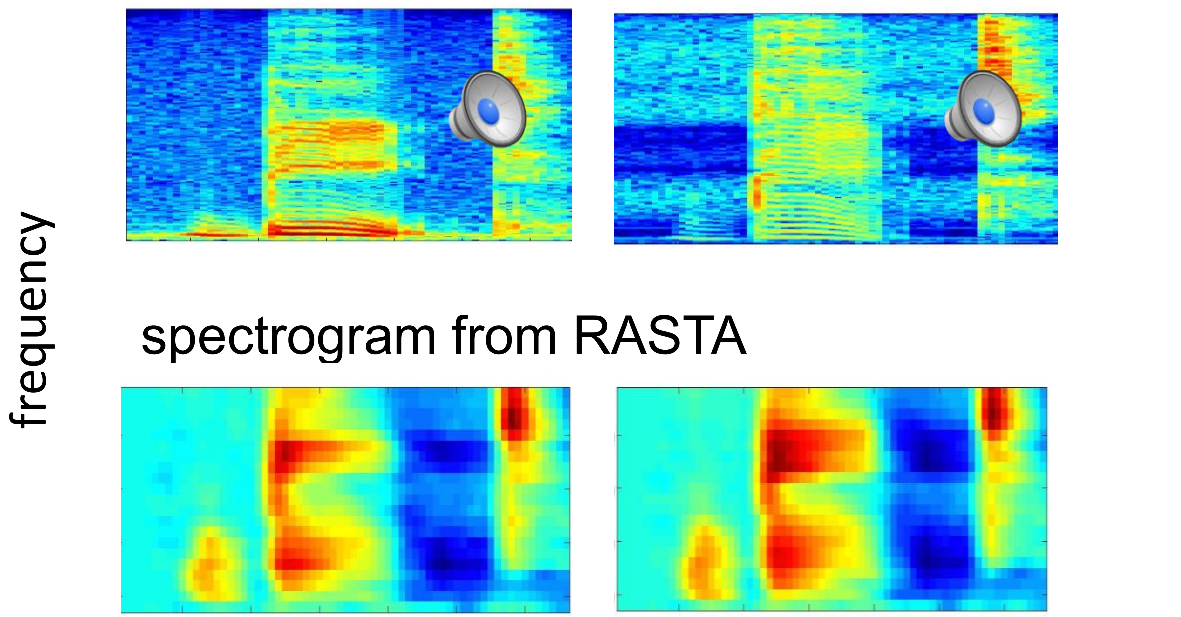


Hermansky and Morgan 1994

original speech filtered speech

spectrogram

spectrogram from RASTA

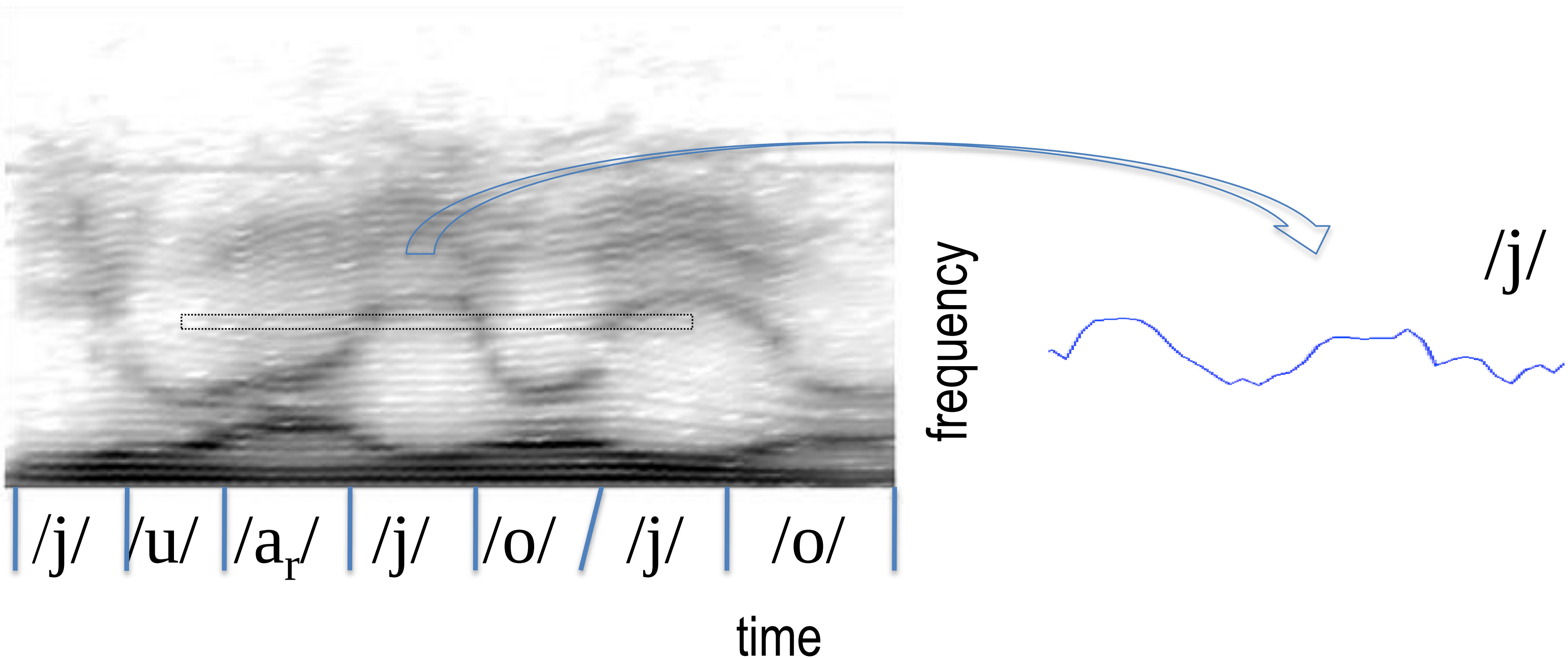


Source: Hermansky, Hynek, Jordan R. Cohen, and Richard M. Stern. "Perceptual properties of current speech recognition technology." Proceedings of the IEEE 101, no. 9 (2013): 1968-1985; DOI: 10.1109/JPROC.2013.2252316.

Environmental mismatch in training and in test

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

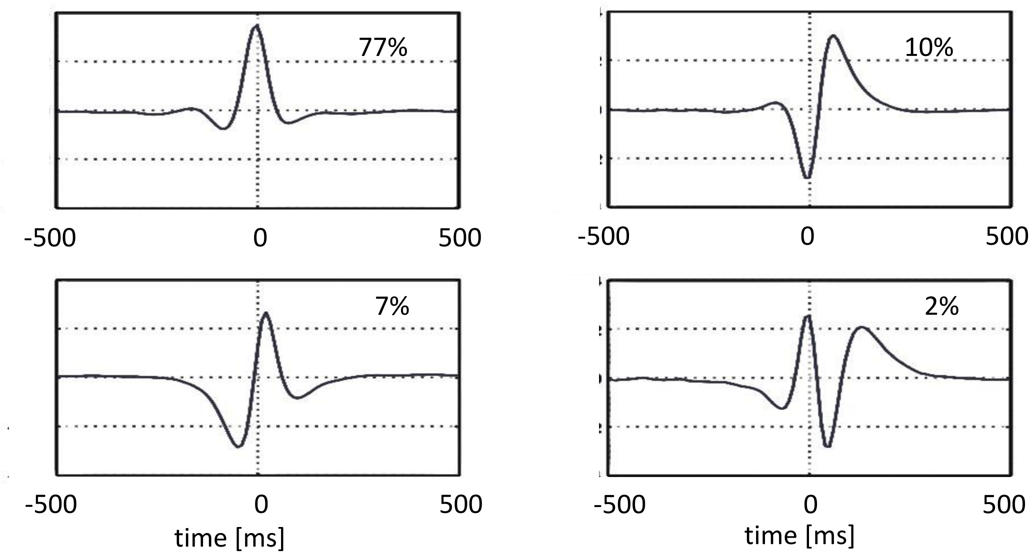
Linear Discriminant Analysis on Temporal Vectors



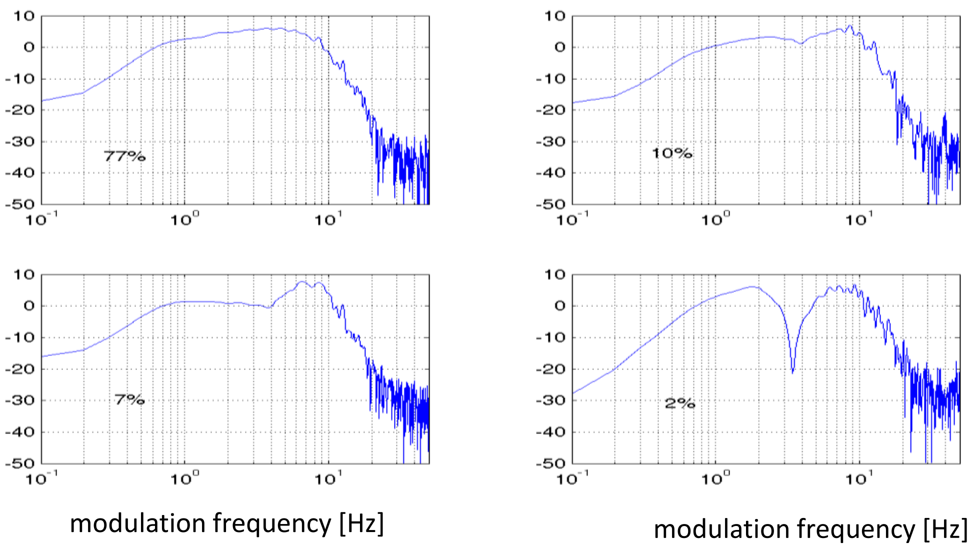
labeled data – labeled temporal vector space – LDA FIR FILTER IMPULE RESPONSES

impulse responses

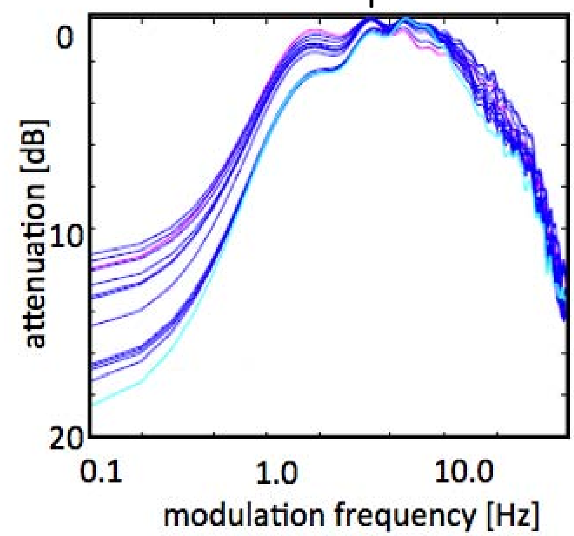
van Vuuren and Hermansky 1997,
Valente and Hermansky 2006



frequency responses

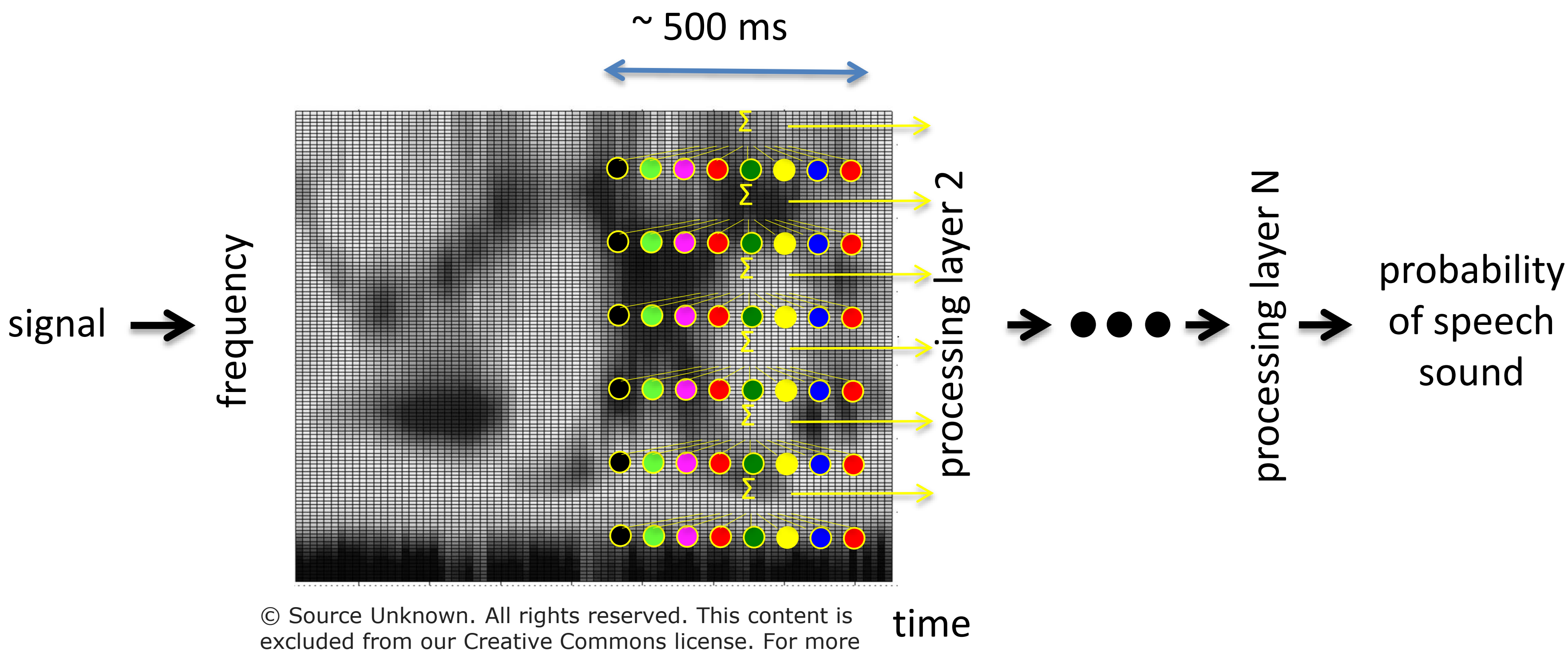


similar filters at all carrier frequencies



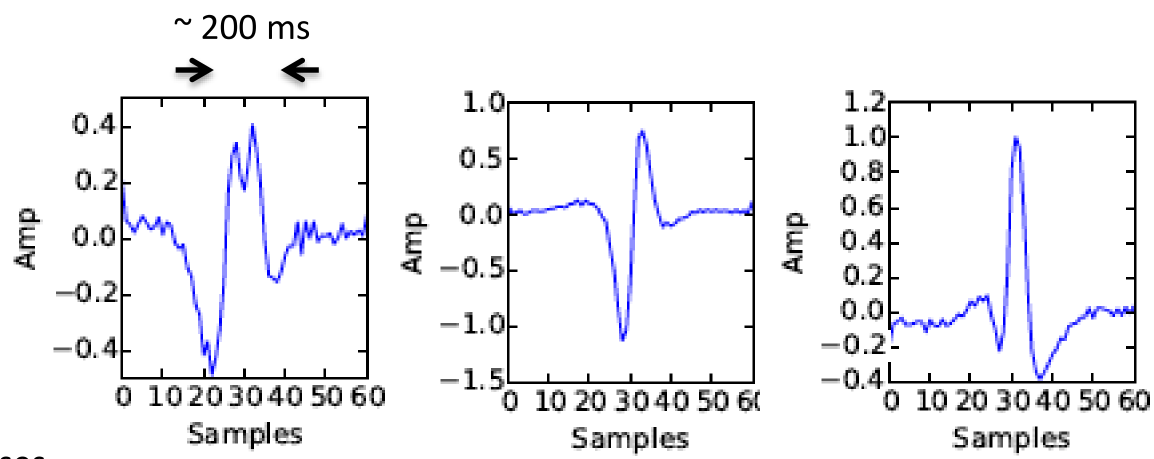
Source: Hermansky, Hynek, Jordan R. Cohen, and Richard M. Stern. "Perceptual properties of current speech recognition technology." Proceedings of the IEEE 101, no. 9 (2013): 1968-1985; DOI: 10.1109/JPROC.2013.2252316.

DNN with convolutions in time

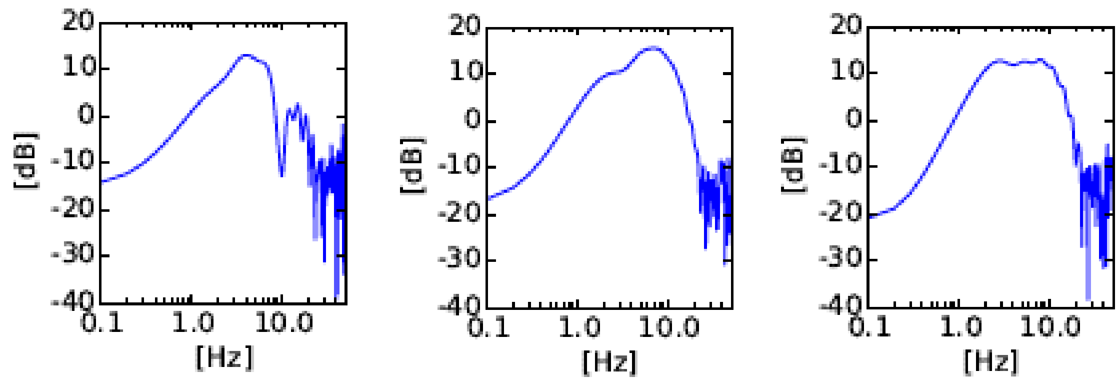


with Peddinti, Pesan, Vesely and Burget

filter impulse responses



filter frequency responses

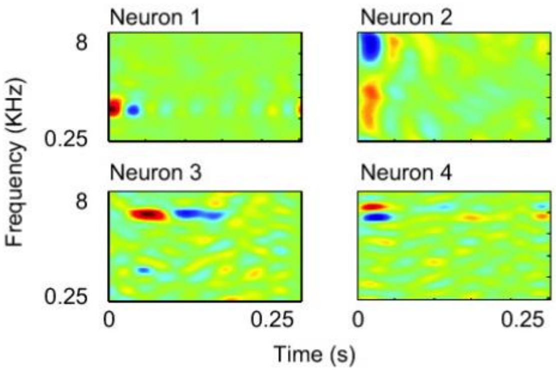


Courtesy of Interspeech. Used with permission.

Source: Pešán, Jan, Lukáš Burget, Hynek Hermanský1, and Karel Veselý. "DNN derived filters for processing of modulation spectrum of speech." In Sixteenth Annual Conference of the International Speech Communication Association. 2015.

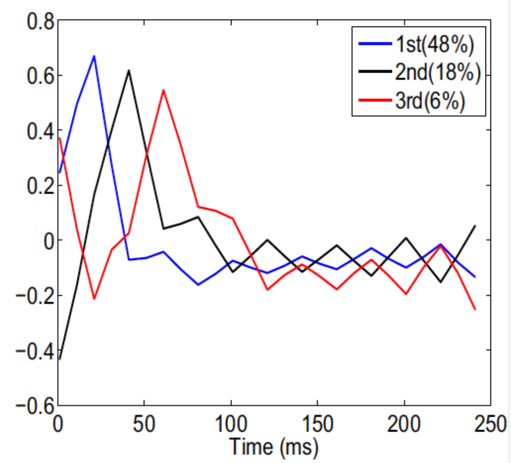
Pešan, Burget, Hermansky, Vesely *Interspeech 2015*

Auditory cortical receptive fields



Thomas et al INTERSPEECH 2010

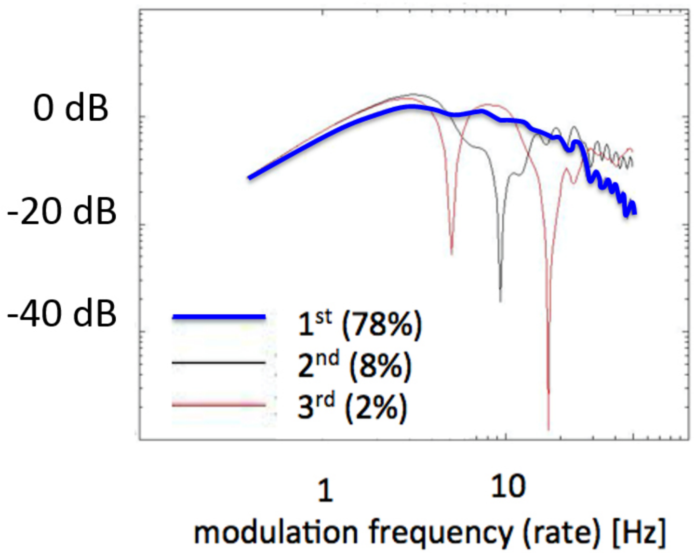
Source: Thomas, Samuel, Sriram Ganapathy, and Hynek Hermansky. "Cross-lingual and multi-stream posterior features for low resource LVCSR systems." In Interspeech, pp. 877-880. 2010.



Source: Mahajan, Nagaraj, Nima Mesgarani, and Hynek Hermansky. "Principal components of auditory spectro-temporal receptive fields." In INTERSPEECH, pp. 1983-1987. 2014.

Temporal principal components from about 2000 cortical receptive fields

Mahajan, Mesgarani, Hermansky, INTERSPEECH 2014

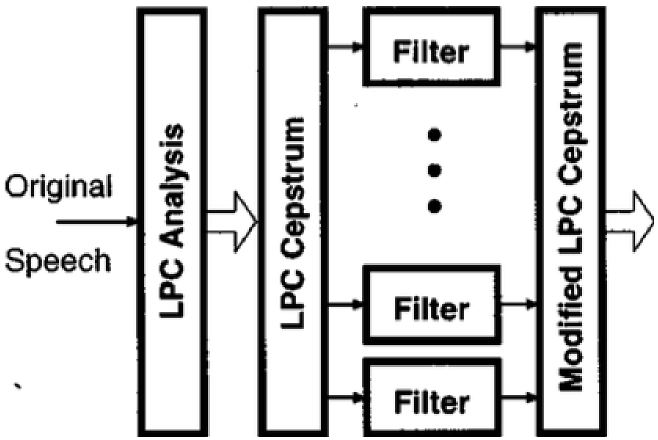


ignoring phase shifts

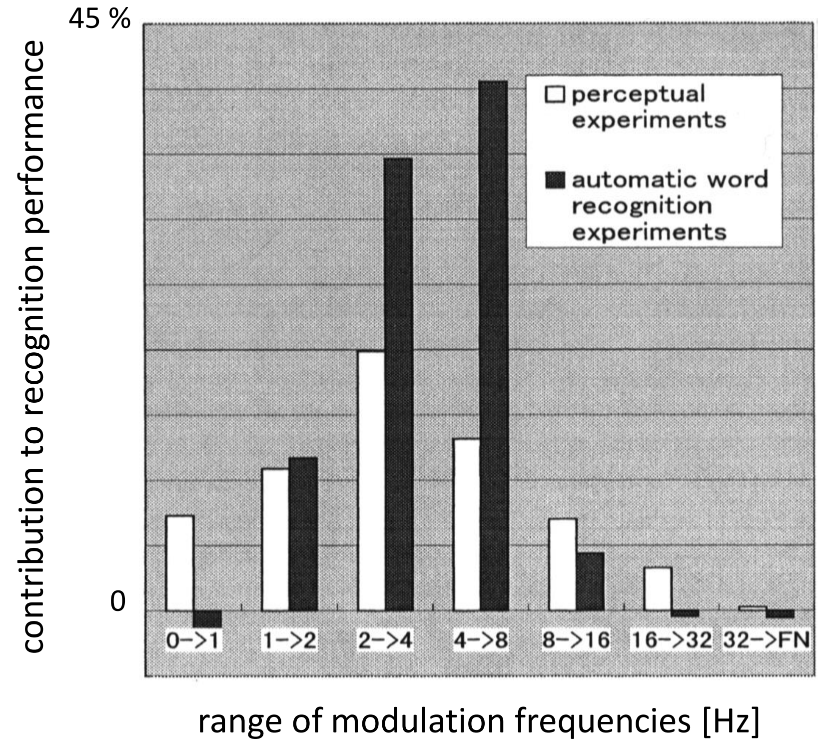
(principal components of magnitudes of temporal components of STRFs)

Mahajan and Hermansky, *in preparation*

Slow Modulations and Speech Communication



Human and machine recognition experiments
(with Kanedera, Arai, and Pavel 1999)



Slow Modulations and Speech Communication

Inaudible **message** in slow motions of vocal tract is made audible by **modulating** the audible carrier

-Dudley 1940

Flow chart of sound filtering removed due to copyright restrictions. Please see the video.

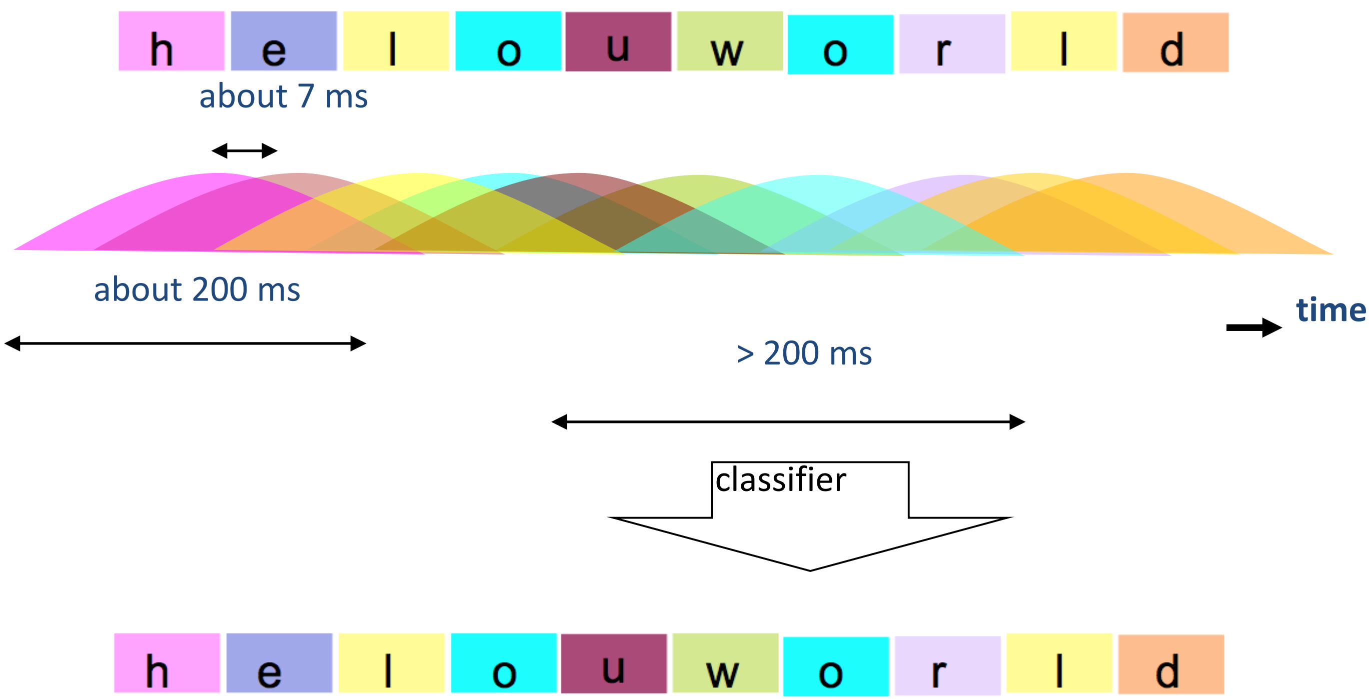
Information about a message is in slow changes of speech signal in individual frequency bands

Slow modulations – long time spans ! (5 Hz - > 200 ms)

- 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
- sub-threshold integration of speech sounds within 200 ms
- e.t.c.

syllable-length buffer of human hearing ?

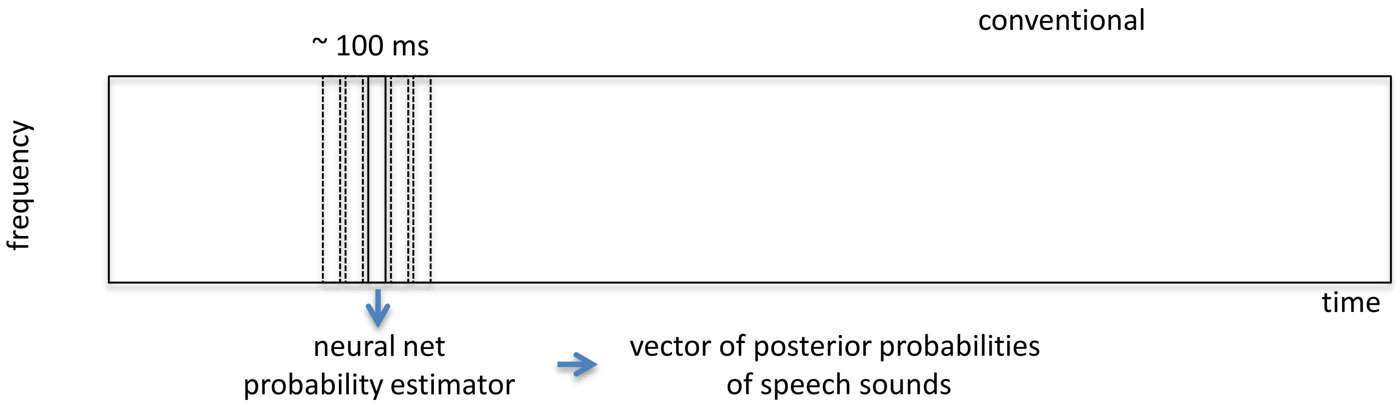
Where are speech sounds (phonemes) ?



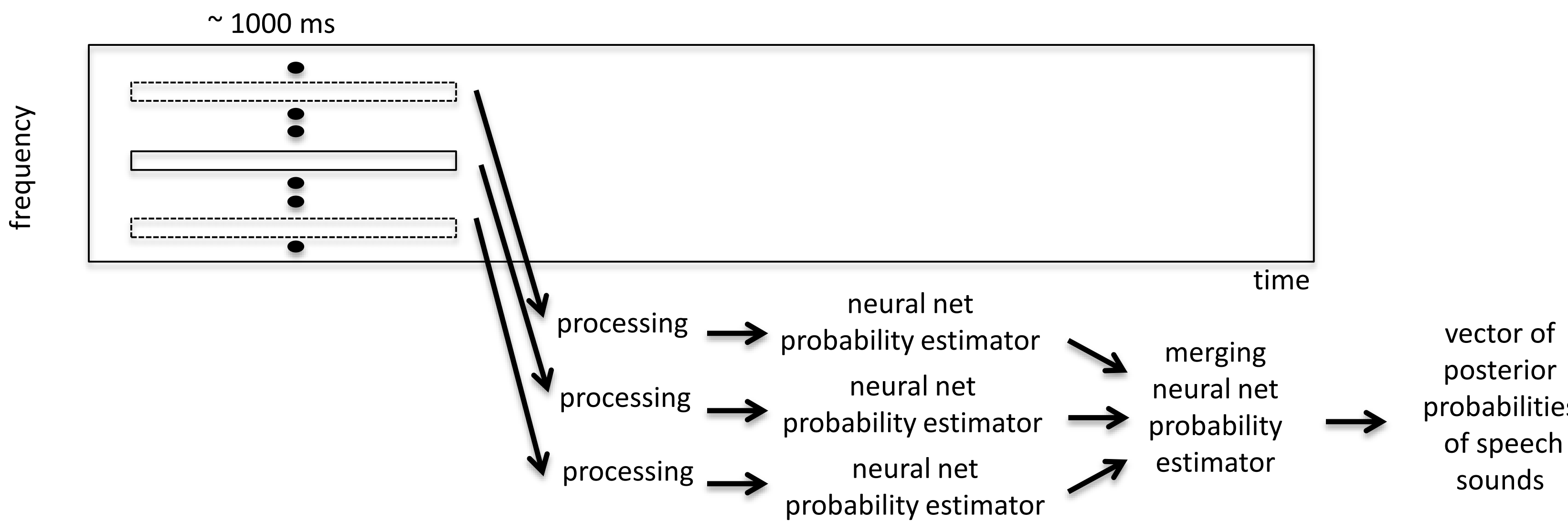
Source: Hermansky, Hynek, Jordan R. Cohen, and Richard M. Stern. "Perceptual properties of current speech recognition technology." Proceedings of the IEEE 101, no. 9 (2013): 1968-1985; DOI: 10.1109/JPROC.2013.2252316.

TRAPS

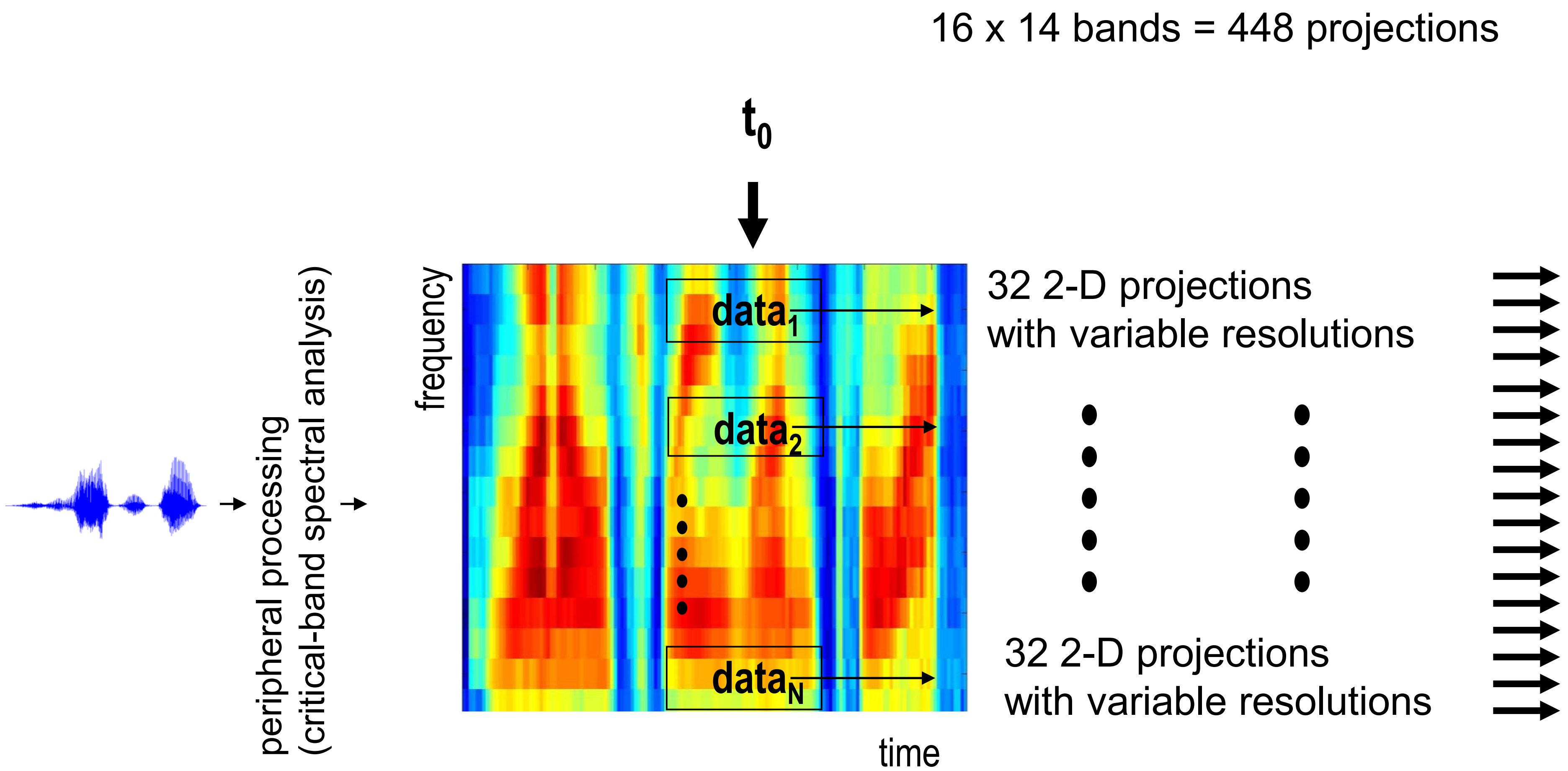
Hermansky and Sharma, ICSLP 1998



Classifying **T**empo**RA**l **P**atterns of **S**pectral Energies



Emulation of cortical processing (MRASTA)

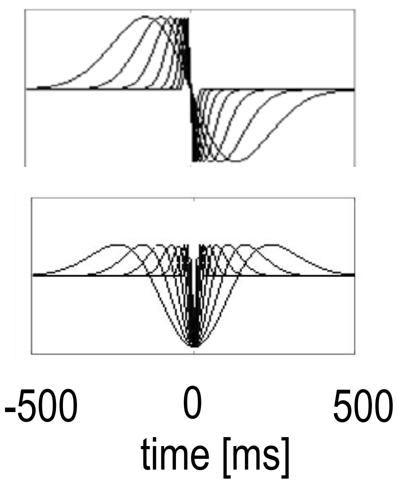


Multi-resolution RASTA (MRASTA)

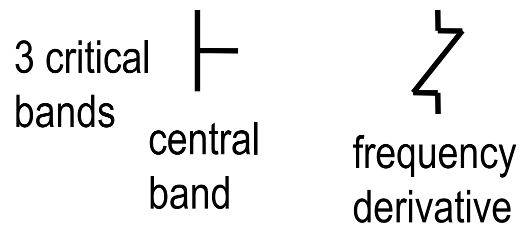
(Interspeech 05)

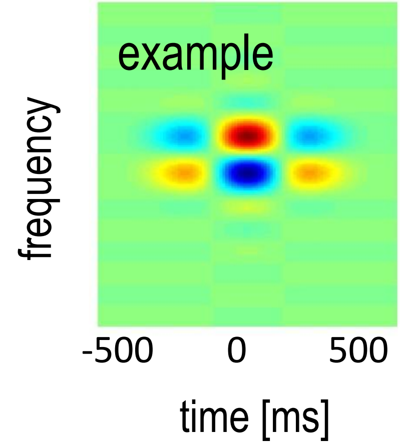
Spectro-temporal basis formed by outer products of

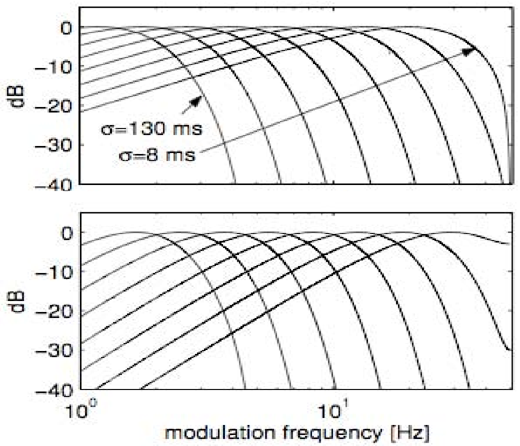
time



frequency







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

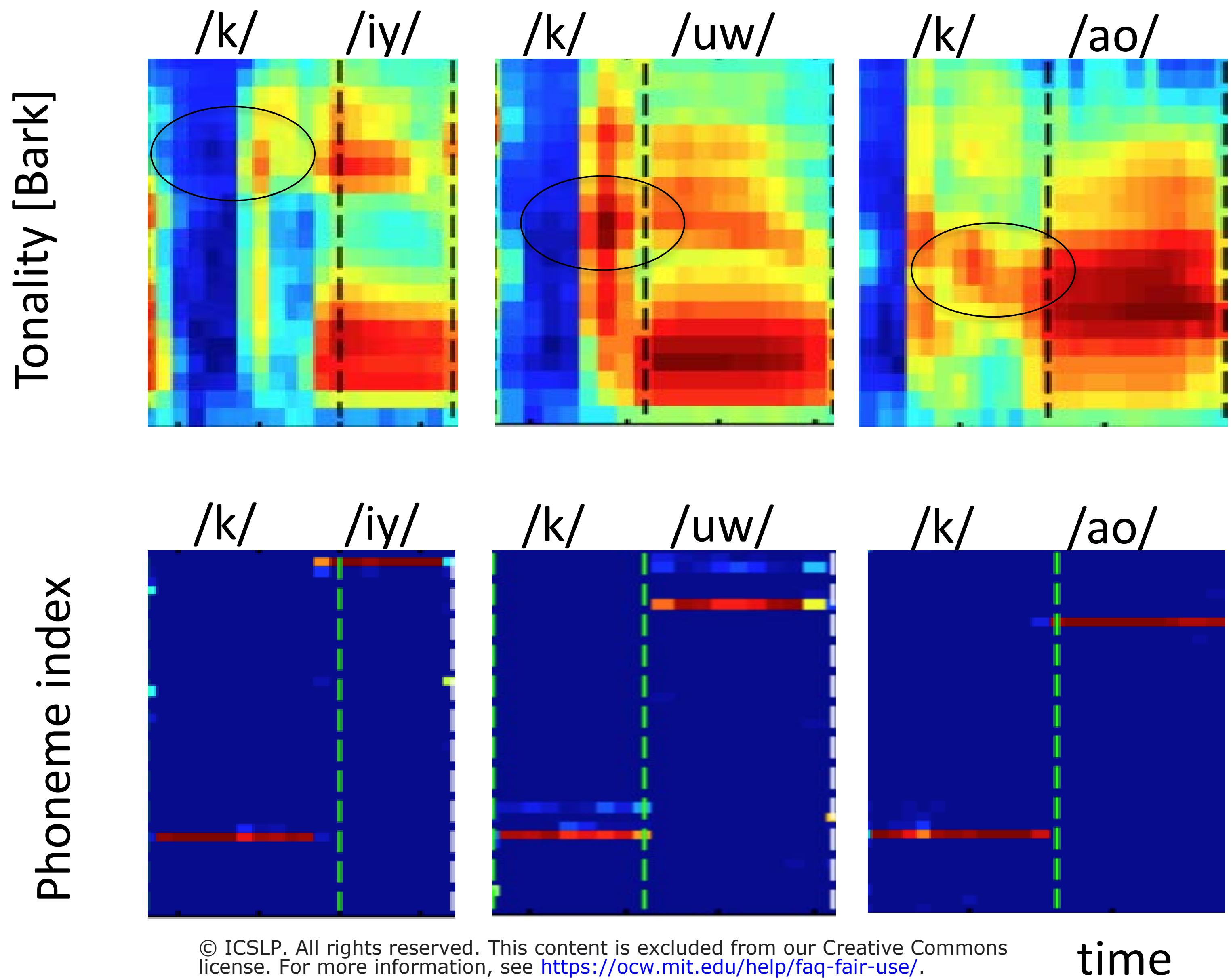
1.RASTA-like: alleviates stationary components
2.multi-resolution in time

Source: Hermansky, Hynek, and Petr Fousek. Multi-resolution RASTA filtering for TANDEM-based ASR. No. EPFL-REPORT-83199. IDIAP, 2005.

Some “novel” (in 1998) elements of TRAPS

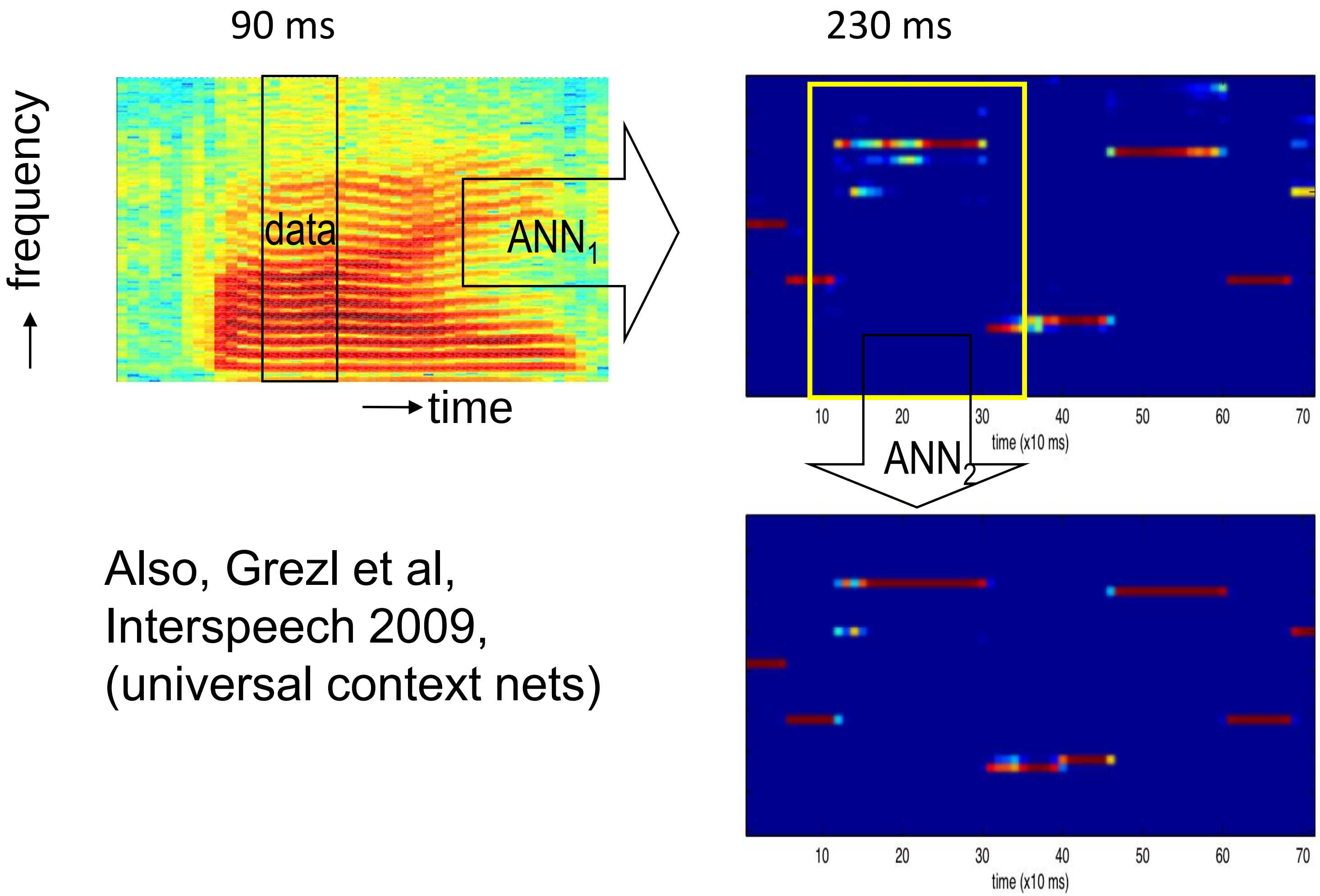
- Rather long temporal context of the signal as input
- Hierarchical structured neural net (“deep neural net”)
- Independent processing in frequency-localized parallel neural net estimators
  - most of these elements typically found in current state-of-the-art speech recognition systems

However, parts of TRAPS DNN trained individually, while today’s DNNs are optimized jointly


/k/
/iy/
/k/
/uw/
/k/
/ao/
Tonality [Bark]
/k/
/iy/
/k/
/uw/
/k/
/ao/
Phoneme index
time

Serial hierarchical estimation

(Pinto et al, Interspeech 2008)



Results (CTS) : Phoneme recognition accuracy **55.3%**

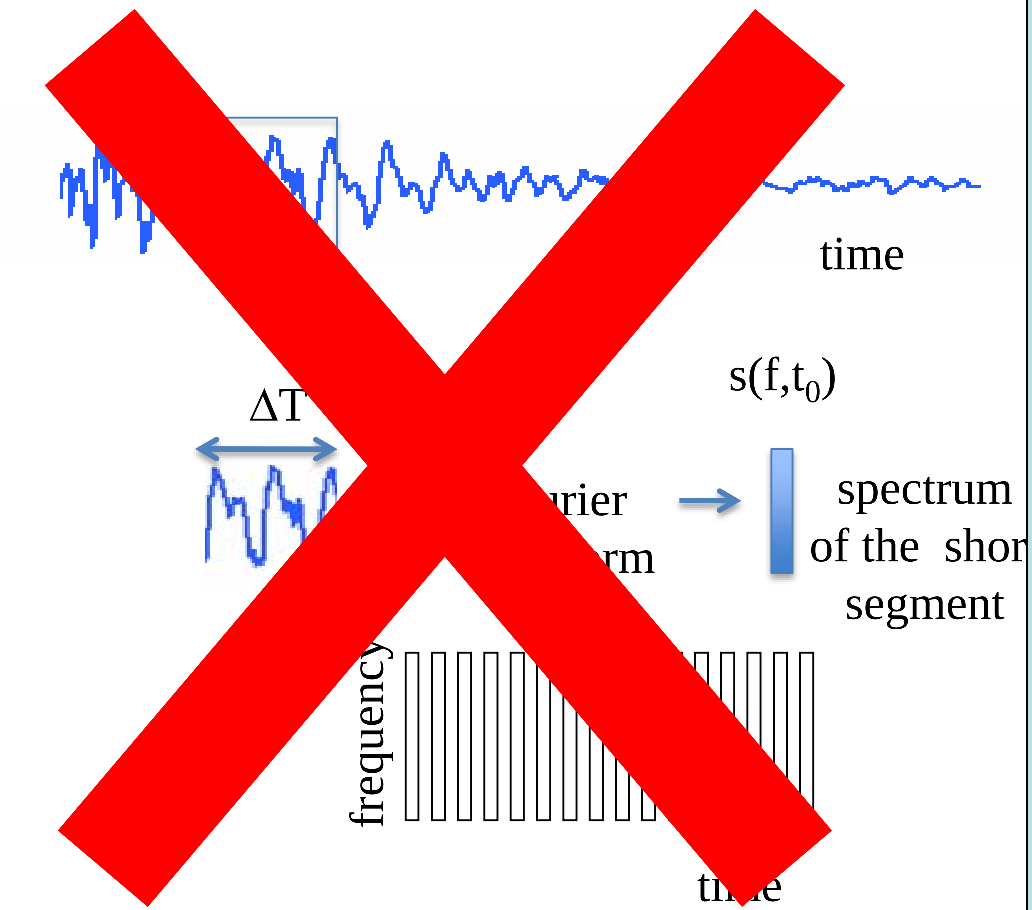
63.6% accuracy

Also, Grezl et al, Interspeech 2009, (universal context nets)

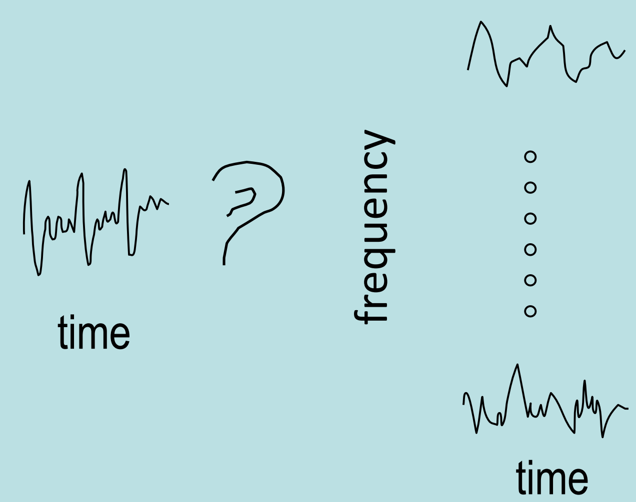
Picture of Columbo removed due to copyright restrictions. Please see the video.

- Processing of frequency-localized temporal trajectories of spectral energies (rather than short-time spectral envelopes) appears to offer a number of advantages

Away from Short-Term Spectrum

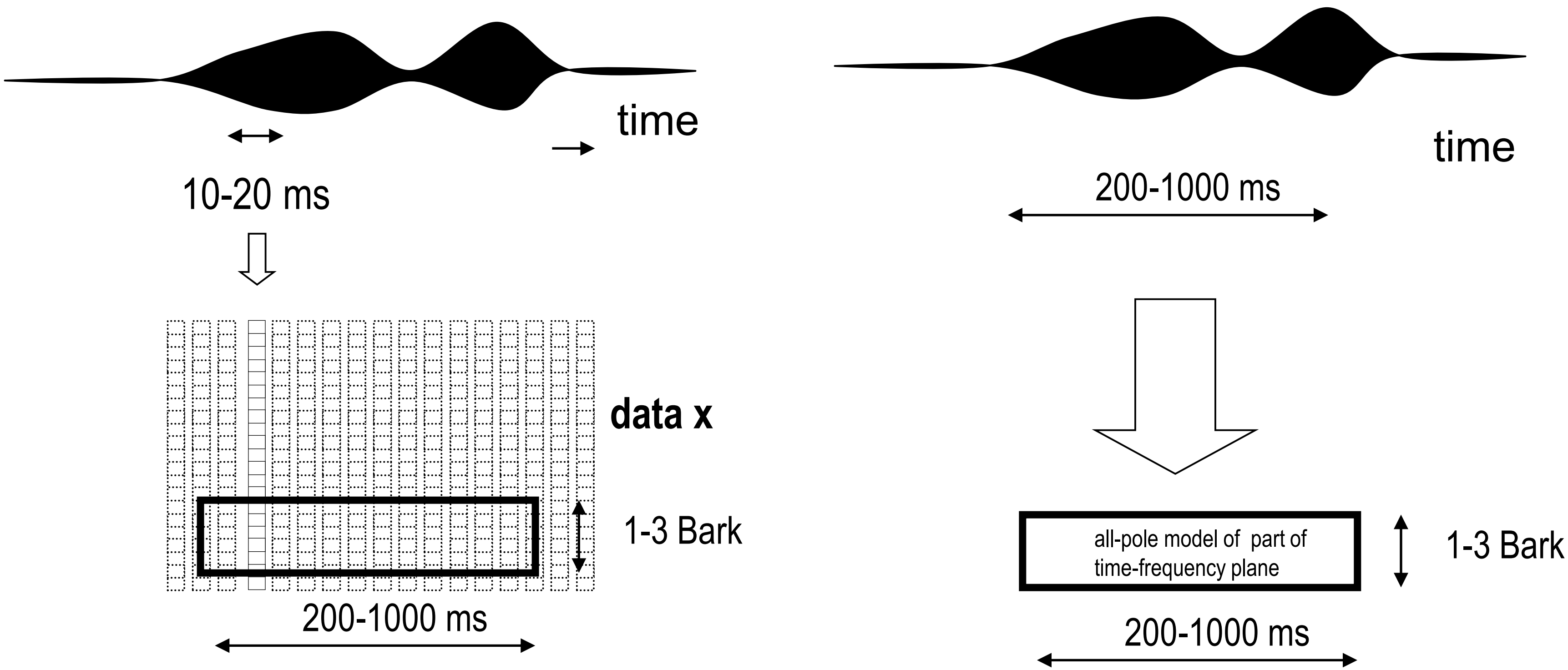


back to human hearing



How to Get Estimates of Temporal Evolution of Spectral Energy ?

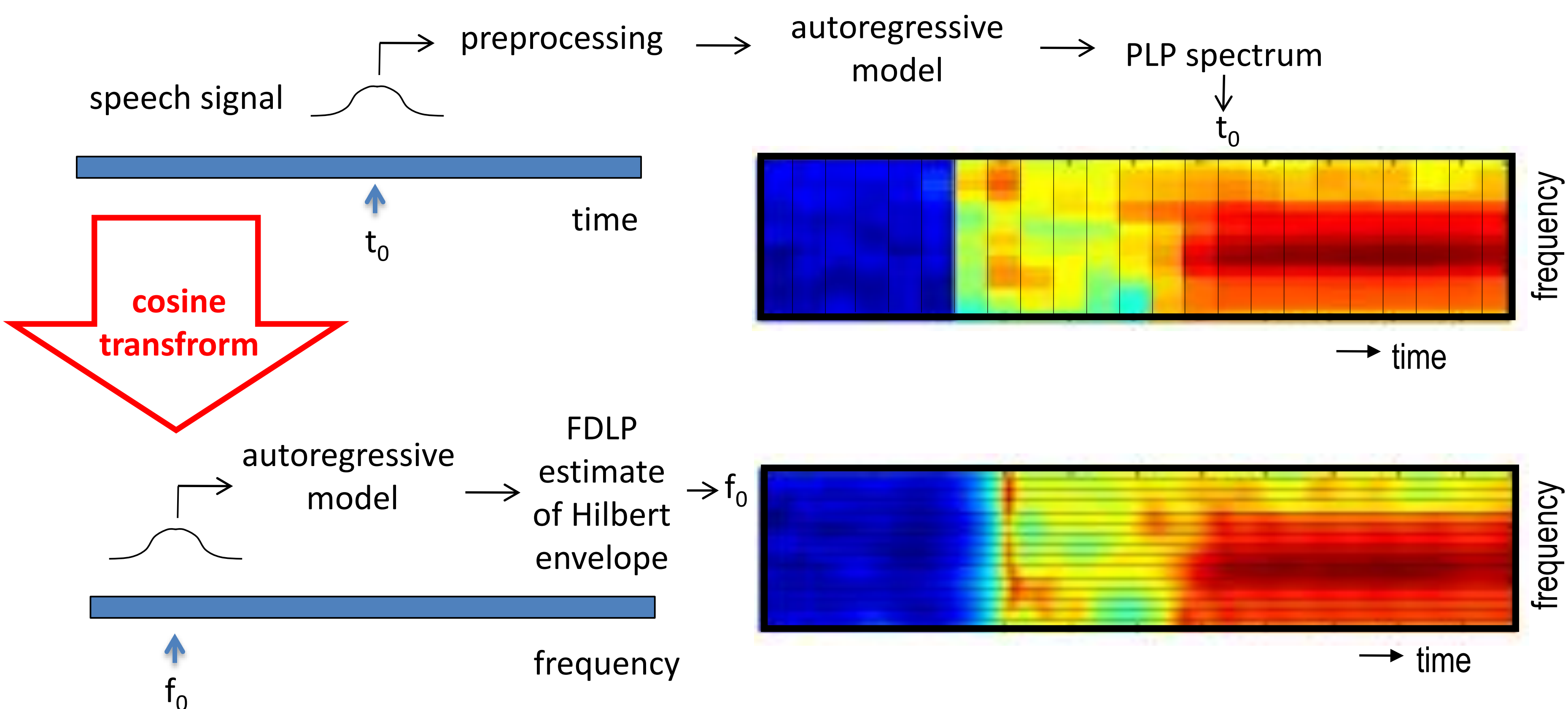
- with M. Athineos, D. Ellis (Columbia Univ), and P. Fousek (CTU Prague)



Frequency Domain Linear Prediction (FDLP)

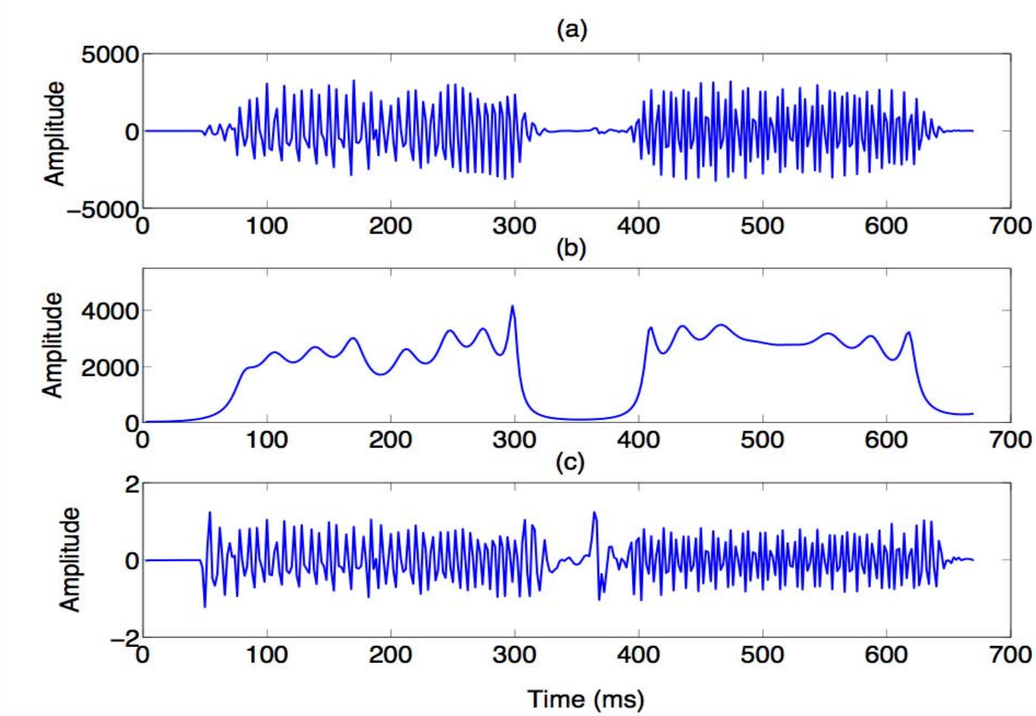
FDLP

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



Source: Athineos, Marios, Hynek Hermansky, and Daniel PW Ellis. "PLP $^ 2$: Autoregressive modeling of auditory-like 2-D spectro-temporal patterns." In Workshop on Statistical and Perceptual Audio Processing (SAPA), no. EPFL-CONF-83126. 2004.

Autoregressive model of Hilbert envelope of the signal

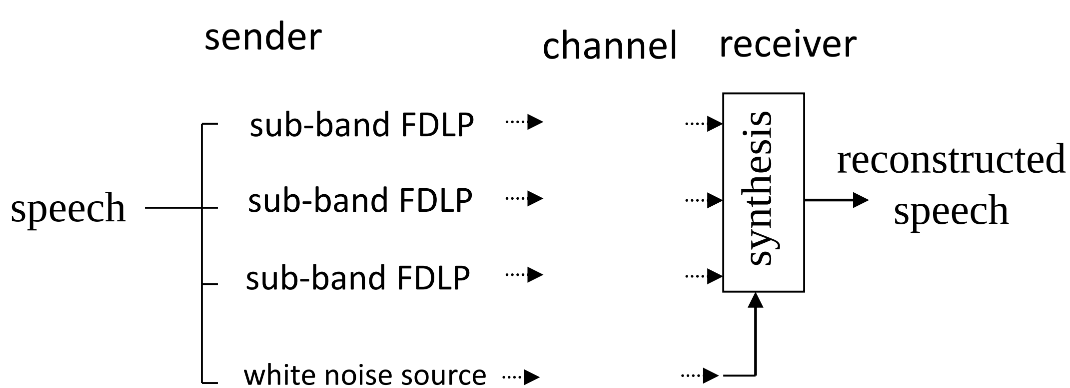


signal

AM component
(temporal envelope)

FM component
(carrier)

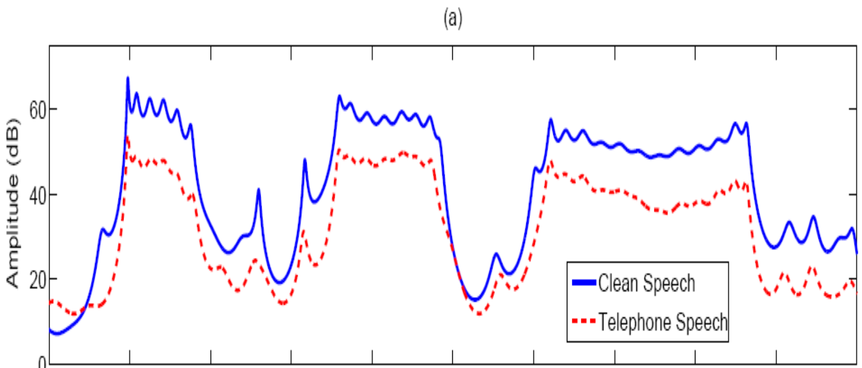
Uses channel vocoder (similar to the original H. Dudley design)



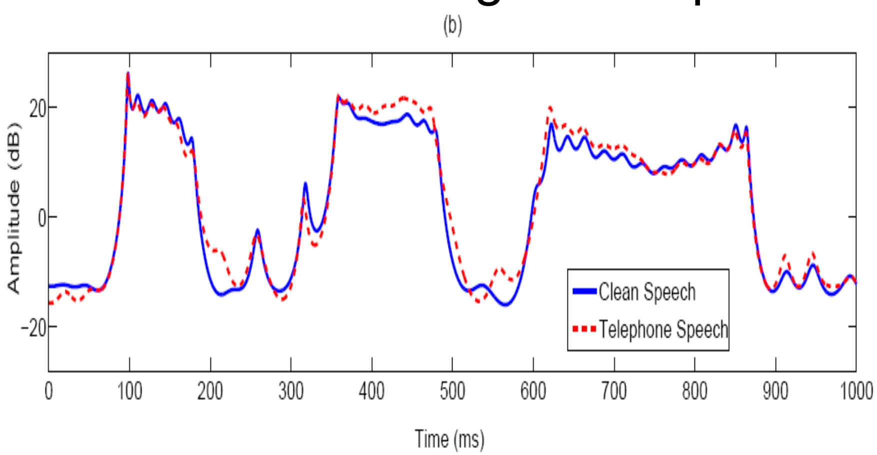
Varying communication channels
(convolution with a short impulse response of a channel)

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.

Source: Ganapathy, Sriram, Samuel Thomas, and Hynek Hermansky. "Temporal envelope compensation for robust phoneme recognition using modulation spectrum." The Journal of the Acoustical Society of America 128, no. 6 (2010): 3769-3780.

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

Gain of the AR model included

Figure removed due to copyright restrictions. Please see the video.
Source: Thomas, Samuel, Sriram Ganapathy, and Hynek Hermansky. "Recognition of reverberant speech using frequency domain linear prediction." IEEE Signal Processing Letters 15 (2008): 681-684.

Recognition accuracy [%]
-clean and reverberated (8 different room responses)
Aurora digits

	PLP	FDLP
clean	99.68	99.18
reveb	80.12	89.48

Improvements on real reverberations similar (Thomas, Ganapathy, Hermansky, IEEE Signal Processing Letters, Dec 2008)

Known noise with unknown effects

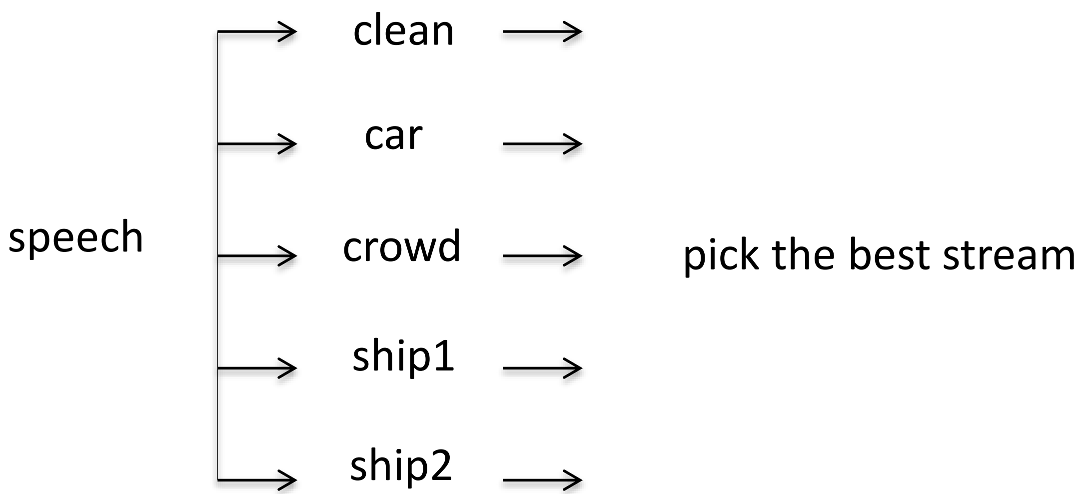
Dealing with unknown effects of known noise

training data with all noises that we do not know how to handle

speech signal → features → **machine** → wanted information phoneme error rates noisy TIMIT

train / test	clean	car	crowd	ship1	ship2
clean	**20.7**	34.2	59.2	65.7	64.9
car	23.8	**22.7**	58.1	65.2	64.6
crowd	30.8	33.1	**36.0**	38.1	44.9
ship 1	35.4	41.3	53.7	**35.6**	44.9
ship 2	37.0	45.4	58.3	45.0	**35.2**
multi-style	23.0	24.9	36.8	39.0	39.7

Mallidi et al *in preparation*



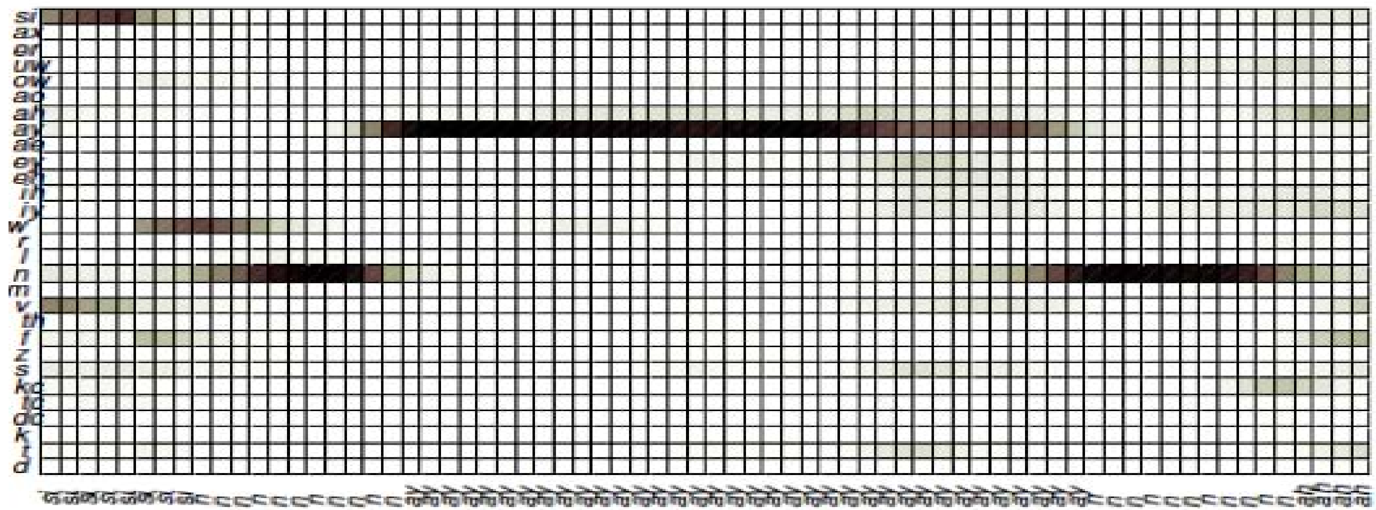
pick the best stream based on input

- recognize type of noise

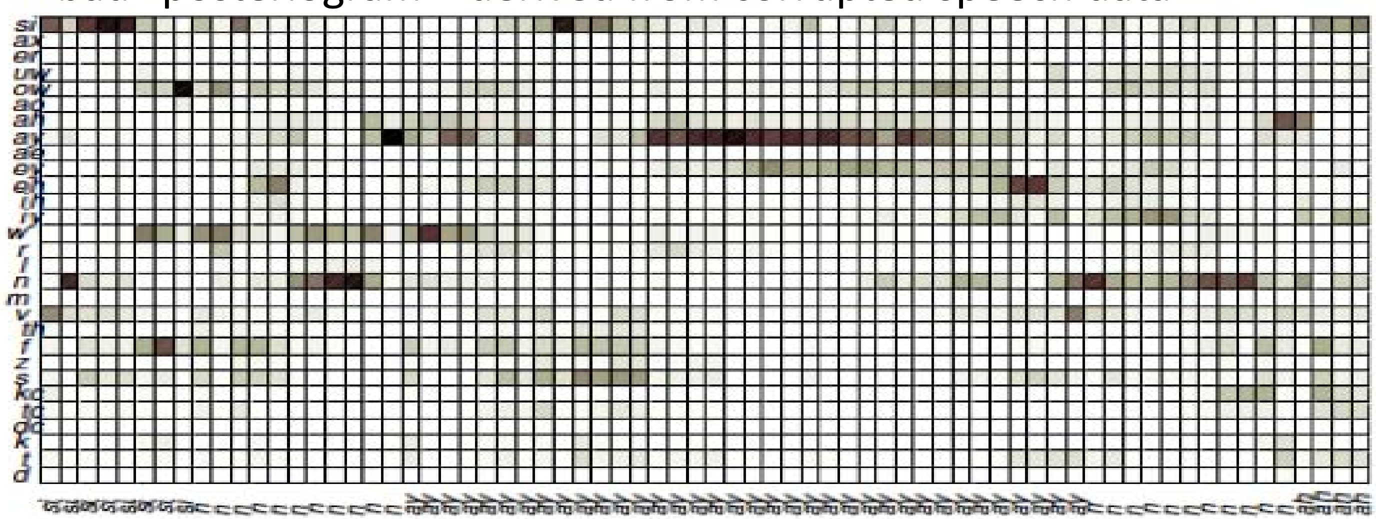
pick “the best” output

- what does “the best” mean ?

Do it fast (based on short segment of test data) "good" posteriogram – derived from speech data similar to its training



"bad" posteriogram – derived from corrupted speech data



The “best” probability estimates?

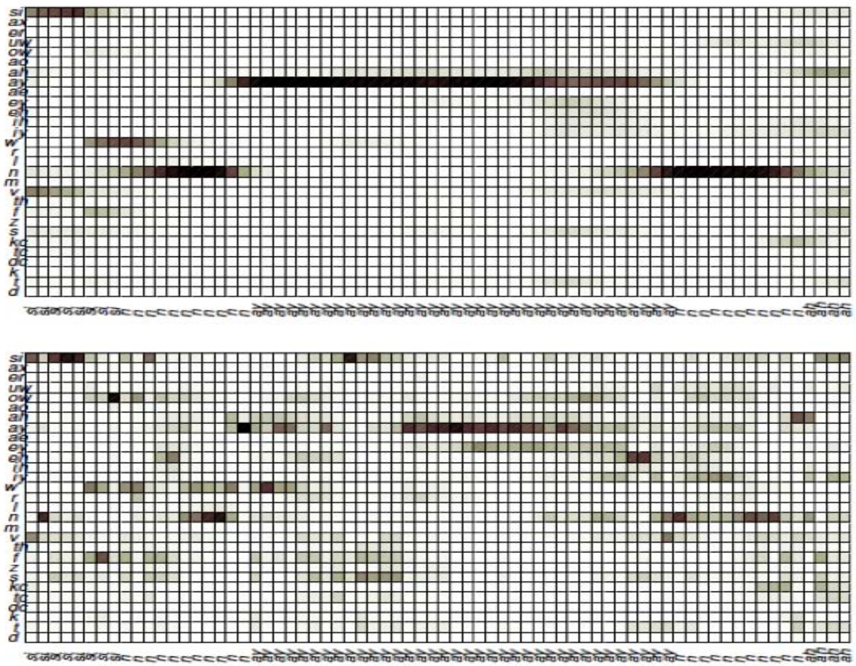
Ideally the ones which yield the lowest error

- do not know the correct answer so do not know the error

1. Estimates which yield “clean” posteriograms

2. “Similar” to ones derived on training data of the estimator

How “clean” is a posteriogram ?

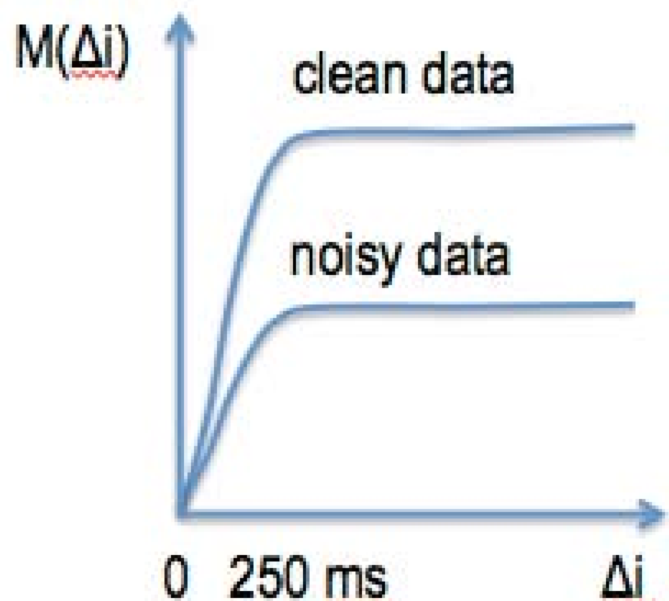


Δτ

$$M(\Delta\tau) = \frac{\sum_{i=0}^{N-\Delta\tau} D(\mathbf{p}_i, \mathbf{p}_{i+\Delta\tau})}{N-\Delta\tau}$$

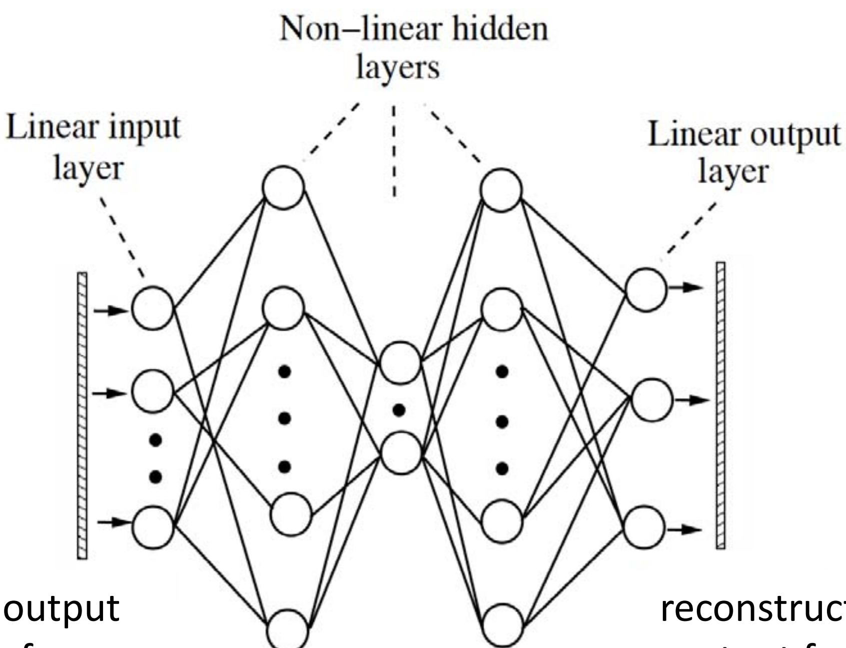
Δi – time delay

D(.) – symmetric Kl divergence



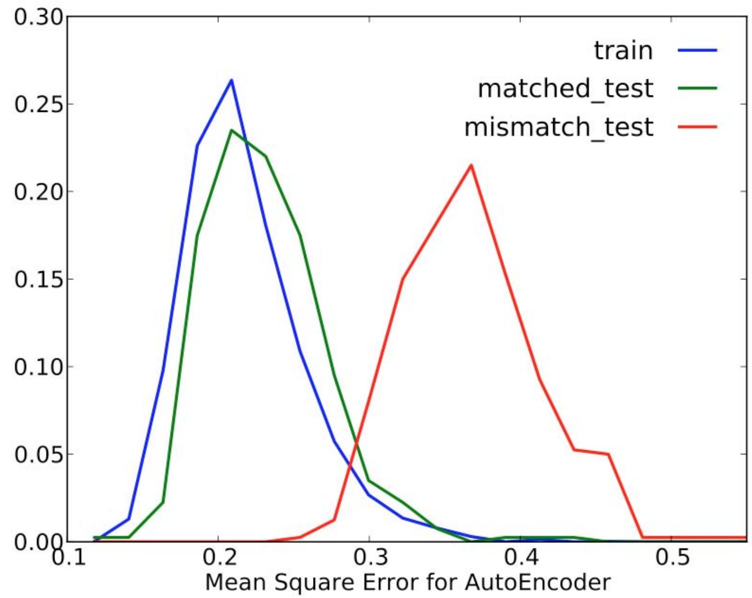
How “similar” is the estimator performance on its training data and in the test?

DNN autoencoder trained on output of the estimator when applied to its training data



output from probability estimator

reconstructed output from probability estimator



picking up good streams

phoneme error rates noisy TIMIT

train / test	clean	car	crowd	ship1	ship2
multi-style	23.0	24.9	36.8	39.0	39.7
matched	20.7	22.7	36.0	35.6	35.2
oracle	17.7	19.9	31.8	31.1	31.4
multi-stream with ` performance monitoring	**20.9**	**24.3**	**35.0**	**34.8**	**37.2**

Mallidi et al *in preparation*

Previously unseen noise

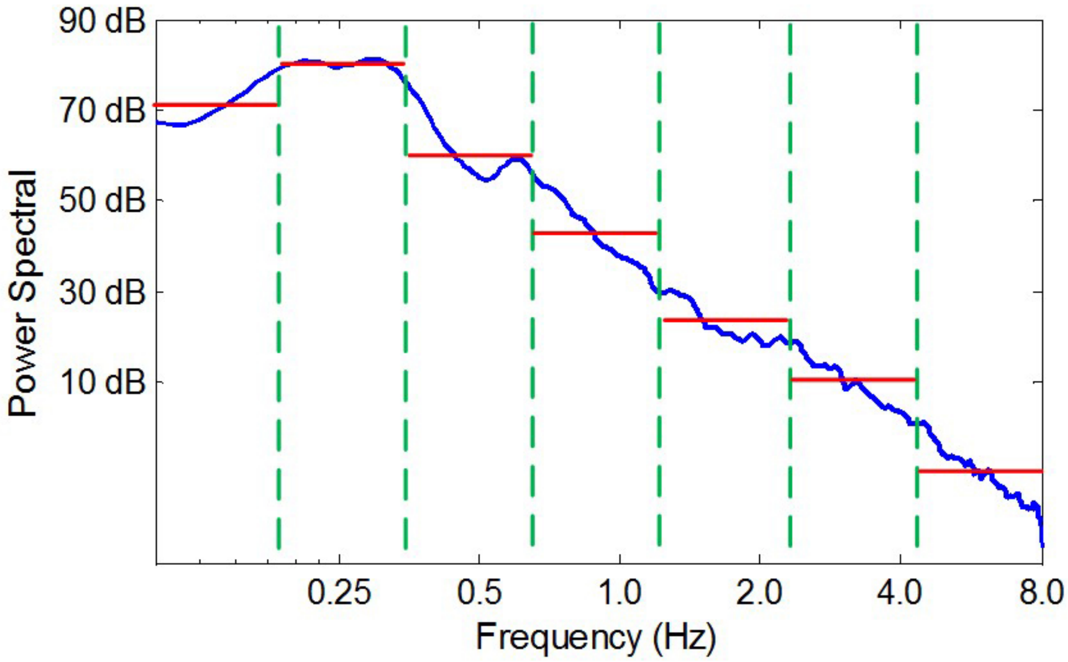
extrapolate from known noise training ?

phoneme error rates noisy TIMIT

train / test	clean	car	crowd	ship1	ship2	unseen noise f16 fighter
clean	**20.7**					**62.9**
car		**22.7**				62.7
crowd			**36.0**			41.4
ship 1				**35.6**		40.8
ship 2					**35.2**	44.8
multi-style	23.0	24.9	36.8	39.0	39.7	**36.3**
oracle	*18.4*	*20.5*	*34.7*	*34.5*	*34.8*	*29.1*
multi-stream	**20.9**	**24.3**	**35.0**	**34.8**	**37.2**	**32.5**

Mallidi et al *in preparation*

Divide et Impera



- unknown noise of arbitrary shape can be approximated by white noise of appropriate levels in individual frequency sub-bands.

1-3 Bark DNN →

4-6 Bark DNN →

7-9 Bark DNN →

speech → 10-12 Bark DNN → fusion DNN → *final posterior estimates*

13-15 Bark DNN →

16-18 Bark DNN →

19-21 Bark DNN →

all neural nets (DNNs) trained on clean, 20 dB, 10 dB , 5 dB SNR **white** noise

Word error rates (Aurora 4)

	test > 30 dB SNR	test 10 dB SNR	test 5 dB SNR	unseen test noise (car)
training > 30 dB SNR	**3.10 %**	15.65 %	36.60 %	13.62 %
training 10 dB SNR	5.06 %	**4.35 %**	14.70 %	7.47 %
training 5 dB SNR	9.04 %	4.73 %	**7.73 %**	7.86 %
multistyle training >30, 15, 10 ,5 dB	4.28 %	5.17 %	11.86 %	8.11 %
sub-band multistream	**2.99 %**	**3.23 %**	**10.18 %**	**4.30 %**

Unexpected noise

Adaptation

- Modify classifier during its operation to better deal with new previously unseen conditions
  - Assemble new classifier on-line from reliable parts of the old one to improve performance on new data?
  - Assumptions
    - some parts of the old classifier remain reliable
    - measure of classifier performance is available

Multi-band processing

Subdivide speech spectrum into independent processing streams for further processing

speech →

1st frequency range →

• →

• →

• →

• →

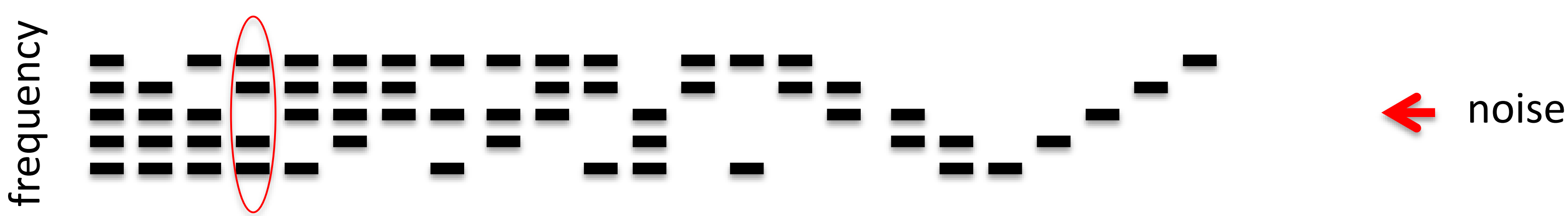
• →

Ntht frequency range →

performance monitor chooses the best stream combination → *final posterior estimates*

5 frequency bands - 31 ways to combine them

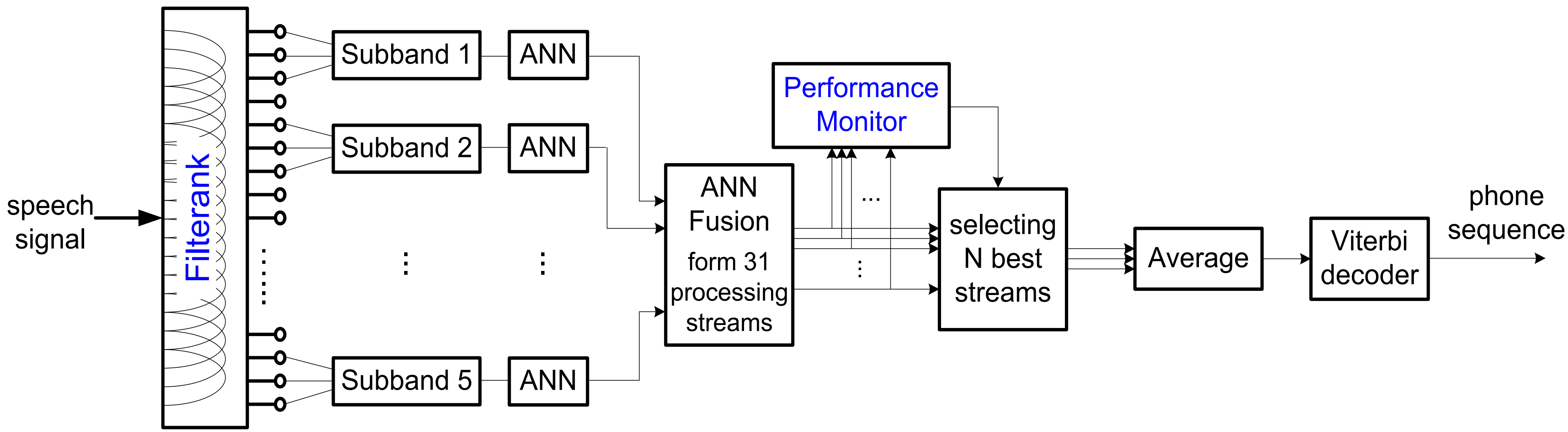
– 31 processing streams, each covering different frequency ranges of the full spectrum



Multi-band processing with performance monitoring

Variani et al, Interspeech 2013

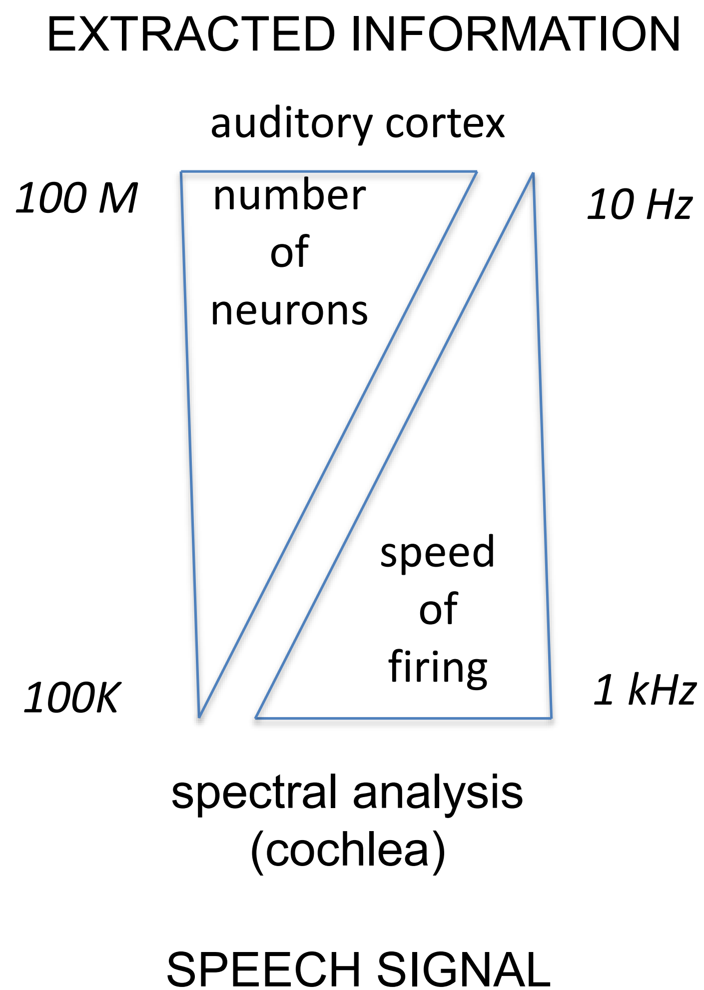
- All processing streams trained on clean speech



Phoneme recognition error rates

environment	conventional	PM	oracle
clean (matched training and test)	31 %	28 %	25 %
TIMIT with car noise at 0 dB SNR (training on clean)	54 %	38 %	35 %

human auditory processing



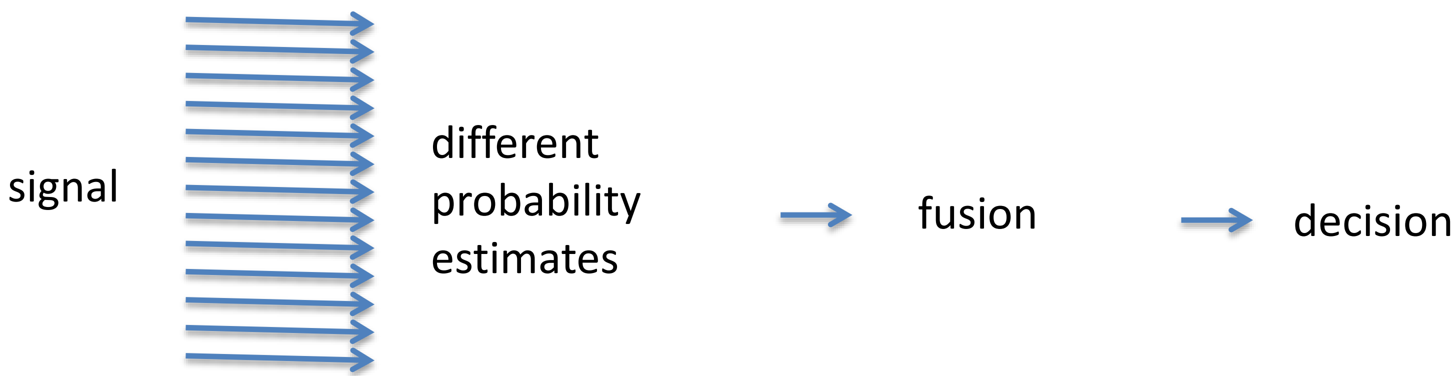
linguistic code (~ 50 b/s)

perceptual and cognitive processes

SPEECH SIGNAL (> 50 kb/s)

many ways of describing the information on higher levels of perception !

Multi-stream Processing



Stream formation

- differently trained probability estimators
- different aspects of the signal
- different modalities
- different strenghts of priors

Fusion

- select “the best” probability estimates

Conclusions

- Predictable effects of noise (e.g., linear distortions) are relatively easy to deal with by signal processing techniques that emulate perception of modulations in signal

- Unpredictable effects of noise, typically handled by multi-style training, could be better handled by a bank of parallel “expert” processing streams that emulate hypothetical parallel processing channels in hearing