# Cognitive Component Analysis

Ling Feng

Ph. D. Defense

October 31, 2008

Intelligent Sound Project
Intelligent Signal Processing
DTU Informatics

# Outline

- Introduction of COgnitive Component Analysis
  - definition & hypothesis
- Human cognition
  - human auditory system
- COCA
  - the preprocessing pipeline
- Hidden variable models
- Cognitive components of phoneme and Identity
- High-level COCA
- Conclusions

# Cognitive Component Analysis-definition

- Theoretical Issue:

  *Do the characteristics of human processors reflect statistical regularities revealed by unsupervised learning of perceptual inputs?*

- What is Cognitive Component Analysis?

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

  It aims at investigating the consistency of statistical regularities in a signaling ecology and human cognitive activity.

# COCA - Hypothesis

- Hypothesis: independence and sparseness

  *What is the source of the extensive and well-organized knowledge of the environment implied by the possession of an cognitive map or working model? - Barlow*

  It is the statistical regularities in the sensory messages which are recorded by the brain, in order to inform the brain what usually happens, and independence is one of the regularities.

  Based on Barlow's *minimum entropy coding*: feature detectors are the result of reduction process on the redundancy of sensory messages, and these detectors are statistically independent. Since the sensory information is encoded by a small number of neurons at a certain point of times, the statistical independent feature detectors are activated as rarely as possible. – sparseness!

  It does not only run for visual system, but also auditory system: The receptive field properties of auditory nerve cells invoke a strategy of sparse independent manner to represent natural sounds.

  The advantages of sparse representations:
  - the most effective means for storing patterns in the associative memories;
  - clears structures in natural inputs;
  - have advantages to designate complex data in a explicit and easy-to-read way;
  - saves the energy required for signaling in cortical neurons w.r.t. the low average firing rates.

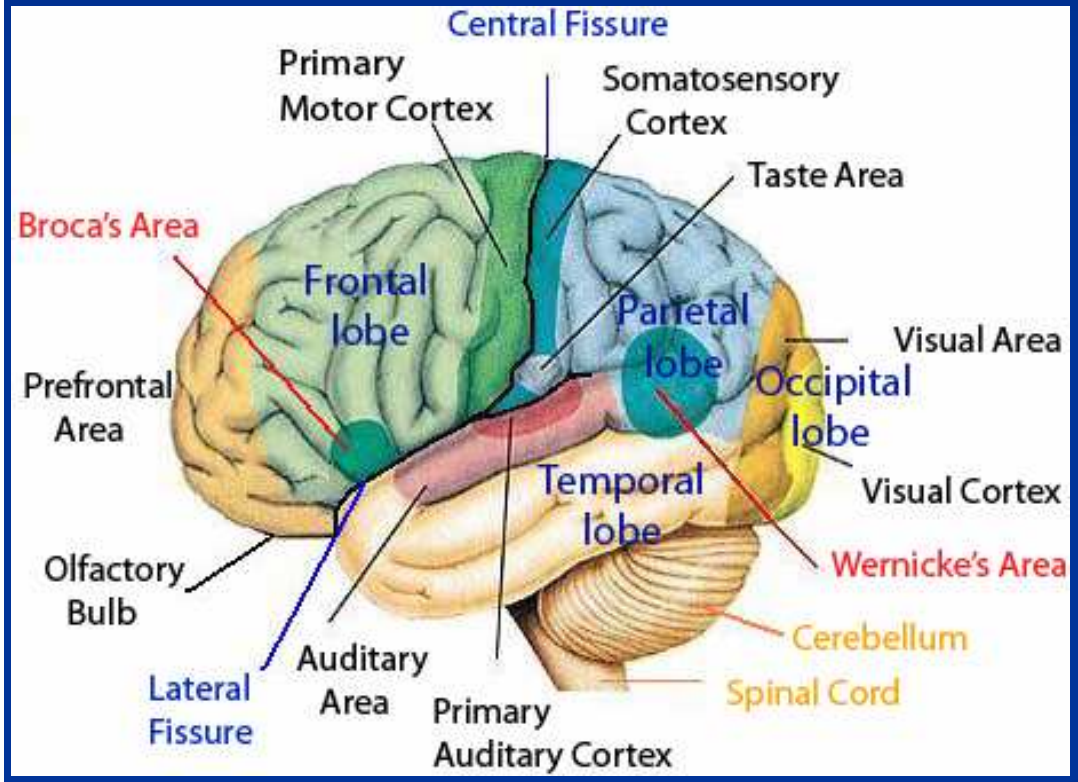
  This hypothesis is ecological: we assume that features that are essentially independent in a context defined ensemble can be efficiently coded using a sparse independent component representation.

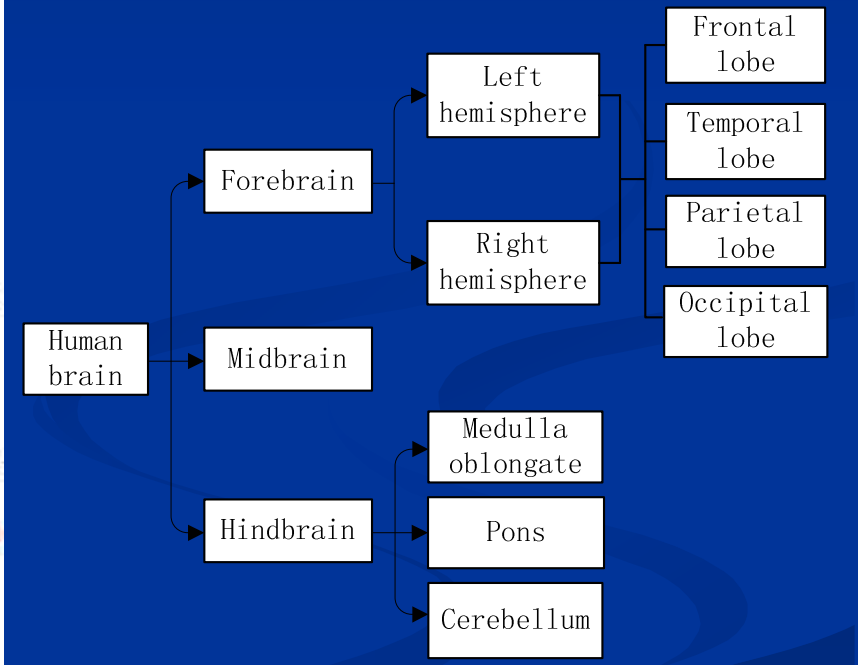
# Outline

- Introduction of COgnitive Component Analysis
  - definition & hypothesis
- Human cognition
  - human auditory system
- COCA
  - the preprocessing pipeline
- Hidden variable models
- Cognitive components of phoneme and Identity
- High-level COCA
- Conclusions

# Human Brain Structure

- *What is cognition?*



# Human Auditory system

- *Peripheral auditory system*

  - Outer ear: Its shape works as a amplifier.

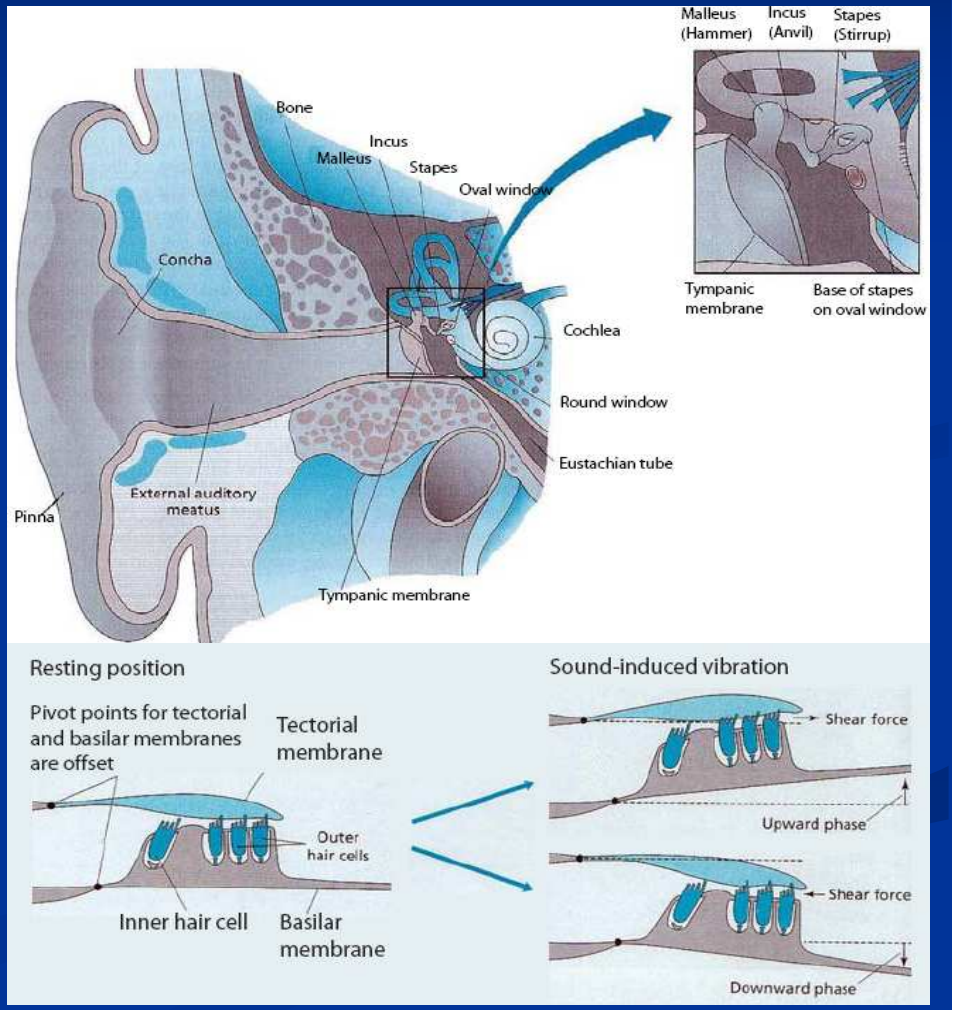
  - Middle ear: It transmits the vibration of the ear drum to the movement of the fluid inside the inner ear; and it maximizes the transmission by increasing the pressure with a ratio of 27 dB.

  - Inner ear: the hearing sense organ, cochlea, is snail-shell shaped structure, and is filled with lymph. 10*mm* in diameter, and 32−34*mm* long if straightened out.

- *Central auditory system*

  - includes a mass amount of neurons in the brainstem and the cerebral cortex.

  - its functionality and mechanism are not fully discovered.



# Human Auditory system

- *Human ear works as a Fourier analyzer?*

  The maximum response of a nerve fibre happens when the sound frequency matches the nerve fibre's *characteristic frequency*. Thus loosely speaking, the ear is behaving like a Fourier analyzer, where each sound can be decomposed to a collection of sine frequency components.

  Based on *Ohm's acoustical law,* humans are able to perceive the harmonics individually from a periodic sound.

- *Non-linear frequency perception*

  Human ears perceive frequency in 'mel' scale, which is linear below 1*k*Hz, and logarithmic above.

- *Critical band*

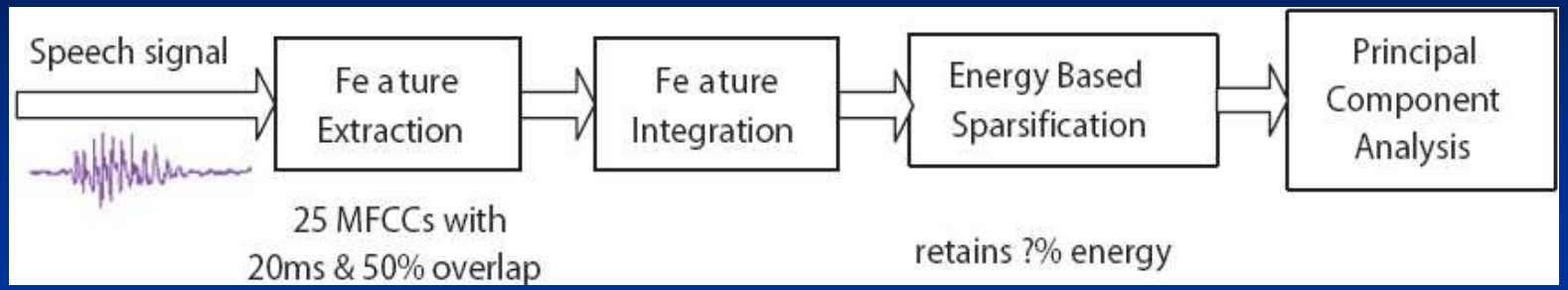
  Human auditory system uses a function like a band-pass filter to perceive signals, select frequencies within the bandwidth, and remove the rest.

  Each point on the basilar membrane can be seen as a band-pass filter with a center frequency corresponding to the characteristic frequency, and a bandwidth.

# Outline

- Introduction of COgnitive Component Analysis
  - definition & hypothesis
- Human cognition
  - human auditory system
- COCA
  - the preprocessing pipeline
- Hidden variable models
- Cognitive components of phoneme and Identity
- High-level COCA
- Conclusions

# COCA – Preprocessing Pipeline

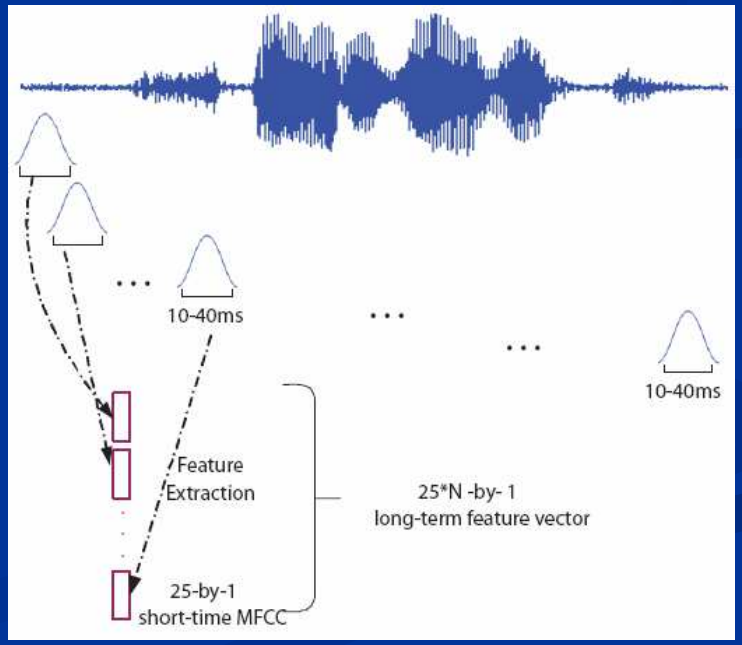



- *Feature: Mel-frequency Cepstral Coefficient (MFCC)*

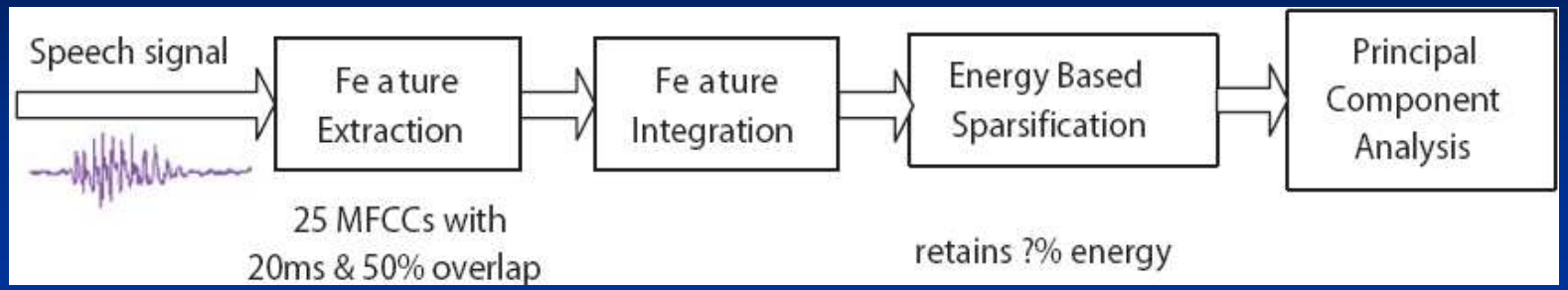
  MFCCs share two aspects with the human auditory system: A logarithmic dependence on signal power and a simple bandwidth-to-center frequency scaling so that the frequency resolution is better at lower frequencies.

  Critical band filters represent the frequency resolution of the peripheral human auditory system, and they also reflect the auditory system in a way that signals passing through different critical bands are processed independently

- *Feature Stacking*



# COCA – Preprocessing Pipeline




- *Energy Based Sparsification (EBS)*

  EBS emulates the detectability and sensory magnitude from perceptual principles.

  The ability of sensory organs to detect the environmental stimulus is reflected by a threshold.

- *Principal Component Analysis (PCA)*

  PCA has human-like performance in text analysis.

$$\mathbf{X} = \mathbf{U}\Lambda\mathbf{V}^T$$

$$\mathbf{Y} = \mathbf{U}_k^T\mathbf{X} = \Lambda_k\mathbf{V}^T$$

# Outline

- Introduction of COgnitive Component Analysis
  - definition & hypothesis
- Human cognition
  - human auditory system
- COCA
  - the preprocessing pipeline
- Hidden variable models
- Cognitive components of phoneme and Identity
- High-level COCA
- Conclusions

# Hidden variable models

$$\mathbf{y} = \Lambda\mathbf{x} + \boldsymbol{\varepsilon}$$

A number of unsupervised learning models share the same form with different constraints on variables. Two classic roles of unsupervised learning are clustering and dimensionality reduction.

- *Principal Component Analysis (PCA)*
  Constraints: **x** is multivariate Gaussian distributed with $N(0, I)$, $I$: identity matrix;
  $\varepsilon$=0; **y** is Gaussian distributed with $N(0, \Sigma)$ where $\Sigma = \Lambda\Lambda^T$ .

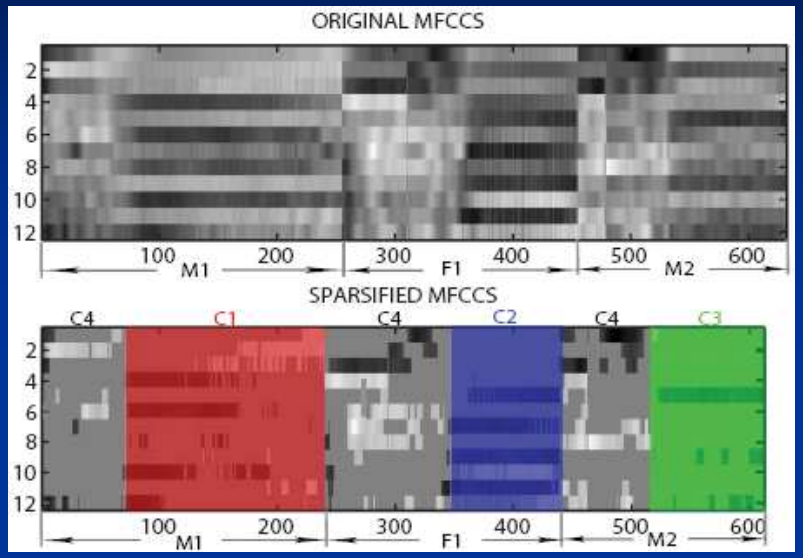
- *Factor analysis (FA)*
  Constraints: **x** is multivariate Gaussian distributed with $N(0, I)$, $I$: identity matrix;
  $\varepsilon$ is multivariate Gaussian noise, $N(0,\Psi)$, $\Psi$ is a diagonal matrix with different entries along the diagonal.
  **y** is Gaussian distributed with $N(0, \Sigma)$ where $\Sigma = \Lambda\Lambda^T + \Psi$.

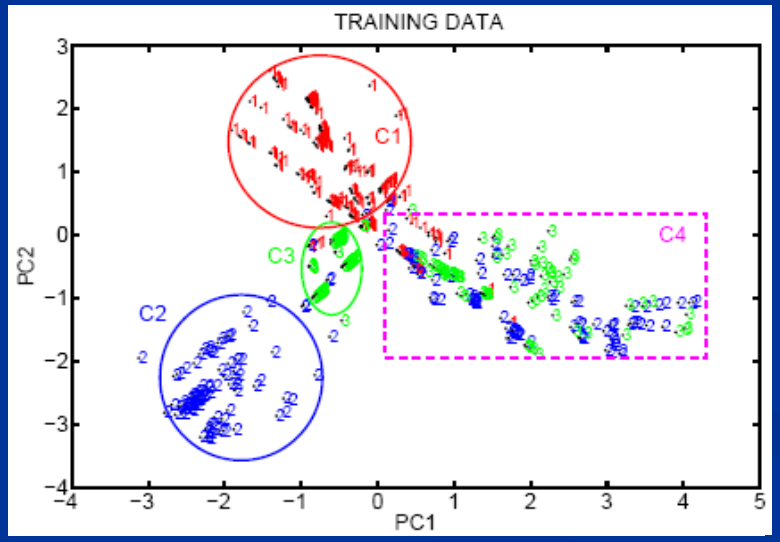
- *Independent Component Analysis (ICA)*
  Hidden variables **x** are assumed as independent *non-Gaussian* distributed sources.

# Outline

- Introduction of COgnitive Component Analysis
  - definition & hypothesis
- Human cognition
  - human auditory system
- COCA
  - the preprocessing pipeline
- Hidden variable models
- Cognitive components of phoneme and Identity
- High-level COCA
- Conclusions

# Cognitive Components of Phonemes



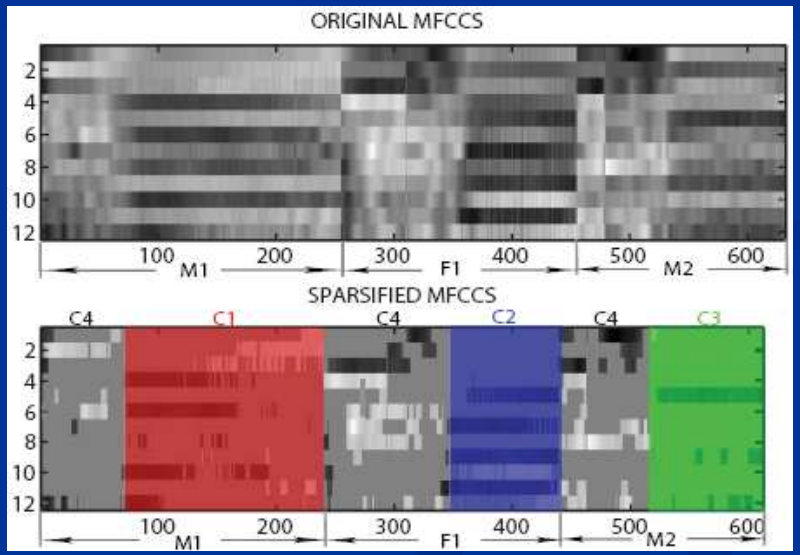

1: M1
2: F1
3: M2

- the letter ‘t’ sound (phonemes included: /t/ + /i:/) from 3 speakers: two male & one female from TIMIT database.
- 12-dimensional MFCCs
- the sparse linear mixture: ‘ray-structure’
- C1 to C3 represent /i:/ from 3 speakers, and C4 is the /t/ sound from all;
- region C2 and C3 follow the same ‘rays’ emanating from the origin with different amplitudes.
- F1 has part of the data locating apart from M1 and M2, which may imply speaker-specific properties.

# Cognitive Components of Phonemes - Invariant Cue

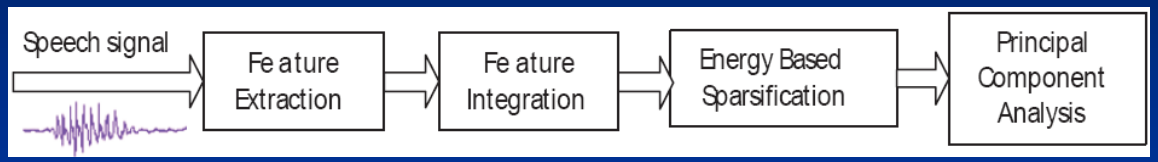
- *Speech signals may vary due to co-articulation, the relation between key features follows a consistent and invariant form*. (Damper)

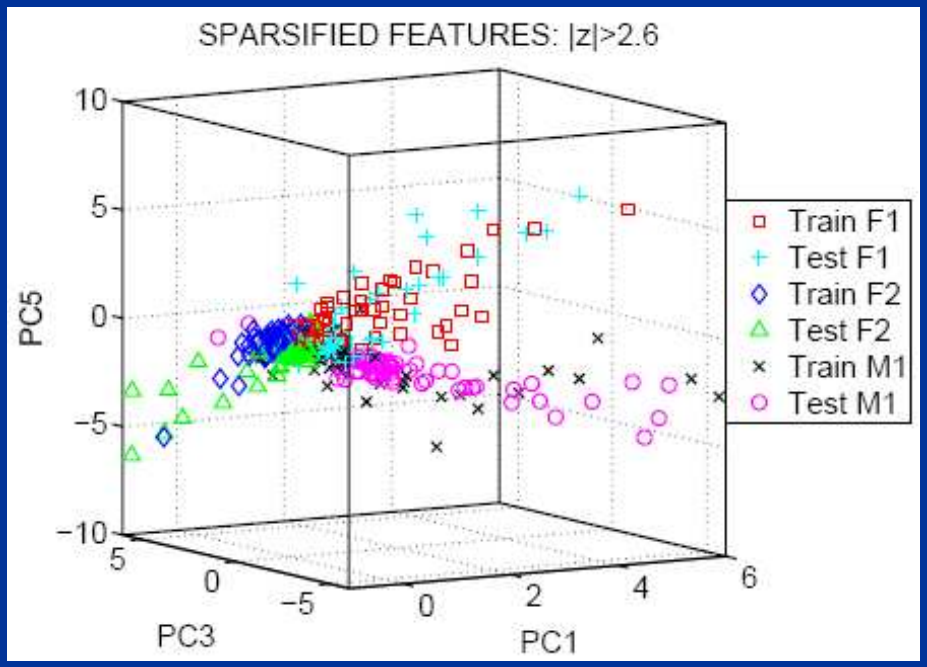
- *Perceived signals are derived as stable phonetic features, despite of different acoustic properties produced by different trials and speakers*.



# Cognitive Components of Speakers

## - From different text



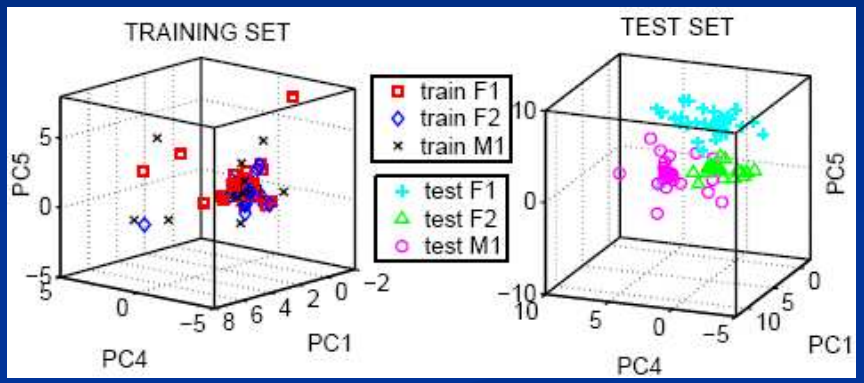



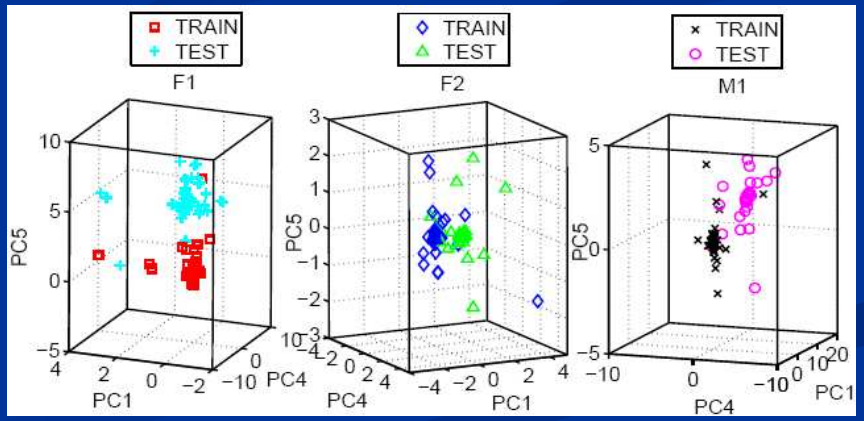

- 12-dimensional MFCCs from 20-ms frames
- Long-term feature at 1 sec time scale
- Sparse components for each individual speaker are evident, and ‘rays’ locate very much separately in the subspace.
- The generalizable ray structures of independent identities emanating from origin of the coordinate system without offsets.

# Cognitive Components of Speakers

## - From same text

- 12-dimensional MFCCs from 20-ms frames
- Long-term feature at 1 sec time scale
- Due to the same-text training and test sets show different patterns: training data from three speakers have large overlaps around origin of the coordinate system; while as ‘rays’ of test data tend to extend along a similar direction.
- A close depiction of the data scatter for each speaker individually, elucidates that the training and test data follow a similar scatter tendency with offsets.
- *We stipulate that it is the interaction between the text content and the speaker identity, which echoed the findings in the previously discussed experiment on ‘invariant cue’ of multi-speaker.*



# Outline

- Introduction of COgnitive Component Analysis
  - definition & hypothesis
- Human cognition
  - human auditory system
- COCA
  - the preprocessing pipeline
- Hidden variable models
- Cognitive components of phoneme and Identity
- High-level COCA
- Conclusions

# High-level COCA – unsupervised vs. supervised

- COCA definition: the process of unsupervised grouping of data such that the ensuing group structure is well-aligned with that resulting from human cognitive activity.
- Theoretical Issue: Do the characteristics of human processors reflect statistical regularities revealed by unsupervised learning of perceptual inputs?
- Human cognition is too sophisticated to model, however as the direct consequence human behavior is easier to access.
- Human cognition is represented by a classification rule, i.e. supervised learning of speech data and corresponding manually obtained labels.
- *The question is then reduced to looking for similarities between representations in supervised learning (of human labels) and unsupervised learning that simply explores statistical properties of the domain.*

# High-level COCA – unsupervised vs. supervised

- When sources are *sparse*, the mixtures have 'ray-structure' corresponding to overlaid lines on a scatter plot!

- Line orientations correspond to columns of the mixing matrix **A**

- The first set, *ICA-like density model*, is based on mixture of factor analyzers model. The modification follows the ideas from Soft-LOST and Hard-LOST (Line Orientation Separation Technique) models.

- Mixture of FA

$$p(\mathbf{y}) = \sum_{\mathbf{x}} p(\mathbf{y},\mathbf{x}) = \sum_{\mathbf{x}} p(\mathbf{y}\,|\,\mathbf{x})\,p(\mathbf{x}) = \sum_{i=1}^{m}\sum_{\mathbf{x}} p(\mathbf{y}\,|\,\mathbf{x})\,p(\mathbf{x}\,|\,i)\,p(i)$$

# High-level COCA – ICA-like MFA

- EM procedure to identify line orientations:

  - E-step, calculate the log posterior probability log $p(i|\mathrm{y})$;

$$\log p(i|\mathbf{y}) = \log p(\mathbf{y}|i) + \log p(i) - \log p(\mathbf{y})$$
$$\propto \log p(\mathbf{y}|i) + \log p(i) = -\frac{1}{2}\log|2\pi(\Lambda_i\Lambda_i^T + \Psi_i)| - \frac{1}{2}\mathbf{y}^T(\Lambda_i\Lambda_i^T + \Psi_i)^{-1}\mathbf{y} + \log p(i),$$
$$-\frac{1}{2}\log|2\pi(\Lambda_i\Lambda_i^T + \Psi_i)| = -\frac{1}{2}\left(\sum_{j=1}^{k}\log(\lambda_{ij} + \sigma_{ij}^2) + \sum_{j=k+1}^{d}\log\sigma_{ij}^2\right) + const.$$
$$-\frac{1}{2}\mathbf{y}^T(\Lambda_i\Lambda_i^T + \Psi_i)^{-1}\mathbf{y} = \frac{1}{2}\mathbf{y}^T(\Psi_i^{-1}\Lambda_i(\Lambda_i^T\Psi_i^{-1}\Lambda_i + I)^{-1}\Lambda_i^T\Psi_i^{-1} - \Psi_i^{-1})\mathbf{y}.$$

  - M-step, adjust lines to match the points assigned to them, i.e. to calculate the covariance matrix of data points within each cluster/FA.

- We reduce the *k* dimensional factor loadings to a single column vector, therefore we assign the eigenvector with the largest eigenvalue of each cluster as the new line vector.

# High-level COCA – ICA-like MFA

- Supervised ICA-like MFA

$$\begin{aligned}\log p(i|\mathbf{y},l) &= \log \frac{p(\mathbf{y},l|i)p(i)}{p(\mathbf{y},l)} \\ &= \log \frac{p(\mathbf{y}|i)p(l|i)p(i)}{p(\mathbf{y},l)} \\ &\propto \log p(\mathbf{y}|i) + \log p(l|i) + \log p(i).\end{aligned}$$

- We train supervised and unsupervised models on the same feature set. For the unsupervised model we first train only using features **y**. When the density model is optimal, we clamp the mixture density model and train only the cluster tables $p(l|i)$, $i = 1, \ldots, m$, using training set labels. This is **unsupervised-then-supervised** learning.

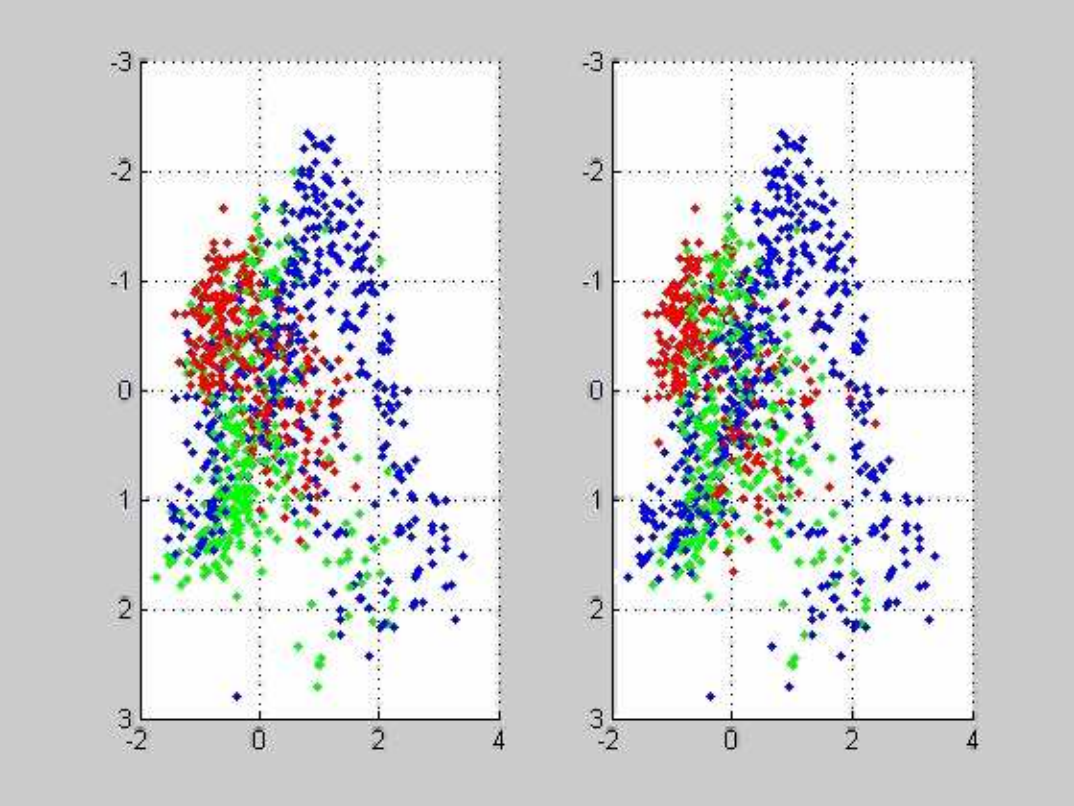
- For supervised learning both feature and label sets are modeled.

- A simple protocol for checking the cognitive consistency: Do we find the same representations when we train them with and without using 'human cognitive labels'?

# High-level COCA – ICA-like MFA

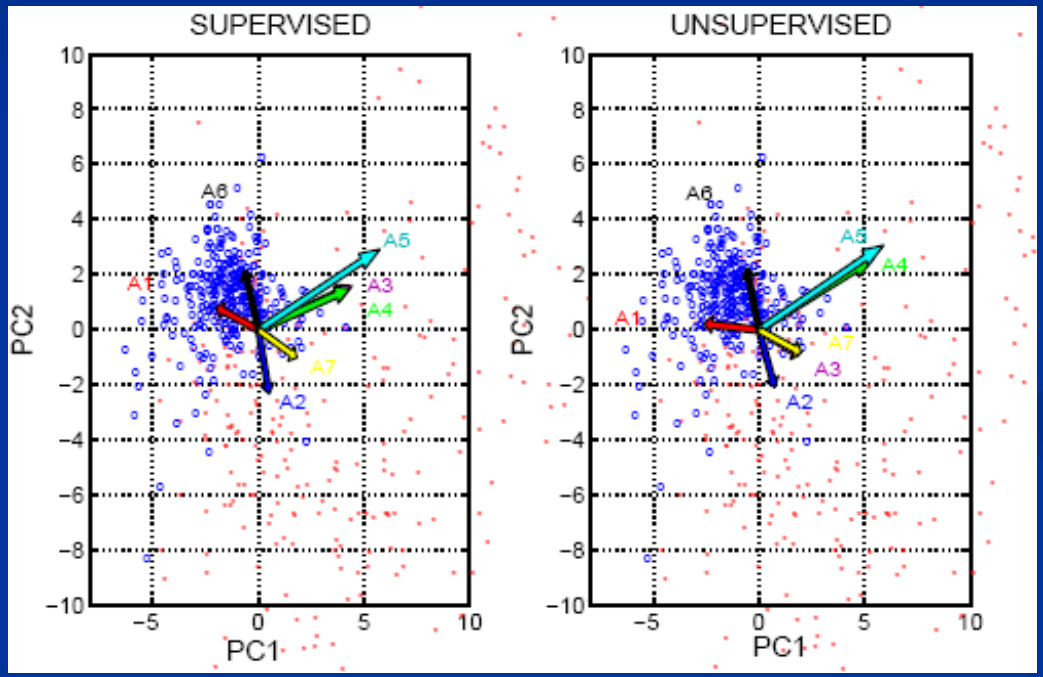
Vowels iy (blue), ay (red) and ow (green)

**Supervised** **Unsupervised**




Err rate: 27.9% (supervised) 30.6% (unsupervised)

# High-level COCA – ICA-like MFA




Gender detection

Male (blue), Female (red)

- Arrows: the first column vectors of the loading matrices for unsupervised and supervised models
-Vectors are normalized
-We align columns vectors from each model based on the correlation coefficients of the recovered factors $\mathbf{X}^{unsup}$ and $\mathbf{X}^{sup}$.

In 2-D space the angles between vector pairs are 37.180, 13.900, 53.380, 8.650, 0.690, 3.970, 8.410.

# High-level COCA
## - ICA+ Bayesian Models

- Unsupervised learning: ICA + naive Bayes classifier

  - ICA on features **y**:

  - Naive Bayes classifier on **s** and labels:

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

$$p(\mathbf{C}_i|\mathbf{s}) = \frac{p(\mathbf{s}|\mathbf{C}_i)p(\mathbf{C}_i)}{\sum_k^i p(\mathbf{s}|\mathbf{C}_i)p(\mathbf{C}_i)}$$

$$p(\mathbf{s}|\mathbf{C}_i) = \prod_{j=1} p(s_j|\mathbf{C}_i)$$

**Unsupervised -then- Supervised learning scheme**

- Supervised learning: Mixture of Gaussian model

Diagonal covariance matrices are assumed. Thus axes of the resulting Gaussian clusters are parallel to the axes of the input data space.

$$p(\mathbf{C}_i|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{C}_i)p(\mathbf{C}_i)}{\sum_i p(\mathbf{y}|\mathbf{C}_i)p(\mathbf{C}_i)}$$

$$p(\mathbf{y}|\mathbf{C}_i) = \sum_j p(\mathbf{y}|j, \mathbf{C}_i)p(j|\mathbf{C}_i)$$

# High-level COCA
## - Experiments

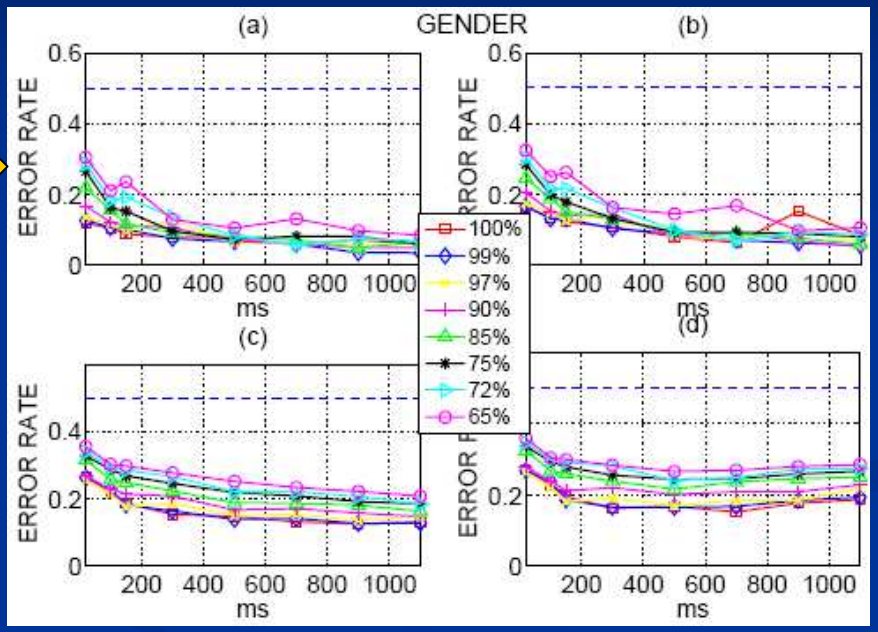
- Five cognitive indicators: phoneme, gender, age, height & speaker identity.
- 46 speakers (23F, 23M) from TIMIT; speech covers 60 phonemes, and age is in [21 72] with 22 values. 6 sentences for training and 4 for testing.
- Stack features into time scales: [20 1100] *ms*; Sparsify features with different thresholds.
- Phonemes: pre-group into 3 classes: Vowels; Fricatives and Others.
- Age: pre-grouped into 4 sets, in order to keep an approximate even population among sets.
- Height: pre-group into 6 classes with ~3 inches range of each class.

# Unsupervised vs. Supervised
## - Comparison Methods

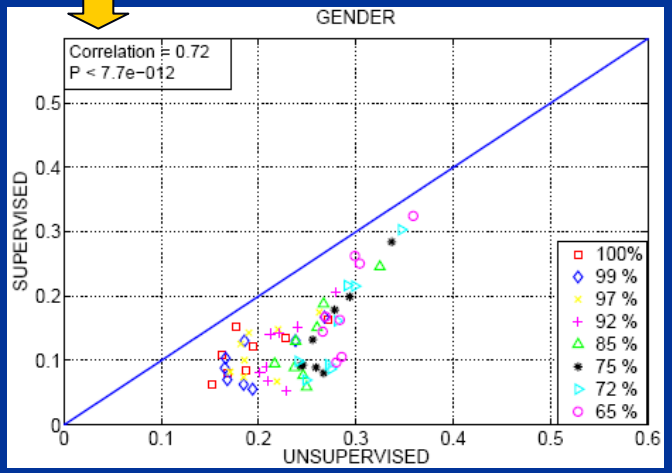
- Error rate comparison

  - The tendency of curves tell us the approximate time scale, at which the cognitive task was best modeled.

  - High correlation between error rates of the paired models indicate similarity of the representations.




The error rates as a function of time scale




The correlation of test error rates in the form of unsupervised vs. supervised.

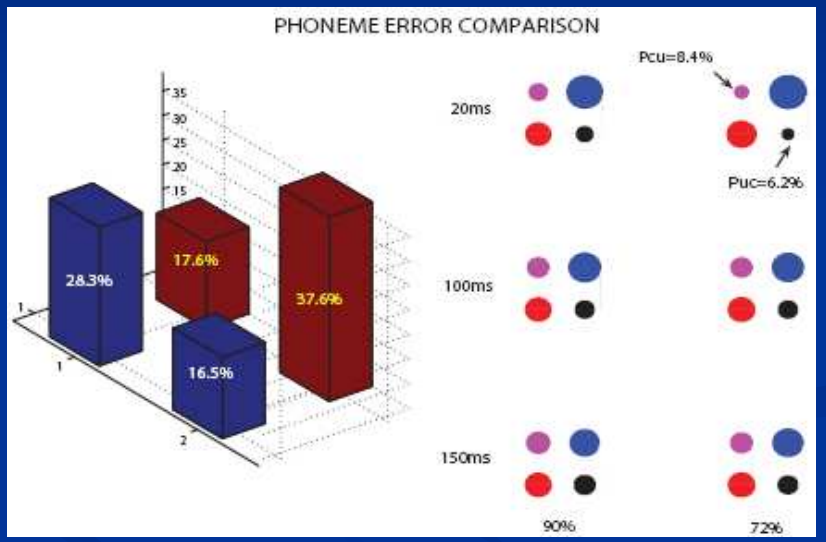
Table 5.1: Recommended time scales for modeling Phonemes, Gender, Age, Height, Identity.

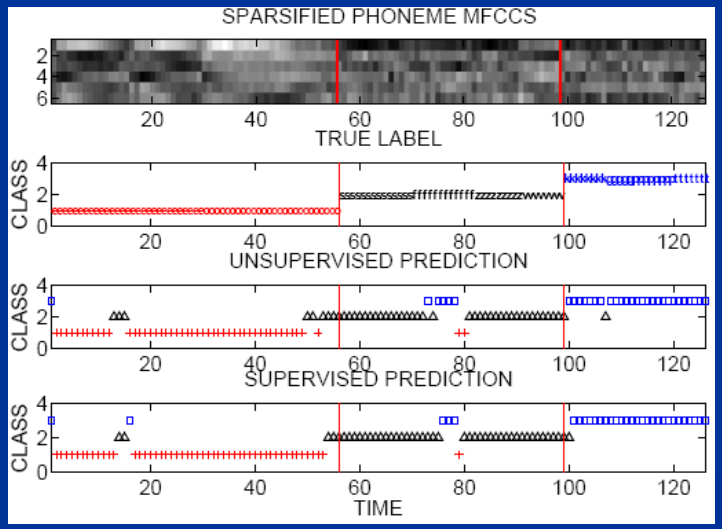
| $(ms)$ | Phoneme | Gender | Age | Height | ID |
|---|---|---|---|---|---|
| Timescale | 20 | 300-500 | $500 < t < 1000$ | $\geq 1000$ | $> 1000$ |

# Unsupervised vs. Supervised
## - Comparison Methods

- Sample-to-sample based comparison

$$R_{cc} = \frac{r_{cc}}{(1-r_{sup})(1-r_{usup})}, \quad R_{uu} = \frac{r_{uu}}{r_{sup}r_{usup}},$$
$$R_{cu} = \frac{r_{cu}}{(1-r_{sup})r_{usup}}, \quad R_{uc} = \frac{r_{uc}}{r_{sup}(1-r_{usup})}.$$
$$P_{ij} = \frac{R_{ij}}{\sum_{mn}(R_{mn})}, \quad m,n = (c,u); \quad i,j = (c,u).$$



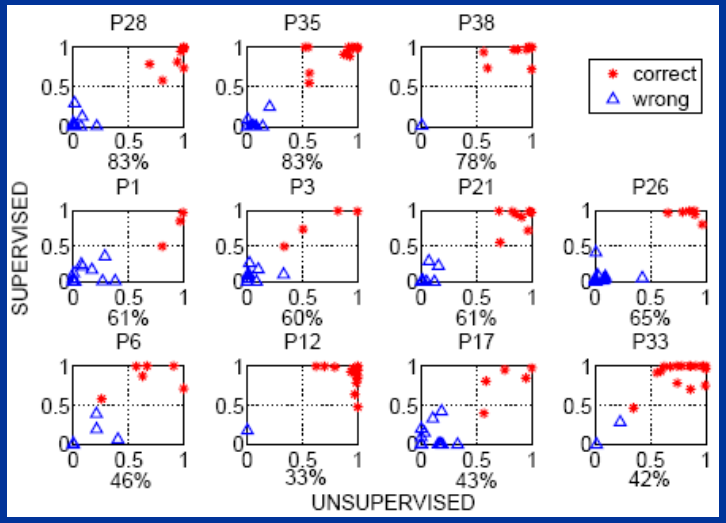




3 groups: vowels eh, ow; fricatives s, z, f, v; and stops k, g, p, t.
- 25-d MFCCs; EBS to keep 99% energy; PCA reduces dimension to 6.
- Two models had a similar pattern of making correct predictions and mistakes, and the percentage of matching between supervised and unsupervised learning was 91%.

# Unsupervised vs. Supervised - Comparison Methods

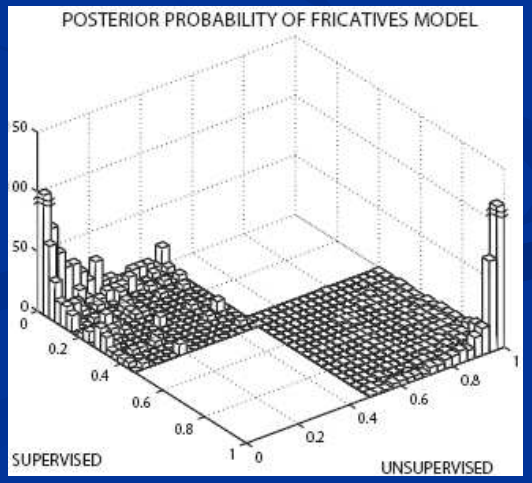
- Posterior probability comparison

When both models making the same predictions, we can measure the certainty of these decisions, and compare them pair to pair between unsupervised and supervised models.




12 models are selected, P1 to P23 are female speakers; the rest is male. Each sub-figure is an unsupervised vs. supervised posterior probability plot on the test set in the matching case. The percentage of matching is given.

The histograms of posterior probabilities provided by unsup and sup-vised *fricatives* models on test set in matching case. Two highest distributions at (1, 1) and (0, 0) are 678.7 and 440.3.



# Conclusion

- An unsupervised learning algorithm is defined as cognitive component analysis if the ensuing group structure is well-aligned with that resulting from human cognitive activity.
- Data analytical processing pipeline was built.
- 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).
- A detailed comparison scheme from classification error rates, sample-to-sample error, to posterior probability level, measures the matching degree of two learning methods.
- Two classifications agree on a majority of scenarios in several cognitive tasks related to speech perception, from low-level to high-level. The unsupervised learning algorithm and supervised learning proxy for a human cognitive activity did lead to comparable classifiers.
- Cognitive components do exist!

# Acknowledgements


- Professor Lars Kai Hansen
- Co-author Andreas Brinch Nielsen
- People who spent their precious time on helping me construct the music database VAPS.
- ISP Stuff
- Secretary Ulla Nørhave
- Professor Te-Won Lee, Lee-Lab members & Jiucang Hao
- The Danish Technical Research Council
- Otto Mønsteds Fond, Reinholdt W. Jorck og Hustrus Fond, Marie & M.B. Richters Fond, Oticon Fonden, and Niels Bohr Legatet for financial support