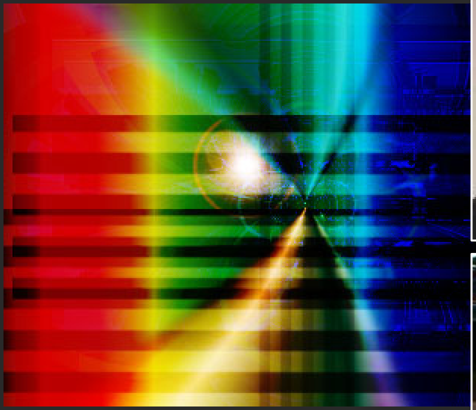
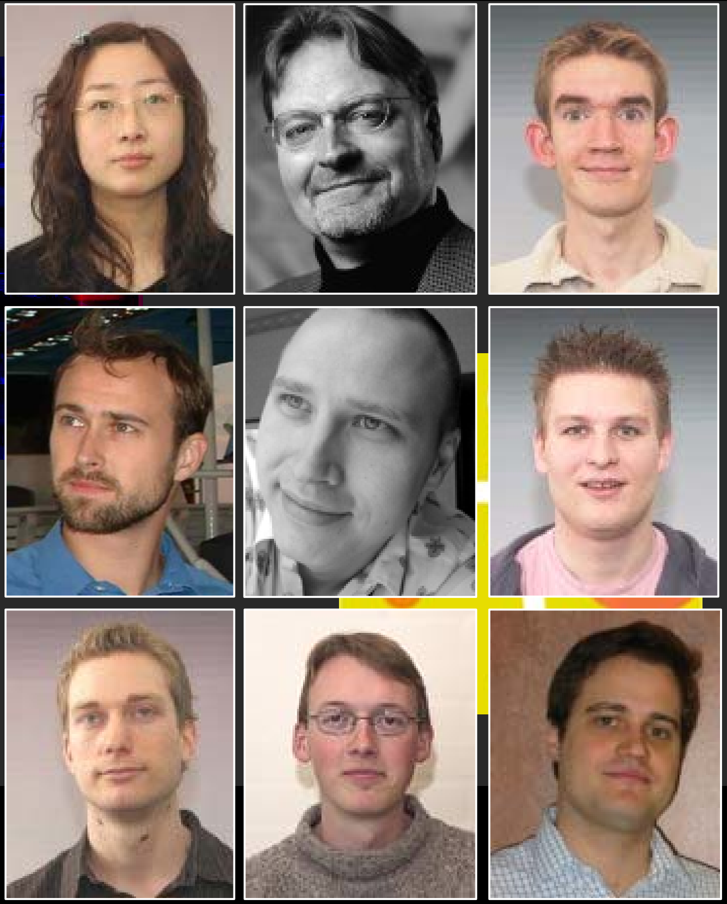
Extracting meaning from audio signals - a machine learning approach

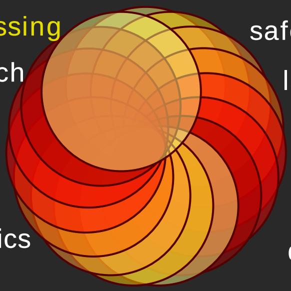
Jan Larsen

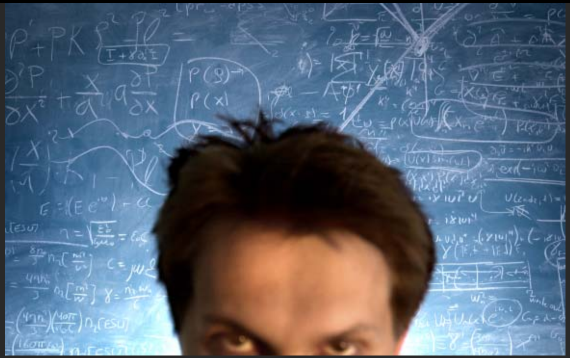




- isp.imm.dtu.dk
- www.intelligentsound.org

Informatics and Mathematical Modelling@DTU – the largest ICT department in Denmark

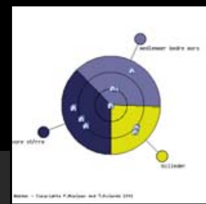




2006 figures

- 11.000 students signed in to courses
- 900 full time students
- 170 final projects at MSc
- 90 final projects at IT-diplom
- 75 faculty members
- 25 externally funded
- 70 PhD students
- 40 staff members
- DTU budget: 90 mill DKK
- External sources: 28 mill DKK

ISP Group



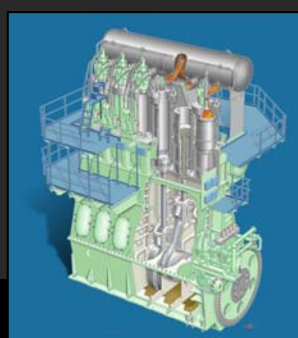
Multimedia

tics

from processing to understanding

extraction of meaningful information by learning

faculty
- 3 postdocs
- 20 Ph.D. students
- 10 M.Sc. students



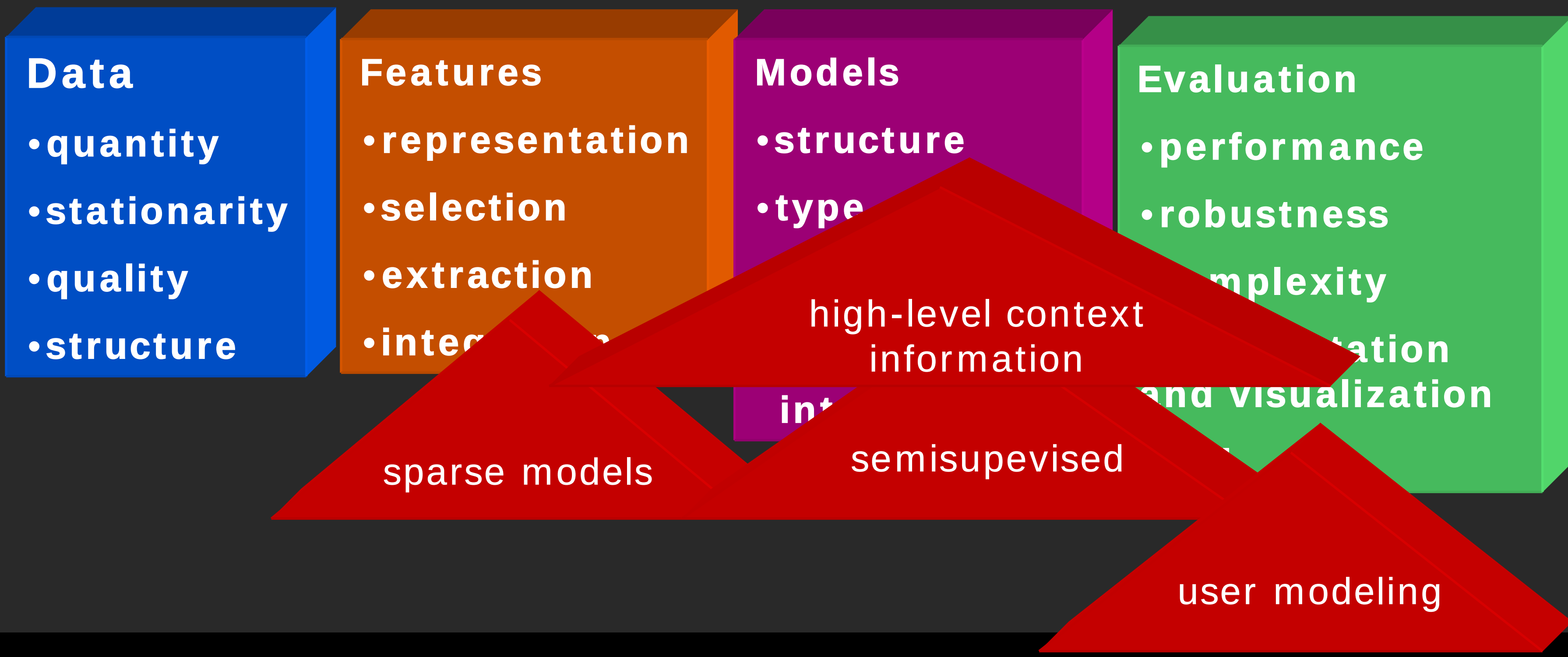
Monitor Systems

Biomedical

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

Issues and trends in machine learning



Outline

- Machine lea ... arch
  - *Involves a ... r modeling*
- Genre class
  - *Involves f ... ration*
  - *Linear and*
- Music and a
  - *Involves c ... rocessing*
  - *NMF and I*
- Wind noise
  - *Semi-supe*

Take home?

•New ways of using semi-supervised learning algorithms

•New ways of incorporating high-level information and users

•New application domains

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," Rob Glaser, RealNetworks chairman and CEO, 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 (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.

- **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 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 is going to play a key role in future systems

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

last.fm the social music revolution

Using social network analysis

AMG allmusic

Explore by Genre, mood, theme, country, instrument

Query by humming

Fraunhofer Institut Digitale Medientechnologie IDMT

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

PANDORA™

search for related songs using the "400 genes of music"

DTU

Sound information data

audio data

Meta data

ags

ontology

cription

level

low

User

co-play data

playlist

communities

user groups

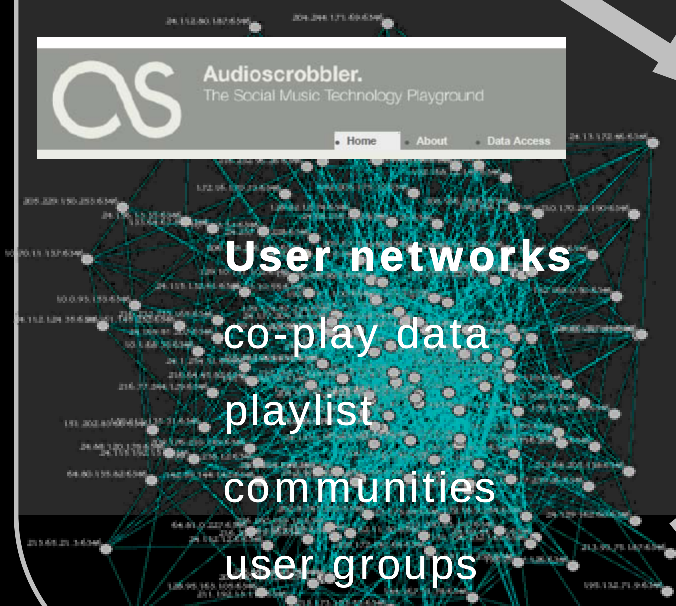
Machine learning in sound information processing

audio data

Meta data

ID3 tags

context



User networks

co-play data

playlist

communities

user groups

machine learning model

Tasks

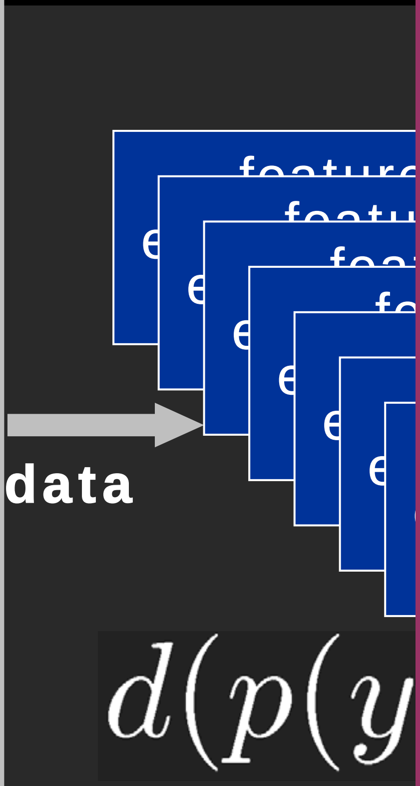
Grouping

Classification

Mapping to a structure

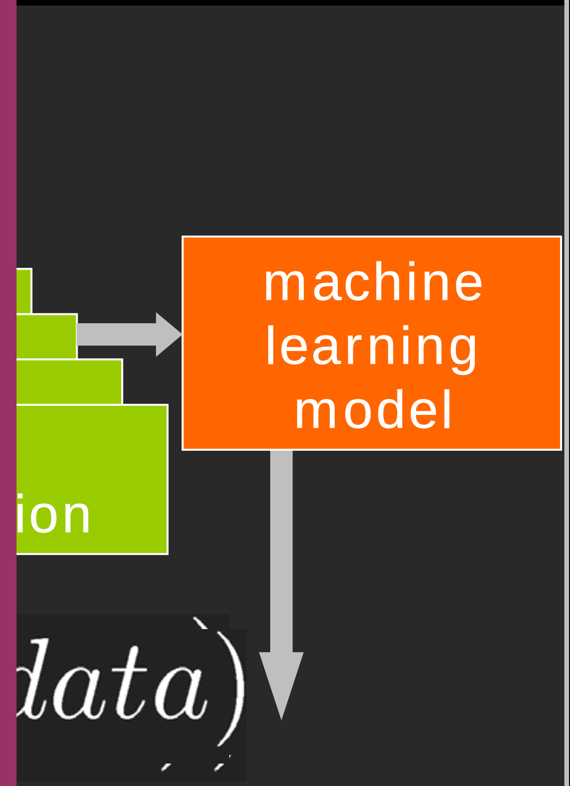
Prediction e.g. answer to query

Machine learning for high level interpretations



Similarity functions

Euclidian, Weighted Euclidian, Cosine, Nearest Feature Line, earth Mover Distance, Self-organized Maps, Distance From Boundary, Cross-sampling, Bregman, KL, Manhattan, Adaptive



Sim

- Lo
  –
- Hi
  –
- Me
  –

Ti

Frequency domain

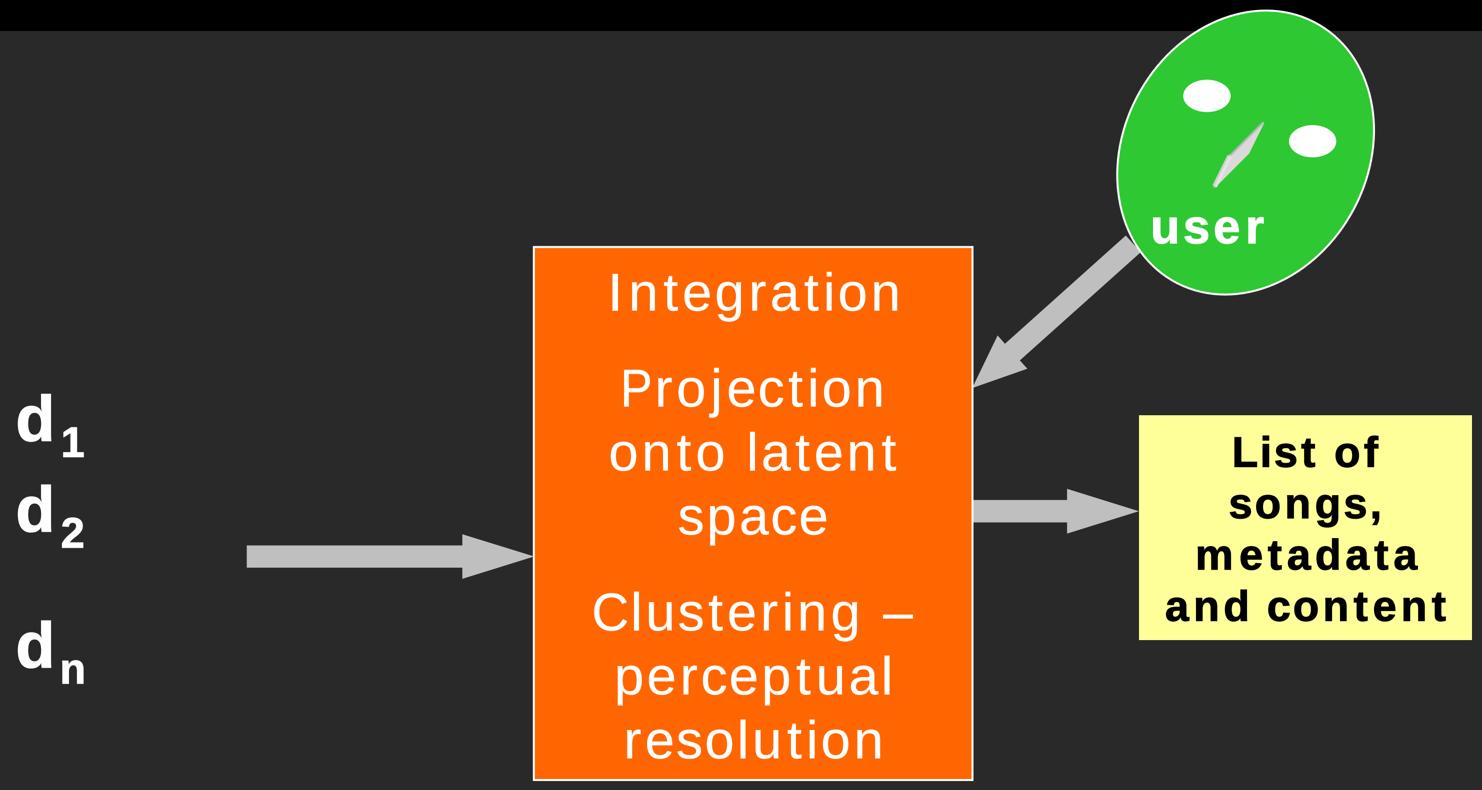
- MFCC
- Gamma tone filterbank
- pitch
- brightness
- bandwidth
- harmonicity
- spectrum power
- subband power
- centroid
- roll-off
- low-pass filtering
- spectral flatness
- spectral tilt
- sharpness
- roughness

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

Search and similarity integration



Similarity fu… …eling

$$P(c_j^{(k)} \ldots |s_l)$$

k'th high-level descriptor quantized in to groups

user specified weights

$$L_k = \sum_{j,l} \tilde{P}(c_j^{(k)}|s_l) \ldots$$

$$\ldots = \sum_{k=1}^{K} \alpha_k L_k$$

- Latent variables can satisfactorily explain all observed similarities and provides a very convenient representation for song retrieval
- Synergy between two descriptors was advatageous
- analogy between documents and songs opens new lines for investigating music structure using the elaborated machinery for web-mining

J. Arenas-García, … …chiøler, L.K. Hansen, J. Larsen… *…milarity fusion*, 2007.

http://www.intelligentsound.org/demos/conceptdemo.swf

Demo of WINAMP plugin

www.intelligentsound.org

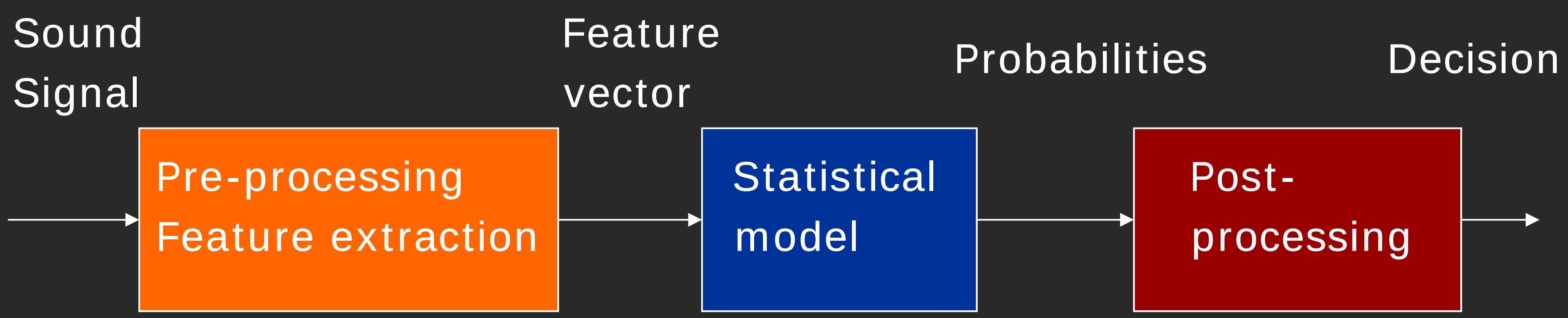
Lehn-Schiøler, T., Arenas-García, J., Petersen, K. B., Hansen, L. K., *A Genre Classification Plug-in for Data Collection*, ISMIR, 2006

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

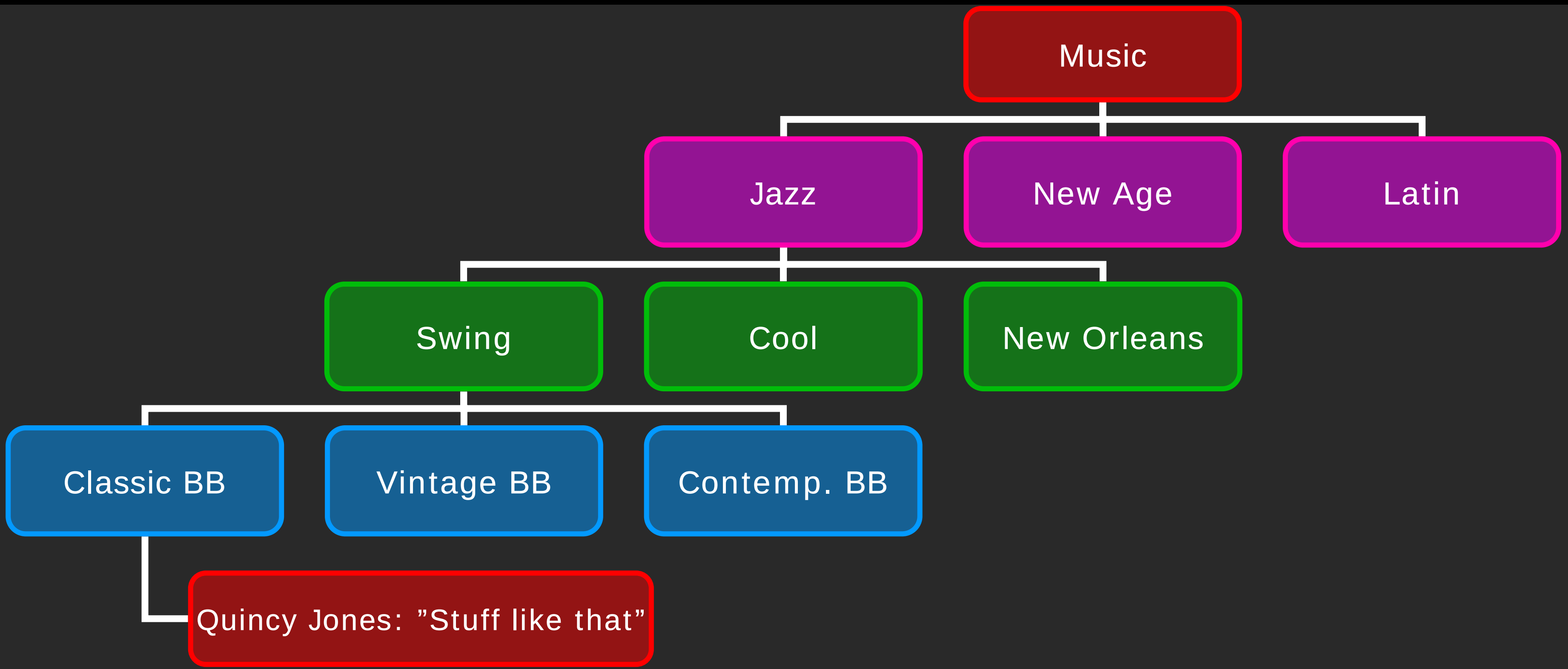
54 - 61 %

Baseline: 9.1%

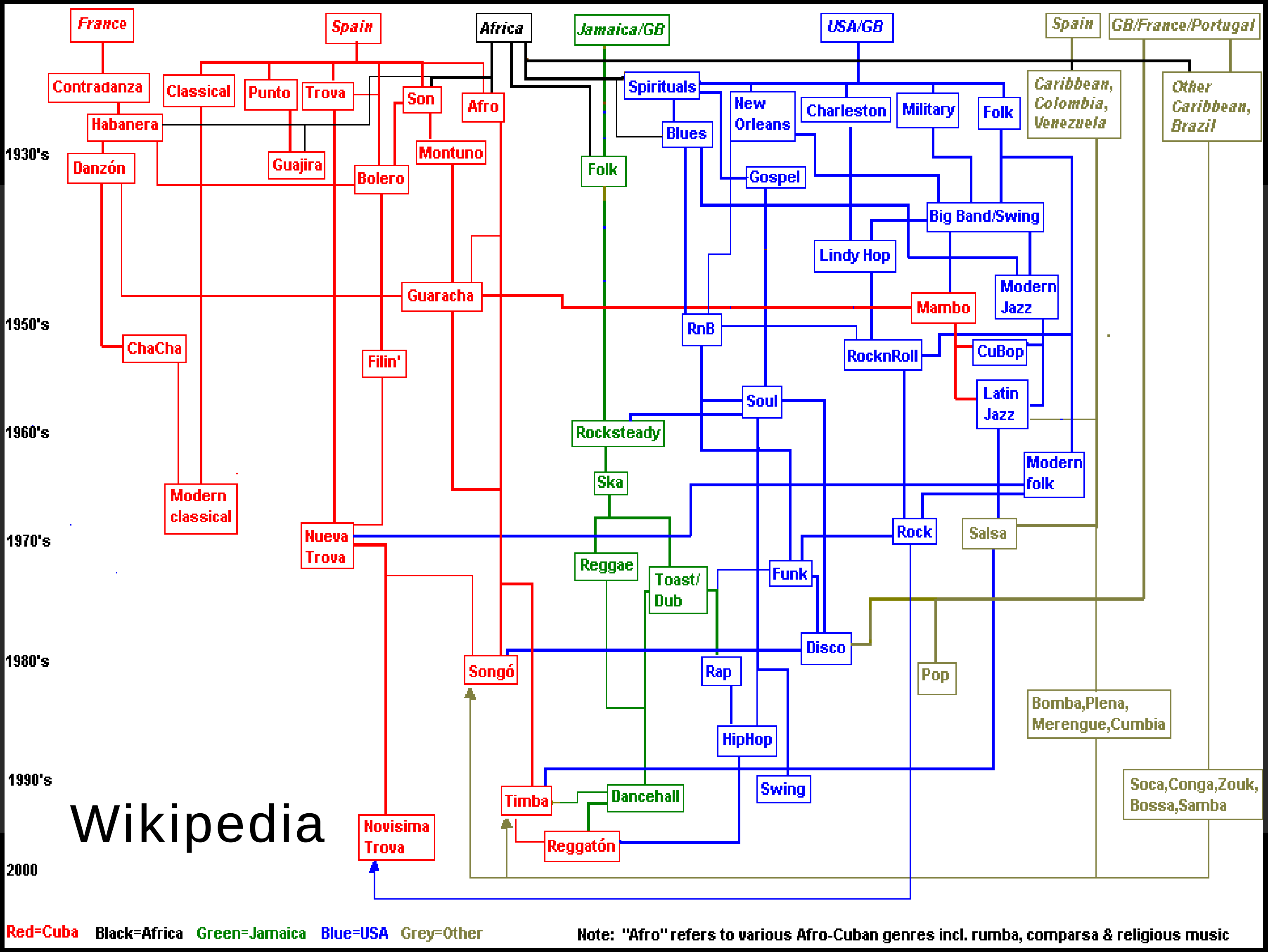
What's the problem ?

- 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

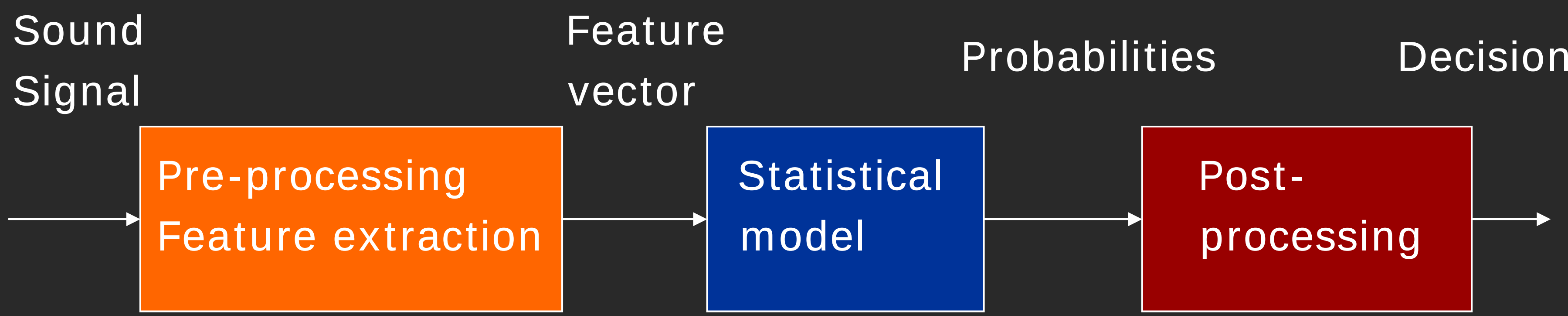
Music genres form a hierarchy



(according to Amazon.com)


France
Spain
Africa
Jamaica/GB
USA/GB
Spain
GB/France/Portugal
Contradanza
Classical
Punto
Trova
Son
Afro
Spirituals
New Orleans
Charleston
Military
Folk
Caribbean, Colombia, Venezuela
Other Caribbean, Brazil
Habanera
Blues
Montuno
Guajira
Bolero
Folk
Gospel
Danzón
Big Band/Swing
Lindy Hop
Guaracha
Modern Jazz
Mambo
RnB
ChaCha
Filin'
RocknRoll
CuBop
Latin Jazz
Soul
Rocksteady
Modern folk
Ska
Modern classical
Nueva Trova
Rock
Salsa
Reggae
Toast/ Dub
Funk
Disco
Songó
Rap
Pop
Bomba,Plena, Merengue,Cumbia
HipHop
Swing
Timba
Dancehall
Soca,Conga,Zouk, Bossa,Samba
Reggatón
Novisima Trova
1930's
1950's
1960's
1970's
1980's
1990's
2000
Wikipedia
Red=Cuba
Black=Africa
Green=Jamaica
Blue=USA
Grey=Other
Note: "Afro" refers to various Afro-Cuban genres incl. rumba, comparsa & religious music

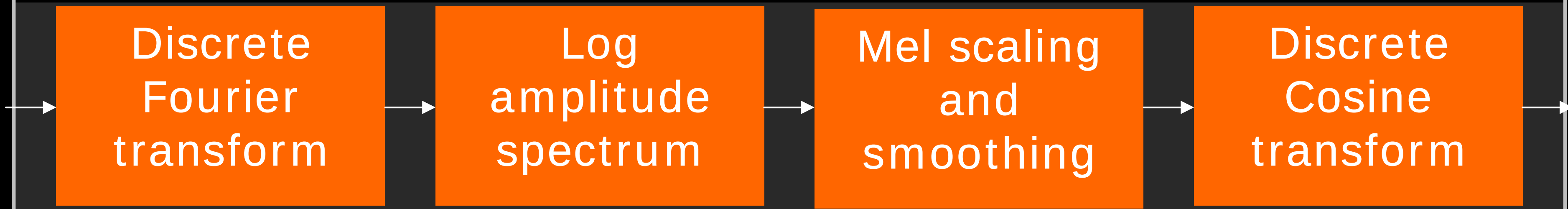
Music Genre Classification Systems



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

On MFCC



- MFCC represents a mel-weighted spectral envelope. The mel-scale models human auditory perception.
- Are believed to encode music timbre

Sigurdsson, S., Petersen, K. B., *Mel Frequency Cepstral Coefficients: An Evaluation of Robustness of MP3 Encoded Music*, Proceedings of the Seventh International Conference on Music Information Retrieval (ISMIR), 2006.

Features for genre classification

30s sound clip from the center of the song

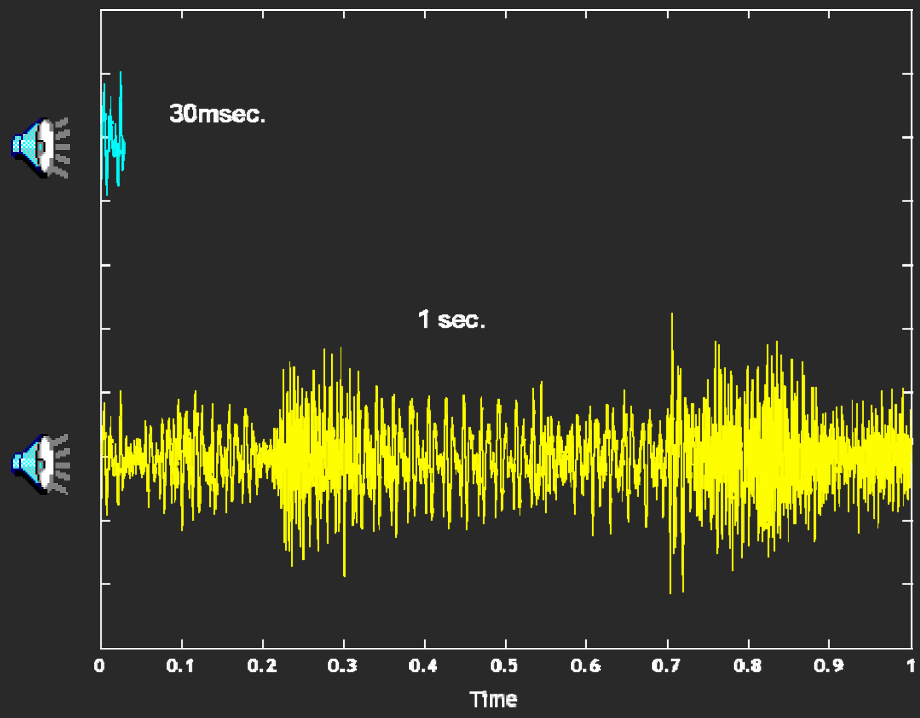
6 MFCCs, 30ms frame

6 MFCCs, 30ms frame

6 MFCCs, 30ms frame

3 ARCs per MFCC, 760ms frame

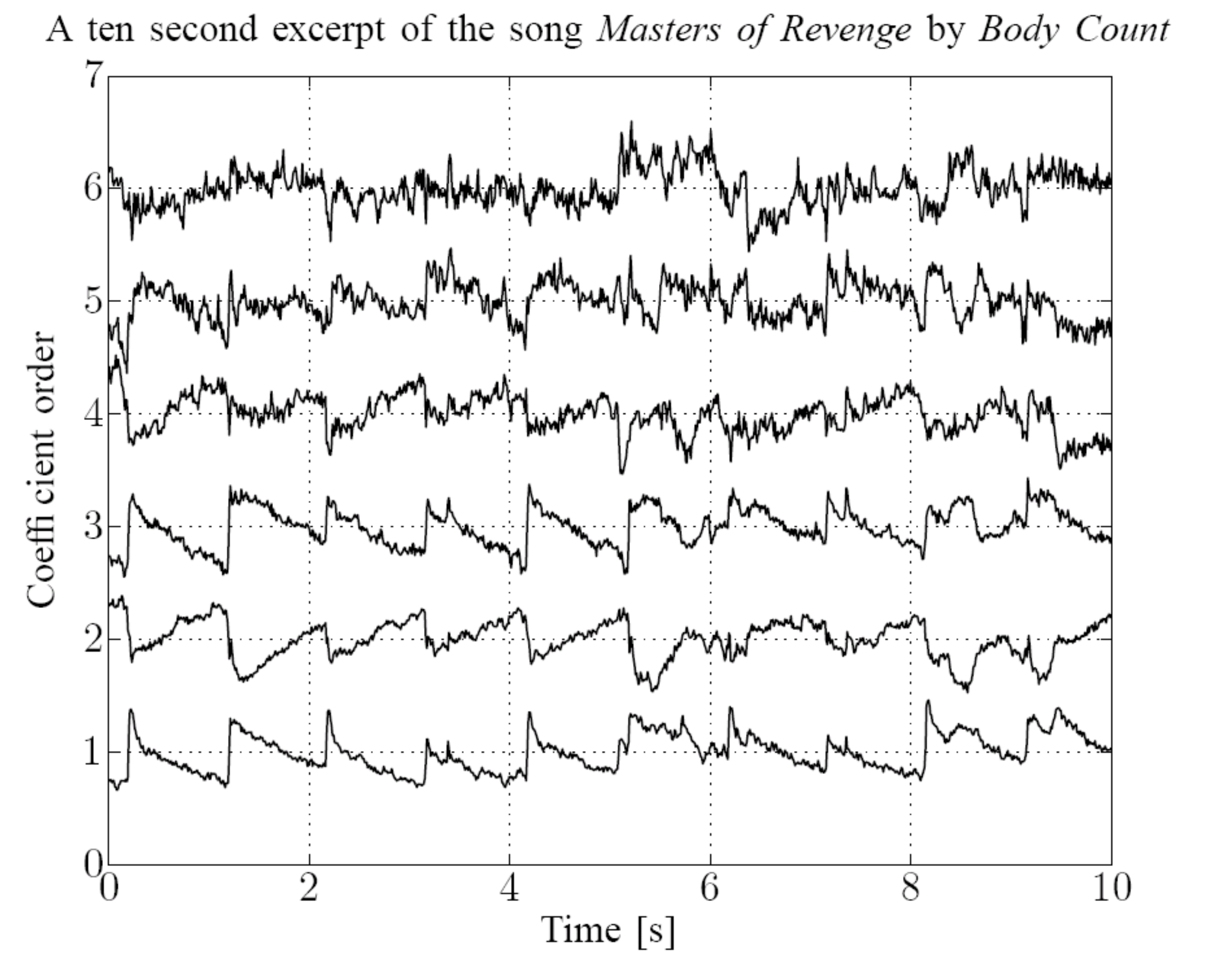
30-dimensional AR features, x_r, $r=1,..,80$


30msec.
1 sec.
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Time

Statistical models

- Desired: $p(c|s)$ (genre class c and song s)
- Used models
  - Intregration of MFCCs using MAR models
  - Linear and non-linear neural networks
  - Gaussian classifier
  - Gaussian Mixture Model
  - Co-occurrence models

Example of MFCC's



- Cross correlation
- Temporal correlation

Results reported in

- Meng, A., Ahrendt, P., Larsen, J., Hansen, L. K., Temporal Feature Integration for Music Genre Classification, IEEE Transactions on Speech and Audio Processing, 2007.
- A. Meng, P. Ahrendt, J. Larsen, *Improving Music Genre Classification by Short-Time Feature Integration*, IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. V, pp. 497-500, 2005.
- Ahrendt, P., Goutte, C., Larsen, J., *Co-occurrence Models in Music Genre Classification*, IEEE International workshop on Machine Learning for Signal Processing, pp. 247-252, 2005.
- Ahrendt, P., Meng, A., Larsen, J., *Decision Time Horizon for Music Genre Classification using Short Time Features*, EUSIPCO, pp. 1293--1296, 2004.
- Meng, A., Shawe-Taylor, J., *An Investigation of Feature Models for Music Genre Classification using the Support Vector Classifier*, International Conference on Music Information Retrieval, pp. 604-609, 2005

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%

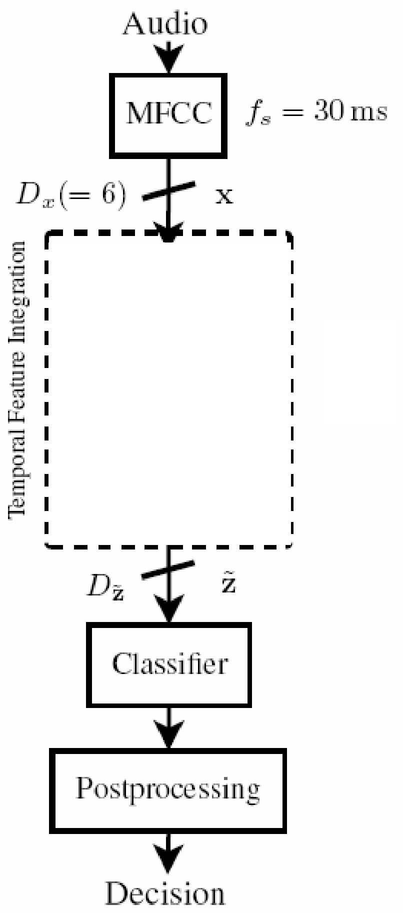
Best 11-genre confusion matrix

Alternative	41.8	6.4	4.5	3.6	3.6	2.7	8.2	2.7	4.5	3.6	18.2
Country	0.9	72.7	7.3	0.0	4.5	2.7	4.5	0.9	2.7	0.0	3.6
Easy-listening	1.8	11.8	61.8	2.7	4.5	2.7	2.7	0.0	2.7	3.6	5.5
Electronica	5.5	0.9	10.9	41.8	8.2	5.5	7.3	10.9	2.7	5.5	0.9
Jazz	0.9	4.5	8.2	10.9	50.0	2.7	3.6	2.7	7.3	6.4	2.7
Latin	3.6	8.2	2.7	4.5	3.6	37.3	8.2	8.2	4.5	11.8	7.3
Pop&Dance	6.4	9.1	6.4	9.1	0.9	11.8	43.6	2.7	3.6	2.7	3.6
Rap&Hiphop	0.0	0.0	0.9	7.3	0.9	4.5	3.6	62.7	1.8	17.3	0.9
RB&Soul	0.9	8.2	9.1	0.9	9.1	11.8	7.3	9.1	29.1	5.5	9.1
Reggae	0.9	0.9	0.0	3.6	4.5	5.5	1.8	17.3	3.6	61.8	0.0
Rock	25.5	16.4	5.5	0.9	5.5	2.7	6.4	0.0	6.4	1.8	29.1

11-genre human evaluation

	Alternative	Country	Easy-listening	Electronica	Jazz	Latin	Pop&Dance	Rap&Hiphop	RB&Soul	Reggae	Rock
Alternative	16.0	2.7	9.3	9.3	1.3	0.0	32.0	0.0	4.0	2.7	22.7
Country	5.3	54.7	9.3	0.0	4.0	1.3	9.3	0.0	4.0	0.0	12.0
Easy-listening	17.3	0.0	34.7	8.0	12.0	0.0	13.3	5.3	2.7	0.0	6.7
Electronica	5.3	0.0	0.0	54.7	1.3	0.0	32.0	1.3	4.0	1.3	0.0
Jazz	5.3	0.0	5.3	4.0	70.7	6.7	2.7	1.3	4.0	0.0	0.0
Latin	2.7	0.0	8.0	5.3	5.3	56.0	14.7	0.0	5.3	2.7	0.0
Pop&Dance	4.0	1.3	10.7	10.7	0.0	1.3	62.7	0.0	5.3	1.3	2.7
Rap&Hiphop	1.3	0.0	5.3	1.3	1.3	1.3	1.3	80.0	6.7	0.0	1.3
RB&Soul	2.7	1.3	13.3	1.3	2.7	0.0	14.7	0.0	57.3	2.7	4.0
Reggae	5.3	0.0	0.0	4.0	0.0	0.0	1.3	5.3	2.7	81.3	0.0
Rock	12.0	1.3	9.3	0.0	1.3	2.7	8.0	1.3	2.7	0.0	61.3

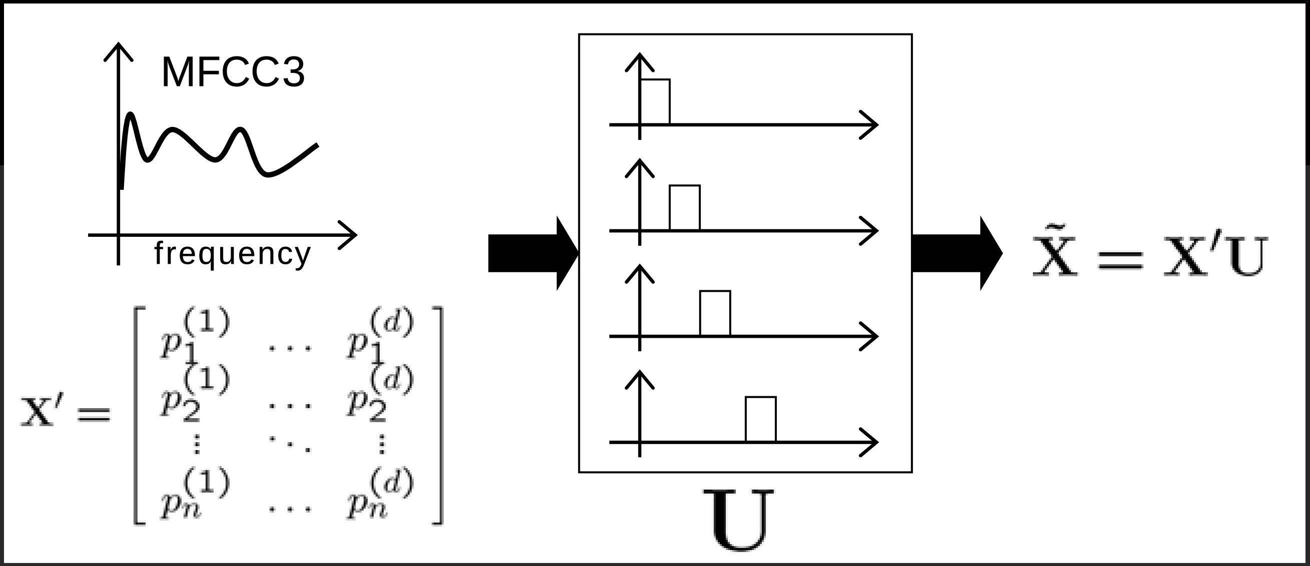
Supervised Filter Design in Temporal Feature Integration



Model the dynamics of MFCCs:

- Obtaining periodograms for each frame of 768ms MFCC
- “Bank-filter” these new features to obtain discriminative data

J. Arenas-Gacía, J. Larsen, L.H. Hansen, A. Meng: *Optimal filtering of dynamics in short-time features for music organization*, ISMIR 2006.



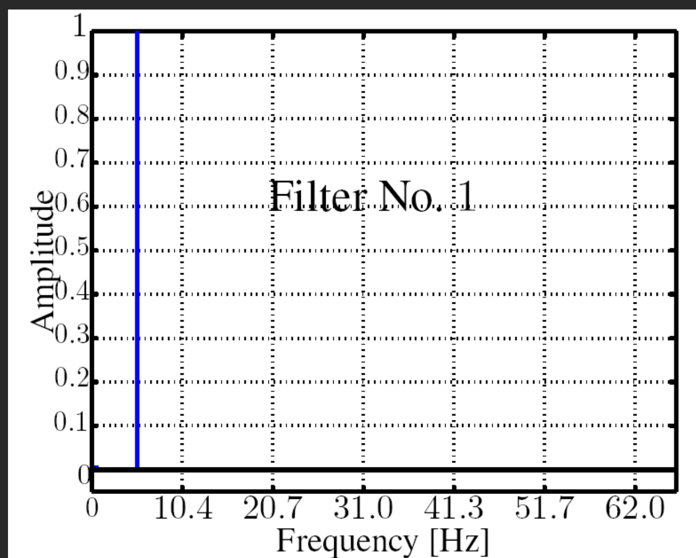
- Periodograms contain information about how fast MFCCs change
- A bank with 4 constant-amplitude was proposed for genre classification
  - 0 Hz : DC Value
  - 1 – 2 Hz : Beat rates
  - 3 – 15 Hz : Modulation energy (e.g., vibrato)
  - 20 – Fs/2 Hz : Perceptual Roughness
- Orthonormalized PLS can be used for a better design of this bank filter. Additional constraint U>0: Positive Constrained OPLS (POPLS)

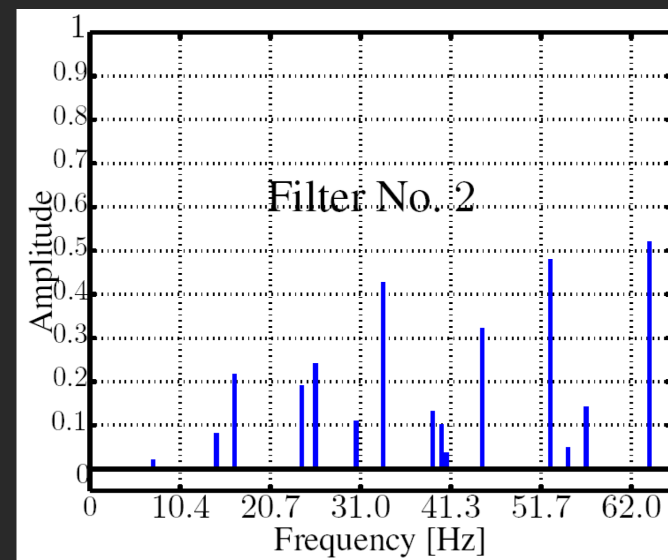
Illustrative example: vibrato detection

- 64 (32/32) AltoSax music snippets in Db3-Ab5
- Only the first MFCC was used

Vib

NonVib

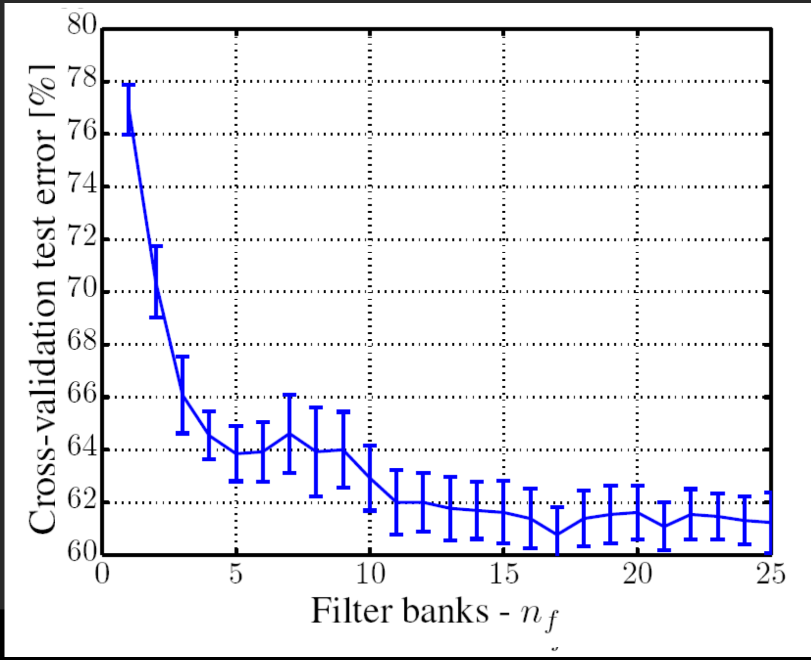




- Leave-one-out CV error: 9,4 % ($n_f = 25$); 20 % ($n_f = 2$) (Fixed filter bank: 48,3 %)

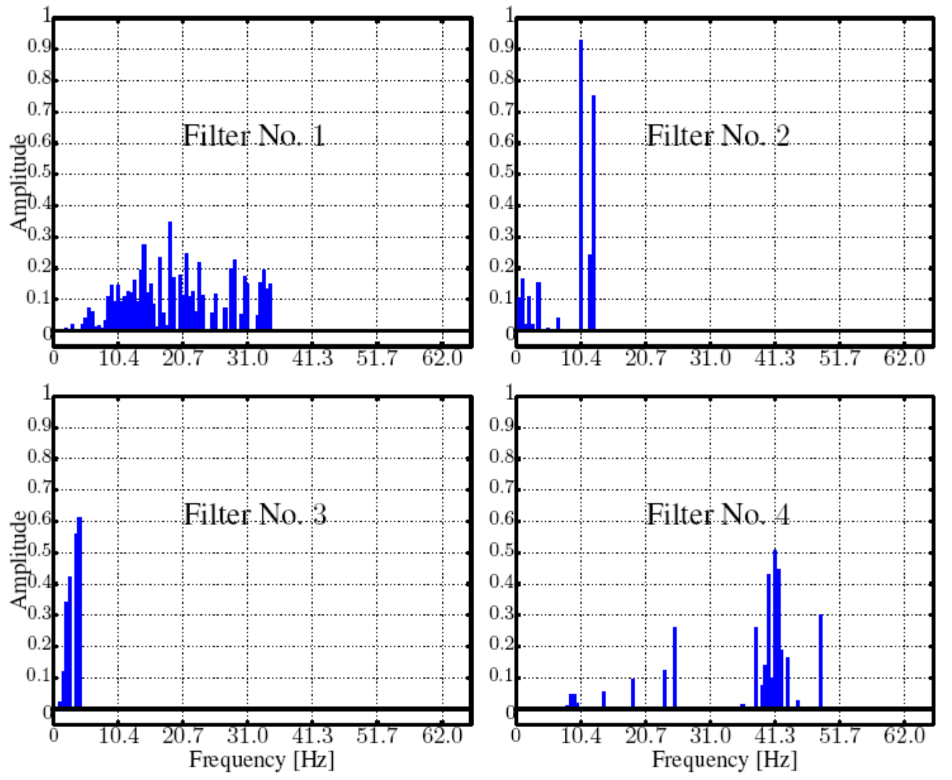
POPLS for genre classification

- 1317 music snippets (30 s) evenly distributed among 11 genres
- 7 MFCCs, but an unique filter bank



- POPLS 2% better on average compared to a fixed filter bank of four filter
- 10-fold cross-validation error falls to 61 % for n_f = 25

Interpretation of filters



- Filter 1: modulation frequencies of instruments
- Filter 2: lower modulation frequency + beat-scale
- Filter 4: perceptual roughness
- Consistent filters across 10-fold cross-validation
  - robustness to noise
  - relevant features for genre

Music separation

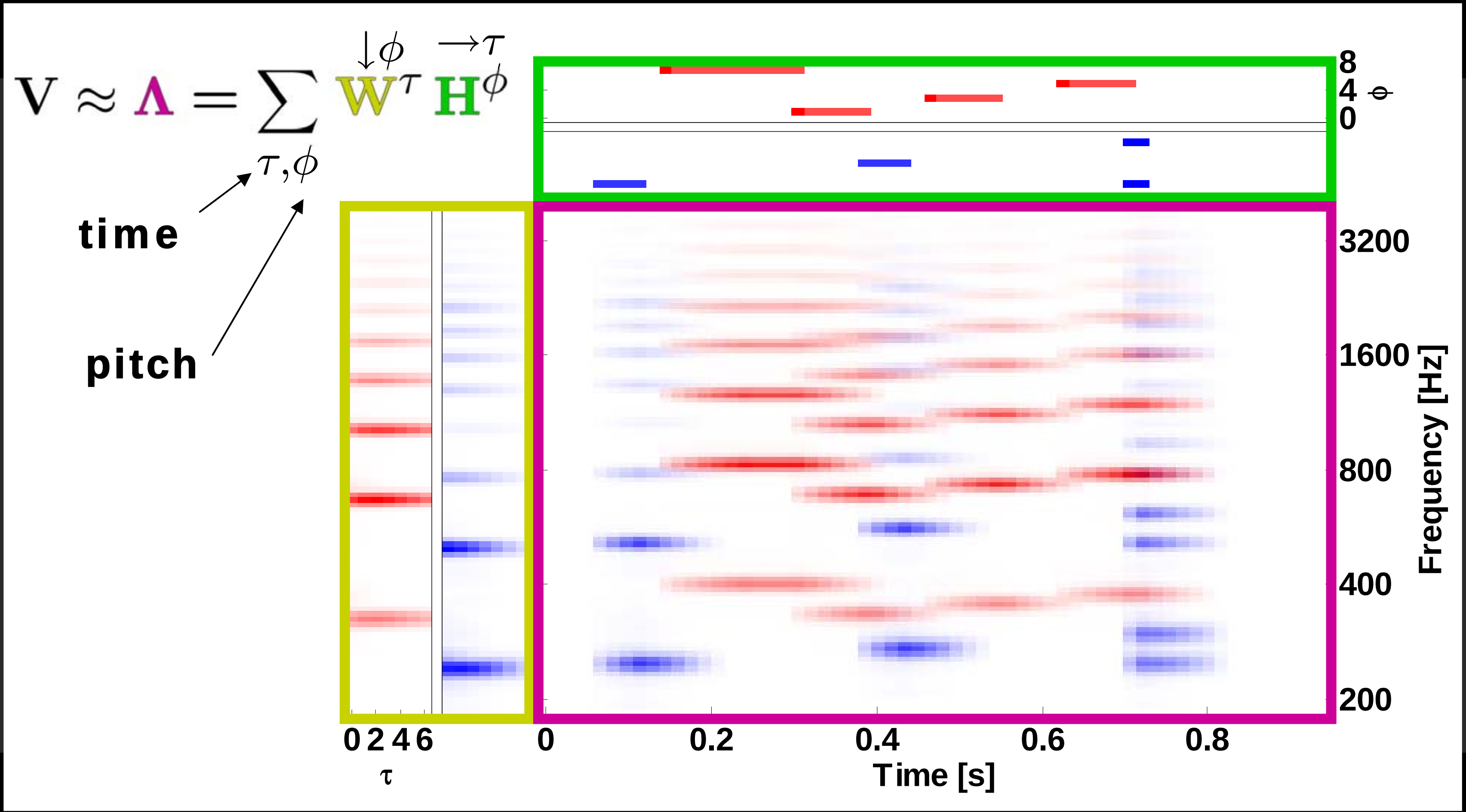
- A possible front end component for the music search framework
- Noise reduction
- Music transcription
- Instrument detection and separation
- Vocalist identification

Semi-supervised learning methods

Pedersen, M. S., Larsen, J., Kjems, U., Parra, L. C., *A Survey of Convolutive Blind Source Separation Methods*, Springer Handbook of Speech, Springer Press, 2007

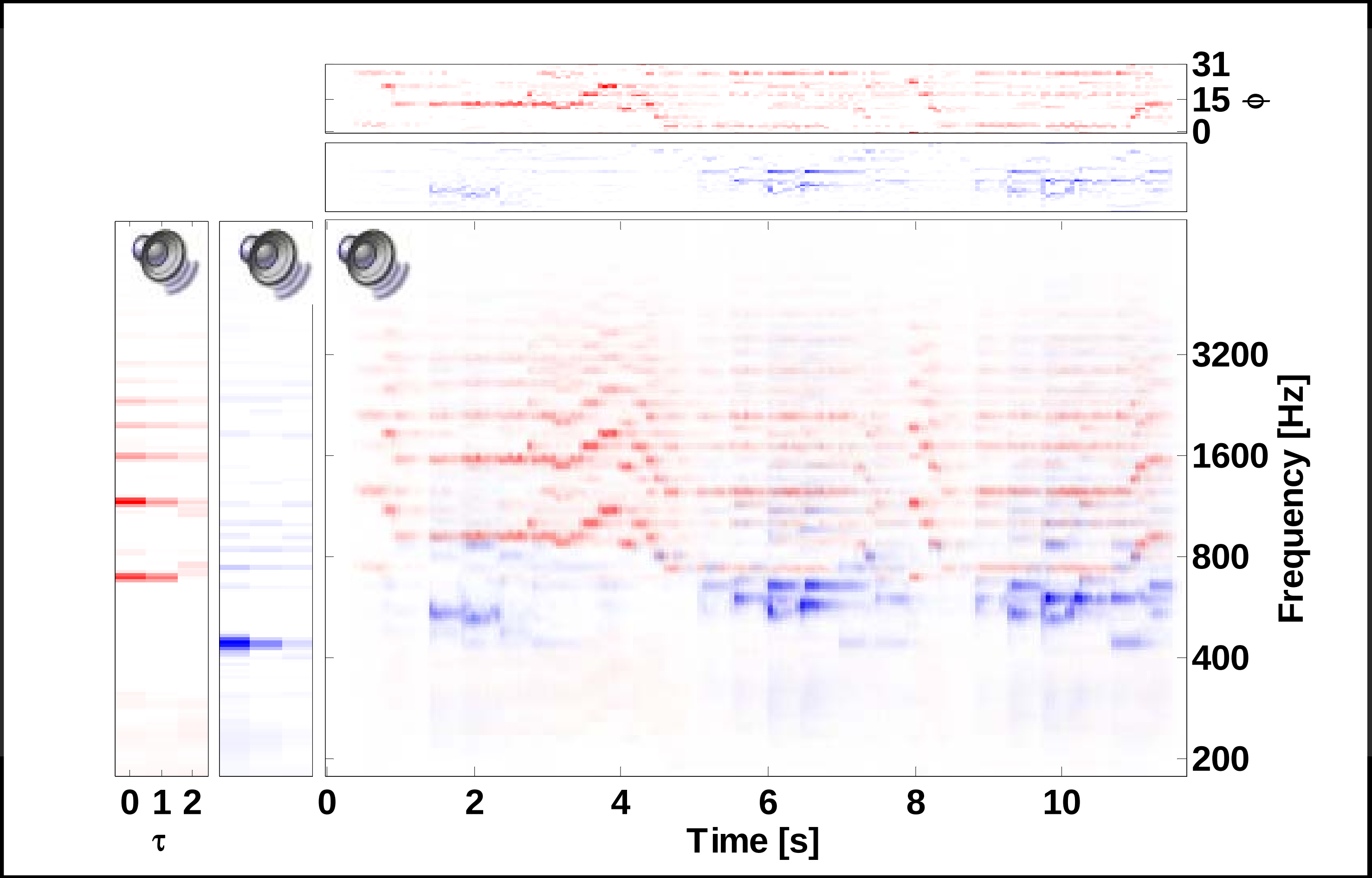
Nonnegative matrix factor 2D deconvolution

$$\mathbf{V} \approx \mathbf{\Lambda} = \sum_{\tau,\phi} \overset{\downarrow\phi}{\mathbf{W}}{}^{\tau} \overset{\rightarrow\tau}{\mathbf{H}}{}^{\phi}$$

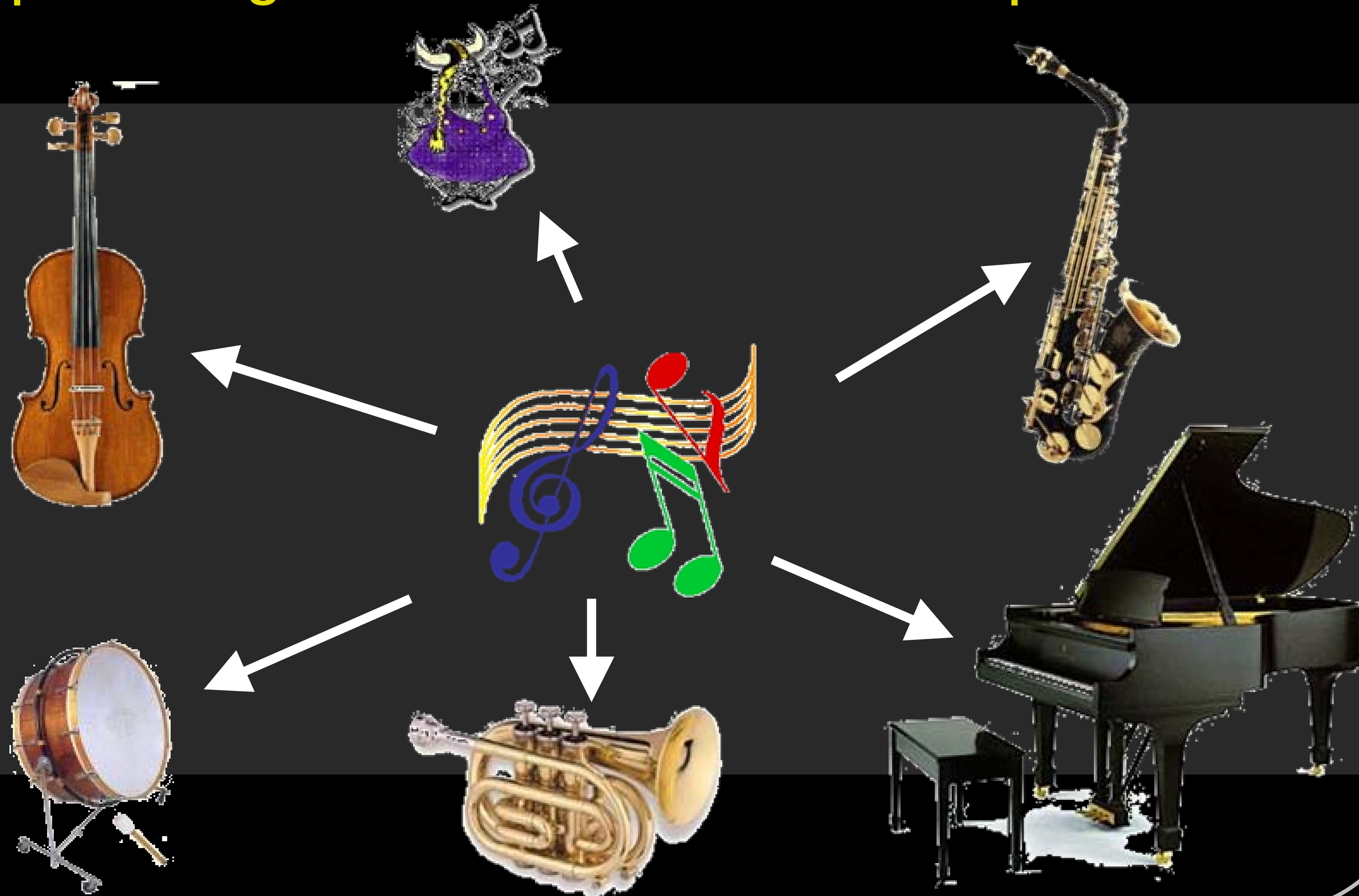


M. N. Schmidt, M. Mørup *Nonnegative Matrix Factor 2-D Deconvolution for Blind Single Channel Source Separation*, ICA2006, 2006. Demo also available.

Demonstration of the 2D convolutive NMF model



Separating music into basic components



Separating music into basic components

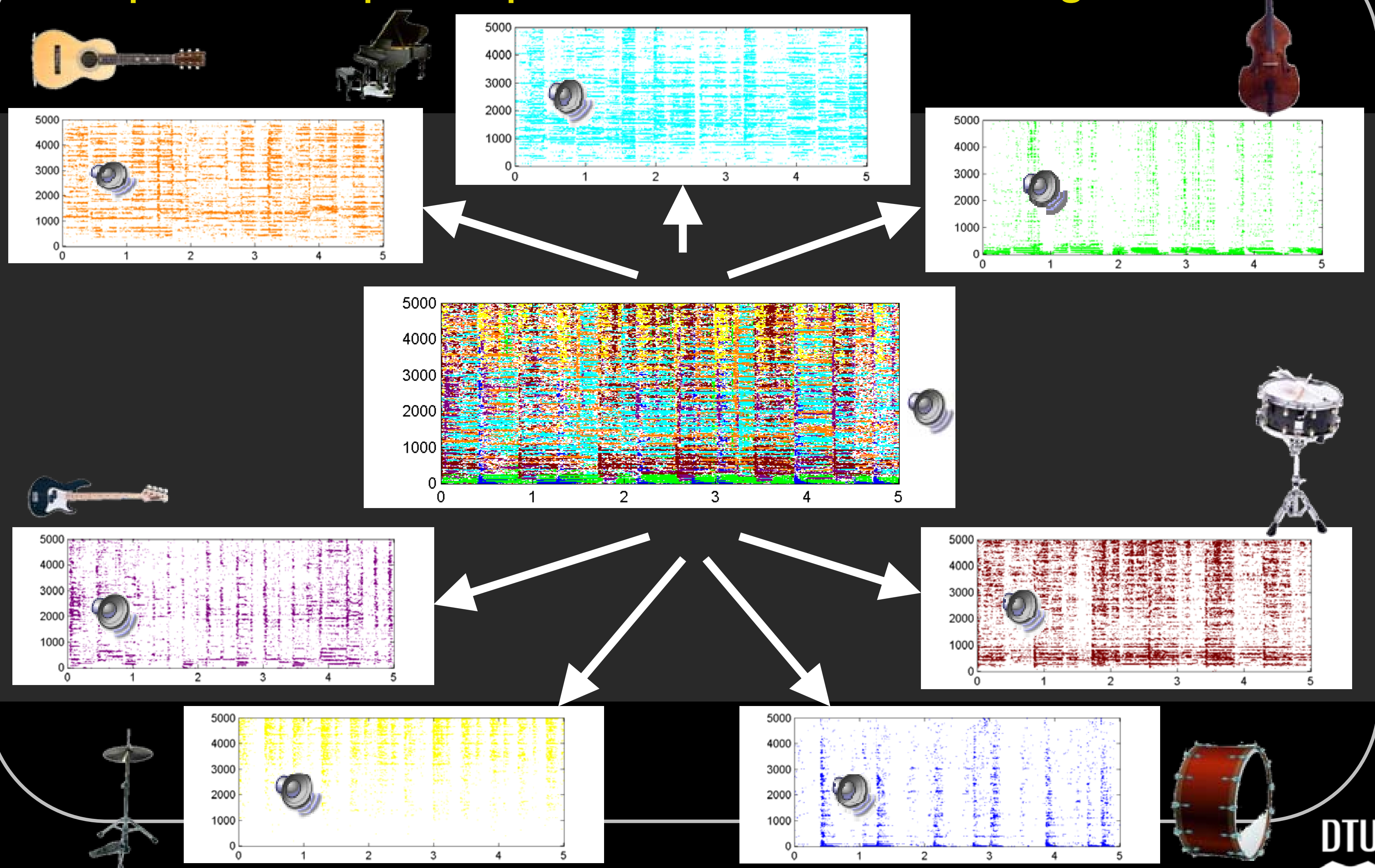
Combined ICA and masking

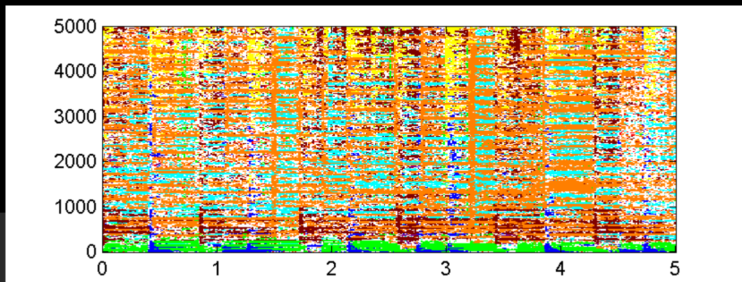
- Pedersen, M. S., Wang, D., Larsen, J., Kjems, U., Two-microphone Separation of Speech Mixtures, IEEE Transactions on Neural Networks, 2007
- Pedersen, M. S., Lehn-Schiøler, T., Larsen, J., *BLUES from Music: BLind Underdetermined Extraction of Sources from Music*, ICA2006, vol. 3889, pp. 392-399, Springer Berlin / Heidelberg, 2006
- Pedersen, M. S., Wang, D., Larsen, J., Kjems, U., *Separating Underdetermined Convolutive Speech Mixtures*, ICA 2006, vol. 3889, pp. 674-681, Springer Berlin / Heidelberg, 2006
- Pedersen, M. S., Wang, D., Larsen, J., Kjems, U., *Overcomplete Blind Source Separation by Combining ICA and Binary Time-Frequency Masking*, IEEE International workshop on Machine Learning for Signal Processing, pp. 15-20, 2005

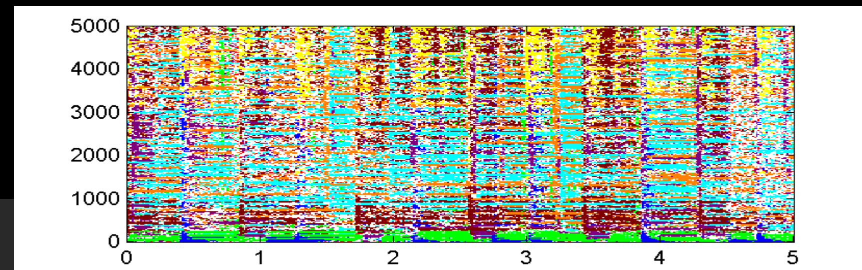
Assumptions

- 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

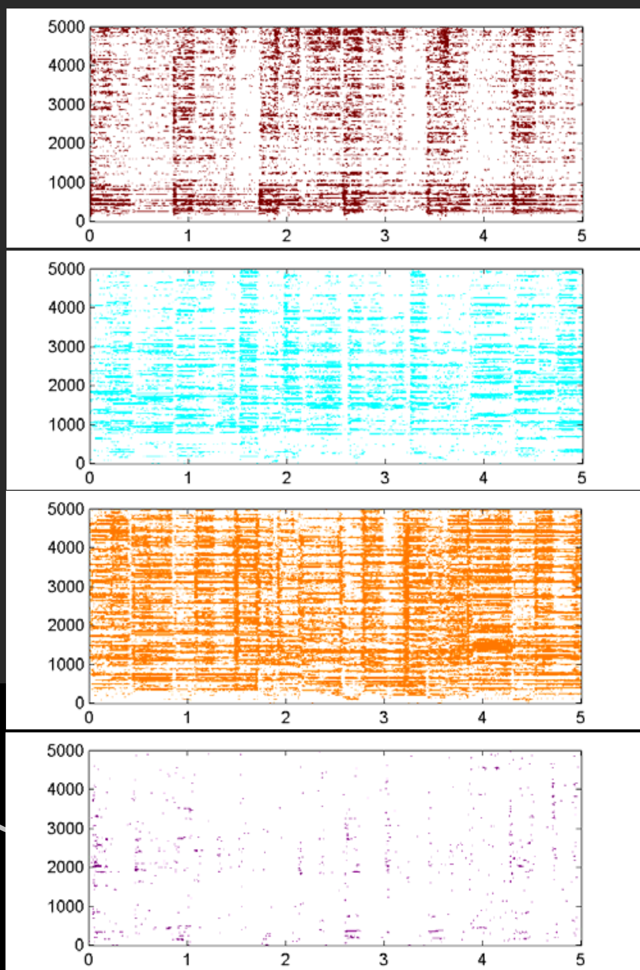


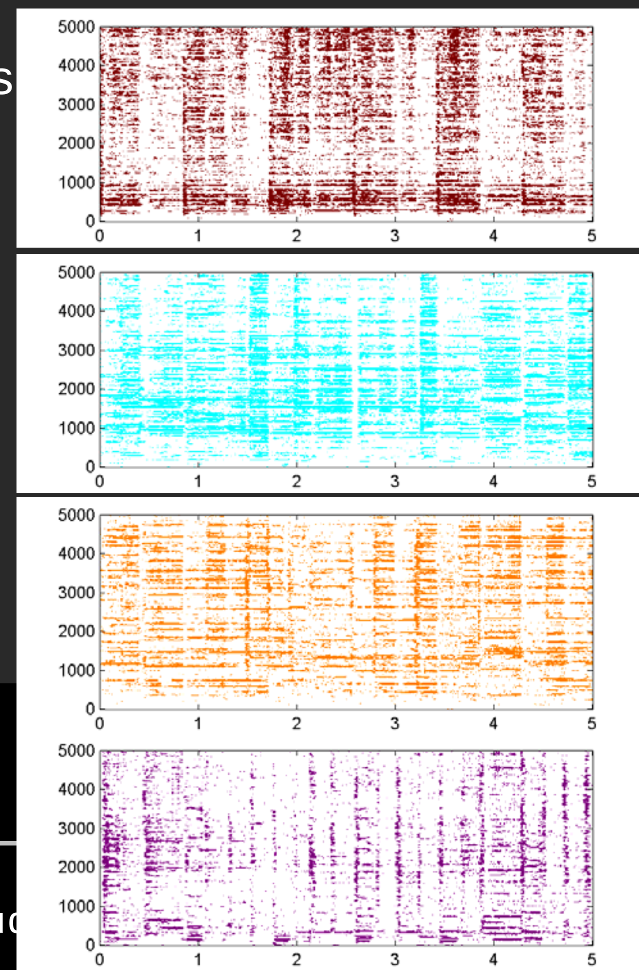




Stereo channel 1

Gain difference between channels





Stereo channel 2

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.,

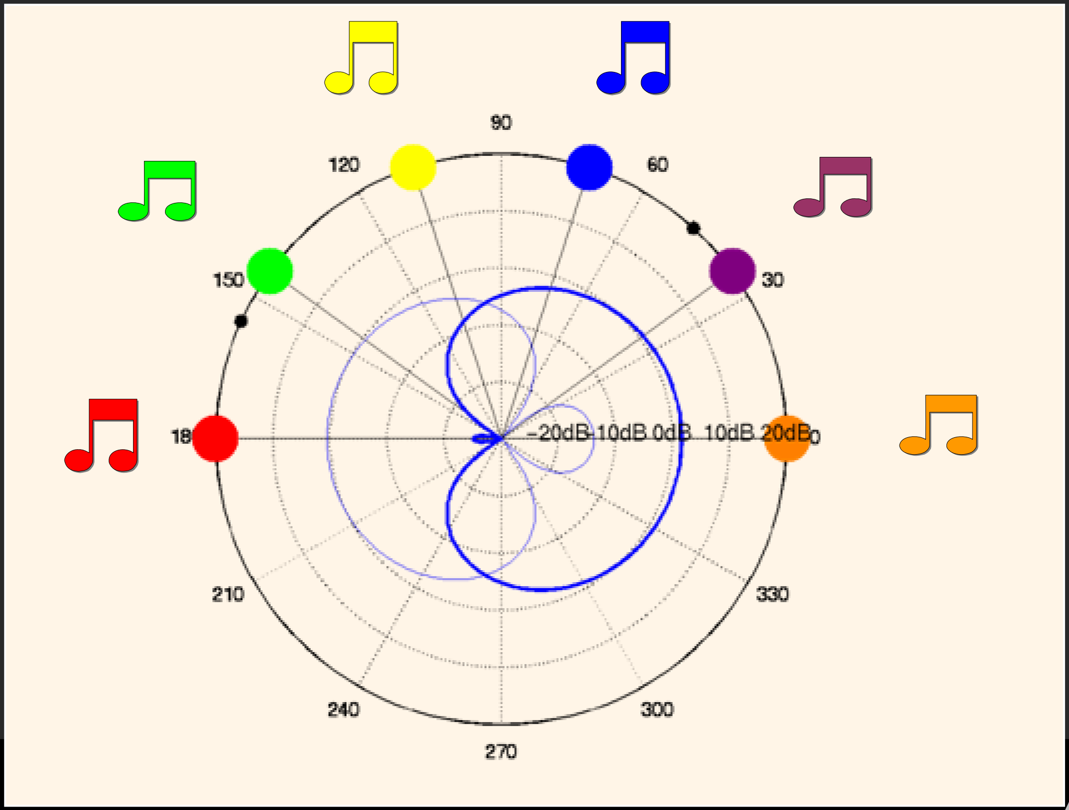
$$x = A(\theta_1, \ldots, \theta_N)s \qquad A(\theta) = \begin{bmatrix} r_1(\theta_1) & \cdots & r_1(\theta_N) \\ r_2(\theta_1) & \cdots & r_2(\theta_N) \end{bmatrix}$$

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

Direction dependent gain

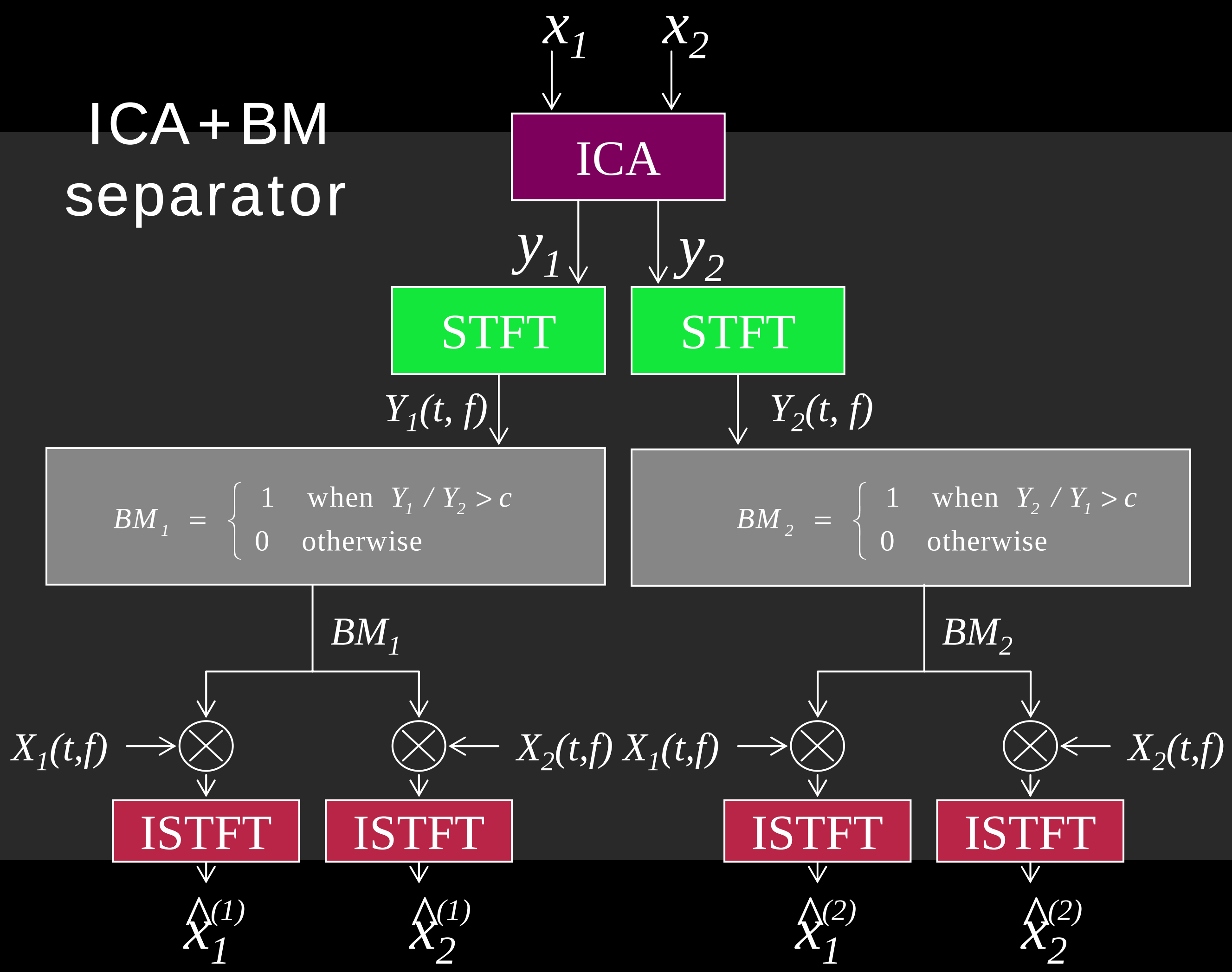
$$\mathbf{r}(\boldsymbol{\theta}) = 20\log|\mathbf{WA}(\boldsymbol{\theta})|$$

When **W** is applied, the two separated channels each contain a *group* of sources, which is as independent as possible from the other channel.

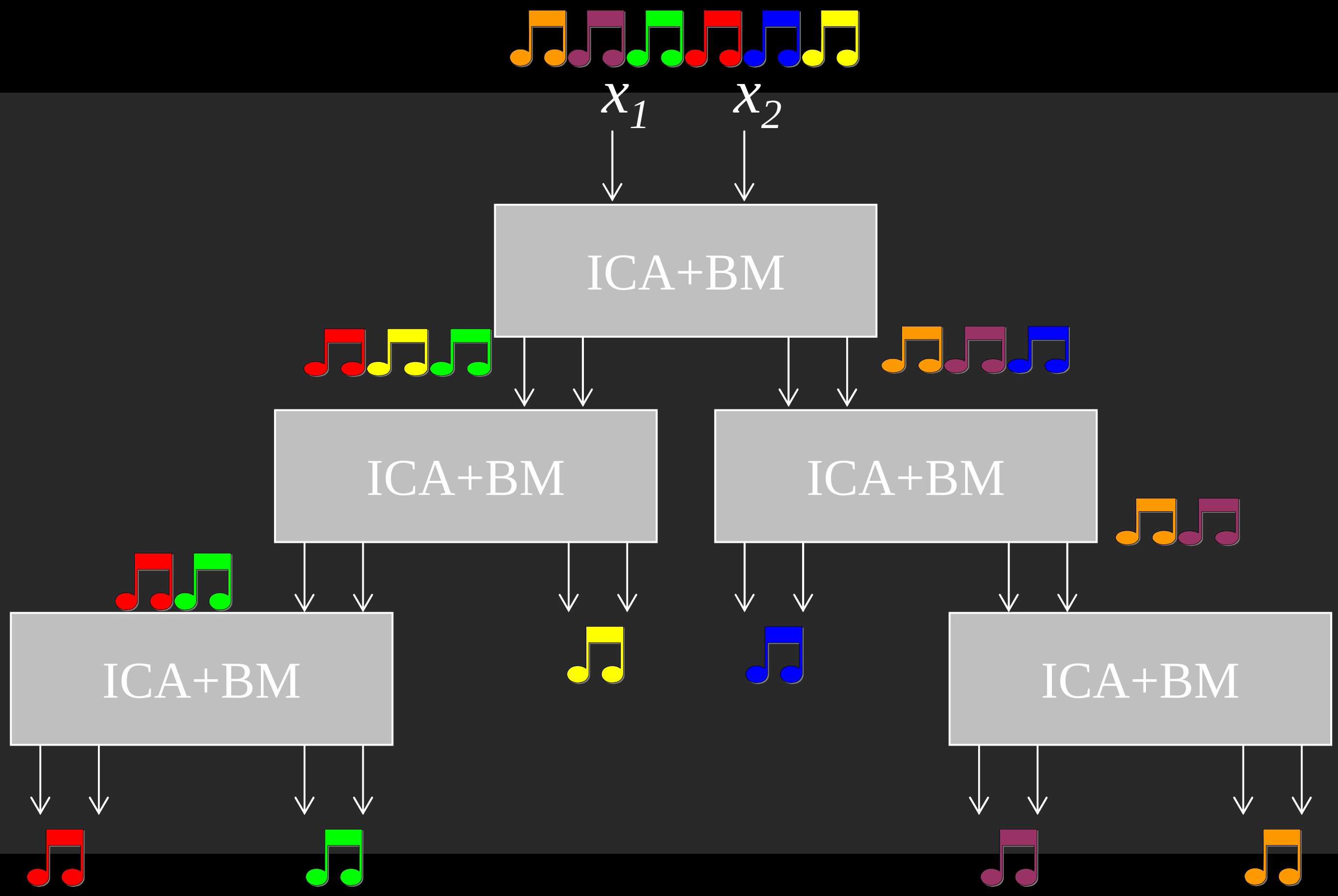


Combining ICA and T-F masking

ICA+BM separator

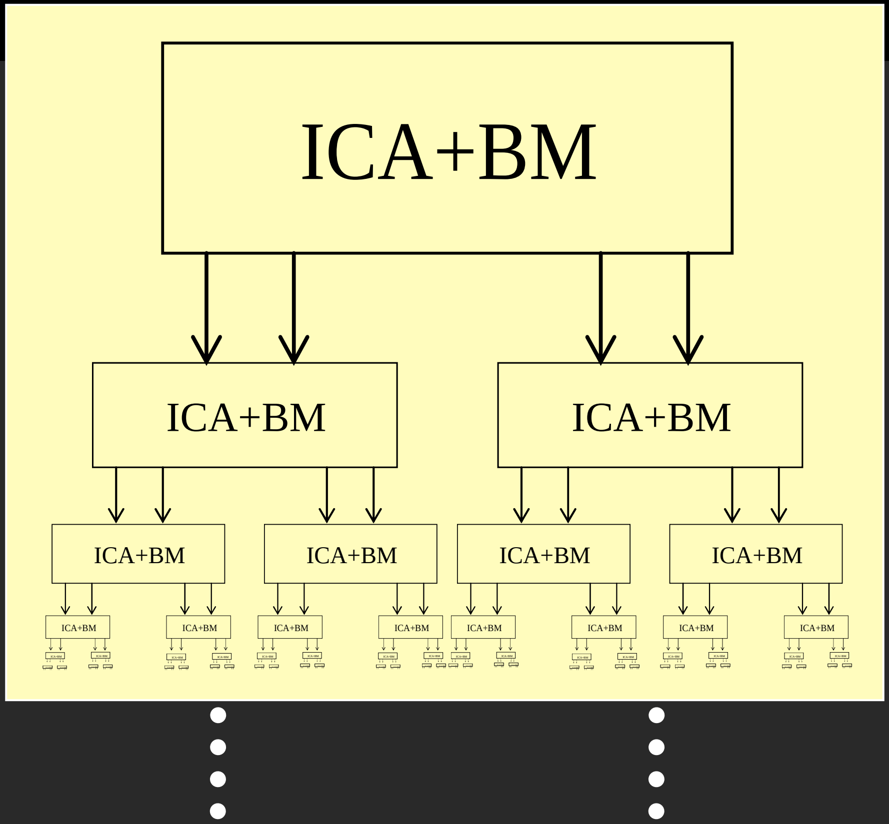


Method applied iteratively

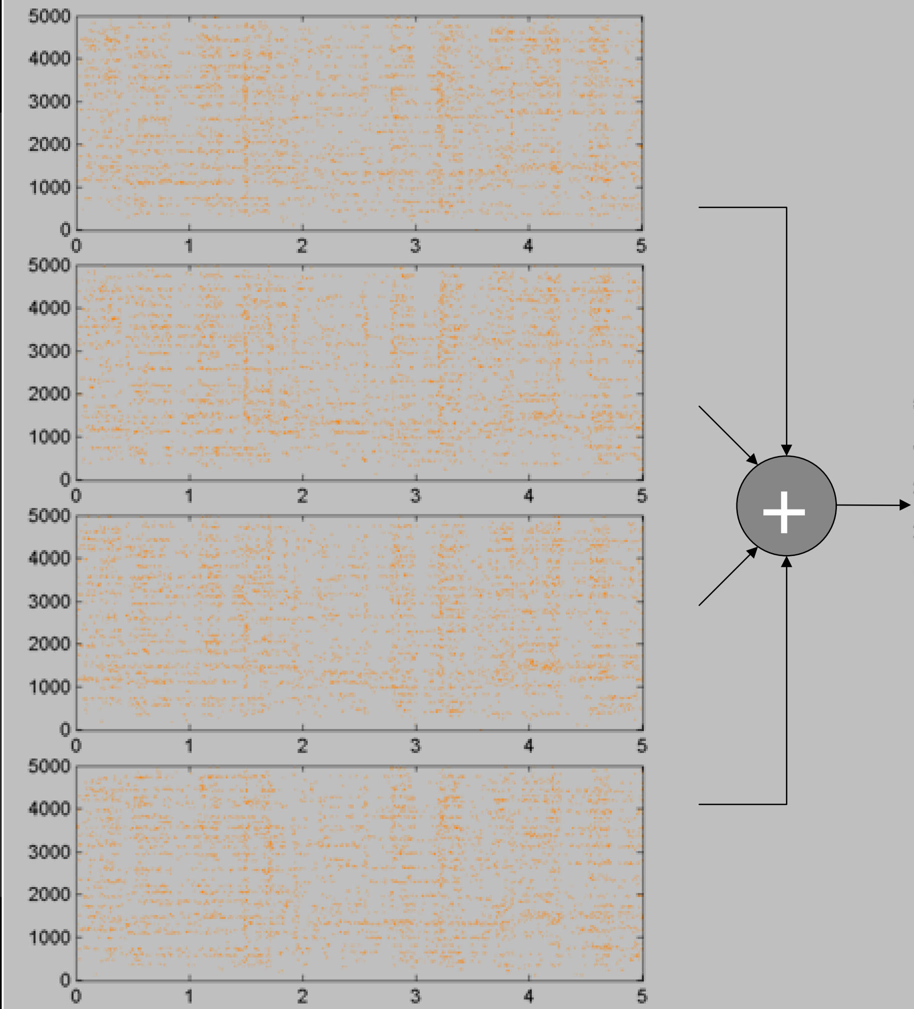


Improved method

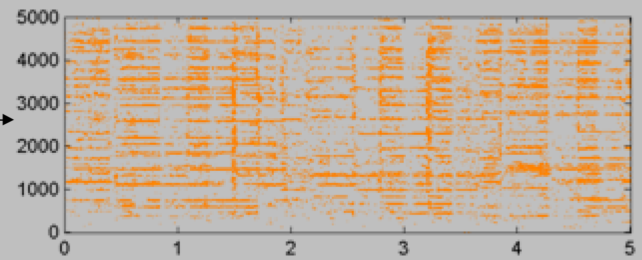
- The assumption of instantaneous mixing may not always hold
- Assumption can be relaxed
- Separation procedure is continued until very sparse masks are obtained
- Masks that mainly contain the same source are afterwards merged



Mask merging



If the signals are correlated (envelope), their corresponding masks are merged.



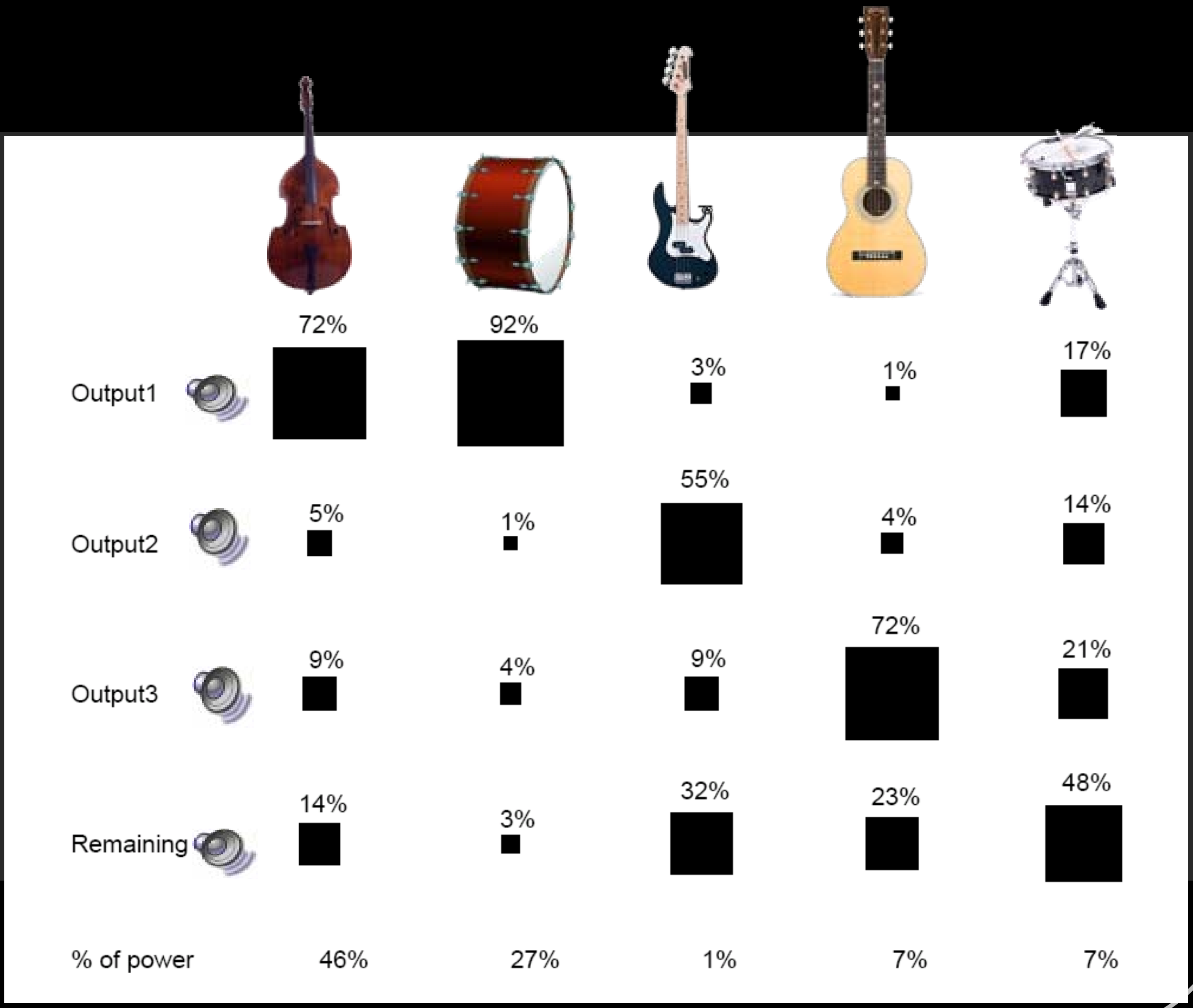
The resulting signal from the merged mask is of higher quality.

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 and code are available online via http://www.imm.dtu.dk

Results

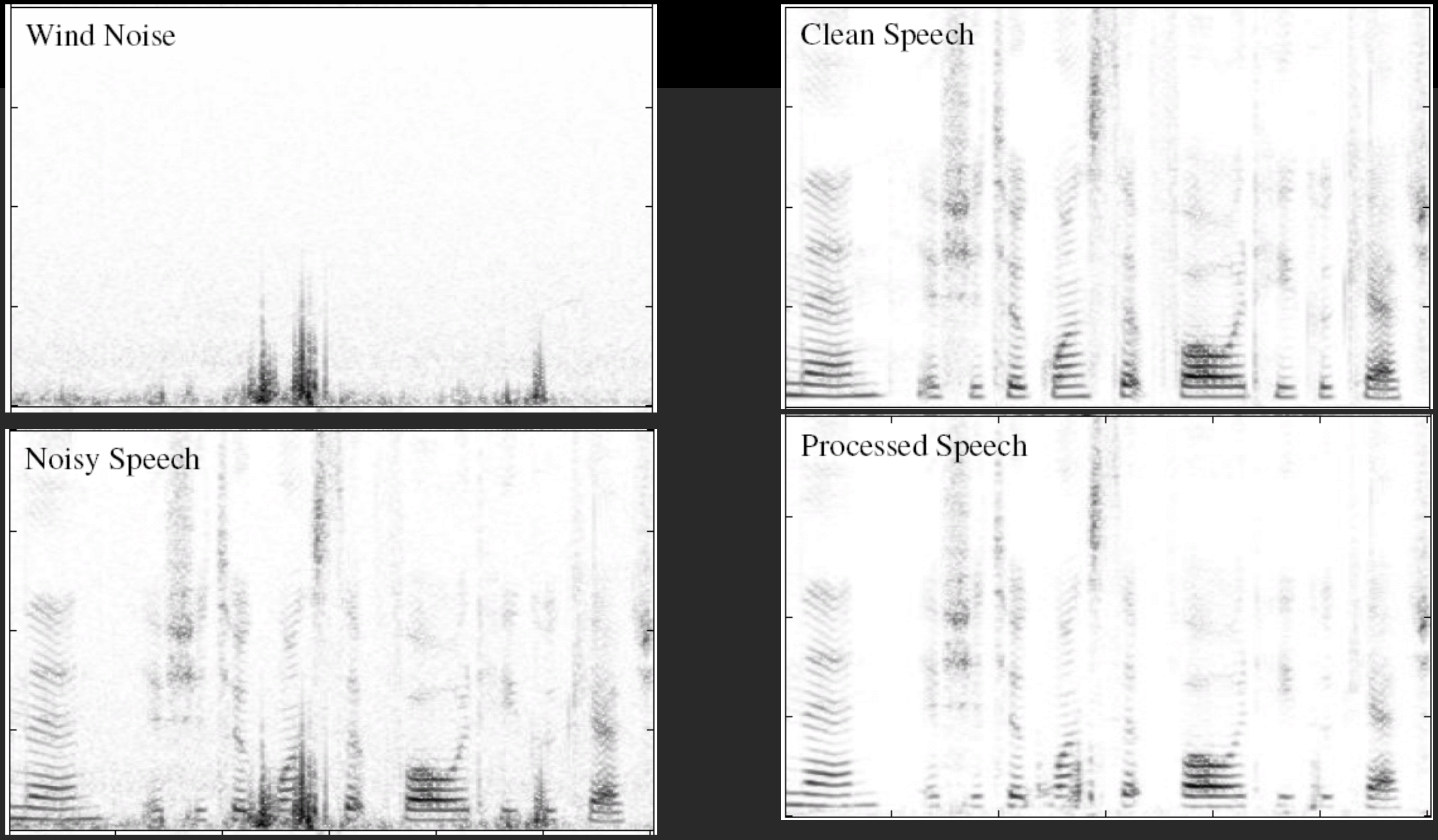
- The segregated outputs are dominated by individual instruments
- Some instruments cannot be segregated by this method, because they are not spatially different.



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



M.N Schmidt, J. Larsen, F.T. Hsiao: Wind noise reduction using non-negative sparse coding, 2007.

Sparse NMF decomposition

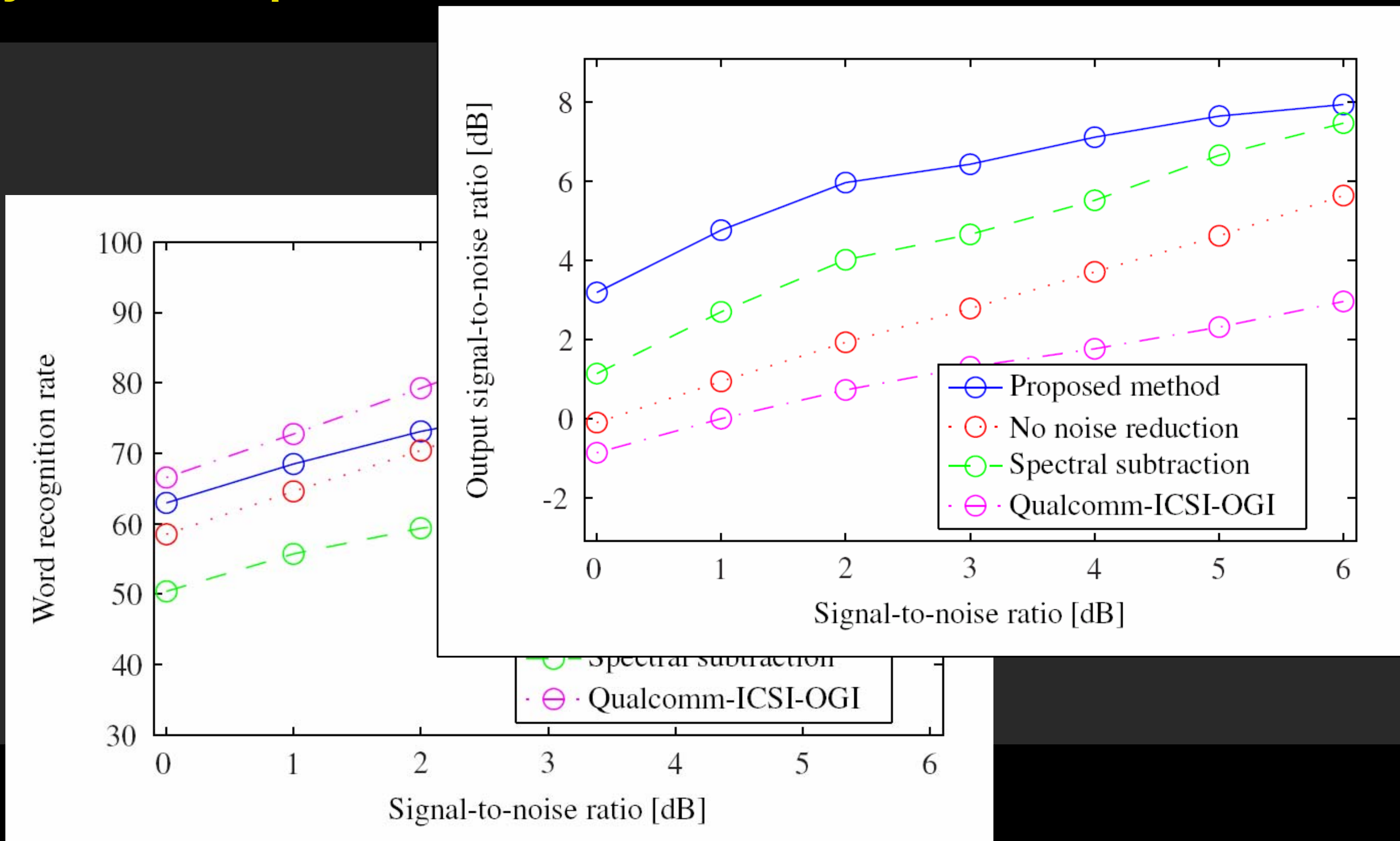
- Code-book (dictionary) of noise spectra is learned
- Can be interpreted as an advanced spectral subtraction technique

original

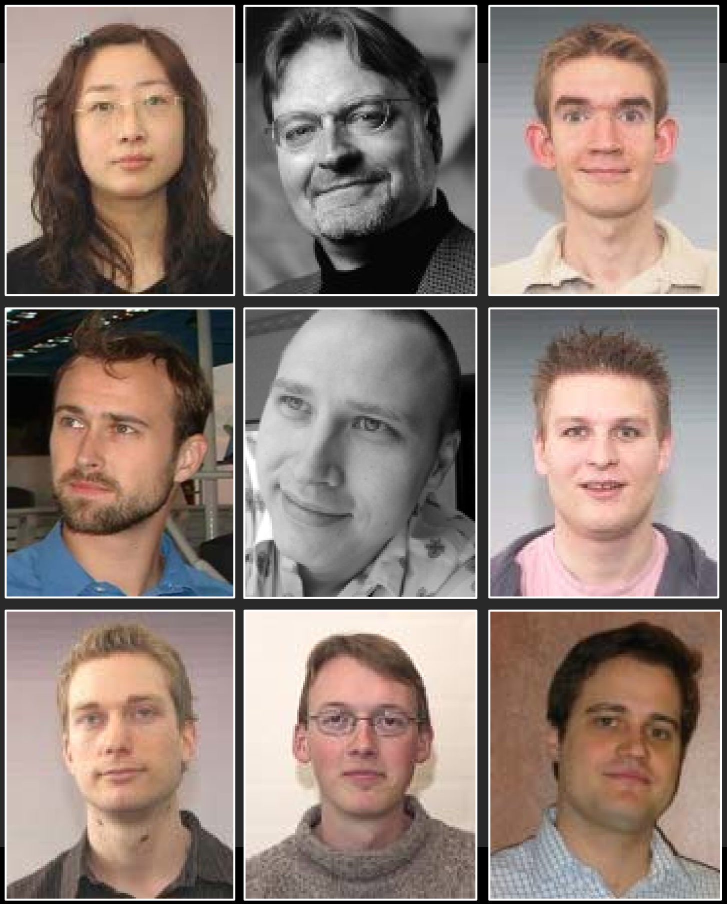
cleaned

alternative method (qualcom)

Objective performance



Summary



- Machine learning is, and will become, an important component in most real world applications
  - Semi-supervised learning
  - Sparse models and automatic model and featutre selection
  - Incorporation of high-level context description
  - User modeling
- Searching in massive amounts of heterogeneous enhances “productivity” simply important to ….quality of life…
- Machine learning is essential for search – in particular mapping low level data to high description levels enabling human interpretation
- Music and audio separation combines unsupervised methods ICA/MNF with other SP and supervised techniques