From ASA to CASA
what does the "C" stand for anyway?

Mathieu Lagrange

ircam
Centre
Pompidou

September 22, 2011

Schedule

- 14h00 -14h30 Introduction
  - Mathieu Lagrange: From ASA to CASA...
- 14h30 - 15h30 Auditory Scene Analysis
  - Trevor Agus: Perceptual learning of novel sounds
  - Josh McDermott: Sound texture perception via statistics of the auditory periphery
- 15h30 - 16h Coffee break (Level -2, orange floor)
- 16h30 - 17h30 Machine Listening
  - Jon Barker: Probabilistic frameworks for Scene understanding
  - Boris Defreville: Machine listening in everyday life
- 17h00 -18h30 Panel

Welcome to the 3rd (and last) DAFX'11 Satellite Workshop

Acknowledgments

CATÓLICA
UNIVERSIDADE CATÓLICA PORTUGUESA | PORTO
Escola das Artes

apoio

People

Organisers

- Luis Gustavo Martins (Portuguese Catholic University, Porto)
- Mathias Rossignol (Ircam)
- Laure Cornu (Ircam)
- Mathieu Lagrange (Ircam)

Invited speakers

- Trevor Agus (ENS)
- Josh McDermott (NYU)
- Jon Barker (Sheffield University)
- Boris Defreville (Orelia)

People

Organisers

- Luis Gustavo Martins (Portuguese Catholic University, Porto)
- Mathias Rossignol (Ircam)
- Laure Cornu (Ircam)
- Mathieu Lagrange (Ircam)

Invited speakers

- Trevor Agus (ENS)
- Josh McDermott (NYU)
- Jon Barker (Sheffield University)
- Boris Defreville (Orelia)

The digital era

A lot of things can be described as a series of 0 and 1.

Important issues

- capture: precise bit
- transmission: efficient bit
- search: relevant bit

Means

- mechanics, biology
- psycho-acoustics
- cognition

The digital era

A lot of things can be described as a series of 0 and 1.

Important issues

- capture: precise bit
- transmission: efficient bit
- search: relevant bit

Means

- mechanics, biology
- psycho-acoustics
- cognition

The digital era

A lot of things can be described as a series of 0 and 1.

Important issues

- capture: precise bit
- transmission: efficient bit
- search: relevant bit

Means

- mechanics, biology
- psycho-acoustics
- cognition

Today's key challenge

What next ?

The main issue here is how to define the notion of relevance.

The target is a human

It is meaningful to understand / replicate how humans perceive and builds their representation of their environment.

Today's key challenge

What next ?

The main issue here is how to define the notion of relevance.

The target is a human

It is meaningful to understand / replicate how humans perceive and builds their representation of their environment.

Opinions

1. The engineer's point of view: no need to understand how a bird moves its wings to send a rocket in the air
2. The biologist's point of view: the effectiveness of a computational implementation does not prove in any way that it effectively replicates a biological behavior
3. Profitable cross-fertilization between
   - perception, cognition
   - engineering, computer science

Opinions

1. The engineer's point of view: no need to understand how a bird moves its wings to send a rocket in the air
2. The biologist's point of view: the effectiveness of a computational implementation does not prove in any way that it effectively replicates a biological behavior
3. Profitable cross-fertilization between
   - perception, cognition
   - engineering, computer science

Opinions

1. The engineer's point of view: no need to understand how a bird moves its wings to send a rocket in the air
2. The biologist's point of view: the effectiveness of a computational implementation does not prove in any way that it effectively replicates a biological behavior
3. Profitable cross-fertilization between
   - perception, cognition
   - engineering, computer science

ASA?

ASA

stands for Auditory Scene Analysis (Bregman [Mit 94])

- the scene can be described as an organized set of atoms
- there exist a set of grouping or segregation rules of those atoms into perceptually meaningful objects

ASA is a Gestaltist theory

Those rules are derived from evidence found in psycho acoustical experiments.

ASA?

ASA

stands for Auditory Scene Analysis (Bregman [Mit 94])

- the scene can be described as an organized set of atoms
- there exist a set of grouping or segregation rules of those atoms into perceptually meaningful objects

ASA is a Gestaltist theory

Those rules are derived from evidence found in psycho acoustical experiments.

CASA?

CASA

stands for Computational Auditory Scene Analysis

Early Ages

- David Mellinger: Event formation and separation in musical sound, [Stanford Phd 91]
- Dan Ellis: Prediction-driven computational auditory scene analysis, [MIT Phd 96]

ASA and CASA as of Google Scholar

ASA

3894 Bregman: Auditory scene analysis: The perceptual organization of sound [Mit 94]

CASA

318 Brown: Computational auditory scene analysis [Csl 94]

325 Ellis: Prediction-driven computational auditory scene analysis [PhD 96]

268 Roweis: One microphone source separation [Nips 01]

262 Wang: Computational auditory scene analysis: Principles, algorithms, and applications [Lavoisier 06]

132 Peltonen & al: Computational auditory scene recognition [Icassp 02]

ircam Centre Pompidou

Reasonable assumptions

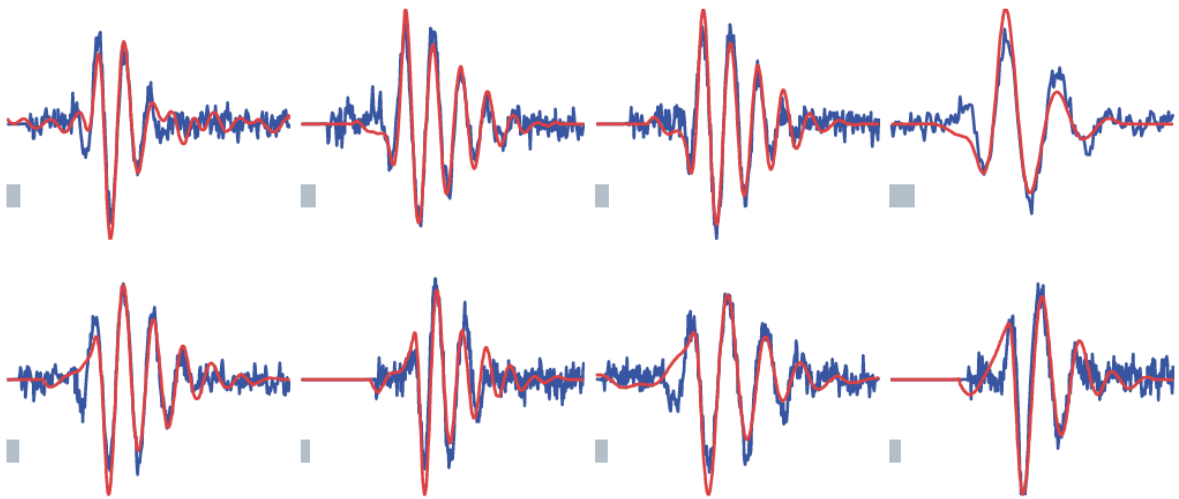
Decomposition of the auditory system

1. Low level: efficient encoding via linear transformation
2. High level: learning of abstractions with high generalization capability via non linear transformation (the quest for invariance)
3. Scene understanding: the Gestalt laws of grouping

First Level

Auditory encoding

Smith and Lewicki [Nature 05] demonstrate that signal bases learned with statistical tools over environmental and speech sounds match auditory nerves filter responses $\Leftarrow$ validation of statistical learning



Red: basis functions learnt with ICA.

Blue: auditory nerve response.

First Level

Auditory encoding

Smith and Lewicki [Nature 05] demonstrate that signal bases learned with statistical tools over environmental and speech sounds match auditory nerves filter responses $\Leftarrow$ validation of statistical learning

Question: What shall the dictionary be ?

- Flat or hierarchical ?
- Fixed or adaptive ?
- Learnt explicitly or implicitly ? (supervised / unsupervised)

First Level

Auditory encoding

Smith and Lewicki [Nature 05] demonstrate that signal bases learned with statistical tools over environmental and speech sounds match auditory nerves filter responses $\Leftarrow$ validation of statistical learning

Question: What shall the dictionary be ?

- Flat or hierarchical ?
- Fixed or adaptive ?
- Learnt explicitly or implicitly ? (supervised / unsupervised)

First Level

Auditory encoding

Smith and Lewicki [Nature 05] demonstrate that signal bases learned with statistical tools over environmental and speech sounds match auditory nerves filter responses $\Leftarrow$ validation of statistical learning

Question: What shall the dictionary be ?

- Flat or hierarchical ?
- Fixed or adaptive ?
- Learnt explicitly or implicitly ? (supervised / unsupervised)

Implicit learning

Evidence from biology

- childhood: a 3 months old baby has expectations regarding its specific cultural background that have been built implicitly (Tillman)
- adults: able to efficiently build persistent features to better recognize acoustic events, in an implicit manner (Agus [Neuron 10]) $\Leftarrow$ presentation of Trevor Agus

Implicit learning

Evidence from engineering

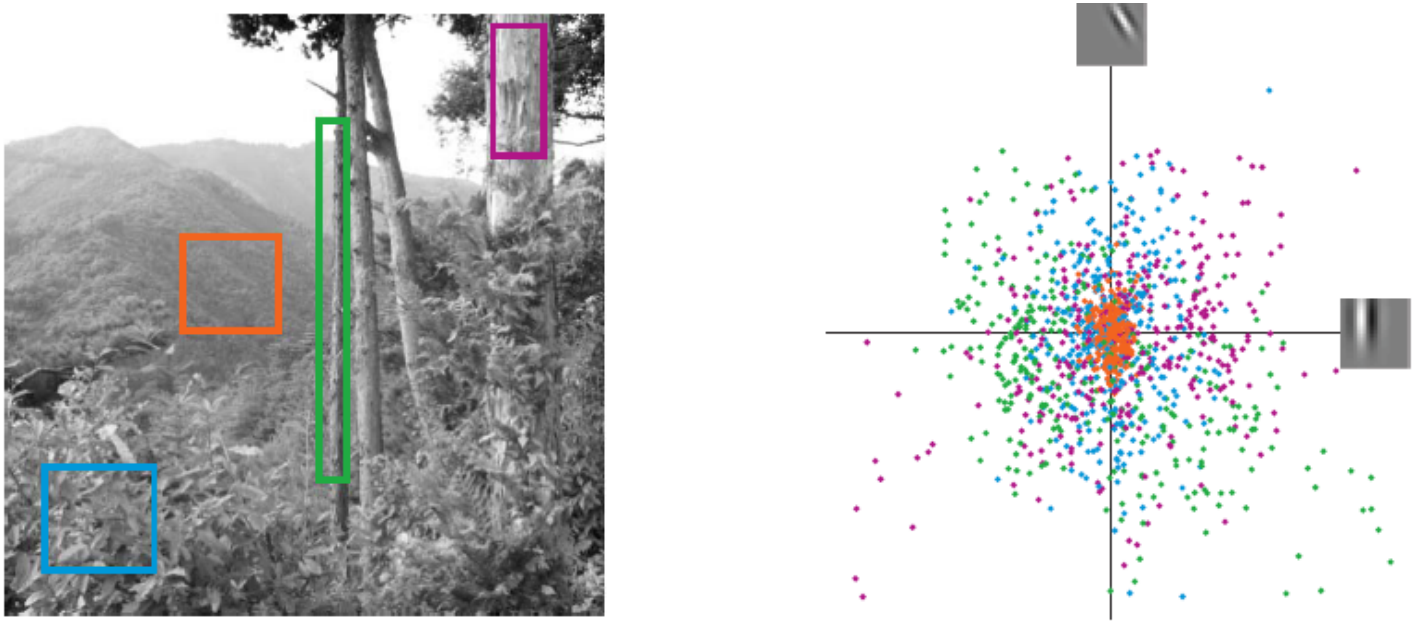
Deep belief Networks (Hinton [Science 06]) are able to build powerful features

- in an unsupervised way
- over raw data (noisy, high dimensional)
- provided that
  - enough data is available
  - the geometry of the network fits the structure of the data

Second Level

Abstraction

Karklin and Lewicki [Nature 08] propose that invariance is obtained via the encoding of statistical variations.

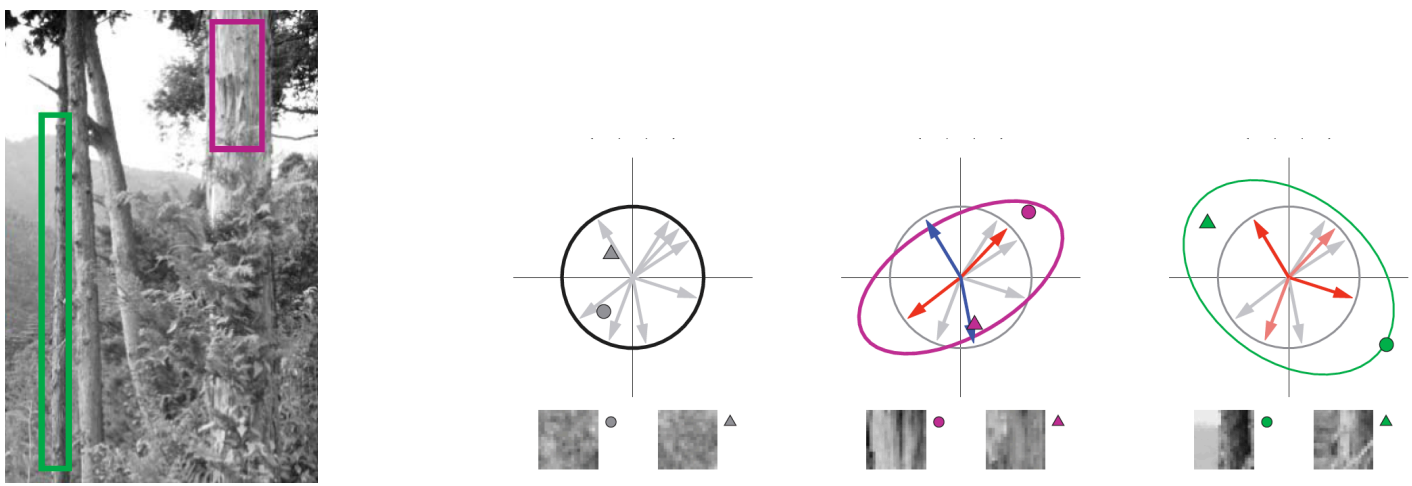


Feature space.

Second Level

Abstraction

Karklin and Lewicki [Nature 08] propose that invariance is obtained via the encoding of statistical variations.

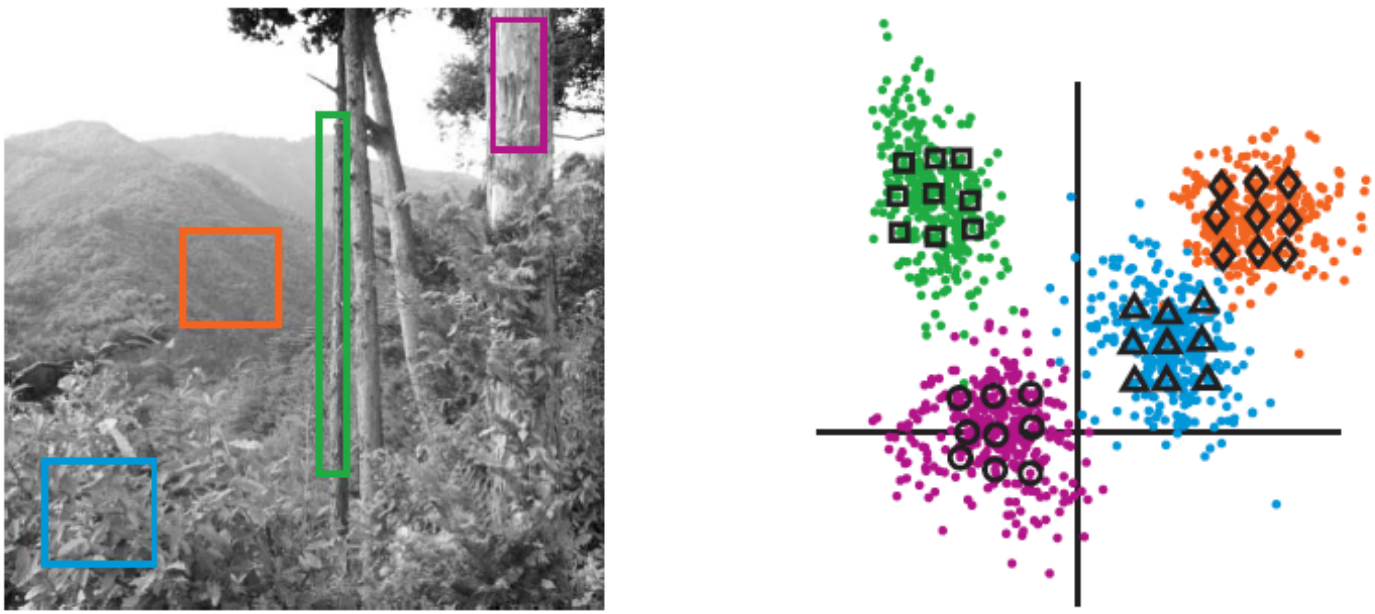


Distribution Coding Model.

Second Level

Abstraction

Karklin and Lewicki [Nature 08] propose that invariance is obtained via the encoding of statistical variations.



Distribution Coding Model.

Second Level

Abstraction

Karklin and Lewicki [Nature 08] propose that invariance is obtained via the encoding of statistical variations.

To model the auditory system

- What kind of statistics are important ? $\Leftarrow$ presentation of Josh McDermott
- What kind of frameworks are powerful ? $\Leftarrow$ presentation of Jon Barker

Third Level: Scene understanding

Scene understanding

is usually approached as a segmentation problem, each relevant object being formed using the Gestalt laws of grouping.
Those laws can be:

- static ones: proximity, similarity
- dynamic ones: closure, good continuation, common fate

What are the most successful computational frameworks:

- for expressing and enforcing such constraints ? ⇐ presentation of Jon Barker
- for dealing with scene understanding as implemented in everyday life use cases taken from the industry ? ⇐ presentation of Boris Defreville

ircam Centre Pompidou

Third Level: Scene understanding

Scene understanding

is usually approached as a segmentation problem, each relevant object being formed using the Gestalt laws of grouping.
Those laws can be:

- static ones: proximity, similarity
- dynamic ones: closure, good continuation, common fate

What are the most successful computational frameworks:

- for expressing and enforcing such constraints ? $\Leftarrow$ presentation of Jon Barker
- for dealing with scene understanding as implemented in everyday life use cases taken from the industry ? $\Leftarrow$ presentation of Boris Defreville

ircam Centre Pompidou

Third Level: Scene understanding

Scene understanding

is usually approached as a segmentation problem, each relevant object being formed using the Gestalt laws of grouping.
Those laws can be:

- static ones: proximity, similarity
- dynamic ones: closure, good continuation, common fate

What are the most successful computational frameworks:

- for expressing and enforcing such constraints ? $\Leftarrow$ presentation of Jon Barker
- for dealing with scene understanding as implemented in everyday life use cases taken from the industry ? $\Leftarrow$ presentation of Boris Defreville

ircam Centre Pompidou

Questions

From ASA to CASA: only insights ?

- Is the knowledge transfer from ASA to CASA only qualitative ?
- Are there other approaches in scientific fields such as biology, cognition, etc. that are also potentially meaningful for building powerful computational systems ?

What is CASA ?

- Is CASA a goal in itself ?
- Can it be decomposed into well defined tasks ?

Is CASA worth pursuing ?

- What are the major locks in contemporary CASA ?
- How does it relates to other sound processing areas such as Blind Source Separation (BSS) or Music Information Retrieval (MIR) ?

ircam Centre Pompidou

Thank you for attending !!

Next...

Stay tuned

- videos of the workshp soon available at: http://anasynth.ircam.fr/home/blogs/lagrange/casa-workshop-dafx
- possibly a continuation of the discussion at ISMIR'12 in Porto, Portugal