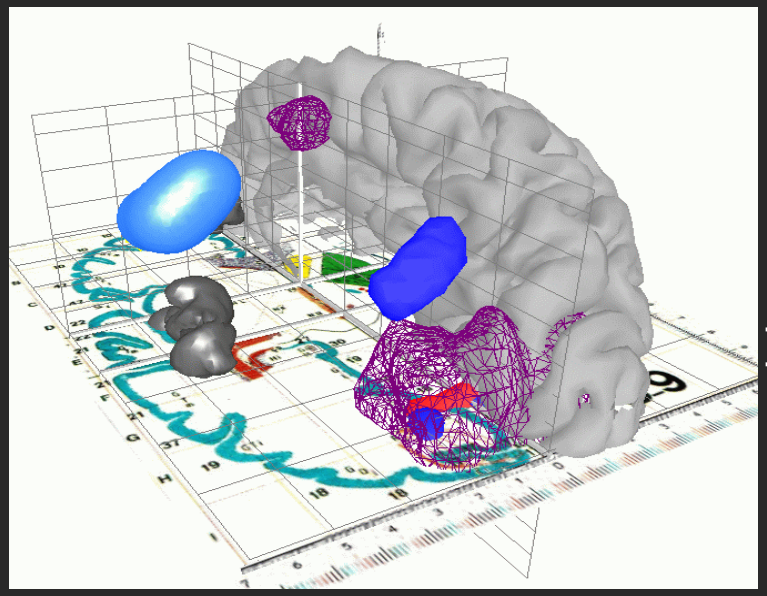
On Low-level Cognitive Components of Speech



Ling Feng

Intelligent Signal Processing
Informatics and Mathematical Modelling
Technical University of Denmark
www.imm.dtu.dk/~lf/

Outline

- **Introduction**
  - Cognitive Components Analysis: A definition
  - Machine Learning Tools
- **Cognitive components**

  Examples:
    - Text analysis
    - music genre
    - Speech (phoneme & speaker)
- **Summary and Outlook**

Cognitive Components Analysis

- What is Cognition?

  Cognition is the process involved in knowing, or the act of knowing, including perception and judgment. It includes every mental process that can be described as an experience of knowing as distinguished from an experience of feeling or of willing. [Encyclopædia Brittanica]

- 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.

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.

Cognitive Components Analysis

- COCA: an intermediate level tool
  - Source separation — low level
  - COCA — intermediate level
  - Content detection — high level

- Theoretical main points:

  The relation between supervised and unsupervised learning. Related to the discussion of the utility of unlabeled samples in supervised learning.

Machine Learning Tools

- Unsupervised Learning

  No separation of the training set into inputs and outputs pairs.

  `Self-organization´

  - LSA $\Sigma = U\Gamma U^{T}$ $Z = U_{L}^{T} X$
  - ICA $X_{j,t} = \sum_{k=1}^{K} A_{j,k} S_{k,t}$

- Supervised Learning

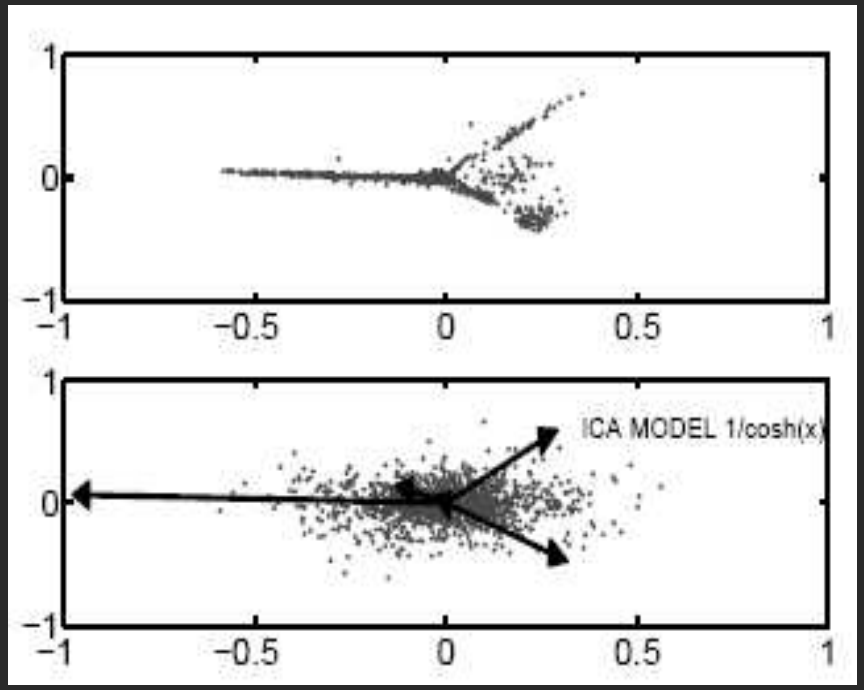
  The model includes mediating variables between inputs x and outputs y.

Text analysis

- Vector Space Representation
- LSA:

  a sparse linear mixture of independent topics in term-document scatter plots
- ICA:

  less than 10% classification error rate

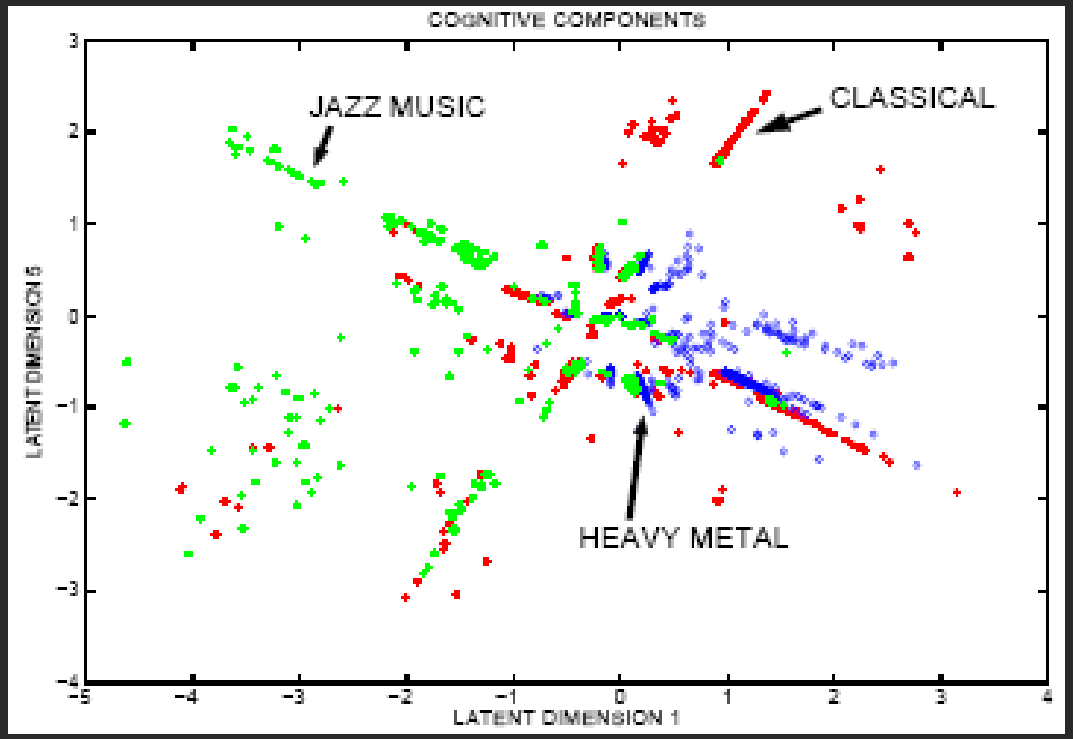


If the ”structure” in the feature space is well aligned with the label structure we expect high utility of unlabeled data.

DTU

Music genre

- Feature
  - 13d MFCC
  - frame =30ms
  - overlap = 10ms
- LSA
  - A sparse linear mixture of independent context from a music database.



If the ”structure” in the feature space is well aligned with the label structure we expect high utility of unlabeled data.

Speech

- Feature
  - Short-term feature: frame size [10ms, 40ms]
  - Long-term feature

  `**To reveal the semantic meaning of a signal, analysis over a much longer period is necessary, usually from one second to several tens seconds**´ [Wang, Liu, Huang, 2000].
- Energy Based Sparsification (EBS) – *ATTENTION!*
  - EBS: retain the upper ? % of normalized magnitudes
  - Attention: the process which gives rise to conscious awareness [Braisby, Gellatly, 2005].

`Attention appears to have surprising similarity with the development of invariant feature.´

Wang, Y., Liu, Z., Huang, J.-C., *Multimedia Content Analysis, IEEE Signal Processing Magazine,* Nov. 2000, 12-36 (2000).
Braisby, N., Gellatly, A., *Cognitive Psychology*, OXFORD University Press, 2005

DTU

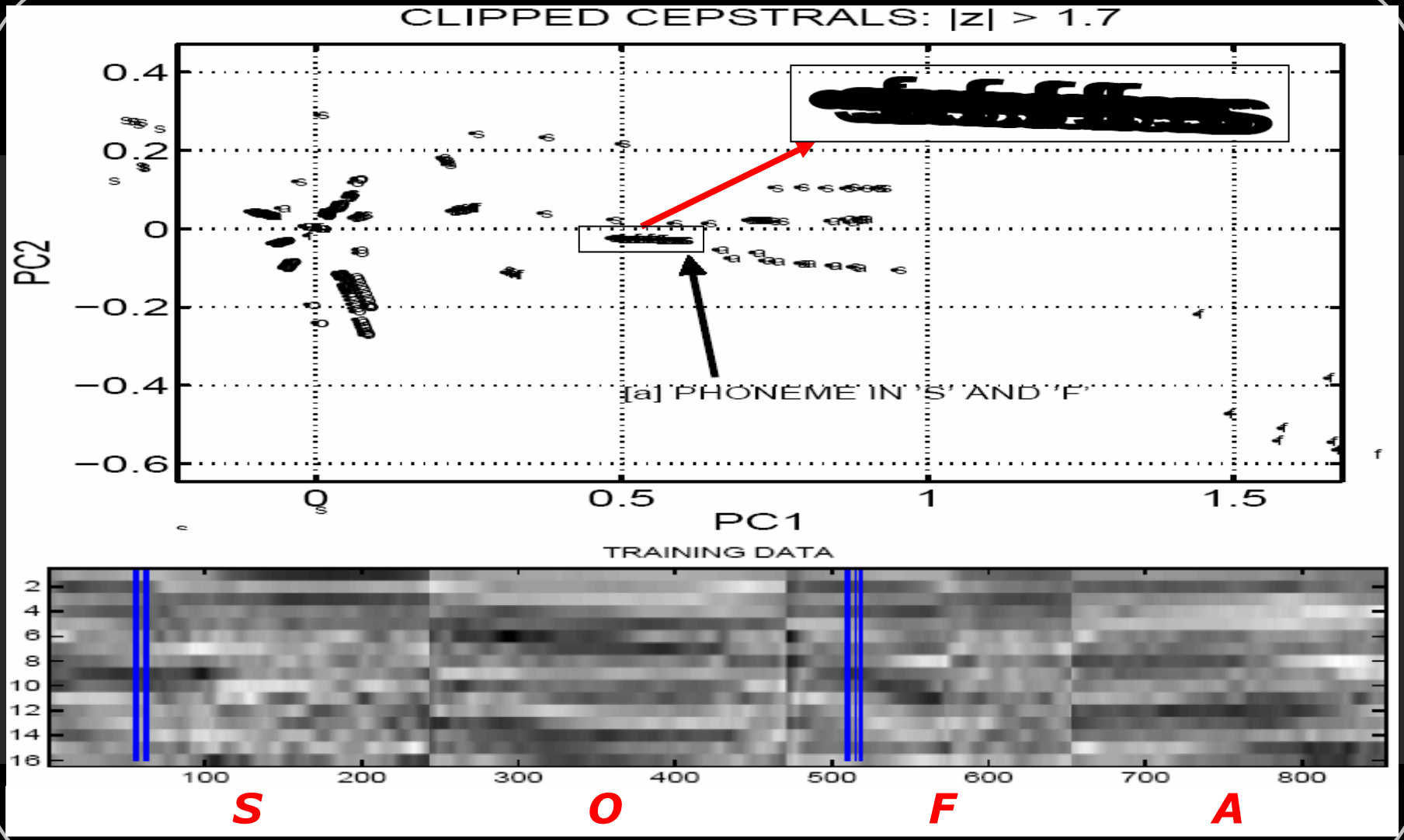
Speech - phoneme

- Phoneme

  The class of sounds that are consistently perceived as representing a certain minimal linguistic unit [Deller, Hansen, Proakis, 2000].
- COCA on Phoneme
  - Feature

    16d Cepstral Coefficients, frame=20ms, overlap=10ms
  - Energy based sparsification:

    retain upper 35% magnitude fractile
  - LSA (PCA) on sparsified features


CLIPPED CEPSTRALS: |z| > 1.7
PC2
PC1
[a] PHONEME IN 'S' AND 'F'
TRAINING DATA
S
O
F
A

Speech - speaker

- Features
  - Basic feature: 12d MFCC
  - Long-term feature: 1 sec
- Energy based sparsification
  retain upper 1% magnitude fractile
- Data source
  Three speakers, F1, F2, M1 from ELSDSR.
  - Text-dependent
  - Text-independent

Text-dependent Speaker Recognition

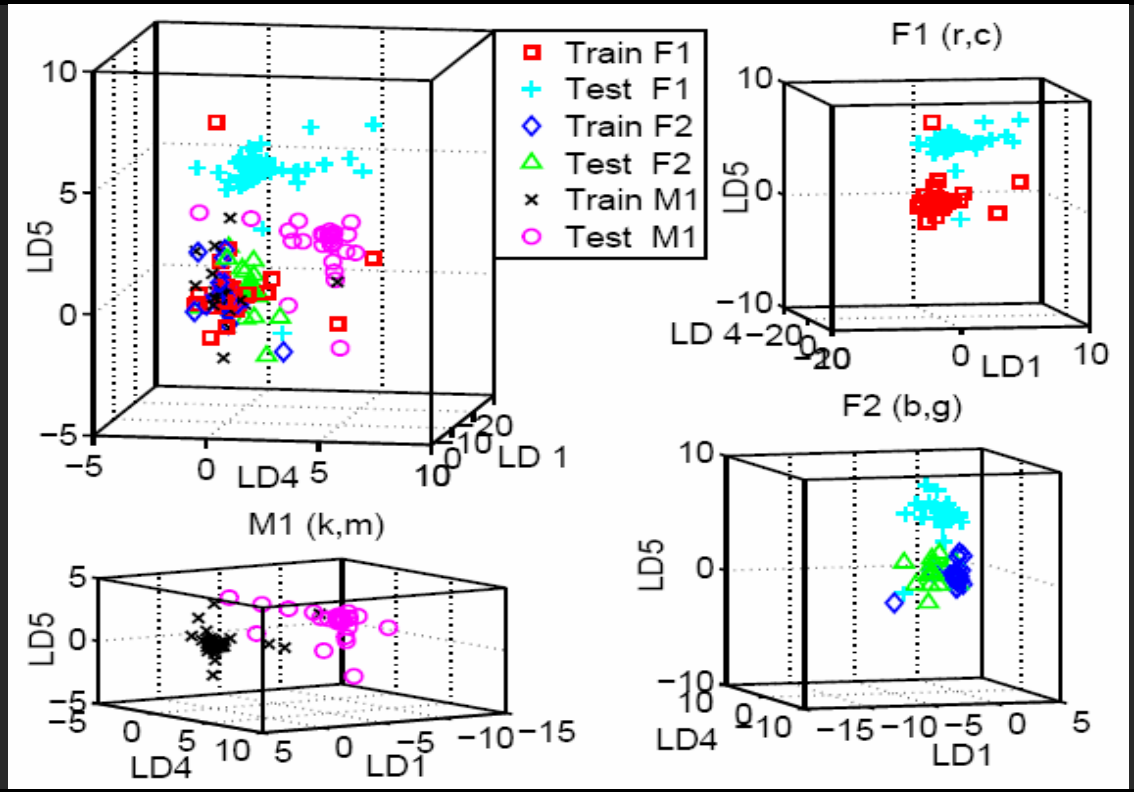
Training set: 52.5s

Test set: 35.5s

Phenomena:

- The phoneme like sparse linear structure;
- Offset between training and test sets;

An interaction between the text content and the speaker identity!



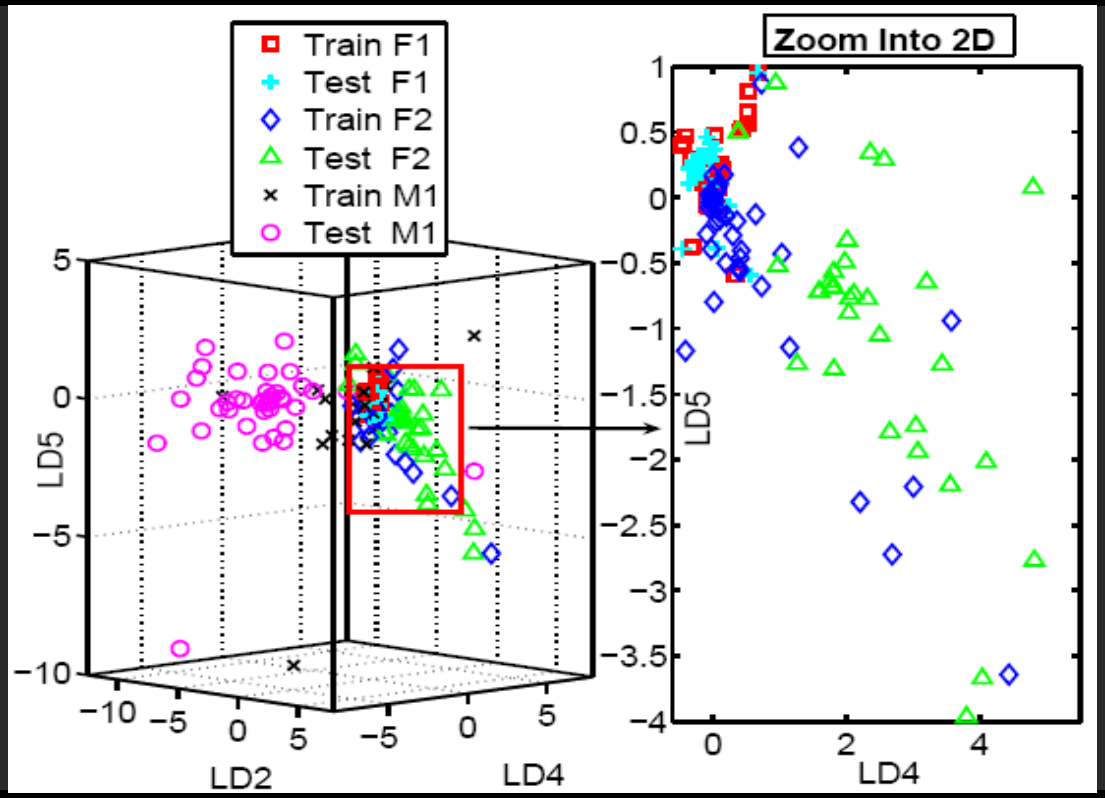
Text-independent Speaker Recognition

Training set: 32 s

Test set: 20 s

Phenomena:

The generalizable ray structures of independent identities emanating from origin of the coordinate system without offsets.



DTU

Summary and Outlook

- Summary
  - Cognitive components can be found by unsupervised learning!
    - generalizable features for phonemes with short-term features
    - generalizable speaker specific sparse components with long-time features
- Outlook

  Should the ray structures likely be based on *independence*?
  - (Labeled) mixture of Factor Analysis

$$p(x,y)=\sum_{k=1}^{K} p(x\mid k)\,p(y\mid k)\,p(k)$$