Learning about hearing from speech data

Hynek Hermansky
IDIAP Research Institute
Martigny, Switzerland

What Kinds of Knowledge about Humans Are Useful for Designing Machine Systems?

- Automatic recognition of speech

  effect (signal) = action
- Human auditory perception

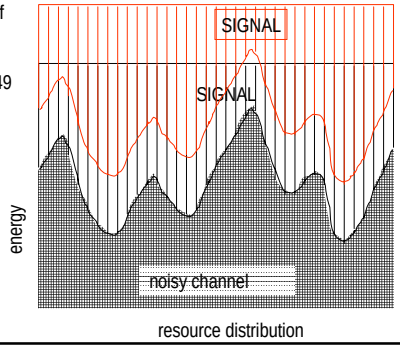
  effect (signal) = action
- Knowledge of human auditory perception !

Where to get the knowledge from?

- By studying biological systems
  - which properties are relevant ?
- From speech data
  - optimized machine processing could (and should) be consistent with **relevant** properties of human hearing

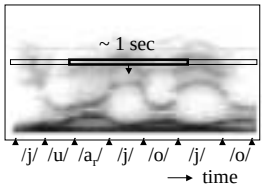
Knowledge *from* Data

Optimal distribution of signal energy in a noisy channel
Shannon 1949

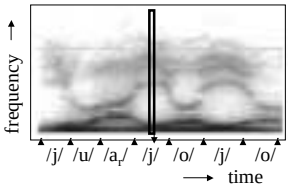


Linear Discriminant Analysis (LDA)

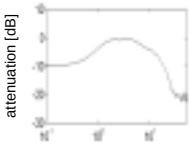
LDA gives FIR filters for processing time trajectories of spectral energies



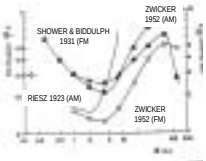
LDA gives basis for projection of spectral space



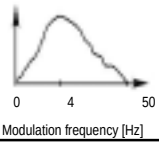
LDA-derived optimal filter for temporal processing of speech features



Perception of modulation

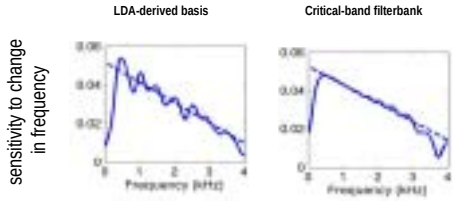


Modulation spectrum of speech



Optimizing spectral basis for speaker-independent ASR

Malayath and Hermansky, Speech Communication 2003



Non-uniform frequency sensitivity of hearing (and other implications)

- Fletcher 1930 (simultaneous masking)
  - Critical bands of hearing (increasing with frequency)
- **What happens outside the critical band does not affect detection of events within the band !**
- Recognition of nonsense CVC syllables [Fletcher/Allen]
  - final error in human phoneme recognition is given by product of errors in (articulatory) sub-bands

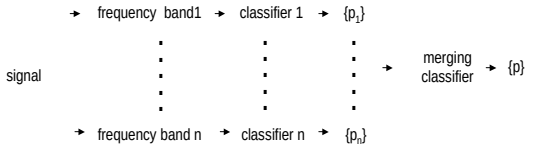
Independent processing of parts of signal spectrum?

Poor man's scene analysis ?

- Subdivide stimulus into a number of information sub-streams
  - ears, eyes, nose, fingers, mouth
  - further sub-division within each sense (e.g. frequency selectivity, sensitivity to rate-of-change,...)
- Select sub-streams with most favorable SNRs, alleviate the rest
- Get the information (likelihoods of events?) from the selected information sub-streams

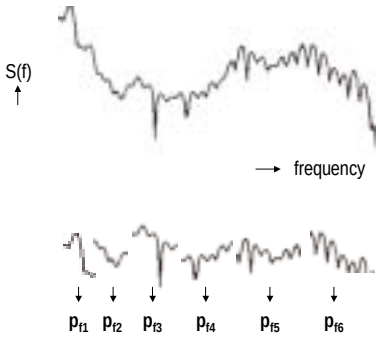
Multi-band recognition of speech

– Hermansky, Tibrewala, and Pavel, ICSLP96



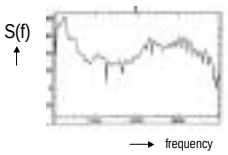
Efficient in recognition of partially corrupted speech

Goodbye to spectral shape



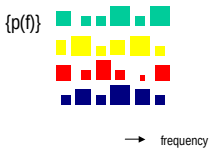
Spectral Envelope

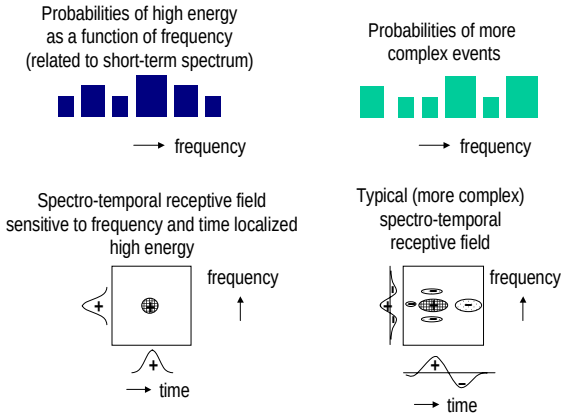
Vector of spectral energies derived from short segment of signal



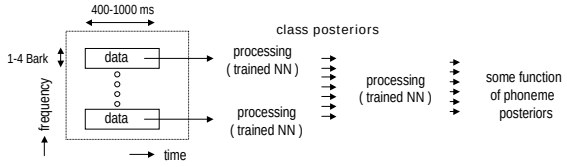
Matrix of posteriors

Matrix of posterior probabilities of relevant sound events, derived from any relevant evidence





Putting it all together

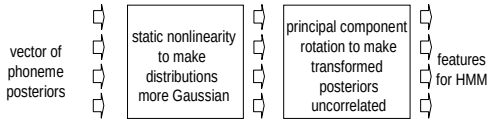


- frequency-localized to 1-4 Bark
  - cortical spectro-temporal patterns (e.g. Shamma and colleagues,…)
  - because it works (multi-band ASR, optimization studies)
- rather long temporal spans 0.4-1 s
  - cortical spectro-temporal patterns (e.g. Shamma and colleagues,…)
  - because that is where the information is (coarticulation, mutual info studies, …)
  - because it works (features for cellular industry, DARPA,…)

How to use estimates of posterior probabilities in current HMM systems?

(TANDEM technique)

Hermansky, Ellis and Sharma, ICASSP-2000



Some results

- about the same (likely somehow better) performance as conventional features in ASR
- performs well in combination with conventional system
  - about 8% relative error improvement in DARPA EARS program
  - part of the most accurate system in AURORA European Telecommunication Standards Institute initiative (more than 50 % relative error improvement on noisy data)

Conclusions

- **data-guided processing (trained on dev data) can be consistent with properties of hearing**
  - features as a function of posterior probabilities of classes
  - longer time spans (300-1000 ms) in feature extraction
  - hierarchical processing
    - frequency-localized features first
    - information fusion of frequency-localized features