Extracting meaning from audio signals – a machine learning and signal processing approach

Jan Larsen
Cognitive Systems Section
Dept. of Informatics and Mathematical Modelling
Technical University of Denmark
jl@imm.dtu.dk, www.imm.dtu.dk/~jl
Potential of technological contributions

- Involvement of people and the inclusiveness goal
- Handling of massive amounts of often conflicting data
- Enabling user-centric crowd computing
- Context detection and adaptation
- New intelligent tools eliminating trivial work - enhancing experience

It takes a cross-disciplinary effort to release the potential
Group profile

- 5 faculty
- 1 adj. prof.
- 3 postdocs
- 4 adm
- 20 Ph.D. students
- 10 M.Sc. students

Machine learning
Signal processing
Cognitive modeling
Systems neuroscience

Digital economy
Mobile services
Multimedia
Biomedical

extraction of meaningful and actionable information by ubiquitous learning from data
The legacy of Allan Touring and Nobert Wiener

- theory of computing
- cybernetics

processing → adaption → understanding → cognition

information and data → people
Transformation of sound technologies

The transformationen happens across business areas, sectors and disciplines

Stand alone P&S to systems and netværk of P&S

Information sources, sensors, and transducers

Adaptive, multimodal interfaces

Interaction and adaption to environment and context

Acoustics

Signal processing

Transducers

Information sources, sensors, and transducers

Adaptive, multimodal interfaces

Interaction and adaption to environment and context

Acoustics

Signal processing

Transducers

Information sources, sensors, and transducers

Adaptive, multimodal interfaces

Interaction and adaption to environment and context

Acoustics

Signal processing

Transducers

Information sources, sensors, and transducers

Adaptive, multimodal interfaces

Interaction and adaption to environment and context

Acoustics

Signal processing

Transducers

Information sources, sensors, and transducers

Adaptive, multimodal interfaces

Interaction and adaption to environment and context
Information processing pipeline

Physical domain
- objects
- environment

Technical domain
- Data modeling
  - Domain knowledge and other data sources
- Sensors/measurements

User/cognitive domain
- HCI
  - Perception
  - Interpretation
  - Interaction

- Quantification
- Detection
- Discrimination
- Prediction
- Description
Technical 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
Learning from massive data sets

Disentanglement of confusing, ambiguous, conflicting and vast amounts of information

Perform specific tasks

- Exploration
- Retrieval
- Search
- Physical operation and manipulation
- Information enrichment
- Making information actionable
- Navigation and control

- Decision support

Examples

- Detecting topics in large text corpora
- Automatic annotation/labeling of songs with genre, mood, etc.
- Speech and image recognition
The unreasonable effectiveness of data

- E. Wigner 1960: The unreasonable effectiveness of mathematics in the natural sciences
- There is often a sufficient number of data such that simple methods performs better than complex methods
- The power of learning with from unlabeled data which are abundant
- The power of linking many different sources
- Bridging semantic gaps
  - The same meaning can be expressed in many ways – and the same expression can convey many different meanings
  - Shared cognitive and cultural contexts helps the disambiguation of meaning
  - Ontologies: a social construction among people with a common shared motive
  - Classical handcrafted ontology building is infeasible – crowd computing / crowd sourcing is possible!

The potential of learning machines

- Most real world problems are too complex to be handled by classical physical models and systems engineering approach
- In most real world situations there is access to data describing properties of the problem
- Learning machines can offer
  - Learning of optimal prediction/decision/action
  - Adaptation to the usage environment
  - Explorative analysis and new insights into the problem and suggestions for improvement
Intelligent Sound Project

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

www.intelligentsound.org
Huge demand for tools

Organization, search and retrieval

- Recommender systems ("taste prediction")
- Playlist generation
- Finding similarity in music (e.g., genre classification, instrument classification, etc.)
- Hit prediction
- Newscast transcription/search
- Music transcription/search
Machine learning in sound information processing

- **Audio data**
- **Meta data**
  - ID3 tags
  - Context
- **Tasks**
  - Grouping
  - Classification
  - Mapping to a structure
  - Prediction (e.g., answer to query)

**User networks**
- Co-play data
- Playlist
- Communities
- User groups

**Machine learning model**
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.

Query by humming.

Using social network analysis.

search for related songs using the “400 genes of music”
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 and blues.
How do humans do?

- Sounds – loudness, pitch, duration and timbre
- Music – mixed streams of sounds
- Recognizing musical genre
  - physical and perceptual: instrument recognition, rhythm, roughness, vocal sound and content
  - cultural effects
How well do humans do?

- Data set with 11 genres
- 25 people assessing 33 random 30s clips

**Accuracy**
54 - 61%

Baseline: 9.1%
What’s the problem?

- Technical problem: Hierarchical, multi-labels
- Real problems: Musical genre is not an intrinsic property of music
  - A subjective measure
  - Historical and sociological context is important
  - No Ground-Truth
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, $x_r, r=1,\ldots,80$
Example of MFCC’s

- Cross correlation
- Temporal correlation

A ten second excerpt of the song Masters of Revenge by Body Count
Results reported in


Best results

- 5-genre problem (with little class overlap) : 2% error
  - Comparable to human classification on this database
- Amazon.com 6-genre problem (some overlap) : 30% error
- 11-genre problem (some overlap) : 50% error
  - Human error about 43%
## Best 11-genre confusion matrix

![Confusion Matrix](image)

- **Alternative**
- **Country**
- **Easy-listening**
- **Electronica**
- **Jazz**
- **Latin**
- **Pop&dance**
- **Rap&hiphop**
- **RB&Soul**
- **Reggae**
- **Rock**

### Table

```
<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;hiphop</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>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><strong>Easy-listening</strong></td>
<td>1.8</td>
<td>11.8</td>
<td>61.8</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><strong>Electronica</strong></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><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>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><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>43.6</td>
<td>2.7</td>
<td>3.6</td>
<td>2.7</td>
<td>3.6</td>
</tr>
<tr>
<td><strong>Rap&amp;hiphop</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>62.7</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>29.1</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>61.8</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>
```
Music separation

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

Semi-supervised learning methods

Nonnegative matrix factor 2D deconvolution

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

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

• Evaluation on real stereo music recordings, with the stereo recording of each instrument available, before mixing.
• We find the correlation between the obtained sources and the by the ideal binary mask obtained sources.
• Other segregated music examples and code are available online via http://www.imm.dtu.dk
Results

• The segregated outputs are dominated by individual instruments

• Some instruments cannot be segregated by this method, because they are not spatially different.
Conclusion on combined ICA T-F separation

• An unsupervised method for segregation of single instruments or vocal sound from stereo music.
• The segregated signals are maintained in stereo.
• Only spatially different signals can be segregated from each other.
• The proposed framework may be improved by combining the method with single channel separation methods.
Wind noise reduction

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)
Objective performance
A cognitive search engine - Muzeeker

- Wikipedia based common sense
- Wikipedia used as a proxy for the music users mental model
- Implementation: Filter retrieval using Wikipedia’s article/categories

Muzeeker.com
A cognitive search engine – CASTSEARCH: Context based Spoken Document Retrieval

... 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'.
Vertical search

• Deep web databases
  – Digital media
  – For profit: DMR issues

• Specialized search engines
  – Professional users
  – Modeling deep structure

• Key role in Web 2.0
  – User generated content
  – Bioinformatics
  – Neuroinformatics:
    • BrainMap, Brede search engine

Horizontal search

• Google
  – Volume
  – Ranking
  – Explorative vs retrieval
  – Adword business model

• Semantic web
  – Wikipedia
  – User generated content
Crowd computing and user involvement

Challenges: There is a social/psychological inertia towards traditional solutions

1. The Retarding Power (or Inertia) of a Word
2. A Partial Restriction Becomes a Blanket Restriction
3. Tradition Cannot be Broken
4. Words and Their Assumed Properties or Characteristics
5. Inadmissible Range of Data
6. Association of Objects with Senses
7. All Information Given is Valid

Victor Fey and Eugene Rivin: Innovation on Demand, 2005.
Samer Abdallah, Mark Plumbley: Information dynamics: patterns of expectation and surprise in the perception of music, [Connection Science, vol. 21, issue 2, p. 89, 2009]( Publication)
• Guessing tags - fun and useful
• Conceived by Luis von Ahn of Carnegie Mellon University
reCAPTCHA is a free CAPTCHA service that helps to digitize books and shows. Check out our paper in Science about it (or read more below).

A CAPTCHA is a program that can tell whether its user is a human or not: seen them — colorful images with distorted text at the bottom of them — are used by many websites to prevent abuse from "bots," or automated programs that generate spam. No computer program can read distorted text as easily as humans can and is known not to be able to navigate sites protected by CAPTCHAs.
Research based vs user-driven knowledge and folksonomy

- user driven knowledge is often inaccurate and misleading
- how do we avoid dominance by the popular (music recommendation systems)
- sufficient amount of contributions ensures the quality (wikipedia)
Measurement systems for ethical capital in the experience economy
socio-economic value of online communication

• New research 3-year research project starting Aug. 2009 (CBS, DTU, Univ. Milan)
• Forrester Research Report shows web2.0 marked grows enormously
• The assumption is that on-line spontaneous communication processes are predictable as they appear in networks and patterns which can be revealed by combining socio-economic studies, linguistics, text and network modeling

Responsible Business in the Blogosphere
• Google only works if you know what you are searching for
• We need to integrate with common knowledge sources (wikipedia)
• We need to use learning to annotate meta data
• We need users to create additional content, collaborate and interact with data
A cognitive architecture for search

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
CoSound architecture

**Bottom-up**
- Primary audio sources
- Sampling
- Feature extraction
- Temporal inference engine
- Domain prior information database
- Data warehouse
- Common knowledge sources

**Top-down**
- Users
  - Interaction and communication module
- User representation
  - User aspect 1
  - User aspect 2

**Cognitive domain representation**
Summary

• A cross-disciplinary effort is required to make research, innovation and commercial products and services
• Massiveness of data requires learning and cognitive modeling but has huge potential for new capabilities
• Integration of multiple information sources helps context detection and adaptation
• Internet penetration makes crowd sourcing possible and ensures inclusiveness
  – a window for the creative common
  – a way to bridging the semantic gap