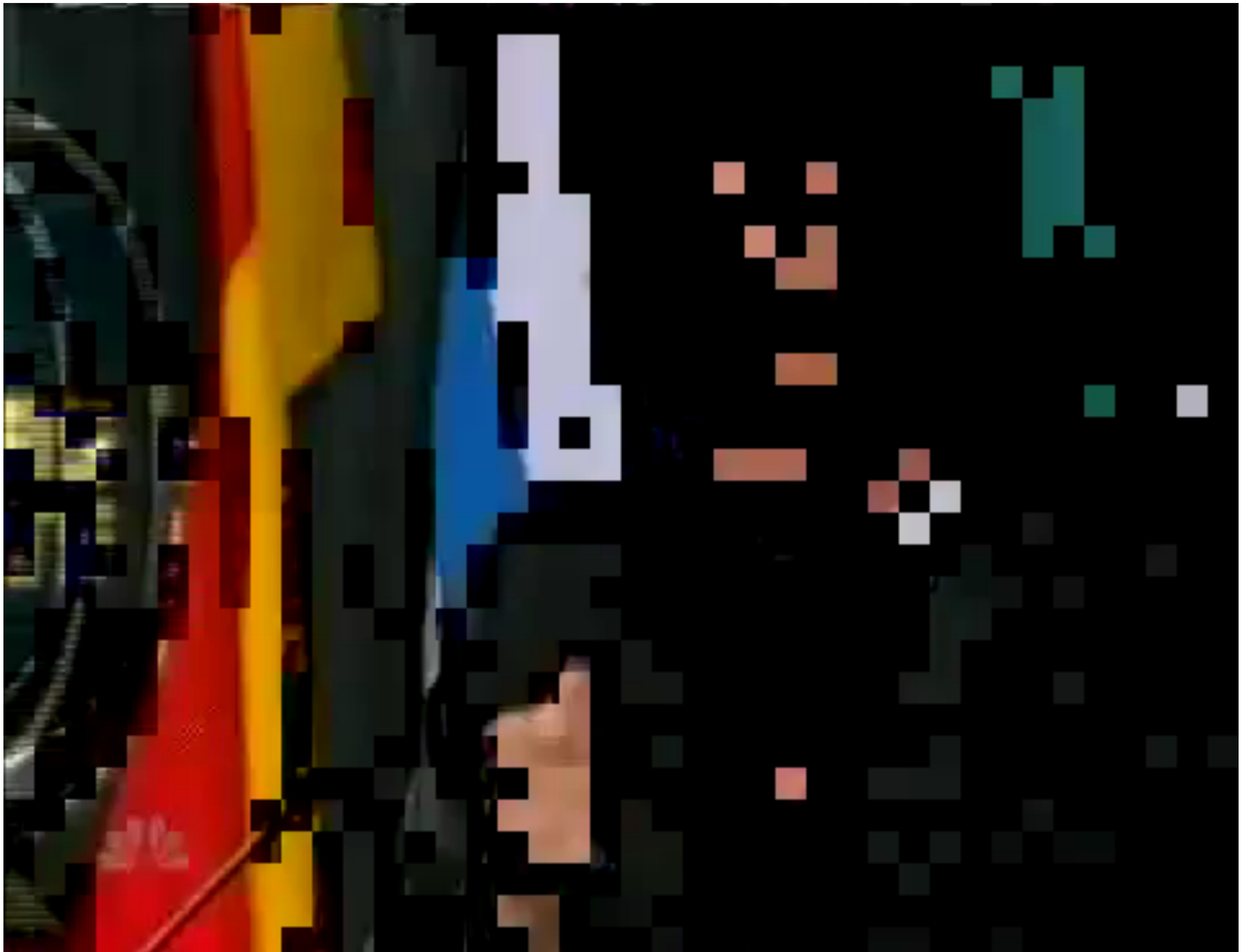
The Center For Language and Speech Processing
at the Johns Hopkins University

Dealing With Unknown Unknowns In Speech

Hynek Hermansky, Nima Mesgarani,
Samuel Thomas, Ehsan Varianni, Feipeng Li and others
The Johns Hopkins University, Baltimore MD



Former Secretary of Defense Ronald Rumsfeld

information value of surprise

$$H = -\sum_i p_i \log p_i$$

noise (unwanted information)

$$C = W \log_2 \frac{S+N}{N}$$

acoustic model (data)

language & lexicon (priors)

$$P(W \mid x) = \max_{w}\{P(x \mid W)P(W)\} / P(x)$$

x - signal $\quad W$ - model of an utterance

Works very well as long as the test data is similar to the training

Problems with unexpected data

- words not in the lexicon (OOVs)
- acoustic data not seen in training (noise)

Unknown unknown

unexpected data → system (prior experience) → unpredictable response

outlier – a data item that does not fit the rest of the **data**
unexpected – a data item that was not seen by the **system**

- How the unseen data affect the system

Noise

~~White noise, car noise, babble noise, factory noise, destroyer noise, machine-gun noise,… ?~~

- Unpredictable and previously unseen distortions of a signal
- Ultimate destroyer of an information (Shannon)

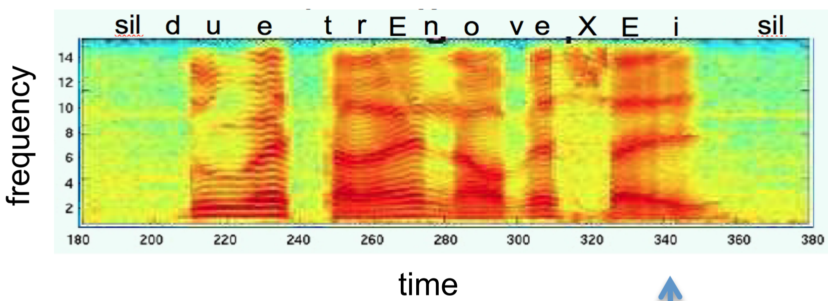
$$C = W \log_2 \frac{S+N}{N}$$

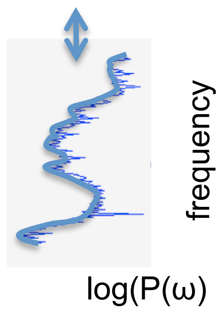
Shannon 1949

The best way to combat noise is through redundancy.

In low SNRs it may be better to ignore parts of the spectrum where noise dominates

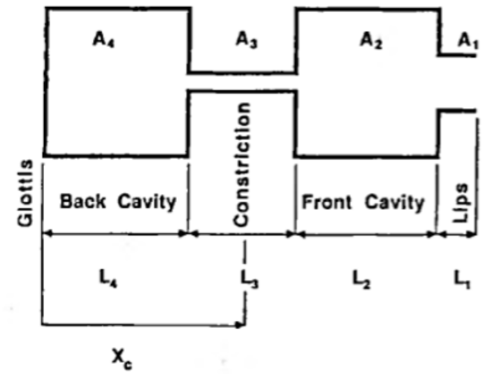
$\boldsymbol{x}$ – typically based on short-term spectrum

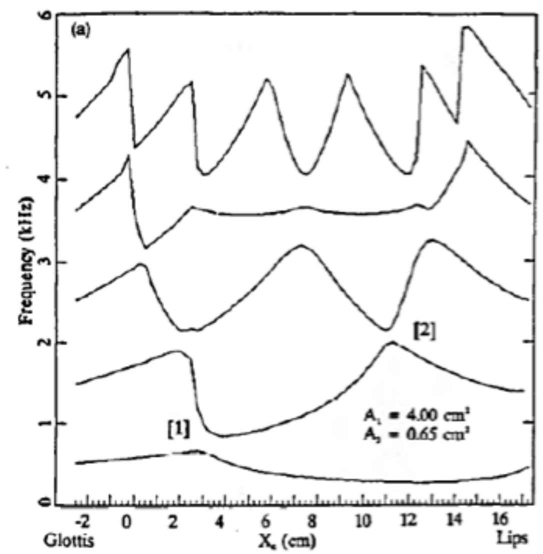




- break the spectrum into parts ?
- figure out how to de-emphasize unreliable elements ?

The best way to combat noise is through redundancy.

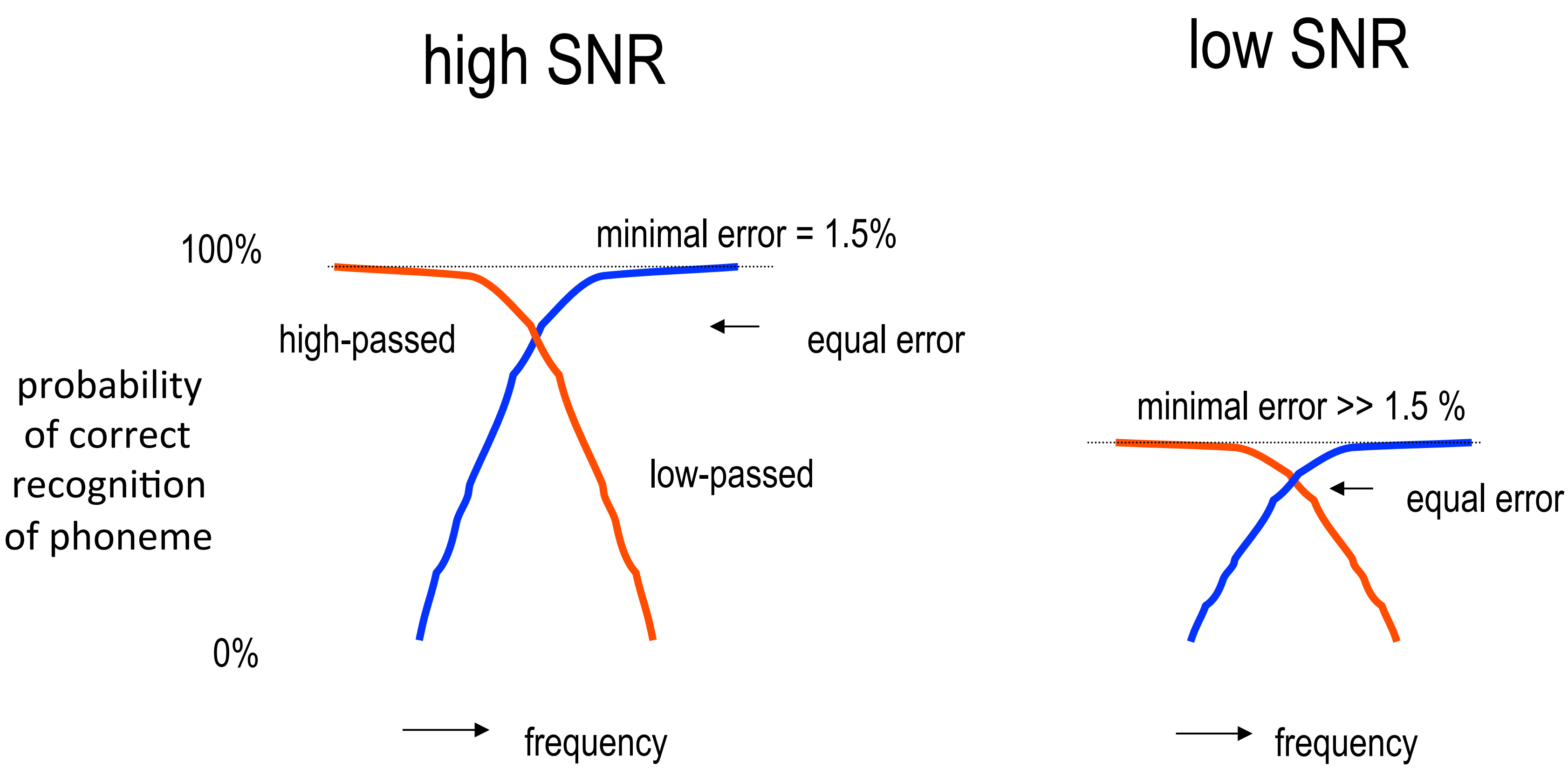




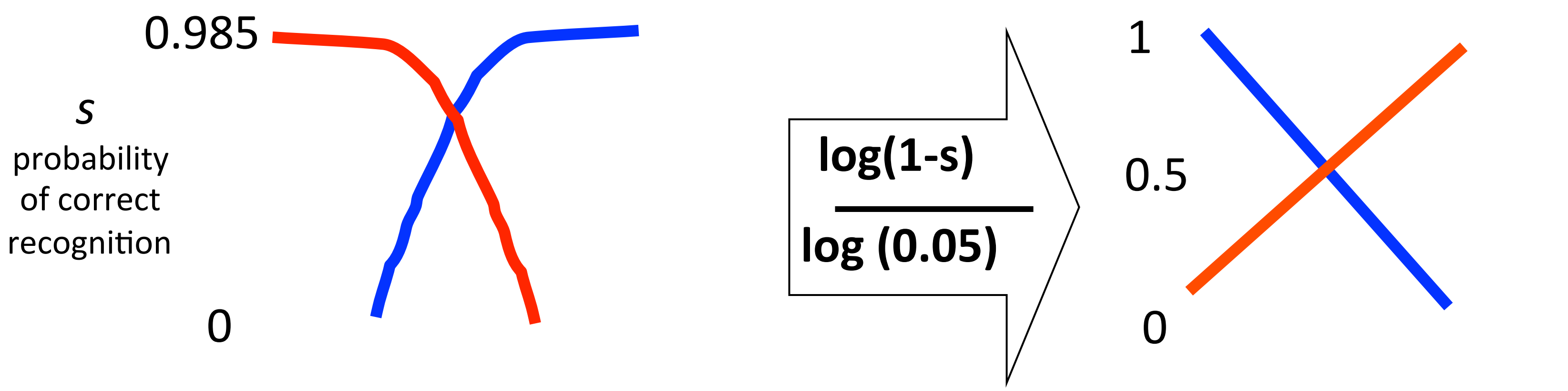
Change in shape of the vocal tract affects all frequencies of the spectrum.

Fletcher and colleagues (1920-1950)

- nonsense CV,VC, and CVC in carrier sentences, well-trained listeners
- low-pass and high-pass filtering
- varying SNR



Make the equal error at 0.5



transformation

$$A(s) = \frac{\log_{10}(1-s)}{\log_{10}(1-s_{\max})}$$

Since (1-s) = p(error), the logarithms of probabilities of errors are additive, i.e.

$$p(error) = p(error_{highband})p(error_{lowband})$$

makes the contributions from high and low band additive for all conditions

True for up to 20 bands

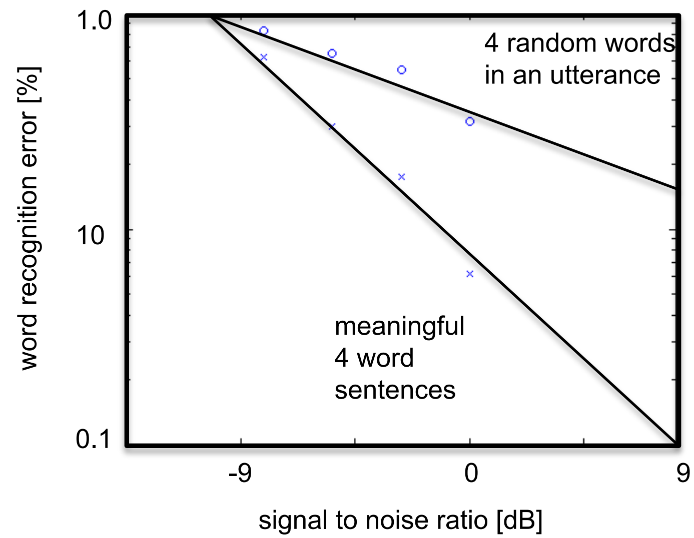
$$p(\varepsilon) = \prod_i p(\varepsilon_i)$$

How do Human Listeners Recognize Words in Context?

J.B. Allen: *Articulation and Intelligibility*, (2005)

…the context is qualitatively equivalent to adding statistically independent channels of sensory data to those already available from the speech units themselves.

(Boothroyd and Nittrouer 1988)



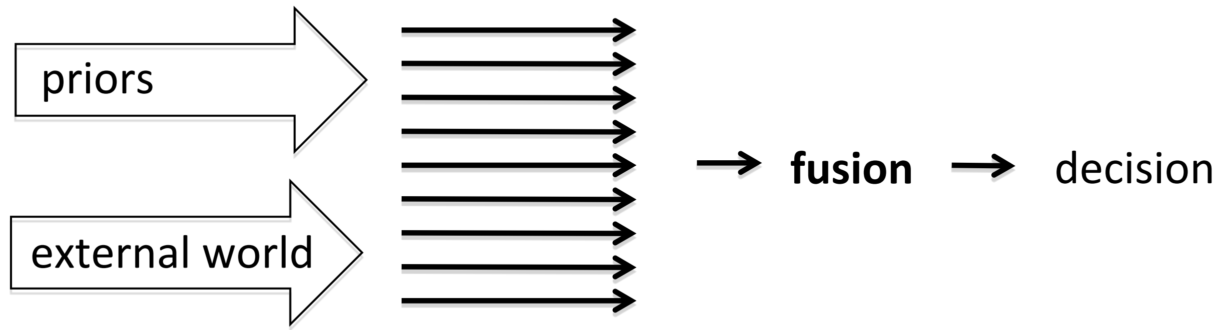
$p(\text{error}_{\text{context}}) = p(\text{error}_{\text{no context}})^k$

$k > 1$ ($k \cong 2.7$)

$p(\text{error}_{\text{context}}) = p(\text{error}_{\text{no context}})\ p(\text{error}_{\text{context channel}})^{(k-1)}$

Final error is dominated by the error in the more efficient channel

Multistream Information Processing



different projections of the signal

unexpected input corrupts only some streams

fusion

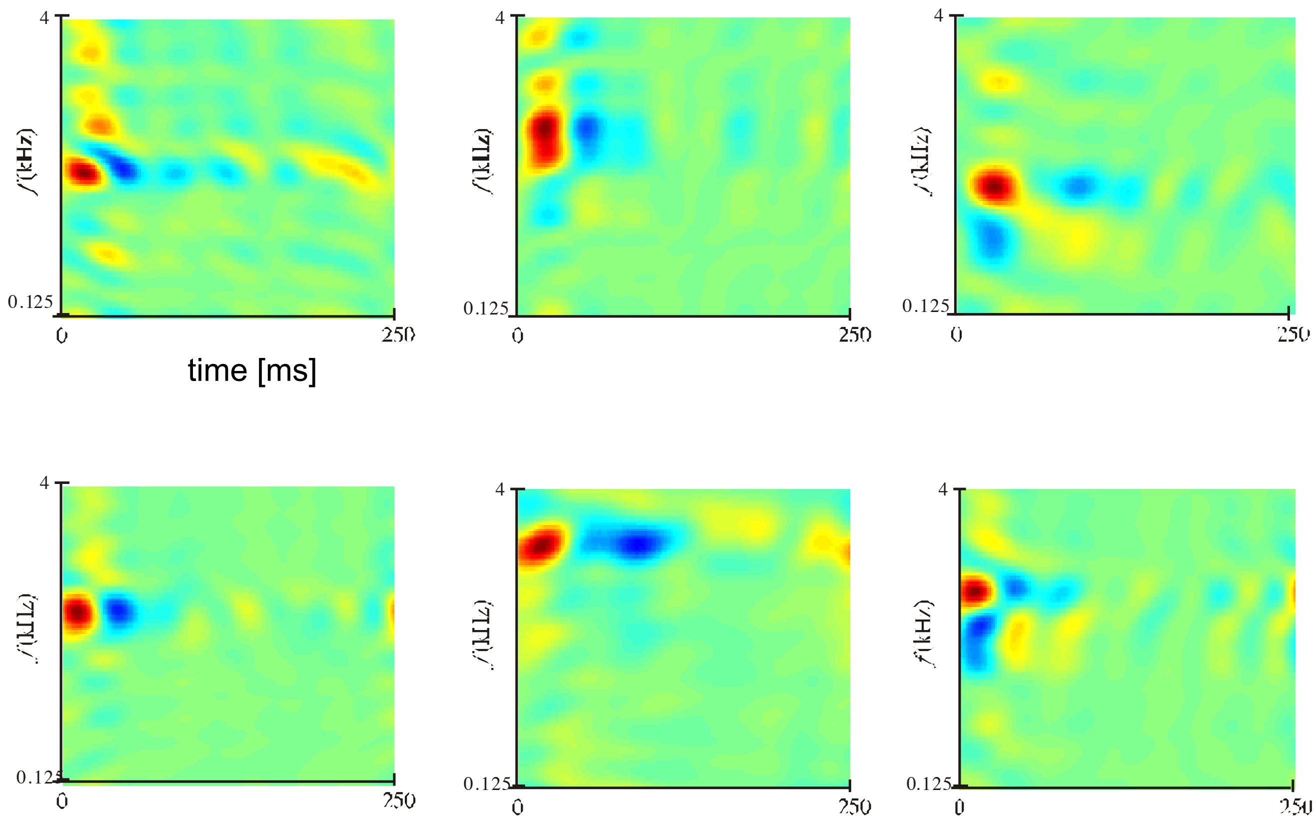
compare

is the signal corrupted (unexpected data) ?

combine

alleviate corrupted streams (product of error probabilities) stream formation in auditory perception ?

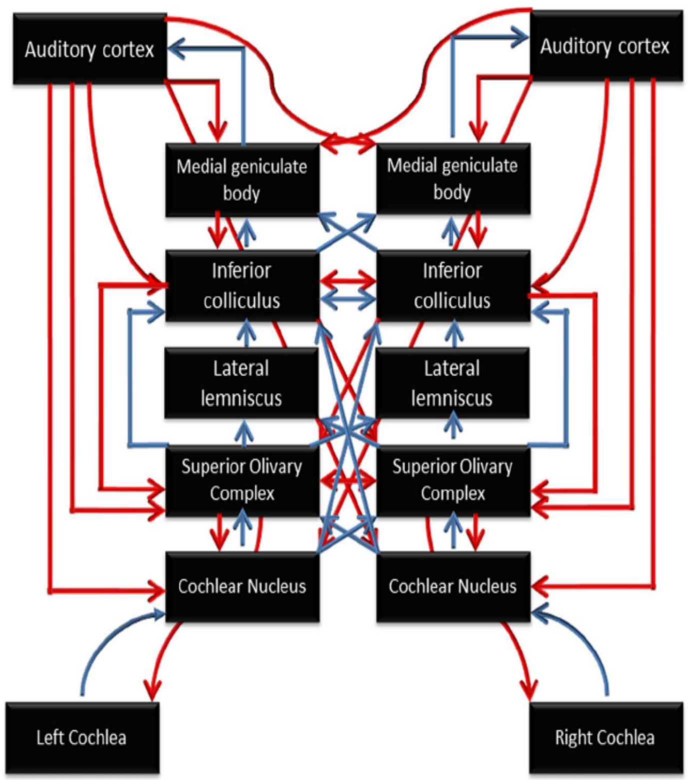
Examples of Different STRF Shapes

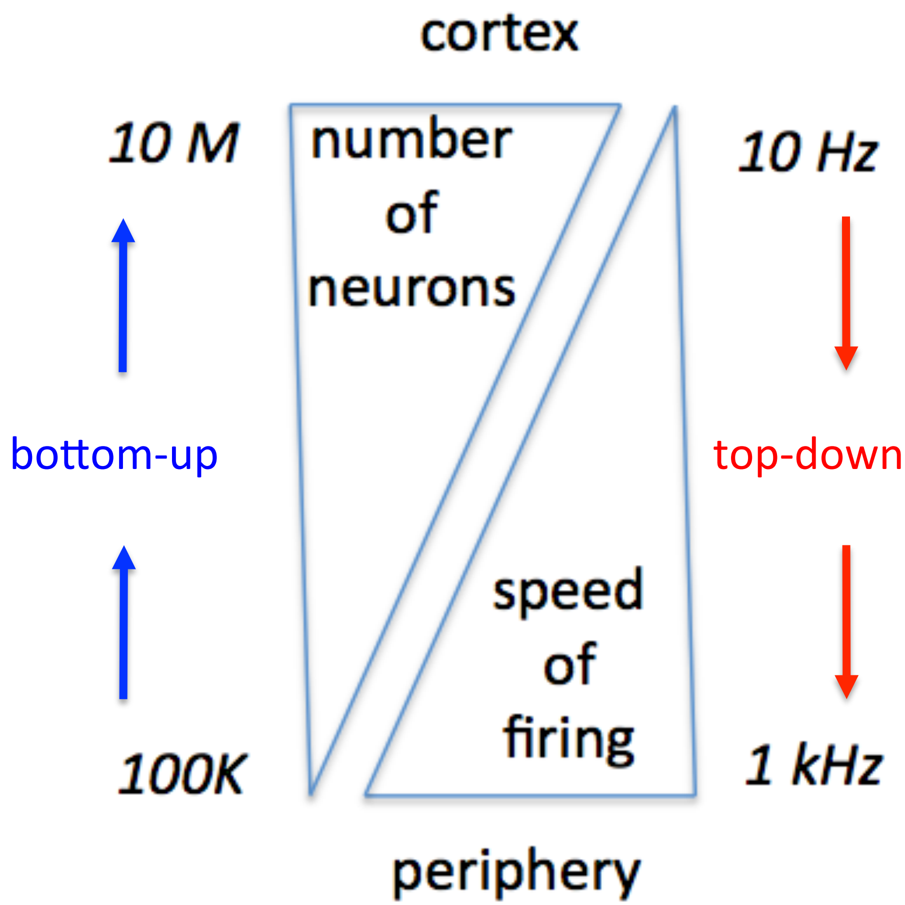


from S. Shamma's lab, U. of Maryland

Typically frequency localized and **quite long (250 ms?)**

Architecture of human auditory perception

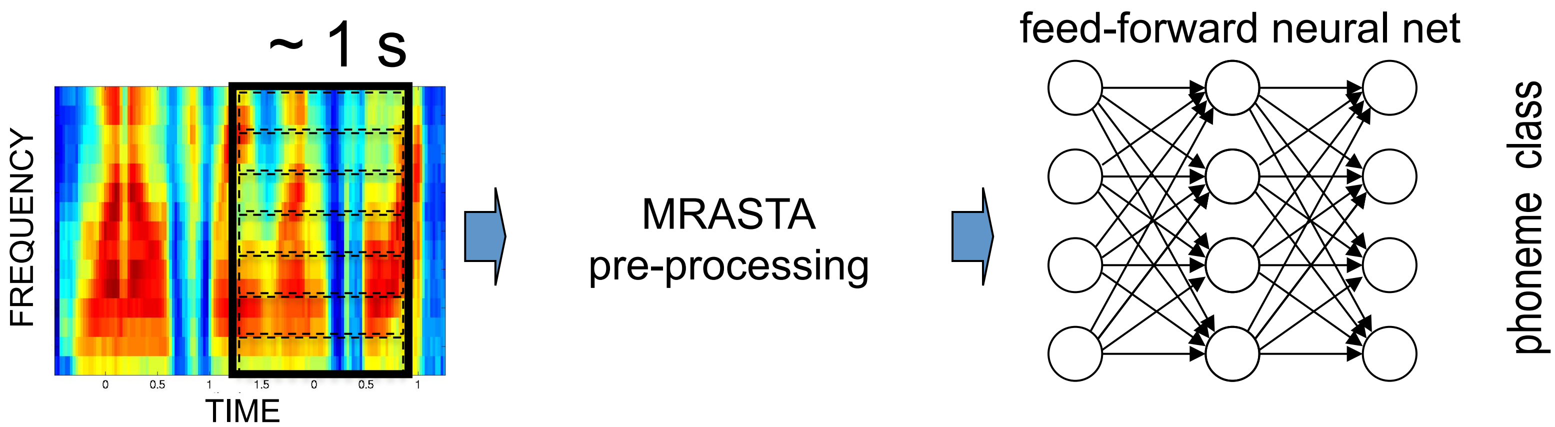




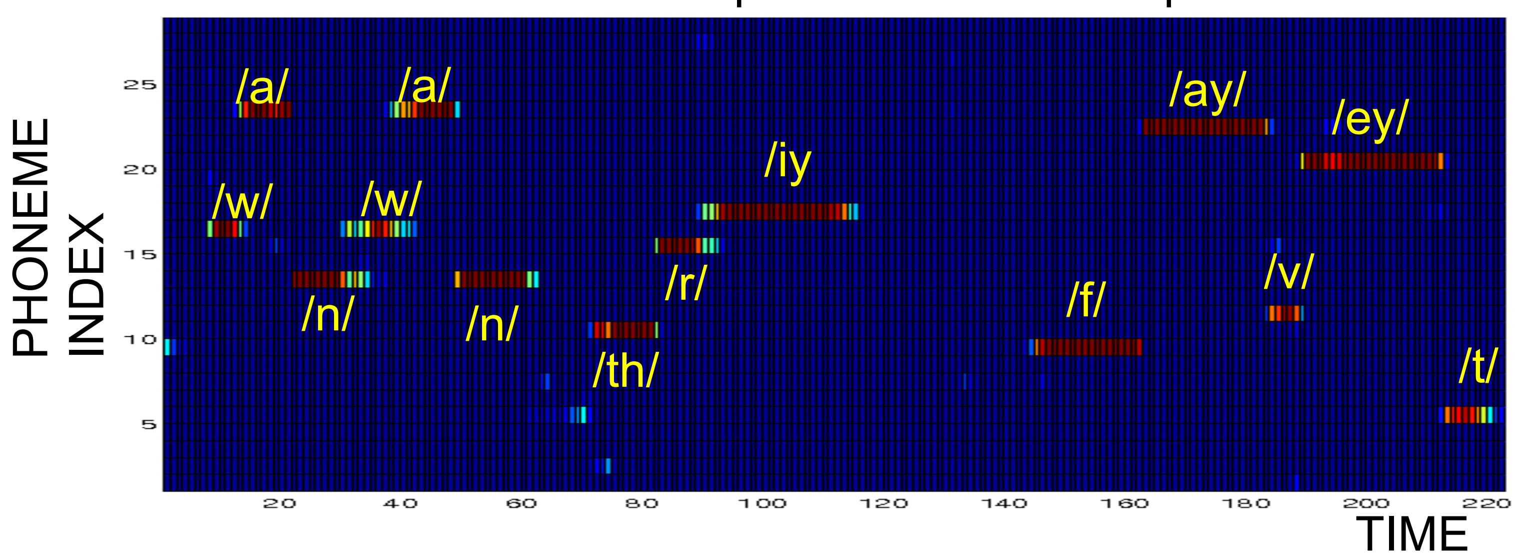
Engineering

Multi-stream recognition of phonemes

Bottom-up Estimates of Posterior Probabilities of Phonemes



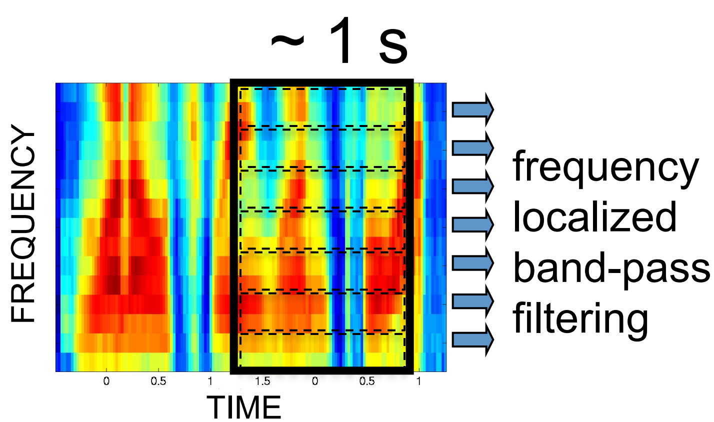
POSTERIOGRAM – a sequence of vectors of posteriors

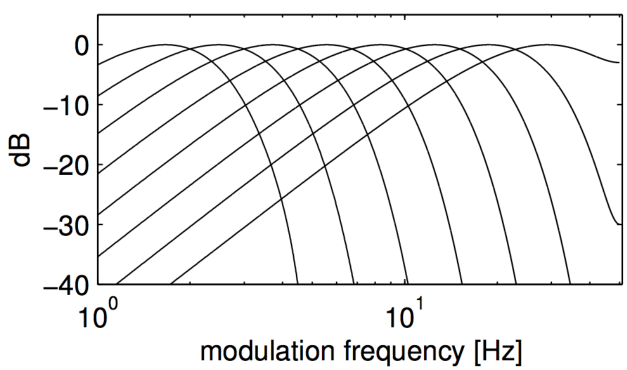


Multi-resolution frequency-localized filtering

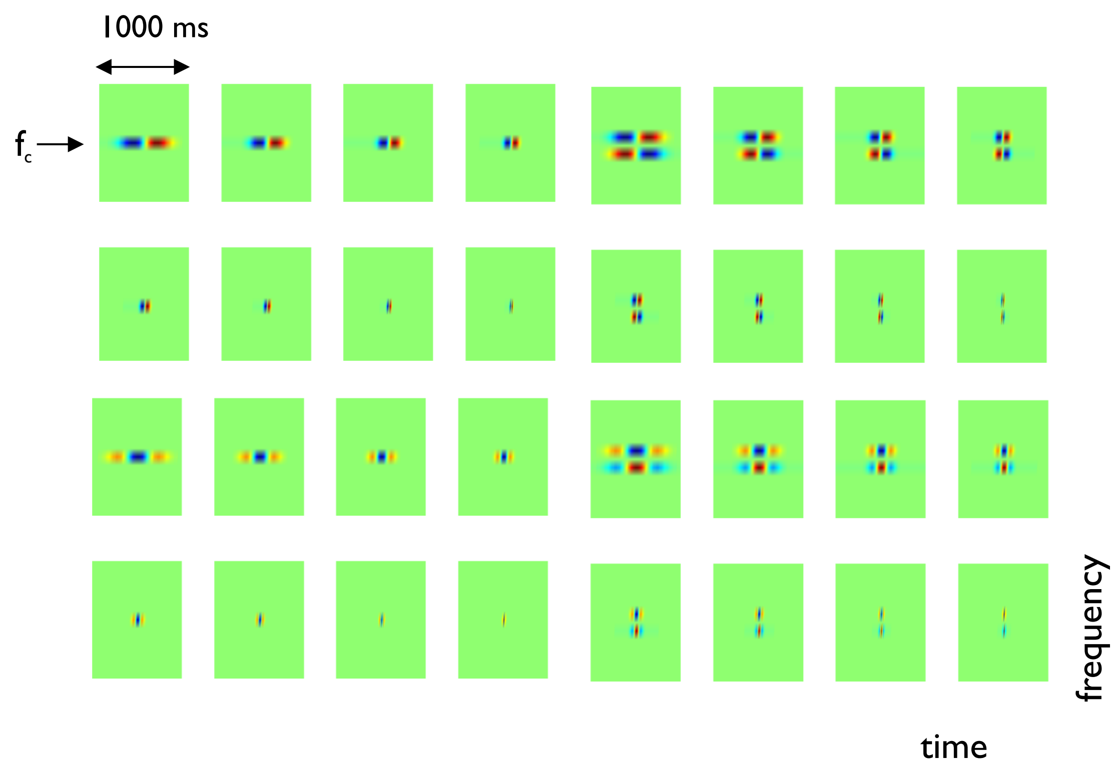
MRASTA

Hermansky and Fousek 2005





impulse responses of 2-D time-frequency filters at each critical band f_c



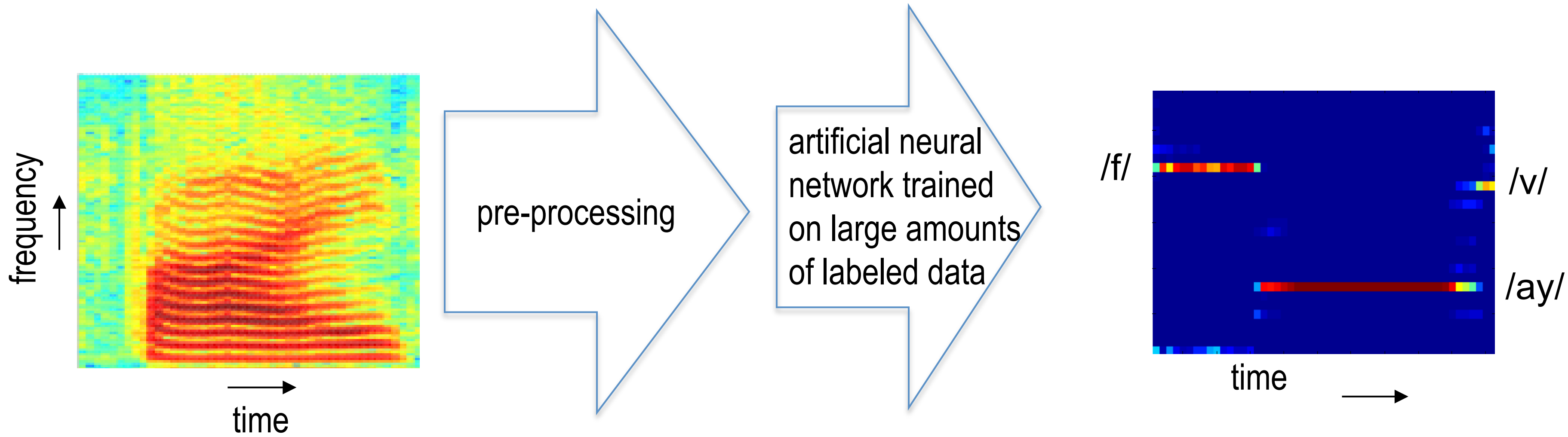
32 features from each of 14 critical bands

448 dimensional vector of features every 10 ms

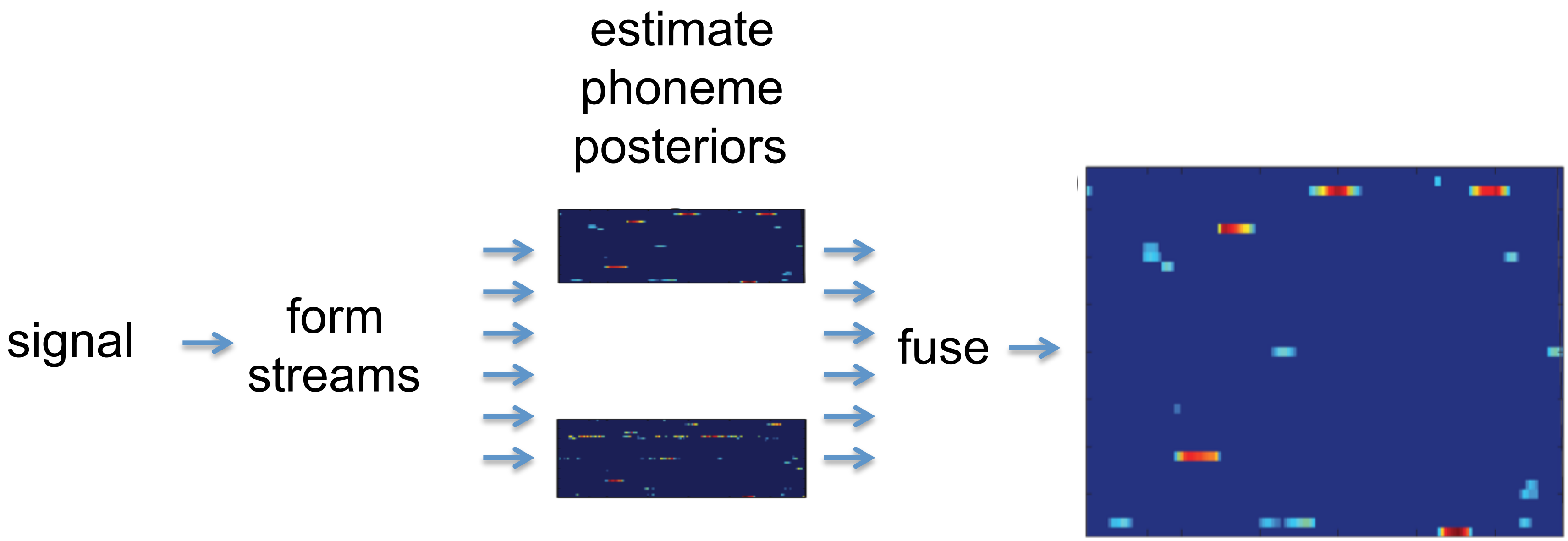
multi-resolution band-pass filtering of modulation spectrum

remove mean value of log spectral trajectoriy at each critical band

Well-trained artificial neural net



Reasonable emulation of categorical perception in ideal conditions.



How to fuse ?

How good is the result of the fusion ?
Does the result make sense ?

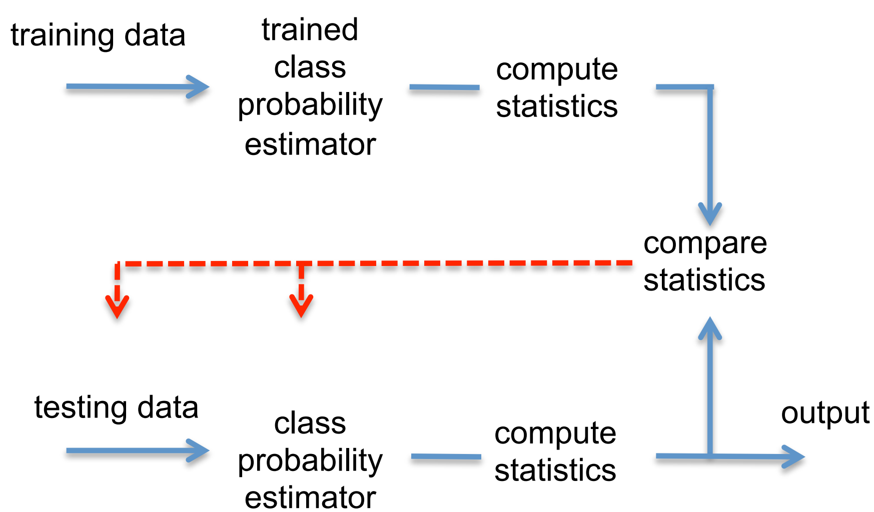
Result that makes sense

 We know what information we should get

 We know some properties of the code

Statistics of the classifier output derived on its training data and during the operation ?

Classifier with performance monitoring

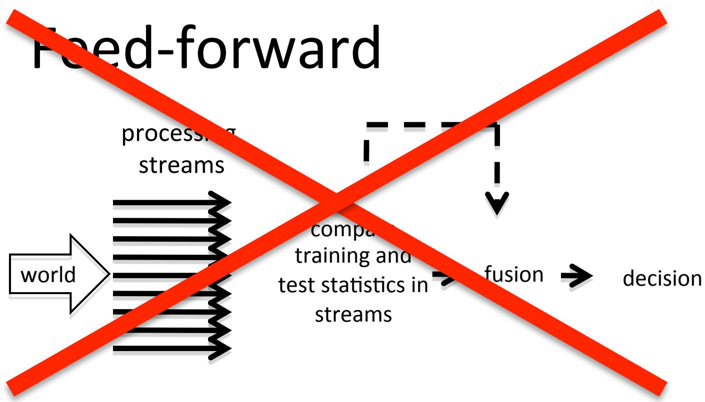


Engineering assumptions

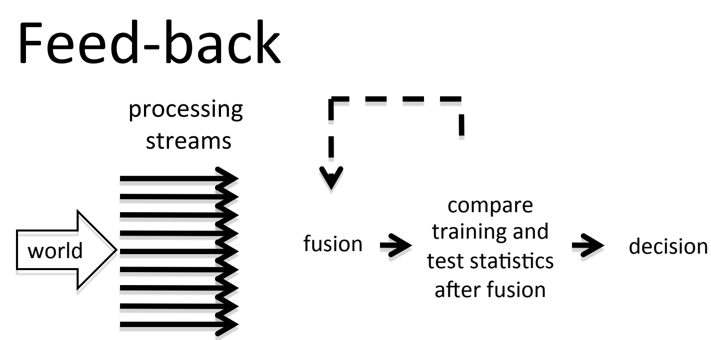
- A classifier will never work better than it does on its training data
- System performance can be summarized by statistics of the classifier **output**
- Corruptions of the data show in the statistics of the classifier output
  - **Modify the classifier (an/or data) to output training-like statistics**

Modifying multi-stream classifier

Evaluate performance of individual streams and alleviate unreliable streams



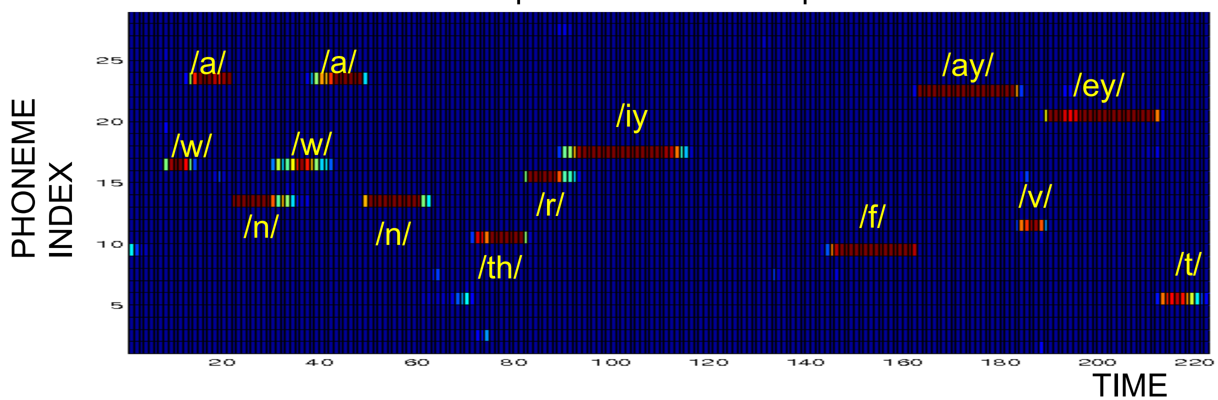
Evaluate performance of whole classifier and modify the fusion to improve the system



Statistics of classifier output: autocorrelation of posteriogram

Mesgarani et al, JASA Acoustic Letters 2011
Varianni and Hermansky, Interspeech 2012

POSTERIOGRAM – a sequence of vectors of posteriors



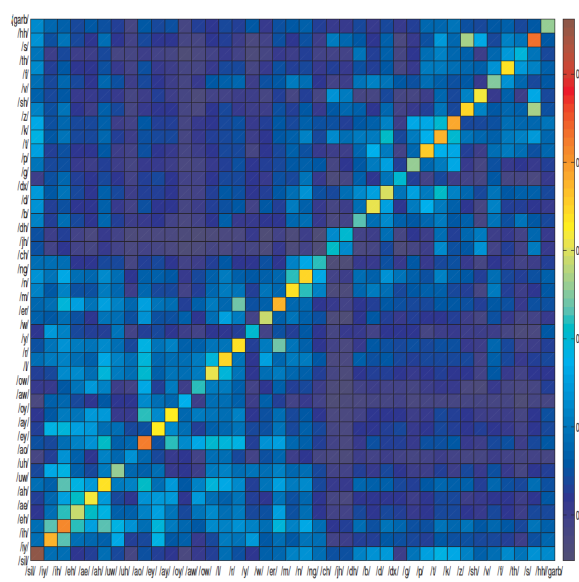
$$AC = \frac{1}{N}\sum_{i=1}^{N}\mathbf{P}(i)\mathbf{P}(i)^{T},$$

where

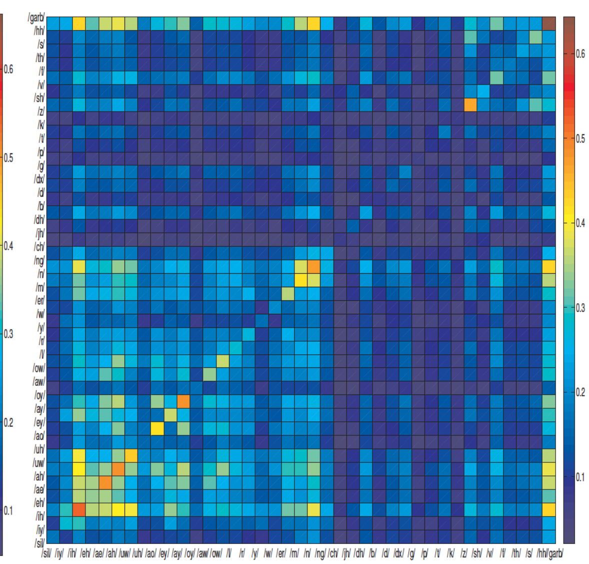
$\mathbf{P}(i)$ – posterior probability vector at time i,

N - length of the data to be described

clean data



noisy data

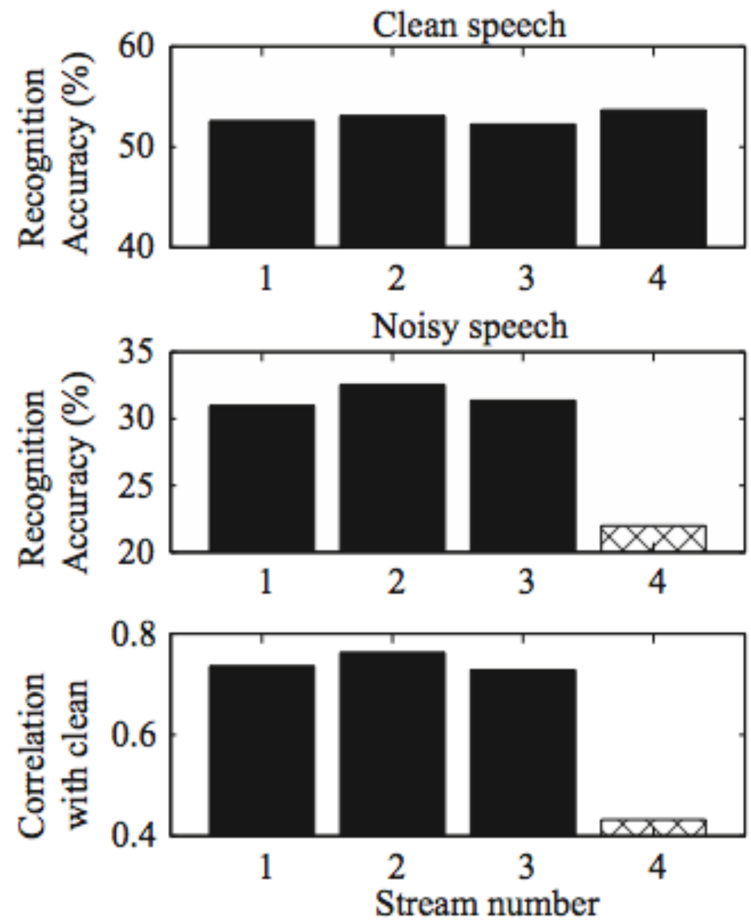


Estimate of "quality" of classification

Mesgarani et al, JASA Acoustic Letters 2011

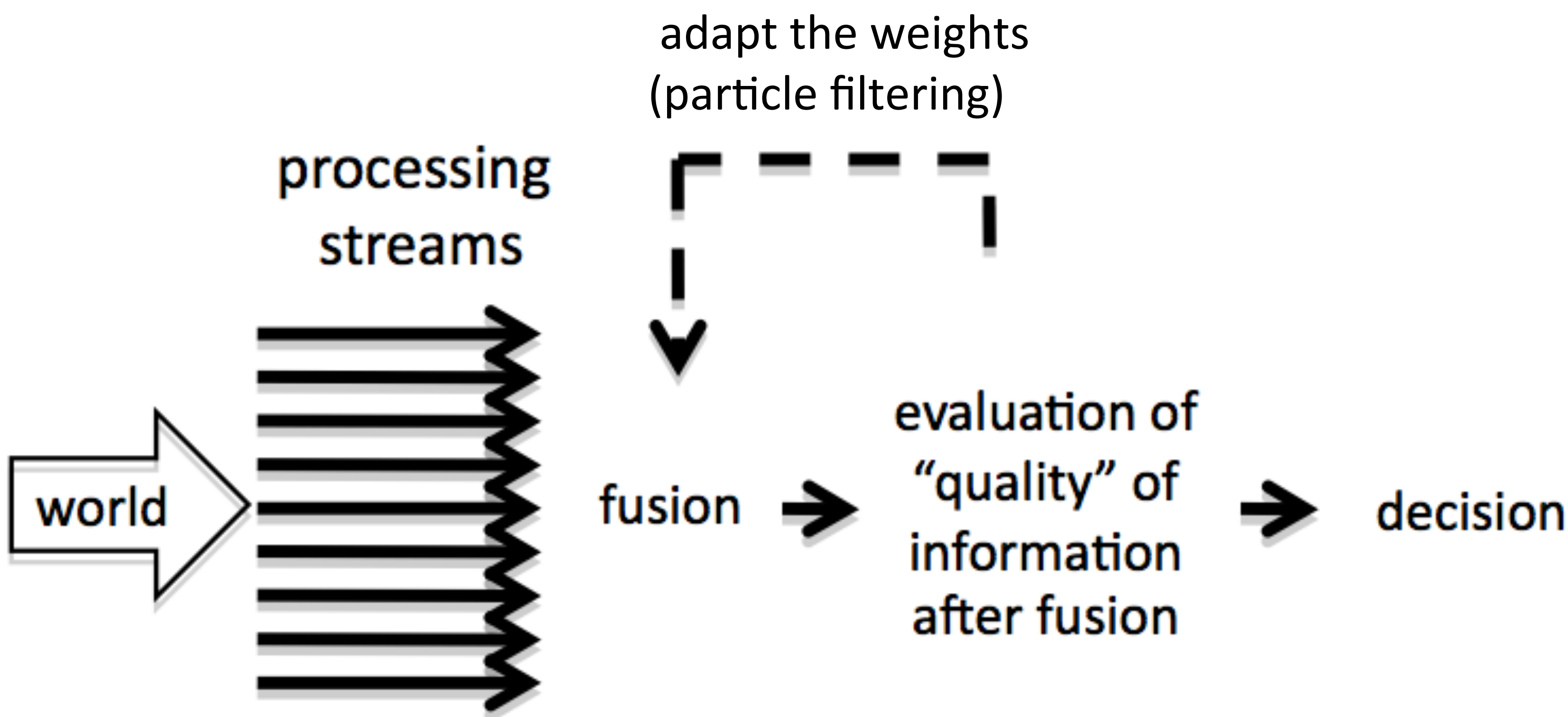
- training data autocorrelation matrix from all training data
- in the test about 4 s of data yield useful autocorrelation matrix
- matrix comparison

$$r = \frac{AC_{clean} AC_{noisy}}{\left\| AC_{clean} \right\| \left\| AC_{noisy} \right\|}$$



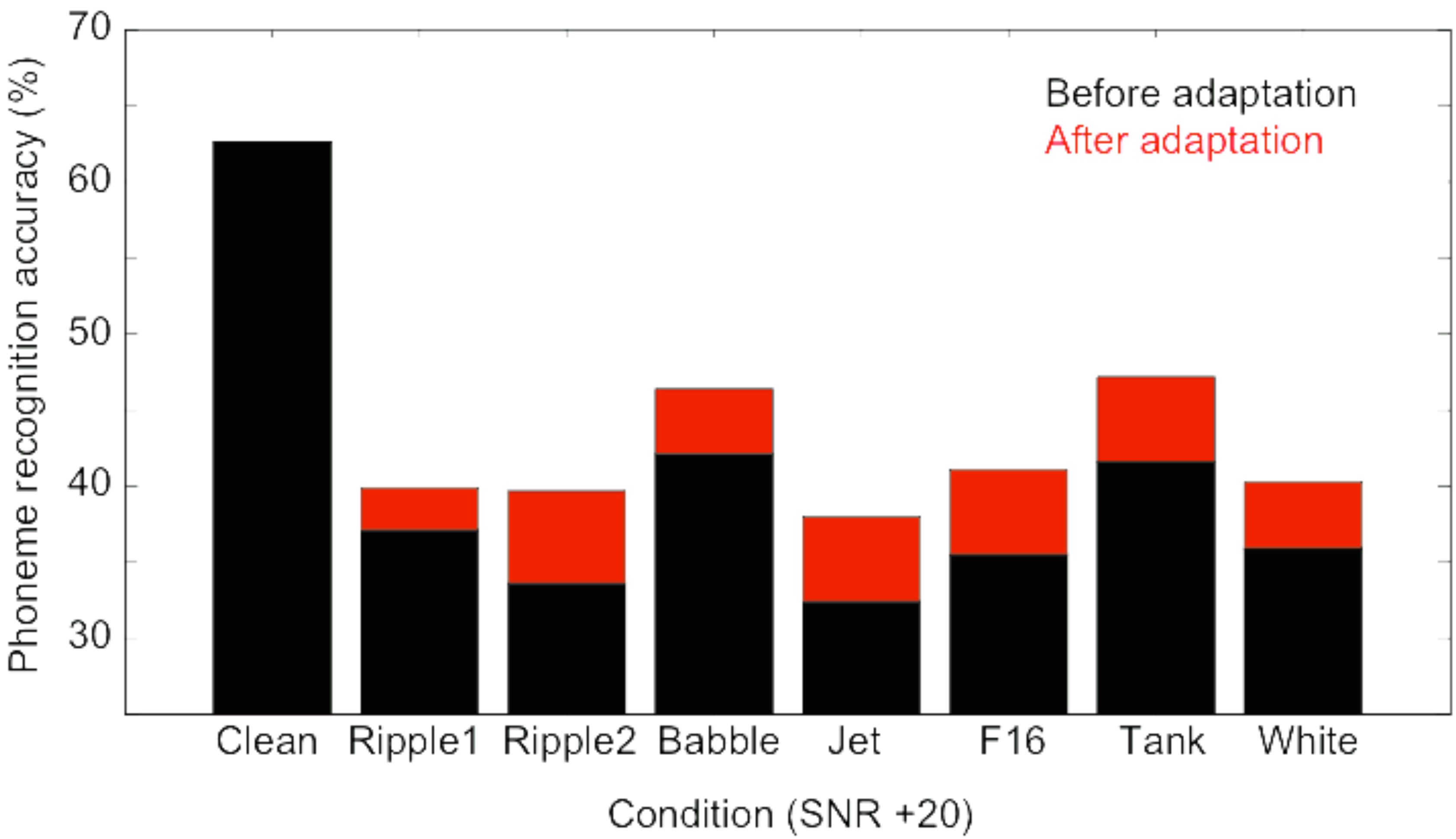
Adaptation

Mesgarani et al, INTERSPEECH 2011

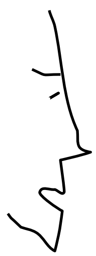


Result

Mesgarani et al, INTERSPEECH 2011



Boothroyd's model of human speech recognition



sensors and peripheral processing

prediction using past experience

top-down

bottom-up recognition

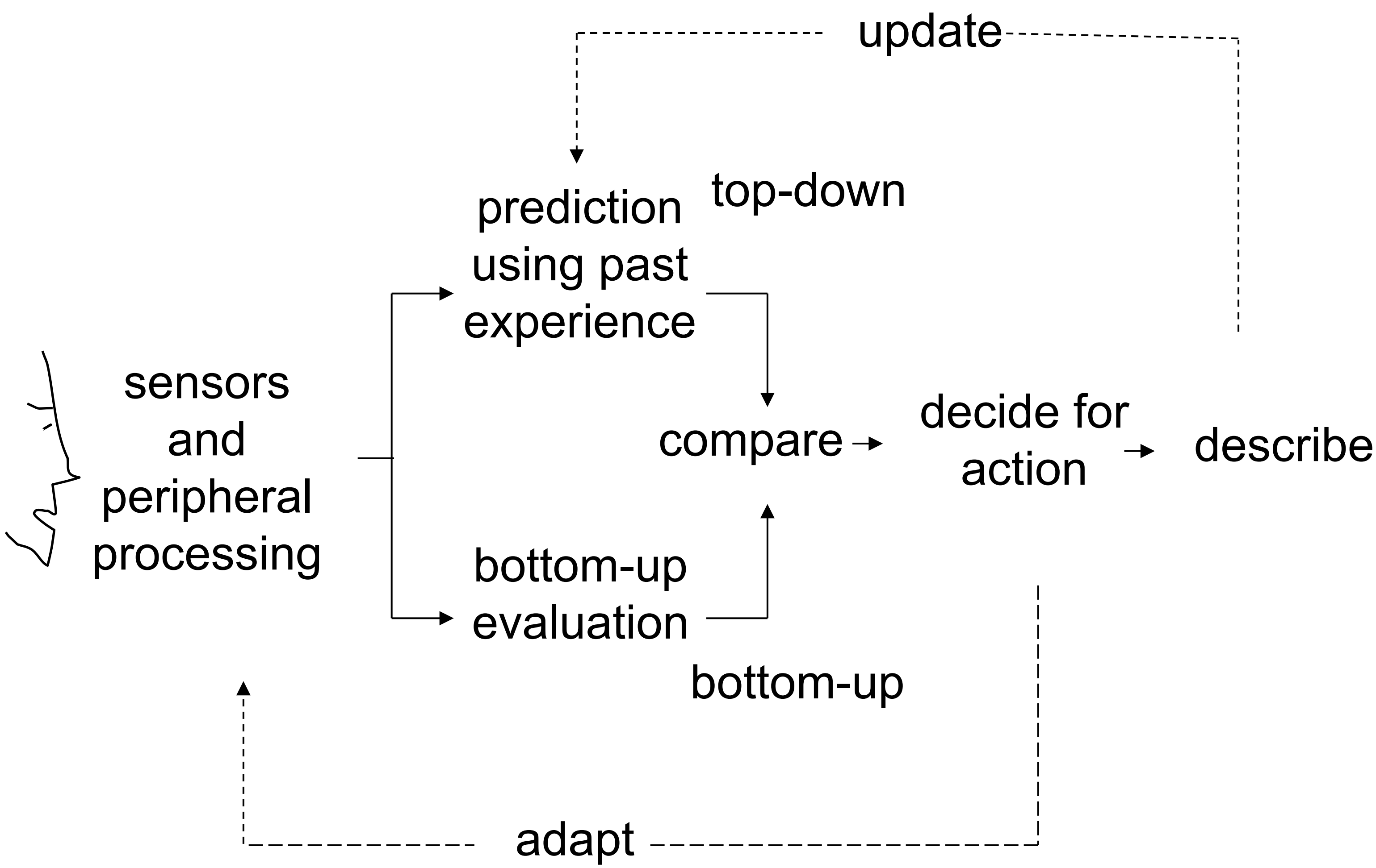
bottom-up

fuse → result

clean signal – streams with weak priors dominate
corrupted signal – streams with strong priors dominate

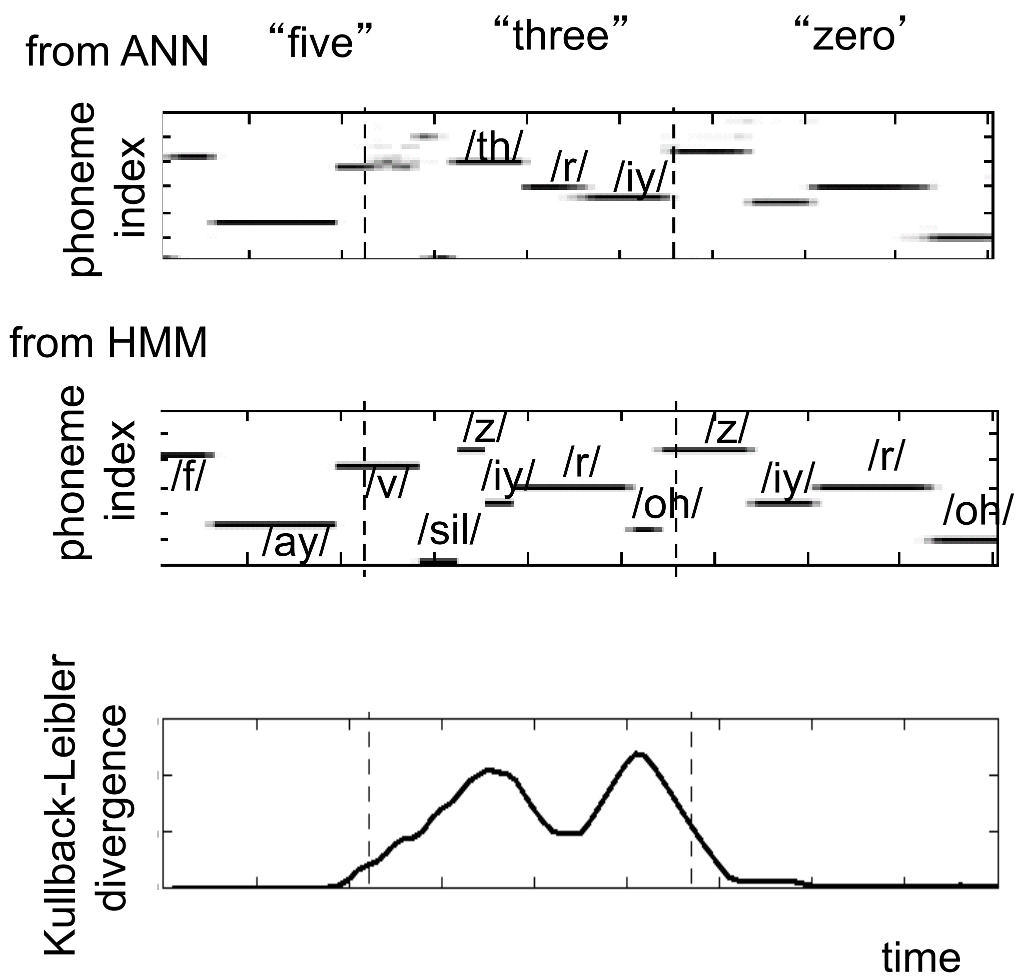
Ketabdar, PhD Thesis 1990

Dealing with unexpected words ?



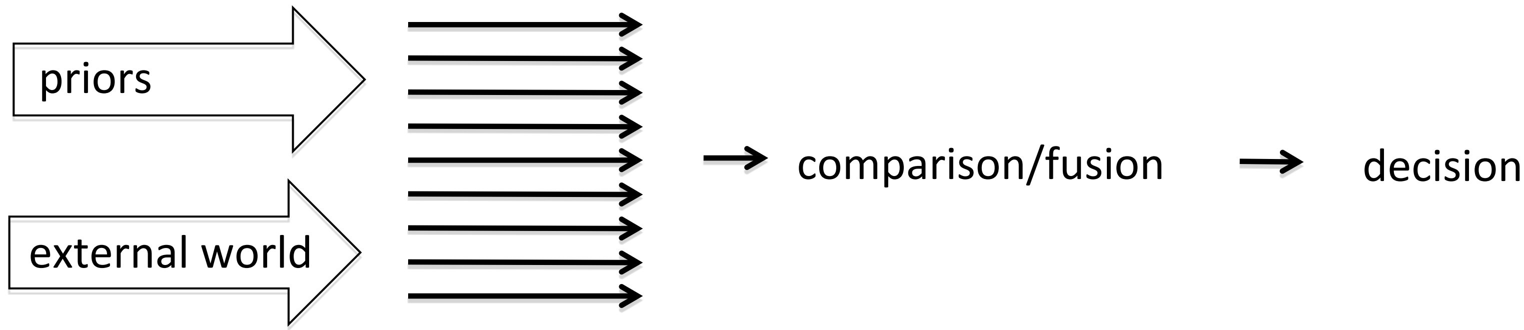
Indicate Out-of-Vocabulary (OOV) Word

- telephone quality continuous digits
- one digit (here “three”) left out from the lexicon (OOV word)



Ketabdar et al 2007

Conclusion



Multistream recognition:

a way towards human-like robustness to unexpected acoustic inputs

- unseen acoustic distortions (noises)
- unexpected words