Cognitive Components of Speech
-On Phonemes as Cognitive Components of Speech

Ling Feng
Lars Kai Hansen

Intelligent Signal Processing
Department of Informatics and Mathematical Modeling
Technical University of Denmark
Outline

1. Cognitive component analysis
2. Pre-processing
3. Unsupervised vs. Supervised
4. Timescales and meaning
5. Conclusion
What is Cognitive Component Analysis (COCA)?

COCA is the process of unsupervised grouping of data such that the ensuing group structure is well-aligned with that resulting from human cognitive activity.

- Unsupervised learning discovers statistical regularities;
- Human cognition is a supervised on-going process.

Human Behavior

Cognition is hard to quantify – its direct consequence: human behavior is easy to access and model.

**COCA - Definition**

- **Key Point**
  
  To investigate the consistency of statistical regularities in a signaling ecology and human cognitive activity! ...  
  
  **We are interested in the performance of unsupervised learning** $p(x|\theta)$ **and supervised learning** $p(y|x,\theta)$ **under equivalent representations.**

- **Hypothesis: independence and sparseness**

  Independence reduces perception-to-action mapping;  
  Optimal representation by sparse distributed codes.

---

Independence Hypothesis

- Independence dramatically reduces perception-to-action mapping by using factorial codes.
- Low level cognition is based on independence in natural ensemble statistics, e.g. visual feature extraction, color imagery, natural sound coding, even video data, etc. in primary sensory systems.
- The activation of each visual cortical feature detector is supposed to be as statistically independent from the others as possible.
- The receptive field properties of auditory nerve cells invoke a strategy of sparse independent manner to represent natural sounds.

Linear mixture of independent topics in text analysis

Sparse ‘ray-structure’

One-to-one correspondence

Using the magnitude of the source signals as a classification scheme, we get more than 90% classification accuracy.
Preprocessing pipeline

- MFCC
  - *Does ear work as a Fourier analyzer?*
  - Non-linear frequency perception
  - Critical band
- Stacking
  - The simplest method for feature integration.
- Energy Based Sparsification
  - Filter out the small (weak) signals
  - Emulate the **detectability** and **sensory magnitude**
- PCA (LSI)
  - The basis of cognitive processes

Phonemes-LSI

TRAINING DATA

TEST DATA

S O F A

CLIPPED CEPSRALS: |z| > 1.7

[a] PHONEME IN 'S' AND 'F'
The stable phoneme-relevant cognitive components (e.g. /e/ sound) are understood as ‘invariant cue’ characteristics of speech.

The perceived signals are derived as stable phonetic features despite of the different acoustic properties produced in different trials and different speakers.
Unsupervised vs. Supervised

We are interested in the performance of unsupervised learning and supervised learning under equivalent representations.

- ICA+Naive Bayes classifier vs. Mixture of Gaussian

Unsupervised learning: Unsupervised-then-supervised learning scheme to represent the ‘ecological’ grouping.

- ICA

- Naive Bayes

- Mixture of Gaussian

\[
x = As
\]

\[
p(C_i | s) = \frac{p(s | C_i)p(C_i)}{\sum_i p(s | C_i)p(C_i)}
\]

\[
p(s | C_i) = \prod_{j=1}^{k} p(s_j | C_i)
\]

\[
p(C_i | x) = \frac{p(x | C_i)p(C_i)}{\sum_i p(x | C_i)p(C_i)}
\]

\[
p(x | C_i) = \sum_j p(x | j, C_i)p(j | C_i)
\]
Time scales and meaning

- Music features are categorized into 3 time scales:
  - short time scale (30ms): instant frequency, e.g. harmonics and pitch;
  - medium time scale (~700ms): timbre, modulation;
  - long time scale (~10s): perceptual information, e.g. beat and mood.

- In COCA experiments:
  - at 10-40ms, there are generalizable ‘fingerprint’ of phonemes;
  - at 1 s, there are generalizable speaker specific sparse components.
  - We are interested in what we can discover with different time scales: gender? Age? Height?...

Experiments - Phonemes

Data: TIMIT database
Data preparation:
- Speech from 46 speakers (23 male, 23 female), reading 10 sentences
- Group phonemes into 3 classes: Vowels; Fricatives and Others;
- Stack features with a variety of time scales: from 20ms to 1100ms;
- Sparsify features with diverse thresholds $z$: to keep the retained energy from 100% to 65%.
**Error rate comparison**

For the given time scales and thresholds, data locate around \( y = x \), and the correlation coefficient \( \rho = 0.67, \ p < 1.38 \times 10^{-9} \).

**Sample-to-sample correlation**

- Three groups: vowels eh, ow; fricatives s, z, f, v; and stops k, g, p, t.
- 25-d MFCCs; EBS to keep 99% energy; PCA reduces dimension to 6.
- Two models had a similar pattern of making correct predictions and mistakes, and the percentage of matching between supervised and unsupervised learning was 91%.
**Sample-to-sample correlation**
- Three groups: vowels eh, ow; fricatives s, z, f, v; and stops k, g, p, t.
- 25-d MFCCs; EBS to keep 99% energy; PCA reduces dimension to 6.
- Two models had a similar pattern of making correct predictions and mistakes, and the percentage of matching between supervised and unsupervised learning was 91%.

**posterior probability comparison**
- One experiment: 100 ms with 97% remaining energy.
- If two models are the exact match, we should expect that the posterior probabilities locate along the diagonal of the histograms with high distribution at (1, 1) and (0, 0).
- The matching in this case is around 57%.
Conclusion

- COCA is the process of unsupervised grouping of data such that the ensuing group structure is well-aligned with that resulting from human cognitive activity.

- Unsupervised vs. Supervised learning
  A devised protocol to test the consistency of statistical regularities (unsupervised learning) and human cognitive processes (supervised learning of human labels).

- The comparison has been carried out at different levels: error rate comparison; sample-to-sample correlation; posterior probability comparison.

- The protocol has successfully revealed the consistency of two classifications.