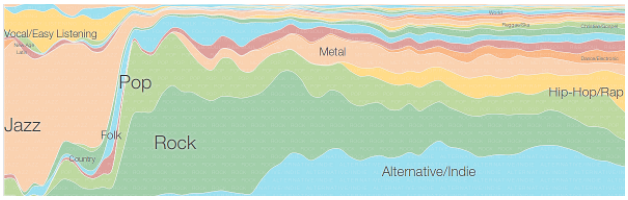
Music Information Retrieval (MIR)

Anders Friberg
Speech, Music and Hearing, CSC, KTH
https://www.kth.se/profile/afriberg/
afriberg@kth.se



Outline

- Automatic analysis of music
  - Background
  - General model
  - Examples of specific techniques
  - Current project at KTH
  - Appendix: Some tools & applications

Music analysis: why?

- Digital music + Internet = explosion of music availability (songs on Spotify > 30 m)
- Practical problems
  - Categorize (user labels too subjective)
  - Search (even without metadata)
  - Enforce copyrights
- More scientific reasons
  - Understand musical communication and perception

Music Information Retrieval (MIR)

- Interdisciplinary subject (engineering, psychoacoustics, music, social sciences, ...)
- Rapidly growing community around the Int. Conf. on Music Information Retrieval (ISMIR, www.ismir.net)
- Strong interest from information and entertainment industries

MIR applications

- MIR systems
  - Classification (genre/style/mood)
  - Recommendation/Playlist generation
  - Content-based querying
  - Summarization/Fingerprinting
  - Transcription/Score following
  - ... and many more!

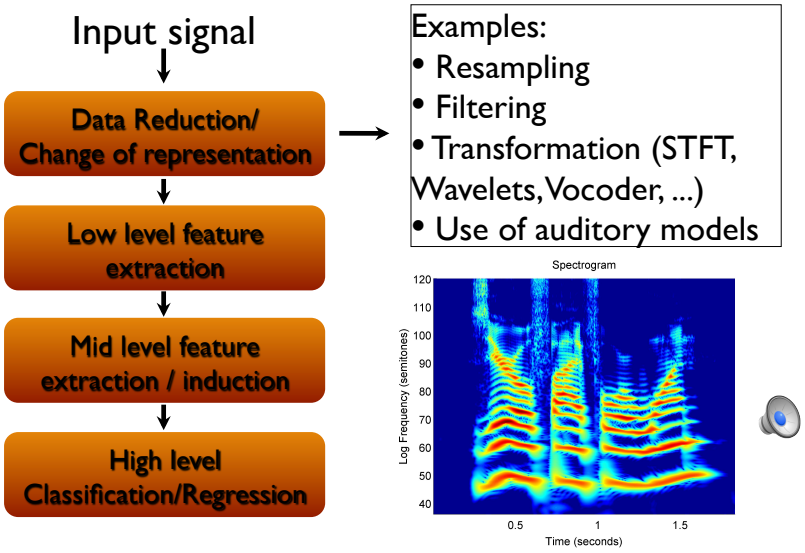
MIR meta-data

- All auxiliary information:
  - Lyrics
  - Publication data (artist , recording date, genre, etc)
  - Expert labels
  - User tags (“sounds like …, “groovy”, …”)
  - Collected listening habits

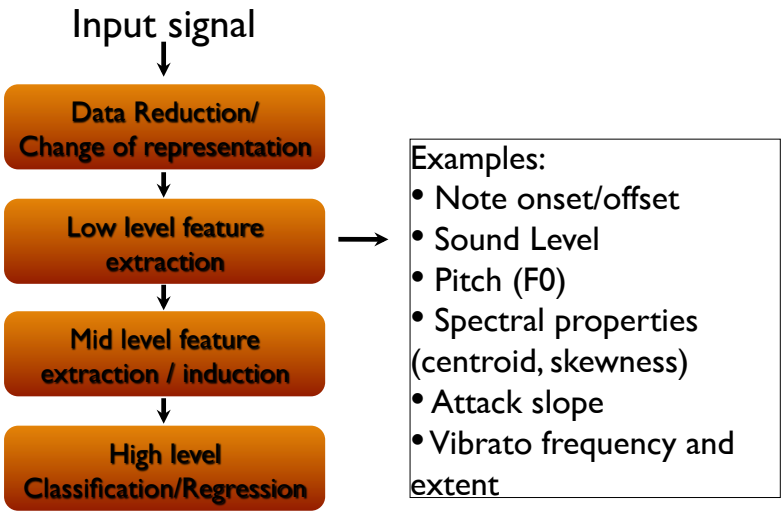
MIR content analysis (focus here)

- Melody and harmony
- Rhythm, beat, tempo and form
- Timbre, instrumentation and voice
- Genre, style, mood
- Performance

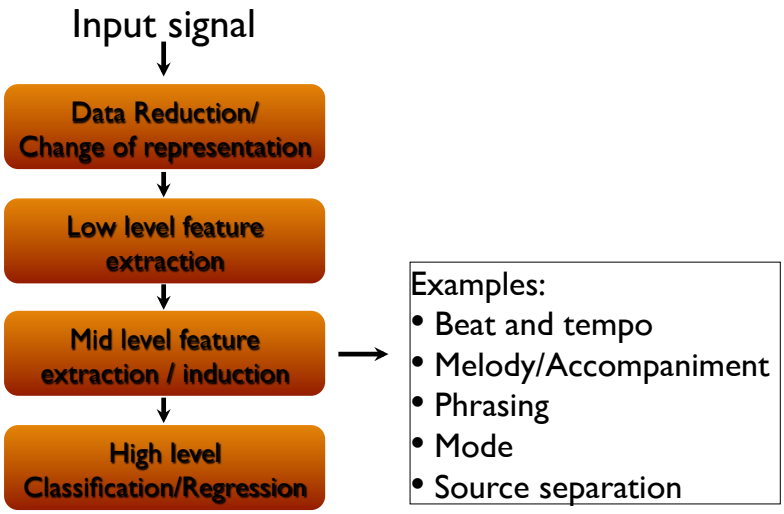
General analysis model



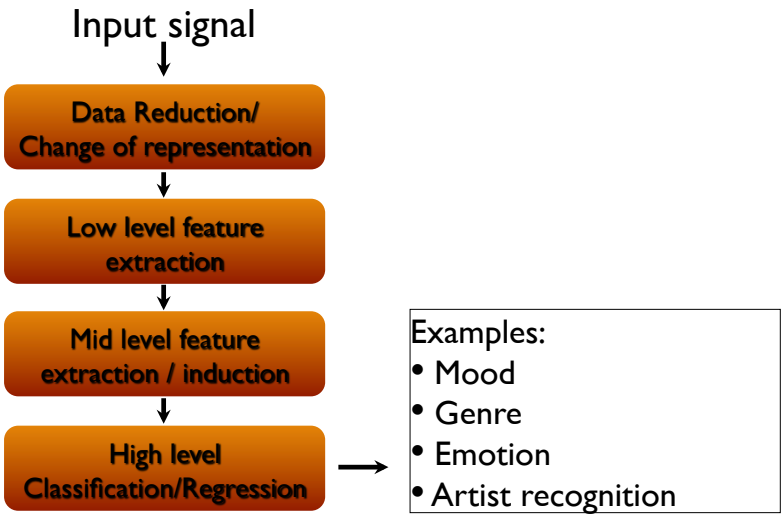
General analysis model



General analysis model



General analysis model



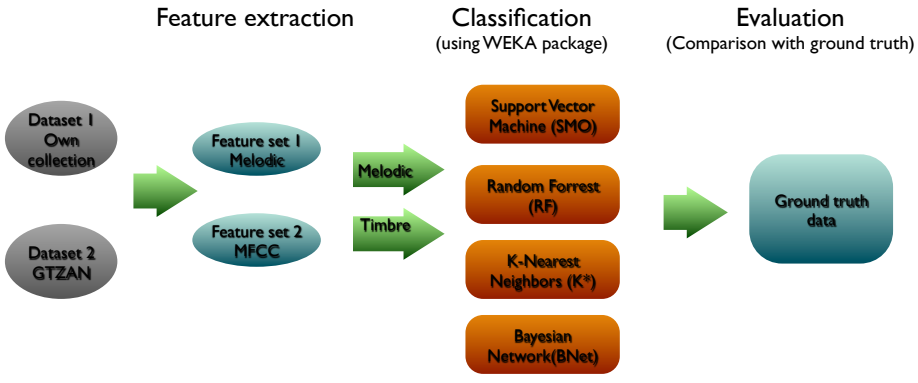
Case study:
Genre classification from audio

- Useful for categorization and playlist generation
- A classic MIR problem
- Often some ground truth already available

Genres from audio

1. Choose relevant features
2. Create a ground-truth data set
3. Extract acoustical features
4. Train the system
5. Test the system

Procedure



Salamon, J., Rocha, B. & Gomez, E. (2012) Musical Genre Classification Using Melody Features Extracted From Polyphonic Music Signals, ICASSP 2012.

Melodic features extracted from the dominant melody:

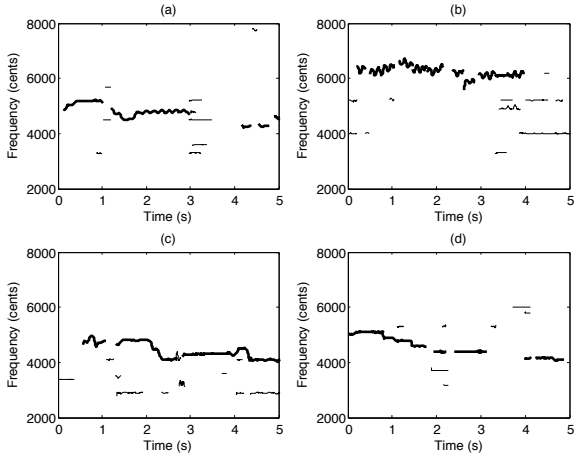


Fig. 1. Pitch contours extracted from excerpts of different genres: vocal jazz (a), opera (b), pop (c) and instrumental jazz (d).

Results own dataset

opera, pop, flamenco, vocal jazz and instrumental jazz

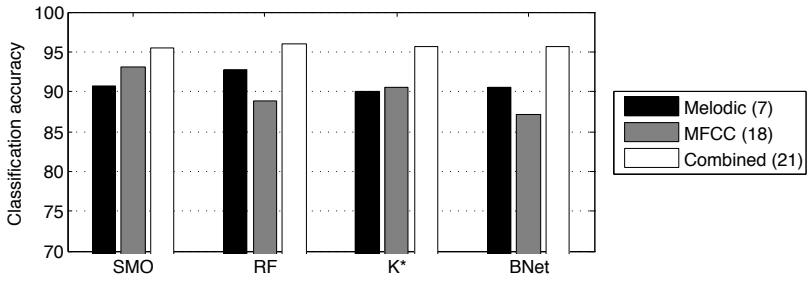


Fig. 5. Classification results for the extended 500 excerpt data-set.

Results external dataset
10 genres

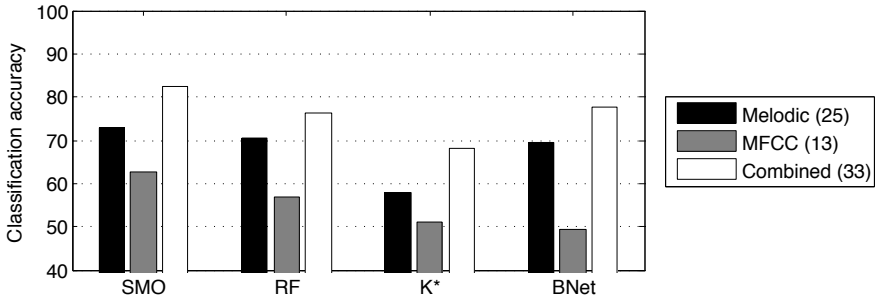


Fig. 6. Classification results for the GTZAN data-set.

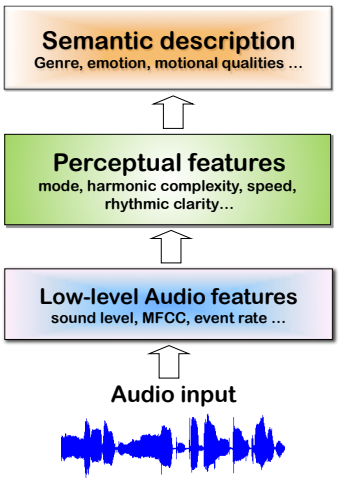
Conclusions

- A large number of features has been used and is available in toolboxes
- Similarly with prediction methods
- Results very good for selected datasets (>90% correct prediction)
- However, for large music collections the methods still don't have a accuracy required for successful commercial applications (≈50-70%)

Future improvements – two alternative paths:

- Alternative 1: Develop and use more advanced machine learning and let the system learn also intermediate levels from data
- Alternative 2: Develop features that corresponds better to human perception

Current research project at KTH:
Using perceptually derived features in music information retrieval



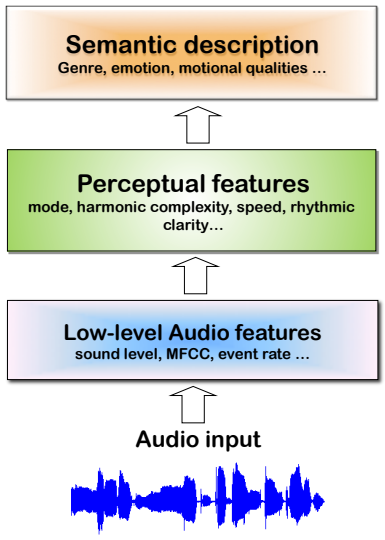
Perceptually determined features

- Our long-term aim: Try to understand which features we use when we listen to music in a casual way.
- Current method: Derive features perceptually in listening tests and then try to model them

Friberg, A., Schoonderwaldt, E., Hedblad, A., Fabiani, M., & Elowsson, A. (2014). Using listener-based perceptual features as intermediate representations in music information retrieval. *Journal of the Acoustical Society of America*, 136(4), 1951-1963.

Specific research questions

- Can we reliably estimate perceptual features in listening experiments?
- Can semantic descriptions (emotions) be modelled from perceptual features?
- Can we make computational models of perceptual features?



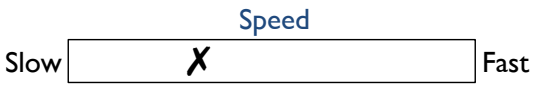
Selected Perceptual features

- Speed (slow-fast) The general speed of the music disregarding any musical analysis such as the tempo.
- Rhythmic clarity (flowing-firm) Indication of how well the rhythm is accentuated (c.f. Lartillot et al., 2008).
- Rhythmic complexity (simple-complex) A companion to rhythmic clarity.
- Articulation (staccato-legato) The duration of tones.
- Modality (minor-major) Modality as a continuous scale.
- Overall Pitch (low-high) The overall pitch height of the music.
- Harmonic complexity (simple-complex) A measure of how complex the harmonic progression is.
- Dissonance (consonant-dissonant) (exp.3)
- Dynamics (soft-loud) The played dynamic level.
- Brightness (dark-bright) (exp. 1). Timbre (exp. 2)

Method

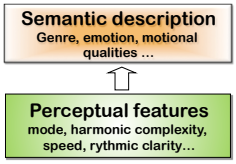
- All features rated on semi-continuous scales in 3 experiments with about 20 subjects each:
- Experiment 1 - Ringtones. 100 ringtones selected from pilot experiment regarding spread in features, both audio and MIDI.
- Experiment 2 – Film music clips. 110 film clips provided by U. Jyväskylä selected from pilot experiment regarding emotional expression, only audio.
- Experiment 3 – K-pop. 98 examples of different Korean pop genres provided by U. Illinois, only audio.

- Prediction methods: Linear regression, Partial Least-Square regression (PLS), Support Vector Regression (SVR)



Accuracy of the mean estimation: Cronbach's alpha

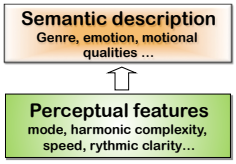
	Experiment 1 Ring tones	*Experiment 2 Film clips*	*Experiment 3 K-Pop*
Feature	*alpha*	*alpha*	*alpha*
Speed	**0.98**	**0.97**	**0.98**
Rhythmic complexity	0.89	**0.91**	0.80
Rhythmic clarity	**0.90**	**0.95**	0.85
Articulation	**0.93**	**0.97**	**0.95**
Dynamics	**0.93**	**0.93 (0.95)**	
Modality	**0.93 (0.94)**	**0.96**	0.85
Harmonic complexity	0.83	0.85 (0.87)	0.68
Dissonance			**0.92**
Pitch	**0.93**	**0.94**	0.88
Brightness/Timbre	0.88	**0.90 (0.91)**	



Predicting emotion ratings from perceptual features using linear regression

	Experiment 1		*Experiment 2*							
	Energy	*Valence*	*Energy*	*Valence*	*Tension*	*Anger*	*Fear*	*Happiness*	*Sadness*	*Tenderness*
R^2	0.94	0.90	0.92	0.80	0.80	0.74	0.67	0.83	0.77	0.65
Adjusted R^2	0.93	0.88	0.91	0.78	0.79	0.72	0.64	0.81	0.75	0.62
Feature	sr^2	sr^2	sr^2	sr^2	sr^2	sr^2	sr^2	sr^2	sr^2	sr^2
Speed	0.36***	0.09*	0.14***	0.10*				0.10*		
Rhy.comp.										
Rhy.clarity	0.08**			0.10*			(-)0.13*			
Articulation		0.07*	0.11***	(-)0.10*	0.15**		0.17**		0.18***	(-)0.18**
Dynamics	0.20***	(-)0.13***	0.39***	(-)0.27***	0.37***	0.50***	0.25***		(-)0.18***	(-)0.37***
Modality	0.10**	0.49***	0.13***	0.27***	(-)0.18***			0.37***	(-)0.44***	0.17**
Harm.comp.				(-)0.21***	0.21***	0.17**	0.30***	0.10*	(-)0.22***	
Pitch			0.07*							
Brightness Timbre		0.10**	(-)0.12***	0.15**	(-)0.16***	(-)0.15**	(-)0.23***		0.11*	

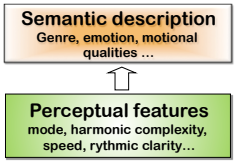
Adjusted R^2 – overall explained variation
sr^2 - semipartial correlation coefficient - the independent contribution of each feature
p-values: * < 0.05; ** < 0.01, ***<0.001.



Overall prediction power

	Experiment 1		*Experiment 2*							
	Energy	***Valence***	***Energy***	***Valence***	***Tension***	***Anger***	***Fear***	***Happiness***	***Sadness***	***Tenderness***
R²	0.94	0.90	0.92	0.80	0.80	0.74	0.67	0.83	0.77	0.65
Adjusted R²	**0.93**	**0.88**	**0.91**	**0.78**	**0.79**	**0.72**	**0.64**	**0.81**	**0.75**	**0.62**
Feature	***sr²***	***sr²***	***sr²***	***sr²***	***sr²***	***sr²***	***sr²***	***sr²***	***sr²***	***sr²***
Speed	0.36***	0.09*	0.14***	0.10*				0.10*		
Rhy.comp.										
Rhy.clarity	0.08**			0.10*			(-)0.13*			
Articulation		0.07*	0.11***	(-)0.10*	0.15**		0.17**		0.18***	(-)0.18**
Dynamics	0.20***	(-)0.13***	0.39***	(-)0.27***	0.37***	0.50***	0.25***		(-)0.18***	(-)0.37***
Modality	0.10**	0.49***	0.13***	0.27***	(-)0.18***			0.37***	(-)0.44***	0.17**
Harm.comp.				(-)0.21***	0.21***	0.17**	0.30***	0.10*	(-)0.22***	
Pitch			0.07*							
Brightness		0.10**								
Timbre			(-)0.12***	0.15**	(-)0.16***	(-)0.15**	(-)0.23***		0.11*	

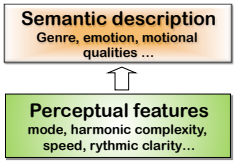
Adjusted R^2 – overall explained variation
sr^2 - semipartial correlation coefficient - the independent contribution of each feature
p-values: * < 0.05; ** < 0.01, ***<0.001.



Energy

	Exp 1 Ring tones		*Exp. 2 Film clips*	
	Energy	***Valence***	***Energy***	***Valence***
R²	0.94	0.90	0.92	0.80
Adjusted R²	**0.93**	**0.88**	**0.91**	**0.78**
Feature	***sr²***	***sr²***	***sr²***	***sr²***
Speed	0.36***	0.09*	0.14***	0.10*
Rhy.comp.				
Rhy.clarity	0.08**			0.10*
Articulation		0.07*	0.11***	(-)0.10*
Dynamics	0.20***	(-)0.13***	0.39***	(-)0.27***
Modality	0.10**	0.49***	0.13***	0.27***
Harm.comp.				(-)0.21***
Pitch			0.07*	
Brightness		0.10**		
Timbre			(-)0.12***	0.15**

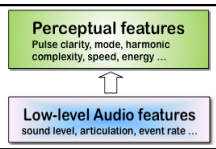
Adjusted R^2 – overall explained variation
sr^2 - semipartial correlation coefficient - the independent contribution of each feature
p-values: * < 0.05; ** < 0.01, ***<0.001.



Valence

	Exp 1 Ring tones		*Exp. 2 Film clips*	
	Energy	***Valence***	***Energy***	***Valence***
R²	0.94	0.90	0.92	0.80
Adjusted R²	0.93	0.88	0.91	0.78
Feature	***sr²***	***sr²***	***sr²***	***sr²***
Speed	0.36***	0.09*	0.14***	0.10*
Rhy.comp.				
Rhy.clarity	0.08**			0.10*
Articulation		0.07*	0.11***	(-)0.10*
Dynamics	0.20***	(-)0.13***	0.39***	(-)0.27***
Modality	0.10**	0.49***	0.13***	0.27***
Harm.comp.				(-)0.21***
Pitch			0.07*	
Brightness		0.10**		
Timbre			(-)0.12***	0.15**

Adjusted R^2 – overall explained variation
sr^2 - semipartial correlation coefficient - the independent contribution of each feature
p-values: * < 0.05; ** < 0.01, ***<0.001.



Prediction of perceptual features from audio features using existing toolboxes

	Exp. 1 Ringtones				**Exp. 2 Film clips**			
Perceptual Feature	*Audio features*	*PLS comp.*	*R^2 PLS*	*R^2 SVR*	*Audio features*	*PLS comp.*	*R^2 PLS*	*R^2 SVR*
Speed	14	3	0.55	0.61	17	2	0.35	0.51
	45	2	0.51	0.54	49	2	0.39	0.57
Rhythmic complex.								
Rhythmic clarity	2	1	0.14	0.00	2	1	0.23	0.30
	45	1	0.25	0.20	49	2	0.34	0.40
Articulation	45	2	0.40	0.34	49	2	0.47	0.48
Dynamics	25	3	0.31	0.34	25	3	0.61	0.74
	45	3	0.45	0.45	49	3	0.61	0.67
Modality	3	2	0.34	0.30	3	2	0.47	0.47
	45	2	0.21	0.08	49	4	0.53	0.52
Pitch								
Timbre					25	4	0.39	0.35
					49	4	0.41	0.35
Brightness	25	2	0.10	0.03				
	45	2	0.22	0.09				

R^2 squared correlation coefficient (explained variation)
PLS Partial Least-square Regression
SVR Support Vector Regression
10-fold cross validation.

Modelling the Speed of Music Using Features from Harmonic/Percussive Separated Audio (ISMIR 2013)

A. Elowsson, A. Friberg, G. Madison, J. Paulin

- ***Ground truth:*** Listener ratings of speed for a set of music examples

Method

- Harmonic/Percussive separation
- Extract audio features such as
  - Onset densities
  - Spectral flux
  - Tempo

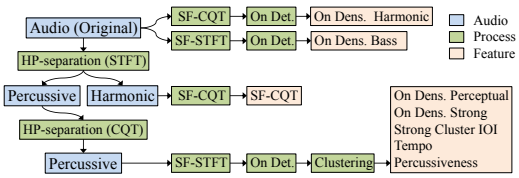
- Map listener ratings to audio features in MLR

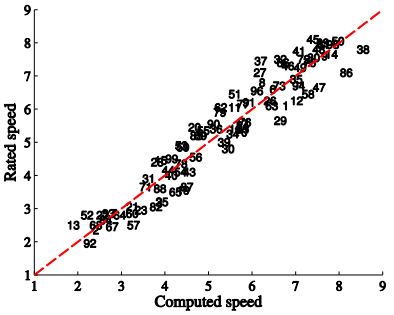
Results

- 93,4 % explained variance (R^2) in independent test set

Application Area

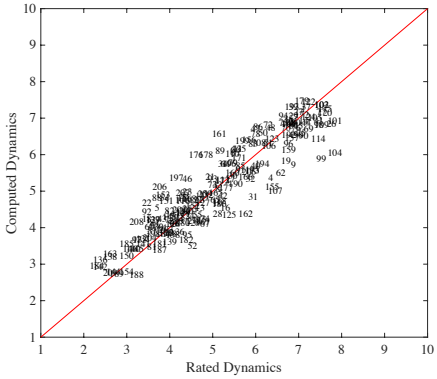
- Modelling high-level features (e.g. valence)
- Finding correct tempo octave

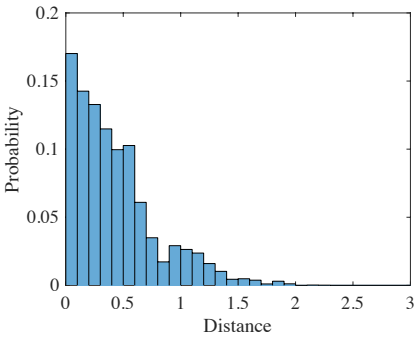




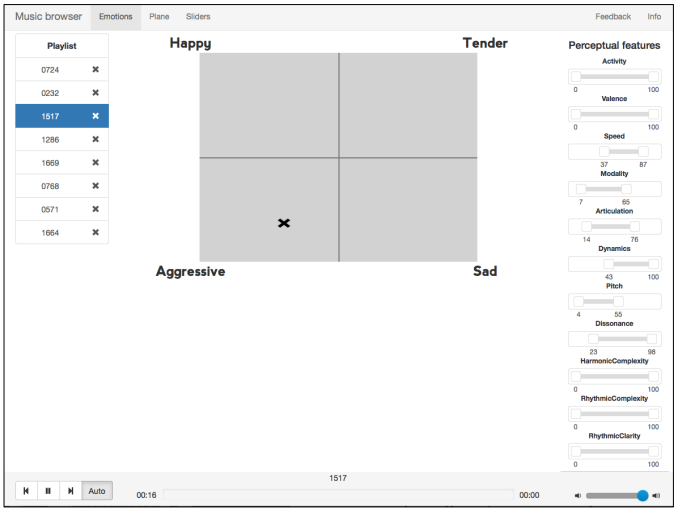
Modeling Dynamics

$R^2 = 84.0$





Music browsing using the perceptual features - an experimental prototype



http://musicdiscovery.se

Richard Nysäter (2016) Master thesis, KTH (forthcoming)

Evaluators are welcome!

Appendix

Available tools

- Analysis
  - MIR Toolbox (Univ. of Jyväskylä)
  - CUEX (TMH - KTH)
  - SonicVisualizer (Queen Mary Univ. London)
  - ... and more

- Analysis and synthesis
  - Marsyas (G. Tzanetakis, Univ. Victoria, CA)
  - CLAM (UPF, Barcelona)
- Auditory models
  - Auditory toolbox (M. Slaney, MATLAB)
  - IPEM (Gent Univ., real-time pD version)
- Score analysis
  - MelodicMatch (Univ. of Melbourne)

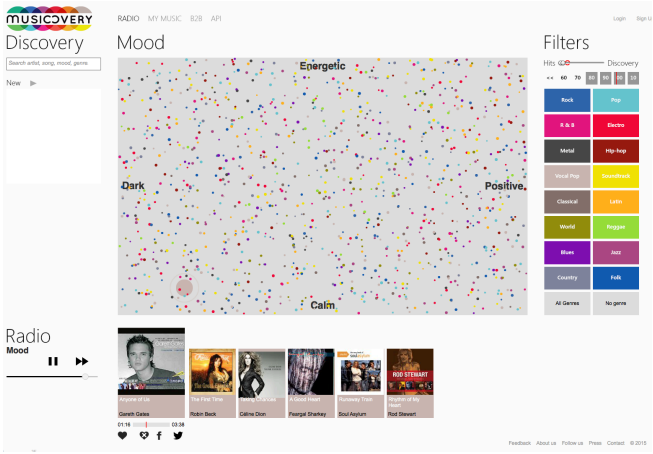
Databases

- Million song dataset
  - http://labrosa.ee.columbia.edu/millionsong/
  - Preanalysed including many audio features and labels (280 GB data)
  - Audio snippets (30s) can be retrieved
- MIREX competition
- List of datasets: http://grh.mur.at/sites/default/files/mir_datasets_0.html (updated 2009)

Popular applications

- Shazam (QBE)
- Midomi (QBH)
- Pandora, last.fm, ... (recommendation)
- Tangerine! (playlist generation)
- Musicovery (mood playlist & recomm.)
- ... suggestions?

Musicovery mood radio

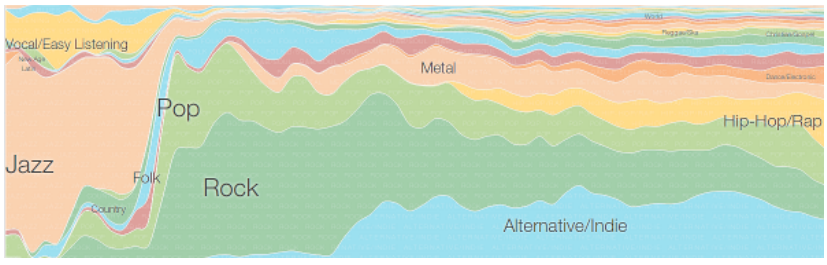


Analysis based on 40 hand-annotated labels by experts

http://musicovery.com/

Example of meta-data analysis: Google play graphs

Meta-data analysis of what Google play users have in their collections



Release year

http://research.google.com/bigpicture/music