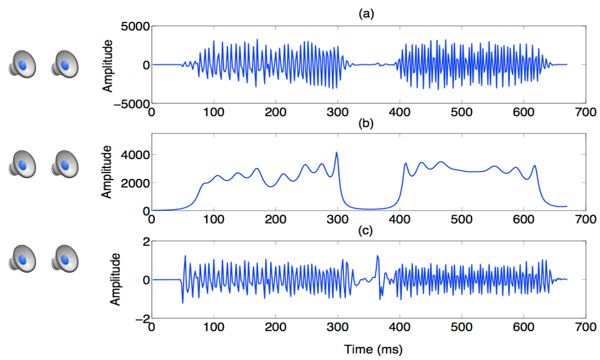
Autoregressive model of Hilbert envelope of the signal



signal

AM component
(temporal envelope)

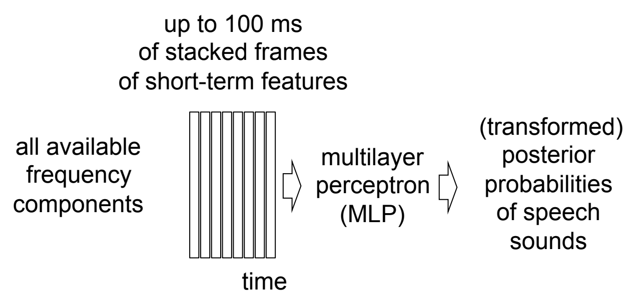
FM component
(carrier)

channel vocoder based on AM or FM components

Artificial Neural Nets for Deriving Speech Features

112

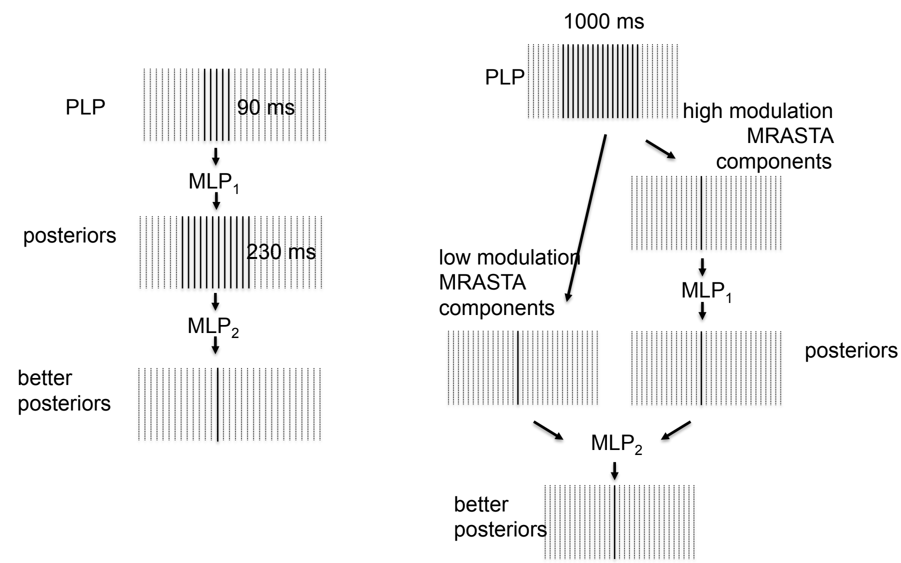
conventional artificial neural net



DEEP: Some Hierarchical Nets

serial hierarchy (with Joel Pinto)

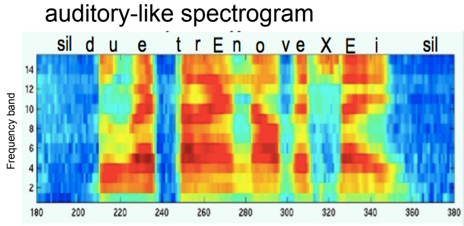
serio-parallel hierarchy (with Fabio Valente)



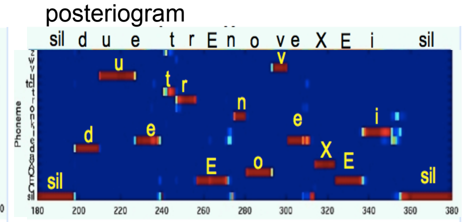
Artificial Neural Nets

signal → pre processing → neural network estimating posteriors of speech sounds →

auditory-like spectrogram



posteriogram



TANDEM

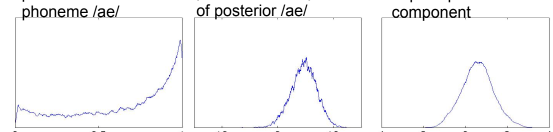
signal → pre processing → neural network estimating posteriors of speech sounds → pre-softmax outputs → principal component projection → to HMM

histograms of one typical element of feature vectors

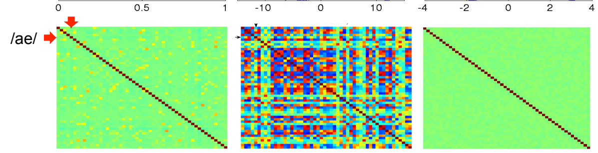
posterior of phoneme /ae/

pre-softmax output of posterior /ae/

4th principal component

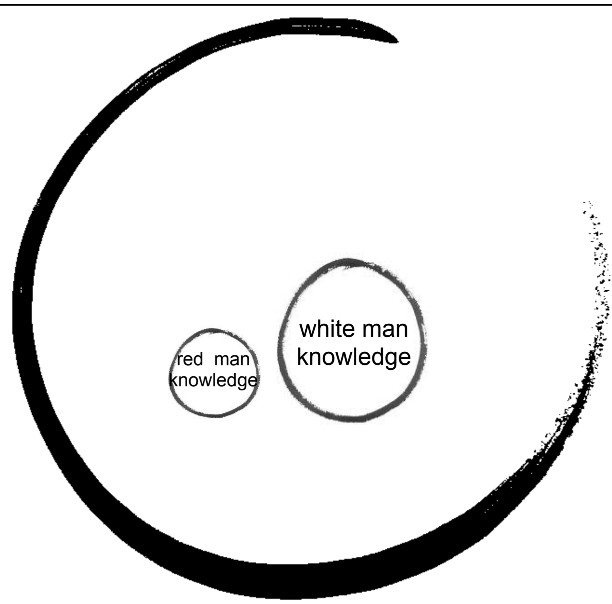


correlation matrices of whole feature vectors



Unknown Unknowns

117



The problem is not what you do not know,
the problem is what you do not know that you do not know

Machine Learning

Create models of the "world"

1. from labeled (annotated) training data
2. **from prior knowledge of what is possible and how likely it is**

Find the model that best accounts for the observed data

Assumption: the future is the same as the past

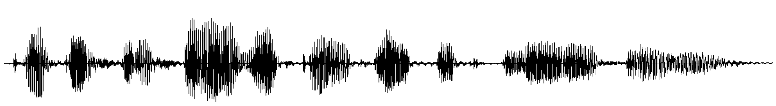
- both the training and the test data are independently and identically obtained samples from the same probability distribution

Unexpected events are hard to deal with because

1. not seen in the training
2. **low (zero) prior probability**

Successfully surviving natural systems attend well to the unexpected

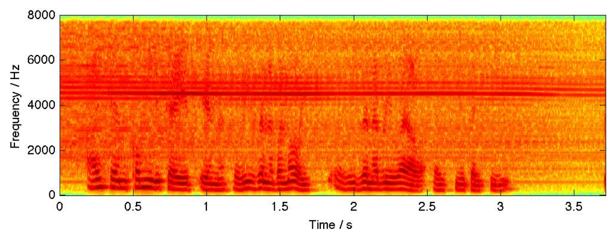
Power of Priors (Language Model)

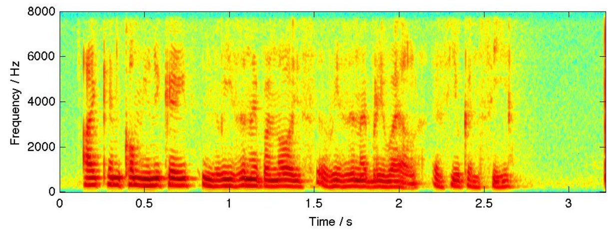


although some sort of the **computer** can either way

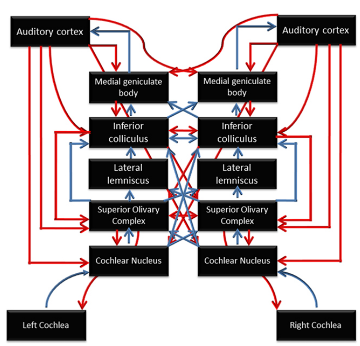
hopefully cin-cin o-bi **computer** connected with

Unexpected Noise





	inter-spike interval		number of spiking neurons
slower complex events	100 ms		10,000,000
	10 ms		1,000,000
faster simpler events	1 ms		100,000



Deep: many layers
Long: cortical event every 100 ms or so
Wide: many possible descriptions of an event in auditory cortex

DEEP: Information in the signal should be extracted in stages, from description of signal features to description of phonetic events.

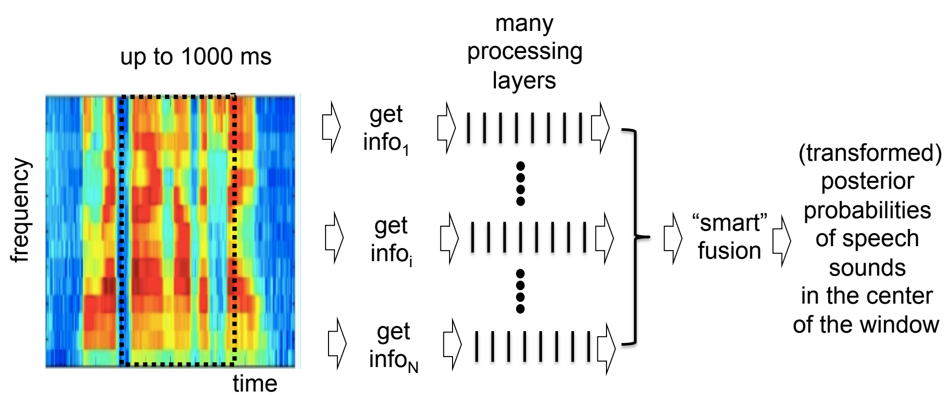
LONG: Information about underlying speech sounds is spread in time for more than 200 ms

WIDE: There are many ways to form parallel processing streams using different signal projections and different prior assumptions.

Not all processing streams get always corrupted and we need to find ways to find the uncorrupted processing streams.

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

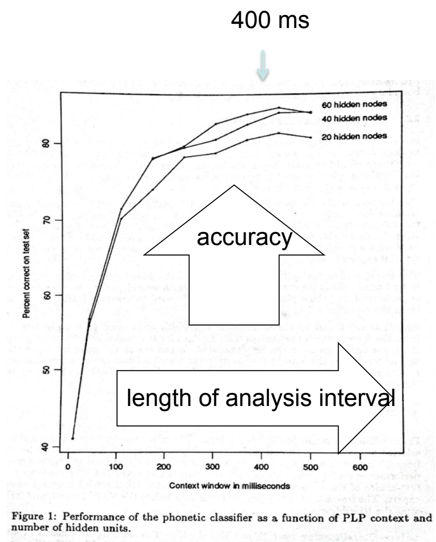
Deep, Long, and Wide Neural Nets



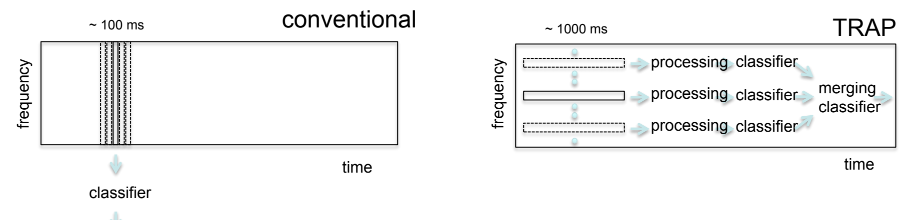
Longer is Better

Phonetic classifier accuracy as a function of a time span of an analysis

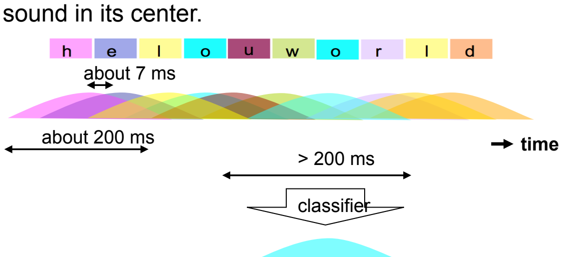
Fanty, Cole, Roginski NIPS 1992



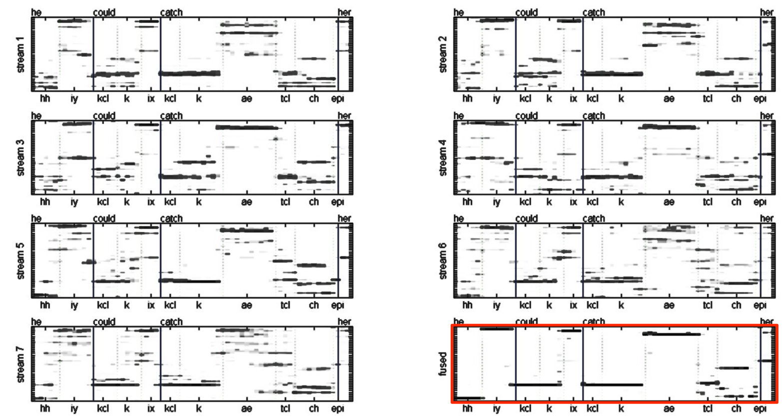
LONG: Classifying **T**empo**RA**l **P**atterns of Spectral Energies
with Sangita Sharma, Pratibha Jain, Honza Cernocky, Pavel Matejka, Petr Schwartz ….



Each temporal pattern contains most of coarticulation span of speech sound in its center.

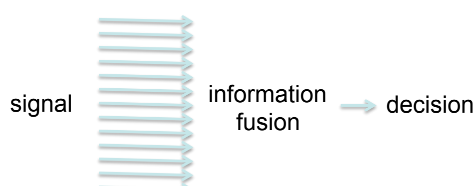


Fusion of streams of different carrier frequencies



Wide: Multi-stream Processing

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



- Parallel information-providing streams, each carrying different redundant dimensions of a given target.
- A strategy for comparing the streams.
- A strategy for selecting “reliable” streams.

Stream formation

- Different perceptual modalities
- Different processing channels within each modality
- Bottom-up and top-down dominated channels

Comparing the streams ?

- various correlation (distance) measures

Selecting reliable streams ?????

Early Attempts for Multi-Stream Recognition

with Sangita Sharma and Misha Pavel

signal → sub-band 1 → MLP
sub-band 2 → MLP
sub-band 3 → MLP
sub-band 4 → MLP → form all nonempty combinations of band-limited streams → MLP, MLP, ..., MLP, MLP, MLP → find reliable streams
sub-band 5 → MLP
sub-band 6 → MLP
sub-band 7 → MLP

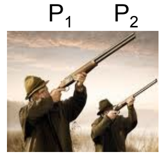
- SNR in sub-bands
- classifier confidence
- majority vote
- supervised adaptation

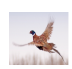
Monitoring Performance

Fletcher et al
Boothroyd ann Nittrouer
Allen

$$P(\varepsilon) = \prod_i P(\varepsilon_i)$$

P_1 P_2





$P_{miss} = (1-P_1)(1-P_2)$

observer - false positives and negatives are possible

$P_{miss\_observed} \neq (1-P_1)(1-P_2)$

Do listeners know when they know ?

How to make machine know when it knows ?

performance on training data → compare

modify ← compare

performance in test → compare

Finding Reliable Streams

Streams which yield the best performance on the test data

Classifier can never work better that it does on the data on which it was trained

performance
on training data

choose
the best
stream
combinations

compare

performance
in test

Evaluating Performance

How often sound classes occur and how often do they get confused?

$$AC = 1/N\sum_{i=1}^{N}(\mathbf{p}_i)^r(\mathbf{p}_i^T)^r$$

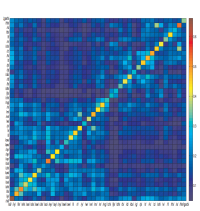
$\boldsymbol{p_i}$ – vector of sound posteriors at i-th time instant
N – time interval of the evaluation
r – th power element-by-element (currently r=0.1)

How much sounds classes differ and how fast do they change?

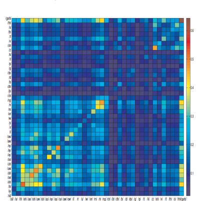
$$M(\Delta i) = \frac{\sum_{i=0}^{N-\Delta i} D(\mathbf{p}_i, \mathbf{p}_{i+\Delta i})}{N-\Delta i}$$

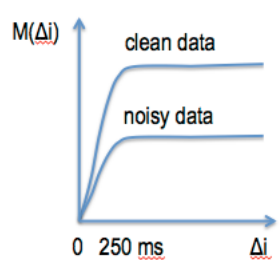
Δi – time delay
D(.) – symmetric Kl divergence

clean data

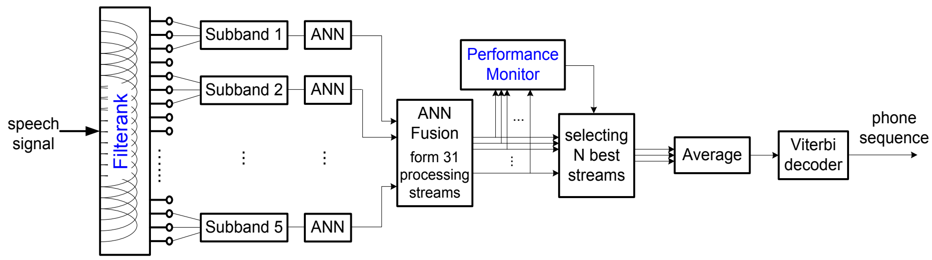


noisy data





Multi-stream speech recognition

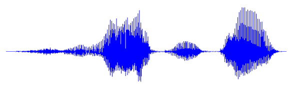


Phoneme recognition error rates

environment	conventional	proposed	best by hand
clean (matched training and test)	**31 %**	**28 %**	**25 %**
TIMIT with car noise at 0 dB SNR (training on clean)	**54 %**	**38 %**	**35 %**
RATS data (Channel E – matched training and test)	**70 %**	**57 %**	**49 %**

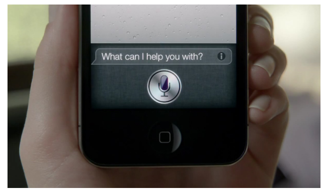
Towards Increasing Error Rates

134



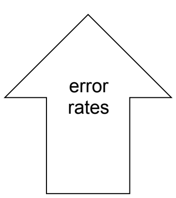
signal processing
↓
pattern classification
↓
decoder
↓
message

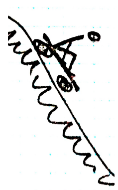
Signal processing, information theory, machine learning, …

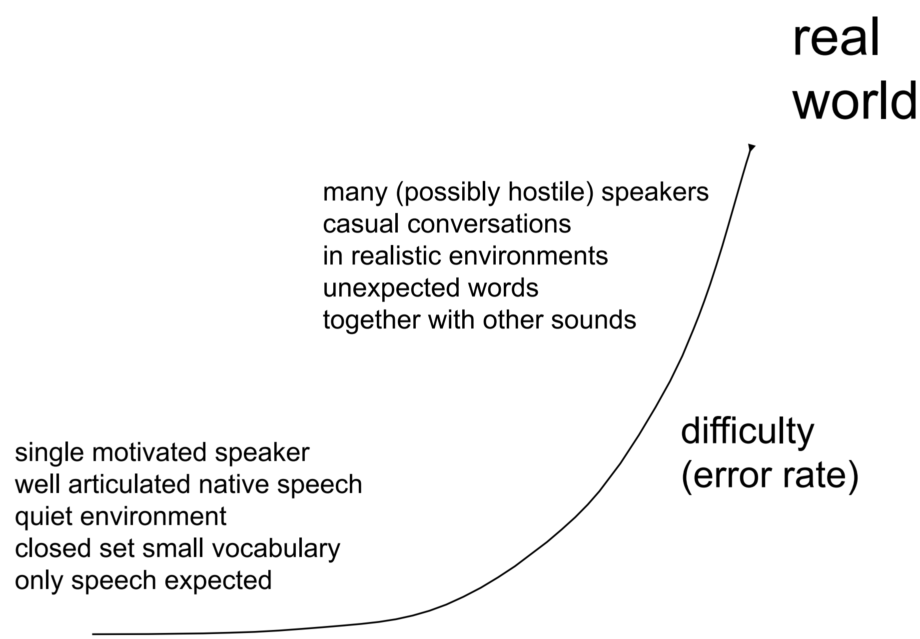


Why to rock the boat? We have good thing going.

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 !

How to Get There ?

Fred Jelinek

Speech recognition
…a problem of maximum likelihood decoding

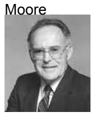
information and communication theory, machine learning, large data,….

Roman Jakobson

We speak, in order to be heard, in order to be understood

human communication, speech production, perception, neuroscience, cognitive science,..

Gordon Moore



The complexity for minimum component costs has increased at a rate of roughly a factor of two per year…

tools

John Pierce

..devise a clear, simple, definitive experiments. So a science of speech can grow, certain step by certain step.

However, also John Pierce:

(Speech recognition is so far (1969) field of) mad inventors or untrustworthy engineers *(because machine need*s) **intelligence and knowledge of language comparable to those of a native speaker .**

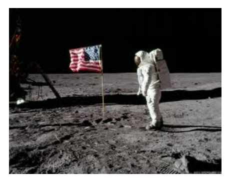
…. should people continue work towards speech recognition by machine ? Perhaps it is for people in the field to decide.

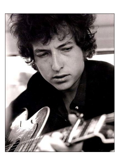
Why Am I Working in Machine Recognition of Speech?

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.

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





Don't Follow Leaders,
Watch the Parking Meters…