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

COCA - Definition

- **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.

L.K. Hansen, P. Ahrendt, and J. Larsen, “Towards cognitive component analysis”. In *AKRR’05* –International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning. Jun 2005.

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 $p(\mathbf{x} \mid \theta)$ **and supervised learning** $p(\mathbf{y} \mid \mathbf{x}, \theta)$ **under equivalent representations.**

- **Hypothesis: independence and sparseness**

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

-D. J. Field, “What is the goal of sensory coding?, ” Neural Computation, vol. 6, pp. 559–601, 1994.
-B. A. Olshausen and D. J. Field, "Sparse coding of sensory inputs,” Current Opinion in Neurobiology, vol. 14(4), pp.481-487, 2004

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.

-A. J. Bell and T. J. Sejnowski, "The 'independent components' of natural scenes are edge filters," *Vision Research*, vol. 37, pp.3327–3338, 1997.

-P. Hoyer and A. Hyvrinen, "Independent component analysis applied to feature extraction from colour and stereo images," *Network: Comput. Neural Syst.*, vol. 11, pp. 191–210, 2000.

-M. S. Lewicki, "Efficient coding of natural sounds," *Nature Neuroscience*, vol. 5, pp. 356–363, 2002.

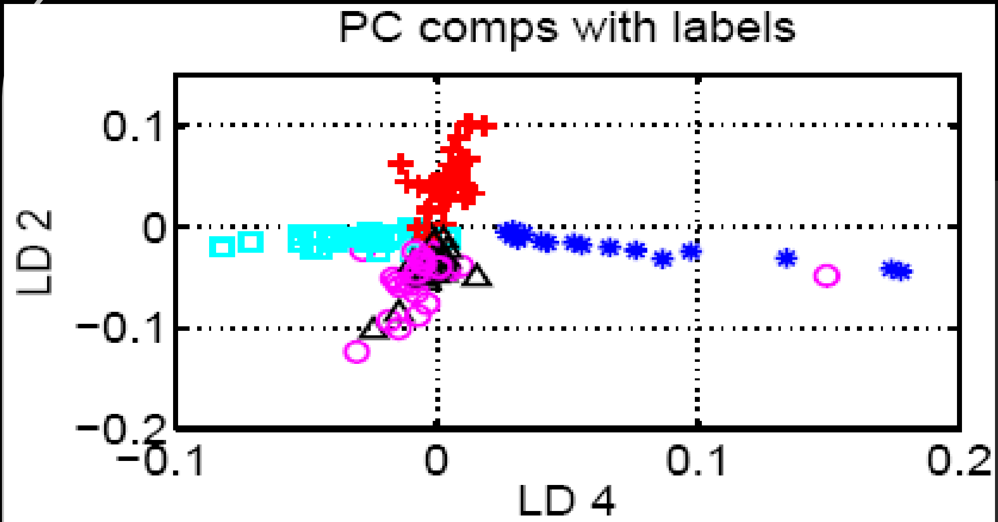
-E. Doi and T. Inui and T. W. Lee and T. Wachtler and T. J. Sejnowski , " Spatiochromatic Receptive Field Properties Derived from Information-Theoretic Analyses of Cone Mosaic Responses to Natural Scenes, " Neural Comput., vol. 15(2), pp. 397-417, 2003.

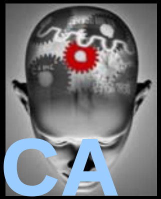
-J. H. van Hateren and D. L. Ruderman, "Independent Component Analysis of Natural Image Sequences Yields Spatio-Temporal Filters Similar to Simple Cells in Primary Visual Cortex," Proc. Biological Sciences, vol. 265(1412), pp. 2315-2320, 1998.

-B. A. Olshausen and K. N. O'Connor , "A new window on sound ," Nature Neuroscience, vol. 5, pp. 292-294, 2002.

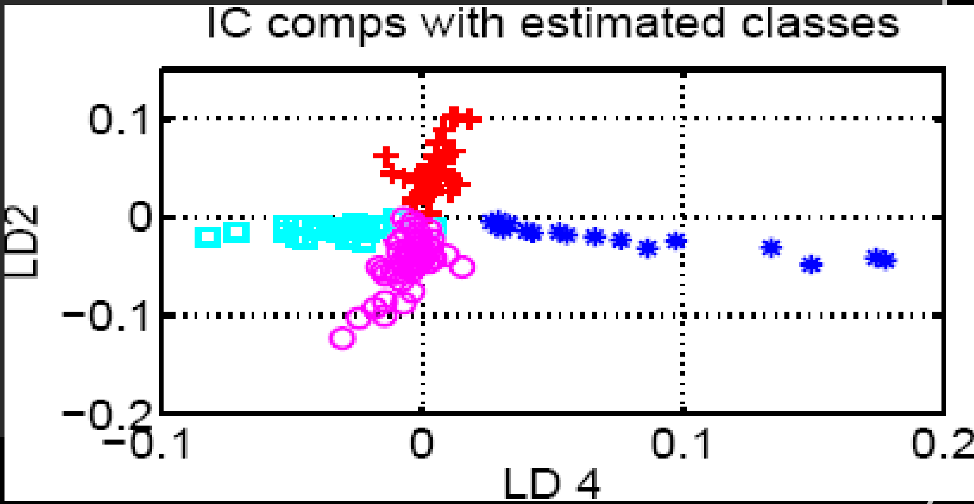
-H.B. Barlow, "Unsupervised learning," Neural Computation, vol. 1, pp. 295–311, 1989.

Example of ICA

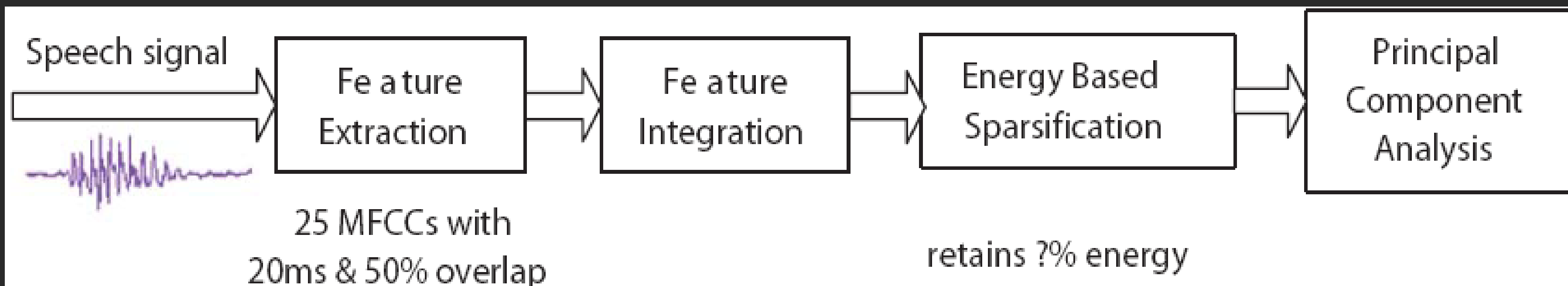




- Linear mixture of independent topics in text analysis
- Sparse ‘ray-structure’
- One-to-one correspondence
- Using the magnitude of the source signals as a classification scheme, we get more than 90% classification accuracy.



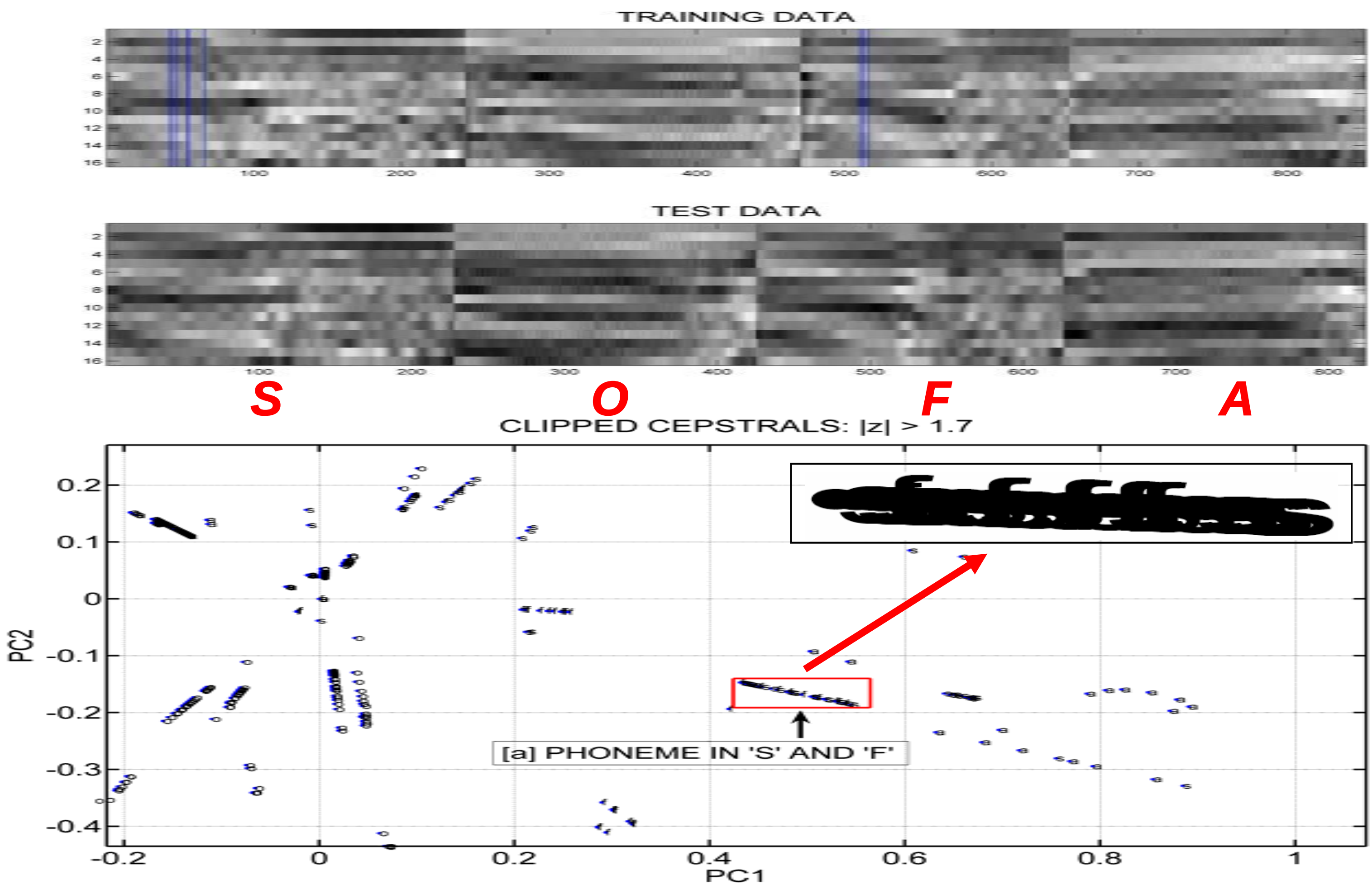
Preprocessing pipeline



- **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

W. Kintsch, “Predication,” *Cognitive Science*, vol. 25, pp.173–202, 2001.

Phonemes-LSI



Invariant Cue

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

$$\mathbf{x} = \mathbf{A}\mathbf{s}$$

- Naive Bayes

$$p(C_i \mid \mathbf{s}) = \frac{p(\mathbf{s} \mid C_i)\, p(C_i)}{\sum_i p(\mathbf{s} \mid C_i)\, p(C_i)}$$

$$p(\mathbf{s} \mid C_i) = \prod_{j=1}^{k} p(s_j \mid C_i)$$

Unsupervised -then- Supervised learning scheme

- Mixture of Gaussian

$$p(C_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C_i)\, p(C_i)}{\sum_i p(\mathbf{x} \mid C_i)\, p(C_i)}$$

$$p(\mathbf{x} \mid C_i) = \sum_j p(\mathbf{x} \mid j, C_i)\, p(j \mid C_i)$$

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?...

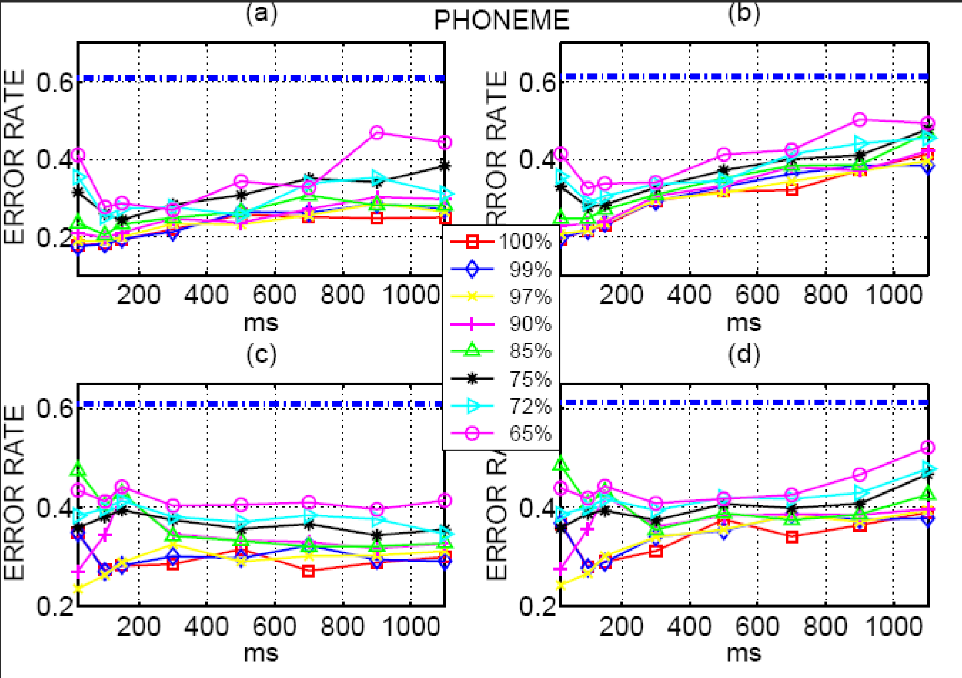
Meng, A., Ahrendt, P., & Larsen, J. (2005). Improving Music Genre Classification by Short-Time Feature Integration. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, pp. 497-500.

Experiments-Phonemes

Data: TIMIT database

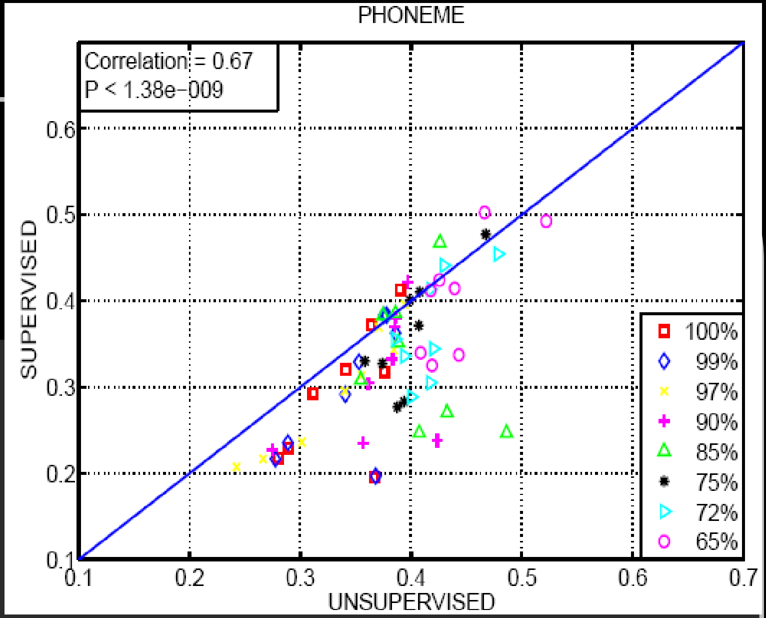
Data preparation:

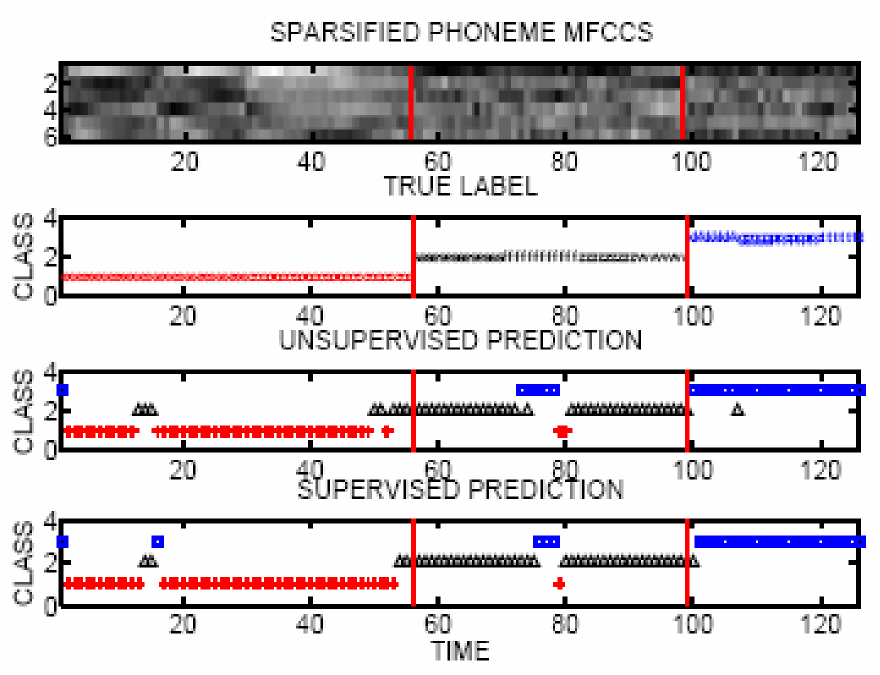
- Speech from 46 speakers (23 male, 23 female), reading 10 sentences
- Group phonemes into 3 classes: Vowels; Fricatives and Others;
- Stack features with a variety of time scales: from 20ms to 1100ms;
- Sparsify features with diverse thresholds z: to keep the retained energy from 100% to 65%.



Error rate comparison

For the given time scales and thresholds, data locate around $y = x$, and the correlation coefficient ρ=0.67, p<1.38e−09.



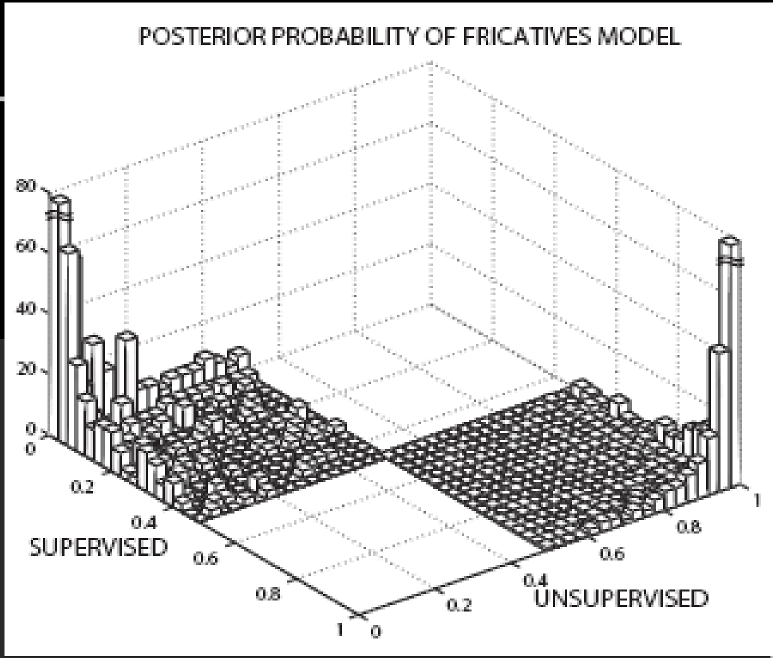


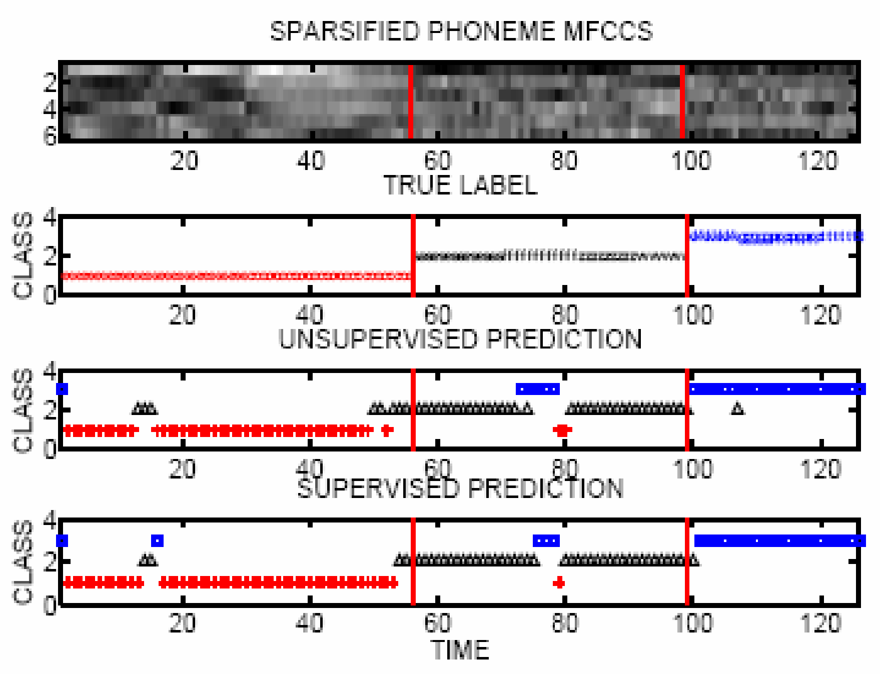
Sample-to-sample correlation

- 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%.





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%.

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.