# Dealing with unexpexted acoustic inputs in ASR

Hynek Hermansky

IDIAP Research Institute Martigny

Swiss Federal Institute of Technology in Lausanne, Switzerland

Introduction to a panel discussion at ASRU 2007, Kyoto, Japan

December 2007

# The Problem

- Acoustic inputs
  - not seen in the training
  - not expected by a prior knowledge
- Out-of-vocabulary, out-of-language, out-of-domain words, accented speech, children speech, accented speech, unexpected noises, e.t.c.

- Typically replaced by a high probability (sometimes) acoustically similar words

- Is this a inherent problem of the current stochastic approach to ASR ?

# A way of dealing with lousy acoustic modeling

$$w \propto \arg\max_{i}( p(x \mid M(w_i)P(M(w_i)^{\gamma})$$

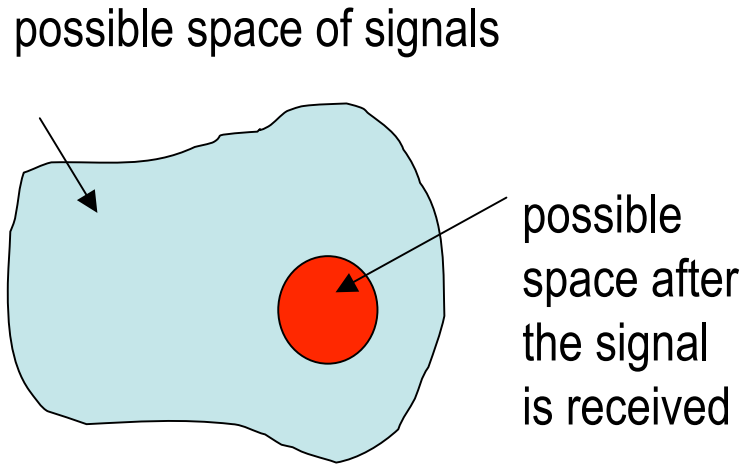
$M( w_i )$ – model of the whole utterance

**Good:** parts of the utterance can be corrupted and the utterance can still be correctly recognized

**Bad:** low prior probability items in the utterance may be substituted by wrong ones

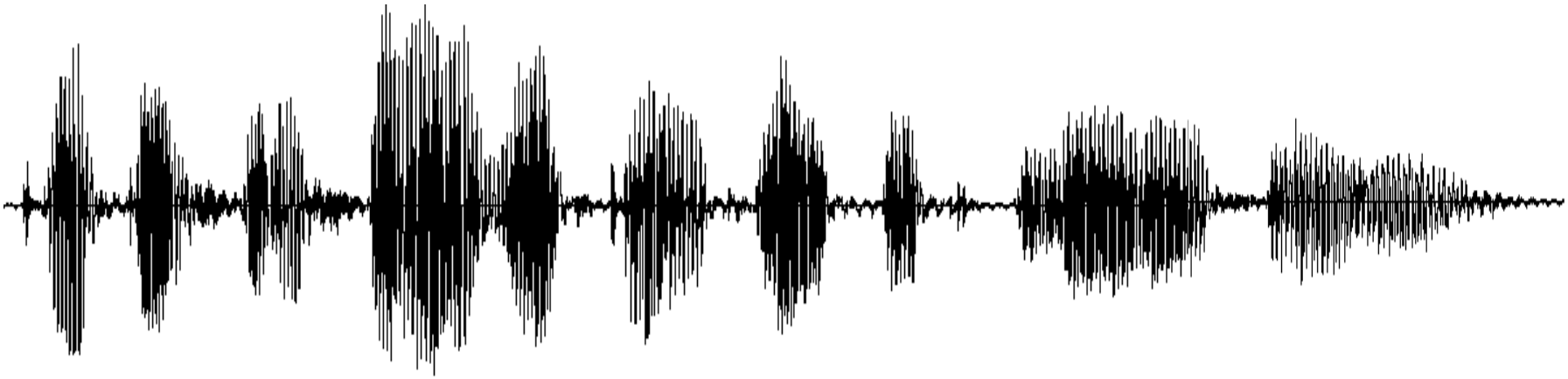
## Low prior probability words

- rare
  - lower impact of the final WER ☺
- unexpected
  - **therefore information-rich** ☹




- The amount of information gained by receiving the signal is proportional to ratio of these two areas
- The less probable the signal, the more information is gained

Czech sentence: Koupil jsem si novy **computer**, ktery nefunguje.



Recognized as:

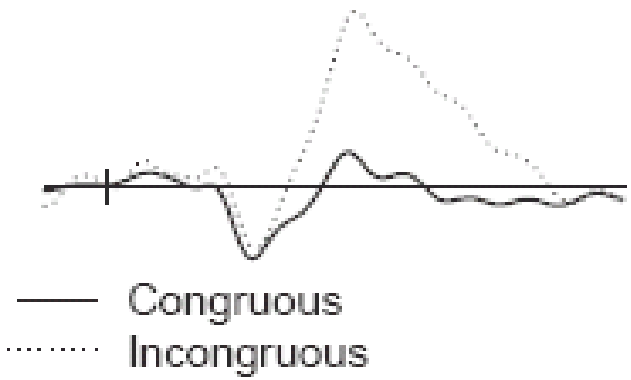
|although|some| sort |of|the| **computer** |can|either| way |

# Electrophysiology and speech comprehension

- Event-related potentials
  - brain electrical activity (neocortex ?)
  - negative potential activity (N400) indicates "difficulty" in processing of the information (Kutas et al, since 1980)

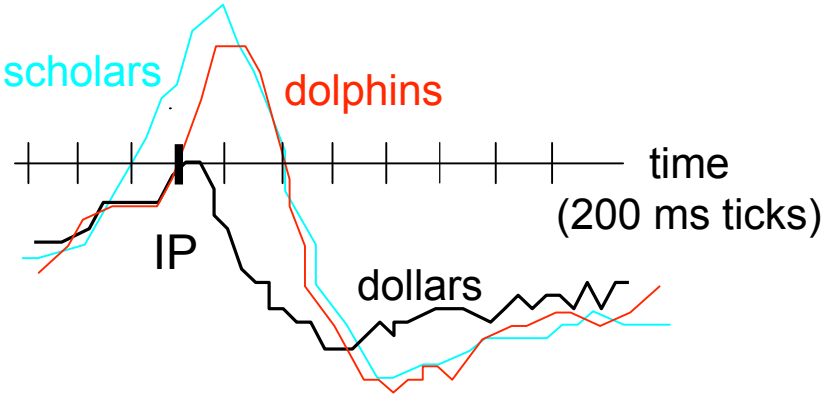
Words in sentence

-van Petten at al., credit to J.B. Allen



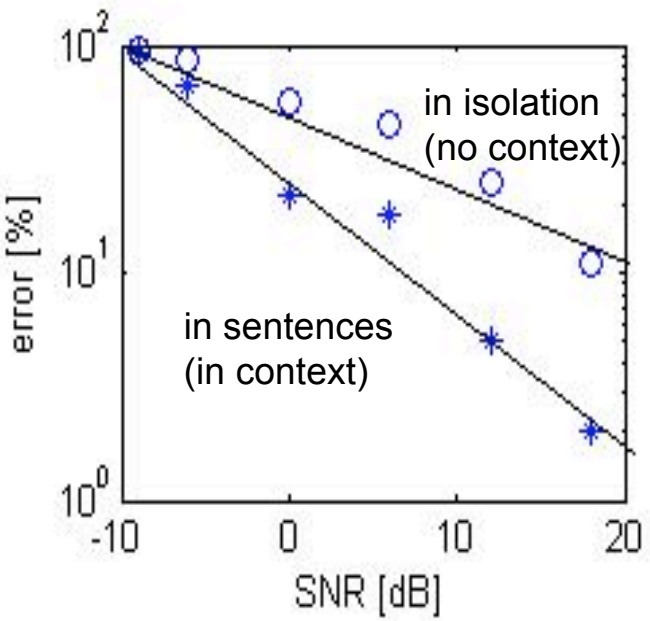

Pay with .....

negative magnitude of averaged EEG




Context of the sentence is used simultaneously (**in parallel)** with the recognition of the word

# Word errors in human recognition of speech




Miller 1962

-interpretation by Boothroyd and Nittrouer 1998

-credit to J. B. Allen

$$error_{context} = error^{k}_{no\ context}$$

$$\boldsymbol{error_{context} = error_{no\ context}\ error_{context\ channel}}$$

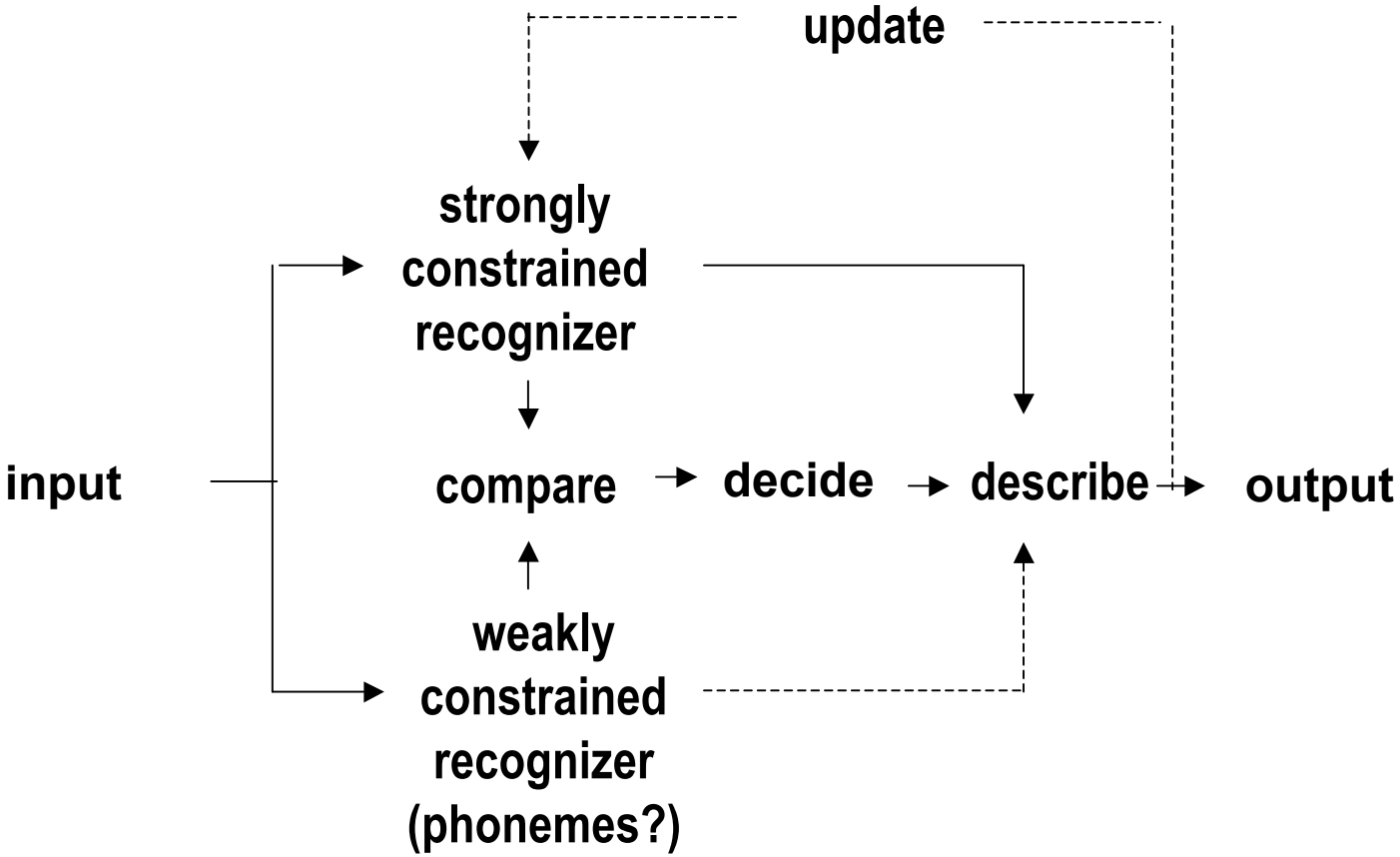
**errors multiply**

context (top-down) channel **is acting in parallel** with the acoustic (bottom-up) channel

Three ways of getting the word right

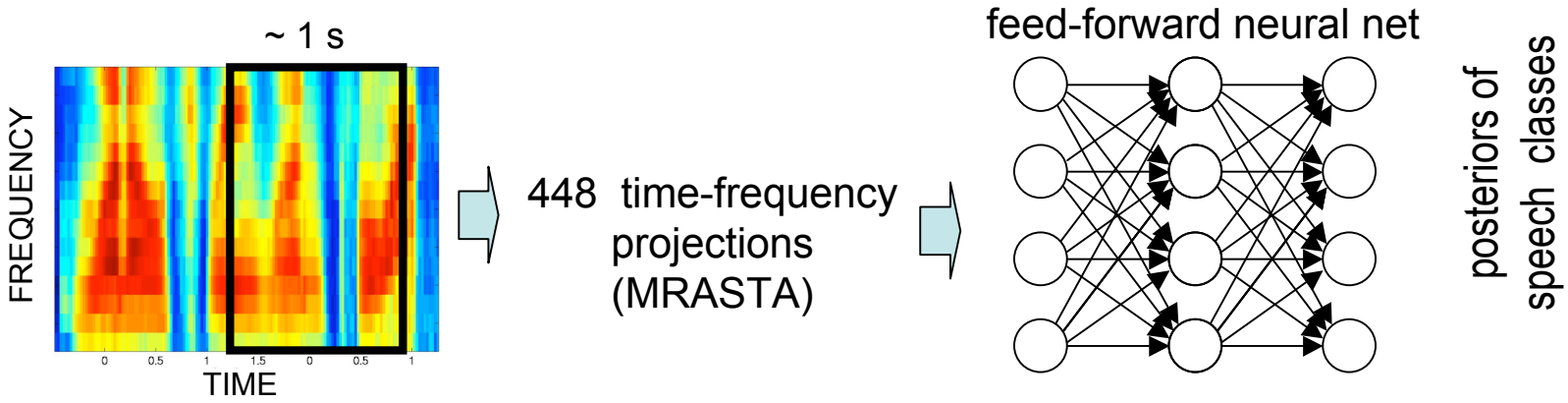
1. From both the sensory data and the context
2. From strong context cues when the sensory data impoverished
3. **From reliable sensory data even when the context suggest otherwise**

# One proposed solution
## (2007 JHU Summer Workshop)




Requires development of the “weakly constrained” recognizer

# Towards better “weakly constrained” recognizer




Posteriors can be also used for deriving features for a conventional HMM-based recognizer

# Posteriors in Conventional HMM System (TANDEM)