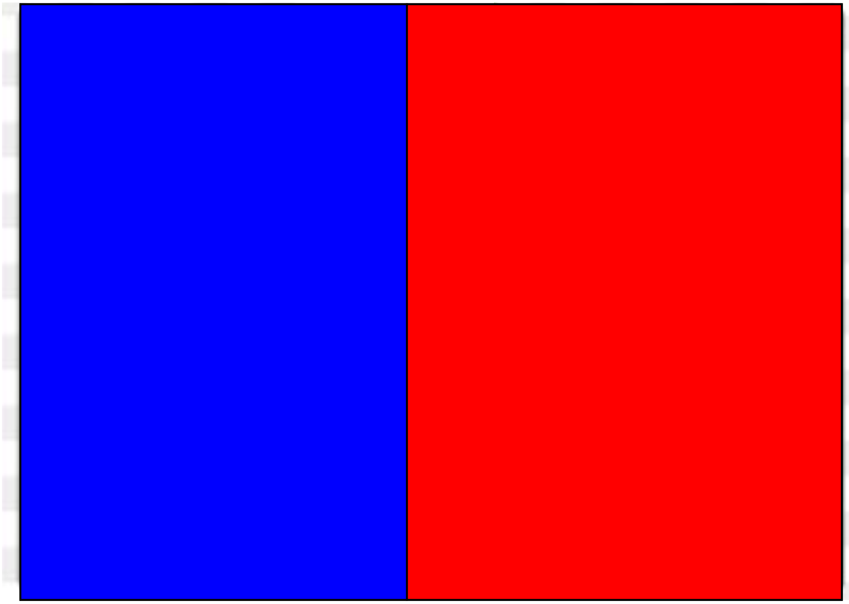
Multistream recognition of speech

Hynek Hermansky
Center for Language and Speech Processing
The Johns Hopkins University, Baltimore, USA
and
FIT VUT Brno Czech Republic

Maxwell demon



LOW ENTROPY

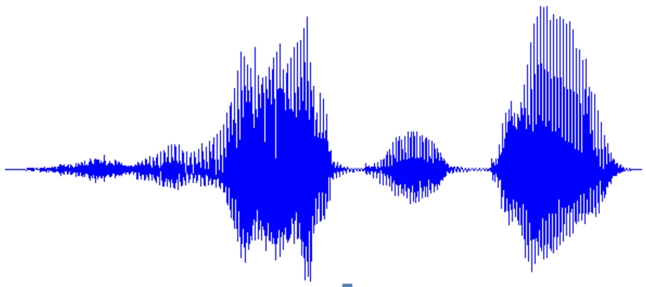
The Demon closes door when a slow air molecule comes and lets the fast air molecules to go through

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

knowledge comes from

- magic
- measurements

When decreasing entropy, one should use knowledge !



machine

message

> 50 kb/s

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

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

< 50 b/s

< 3bits/phoneme, < 15 phonemes/s

linguistic message

Information rate (entropy) reduction

- requires **knowing** what to leave out and how

KNOWLEDGE

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

HARDWIRED

– reusable permanent knowledge
  - no need to re-learn known facts

but

– experts and beliefs can be wrong

DATA

– no knowledge better than wrong knowledge
  - data do not lie

but

– transcribed data are expensive

REUSEABLE AND HARDWAREABLE KNOWLEDGE FROM DATA !

Acoustic Processing in ASR

signal → features → probability estimator (classifier)

features (signal processing)

- what we already know (general knowledge)
- alleviate unwanted information
    - wanted information, which is left out is gone forever

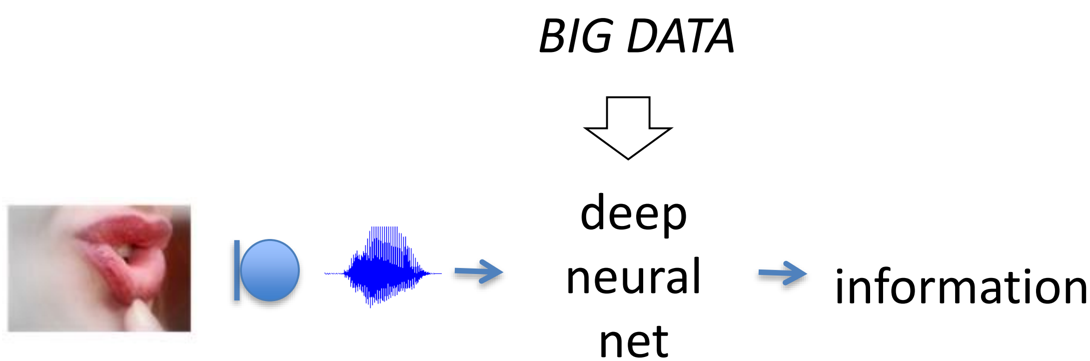
classifier (machine learning)

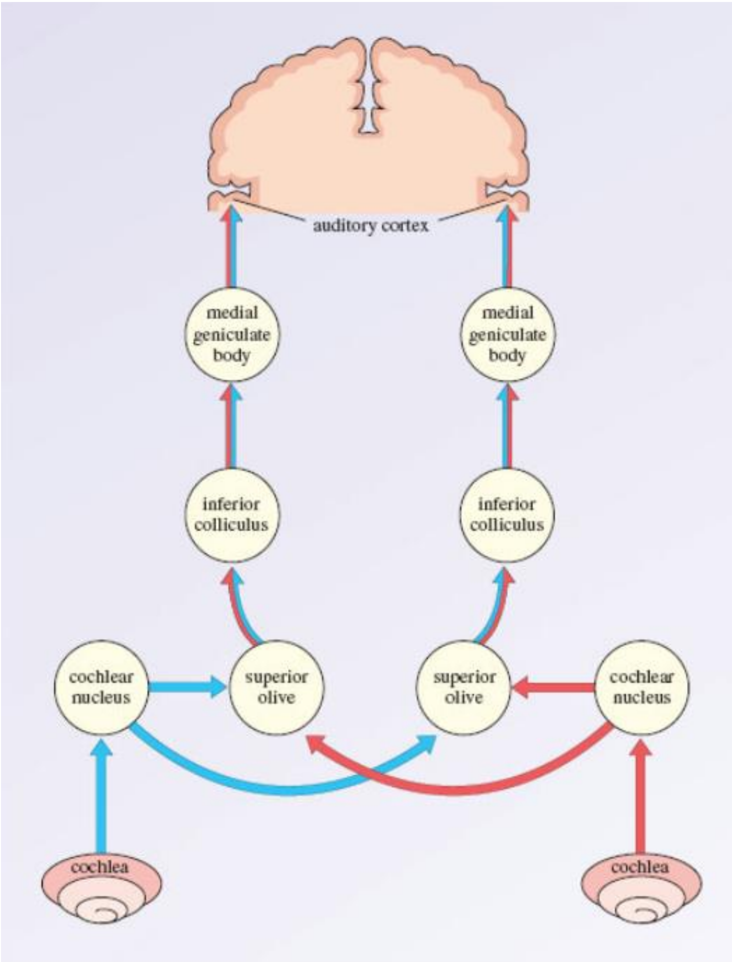
- what we yet do not know (task-specific knowledge)
- typically stochastic (trained on data)
    - unwanted information, which is kept, requires more complex classifiers, trained on more data

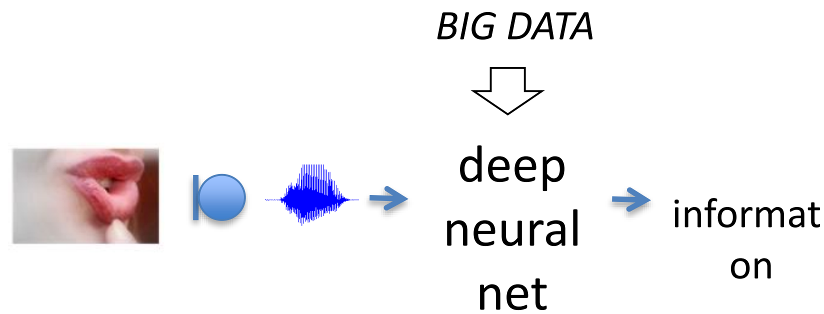
Data-driven approaches dominate ASR field

Artificial Neural Networks

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







Deep Neural Net:

Hierarchical convolutional long-short-memory highway-connected attention-based bi-directional-gated pyramidal temporal-classifying recurrent DNN.

New DNN structures and their parameters

New opportunities to verify existing knowledge and to learn new things .

Data-derived knowledge should be hardwired into future designs !

Deep Neural Network Based ASR from Raw Speech Signal

Tüske, Golik, Schlüter and Ney 2015

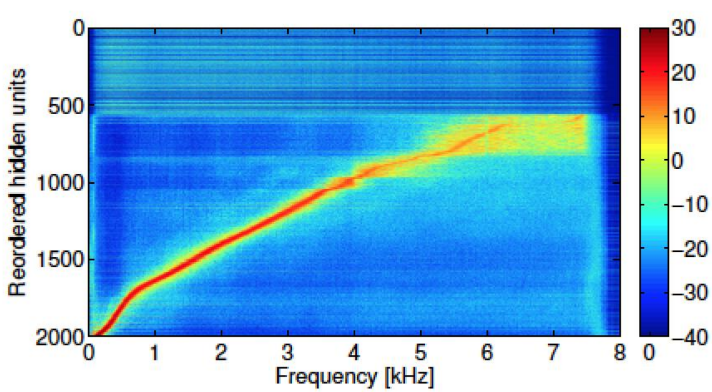
power spectrum

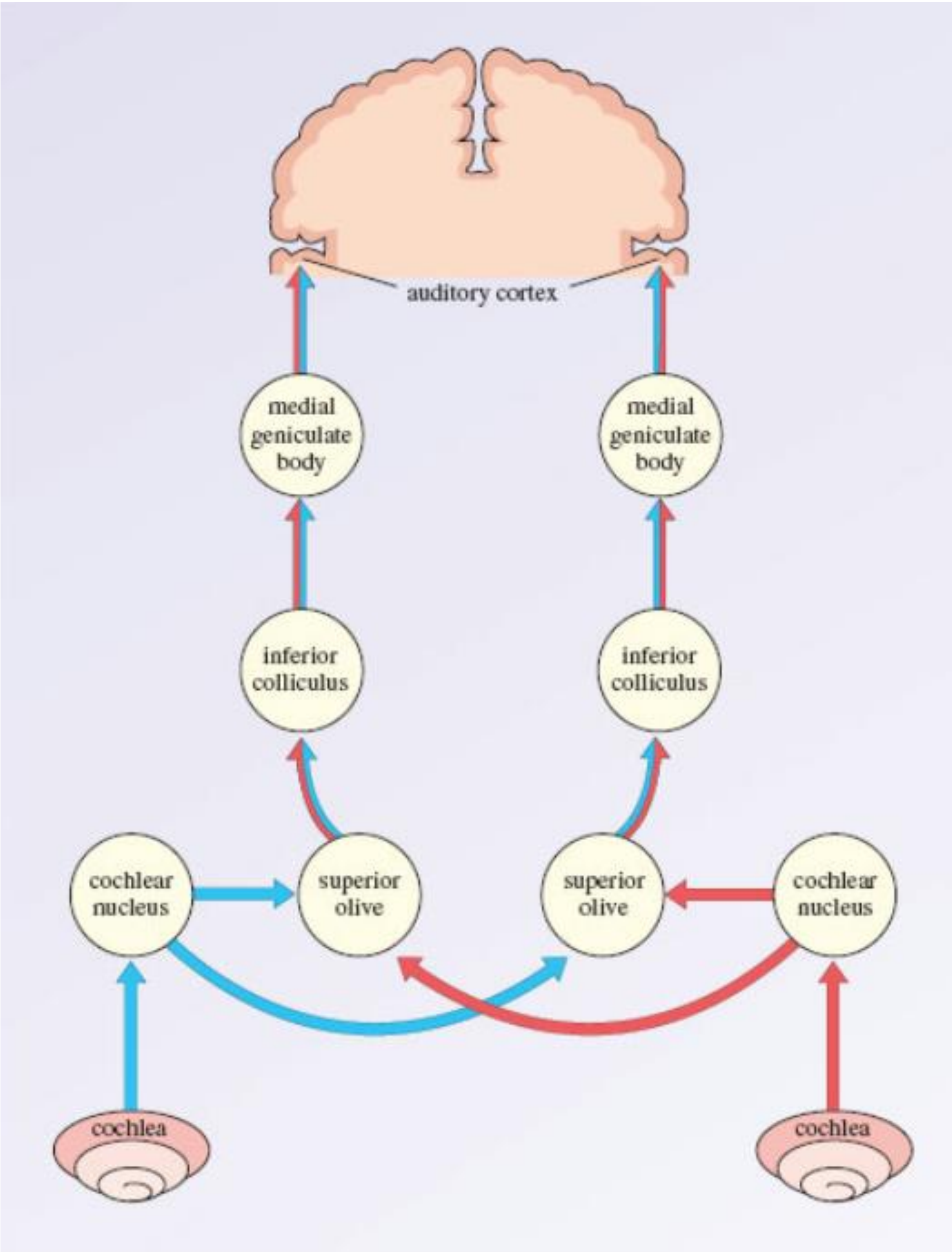
speech

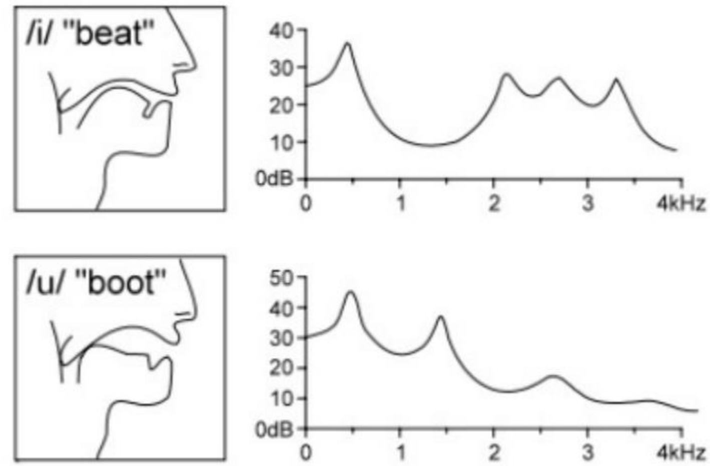
convolutions with input speech signal

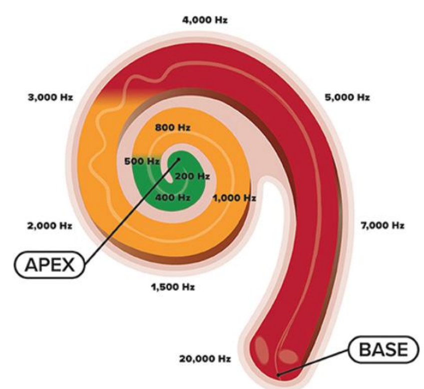
remaining fully connected hidden layers of the deep neural networks

posterior probabilities of generalized tied triphones




auditory cortex
medial geniculate body
medial geniculate body
inferior colliculus
inferior colliculus
cochlear nucleus
superior olive
superior olive
cochlear nucleus
cochlea
cochlea


/i/ "beat"
40
30
20
10
0dB
0
1
2
3
4kHz
/u/ "boot"
50
40
30
20
10
0dB
0
1
2
3
4kHz


4,000 Hz
3,000 Hz
5,000 Hz
800 Hz
500 Hz
200 Hz
400 Hz
1,000 Hz
2,000 Hz
7,000 Hz
APEX
1,500 Hz
20,000 Hz
BASE

Data-driven two-stage acoustic processing of raw speech signal (spectrum and time-frequency cortical-like filters)

Golik, Tüske, Schlüter and Ney 2015

power spectrum

time-frequency speech representation

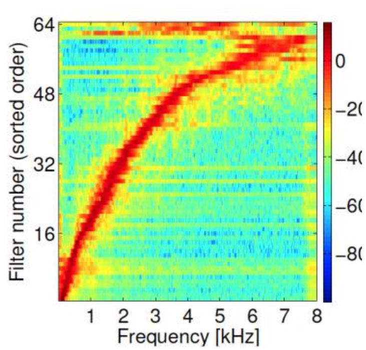
speech

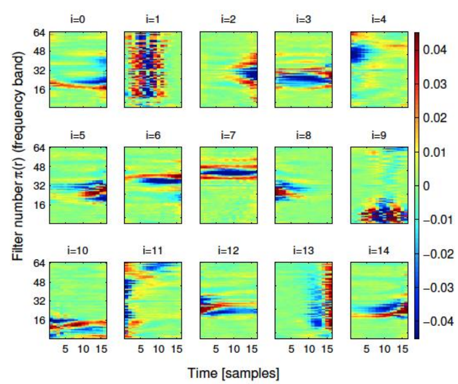
convolutions with input speech signal

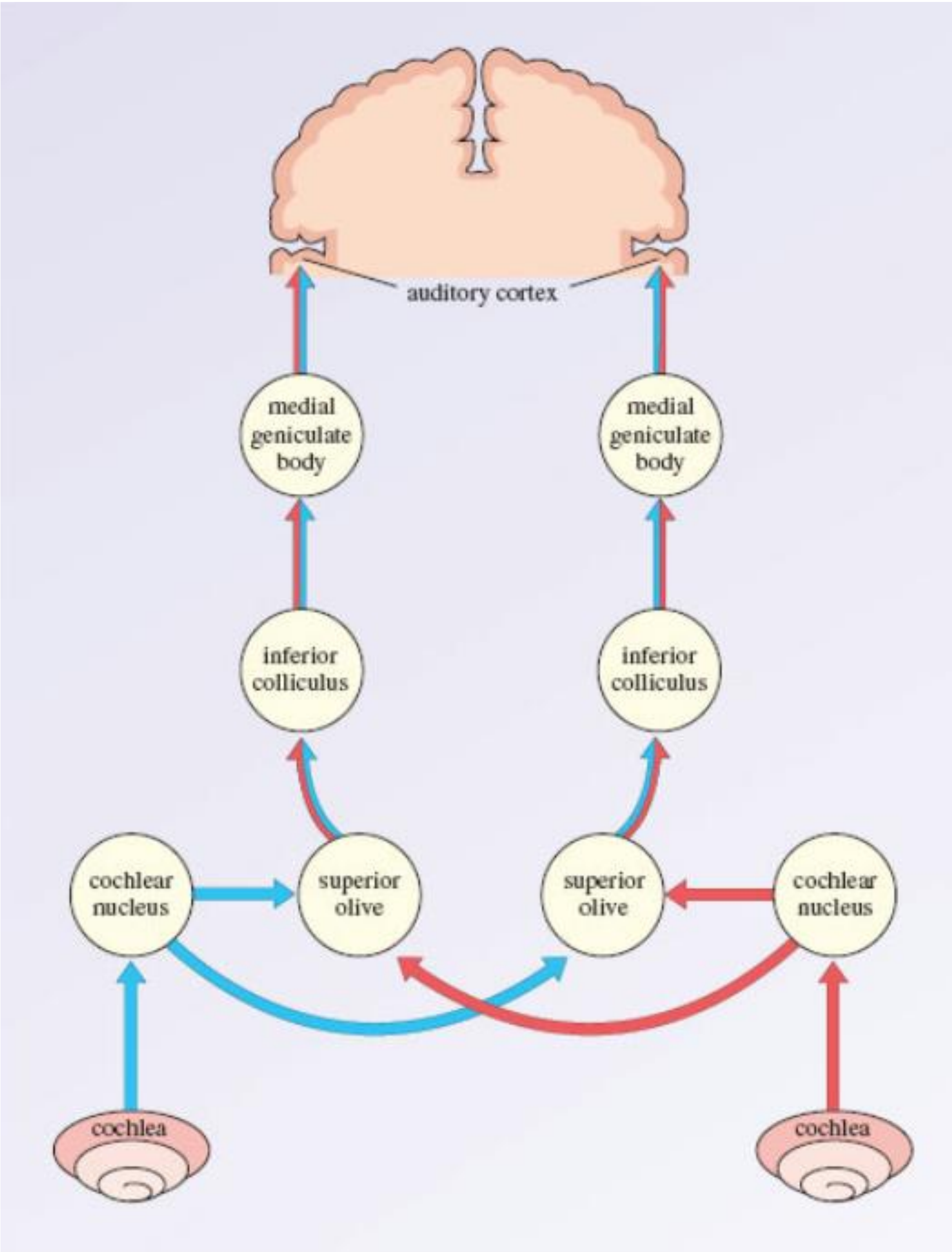
convolutions with time trajectories of power spectra

remaining fully connected hidden layers of the deep neural networks

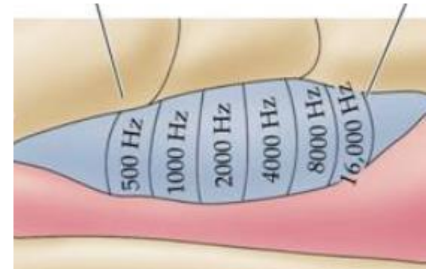
posterior probabilities of generalized tied triphones

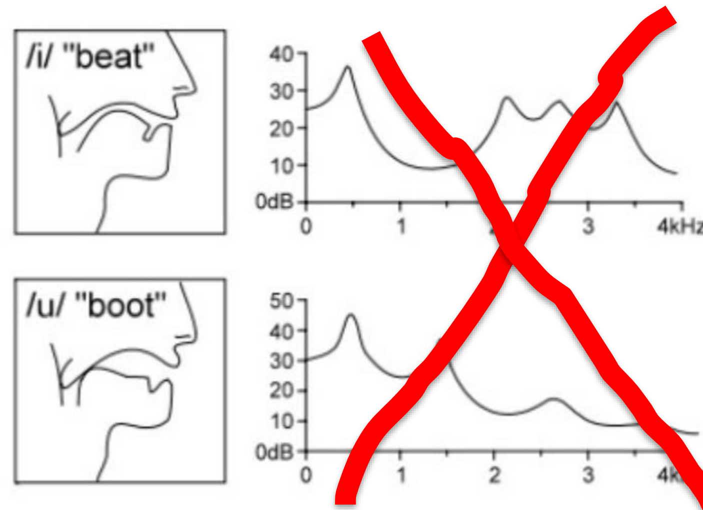


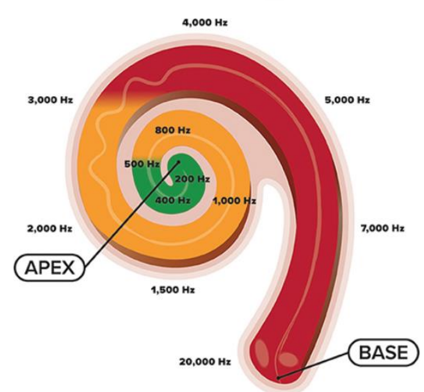


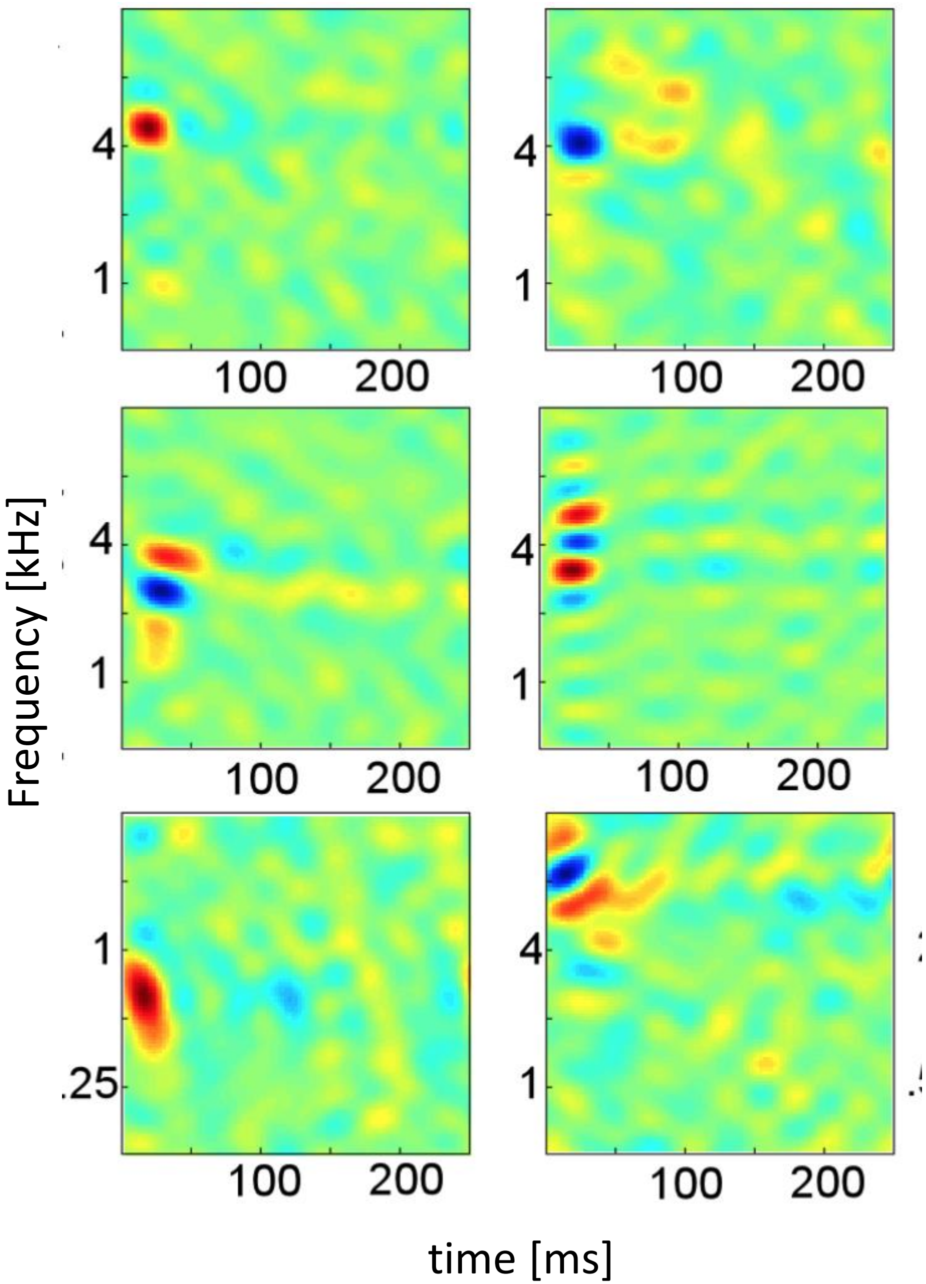

auditory cortex
medial geniculate body
medial geniculate body
inferior colliculus
inferior colliculus
cochlear nucleus
superior olive
superior olive
cochlear nucleus
cochlea
cochlea

APEX BASE


500 Hz
1000 Hz
2000 Hz
4000 Hz
8000 Hz
16,000 Hz


/i/ "beat"
40
30
20
10
0dB
0
1
3
4kHz
/u/ "boot"
50
40
30
20
10
0dB
0
1
2
3


4,000 Hz
3,000 Hz
5,000 Hz
800 Hz
500 Hz
200 Hz
400 Hz
1,000 Hz
2,000 Hz
7,000 Hz
APEX
1,500 Hz
20,000 Hz
BASE



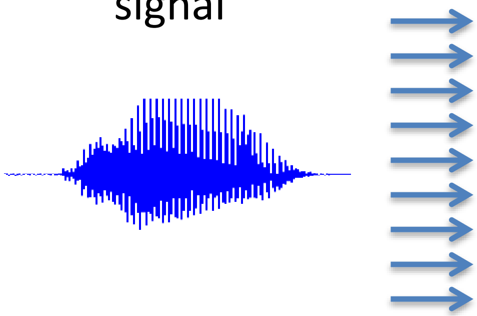
Some examples of mammalian auditory cortical receptive fields

Patil et al 2012

Spectral (simultaneous) masking

signal

high frequencies



spectral
energies in
critical bands

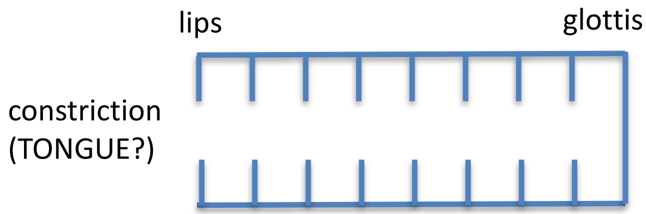
low frequencies

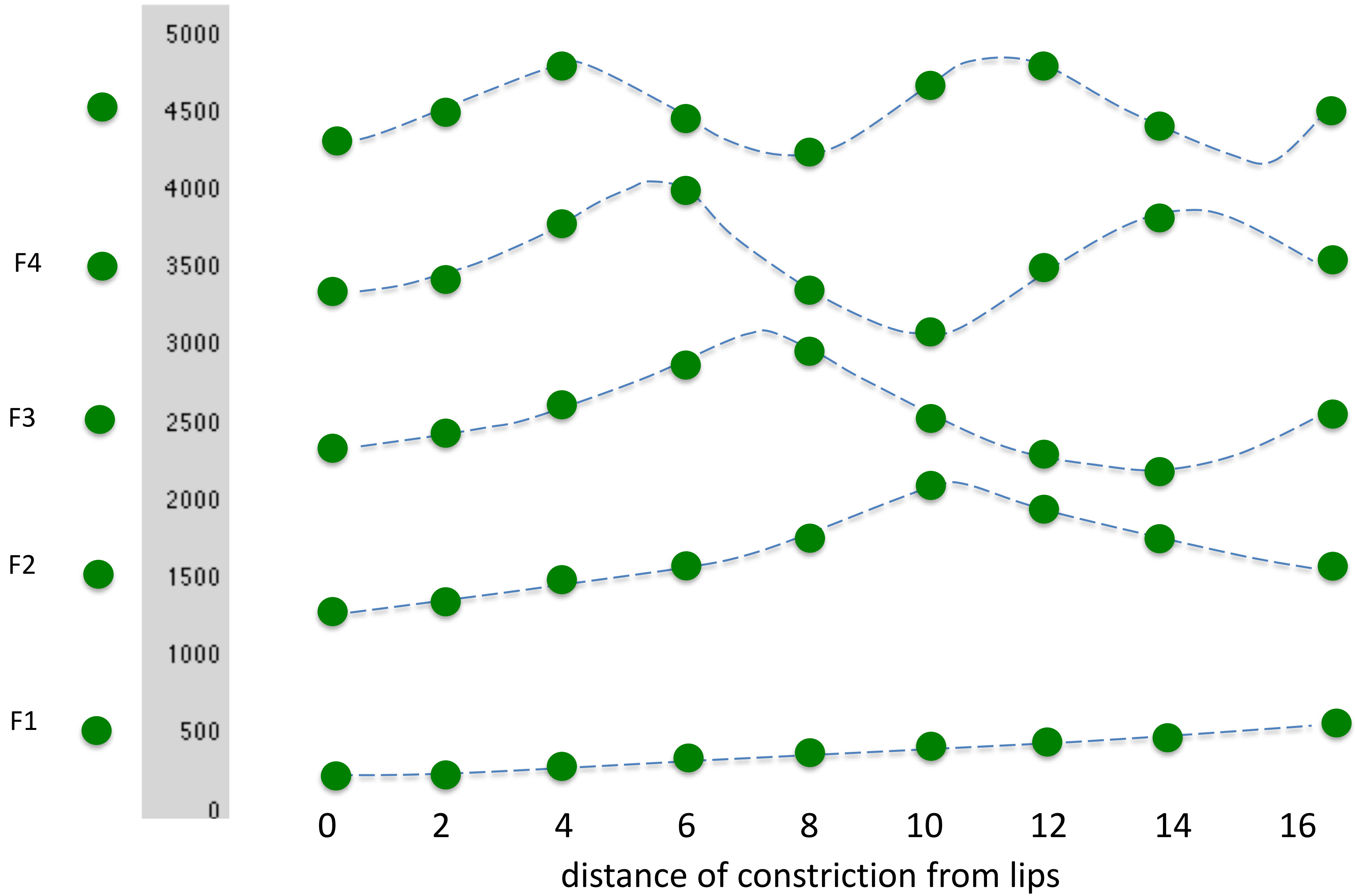
spectral masking:

detection of signal in one critical band is not influenced by signal in another critical band

Fletcher 1933

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



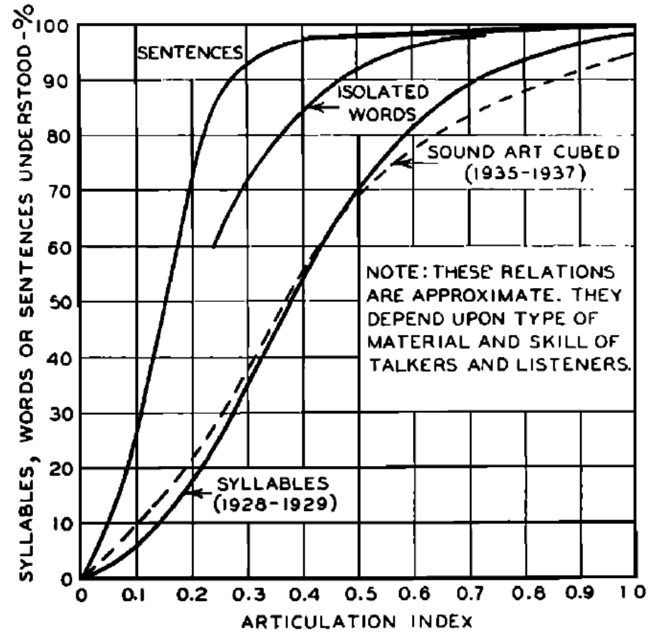


Articulatory Bands

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

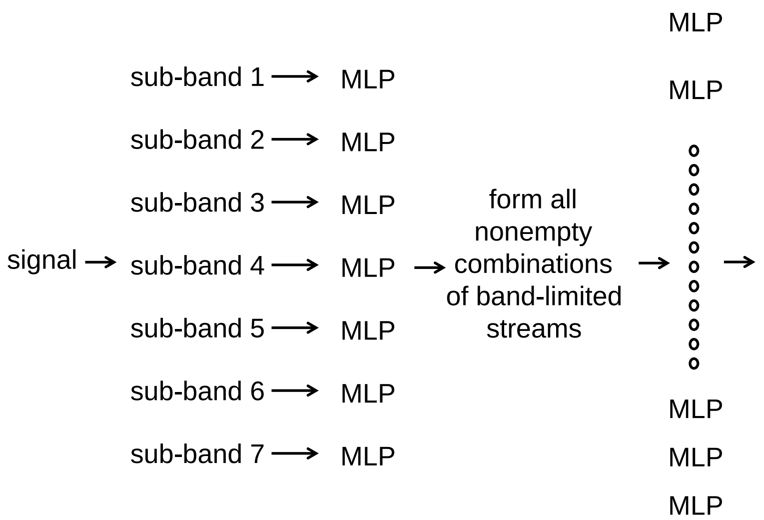
- 20 frequency bands in speech spectral region
- each band contributes about equally to human speech recognition
- any 10 bands sufficient for 70% correct recognition of nonsense syllables, better than 95% correct recognition of meaningful sentences [Fletcher and Steinberg 1929]

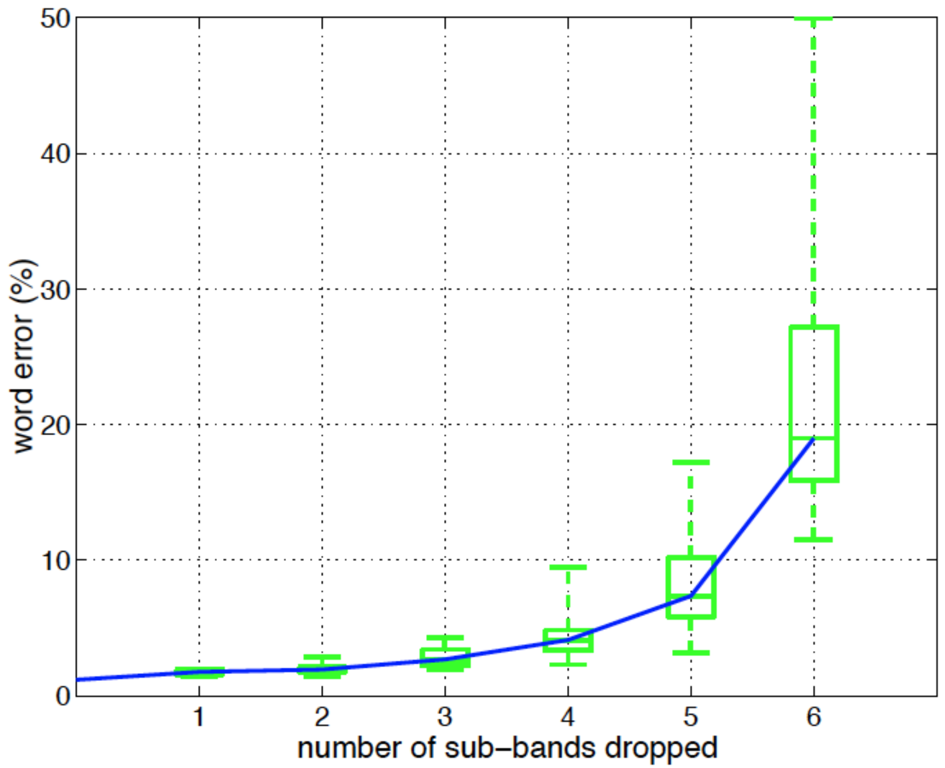


127 different stream combinations in hierarchical MLP structures

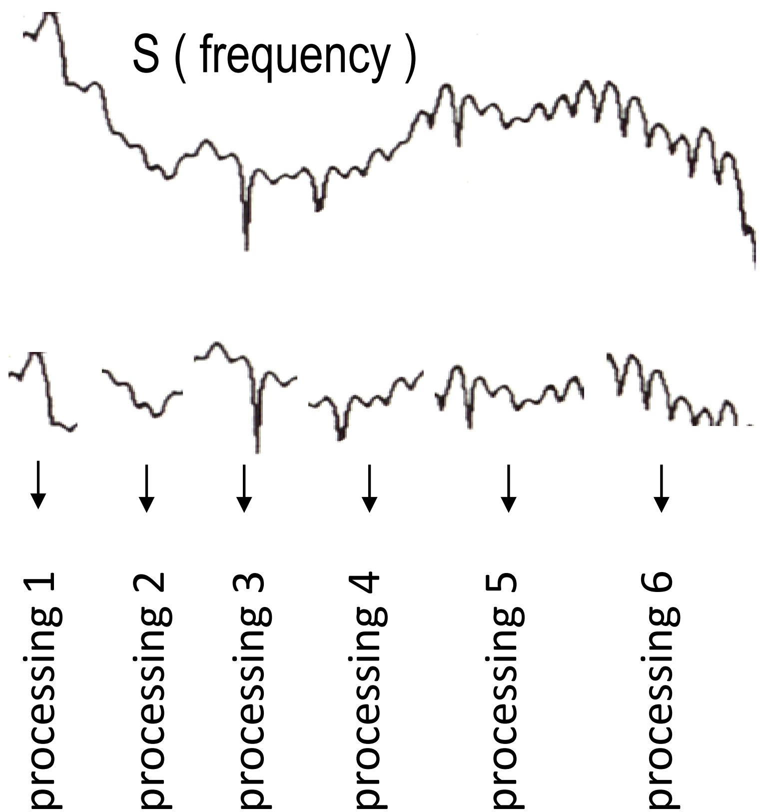
evaluate word error for different stream combinations

Hermansky et al 1996

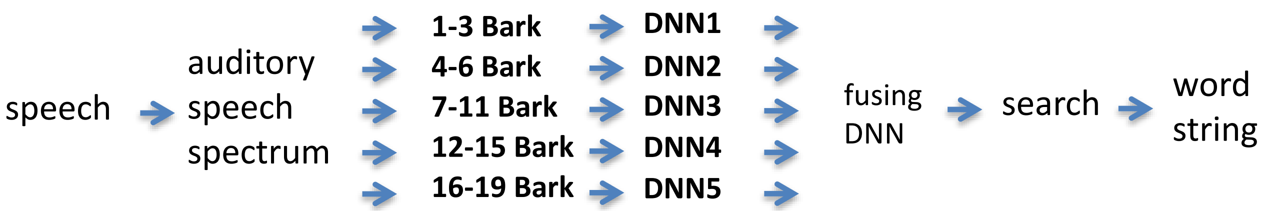




Human Recognition Strategy (and eventually also machines) ? Divide et Impera



- colored noise can be seen as close to white noise in individual bands
- corrupted frequency bands could be left out from further processing



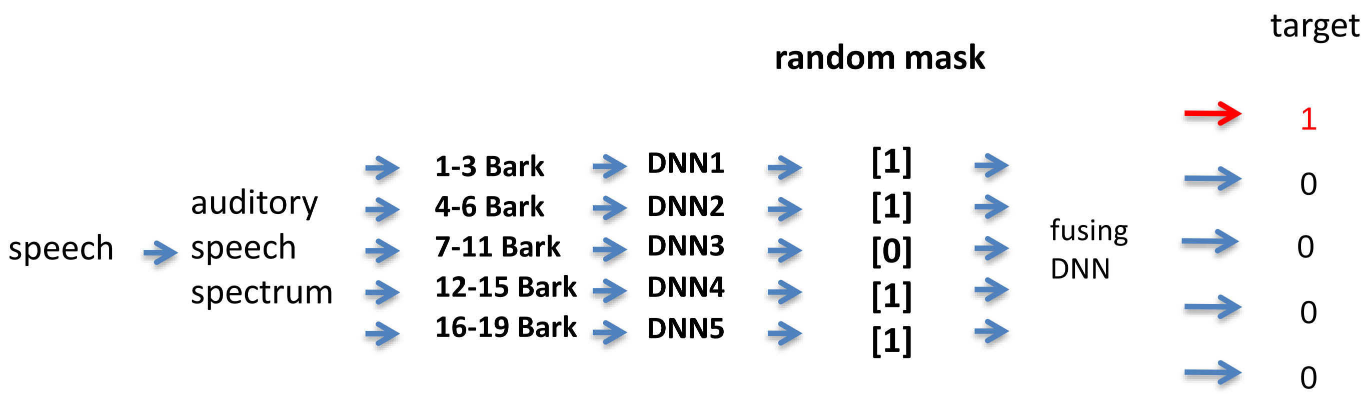
Word error rates of DNN recognizer
on Aurora noisy data (relative change in brackets)

auditory spectrum	spectral streams
12.6	11.0 (-12.8)

Sri Harish Mallidi, JHU PhD Thesis,
in preparation

Some of the streams may carry garbage

Train fusing DNN on inputs, which carry no information.
During training, randomly set some stream outputs to all-zero.



Similar to feature dropping but here the whole organized sets of features representing streams are being dropped at any given time.

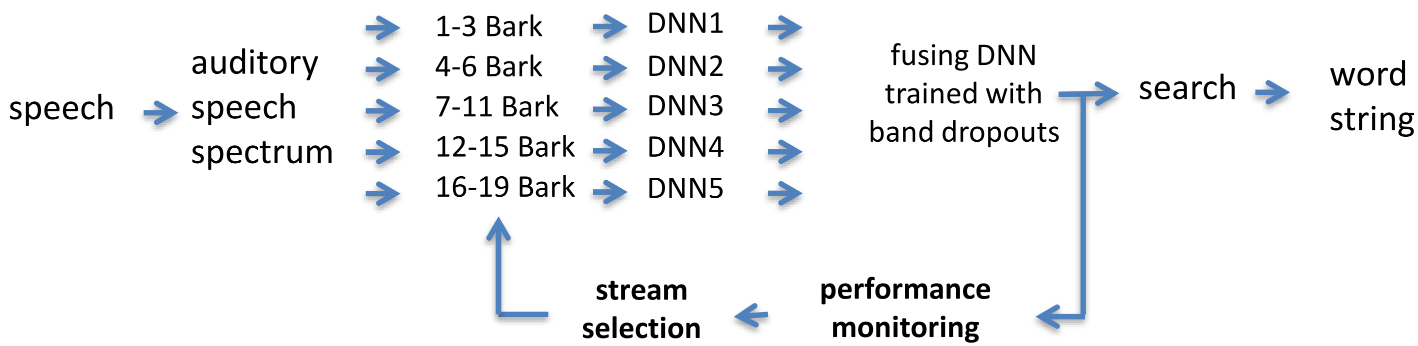
Word error rates of DNN recognizer
on Aurora noisy data (relative change in brackets)

auditory spectrum	spectral streams	stream dropping
12.6	11.0	9.9
	(-12.8)	(-10.1)

Sri Harish Mallidi, JHU PhD Thesis,
in preparation

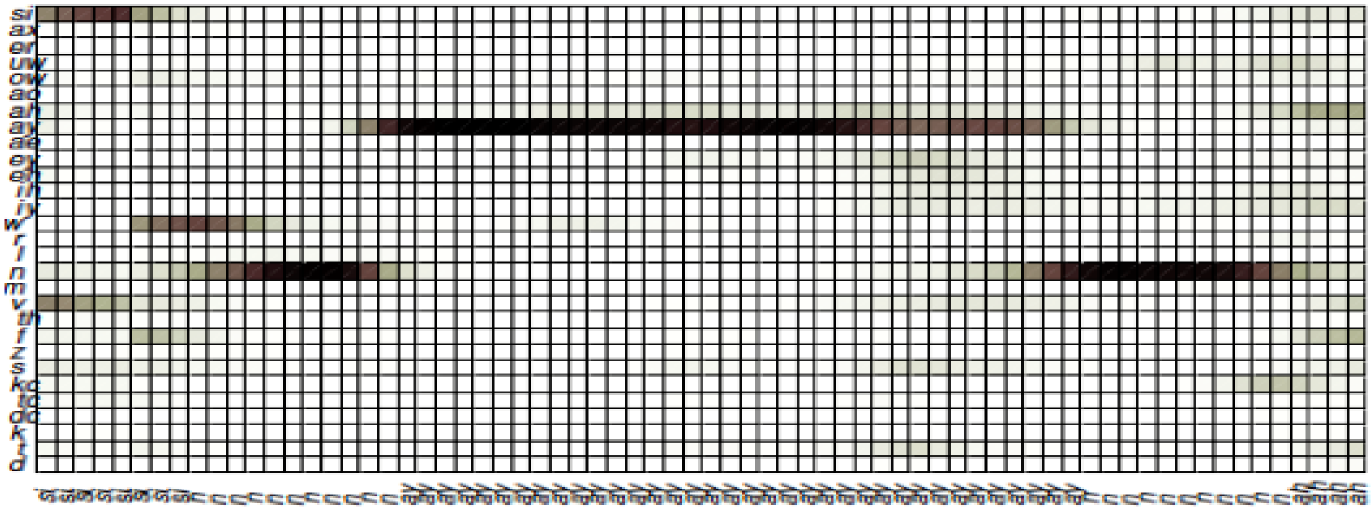
Performance monitoring

Knowing when the result in probability estimation is in error would allow for the selection of the best performing stream combination

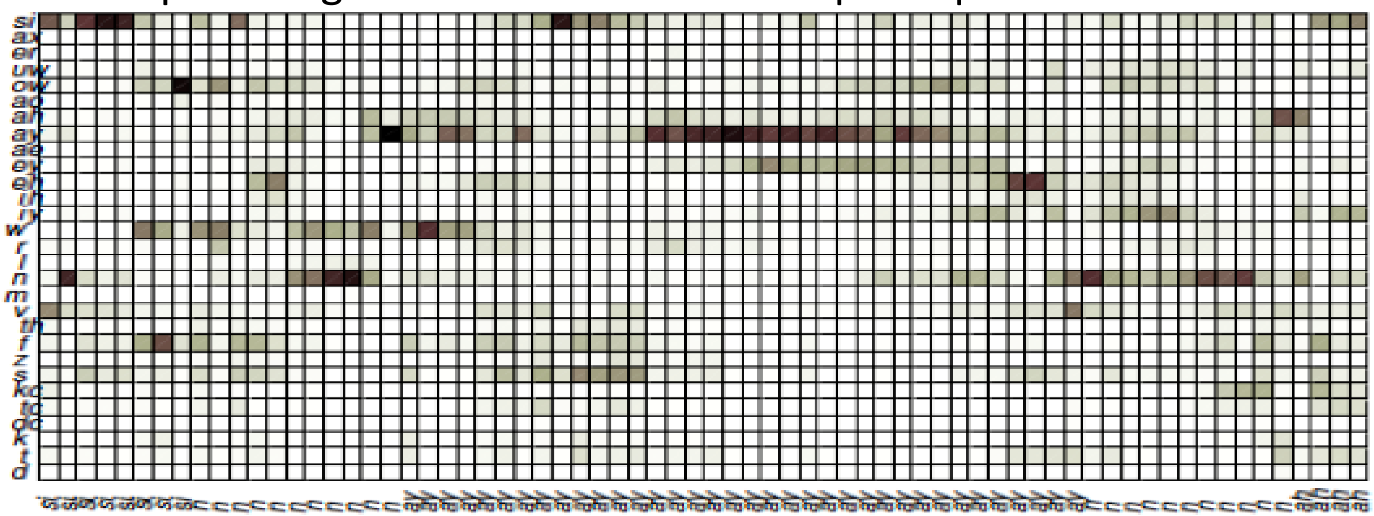


Performance monitoring :

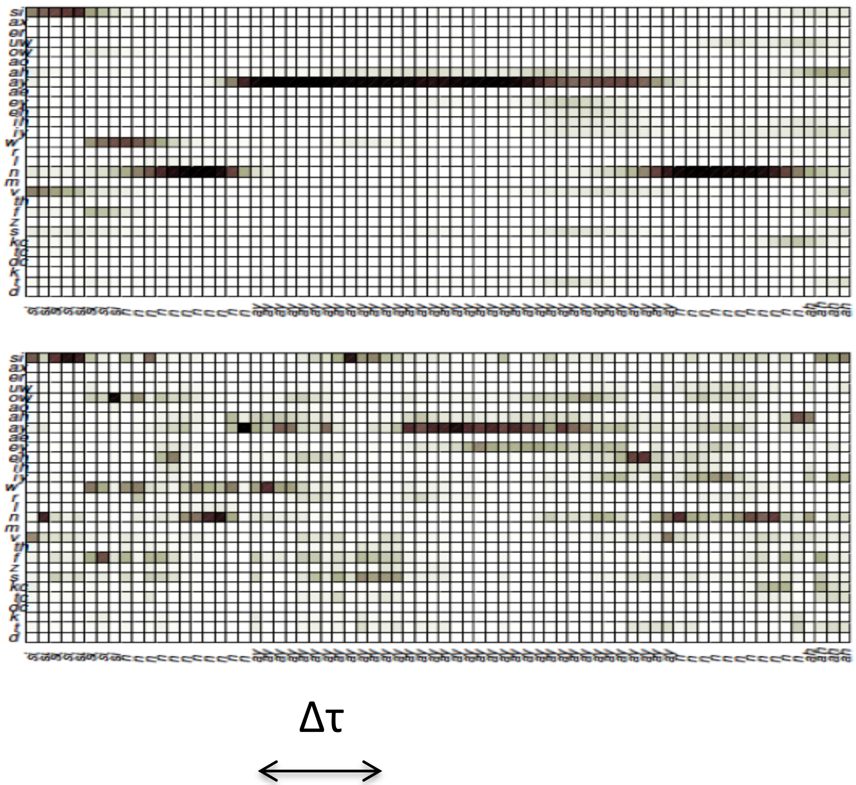
requires estimation of performance of a classifier without knowing what the correct result is “good” posteriogram – derived from speech data similar to its training



“bad” posteriogram – derived from corrupted speech data



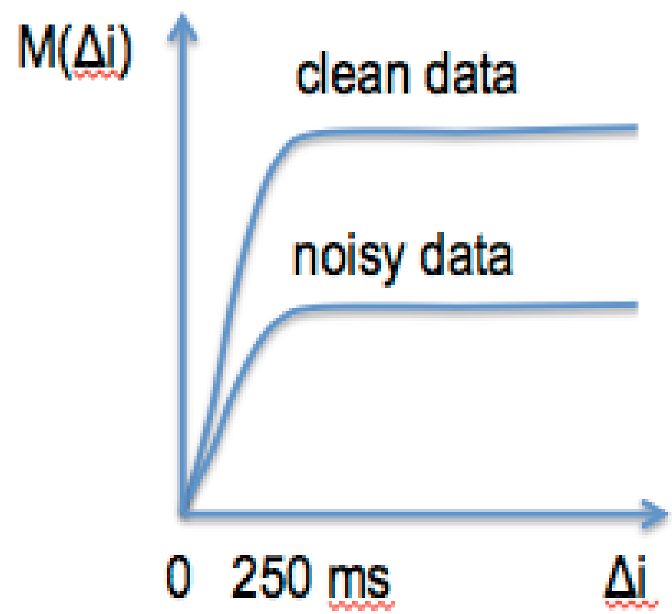
How “clean” is a posteriogram ?



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

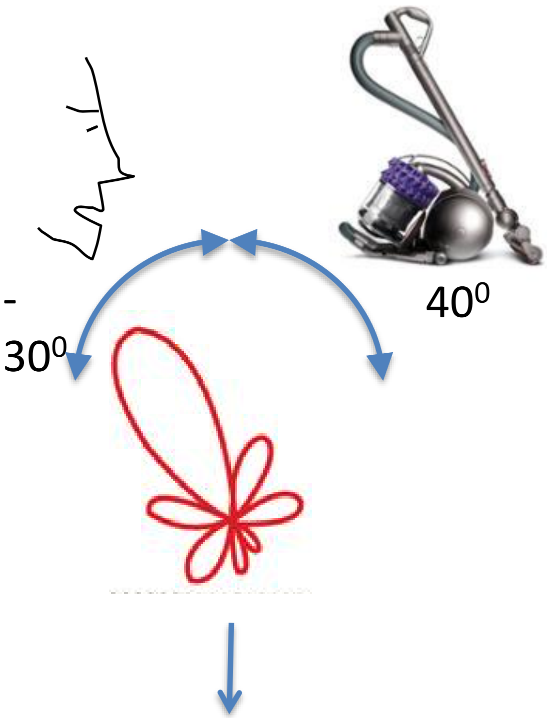
Δi – time delay

D(.) – symmetric Kl divergence



Quality of speech signal from microphone array

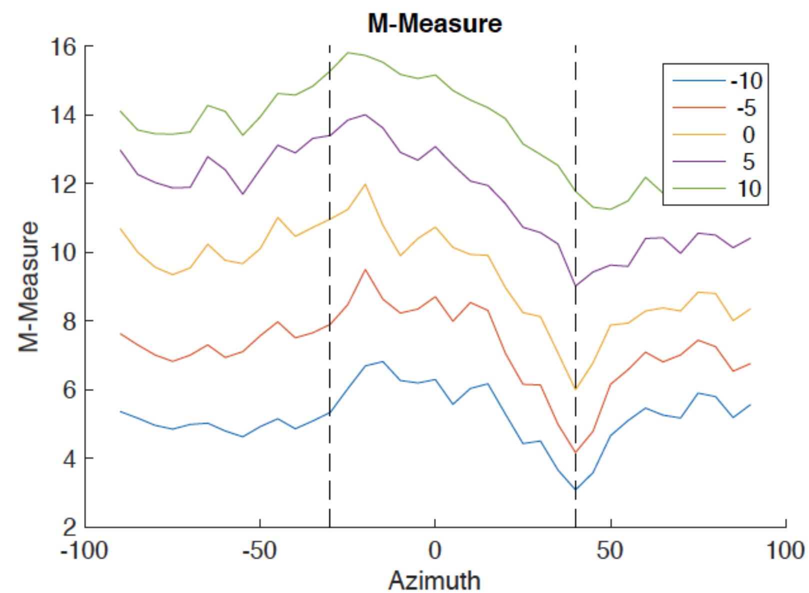
from Bernd T. Meyer



performance
monitoring
module

speaker

noise
source



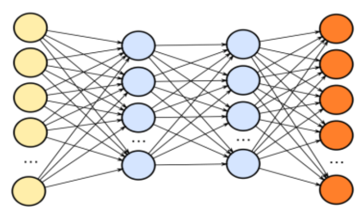
How “similar” is the estimator performance on its training data and in the test?

Mesgarani et al 2011

DNN auto-encoder, trained on output of the estimator when applied to its training data

training of probability estimator

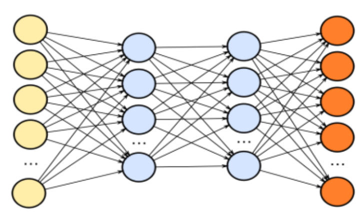
training data



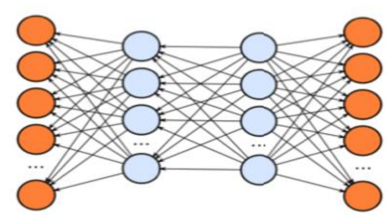
p targets from labels

training of performance monitor

training data



p'

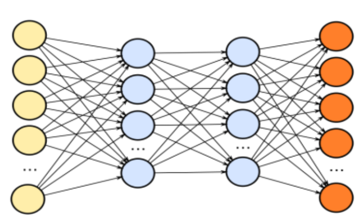


p''

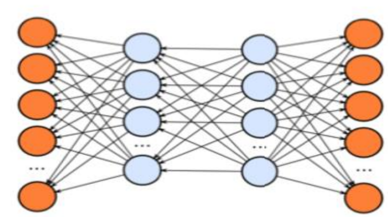
train to minimize $(p' - p'')^2$

performance monitor in use

test data

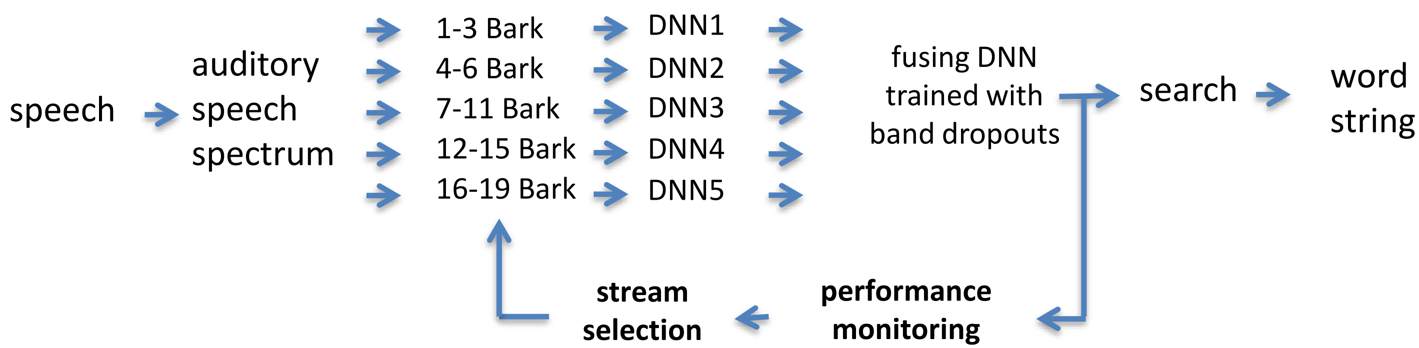


p'_{test}



p''_{test}

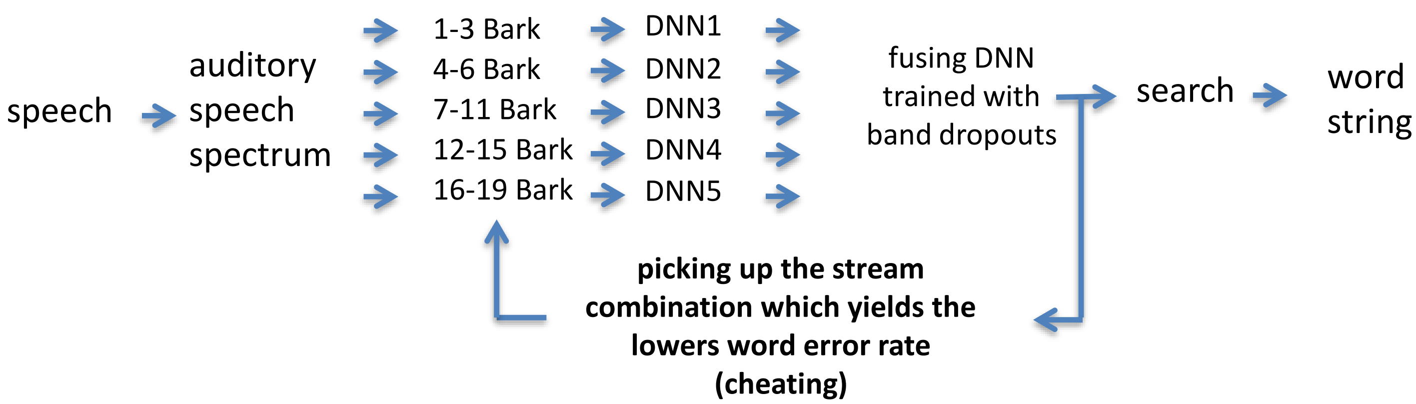
evaluate $(p'_{test} - p''_{test})^2$



Word error rates of DNN recognizer
on Aurora noisy data (relative change in brackets)

auditory spectrum	spectral streams	stream dropping	performance monitoring
12.6	11.0	9.9	9.6
	(-12.8)	(-10.1)	(-2.8)

Sri Harish Mallidi, JHU PhD Thesis,
in preparation

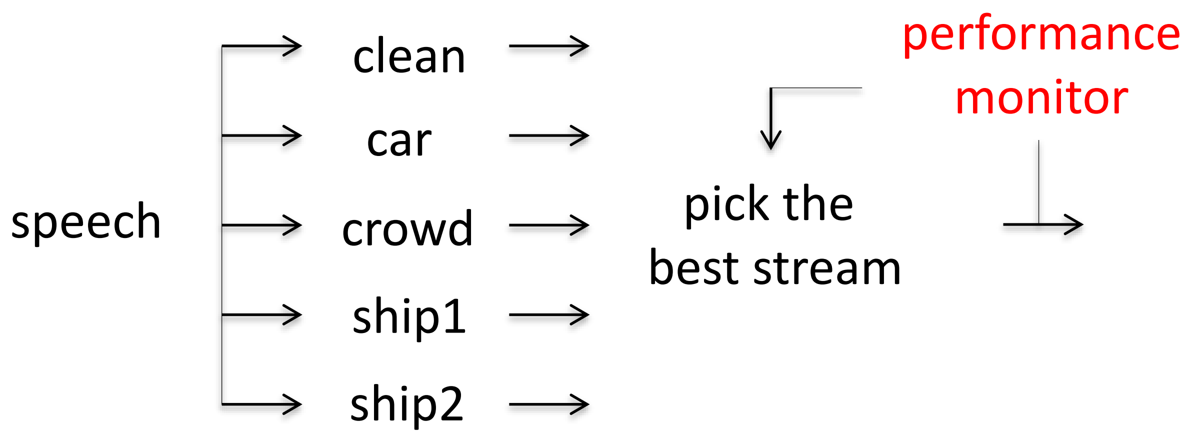


Word error rates of DNN recognizer
on Aurora noisy data (relative change in brackets)

auditory spectrum	spectral streams	stream dropping	performance monitoring	**oracle band selection**
12.6	11.0 (-12.8)	9.9 (-10.1)	9.6 (-2.8)	7.9 (-18.0)

Sri Harish Mallidi, JHU PhD Thesis,
in preparation

Multiple parallel noise-specific streams

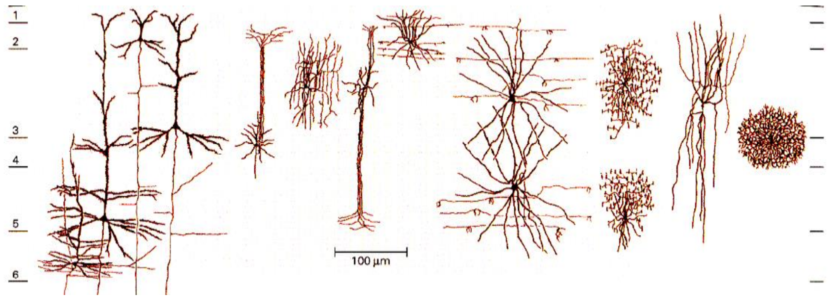


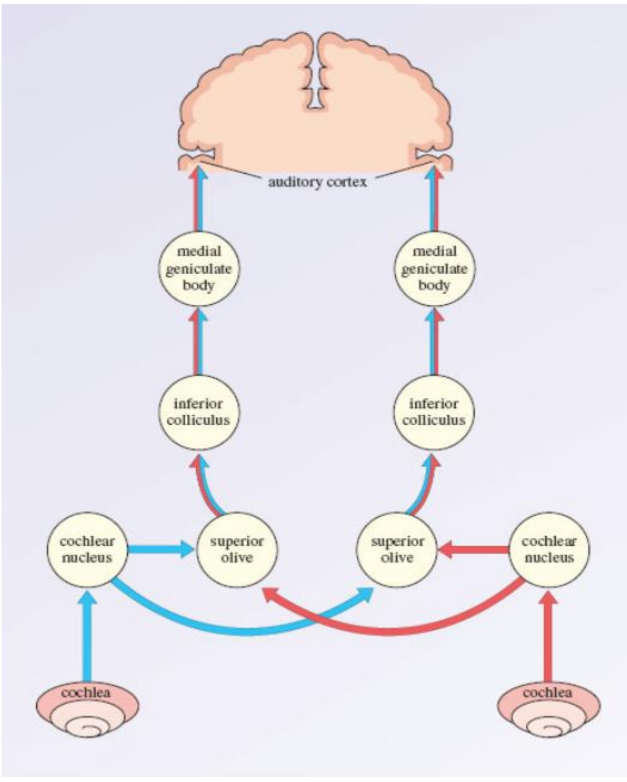
phoneme error rates noisy TIMIT

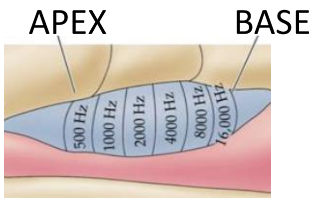
train / test	clean	car	crowd	ship1	ship2
multi-style	23.0	24.9	39.4	42.0	43.0
matched	20.7	22.8	37.0	38.1	37.6
oracle (cheating)	18.4	20.5	34.7	34.5	31.8
multi-stream with	**20.9**	**22.9**	**36.8**	**36.6**	**36.8**

Mallidi et al ASRU 2015

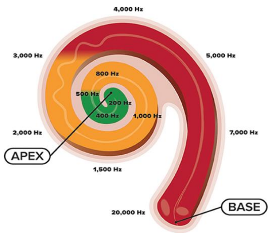
Many ways of seeing the signal



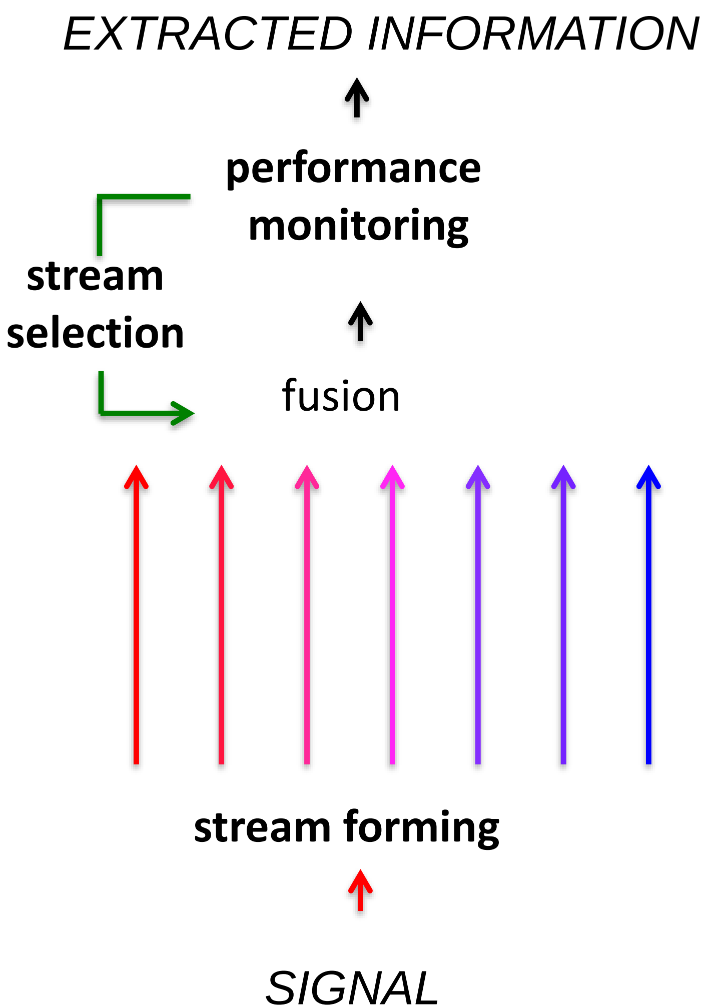




number of neurons	speed of firing
100 M	*10 Hz*
100K	*1 kHz*



Concept of multi-stream recognition



different streams

- modalities,
- frequency bands,
- spectral and temporal resolutions,
- levels of prior knowledge

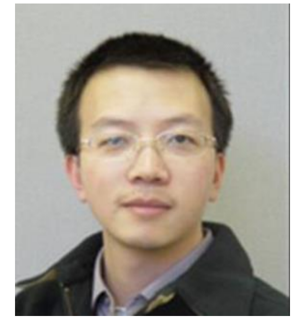
THANKS


Sri Harish Mallidi


Nima Mesgarani


Tetsuji Ogawa


Samuel Thomas


Feipeng Li


Ehsan Variani


Phani Nidadavolu


Vijay Peddinti


Bernd T Meyer

Regarding the database:

The training set consists of 14 hours of multi-condition data, sampled at 16 kHz.

Total 7137 utterance from 83 speakers.

Half of the utterances were recorded by the primary Sennheiser microphone and the other half were recorded using one of a number of different secondary microphones.

Both halves include a combination of clean speech and speech corrupted by one of six different noises (street traffic, train station, car, babble, restaurant, airport) at 10-20 dB signal-to-noise ratio.

The test set consist of 14 conditions, with 330 utterances for each condition. The conditions include clean set recorder with primary Sennheiser microphone, clean set with secondary microphone, 6 additive noise conditions which include airport, babble, car, restaurant, street and train noise at 5-15 dB signalto-noise ratio (SNR) and 6 conditions with the combination of additive and channel noise

Regarding the features:

From signal extract 63 Mel filterbank energies

At a given frame, take 11 frame context (-5 , +5)

In each subband project the 11 frame context onto 6 dct basis