Extracting Meaning from Sound Signals
*a machine learning approach*

Jan Larsen, Associate Professor PhD
Cognitive Systems Section
Dept. of Informatics and Mathematical Modelling
Technical University of Denmark
jl@imm.dtu.dk, www.imm.dtu.dk/~jl
DTU, Lyngby Campus

**Education**
- 6,270 BSc, MSc and BEng students, including
- **654 international MSc students**
  - 759 PhD fellows (3 years)
  - 560 Exchange students (3–6 months)
  - 162 DTU students abroad
  - 419 Paying students in open education and part-time education

**Research**
- 3,144 Research publications
- 157 PhD dissertations

**Innovation**
- 67 Inventions reported
- 39 Patent applications submitted

**Employees (head counts)**
- 1,447 Faculty and research staff
- 1,583 Technical and administrative personnel in departments
- 1,491 Administration, campus service and PhD fellows

**Finances**
- Income (2008): €469.2 million
Section for Cognitive Systems

Why do we do it? VISION

What do we do? MISSION

machine learning
media technology
cognitive science

- 5 faculty
- 1 adj. prof.
- 3 postdocs
- 4 admin
- 17 Ph.D. students
- 10 M.Sc. students
Vision

Cognition refers to the representations and processes involved in thinking and decision making. Cognitive systems integrate information processing in brains and computers for collaborative problem solving.

Our vision is to design and implement profound cognitive systems for augmented human cognition in real-life environments.

Our research is driven both by curiosity and by an engineering desire to do good: To better understand human behaviors and to create engineering solutions with a positive impact on human well-being and productivity.

We will contribute to DTU's vision of excellence and strive to be a highly valued partner for our national and international networks.
Legacy of cognitive systems

Allan Touring
Theory of computing
1940’s

Norbert Wiener
Cybernetics
1948

machine learning

processing
adaption
understanding
cognition

information and data
media technology
cognitive science

people
Mission

To measure, model, and augment cognition from neuron to internet scale systems

A cognitive system should optimize itself according to:

The statistical model of the domain, the psycho-physical model of the users, the social context, and the computational resources in time and space
Interplay and Synergy

- Research Competences
- Innovation
- Society Challenges
- Education
Society challenges

Future improvement in productivity and quality of life requires **organization and integration of internet-size data sets**

**Digital media modeling** enables ubiquitous access to actionable information for personal development and organization of interpersonal relations

**Brain modeling and mental decoding** are crucial for augmented cognition, lifelong learning, and may revolutionize health services
extraction of meaningful and actionable information from audio by ubiquitous learning from data
Research Competences

Media technology: mobile platforms, digital media, social networks, search, navigation, and semantics

Machine learning: statistical modeling, signal processing, and complex networks

Cognitive science: perception, cognition, psycho-physics, and human computer interfacing
Statistical machine **learning abstracts data to active knowledge by identifying predictive relations** and has become a major driver of the knowledge society. Machine learning drives the Google economy, empowers bioinformatics, and enables mind reading in neuroimaging.

Our research in machine learning is rooted in statistics, including Bayesian and in resampling based methods, and has a strong algorithmic component. Past developments include ensembles, approximate inference, blind signal separation, and multi-way methods.

Current theoretical work concerns sparse representations, infinite models, multiway methods, and complex networks.

[https://ml.imm.dtu.dk/](https://ml.imm.dtu.dk/)
Data modeling framework

Evaluation, interpretation and visualization
Performance, robustness, complexity, interpretation and visualization, HCI

Data preparation
- quantity
- modality
- stationarity
- quality
- structure

Features extraction
- representation
- selection
- construction
- integration

Modeling
- structure
- type
- learning
- selection and integration

Result
Decision
Dissemination

Domain knowledge
**Unsupervised learning**

- Probabilistic modeling of structure in multivariate data
- Preprocessing, data reduction, outlier detection

- Clustering
- Linear factor models (ICA, NMF)
- Kernel method
Supervised learning

- Mapping between domains – from features to decision
- Based on a data set of simultaneous observations of X and Y

- Neural networks
- Kernel machines
- Bayesian learning
Semi-supervised learning

- Learning from labeled and unlabeled data
- Optimal use of inexpensive unlabeled data
- Quantification of robustness

Active learning

- Active learning - related method in which samples are initially unknown
  - Labelling may be expensive or laborsome
  - Methods should decide which samples help learning most
Huge demand for tools: organization, search, information enrichment

- Recommender systems ("taste prediction")
- Playlist generation
- Finding similarity in music (e.g., genre classification, instrument classification, etc.)
- Meta data generation (emotional tags, labels)
- Newscast transcription/search
- Music transcription/search
- Audio separation
Intelligent Sound Project

- FTP project 2005-2009
- 14 mil DKK
- Participants: DTU and Aalborg University

www.intelligentsound.org
Machine learning in sound information processing

- audio data
- Meta data: ID3 tags, context
- machine learning model
- Tasks: Grouping, Classification, Mapping to a structure, Prediction (e.g. answer to query)
- user networks, co-play data, playlist, communities, user groups
- Audioscrobbler: The Social Music Technology Playground

ISO/IEC JTC1/SC29/WG11
Moving Picture Experts Group

MusicBrainz

DTU Informatics, Technical University of Denmark
Specialized search and music organization

The NGSW is creating an online fully-searchable digital library of spoken word collections spanning the 20th century.

Organize songs according to tempo, genre, mood.

Explore by genre, mood, theme, country, instrument.

Using social network analysis.

Query by humming.

search for related songs using the “400 genes of music”.

FindSounds

lost fm the social music revolution

allmusic

The National Gallery of the Spoken Word

PANDORA
<table>
<thead>
<tr>
<th>Genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alternative Rock</td>
</tr>
<tr>
<td>Blues</td>
</tr>
<tr>
<td>Classical</td>
</tr>
<tr>
<td>Country</td>
</tr>
<tr>
<td>Dance</td>
</tr>
<tr>
<td>Folk</td>
</tr>
<tr>
<td>Jazz</td>
</tr>
<tr>
<td>Opera &amp; Vocal</td>
</tr>
<tr>
<td>Pop</td>
</tr>
<tr>
<td>R&amp;B</td>
</tr>
<tr>
<td>Rap &amp; Hip-Hop</td>
</tr>
<tr>
<td>Rock</td>
</tr>
</tbody>
</table>
Meta data generation: genre classification

- Prototypical example of predicting meta and high-level data
- The problem of interpretation of genres
- Can be used for other applications e.g. context detection in hearing aids
Model

- Making the computer classify a sound piece into musical genres such as jazz, techno or blues.
Features for genre classification

30s sound clip from the center of the song

6 MFCCs, 30ms frame
6 MFCCs, 30ms frame
6 MFCCs, 30ms frame

3 ARCs per MFCC, 760ms frame

30-dimensional AR features, $x_r, r=1,..,80$
Results reported in


### Best 11-genre confusion matrix

<table>
<thead>
<tr>
<th></th>
<th>Alternative</th>
<th>Country</th>
<th>Easy-listening</th>
<th>Electronica</th>
<th>Jazz</th>
<th>Latin</th>
<th>Pop &amp; Dance</th>
<th>Rap &amp; Hip Hop</th>
<th>RB &amp; Soul</th>
<th>Reggae</th>
<th>Rock</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Alternative</strong></td>
<td>41.8</td>
<td>6.4</td>
<td>4.5</td>
<td>3.6</td>
<td>3.6</td>
<td>2.7</td>
<td>8.2</td>
<td>2.7</td>
<td>4.5</td>
<td>3.6</td>
<td>18.2</td>
</tr>
<tr>
<td><strong>Country</strong></td>
<td>0.9</td>
<td><strong>72.7</strong></td>
<td>7.3</td>
<td>0.0</td>
<td>4.5</td>
<td>2.7</td>
<td>4.5</td>
<td>0.9</td>
<td>2.7</td>
<td>0.0</td>
<td>3.6</td>
</tr>
<tr>
<td><strong>Easy-listening</strong></td>
<td>1.8</td>
<td>11.8</td>
<td><strong>61.8</strong></td>
<td>2.7</td>
<td>4.5</td>
<td>2.7</td>
<td>2.7</td>
<td>0.0</td>
<td>2.7</td>
<td>3.6</td>
<td>5.5</td>
</tr>
<tr>
<td><strong>Electronica</strong></td>
<td>5.5</td>
<td>0.9</td>
<td>10.9</td>
<td><strong>41.8</strong></td>
<td>8.2</td>
<td>5.5</td>
<td>7.3</td>
<td>10.9</td>
<td>2.7</td>
<td>5.5</td>
<td>0.9</td>
</tr>
<tr>
<td><strong>Jazz</strong></td>
<td>0.9</td>
<td>4.5</td>
<td>8.2</td>
<td>10.9</td>
<td>50.0</td>
<td>2.7</td>
<td>3.6</td>
<td>2.7</td>
<td>7.3</td>
<td>6.4</td>
<td>2.7</td>
</tr>
<tr>
<td><strong>Latin</strong></td>
<td>3.6</td>
<td>8.2</td>
<td>2.7</td>
<td>4.5</td>
<td>3.6</td>
<td><strong>37.3</strong></td>
<td>8.2</td>
<td>8.2</td>
<td>4.5</td>
<td>11.8</td>
<td>7.3</td>
</tr>
<tr>
<td><strong>Pop &amp; Dance</strong></td>
<td>6.4</td>
<td>9.1</td>
<td>6.4</td>
<td>9.1</td>
<td>0.9</td>
<td>11.8</td>
<td><strong>43.6</strong></td>
<td>2.7</td>
<td>3.6</td>
<td>2.7</td>
<td>3.6</td>
</tr>
<tr>
<td><strong>Rap &amp; Hip Hop</strong></td>
<td>0.0</td>
<td>0.0</td>
<td>0.9</td>
<td>7.3</td>
<td>0.9</td>
<td>4.5</td>
<td>3.6</td>
<td><strong>62.7</strong></td>
<td>1.8</td>
<td>17.3</td>
<td>0.9</td>
</tr>
<tr>
<td><strong>RB &amp; Soul</strong></td>
<td>0.9</td>
<td>8.2</td>
<td>9.1</td>
<td>0.9</td>
<td>9.1</td>
<td>11.8</td>
<td>7.3</td>
<td>9.1</td>
<td><strong>29.1</strong></td>
<td>5.5</td>
<td>9.1</td>
</tr>
<tr>
<td><strong>Reggae</strong></td>
<td>0.9</td>
<td>0.9</td>
<td>0.0</td>
<td>3.6</td>
<td>4.5</td>
<td>5.5</td>
<td>1.8</td>
<td>17.3</td>
<td>3.6</td>
<td><strong>61.8</strong></td>
<td>0.0</td>
</tr>
<tr>
<td><strong>Rock</strong></td>
<td>25.5</td>
<td>16.4</td>
<td>5.5</td>
<td>0.9</td>
<td>5.5</td>
<td>2.7</td>
<td>6.4</td>
<td>0.0</td>
<td>6.4</td>
<td>1.8</td>
<td>29.1</td>
</tr>
</tbody>
</table>
Best 11-genre confusion matrix

<table>
<thead>
<tr>
<th></th>
<th>Alternative</th>
<th>Country</th>
<th>Easy-listening</th>
<th>Electronic</th>
<th>Jazz</th>
<th>Latin</th>
<th>Pop&amp;Dance</th>
<th>Rap&amp;Hiphop</th>
<th>RB&amp;Soul</th>
<th>Reggae</th>
<th>Rock</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alternative</td>
<td>41.8</td>
<td>6.4</td>
<td>4.5</td>
<td>3.6</td>
<td>3.6</td>
<td>2.7</td>
<td>8.2</td>
<td>2.7</td>
<td>4.5</td>
<td>3.6</td>
<td>18.2</td>
</tr>
<tr>
<td>Country</td>
<td>0.9</td>
<td>72.7</td>
<td>7.3</td>
<td>0.0</td>
<td>4.5</td>
<td>2.7</td>
<td>4.5</td>
<td>0.9</td>
<td>2.7</td>
<td>0.0</td>
<td>3.6</td>
</tr>
<tr>
<td>Easy-listening</td>
<td>1.8</td>
<td>11.8</td>
<td>81.8</td>
<td>2.7</td>
<td>4.5</td>
<td>2.7</td>
<td>2.7</td>
<td>0.0</td>
<td>2.7</td>
<td>3.6</td>
<td>5.5</td>
</tr>
<tr>
<td>Electronic</td>
<td>5.5</td>
<td>0.9</td>
<td>10.9</td>
<td>41.8</td>
<td>8.2</td>
<td>5.5</td>
<td>7.3</td>
<td>10.9</td>
<td>2.7</td>
<td>5.5</td>
<td>0.9</td>
</tr>
<tr>
<td>Jazz</td>
<td>3.6</td>
<td>8.2</td>
<td>2.7</td>
<td>4.5</td>
<td>3.6</td>
<td>37.3</td>
<td>8.2</td>
<td>8.2</td>
<td>4.5</td>
<td>11.8</td>
<td>7.3</td>
</tr>
<tr>
<td>Latin</td>
<td>6.4</td>
<td>9.1</td>
<td>6.4</td>
<td>9.1</td>
<td>0.9</td>
<td>11.8</td>
<td>43.6</td>
<td>2.7</td>
<td>3.6</td>
<td>2.7</td>
<td>3.6</td>
</tr>
<tr>
<td>Pop&amp;Dance</td>
<td>0.0</td>
<td>0.0</td>
<td>0.9</td>
<td>7.3</td>
<td>0.9</td>
<td>4.5</td>
<td>3.6</td>
<td>62.7</td>
<td>1.8</td>
<td>17.3</td>
<td>0.9</td>
</tr>
<tr>
<td>Rap&amp;Hiphop</td>
<td>0.9</td>
<td>8.2</td>
<td>9.1</td>
<td>0.9</td>
<td>9.1</td>
<td>11.8</td>
<td>7.3</td>
<td>9.1</td>
<td>29.1</td>
<td>5.5</td>
<td>9.1</td>
</tr>
<tr>
<td>RB&amp;Soul</td>
<td>0.9</td>
<td>0.9</td>
<td>0.0</td>
<td>3.6</td>
<td>4.5</td>
<td>5.5</td>
<td>1.8</td>
<td>17.3</td>
<td>3.6</td>
<td>61.8</td>
<td>0.0</td>
</tr>
<tr>
<td>Reggae</td>
<td>25.5</td>
<td>16.4</td>
<td>5.5</td>
<td>0.9</td>
<td>5.5</td>
<td>2.7</td>
<td>6.4</td>
<td>0.0</td>
<td>6.4</td>
<td>1.8</td>
<td>29.1</td>
</tr>
</tbody>
</table>

11-genre problem (some overlap): 50% error

Human error about 43%
Emotional spaces

active

arousal

joyous

passive

valence

afraid

exited

happy

bored

idle

calm

content

unpleasant

distressed

depressed

sad

joyous

happy

joyous

bored

idle

calm

content

unpleasant

distressed

depressed

sad

afraid

exited

active

valence


Emotion modelling

\[ x_{ijk} \approx \mathcal{G} \times A \times B \times C = \sum_{lmn} g_{lmn} a_{il} b_{jm} c_{kn} \]
Semantics and Acoustics Features for Emotional Recognition in Speech

Semantics and Acoustics Features for Emotional Recognition in Speech

![Graph showing emotional recognition in speech with different features]
The valence dimension is more about what we say, while the arousal dimension is more about how we say it.
Audio separation

- A possible front end component e.g. the music search framework
- Noise reduction
- Music transcription
- Instrument detection and separation
- Vocalist identification

Semi-supervised learning methods

Nonnegative matrix factor 2D deconvolution

\[ V \approx \Lambda = \sum_{\tau, \phi} W^\tau H^\phi \]

time

pitch

Demonstration of the 2D convolutive NMF model
Separating music into basic components
Separating music into basic components

- Combined ICA and masking


Assumptions

- Stereo recording of the music piece is available.
- The instruments are separated to some extent in time and in frequency, i.e., the instruments are sparse in the time-frequency (T-F) domain.
- The different instruments originate from spatially different directions.
Separation principle: ideal T-F masking
**Results**

- The segregated outputs are dominated by individual instruments.
- Some instruments cannot be segregated by this method, because they are not spatially different.

<table>
<thead>
<tr>
<th>Output</th>
<th>Bass</th>
<th>Drum</th>
<th>Electric Guitar</th>
<th>Acoustic Guitar</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output1</td>
<td>72%</td>
<td>92%</td>
<td>3%</td>
<td>1%</td>
<td>17%</td>
</tr>
<tr>
<td>Output2</td>
<td>5%</td>
<td>1%</td>
<td>55%</td>
<td>4%</td>
<td>14%</td>
</tr>
<tr>
<td>Output3</td>
<td>9%</td>
<td>4%</td>
<td>9%</td>
<td>72%</td>
<td>21%</td>
</tr>
<tr>
<td>Remaining</td>
<td>14%</td>
<td>3%</td>
<td>32%</td>
<td>23%</td>
<td>48%</td>
</tr>
</tbody>
</table>

% of power: 46% 27% 1% 7% 7%
Wind noise reduction

Single channel separation: Sparse NMF decomposition

- Code-book (dictionary) of noise spectra is learned
- Can be interpreted as an advanced spectral subtraction technique

original          cleaned

alternative method (qualcom)
A cognitive search engine - MuZeeker

Idea is to create a search engine that is not affected by the link structure, but instead based solely on the actual contents of web pages and capability to perform categorizing. This making it possible to filter out any unwanted results.

- Wikipedia used as a proxy for the music users mental model
- Implementation: Filter retrieval using Wikipedia’s article/ categories
- Preference to MuZeeker over Google in task solving
A cognitive search engine – CASTSEARCH: Context based Spoken Document Retrieval

http://castsearch.imm.dtu.dk
Sound segmentation
Retrieved documents:
... california governor arnold's fortson agar inspected the california mexico border by helicopter wednesday to see ...

... but governor orville schwartz wicker denying the request saying...
AV integration

Acoustic epe
+ Visual ete
= perceptual eke / ete

Vision influences auditory perception!
Cognitive AV integration

Purpose

To study AV integration and how it is influenced by physical and cognitive factors

- Behavioral experiments
  - Reveal the subjective audiovisual percept
- EEG
  - reveals the electro-physiological correlates of AV integration
- Mathematical modeling
  - Reveals the brain’s assumptions, goals and flaws in the integration of information across the senses
Research and innovation projects

2009

Danish Sound Technology Network. Supported by DASTI. 14 MDKK + 8 MDKK (15 MDKK)

2012

CoSound - a cognitive systems approach to enriched and actionable information from audio streams. Supported by the Danish Council for Strategic Research. 17.5 MDKK (6 MDKK)

2014

2015
CoSound

CoSound is a multi-disciplinary strategic research project addressing societal challenges related to **productivity, communication and well-being**

Productivity, communication and well-being depends on digital media and the delivery of multimodal media information on many different platforms including TV, social, and mobile media.

**Music and media consumption is in a revolution**

Traditional business models in the music, audio and broadcast sectors are challenged; however, the ubiquitous digitalization of media, localization information, and human behaviors has a huge and disruptive potential to be explored in strategic research.

**Audio information** represents a separate challenge over other modalities (e.g. text or visual information) since it can be sensed and perceived as an abstract, emotional stream.
CoSound

DTU Informatics
Musikzonen
B&O
Royal School of Library and Information Science
UCL
Department of Arts and Cultural Studies, Copenhagen University
State and University Library

DR
Queen Mary University of London
Geckon
Hindenburg Systems
Aalborg University

Syntonetic
University of Glasgow
CoSound

VISION
to develop a flexible modular audio data processing platform for new products and services in the commercial sector; the public service sector; and in educational and cultural research. We will prototype and evaluate solutions in all these areas.
A cognitive architecture

Combine bottom-up and top-down processing
- Top-down user feedback
  • High specificity
  • Time scales: long, slowly adapting
- Bottom-up data modeling
  • High sensitivity
  • Time scales: short, fast adaptation

Courtesey of Lars Kai Hansen, DTU
CoSound

The main hypothesis is that the integration of bottom-up data derived from audio streams and top-down data streams from users can enable actionable cognitive representations, which will positively impact and enrich user interaction with massive audio archives, as well as facilitating new commercial success in the Danish sound technology sector.

We will test the hypothesis at three different functionality levels: 1) personalized audio streams; 2) task driven navigation and organization; 3) sharing of enriched audio streams through editing and co-creation.
Danish Sound Technology Network

What is it?

What do we do?
VISION

The vision of the Danish Sound Technology network is that Denmark is a leading country with regards to sound technology in terms of knowledge, research and education. Danish sound technology will be the epitome of high quality in products and services as well as in physical rooms and social contexts.
MISSION

Danish Sound Technology Network embraces all individuals, organizations and businesses in Denmark in the area of sound technology. We create a new space for innovation, collaboration and dissemination of knowledge across existing.
http://www.lydteknologi.dk/pa2011/

PUBLIC ADDRESS 2011
557 members in 321 companies and organizations

Netværkets 557 medlemmer fordelt på organisationstype

- Enkeltmand svirksomhed: 15
- SMV: 36
- Stor virksomhed: 60
- Freelance: 114
- Universiteter: 185
- Universiteter: 90
321 companies and organizations

321 organisationer i netværket

- 134: SMV
- 57: GTS
- 25: Offentlig virksomhed
- 24: Freelance
- 20: GTS
- 23: Enkeltmandsvirksomheder
- 27: Andre
- 31: Stor virksomhed
Consortium partners in Danish Sound Technology Network

More than 100 researchers at

- Sections for Acoustics and Multimedia Information and Signal Processing, Electronics Systems, AAU
- Section for Media Technology, Dept. of Architecture, Design and Media Technology, AAU
- Acoustics Technology and Hearing Systems groups at Dept. of Electrical Engineering, DTU
- Section for Cognitive Systems at Dept. of Informatics and Mathematical Modelling, DTU
- Institute of Sensors, Signals and Electrotechnics, SDU
- DELTA
### Danish positions of strength

**critical mass and visibility**

<table>
<thead>
<tr>
<th>Area</th>
<th>Key Points</th>
</tr>
</thead>
</table>
| **Sound recording and reproduction**      | • Professional live sound systems  
• HiFi systems  
• Class D amplifier systems |
| **Diagnostic and monitoring systems**     | • Environmental sound analysis  
• Forensics and surveillance  
• Measurement systems |
| **Digital media systems**                 | • Organization and retrieval of music and sound and semantic audio  
• Professional broadcast production systems  
• Home entertainment systems incl. gaming |
| **Designed soundscapes and sound branding** | • Sound communication  
• Sound for electric cars |
| **Assistive technology and medical devices** | • Hearing instruments  
• Assistive sound in the medical care sector |