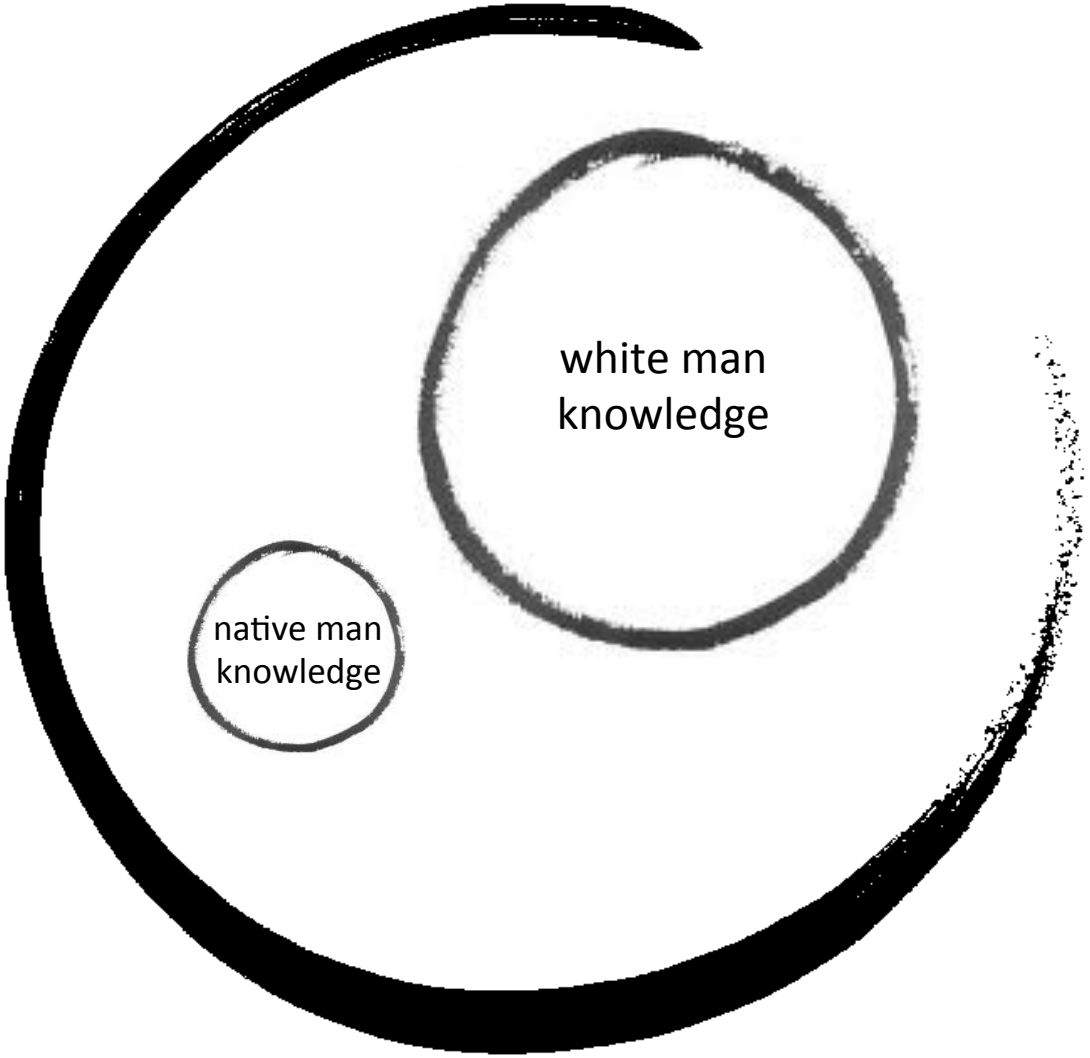
Towards Machines Which Know When They Do not Know

Hynek Hermansky

The Johns Hopkins University

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



speech → sounds → message → meaning

machine recognition of speech

assumption:

- the world did not change since the machine was trained

"unknown unknowns" for the machine:

- data distortions that were not seen in the training
- words that are not in the lexicon of the machine

one possible solution:

- on-line adaptation to new situations

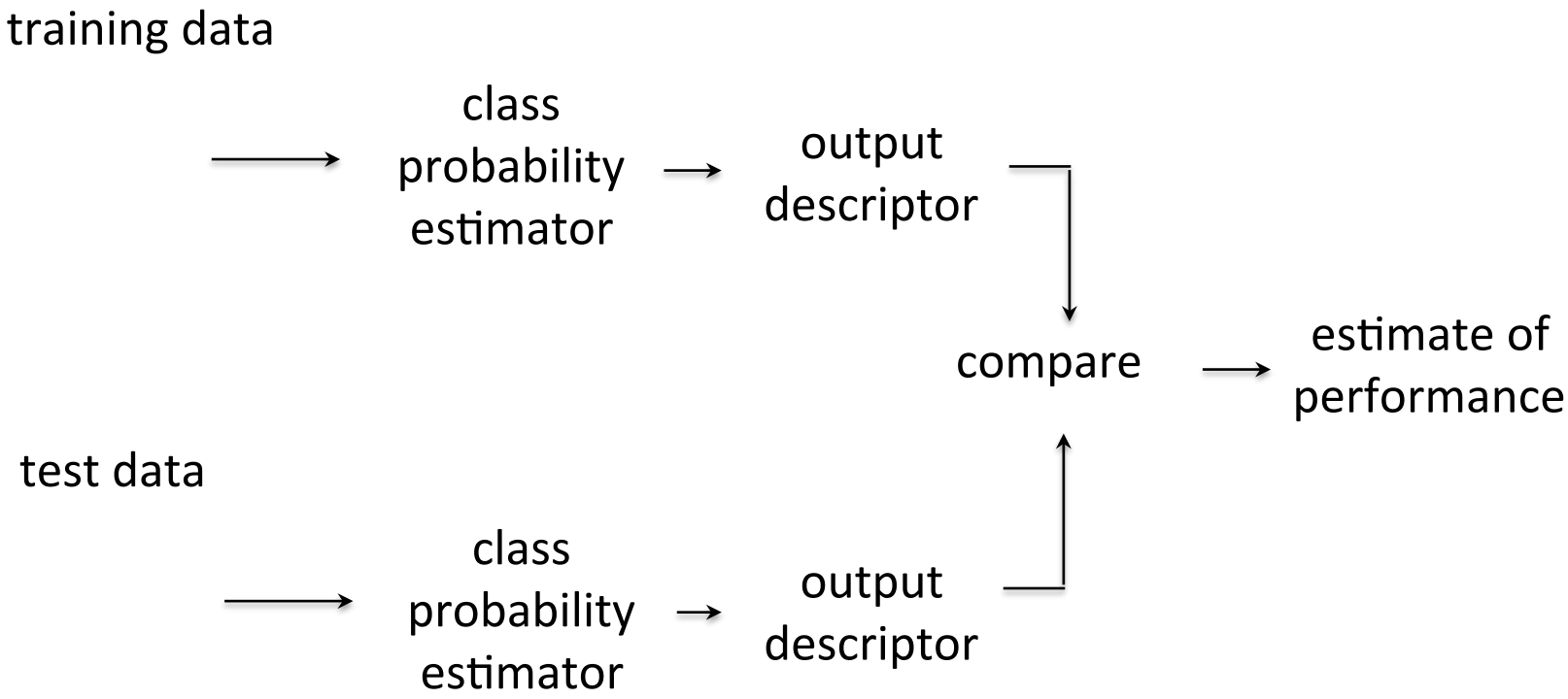
Unsupervised adaptation

- generative models
  - modify models to increase their likelihoods on unseen data
- non-generative models (e.g. neural nets)
  - modify models so that the estimates still make sense ?????

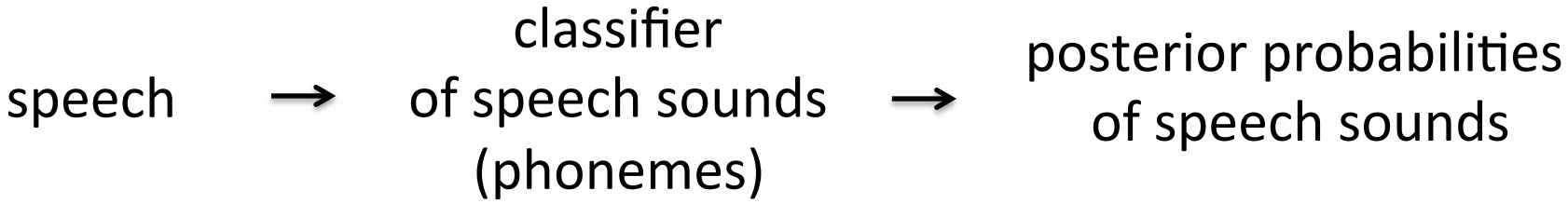
Do the estimates make sense?

(Performance monitoring)

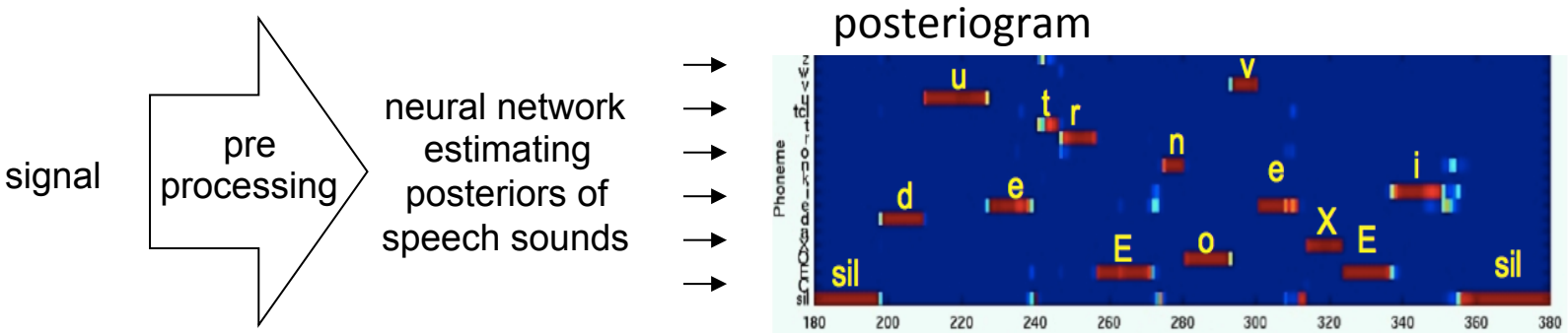
1. we know some characteristics of the expected estimates
2. **estimates that are observed on the data on which the models were trained**


training data
class probability estimator
output descriptor
compare
estimate of performance
test data
class probability estimator
output descriptor

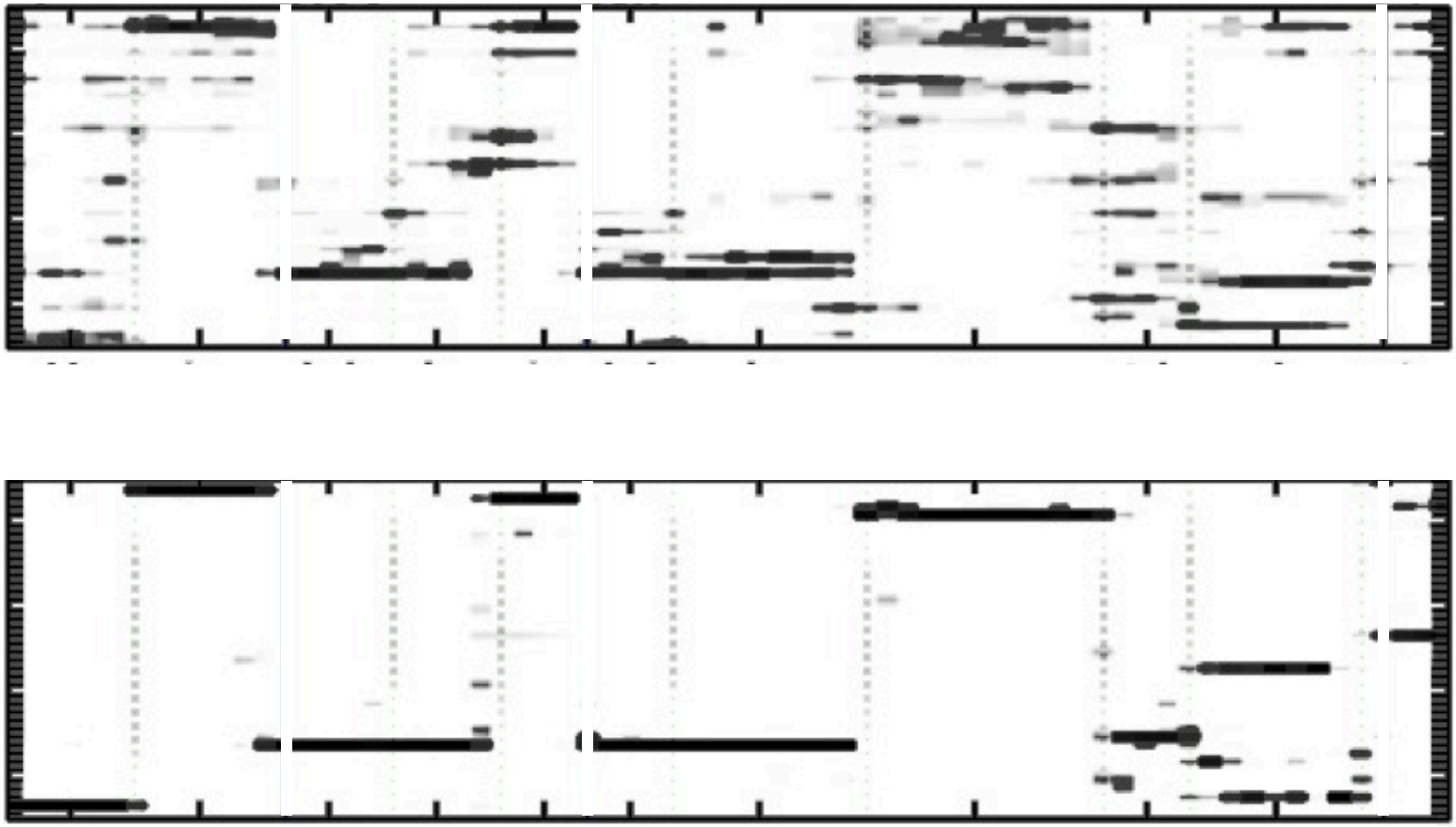
Problem Statement



- **Having a classifier that yields a frame based vectors of posterior probabilities for speech sounds of interest, predict the accuracy of these estimates without knowing the correct probabilities on test data but knowing performance of the classifier on the training data.**



- Do not know the correct answer (ground truth)
  - Is the result good or bad?
  - How long does it take to get this info?



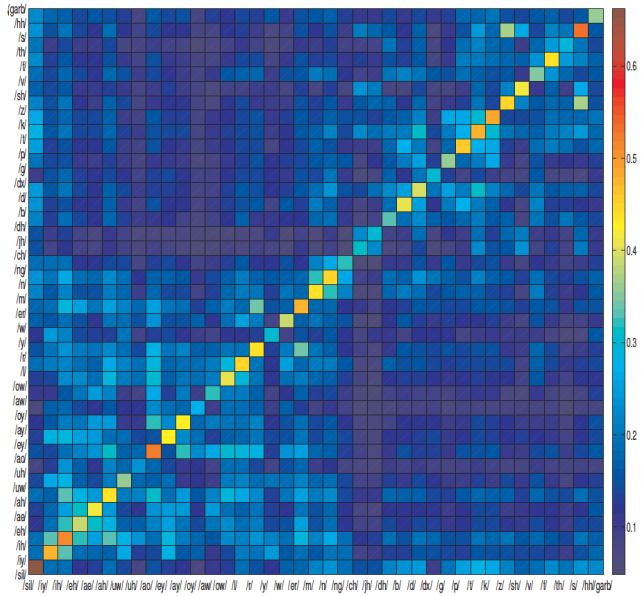
Evaluating Performance

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

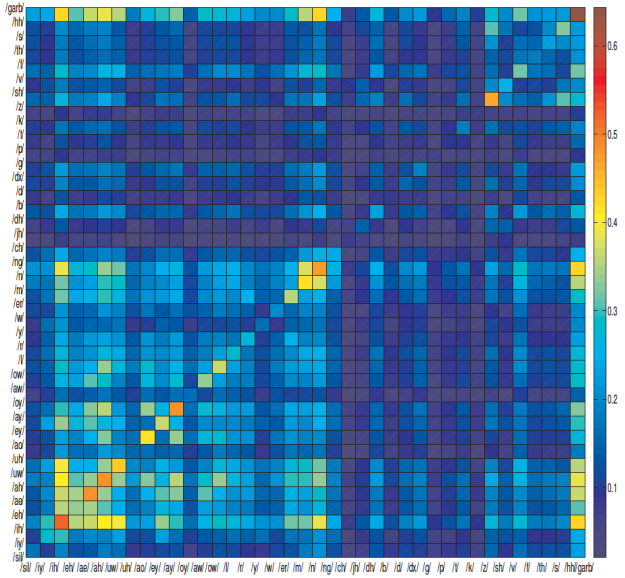
$$AC = 1/N \sum_{i=1}^{N} (\mathbf{p}_i)^r (\mathbf{p}_i^T)^r$$

$\boldsymbol{p}_i$ – 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)

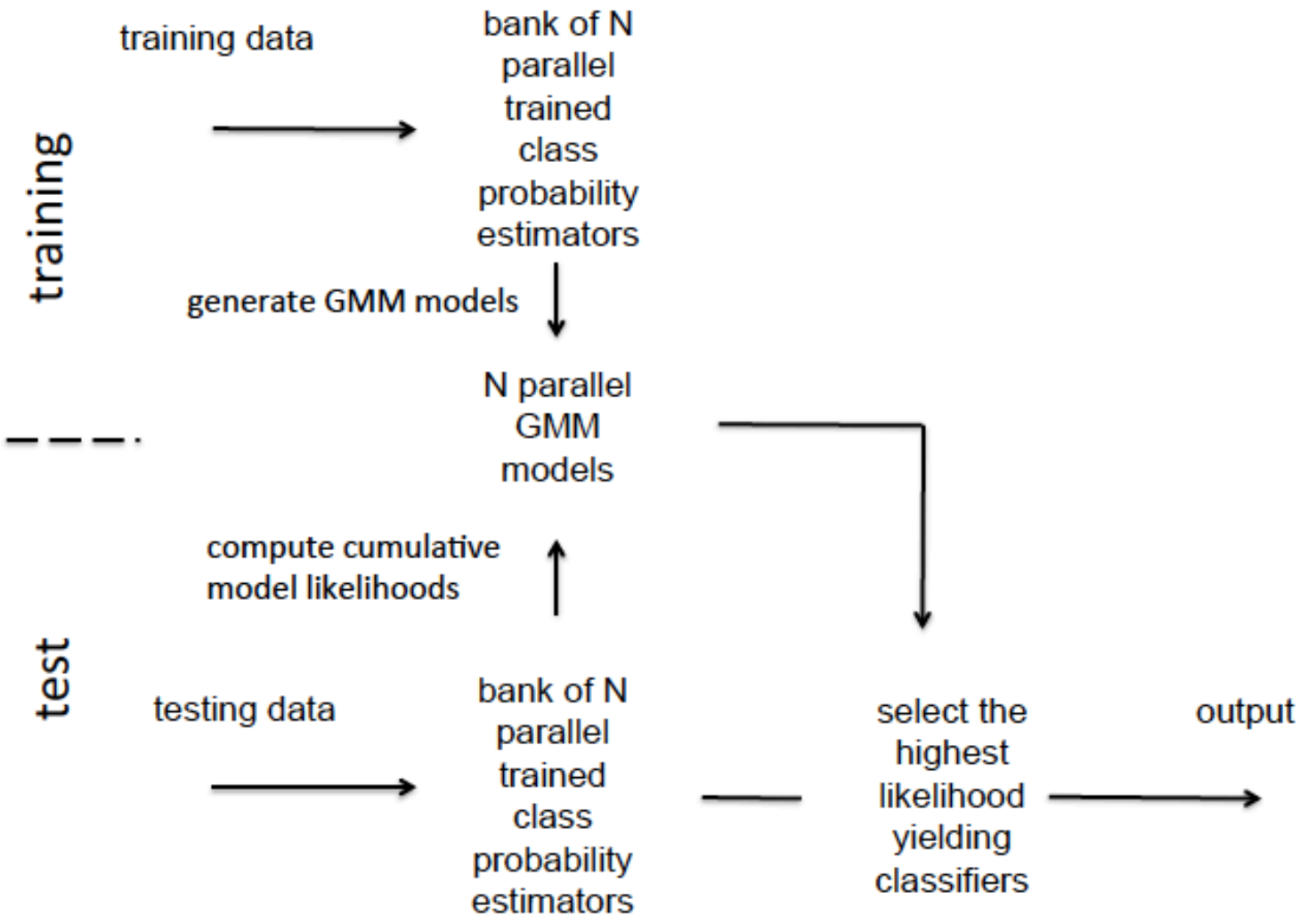
clean data



noisy data



compare matrixes derived on training data and on test

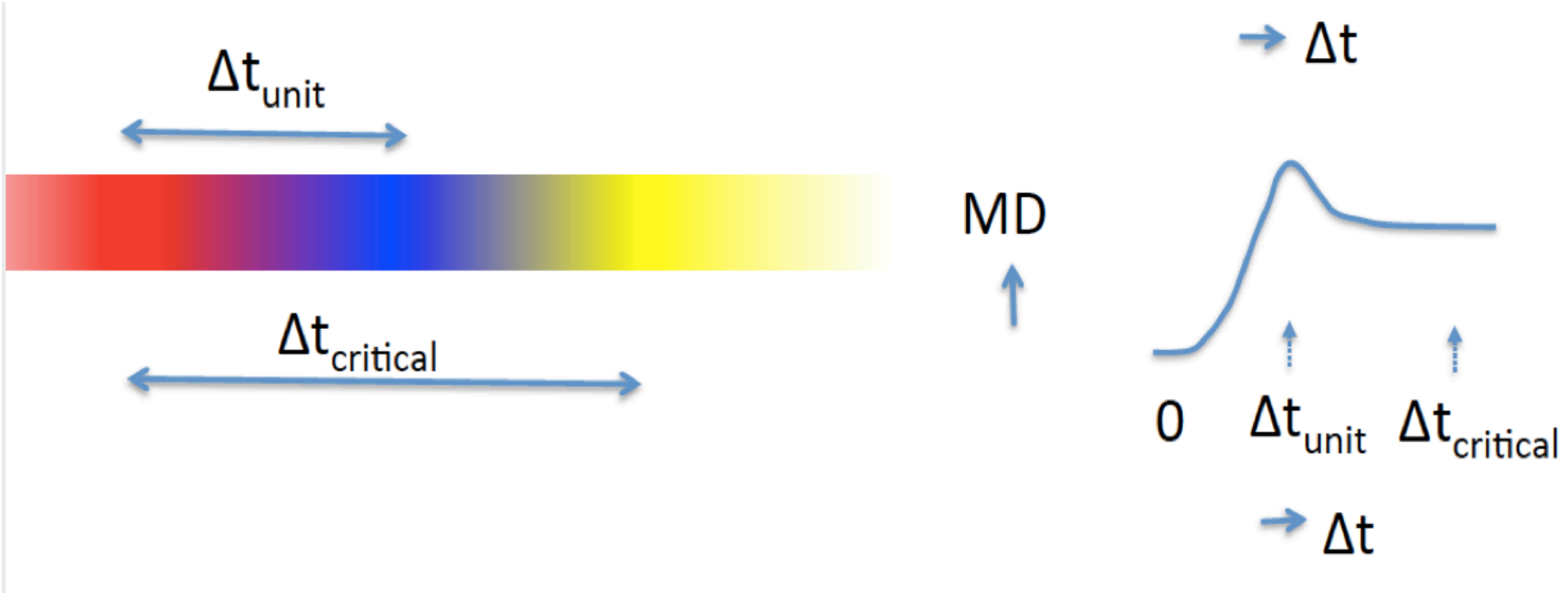

training
training data
bank of N parallel trained class probability estimators
generate GMM models
N parallel GMM models
compute cumulative model likelihoods
test
testing data
bank of N parallel trained class probability estimators
select the highest likelihood yielding classifiers
output

Evaluating Performance

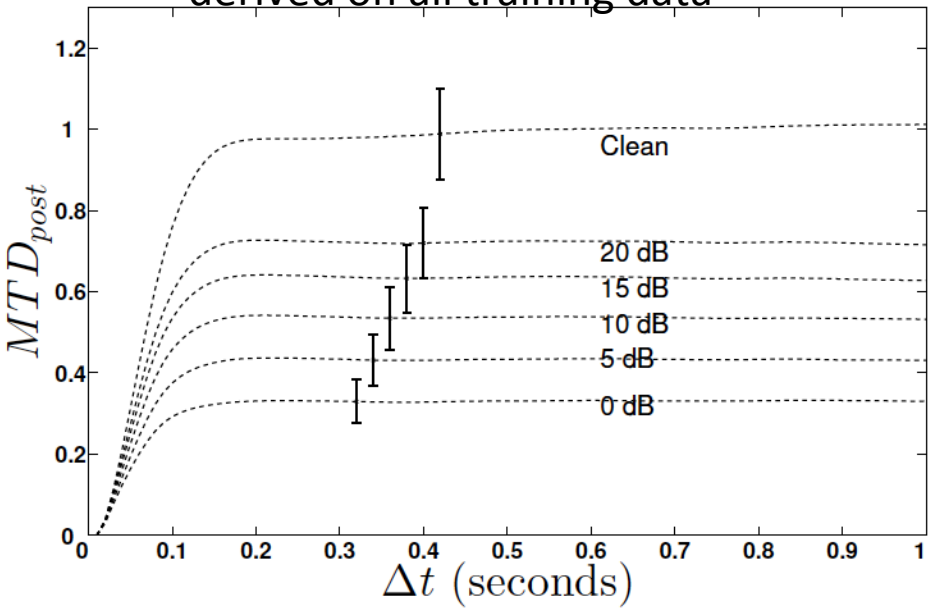
How much sounds classes differ and how fast do they change?

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

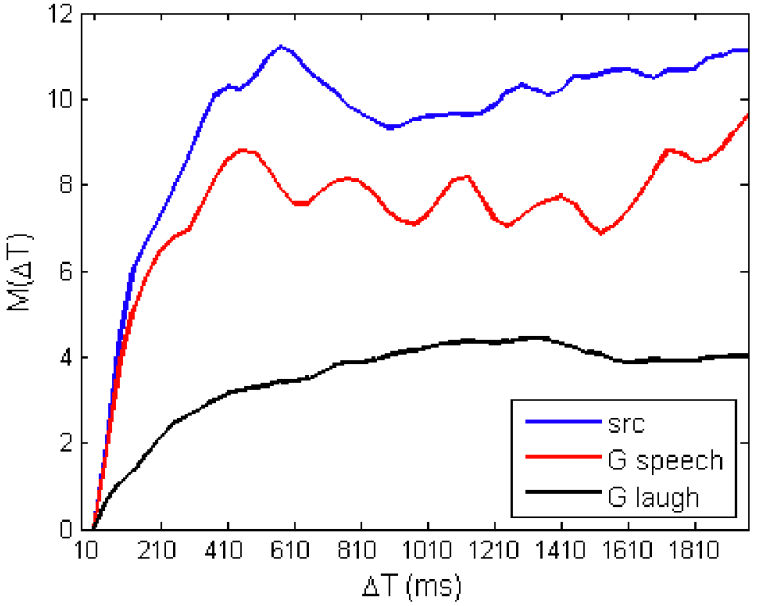
Δi – time delay
D(.) – symmetric Kl divergence



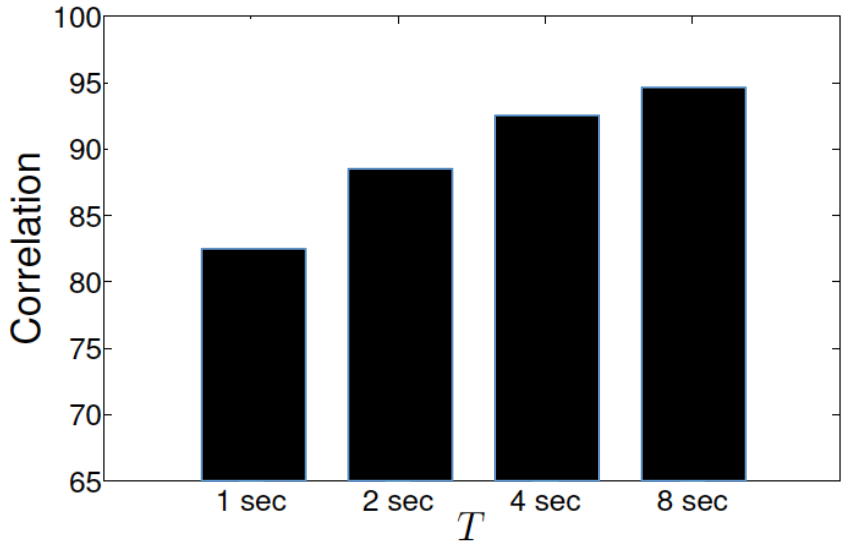
derived on all training data



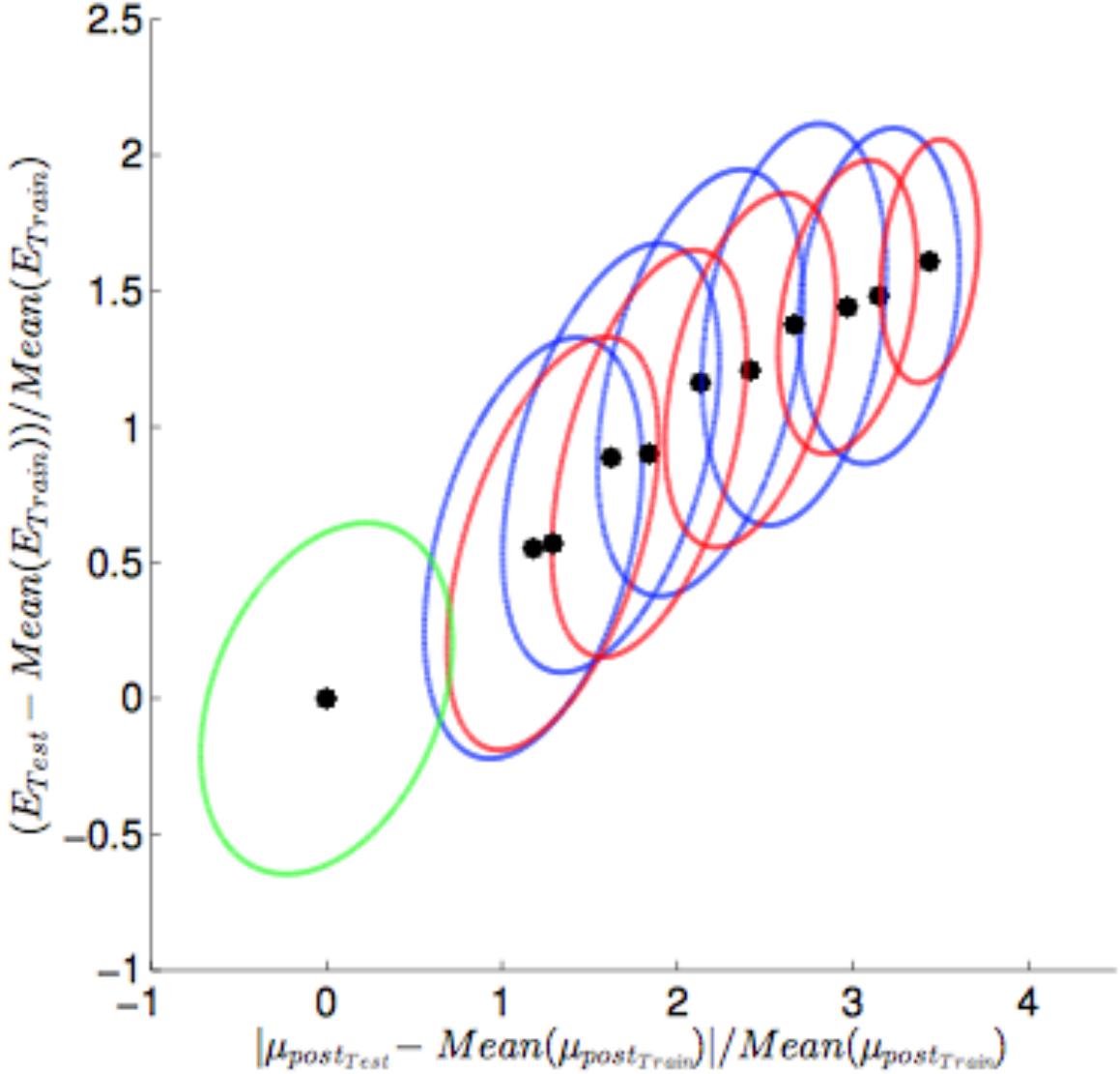
derived on one sentence

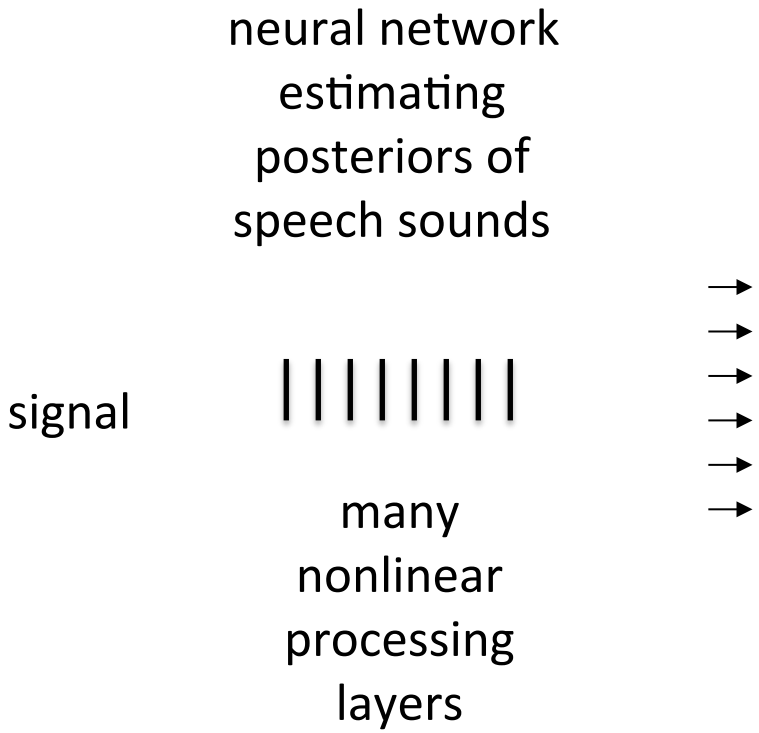


length of interval for the estimation



correlations between predictions of errors and true errors



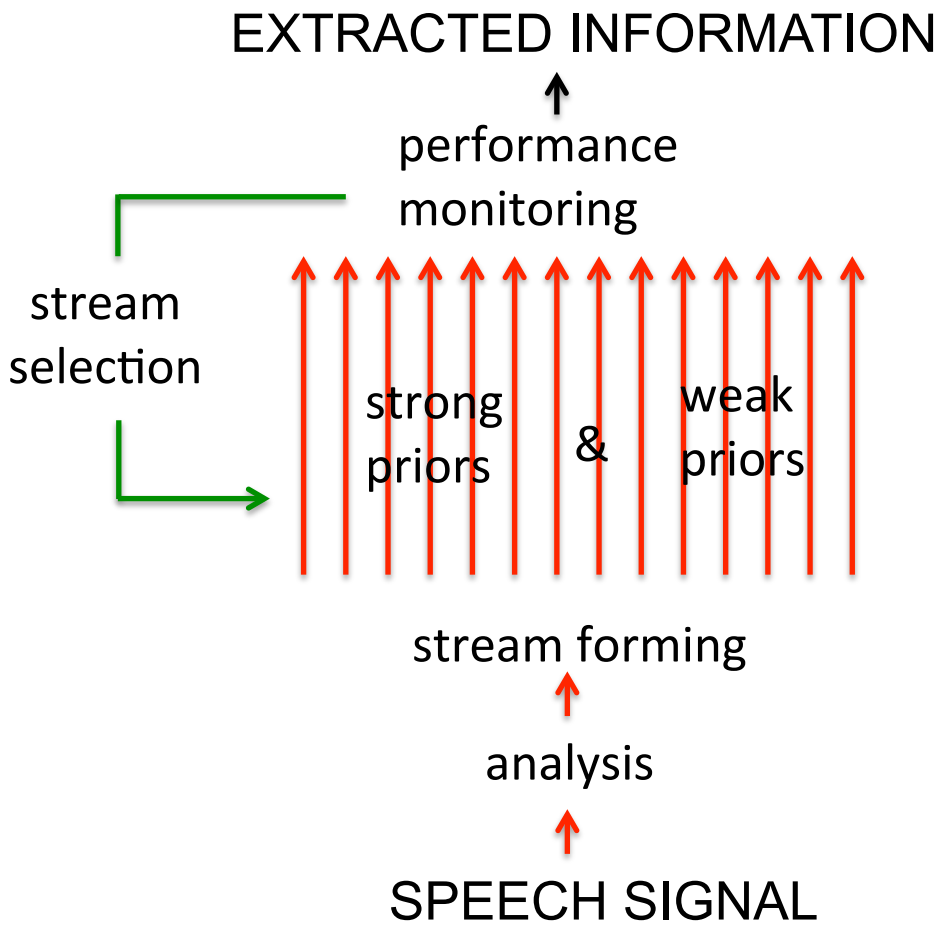


estimate reliability from the output of the classifier

estimate reliability by comparing of values on hidden layers of the network

- 31 processing streams trained on band-limited data
  - clean TIMIT
  - TIMIT corrupted by random level subway noise
- test results (39 phoneme posterior estimates and sentence-level phoneme accuracies) for 8 different noisy conditions for each processing stream

A Way of Dealing with Unknown Unknowns ?



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

Stream formation

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

Select only reliable streams for further processing

N parallel processing streams

$$P(error) = \prod_{i=1}^{N} P_i(error)$$

Observed in

- human recognition of frequency-limited and noisy speech sounds (Fletcher et al)
- human recognition of words in and out of context (Miller et al, Boothroyd and Nittrouer)

Requires reliable identification of correct answers in processing streams (knowing when knowing)

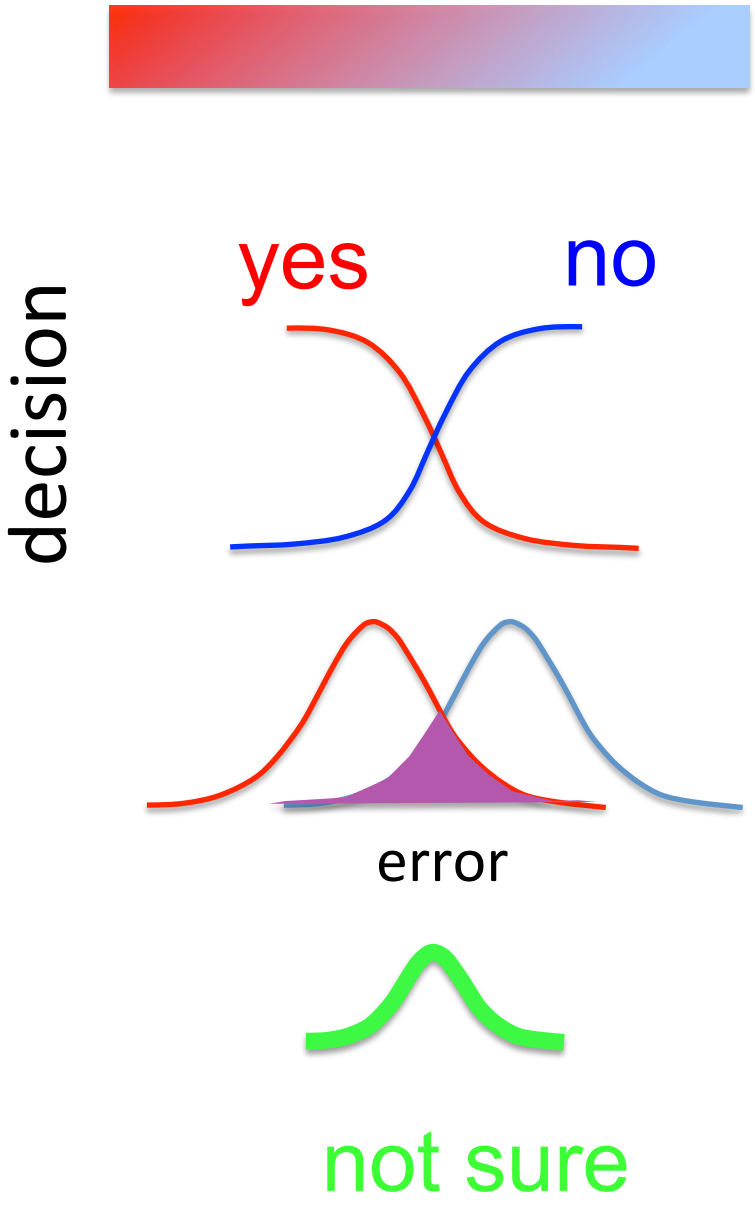
Monitoring Performance

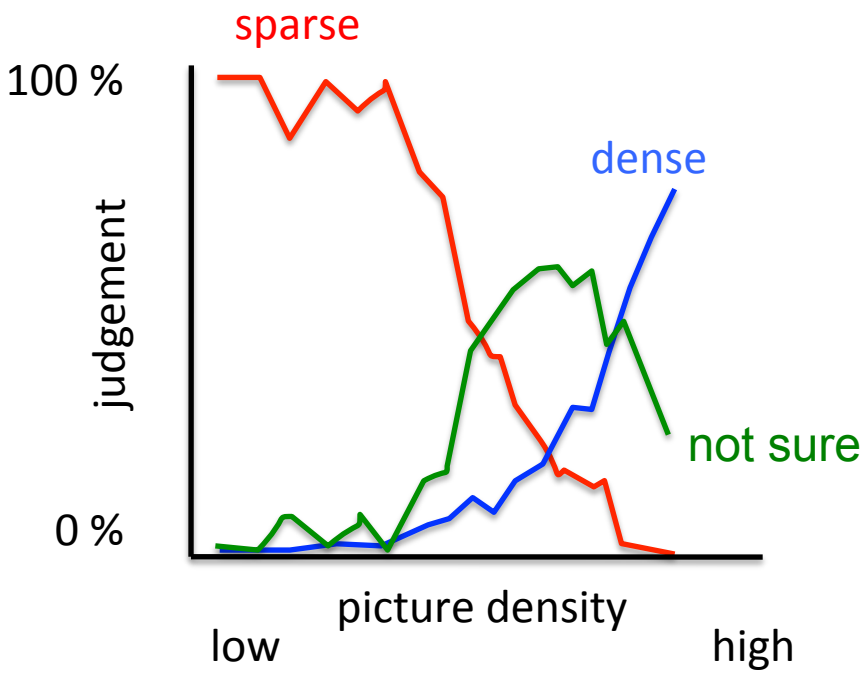
P_{1hit} P_{2hit}

$$P_{miss} = (1-P_{1hit})(1-P_{2hit})$$

observer - false positives and negatives are possible

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



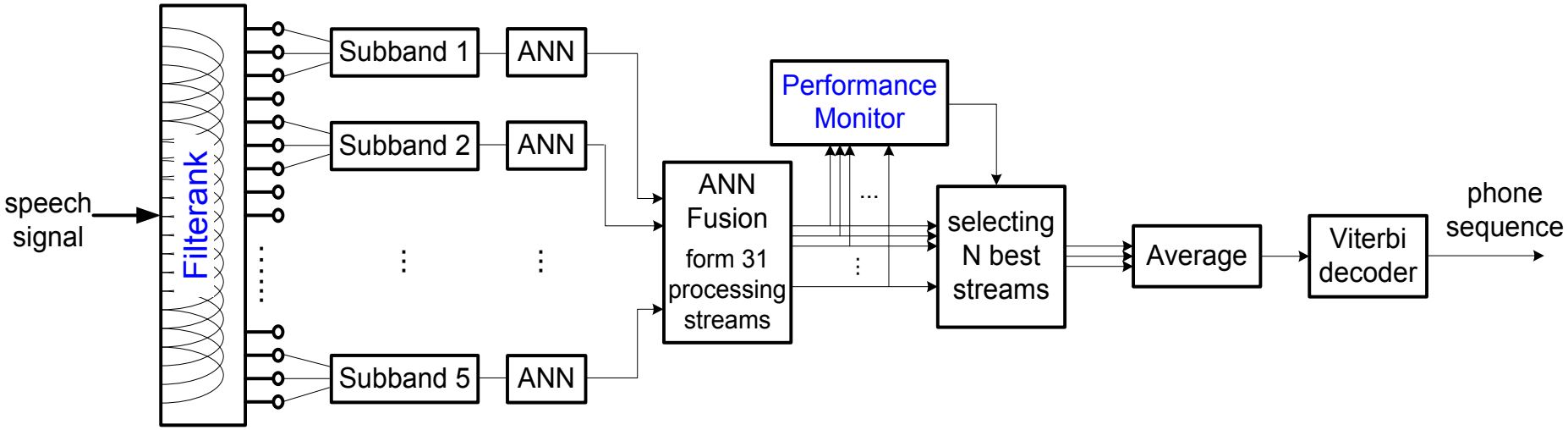


human judgment of density of pattern in a picture
(adopted from Smith et al 2003)

similar data available for monkeys, dolphins, rats,…

Multi-stream speech recognition

Variani, Li and Hermansky 2013



Phoneme recognition error rates

environment	conventional	proposed	best by hand
clean (matched training and test)	**31 %**	**28 %**	**25 %**
TIMIT with car noise at 0 dB SNR (training on clean)	**54 %**	**38 %**	**35 %**