“Biologically-Inspired” Engineering

Hynek Hermansky
IDIAP Martigny, Switzerland
“biologically-inspired” processing

production

\[ \text{signal} \]

\[ \text{effect} \]

perception

engineering

\[ \text{effect (signal)} \]

• good engineering
  – optimal engineering may be consistent with biology
    • speech evolved to be perceived
    • human perception as the optimal receiver of speech
speech signal
high entropy
(high information rate)

needs knowledge

recognition
machine

message
low entropy
(low information rate)

↑ data ? ↑ “knowledge” ?
(opinions)
my journey from engineering to “biology”

\[ s(t) \rightarrow P(\omega) \xrightarrow{r} r(t) \rightarrow \text{LP coefficients} \]

\text{FT} \quad \text{IFT} \quad \text{durbin}

\text{lp fit} \quad \text{root lp fit (root = 0.33)}

\text{loudness} = \text{intensity}^{0.33}

\text{PLP}

- critical-band spectral resolution
- equal loudness sensitivity
- root compression
critical band - like spectral resolution

- MCEP, PLP,…
  - sensitivity to spectral change higher at low frequencies (mel scale, Bark scale, ERB scale,..)

- LDA-derived spectral bases (6 hours of hand-labeled data)
  Malayath and Hermansky, Speech Communication 2003

Cosine basis  LDA-derived bases  Critical-band filterbank
constraints on speech production

- **RASTA processing**
  - eliminate features that change too slow or too fast
  - experimentally-derived band-pass filter

- **LDA-derived filter**
  - 6 hours of hand-labeled data
    - van Vuuren and Hermansky [ICSLP 96]
modulations in acoustics

Relative importance of components of modulation spectrum of speech
[Kaneda et al, Speech Communication 1999]

![Graph showing modulation thresholds and contributions to recognition performance.](image)
Modulation frequency [Hz]

0          4                  50

Modulation spectrum of speech

Optimized processing

Perception of Sinusoidal FM (after G. Green)
data-guided features

• some knowledge should be build-in the ASR system
  – more knowledge means less training is necessary
• no knowledge better than wrong knowledge
  – relevant (speech-specific and task-independent) knowledge is in the data
• hybrid approach
  – knowledge-guided structure of the model
  – data-derived parameters of the model
one simple architecture (TANDEM)

90 ms

neural network (multi-layer perceptron) → gaussianize → whiten → features for HMM

development data (OGI Stories)

WER on SPINE as a function of amount of training

word error rate

amount of training data [%]

PLP

PLP&TANDEM
prior knowledge: auditory neural processing

- ear
- cochlea
- cochlear nucleus
- inferior colliculus
- medial geniculate body
- auditory cortex

acoustic - mechanical conversion

frequency analysis

10^{-3}

10^{-2}

10^{-1}

gradual reduction of “information rate”
(but responding to more complex stimuli)

reverse correlation technique
(which signal most likely triggers response?)
physiology of auditory cortex

Average of the first two principal components (83% of variance) along temporal axis from about 180 cortical receptive fields (from D. Klein, unpublished)
Coincidences?

Temporal principal components
from cortical receptive fields

Impulse response
of optimized RASTA filter
What can cortical receptive fields do?

- estimate of posterior probability of formant at a given frequency and at the given time?
  - CRF as 2-D matched filter?

output of the filter would be estimate of the posterior

- more complex CRF (events) exist!
putting it all together

- TRAP-TANDEM
  - data-guided features based on frequency-independent processing of relatively long spans of signal
details

- up to 1 s long (300 ms minimum)
  - pre-processing by (truncated) cosine transform in time
- up to 3 critical bands
  - pre-processing by integration and by differentiation across frequency
- issue of target classes
  - “events” in sub-bands, context-independent phonemes in information fusion module
- issue of training data
  - task-independent in sub-bands, task-specific in fusion
results

• about the same performance as conventional features on small vocabulary (OGI digits) task
• combines well with conventional features
  – ETSI Aurora DSR
  – DARPA EARS
conclusions

• 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
• data-guided processing (trained on dev data)
• consistent with biology of hearing
“biologically-motivated” engineering

• not because it is “cool” but because it is often the most efficient way of dealing with cognitive signals

• we speak in order to hear …..

– [R. Jacobson]
Should airplanes flap wings?
maybe we just need to flap a bit harder...