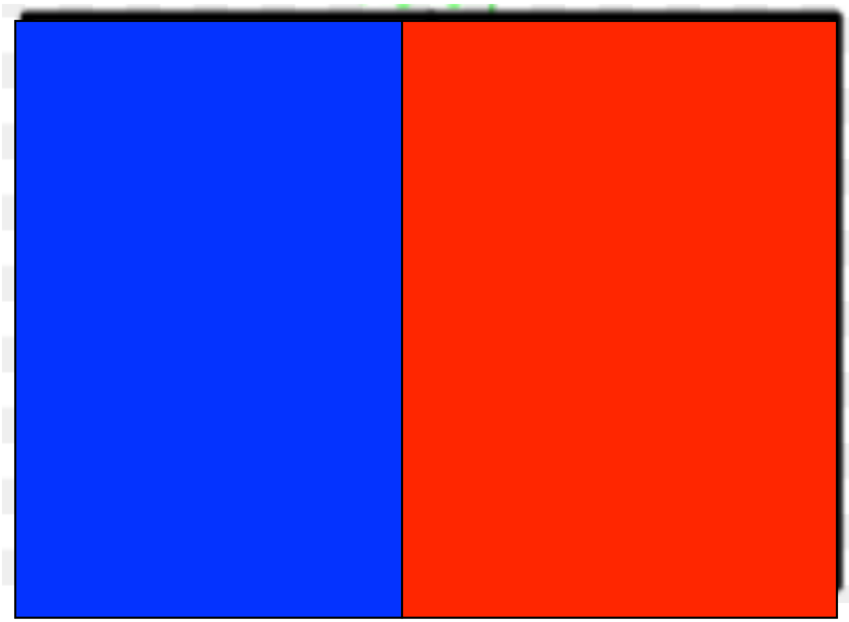
The Center For Language
and Speech Processing
at the Johns Hopkins University

When you can't beat them, join them

How I learned to stop worrying and started to love the machine

Hynek Hermansky

Maxwell demon



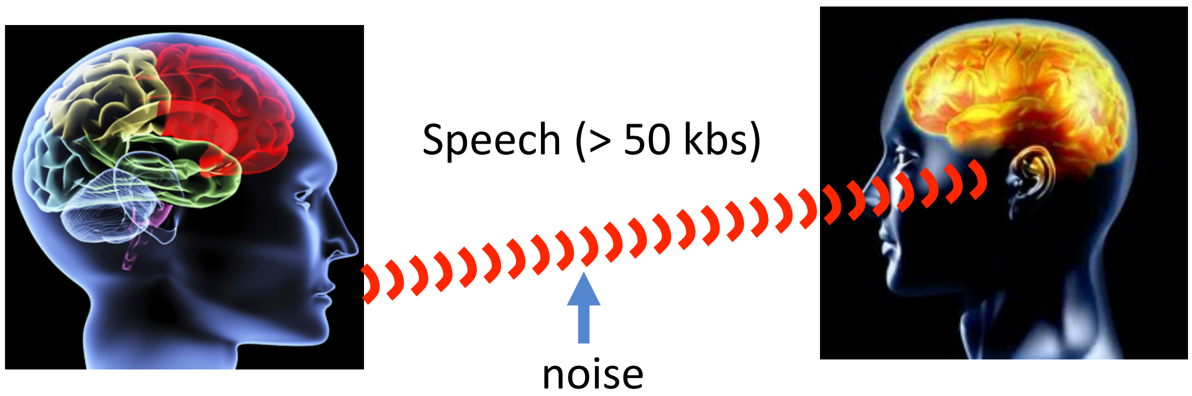
LOW ENTROPY

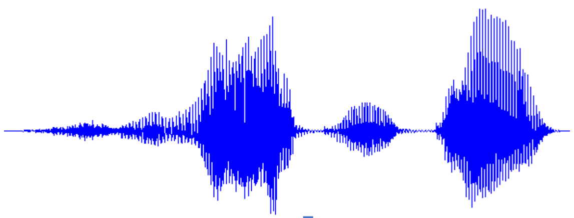
The Demon closes door when a slow air molecule comes and lets the fast air molecules to go through

When decreasing entropy, one needs to know what one is doing!

The Demon must KNOW which molecule is fast and which is slow!

Message (<50 bps) Message (<50 bps)





machine

message

> 50 kb/s $C = W\log_2(S/N+1)$, $W=5kHz$, $S/N+1>10^3$

message and its coding redundancy,

who is speaking, emotions, accent, acoustic environment, ….

< 50 b/s < 3bits/phoneme, < 15 phonemes/s

message

KNOWLEDGE

- magic
- experts, beliefs, previous experience
- measurements (data)

HARDWIRED

Reusable permanent knowledge

but

Experts and beliefs can be wrong

Wrong knowledge is worse that no knowledge

FROM DATA

Data do not lie

but

Transcribed data are expensive

No need to re-learn known facts

Bad data are worse than no data

More reliable knowledge hardwired, less training data needed

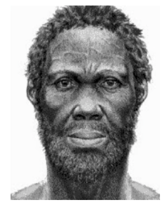
When using ”knowledge”, then **which** knowledge?

Environment
(survival)

Hearing
(communication)

Evolution of hearing

200 000 000 years



Evolution of speech

200 000 000 years

Pristerodon

Homo sapiens

We hear to survive

…. sensory neurons are adapted to the statistical properties of the signals to which they are exposed.

Simoncelli and Olshausen

We speak to hear

We speak in order to be heard and need to be heard in order to be understood.

Jakobson and Waugh p.95

Human speech evolved to fit properties of human hearing

LEARN FROM HUMAN HEARING

WHAT?
(recognize message)

HOW?
(human hearing)

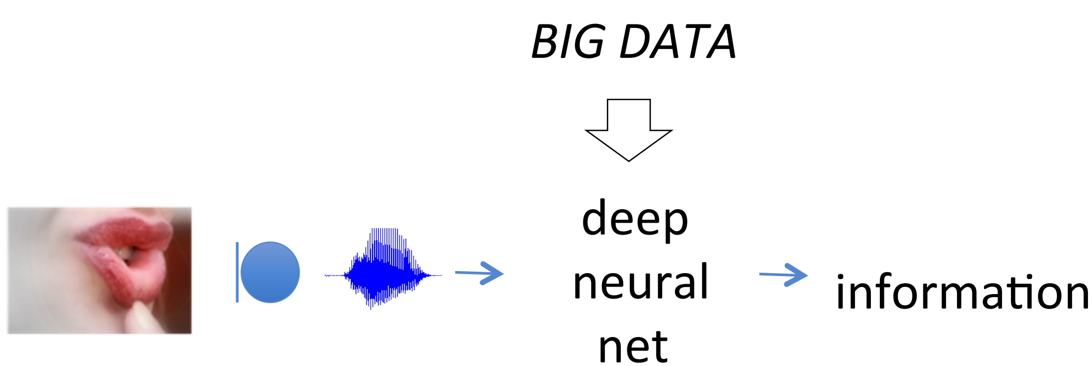
WHY ?

More data is always better than more thinking

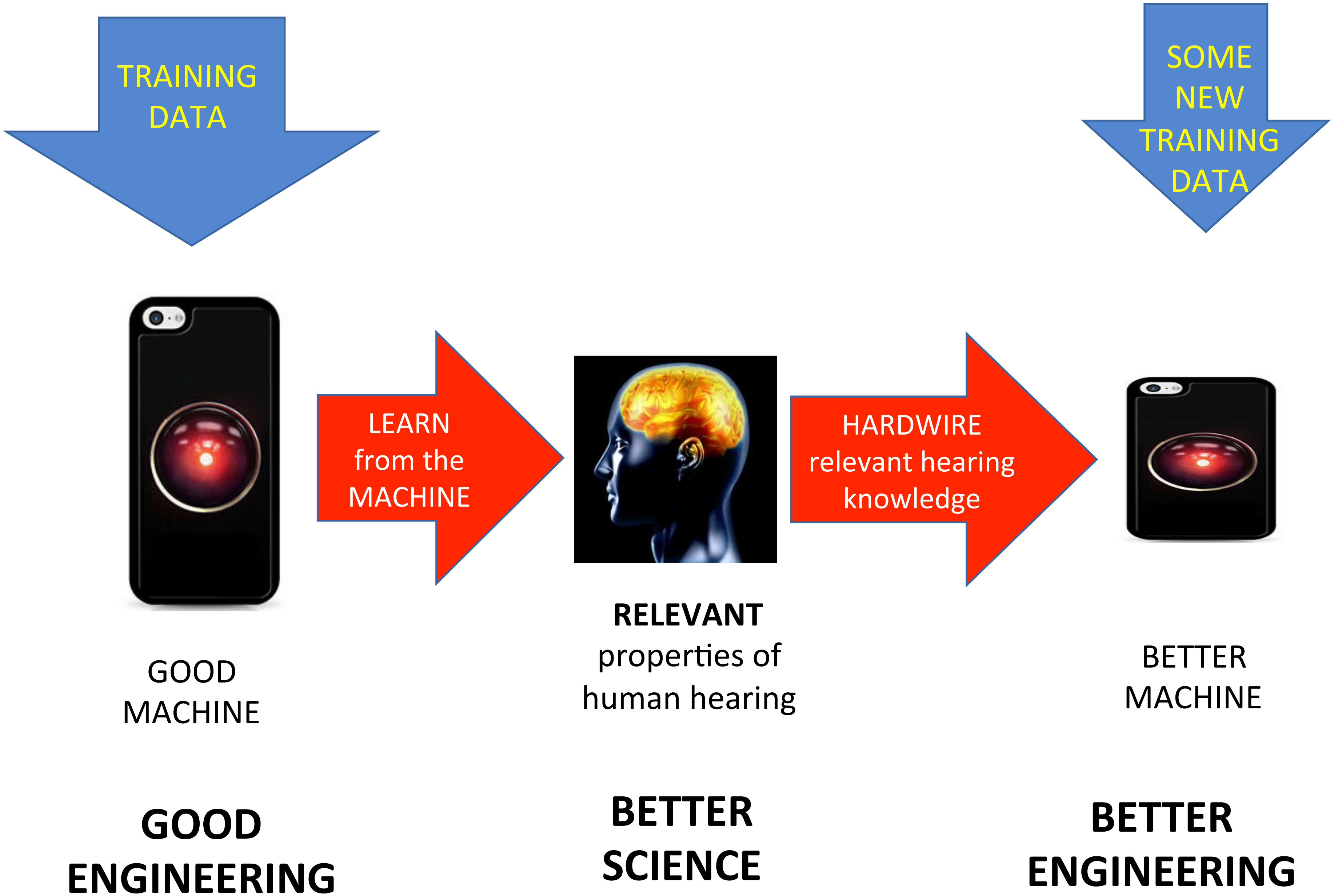
- Fred Jelinek (attributed to Eric Brill)

Artificial Neural Networks

- Discriminative nonlinear classifiers introduced to ASR in late eighties of 20^{th} century
- Fewer restrictions on form of input features
- Current hardware advances allow for new revolutionary approaches to ASR



When you can’t beat them, join them!

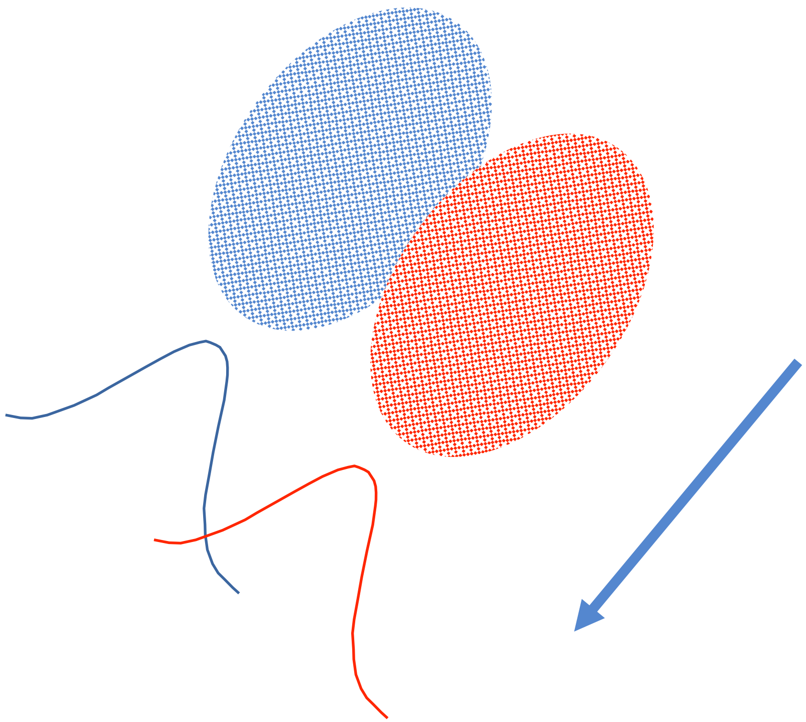

TRAINING DATA
SOME NEW TRAINING DATA
LEARN from the MACHINE
HARDWIRE relevant hearing knowledge
GOOD MACHINE
RELEVANT properties of human hearing
BETTER MACHINE
GOOD ENGINEERING
BETTER SCIENCE
BETTER ENGINEERING

Let’s assume

- Linguistic messages are represented by sequences of speech sounds (context-dependent or context-independent, senones,…)
  - not everybody agrees but …..
- Very large amounts of speech data labeled with speech sounds are available
  - hand labeled, transcribed with force alignment,…

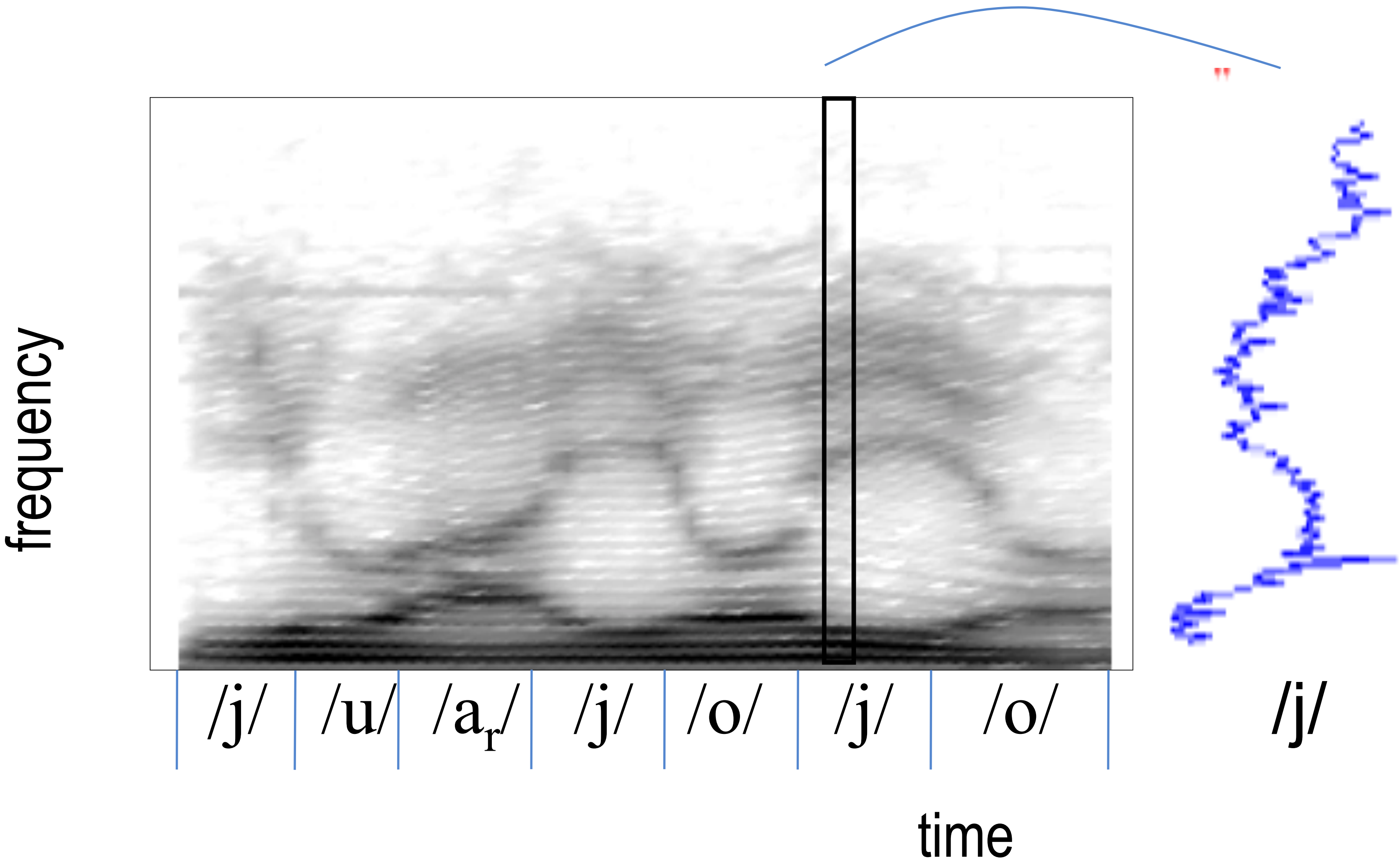
Linear discriminant analysis (Ronald A. Fisher 1936)

- find such projection of vectors of data, which preserves most of the discriminability
- data vectors need to be labeled by classes to be discriminated among
- yields matrix of discriminant vectors, ordered by their discrimination power
- discriminants are linear and therefore can be easily interpreted



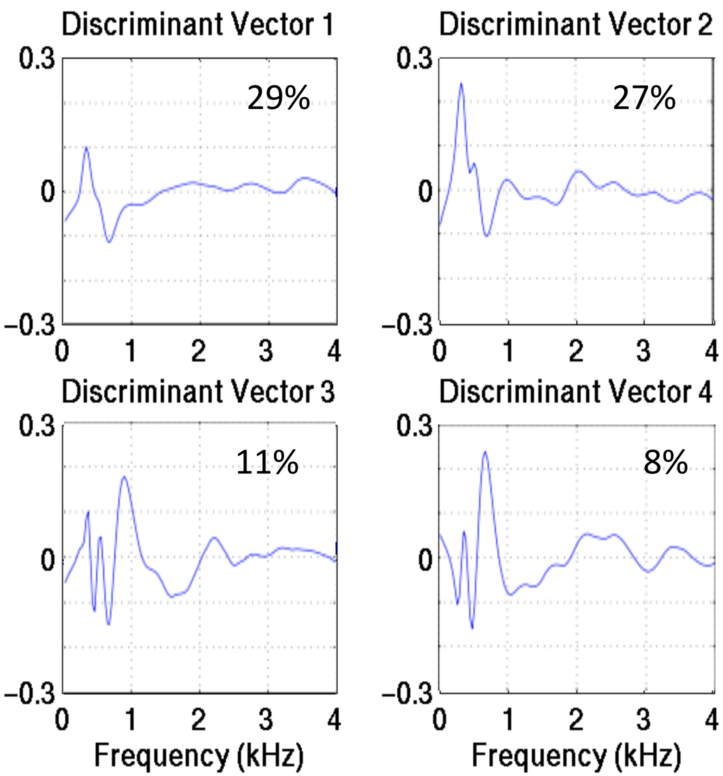
Spectral processing of short-time speech spectrum

(with Naren Malayath 1998)

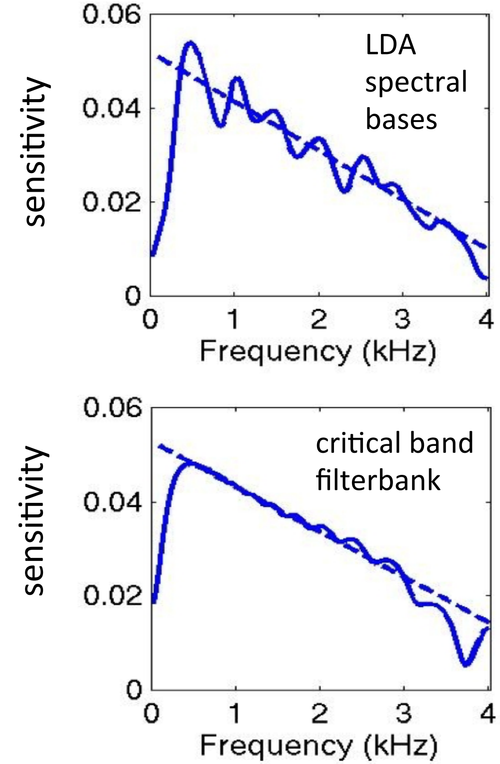


LDA-derived spectral bases

(30 hours of continuous telephone speech database – automatic labeling)



sensitivity to spectral perturbations



Malayath and Hermansky 1998**, Valente and Hermansky 2006**

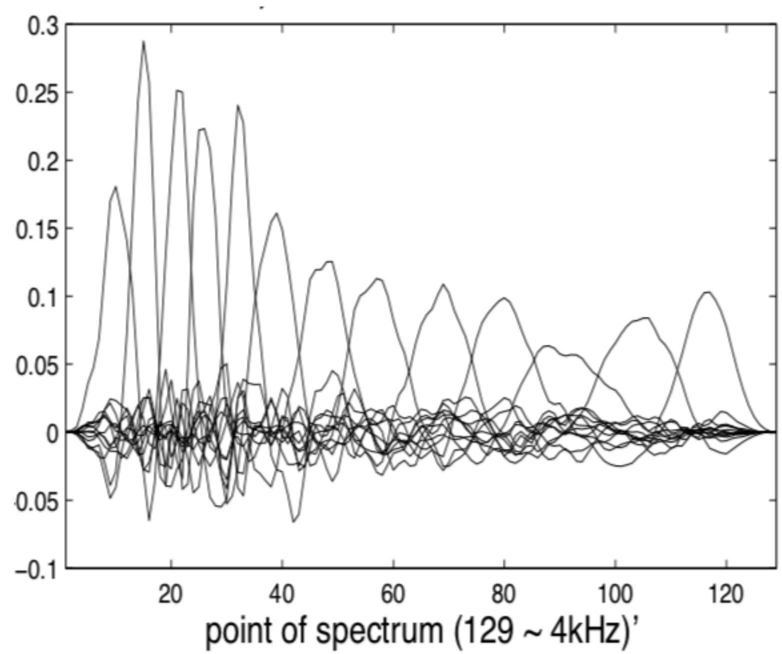
Similar observations using different optimization techniques

Biem and Katagiri 1994, Cohen et al 1996, Kamm et al 1997, Palival et al 1997, Burget and Hermansky 2001

1. Derive truncated matrix M by keeping only the LDA-derived bases with high eigenvalue discriminants

2. Compute the pseudoinverse M^+ of the truncated discriminant matrix M

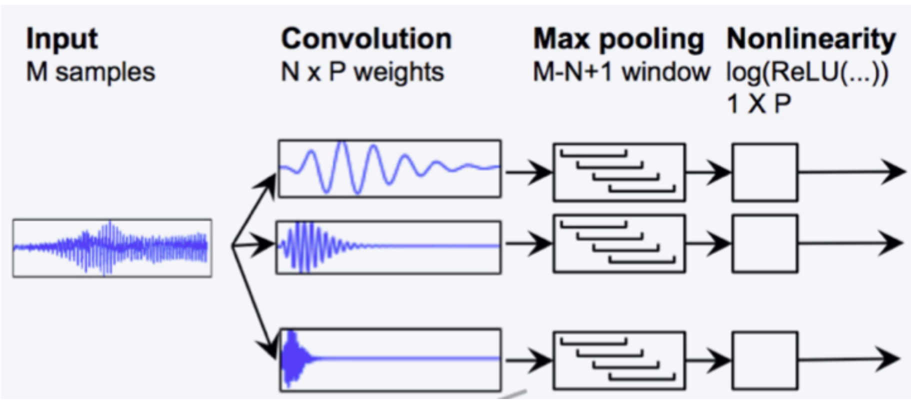
3. The product $M\,M^+$ represents weightings (filters) applied to the spectrum



Burget and Hermansky TSD 2001
Data driven design of filter bank for speech recognition

Using deep neural net classifiers, filters directly from speech signal

Sainath et al ASRU 2013



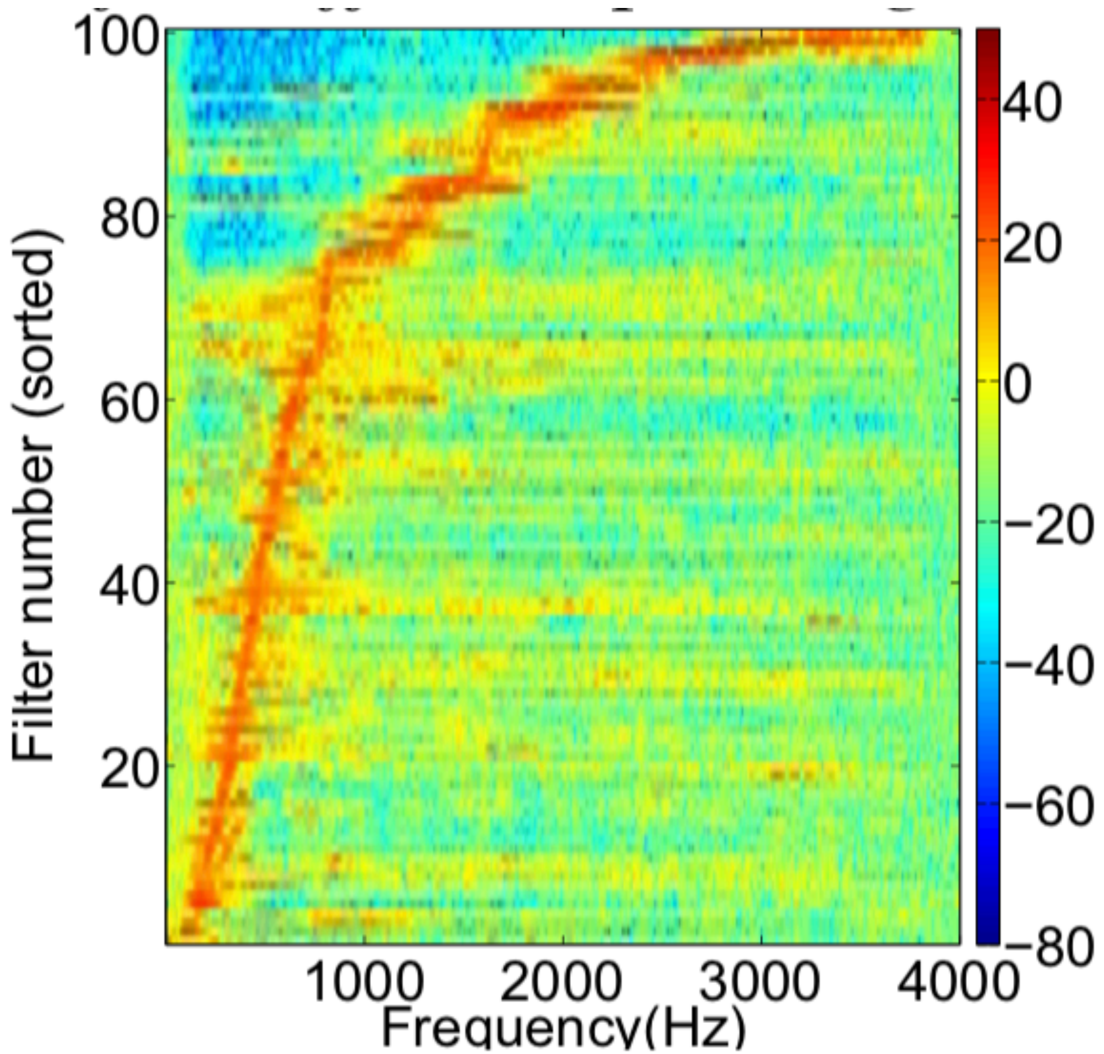
DNN-based ASR

Q: what are the learned weights in the convolution input layer?
A: impulse responses of filters consistent with critical bands of hearing

also Palatz et al 2013, Tueske et al 2014, Golik et al 2015, Gharemani et al 2016, Luo and Mesgarani 2918, …

Magnitude response of learned filters ordered by center frequency

Gharemani et al 2016



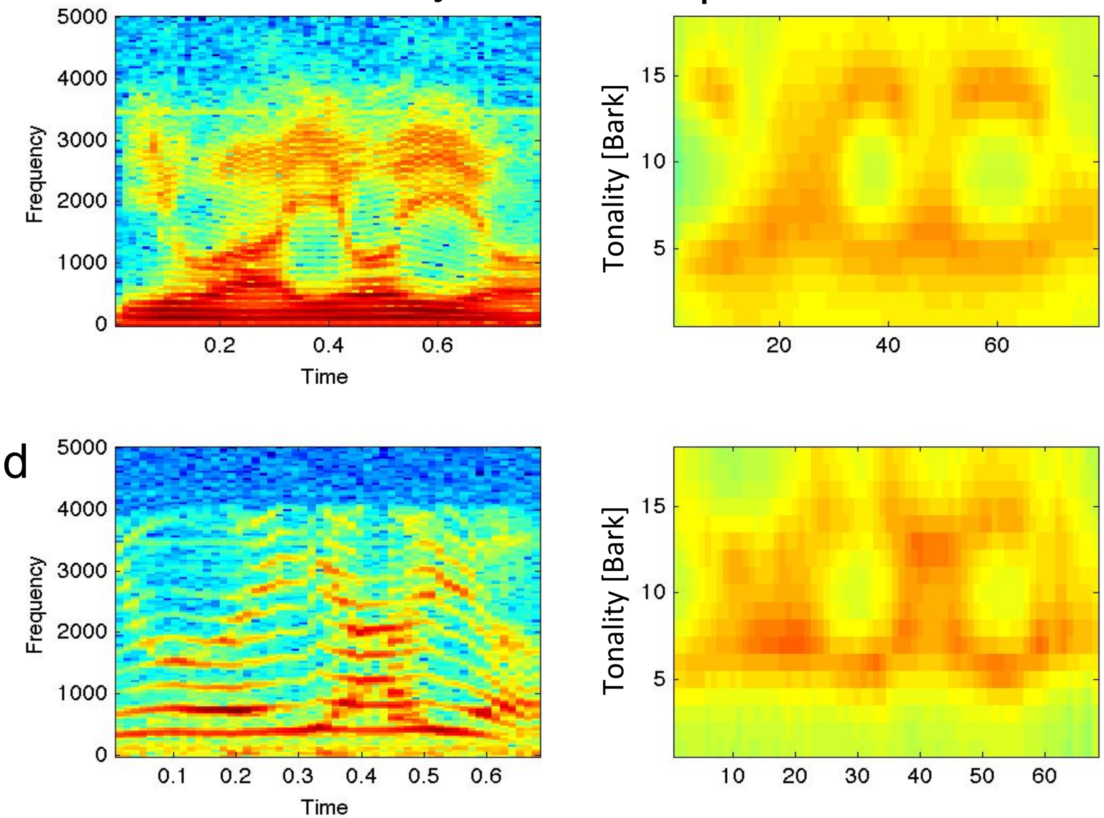
Effect of auditory-like spectral resolution

short-term spectrum from FFT analysis

spectrum with auditory-like spectral resolution

adult male

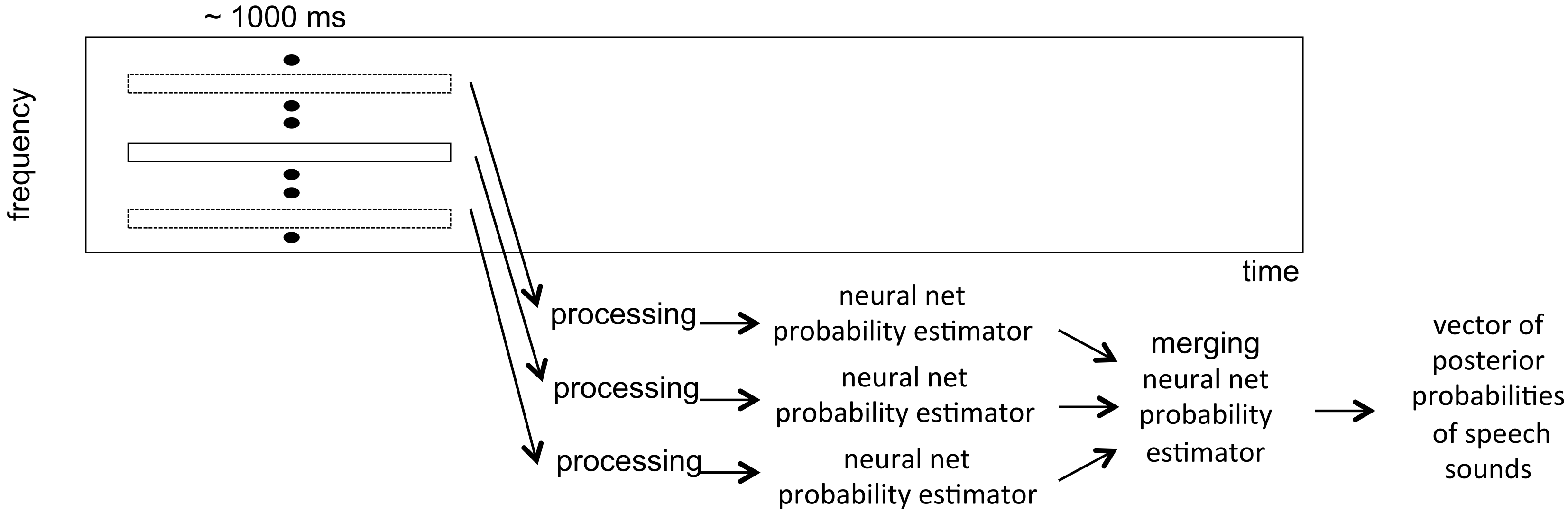
4 year old child



TRAPS

Hermansky and Sharma, ICSLP 1998

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

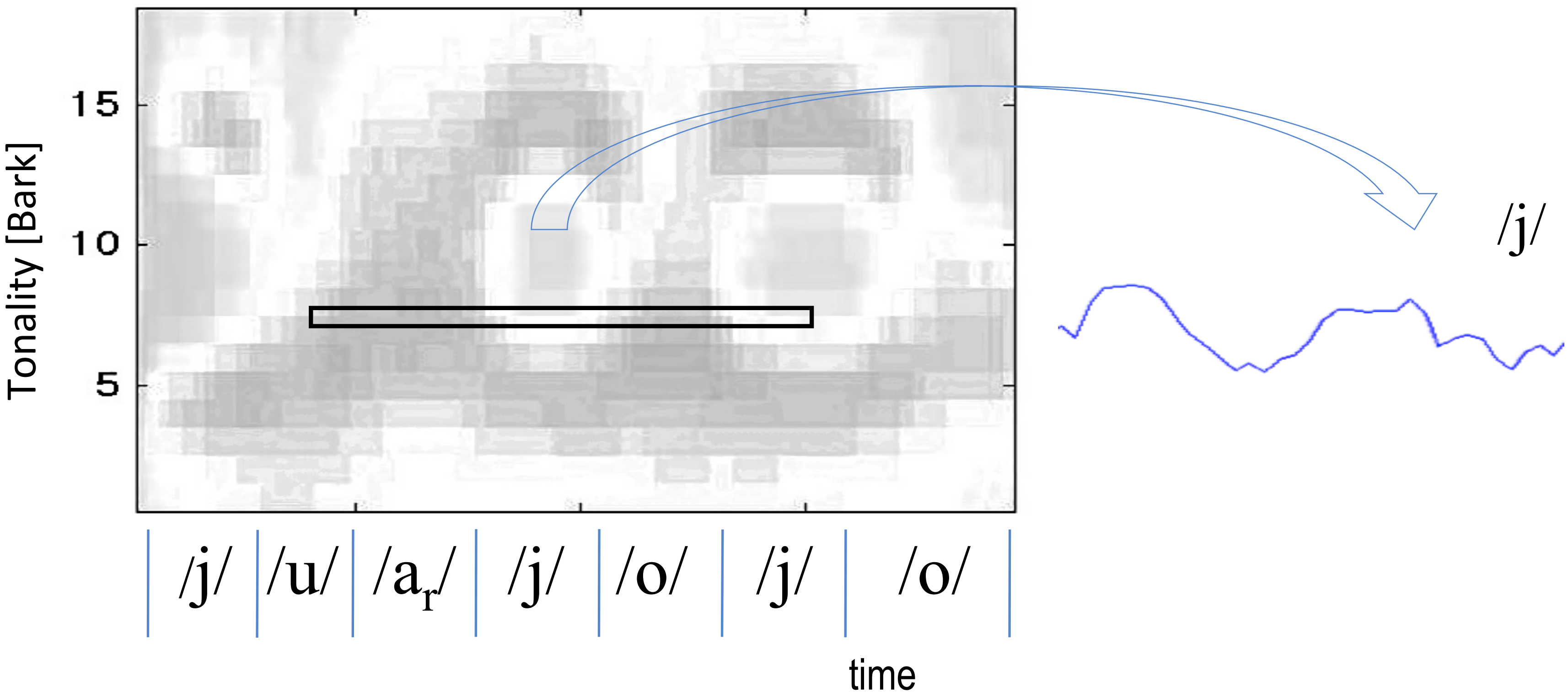


Some "novel" (in 1998) elements of TRAPS

- 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

Temporal processing of auditory-like speech spectrum

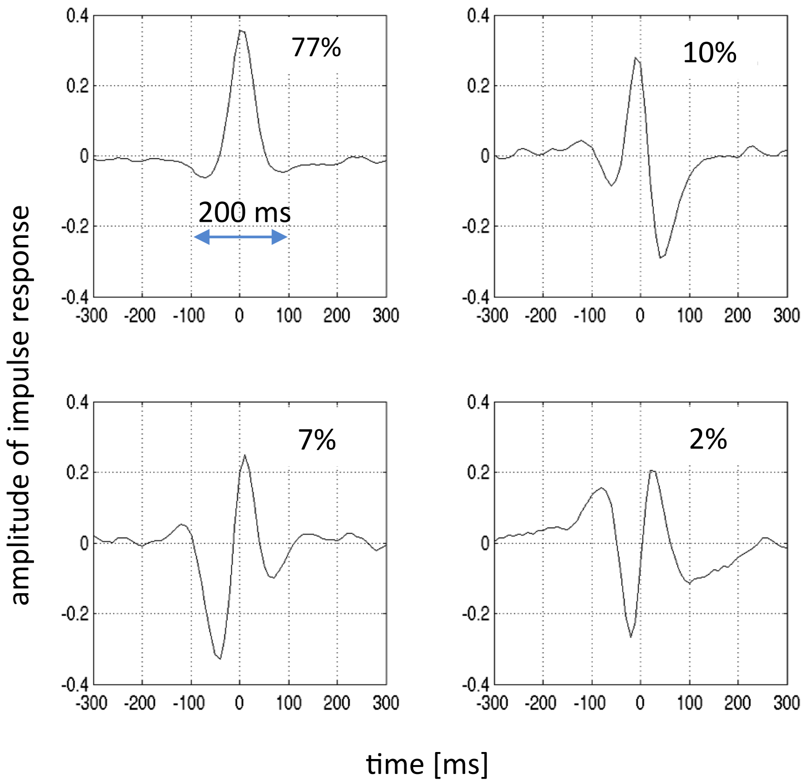
van Vuuren and Hermansky 1997, Valente and Hermansky 2006



LDA-derived FIR filters
(30 hours of continuous telephone speech database – automatic labeling)

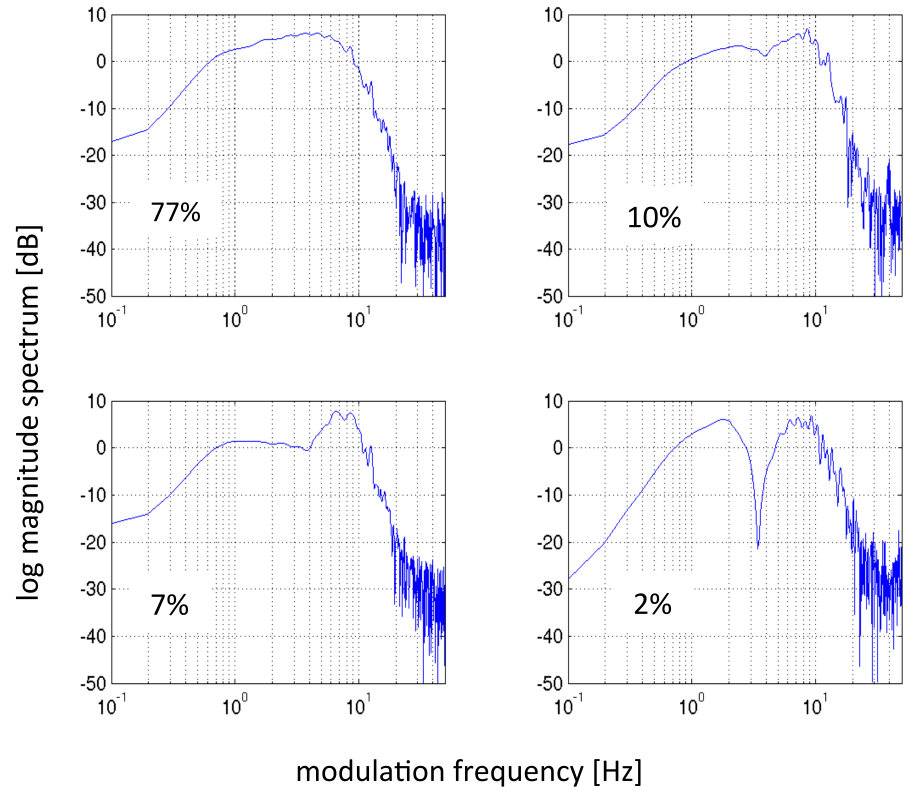
impulse responses

active parts of impulse responses > 200 ms



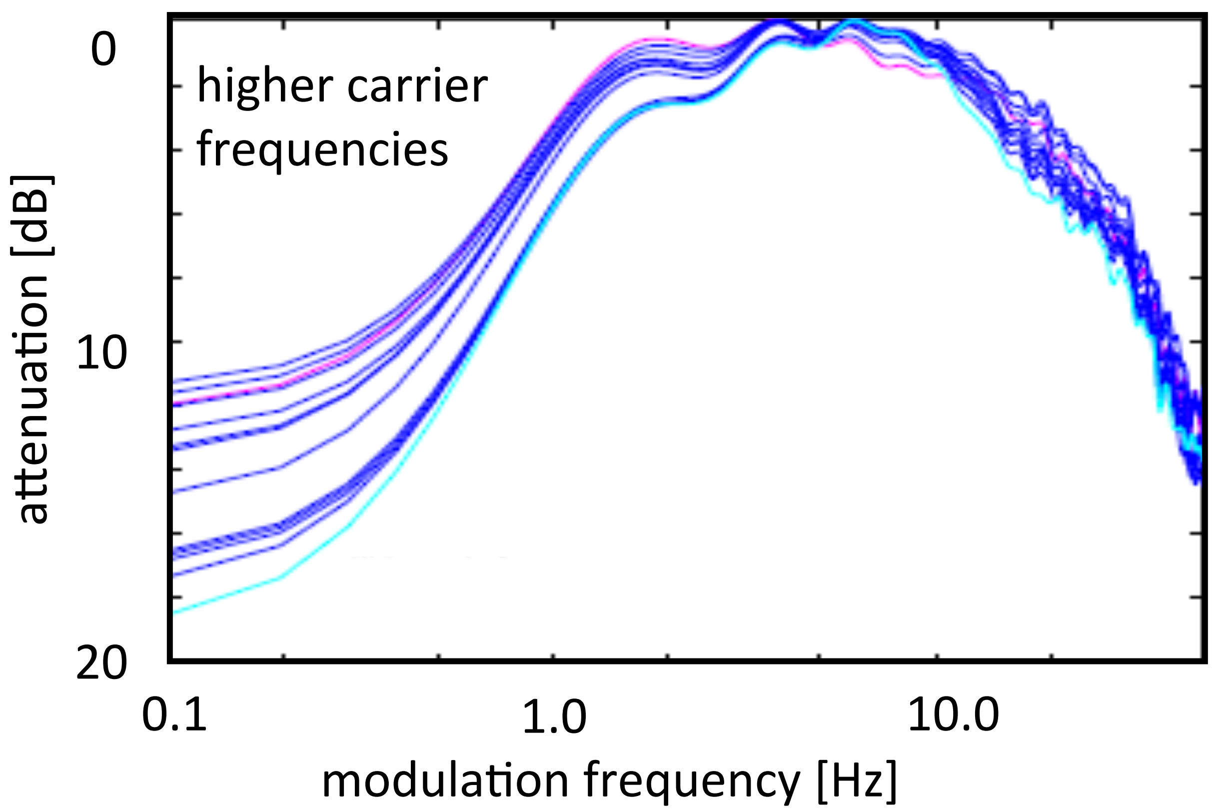
frequency responses

band-pass roughly 1-10 Hz



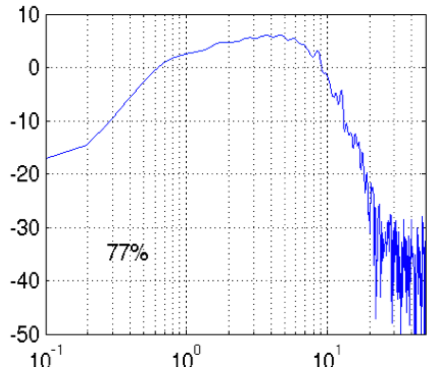
van Vuuren and Hermansky 1997, Valente and Hermansky 2006

frequency responses
(1^{st} discriminant in all frequency channels)

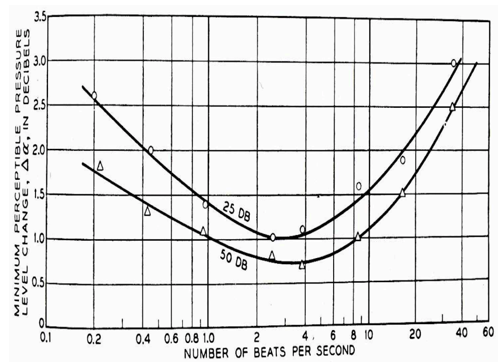


Modulation filters are very similar at all carrier frequencies

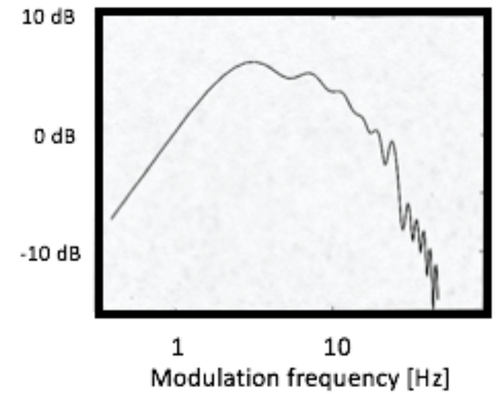
Frequency response of the 1st temporal discriminant



Sensitivity of human hearing to modulations (Riesz 1928)

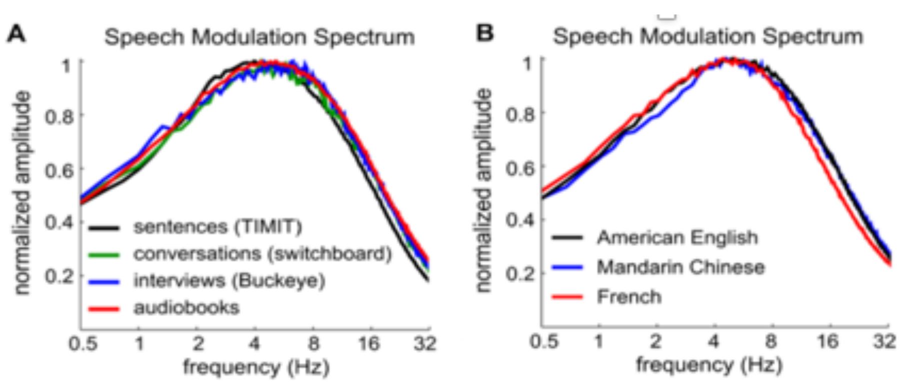


Frequency response of the 1st temporal principal component of about 3000 cortical spectro-temporal receptive fields (ferret)



Mahesan, Mesgarani, Hermansky (in preparation)

Modulation spectra of speech



Ding, Patel and Poeppel 2015

Optimizing temporal processing for discrimination among speech sounds yields filters, which are consistent with temporal properties of mammalian hearing.

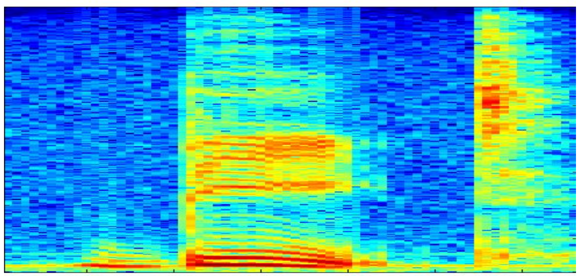
Enhancing message-carrying spectral components

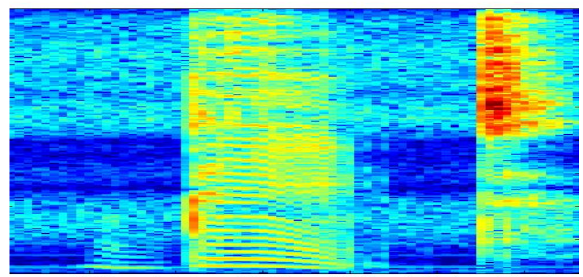
Hermansky and Morgan 1990

original speech

linear distortions
(stationary filter)

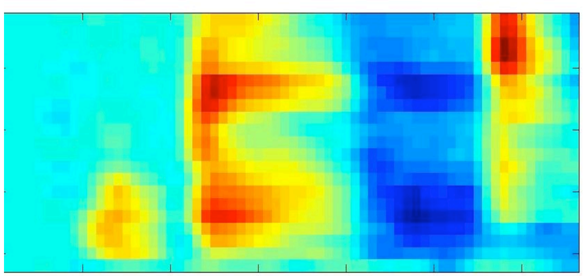
spectrogram

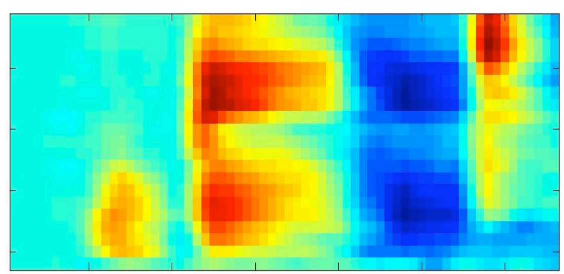




frequency

auditory-like spectrogram after band-pass filtering of spectral trajectories





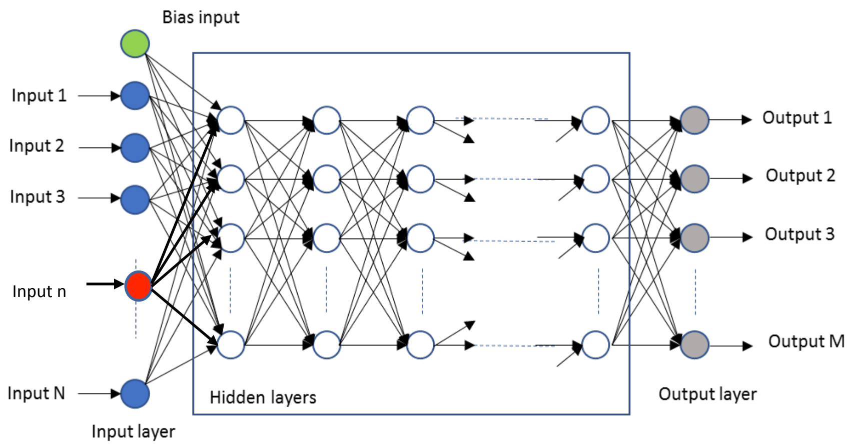
time

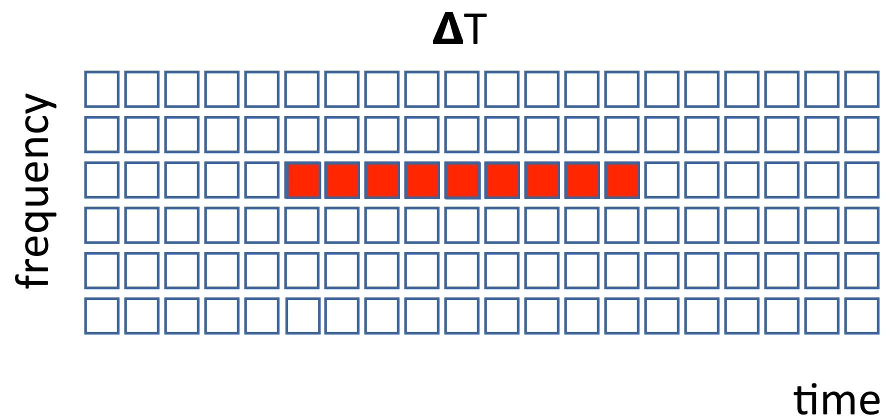
recognizer trained on data from New Jersey Labs

tested in New Jersey
2.8 % error
tested in Colorado
60.7 % error

tested in New Jersey
2.2 % error
tested in Colorado
2.9 % error

DNN-based design of linear pre-processing



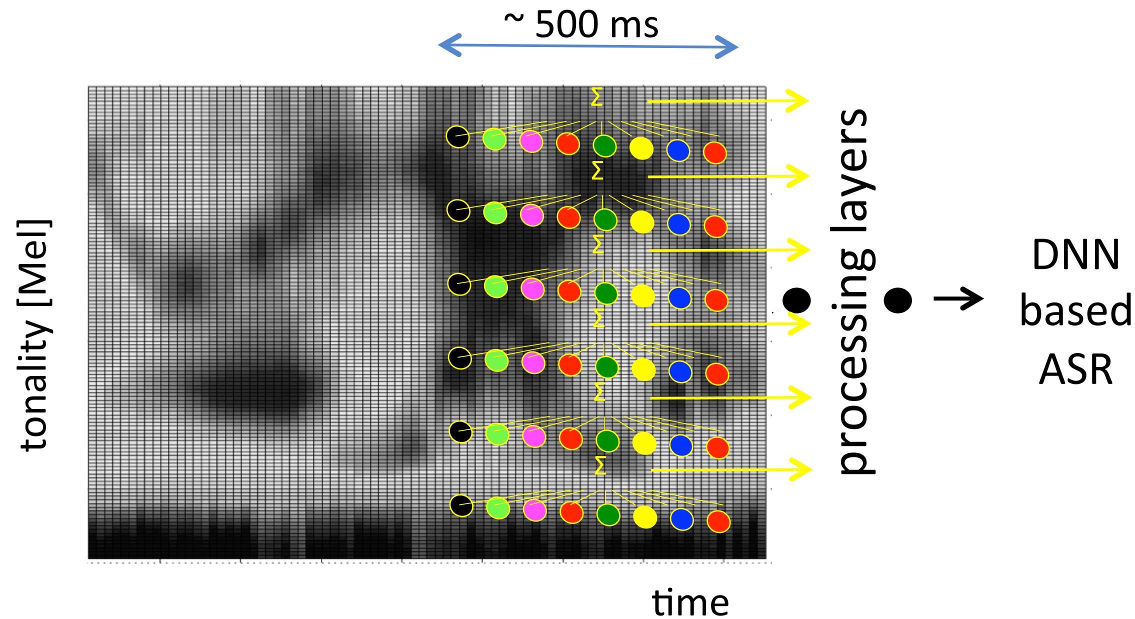


input n   modulation frequency filters

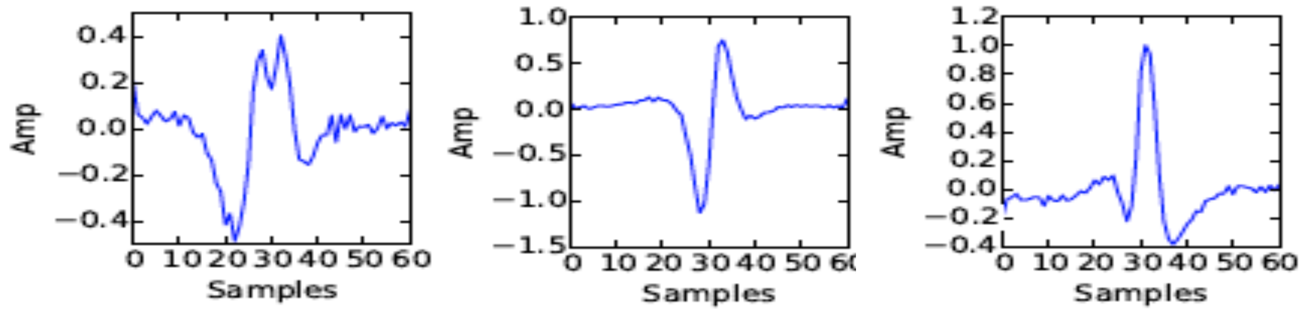
weighted sum of spectral values at frequency n within a time window **Δ**T
(weights optimized with the rest of the DNN weights)

DNN-based learning of modulation frequency filters

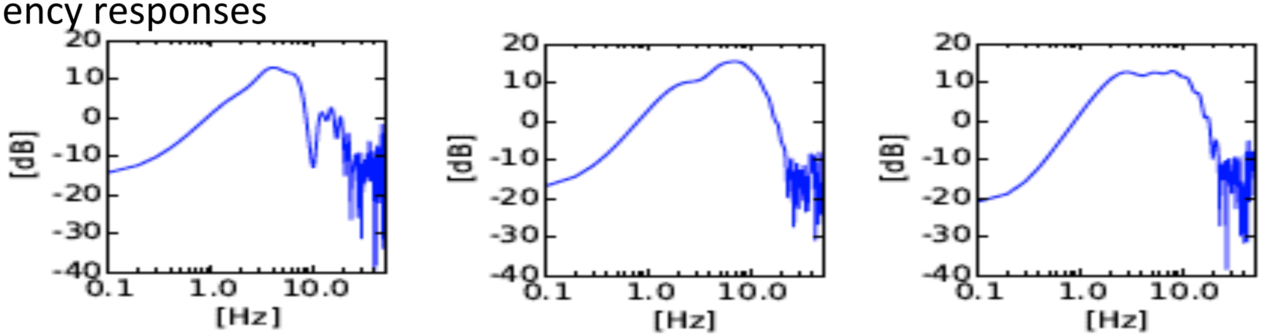
Pesan et al 2015



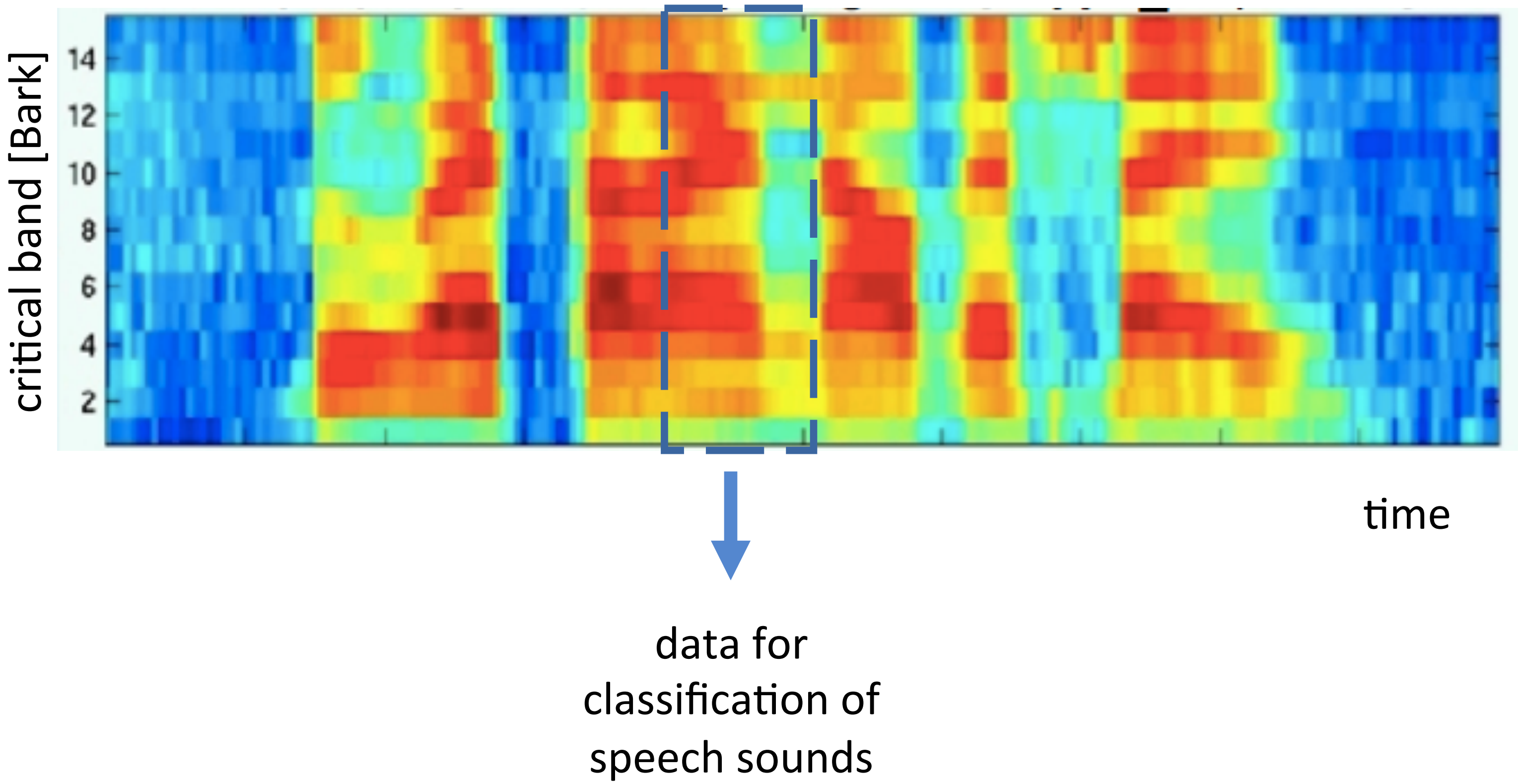
filter impulse responses



frequency responses

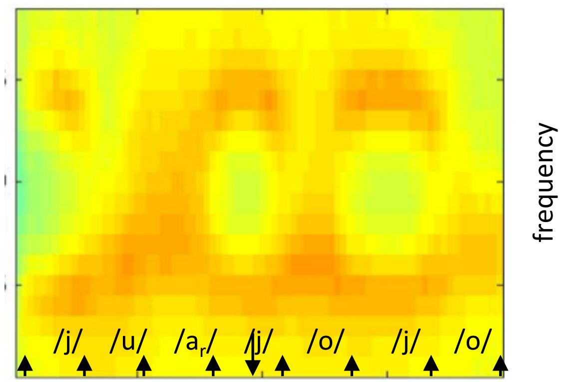


Optimizing for classification of speech sounds suggest critical-band-like spectral resolution and processing within at least 200 ms temporal intervals



Important information about the message is syllable-length time-frequency patterns

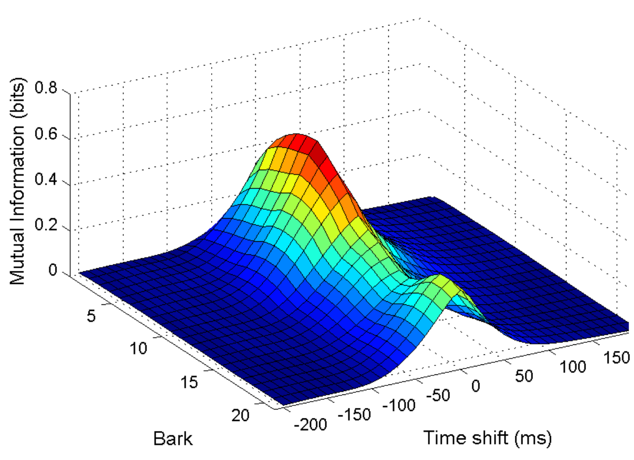
Where is the message in speech?

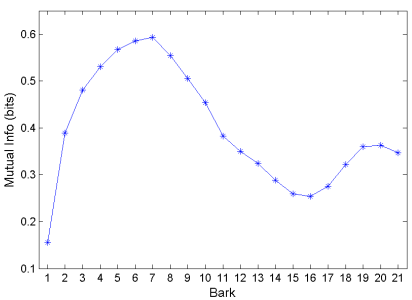


Mutual information between a point *X* in a time-frequency representation of speech (spectrogram) and a phoneme label *Y*

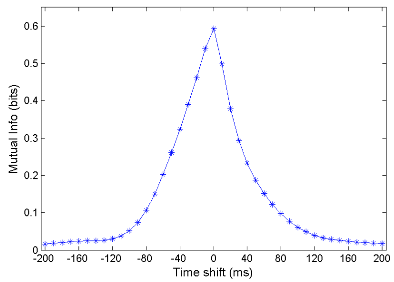
Yang et al, Speech Communication 2000

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



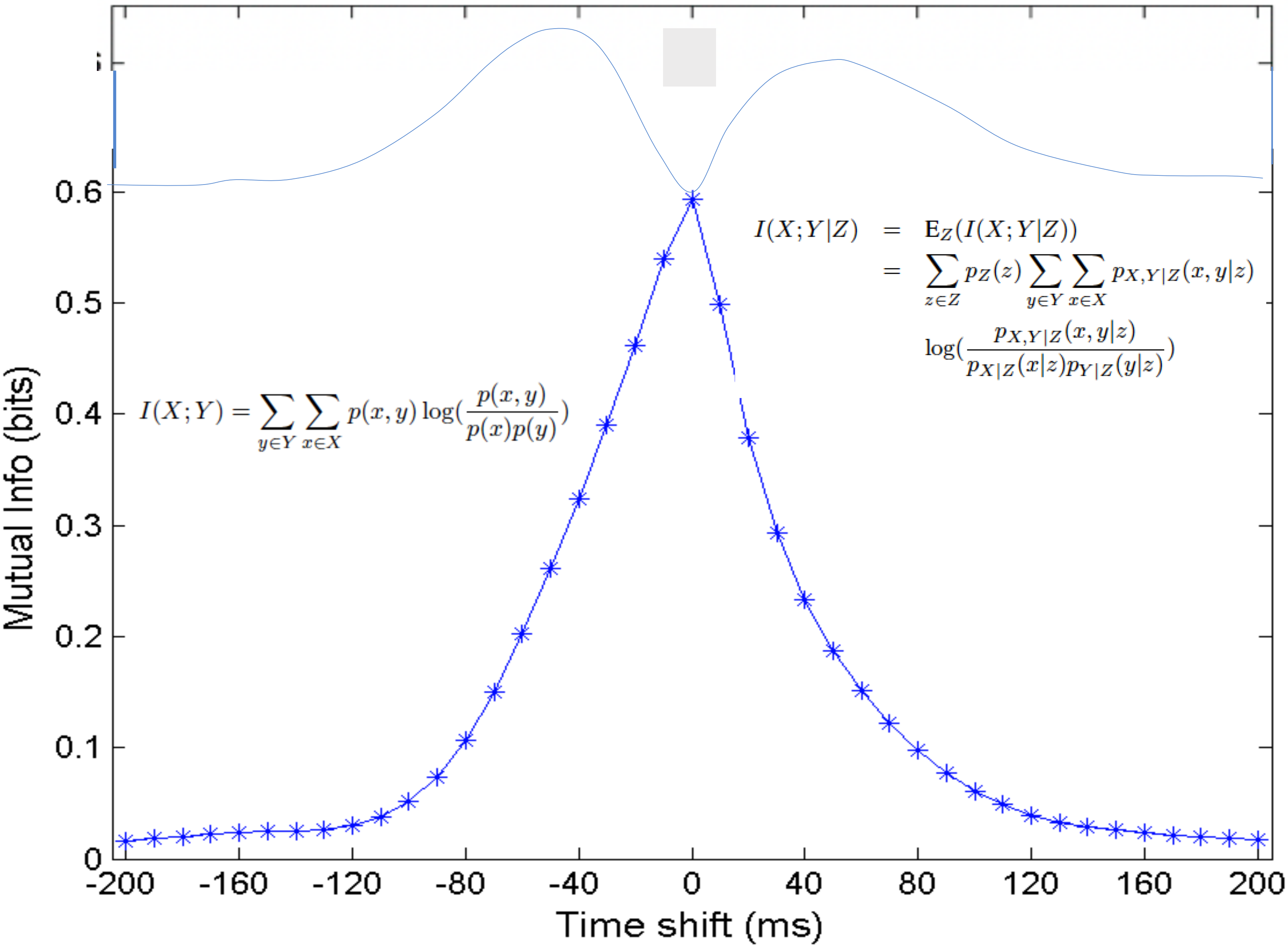


in frequency:
info spread at all frequencies



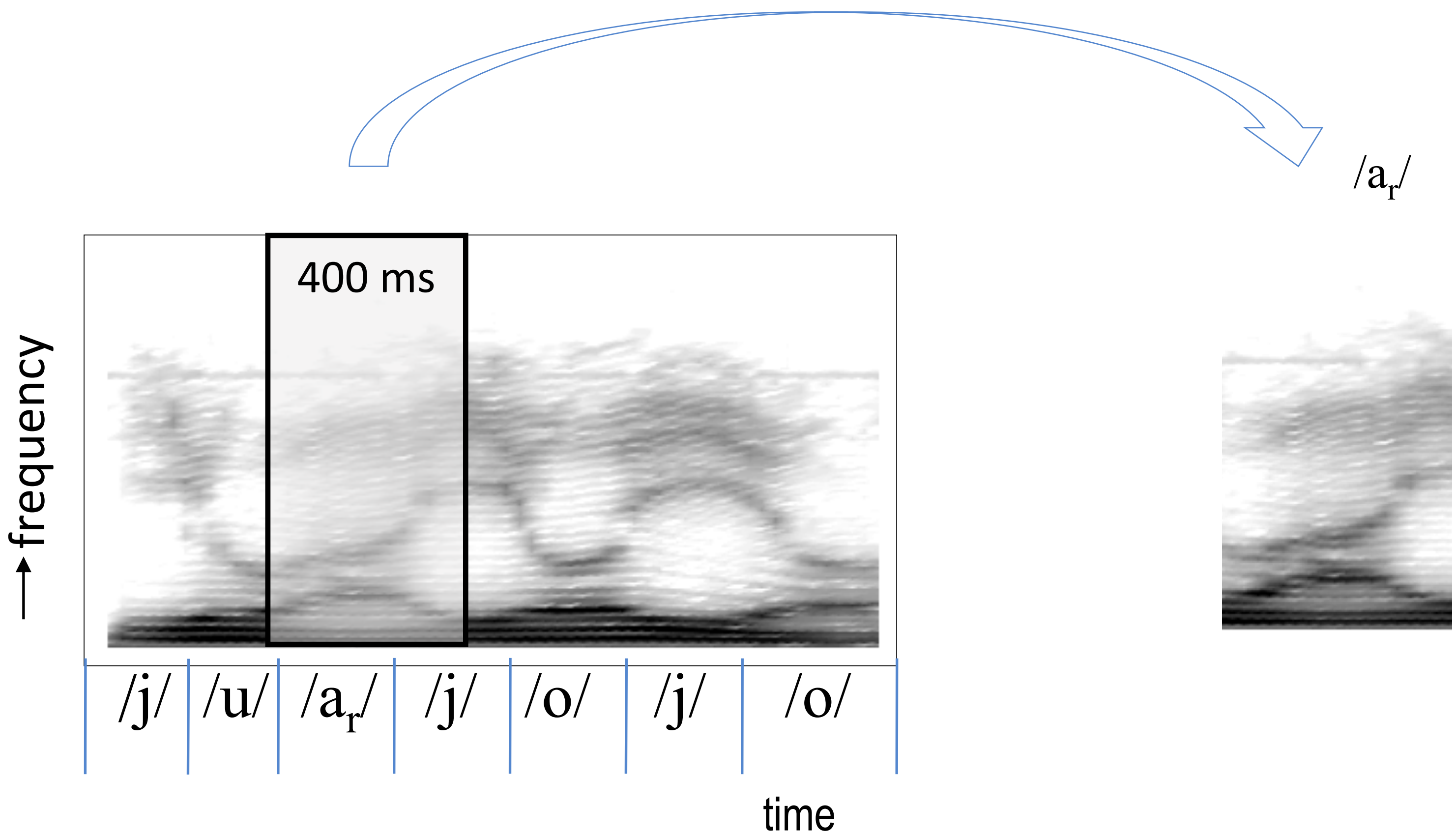
in time:
info spread over about 200 ms

Thanks Feipeng Li (now Apple) for the figures



Thanks Feipeng Li (now Apple) for the figures

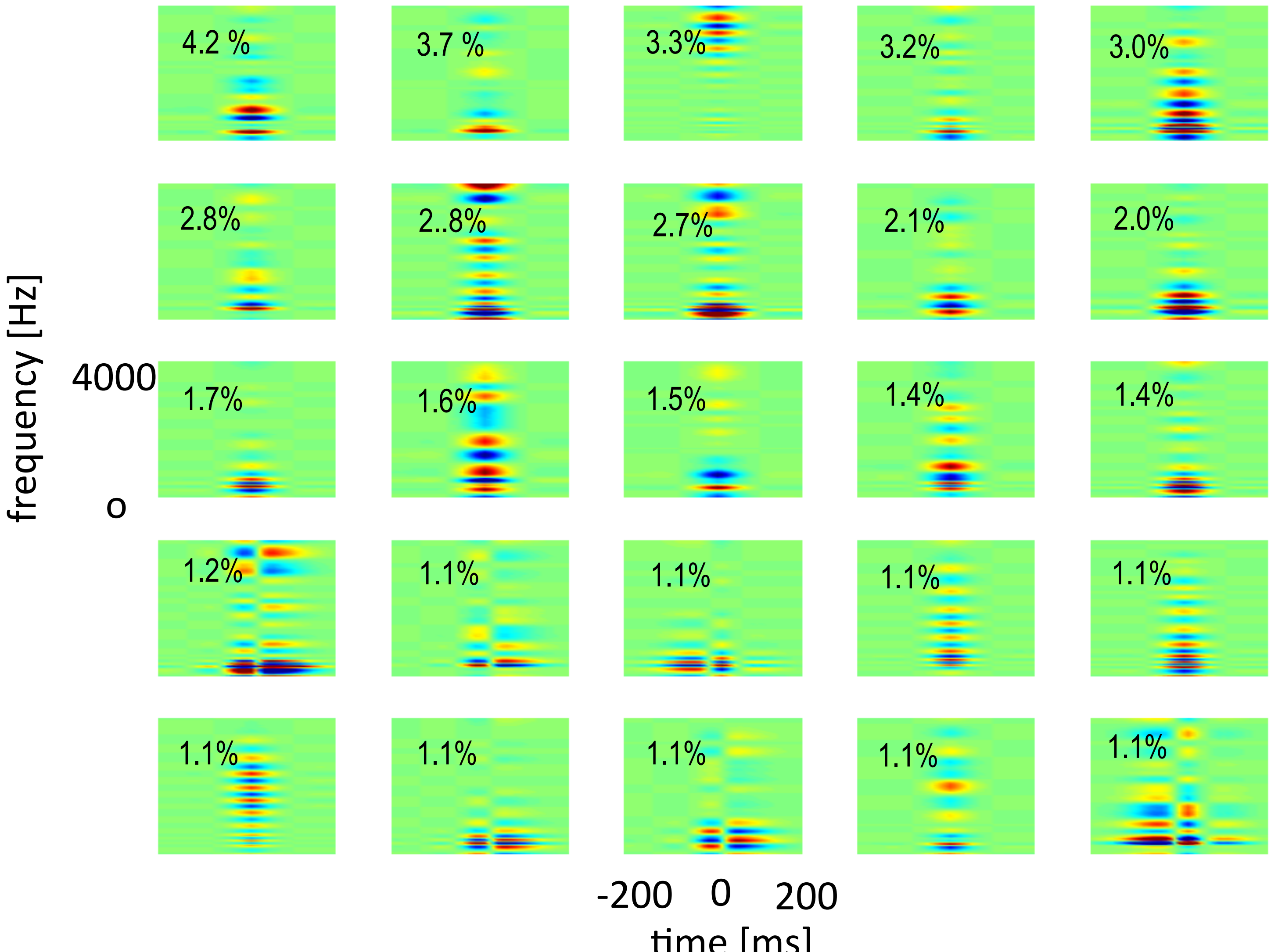
2D (time-frequency) processing of auditory-like speech spectrum



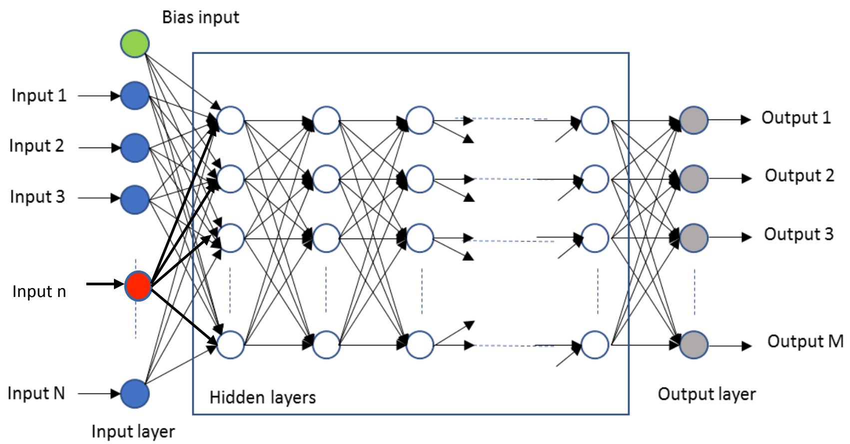
2-D discriminants

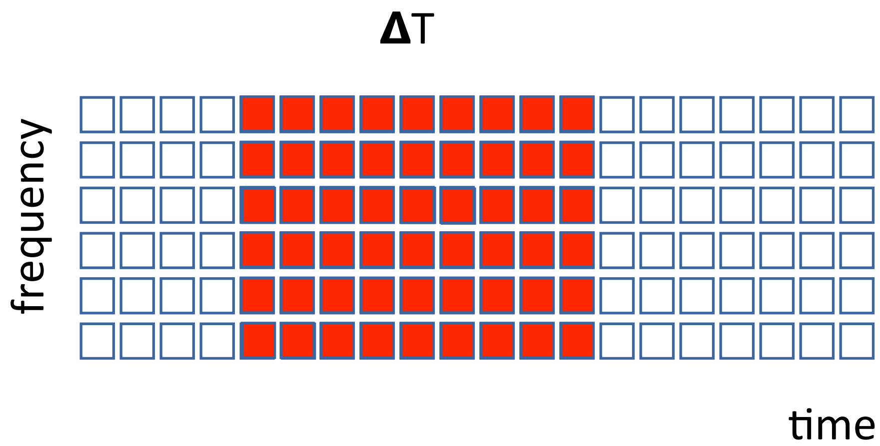
Many 2D discriminants are frequency-selective, emphasizing particular parts of speech spectrum.

Valente and Hermansky 2006



DNN-based design of linear 2D pre-processing

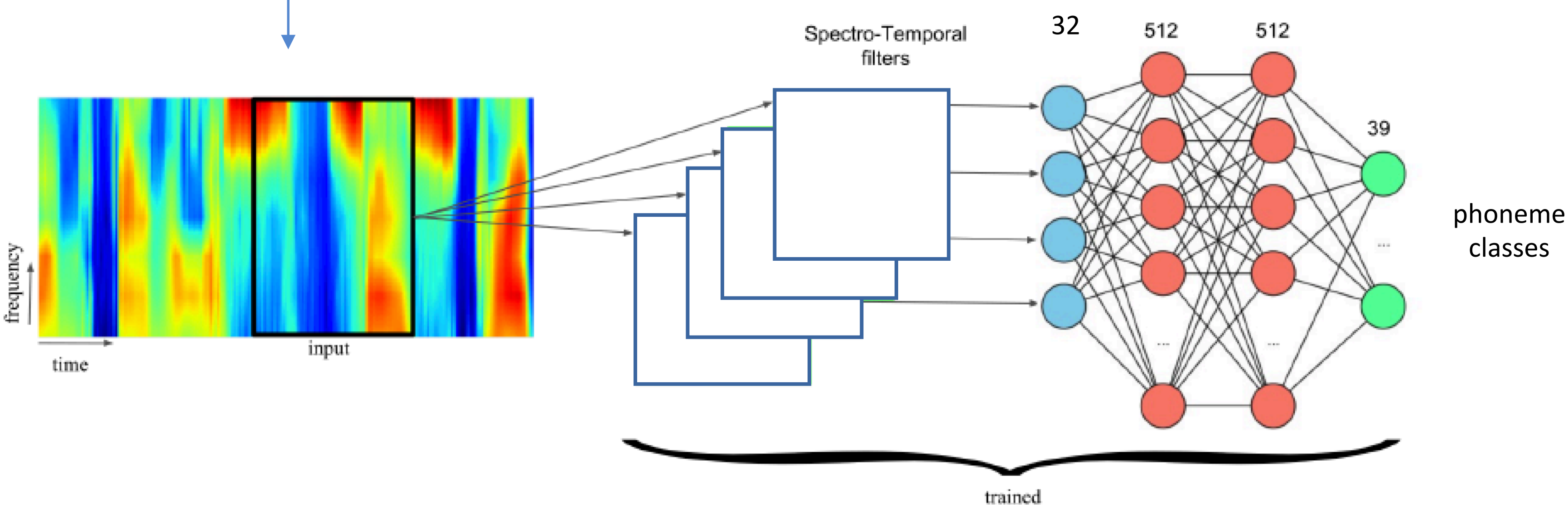




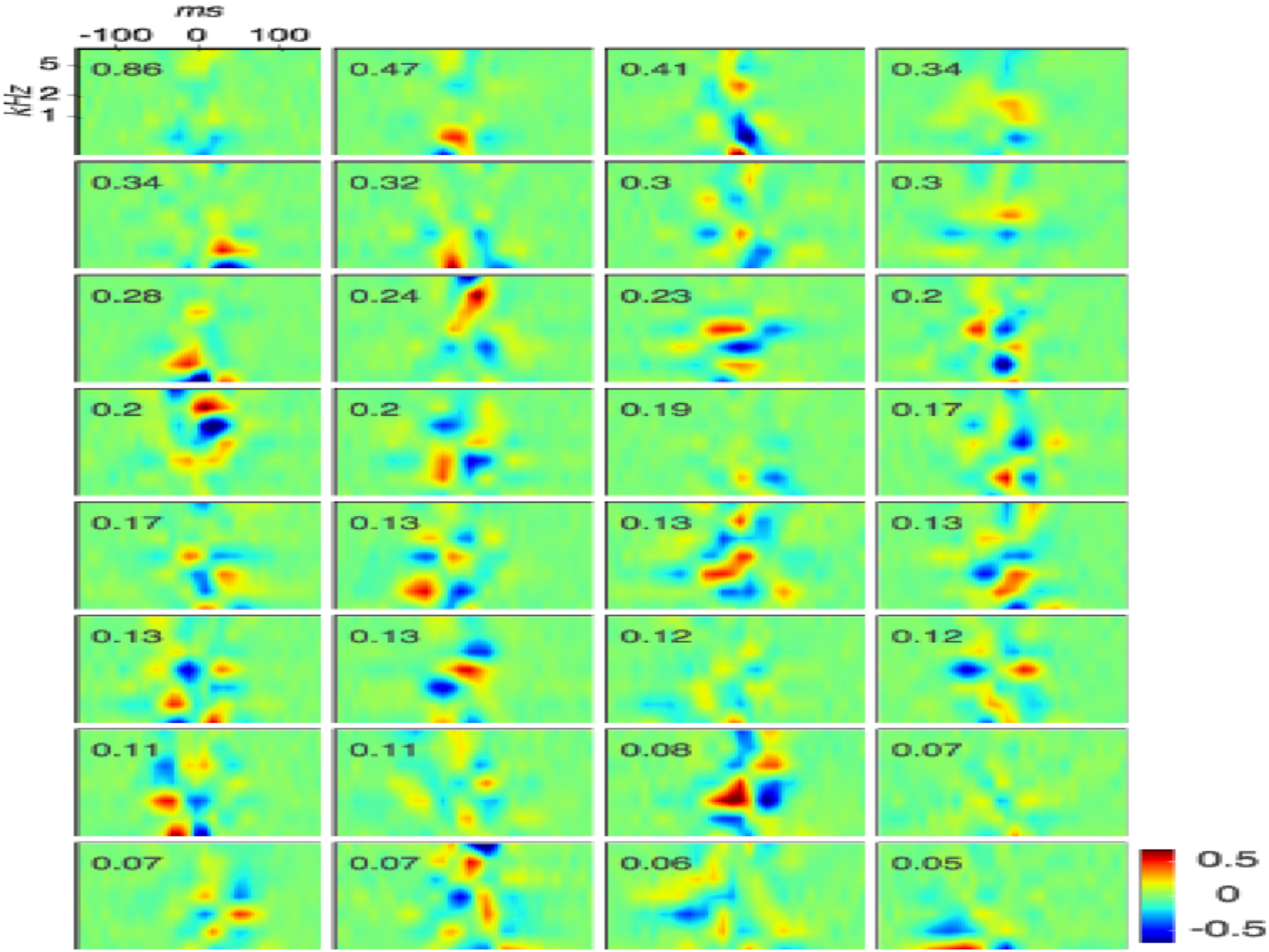
input n spectro-temporal receptive fields

weighted sum of time-frequency values at all frequencies within a time window **Δ**T (weights optimized with the rest of the DNN weights) speech

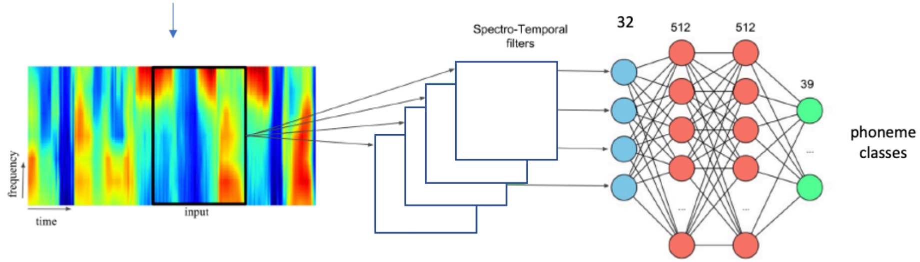
auditory-like spectral analysis



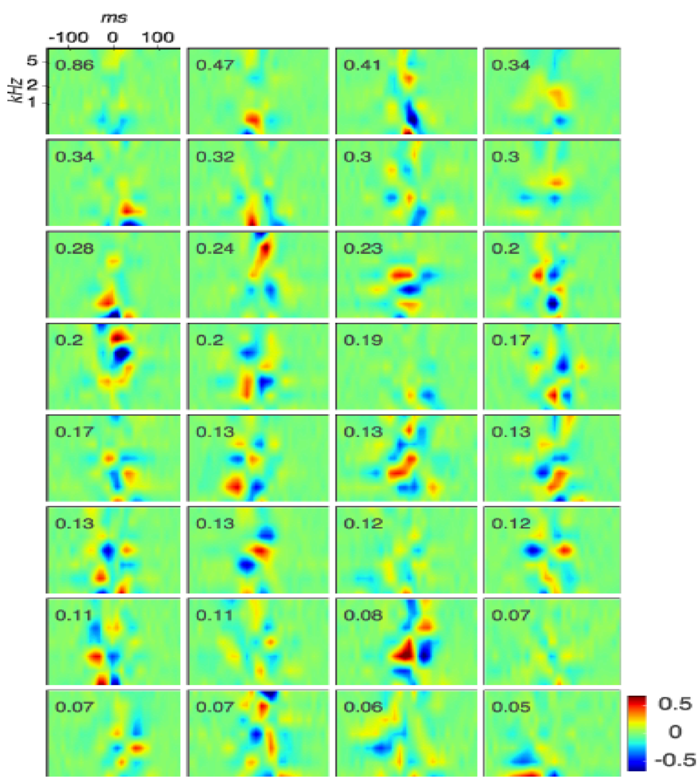
Subset of the phoneme-labeled Wall Street Journal corpus, roughly 37K sentences spoken by 284 speakers for a total of about 62 hours of data, training only on center frames of each phoneme segment


ms
-100 0 100
5
2
1
kHz
0.86
0.47
0.41
0.34
0.34
0.32
0.3
0.3
0.28
0.24
0.23
0.2
0.2
0.2
0.19
0.17
0.17
0.13
0.13
0.13
0.13
0.13
0.12
0.12
0.11
0.11
0.08
0.07
0.07
0.07
0.06
0.05
0.5
0
-0.5

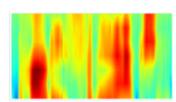
Convolve each temporal trajectory of spectral energy with different FIR filters (rows of the spectro-temporal filters matrices)

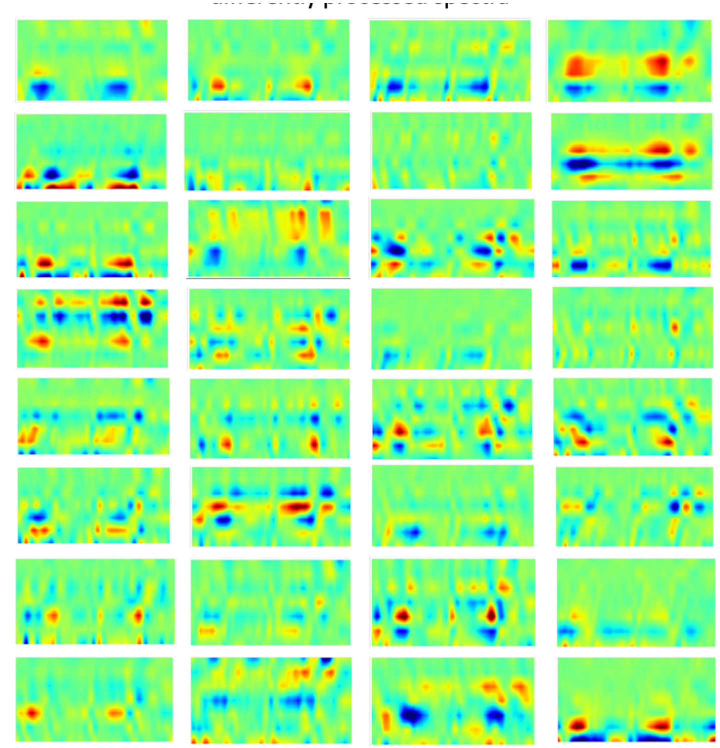


each node in the first hidden layer sees speech with **differently emphasized spectral components (different combinations of spectral channels)**



original spectrum





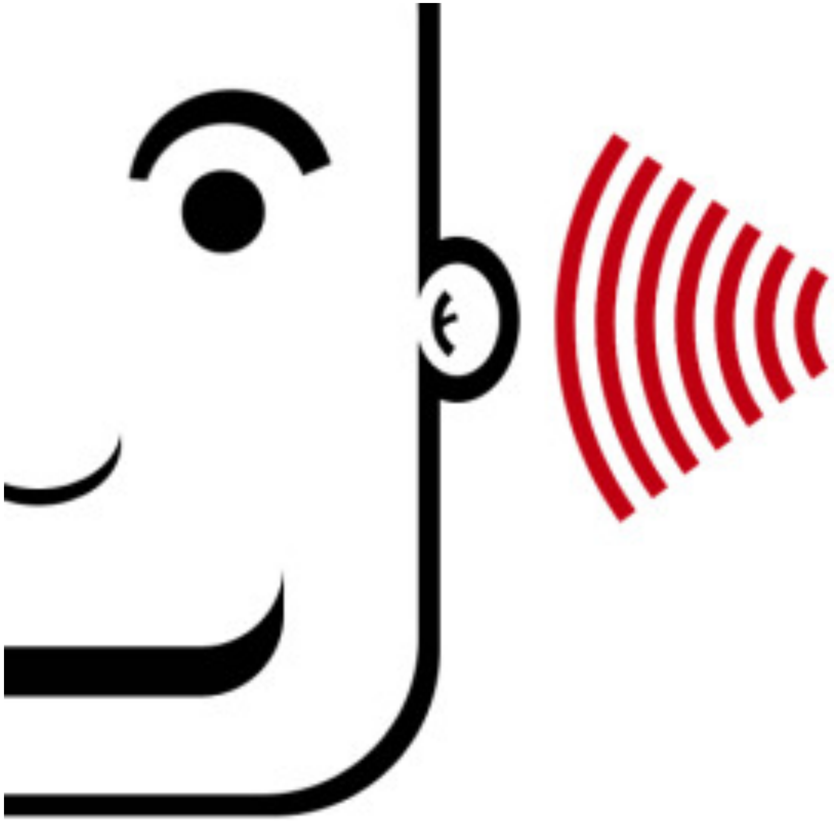
Articulatory Bands

French and Steinberg 1949

250-375-505-654-795-995-1130-1315-1515-1720-1930-2140-2355-2600-2900-3255-3680-4200-4860-5720-7000 Hz

- 20 frequency bands in speech spectral region
- each band contributes about equally to human speech recognition
- any 10 bands sufficient for 70% correct recognition of nonsense syllables, better than 95% correct recognition of meaningful sentences [Fletcher and Steinberg 1929]

HEARING



message? who ? where from?

inter-spike interval

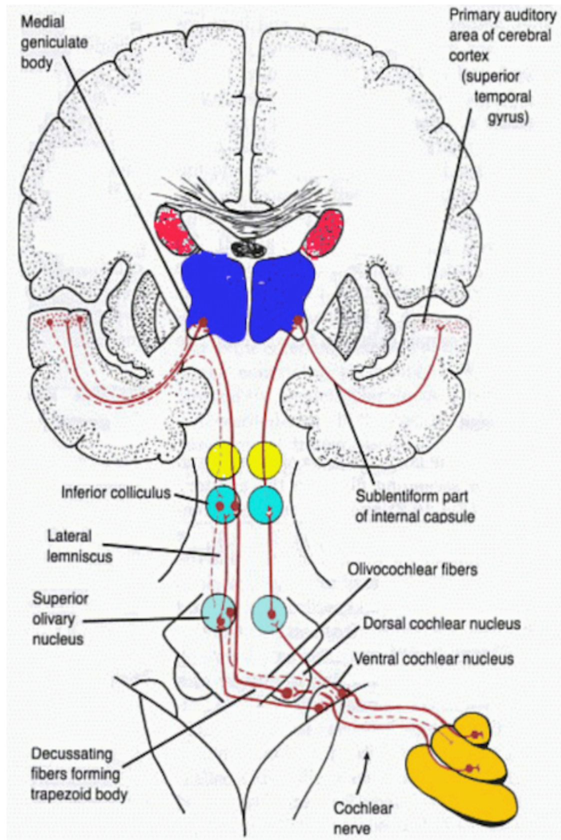
~100 ms

number of spiking neurons

~100,000,000

up to 10, 000,000 active in a given task

top-down connections



bottom-up connections

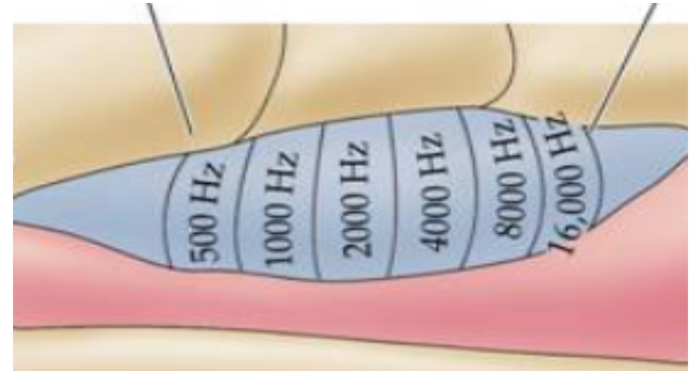
~1 ms

~100,000

speech signal

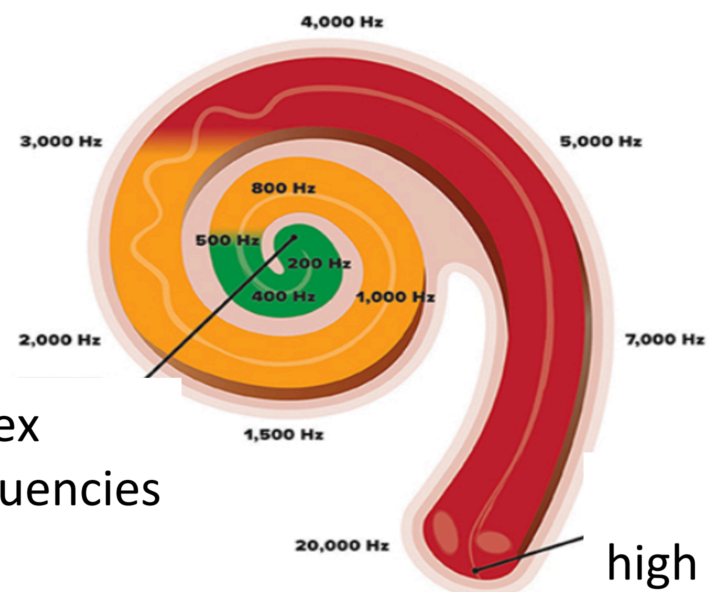
TONOTOPY

different frequencies
excite different parts
of the cortex

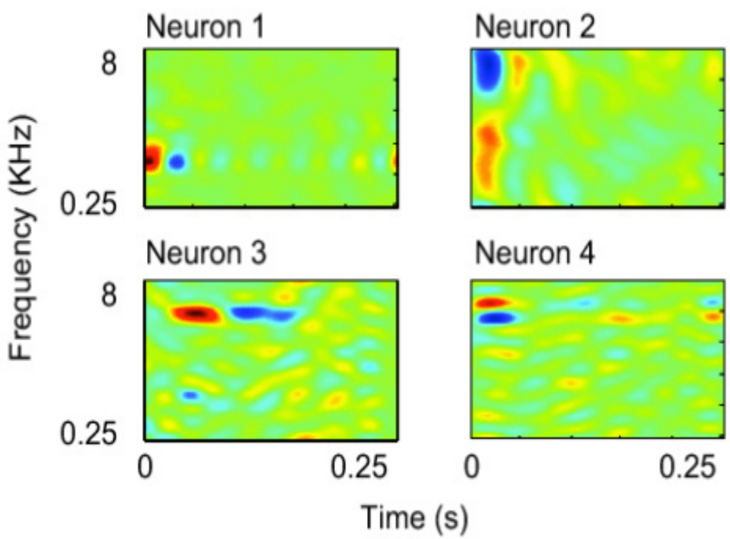


processing
stages

different frequencies
excite different parts
of the cochleaa



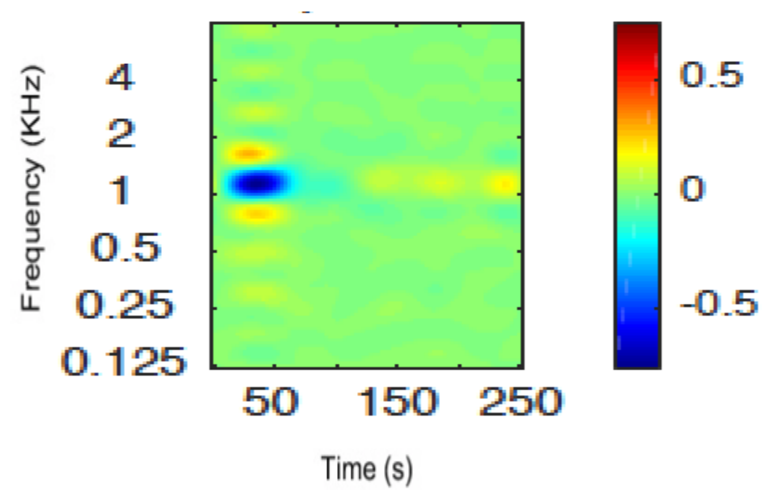
Cortical spectro-temporal receptive fields (STRFs)



Mesgarani et al Interspeech 2010

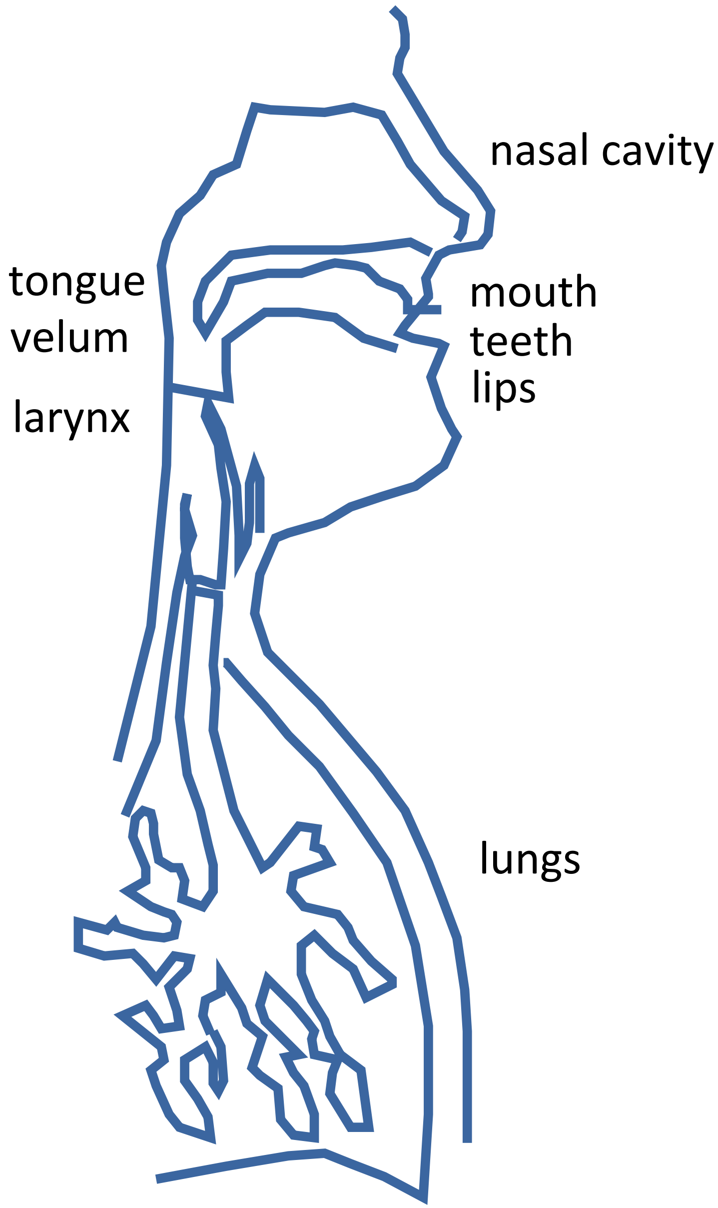
first principal component of about 700 STRFs

– Mesgarani et al (in preparation)



- **Frequency-selective (> 2 octaves) and relatively long in time (> 200 ms)**

SPEAKING



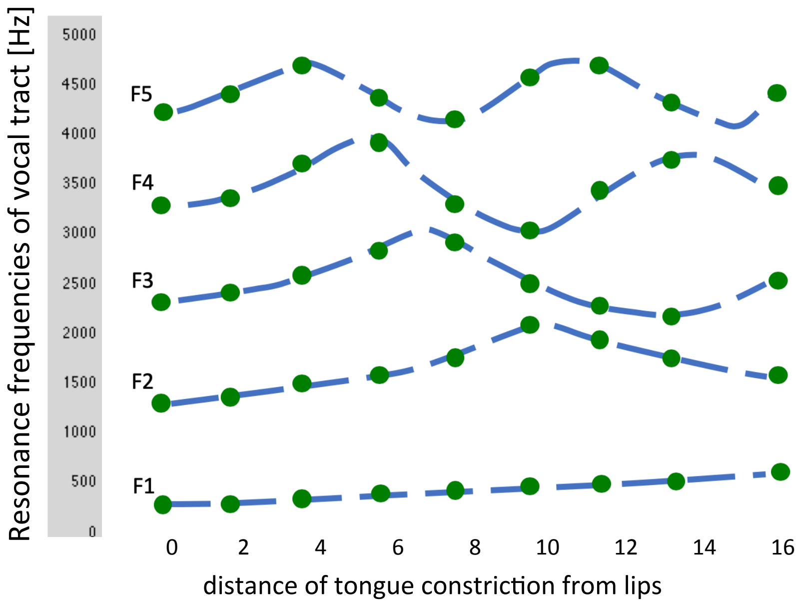
breathing

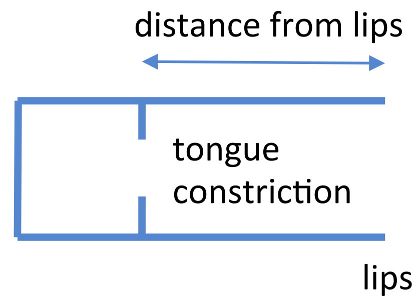
eating

biting

speaking?

INFORMATION ABOUT TRACT SHAPES DISTRIBUTED IN FREQUENCY





any change in the tract shape is reflected at ALL FREQUENCIES of speech spectrum !

Information about vocal tract shape (about linguistic message) is **coded redundantly** in frequency

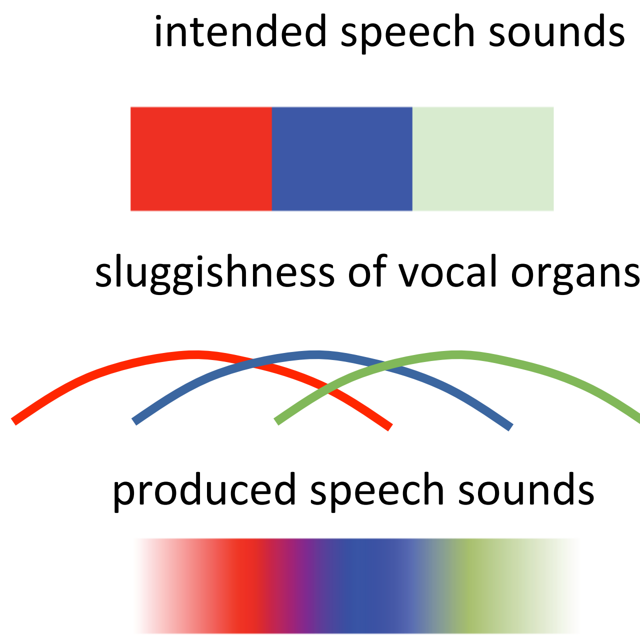
motor control → critical elements (tongue, lips, velum) → shape of the whole vocal tract →

spectrum of speech signal
(redundant contributions of movements of critical elements in different frequency bands)

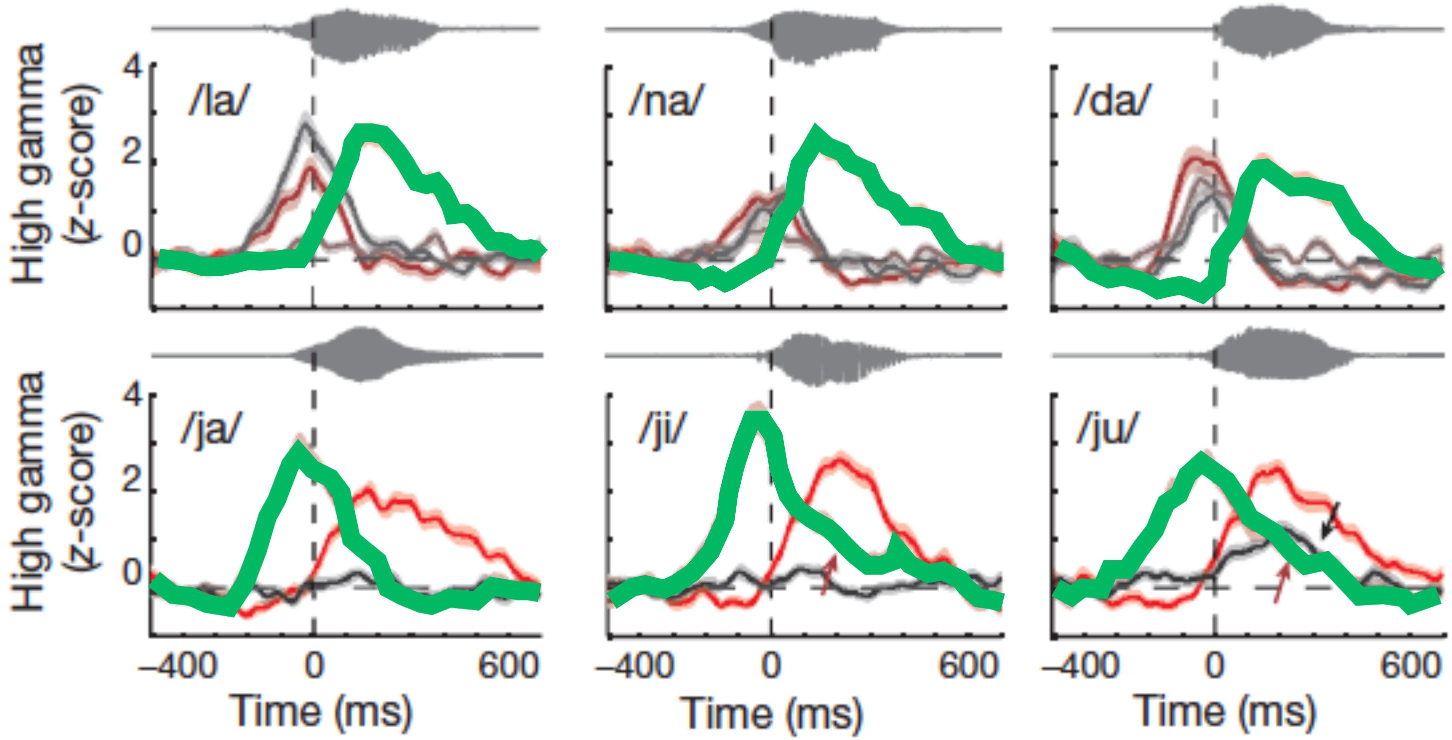
INFORMATION ABOUT TRACT SHAPES DISTRIBUTED IN TIME

from Sri Narajanan

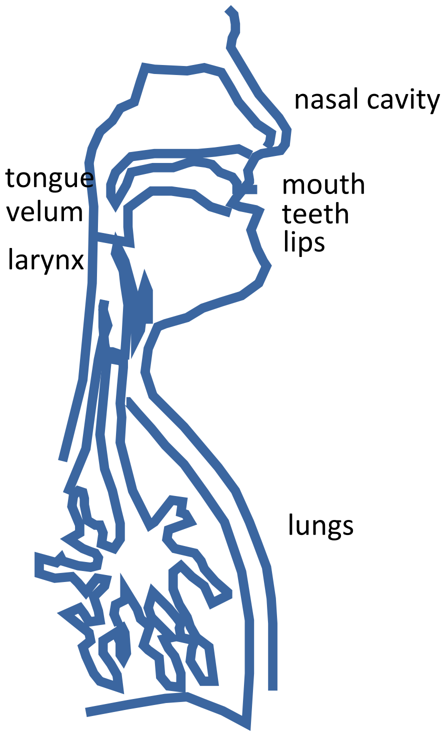
movements of vocal organs are rather sluggish



Where is the corticulation in production?

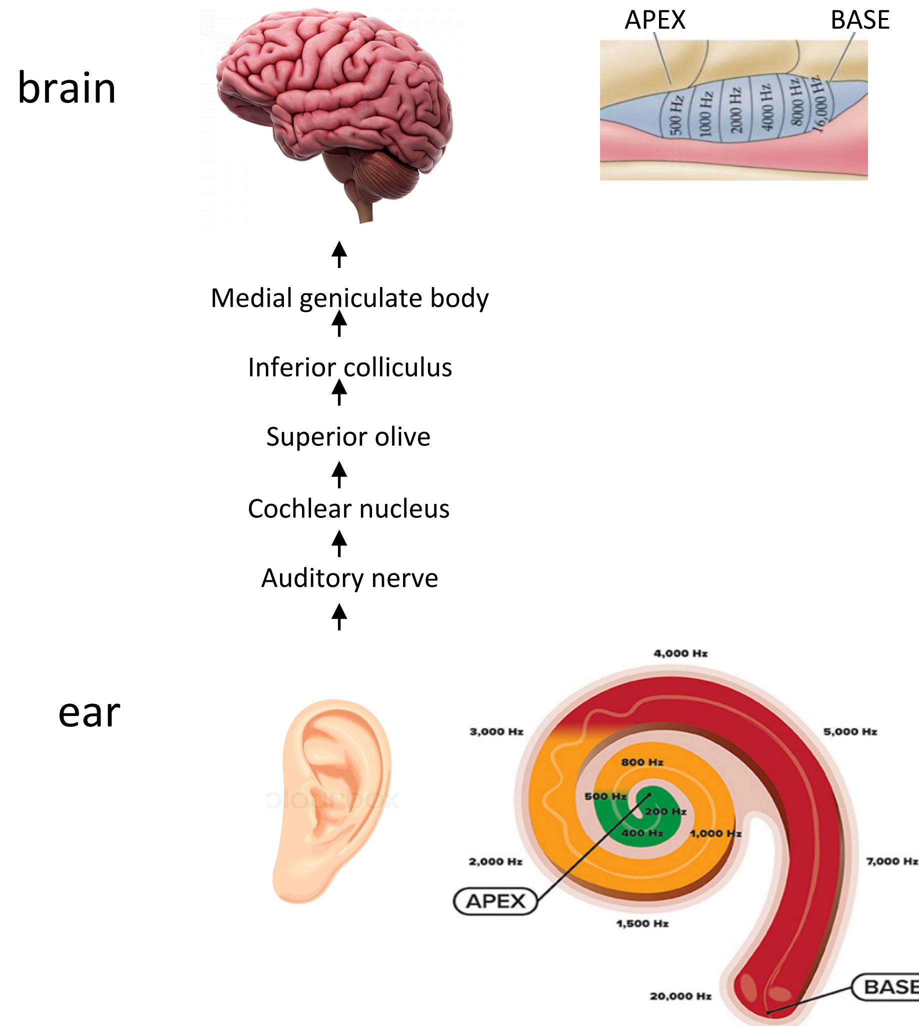


Functional Organization of Human Sensorimotor Cortex for Speech Articulation
Kristofer E .Bouchard, Nima Mesgarani, Keith Johnson, and Edward F. Chang, *Nature*.2013



Redundant spread of information

- every change of the tract shape shows at all frequencies of speech spectrum
- tract shape changes do not happen very fast



- frequency selective (about 20 bands)
- sluggish (tenths of seconds)

< 50 bps		> 50 kbs		< 50 bps
message →	**coding** introduce redundancies in frequency and in time →	**speech signal** ↑ noise →	**decoding** use redundancies for reliable extraction of the message →	**message**

PRODUCTION *TRANSMISSION* **PERCEPTION**

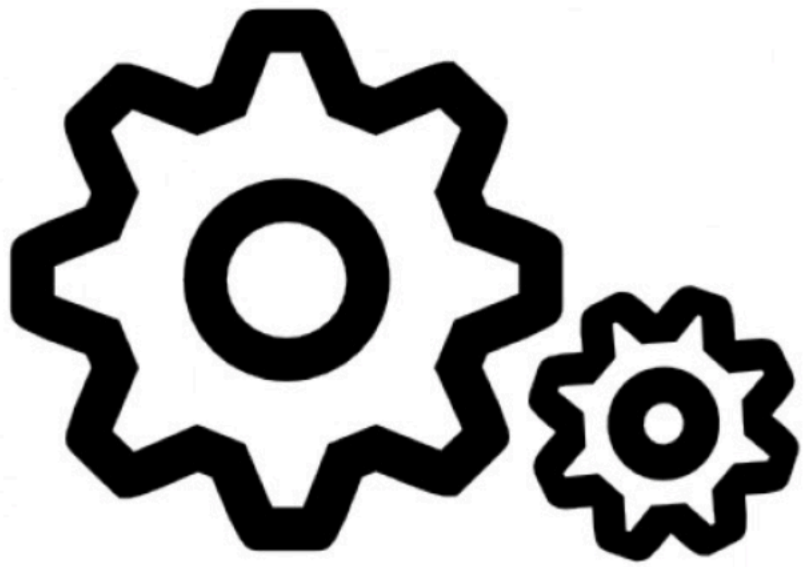
redundancy in frequency

production: tract acoustics distributes the information to all frequencies of the speech spectrum
perception: hearing selectivity allows for decoding the information in separate frequency bands

redundancy in time

production: tract sluggishness (coarticulation) distributes information about each speech sound in time
perception: temporal sluggishness of hearing collect the information distributed in time

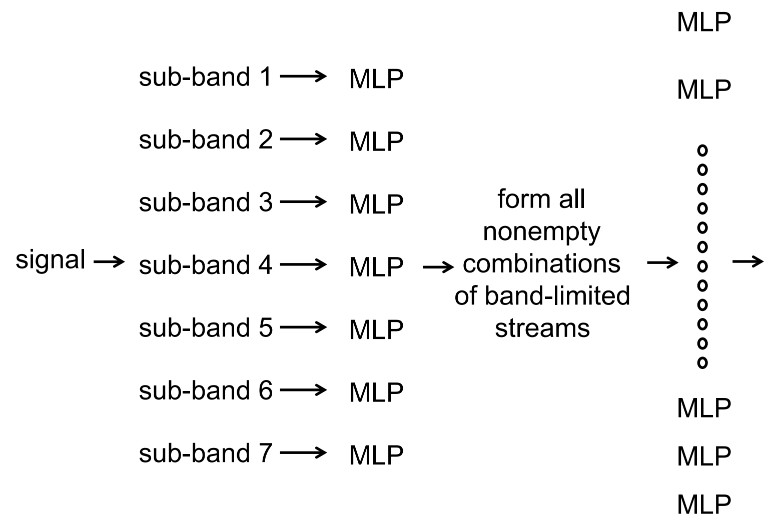
ENGINEERING

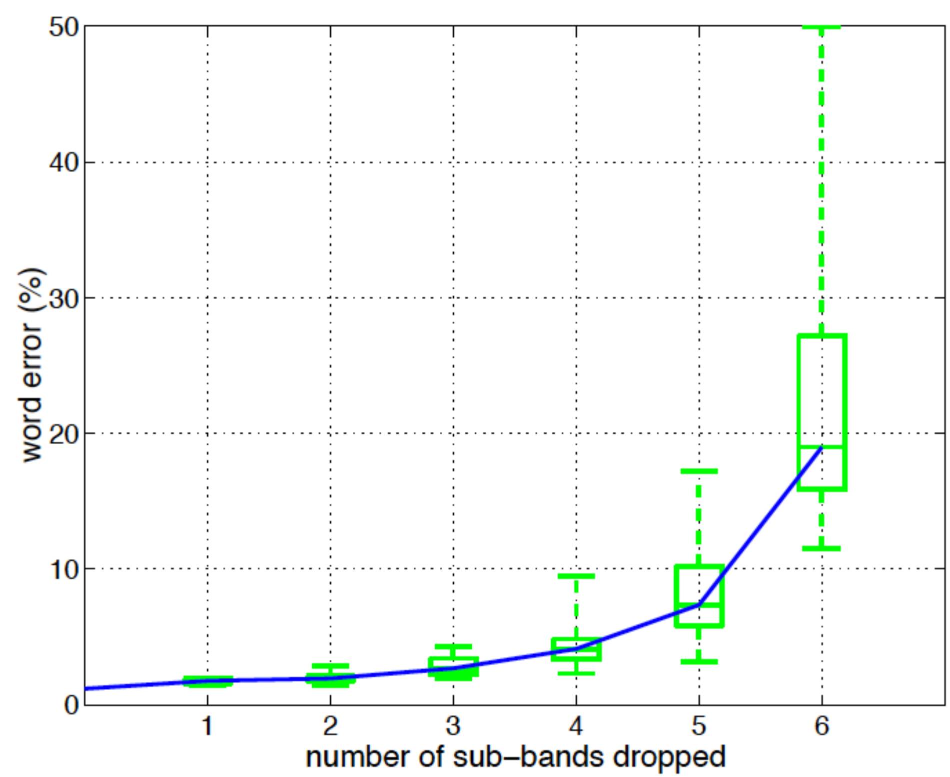


127 different stream combinations in hierarchical MLP structures

evaluate word error for different stream combinations

Hermansky et al 1996

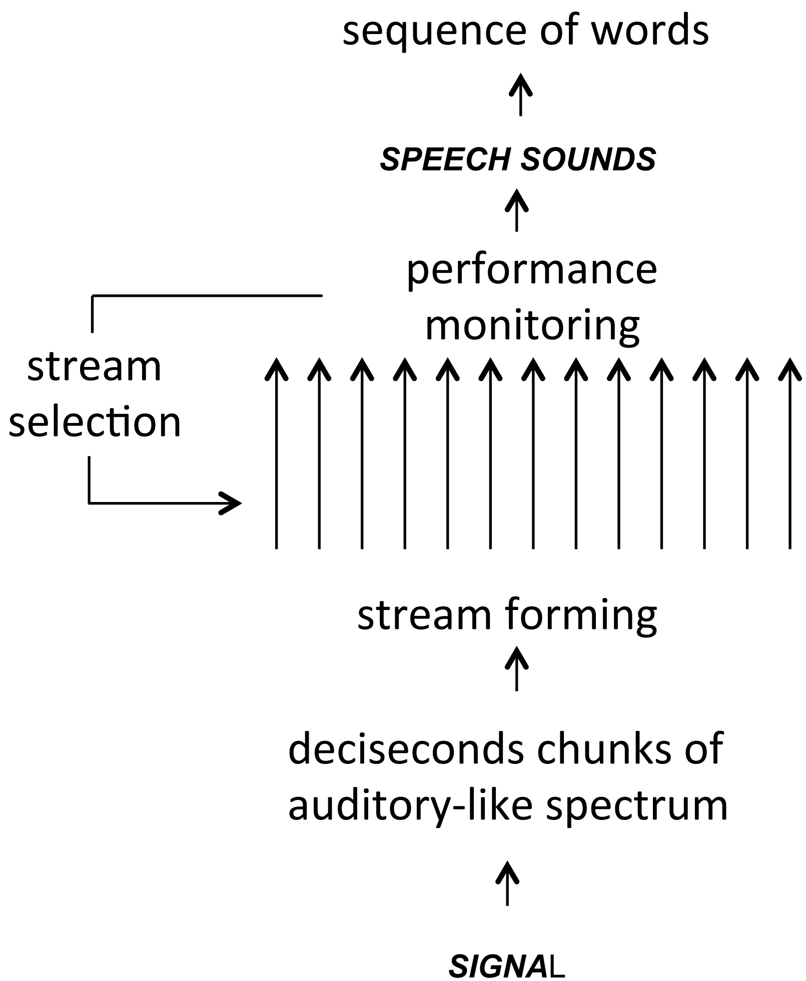




Word error rates on very noisy reverberant speech (Chime 5)

length of the input pattern [s]	word error rate [%]
0.25	83.4
0.5	70.7
1.0	66.7
1.5	64.2

Machine Recognition of Speech ?



streams

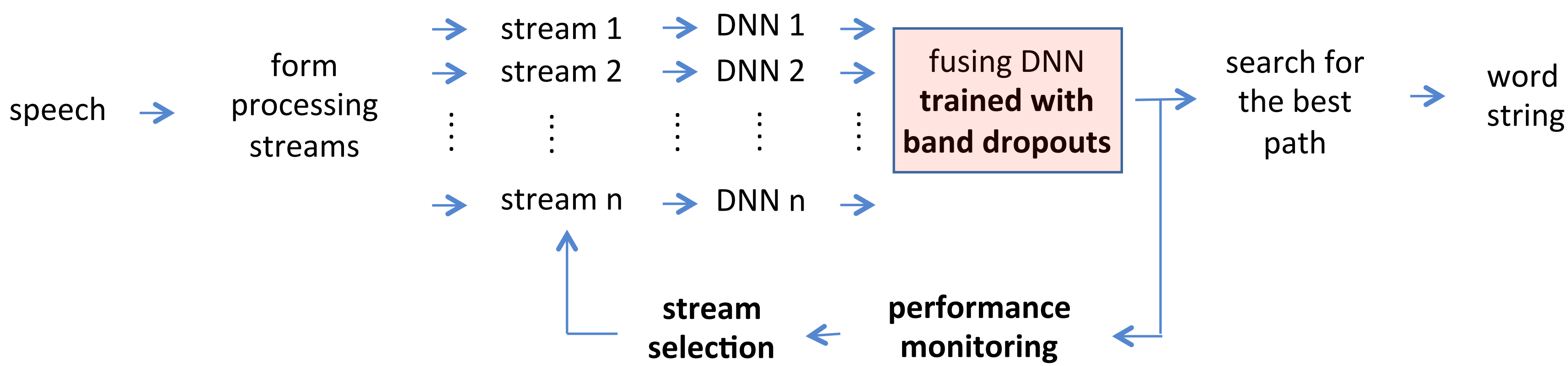
- enhancing different parts of speech spectrum
- enhancing different spectral and temporal modulations

performance monitoring

- **estimate quality of information without knowing the information**

multiband recognizer with stream dropping

Mallidi and Hermansky 2016



word error rates of on Aurora noisy data

auditory spectrum	spectral streams	stream dropping	performance monitoring	**oracle selection**
12.6	11.0	9.9	9.6	7.9

Sri Harish Mallidi, JHU PhD Thesis, 2018

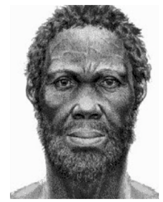
training with stream dropping also applied in Park, Daniel S., et al. "Specaugment: A simple data augmentation method for automatic speech recognition." 2019

Environment (survival)

Hearing (communication)

Evolution of hearing

200 000 000 years



Evolution of speech

200 000 years

Pristerodon

Homo sapiens

We hear to survive

We speak to hear

Human speech evolved to fit properties of human hearing

ergo

Optimizing speech technology on speech data yields relevant hearing knowledge

Supported by the National Science Foundation
EAGER Grant 126289

Prof. Frederick Jelinek says:
"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 first we must figure out how to parameterize it, and **how to estimate the parameter values from speech 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

Received 20 June 1969 9.10, 9.1

Whither Speech Recognition?

Letter to Editor
J.Acoust.Soc.Am.

J.R. Pierce

Bell Telephone Laboratories, Inc., Murray Hill, New Jersey 07971

Implement…. ***intelligence and knowledge of language comparable to those of a native speaker !***

…. should people continue work towards speech recognition by machine ? Perhaps it is for people in the field to decide.

why to work on machine recognition of speech?

useful technology, profits, safe jobs,…..

Why to climb Mount Everest?
Because it is there.
George Leigh Mallory

Why to study speech?

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