Search for sounds -
a machine learning approach

www.intelligentsound.org
The digital music market

- **Wired, April 27, 2005:**
  
  "With the new Rhapsody, millions of people can now experience and share digital music legally and with no strings attached," RealNetworks chairman and CEO Rob Glaser said in a statement. "We believe that once consumers experience Rhapsody and share it with their friends, many people will upgrade to one of our premium Rhapsody tiers."

- **Financial Times**: 12:46 p.m. ET Dec. 28, 2005
  
  LONDON - Visits to music downloading Web sites saw a 50 percent rise on Christmas Day as hundreds of thousands of people began loading songs on to the iPods they received as presents.

- **Wired, January 17, 2006:**
  
  Google said today it has offered to acquire digital radio advertising provider dMarc Broadcasting for $102 million in cash.

- **Huge demand for tools:**
  
  organization, search, retrieval

- **Machine learning will play a key role in future systems**
Outline

- Machine learning framework for sound search
- Genre classification
- Independent component analysis for music separation
Informatics and Mathematical Modelling, DTU

- image processing and computer graphics
- intelligent signal processing
- operations research
- numerical analysis
- geoinformatics
- mathematical statistics
- mathematical physics
- safe and secure IT systems
- languages and verification
- system on-chips
- ontologies and databases
- design methodologies
- embedded/distributed systems

2003 figures
- 84 faculty members
- 28 administrative staff members
- 60 Ph.D. students
- 90 M.Sc. students annually
- 4000 students follow an IMM course annually
ISP Group

from processing to understanding

extraction of meaningful information by learning

Multimedia

Monitor Systems

Biomedical

Humanitarian

Demining

Monitor Systems

System

Machine learning

• 3+1 faculty
• 6+1 postdocs
• 20 Ph.D. students
• 10 M.Sc. students

Search for sounds – a machine learning approach
Machine learning in sound information processing

Tasks
- Grouping
- Classification
- Mapping to a structure
- Prediction
  e.g. answer to query

Meta data
- ID3 tags
- context

Machine learning model

User networks
- co-play data
- playlist
- communities
- user groups

Audio data

Users

Search for sounds – a machine learning approach
Aspects of search

**Specificity**
- standard search engines
- indexing of deep content
- Objective: high retrieval performance

**Similarity**
- more like this
- similarity metrics
- Objective: high generalization and user acceptance
Specialized search and music organization

The National Gallery of the Spoken Word

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

Query by humming

Explore by Genre, mood, theme, country, instrument

search for related songs using the “genes of music”
System overview

- Front-Ends
  - Winamp Users (Plugin)
  - Browser

- Sound Search Engine
  - DATA (Storage Zone) [SQL Tables]
  - Regular updates
  - DATA (Query Zone) [Similarity Structures]

- Maintenance System
  - Request for data
  - Processed data

- External sources
  - Streamripper archive/ripper
  - Internet/Free-Db

Search for sounds – a machine learning approach
Introduction

Financial Times (ft.com) 12:46 p.m. ET Dec. 28, 2005:
"LONDON - Visits to music downloading Web sites saw a 50 percent rise on Christmas Day as hundreds of thousands of people began loading songs on to the iPods they received as presents."

SoundSearch 0.1 combines co-play patterns, expert evaluations and music features to help you retrieve the music you like.

Use these music features to organize your search:

- Co-play
- Beat
- Expert
- Sound

Start the Music:
Storage and query

DATA (Storage Zone)

- SONGS:
  - idSlugs
  - Sound Features
  - ...

- USAGE:
  - userid
  - songid
  - timestamp
  - ...

- SOCIAL
  - ...

Regular Update (Offline)
SQL Search + Machine Learning

DATA (Query Zone)

Similarity Structures

- Title
- Snd.Feat.
- Co-play
- ...
- Genre

Query Process (Online)
Fast Search + Machine Learning

Query Result

- List A
- List B
- ...
- List Z

Search for sounds – a machine learning approach
Similarity structures

- Low level features:
  - Ad hoc from time-domain
  - Ad hoc from spectrum
  - MFCC
  - RCC
  - Bark/Sone
  - Wavelets
  - Gamma-tone-filterbank

- High level features:
  - Basic statistics
  - Histograms
  - Selected subsets
  - GMM
  - Kmeans
  - Neural Network
  - SVM
  - QDA
  - MoHMM

- Metrics:
  - Euclidian
  - Weighted Euclidian
  - Cosine
  - Nearest Feature Line
  - earth Mover Distance
  - Distance From Boundary
  - Cross-sampling

- Other:
  - loudness
  - zero-crossing
  - energy
  - log-energy
  - down sampling
  - autocorrelation
  - peak detection
  - delta-log-loudness
  - pitch
  - brightness
  - bandwidth
  - harmonicity
  - spectrum power
  - subband power
  - centroid
  - roll-off
  - low-pass filtering
  - spectral flatness
  - spectral tilt
  - shaprnness
  - roughness
Predicting the answer from query

\[ p(S_a | S_q, u) \]

- \( S_a \): index for answer song
- \( S_q \): index for query song
- \( u \): user (group index)
- \( C_i \): hidden cluster index of similarity \( i \)
Intelligent Sound Project
IMM (DTU) – CS, CT (AaU)

- Signal processing
- Databases
- Machine learning

Demo: Sound search engine
Demo: Matlab toolbox

Phd projects
Group publications
Joint publications
Workshops/Phd-courses
Research “tasks”

AaU Communication Technology:

TASK i): Features for sound based context modelling - MPEG and beyond
TASK ii): Signal separation in noisy environments: ICA and noise reduction

AaU Computer Science/Database Management:

TASK iii): Multidimensional management of sound as context
TASK iv): Advanced Query Processing for Sound Feature Streams

DTU IMM-ISP

TASK v): Context detection in sound streams
TASK vi): Webmining for sound
<table>
<thead>
<tr>
<th>ISOUND PUBLICATIONS 2005-2006</th>
</tr>
</thead>
<tbody>
<tr>
<td>L. Feng, L. K. Hansen, <em>On low level cognitive components of speech</em>, International Conference on Computational Intelligence for Modelling (CIMCA’05), 2005</td>
</tr>
<tr>
<td>A. B. Nielsen, L. K. Hansen, U. Kjems, <em>Pitch Based Sound Classification</em>, Informatics and Mathematical Modelling, Technical University of Denmark, DTU, 2005</td>
</tr>
</tbody>
</table>
Genre classification

- Prototypical example of predicting meta data
- The problem of interpretation of genres
- Can be used for other applications e.g. hearing aids
- Models
Model

- Making the computer classify a sound piece into musical genres such as jazz, techno and blues.

Flowchart:

Sound Signal → Pre-processing → Feature extraction → Statistical model → Probabilities → Decision → Post-processing
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
Music genres form a hierarchy

Music
- Jazz
  - Swing
  - Classic BB
    - Quincy Jones: “Stuff like that”
  - Vintage BB
  - Contemp. BB
- New Age
- Latin
- Cool
- New Orleans

(according to Amazon.com)
Search for sounds – a machine learning approach
Music Genre Classification Systems

Sound Signal → Pre-processing → Feature extraction → Feature vector → Statistical model → Probabilities → Post-processing → Decision
Features

- **Short time features (10-30 ms)**
  - MFCC and LPC
  - Zero-Crossing Rate (ZCR), Short-time Energy (STE)
  - MPEG-7 Features (Spread, Centroid and Flatness Measure)

- **Medium time features (around 1000 ms)**
  - Mean and Variance of short-time features
  - Multivariate Autoregressive features (DAR and MAR)

- **Long time features (several seconds)**
  - Beat Histogram
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, $x_r, r=1,\ldots,80$
Search for sounds – a machine learning approach
Statistical models

- Desired: \( p(c|s) \) (class \( c \) and song \( s \))
- Used models:
  - Integration of MFCCs
  - Linear and non-linear neural networks
  - Gaussian classifier
  - Gaussian Mixture Model
  - Co-occurrence models
Best results

- 5-class problem (with little class overlap) : 2% error
  - Comparable to human classification on this database
- Amazon.com 6-class problem (some overlap) : 30% error
- 11-class problem (some overlap) : 50% error
  - Human error about 43%
The Clever Jukebox
The Art of Automated Genre Classification

**Theory:**

Automatic musical genre classification can be defined as the science (or art) of finding computer algorithms that take a (digitized) sound clip as input and yield a musical genre as output. The goal of automated genre classification is, of course, that the musical genre which is output should agree with the human classification of the sound into genre.

This demo illustrates an approach to the problem that first extract frequency-based sound features followed by a "linear regression" classifier. The basic features are the so-called mel-frequency cepstral coefficients (MFCCs), which are extracted on a time-scale of 30 msec. From these MFCC features, auto-regressive coefficients (ARs) are extracted along with the mean and gain to get a single (30 dimensional) feature vector on the time-scale of 1 second. These features have been used because they have performed well in a previous study (Meng, Ahrendt, Larsen (2005)). Linear regression (or single-layer linear NN) is subsequently used for classification. This classifier is rather simple; current research investigates more advanced methods of classification.
Nonnegative matrix factor 2D deconvolution

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

Ref: Mikkel Schmidt and Morten Mørup, ICA2006
Demonstration of the 2D convolutive NMF model
Separating music into basic components
Motivation: Why separating music?

- Music Transcription
- Identifying instruments
- Identify vocalist
- Front end to search engine
### 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
Search for sounds – a machine learning approach

Gain difference between channels

Intelligent Signal Processing Group, IMM, DTU / Jan Larsen
Separation principle 2: ICA

What happens if a 2-by-2 separation matrix $W$ is applied to a 2-by-N mixing system?
ICA on stereo signals

- We assume that the mixture can be modeled as an instantaneous mixture, i.e.

\[ x = A(\theta_1, \ldots, \theta_N)s \quad A(\theta) = \begin{bmatrix} r_1(\theta_1) & \cdots & r_1(\theta_N) \\ r_2(\theta_1) & \cdots & r_2(\theta_N) \end{bmatrix} \]

- The ratio between the gains in each column in the mixing matrix corresponds to a certain direction.
Direction dependent gain

\[ r(\theta) = 20 \log |WA(\theta)| \]

When \( W \) is applied, the two separated channels each contain a group of sources, which is as independent as possible from the other channel.
Combining ICA and T-F masking

ICA+BM Separator

ICA

\[
\begin{align*}
BM_1 &= \begin{cases} 
1 & \text{when } Y_1 / Y_2 > c \\
0 & \text{otherwise}
\end{cases} \\
BM_2 &= \begin{cases} 
1 & \text{when } Y_2 / Y_1 > c \\
0 & \text{otherwise}
\end{cases}
\end{align*}
\]

\[X_1(t,f) \xrightarrow{BM_1} X_2(t,f)\]

\[\hat{x}_1^{(2)} \xrightarrow{BM_2} \hat{x}_2^{(2)}\]
Method applied iteratively
Improved method

- The assumption of instantaneous mixing may not always hold.
- Assumption can be relaxed.
- Separation procedure is continued until very sparse masks are obtained.
- Masks that mainly contain the same source are afterwards merged.
Mask merging

If the signals in the time domain are correlated, their corresponding masks are merged.

The resulting signal from the merged mask is of higher quality.
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 are available online.
The segregated outputs are dominated by individual instruments.

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

- We have presented an unsupervised method for segregation of single instruments or vocal sound from stereo music.
- Our method is based on combining ICA and T-F masking.
- 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.
**Conclusions**

- Search is a “productivity engine” simply important to quality of life...
- Generic and specialized search engines: different criteria and challenges
- Machine learning is essential for search!
- Music search based on musical features, meta data, and social network information