The Cognitive Components of Audio Spaces
Outline

Cognitive component analysis:
- Our definition
- Motivation and related ideas in the literature
- Audio signals: Phonemes as cognitive components
- Higher order cognition: Text, indexing media, social cognition

Conclusion and outlook
Cognitive Component Analysis

Cognitive component analysis (COCA)
- Hypothesis: Cognitive information processing is driven by statistical properties of the environment.
- The process of unsupervised grouping of data so that the resulting group structure is well-aligned with grouping based on human cognitive activity (Hansen et al., 2005).

“Rational models of cognition explain human behavior as approximating optimal solutions to the computational problems posed by the environment” (Griffiths et al., 2007)

Cognitive compatibility as "μ-Turing" test...
Ecological modeling approach

Important for engineering proxies for human information processing...
Cf. efficient coding of "context-to-action" mapping
Many generalizations are possible – which ones will make sense to a human?

Can we answer the key question: What is an object in cognition?
Cognitive component analysis and the notion of object

The object is a basic notion in cognitive psychology

- E.g., TVA estimates number of objects in short time memory
- A pragmatic definition of an object could be: An object is a signal source with independent behavior in a given environment
- Cognitive component analysis is a step towards an general purpose definition of an object
- Information theory: Optimality, do brains exploit the coding advantage?

Modeling issues: We are interested in the relation between supervised and unsupervised learning. Related to the discussion of the utility of unlabeled examples in supervised learning and swift learning.

Engineering issues: Can we predict the digital media components that a human will pay attention to? - a key challenge for cognitive systems.
Cognitive Information Processing

**Cognitive Component Analysis**


**Emotion in song lyrics**


**Top-down attention**

Cognitive modeling, mental models

Human cognition is often to act on weak signals, i.e., lack of information or poor signal to noise conditions.

Solve the problem by being very sensitive and post-process alarms with rich context models.

Mental models can be more or less well-aligned with actual physics/ecology, c.f. Friston et al.’s Predictive coding model

Subjective, sensory data

Qualitative data often mapped with MDS multidimensional scaling: low-dimensional, neighbor preserving Euclidean representation

Austen Clark in Sensory Qualities (1993):

“The number of dimensions of the MDS space corresponds to the number of independent ways in which stimuli in that modality can be sensed to resemble or differ, but the dimensions per se have no meaning”
Gärdenfors’ conceptual spaces

Cognitive models:
- Symbolic, associative/connectionist, geometrical

Human cognition ~ similarity judgments ~ Gestalt theory ~ geometrical proximity

How to identify conceptual spaces, i.e., geometrical representations? - Cognitive component analysis?

(Gärdenfors, 2000)
Kemp-Tenenbaum – Discovery of structural form (2008)

Human mind has access only to relatively low complexity modeling tools
Cognitive compatibility

Unsupervised Learning

Hidden variable

$p(s \mid x, w_u) \propto p(x \mid s, w_u)p(s \mid w_u)$

“Cognitive event”: Data, sound, image, behavior

Supervised learning

$p(y \mid x, w_s)$

“Cognitive” label, i.e. provided by a human observer

How well do these learned representations match: $s = y$?
When can COCA be expected to work?

If “statistical structure” in the relevant feature space is well aligned with the label structure we expect high cognitive compatibility.

Unsupervised-then-supervised learning can explain “learning from a single example”.

The Good, the Bad, and the Ugly...
How will COCA help computers understand media content?

Understand = simulate cognitive processing in humans

Help metadata estimation automatic tagging of digital media (sound/images/video/ deep web objects)

Basic signal processing tools are known (perceptual models...)

Amtrak
... Sort by: Schedule Fare. One city is required if a train number is used. Help. Departs: Arrives: Station List. Station List, Or. Train No.: (optional), Date. Time. ...

Official SUBWAY Restaurants’ Web Site
... Go To SUBWAY KIDS. ... Heart Walk. SUBWAY® restaurants is a national sponsor of the American Heart Association’s Heart Walks. ... The Subway Real Estate Corp. ...

www.sncf.com - A nous de vous faire préférer le train
... Réservée ou achetez, avec vos billets de train et tous vos voyages en France comme dans le reste du monde sur notre agence de voyages en ligne. ...
www.sncf.fr/ - 27k - Cached - Lignende side

Die Bahn - Startseite Reiseportal; Auskunft, Fahrkarten ...
Werbung, ...
www.bahn.de/ - 54k - 5 okt 2004 - Cached - Lignende side

National Rail Enquiries Online: train times and fare info for ...
National Rail Enquiries Logo. ... Limit changes to (if possible). Unlimited changes. ...
www.nationalrail.co.uk/plannmyjourney/ - 13k - 5 okt 2004 - Cached - Lignende side
Vector space representation

Abstract representation - can be used for all digital media
A “cognitive event” is represented as a point in a high-dimensional “feature space” - document similarity ~ spatial proximity in a given metric

Text: Term/keyword histogram, N-grams
Image: Color histogram, texture measures
Video: Object coordinates (tracking), active appearance models
Sound: Spectral coefficients, mel cepstral coefficients, gamma tone filters

Contexts can be identified by their feature associations ( = Latent semantics )

S. Deerwester et al. Indexing by latent semantic analysis.
The independent context hypothesis:
The perpetual cocktail party

Challenge: Presence of multiple agents/contexts
Need to "blindly" separate source signals = learn contexts
Independent Component Analysis come to rescue!

\[ x(feature, time) = \sum_k A(feature, k) s(k, time) \]
Linear mixing generative model ICA - “Synthesis”
simplistic model incorporating sparsity and independence

Space-time matrix
Component’s “where”
Vector of “what”

Normal sources

\[ x(\text{loc, time}) = \sum_k A(\text{loc, } k) \ s(k, \text{time}) \]

Dense sources

Sparse sources
Protocol for comparing supervised and unsupervised learning

Use an “unsupervised-then-supervised” classifier:
- Train the unsupervised scheme, eg., ICA
- Freeze the ICA representation (A matrix)
- Train a simple (e.g. Naïve Bayes) classifier using the features obtained in unsupervised learning

Compare with supervised classifier == human proxy
- Error rates of the two systems
- Compare posterior probabilities

Research question: Can statistics of independence account for human object detection/uncertainty?
Phoneme classification

Nasal vs oral: “Esprit project ROARS” (Alinat et al., 1993)

Binary classification

Error rates: **0.23** (sup.), **0.22** (unsup.)

Bitrates: **0.48** (sup.), **0.39** (unsup.)
Cognitive components of speech

Basic representation: Mel weighted cepstral coefficients (MFCCs)
Modeling at different time scales 20 msec – 1000 msec

Phonemes
Gender
Speaker identity

Co-worker: Ling Feng

Figure 3: The latent space is formed by the two first principal components of data consisting of four separate utterances representing the sounds ‘s’, ‘o’, ‘f’, ‘a’. The structure clearly shows the sparse component mixture, with ‘rays’ emanating from the origin (0,0). The ray embraced in a rectangle contains a mixture of ‘s’ and ‘f’ features, a cognitive component associated with the vowel /s/ sound.
Mel weighted cepstral coeff. (MFCC)
Error rate comparison
For the given time scales and thresholds, data locate around $y = x$, and the correlation coefficient $\rho = 0.67, p < 1e^{-09}$.

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. Match between supervised and unsupervised learning = 91%.
Longer time scales

Time integrated (1000ms) MFCC’s: text independent speaker recognition....

Feng & Hansen (CIMCA, 2005)
Error rate correlations for super/unsupervised learning for different cognitive time scales and events

Challenged by degree of sparsity and time averaging

**Fig. 4.** Figure shows test error rates of both supervised and unsupervised learning on four topics: phonemes, gender, height and identity. Solid lines indicate \( y = x \) in the coordinate systems. All data located along this line, meaning high correlation between supervised and unsupervised learning.
“Higher” cognitive representations:
Digital media vector space representation

Abstract representation - can be used for all digital media

Document is represented as a point in a high-dimensional "feature space"
  document similarity ~ spatial proximity in a given metric

Text: Term/keyword histogram, N-grams
Image: Color histogram, texture measures
Video: Object coordinates (tracking), active appearance models
Sound: Spectral coefficients, cepstral coefficients, gamma tone filters
Latent semantics

Document features are correlated, the pattern of correlation reflects "associations".
Associations are context specific
Word sets are activated in concert in a given context
ape ~ zoo, zoo ~ elephant => ape ~ elephant

Latent semantic analysis: Contexts can be identified by term co-variance patterns (PCA)

Factor models for what / where

Represent a datamatrix by a low-dimensional approximation

\[ X \approx \sum_{k=1}^{K} A_k S_k \]

\( X \) represents the original data matrix, \( A \) and \( S \) are matrices of factors, and \( K \) is the number of factors.
Generative model for hidden variables

\[ \mathbf{x} = \mathbf{As} + \mathbf{e} \sim N(\mathbf{0}, \Sigma) \]

\[
p(\mathbf{x} | \mathbf{A}, \mathbf{s}, \Sigma) = \int p(\mathbf{x} | \mathbf{A}, \mathbf{s}, \Sigma) p(\mathbf{A}, \mathbf{s}) d\mathbf{A} d\mathbf{s}
\]

\[
p(\mathbf{x} | \mathbf{A}, \mathbf{s}, \Sigma) = 2\pi \Sigma^{-1/2} e^{-\frac{1}{2}(\mathbf{x} - \mathbf{As})^T \Sigma^{-1} (\mathbf{x} - \mathbf{As})}
\]

S known: GLM
(1-A^-1) sparse: SEM
S,A positive: NMF

Source distribution:
PCA: … normal
ICA: … other

PCA: \( \Sigma = \sigma^2 \cdot \mathbf{1} \)
FA: \( \Sigma = \mathbf{D} \)

Højen-Sørensen, Winther, Hansen, Neural Comp (2002), Neurocomputing (2002)
Figure 1 Non-negative matrix factorization (NMF) learns a parts-based representation of faces, whereas vector quantization (VQ) and principal components analysis (PCA) learn holistic representations. The three learning methods were applied to a database of $m = 2,429$ facial images, each consisting of $n = 19 \times 19$ pixels, and constituting an $n \times m$ matrix $V$. All three find approximate factorizations of the form $V = WH$, but with three different types of constraints on $W$ and $H$, as described more fully in the main text and methods. As shown in the $7 \times 7$ montages, each method has learned a set of $r = 49$ basis images. Positive values are illustrated with black pixels and negative values with red pixels. A particular instance of a face, shown at top right, is approximately represented by a linear superposition of basis images. The coefficients of the linear superposition are shown next to each montage, in a $7 \times 7$ grid, and the resulting superpositions are shown on the other side of the equality sign. Unlike VQ and PCA, NMF learns to represent faces with a set of basis images resembling parts of faces.

---

Learning the parts of objects by non-negative matrix factorization

Daniel D. Lee* & H. Sebastian Seung†

* Bell Laboratories, Lucent Technologies, Murray Hill, New Jersey 07974, USA
† Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

NATURE | VOL 401 | 21 OCTOBER 1999 | www.nature.com
Modeling the generalizability of factorization

Rich physics literature on “retarded” learning

Universality

- Generalization for a “single symmetry breaking direction” is a function of ratio of N/D and signal to noise S
- For subspace models -- a bit more complicated -- depends on the component SNR’s and eigenvalue separation
- For a single direction, the mean squared overlap \( R^2 = \langle (u_1^T u_0)^2 \rangle \) is computed for N,D \( \rightarrow \infty \)

\[
R^2 = \begin{cases} 
  (\alpha S^2 - 1) / S(1 + \alpha S) & \alpha > 1 / S^2 \\
  0 & \alpha \leq 1 / S^2 
\end{cases}
\]

\[
\alpha = N / D \quad S = 1 / \sigma^2 \quad N_c = D / S^2
\]


\( N_c = (0.0001, 0.2, 2, 9, 27, 64, 128, 234, 400, 625) \)
\( \sigma = (0.01, 0.06, 0.12, 0.17, 0.23, 0.28, 0.34, 0.39, 0.45, 0.5) \)
Linear mixture of independent agents in term-document scatterplots

Linear mixture of independent contexts observed in short time features (mel-cepstrum) in a music database.
Complex (social) networks:
Linear mixtures of independent “interest”? 

Genre patterns in expert opinion (tags) on 400 musical artists


"Movie actor network"
- A collaborative small world network

128.000 movies
380.000 actors
Independent contexts in document databases

- $x(j,t)$ is the occurrence of the $j$'th word in the $t$'th document.
- $s(k,t)$ quantifies how much the $k$'th context is expressed in $t$'th document.
- $A(j,k)$ quantifies the typical importance of the $j$'th word in the $k$'th context.

ICA in text
Isbell and Viola (1999)
PCA vs ICA document scatterplots
Independent contexts in dynamic text: Chat room analysis

We logged a day's chat in a CNN "news cafe".

The database involves 120 users chatting during an 8 hour period.
ICA by dynamic decorrelation

The Bayes factor - \( P(M|D) \) - of each model is estimated in the BIC approximation

Source autocorrelations
Independent contexts in multi-media

- Organizing webpages in categories
- Labels obtained from Yahoo’s directory
- Features: Text, color, and texture subsets of MPEG image features

Performance of the system trained by associating unsupervised independent components with labels – generalization based on Yahoo categories

<table>
<thead>
<tr>
<th>Modality</th>
<th>Classification Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Color</td>
<td>23.0%</td>
</tr>
<tr>
<td>Texture</td>
<td>18.0%</td>
</tr>
<tr>
<td>Texture/Color</td>
<td>11.5%</td>
</tr>
<tr>
<td>Text</td>
<td>5.7%</td>
</tr>
<tr>
<td>Combined (texture/color/text)</td>
<td>2.8%</td>
</tr>
</tbody>
</table>

Fig. 3. Scatterplots of the text and image multimedia data, projected to a two-dimensional subspace found by PCA. Grey value of points corresponds to the three classes considered, see Fig. 4. The ray like structure strongly suggest an ICA interpretation, however, the relevance of this representation can only be determined by a subsequent inspection of the recovered source signals. As we will see in section 4.6, it turns out that there is an interesting alignment of the source signals and a manual labeling of the multimedia documents.
### Texture (K=13)

<table>
<thead>
<tr>
<th></th>
<th>69.75</th>
<th>7.75</th>
<th>6.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.5</td>
<td>88.5</td>
<td>5.75</td>
<td></td>
</tr>
<tr>
<td>18.75</td>
<td>3.75</td>
<td>87.75</td>
<td></td>
</tr>
</tbody>
</table>

### Color (K=16)

<table>
<thead>
<tr>
<th></th>
<th>70.75</th>
<th>3.75</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>81.5</td>
<td>11.25</td>
<td></td>
</tr>
<tr>
<td>17.25</td>
<td>14.75</td>
<td>78.75</td>
<td></td>
</tr>
</tbody>
</table>

### Text (K=45)

<table>
<thead>
<tr>
<th></th>
<th>93</th>
<th>2</th>
<th>2.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>94.75</td>
<td>2.5</td>
<td></td>
</tr>
<tr>
<td>6.5</td>
<td>3.25</td>
<td>95.25</td>
<td></td>
</tr>
</tbody>
</table>

### Texture Color (K=26)

<table>
<thead>
<tr>
<th></th>
<th>82</th>
<th>1.75</th>
<th>4.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>93.75</td>
<td>5.75</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>4.5</td>
<td>89.75</td>
<td></td>
</tr>
</tbody>
</table>

### Combined error rate: 2.8%

### Single best error rate: 5.7%
CASTSEARCH - CONTEXT BASED SPEECH DOCUMENT RETRIEVAL

Lasse Lohilahti Mølgaard, Kasper Winther Jørgensen, and Lars Kai Hansen

Informatics and Mathematical Modelling
Technical University of Denmark Richard Petersens Plads
Building 321, DK-2800 Kongens Lyngby, Denmark

Fig. 1. The system setup. The audio stream is first processed using audio segmentation. Segments are then using an automatic speech recognition (ASR) system to produce text segments. The text is then processed using a vector representation of text and apply non-negative matrix factorization (NMF) to find a topic space.

Mølgaard et al. 2007

Fig. 3. Figure 3(a) shows the manual segmentation of the news show into 7 classes. Figure 3(b) shows the distribution $p(k|d^*)$ used to do the actual segmentation shown in figure 3(a). The NMF-segmentation is in general consistent with the manual segmentation. Though, the segment that is manually segmented as 'crime' is labeled 'other' by the NMF-segmentation.
... california governor arnold’s fortson agar inspected the california mexico border by helicopter wednesday to see ...

... the past days president bush asking california’s governor for fifteen hundred more national guard troops to help patrol the mexican border but governor orville schwartz wicker denying the request saying...

Fig. 2. Two examples of the retrieved text for a query on 'schwarzenegger'.
Conclusions & outlook

Evidence that phonemes, gender, identity are independent components ‘objects’ in the (time stacked) MFCC representation.

Evidence that human categorization is based on sparse independent components in social networks, text, digital media.

Conjecture: Objects in digital media can be identified as independent components: The brain uses old tricks from perception to solve complex “modern” problems.
Acknowledgments

- Danish Research Councils
- EU Commission
- NIH Human Brain Project grant (P20 MH57180)

For software and demos:
DTU:ICA toolbox (www.imm.dtu.dk/cisp)
Additional references


