Creating meaning in audio and music signals

Jan Larsen, Associate Professor PhD
Cognitive Systems Section
Dept. of Applied Mathematics and Computer Science
Technical University of Denmark
janla@dtu.dk, www.compute.dtu.dk/~jl

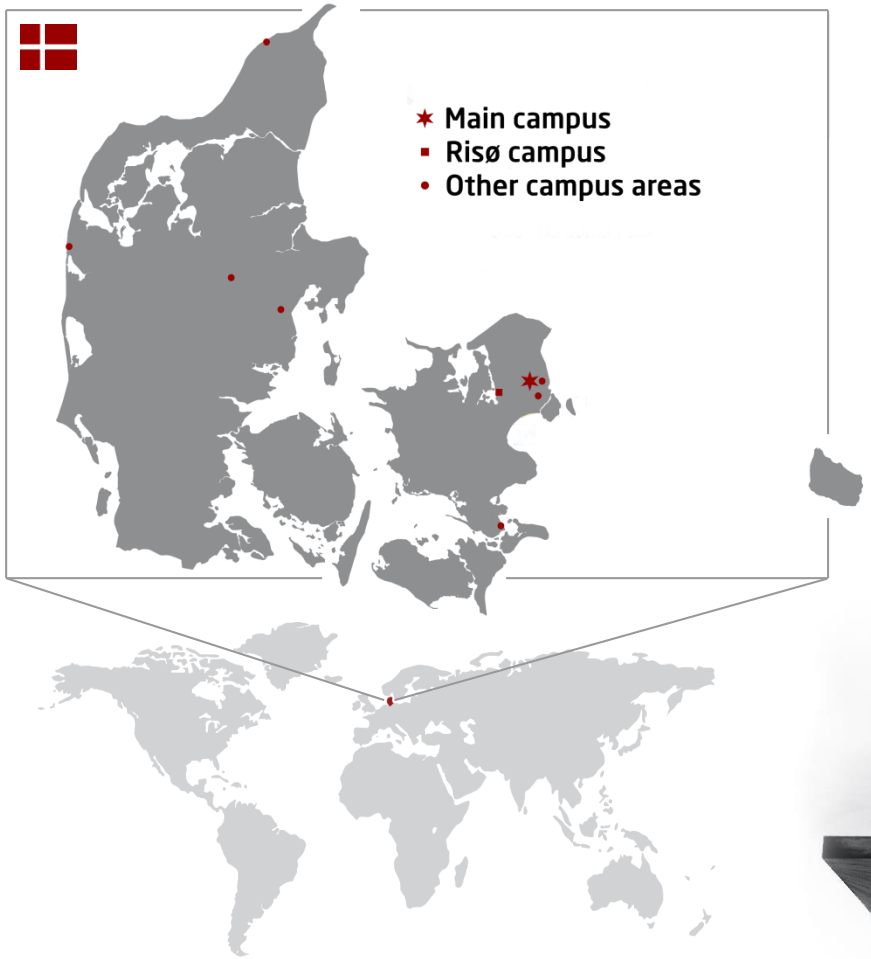
Agenda

- Computational audio
- Cognitive audio information retrieval
- Elicitation of cognitive aspects
  - expressed emotion using pairwise comparisons
  - personalized audio system
- Audio source separation

DTU COMPUTE

Technical University of Denmark

(founded 1829; first rector H.C. Ørsted)



Ranking

Leiden *Crown Indicator* 2010

no. 1 in Scandinavia

no. 7 in Europe

DTU facts and figures

Education

7072 BSc, MSc og Beng students

incl. 626 international MSc students

1197 PhD students

626 exchange studens

296 DTU students at exhange programs

Innovation

87 registered IPR

46 submitted patent applications

Personel

31 DVIP

2657 VIP

2221 TAP

Research

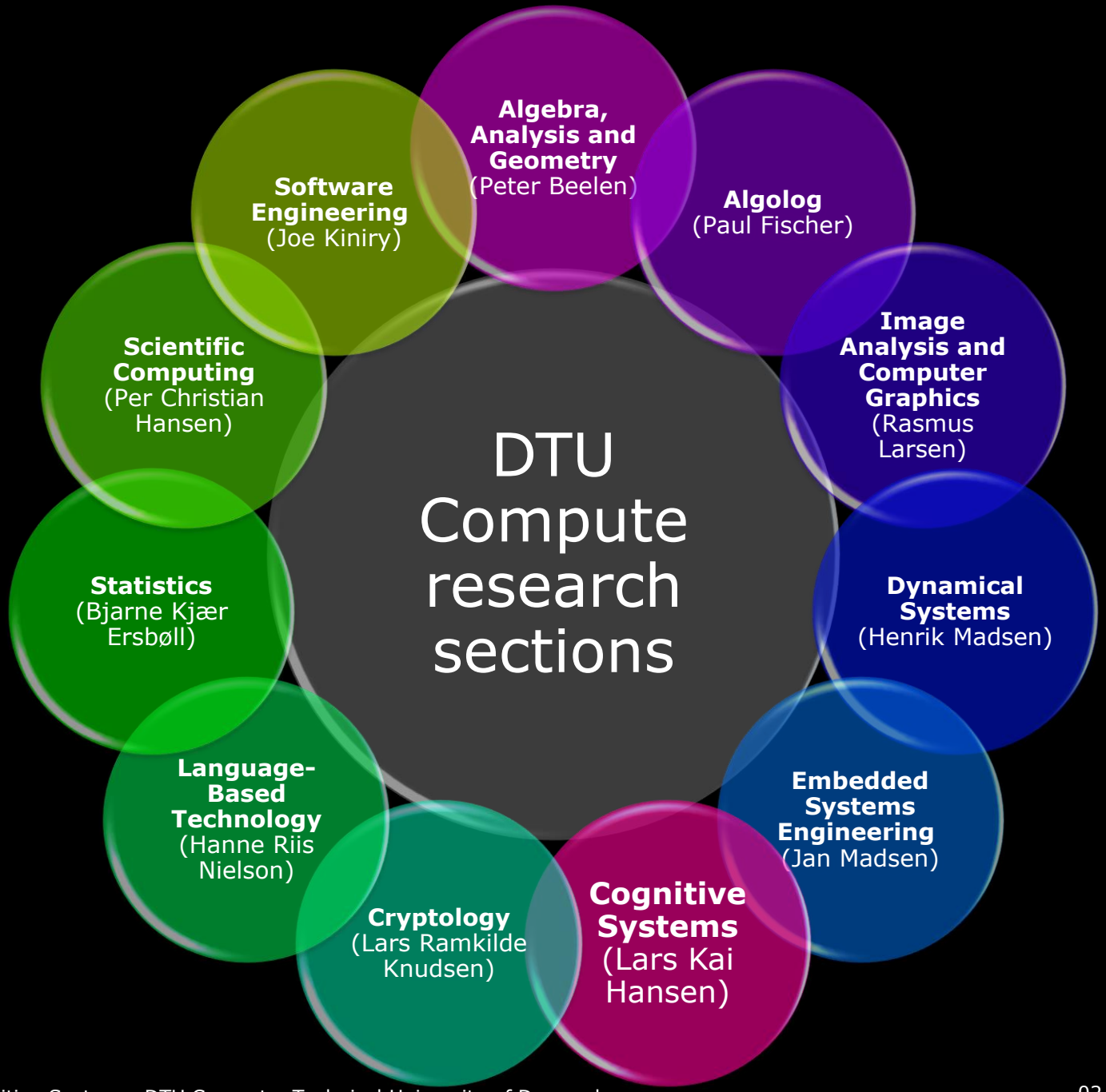
3648 research publications

241 PhD theses

Public sector consultancy
Strategic contract with Danish ministries 338 MDKK

Economy 5.8 bil. DKK

Buildings 454.420 m²


DTU Compute research sections
Algebra, Analysis and Geometry (Peter Beelen)
Algolog (Paul Fischer)
Image Analysis and Computer Graphics (Rasmus Larsen)
Dynamical Systems (Henrik Madsen)
Embedded Systems Engineering (Jan Madsen)
Cognitive Systems (Lars Kai Hansen)
Cryptology (Lars Ramkilde Knudsen)
Language-Based Technology (Hanne Riis Nielson)
Statistics (Bjarne Kjær Ersbøll)
Scientific Computing (Per Christian Hansen)
Software Engineering (Joe Kiniry)

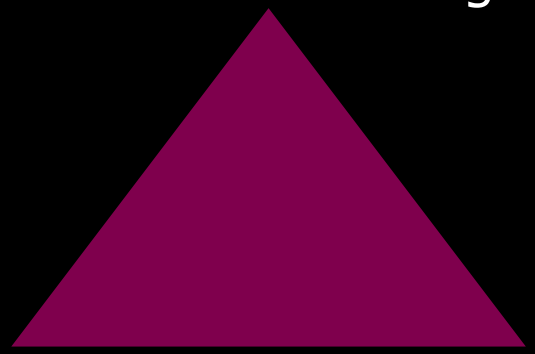
Cognitive Systems Section

Why do we do it? VISION

What do we do? MISSION

- 2 professors
- 7 associate prof.
- 1 assistant prof.
- 1 senior researcher
- 5 postdocs
- 17 Ph.D. students
- 5 project coordinators
- 2 programmers
- 1 admin assistant
- 10 M.Sc. students

machine learning



media technology cognitive science

Bjørn Sand Jensen

Jens Brehm Nielsen

Jens Madsen

Rasmus Troelsgaard

Lars Kai Hansen

Mikkel N. Schmidt

Jerónimo Arenas-García

Ling Feng

Anders Meng

Seliz Karadogan

Letizia Marchegiani

Peter Ahrendt

Michael Kai Petersen

Michael Syskind Pedersen

COMPUTATIONAL AUDIO

Lasse Lohilahti Mølgaard

Tue Lehn-Schiøler

Kaare Brandt Petersen

Cognizant audio systems

fully informed and aware systems

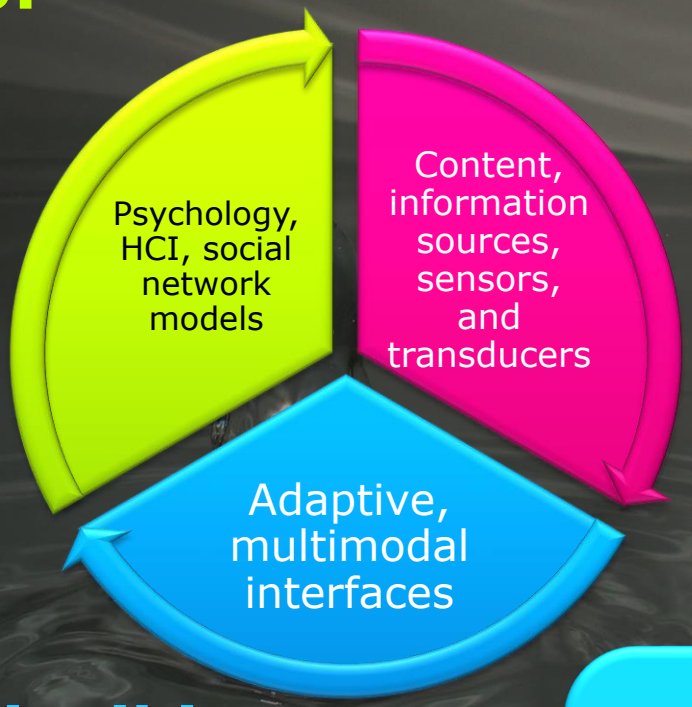
Context:

who, where, wha

Users in the loop:

direct and indirect

Interactive dialog with the user enables long term/continuous behavior tracking, personalization, elicitation of perceptual and affective preferences, as well as adaptation



Listen in on audio and other sensor streams to segment, identify and understand

Flexible integration with other media modalities

Mixed modality experience: Use other modalities to enhance, substitute or provide complementary information

COGNITIVE AUDIO INFORMATION RETRIEVAL

DTU

DTU

DR

Syntonetic

Musikzonen

Geckon

COSOUND

Royal School of Library and Information Science

Hindenburg Systems

UCL

Queen Mary University of London

B&O

Danish Council for Strategic Research Project 2012-2015

Copenhagen University

Aalborg University

State and University Library

University of Glasgow

Hypothesis

Top-down user streams

Learning cognitive representations and interaction

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.

Buttom up audio streams

Vision

The overall vision is to foster truly participatory, collaborative, and cross-cultural tools for enrichment of audio streams which can improve interactivity, findability, experienced quality, ability to co-create, and boost productivity in a broad sense.

Mission

Establish a multi-disciplinary strategic research activity to build a flexible modular audio data processing platform which enables and demonstrates new products and services for the

- commercial sector (Bang&Olufsen, Hindenburg Systems)
- **public service sector (Danish Broadcasting Corporation)**
- education and cultural research (Cultural research at UC)

ELICITATION OF COGNITIVE ASPECTS

Research contributions 2013

- J. Madsen, B. S. Jensen, J. Larsen, Predictive Modeling of Expressed Emotions in Music using Pairwise Comparisons, *CMMR 2012 Post-Proceedings*, vol. 7900, pp. 253-277, Springer-Verlag Berlin Heidelberg, 2013
- B. S. Jensen, J. B. Nielsen, J. Larsen, *Bounded Gaussian Process Regression*, IEEE International Workshop on Machine Learning for Signal Processing, 2013
- J. B. Nielsen, B. S. Jensen, T. J. Hansen, J. Larsen, *Personalized Audio Systems - a Bayesian Approach*, 135th AES Convention, 2013
- Jens Brehm Nielsen, Jakob Nielsen: Efficient Individualization of Hearing and Processers Sound, ICASSP2013.
- Jens Brehm Nielsen, Jakob Nielsen, Jan Larsen: Perception based Personalization of Hearing Aids using Gaussian Process and Active Learning, in preparation for IEEE Trans. ASLP, 2013.
- Jens Brehm Nielsen, PhD Thesis, 2013.

Research contributions 2012

- Bjørn Sand Jensen, Javier Saez Gallego and Jan Larsen. *A Predictive model of music preference using pairwise comparisons*. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2012.
- Jens Madsen, Bjørn Sand Jensen, Jan Larsen and Jens Brehm Nielsen. *Towards Predicting Expressed Emotion in Music from Pairwise Comparisons*, 9th Sound and Music Computing Conference, 2012.
- Jens Madsen, Jens Brehm Nielsen, Bjørn Sand Jensen and Jan Larsen. *Modeling Expressed Emotions in Music using Pairwise Comparisons*. 9th International Symposium on Computer Music Modeling and Retrieval (CMMR) 2012.
- Jens Brehm Nielsen, Bjørn Sand Jensen and Jan Larsen, *Pseudo Inputs For Pairwise Learning With Gaussian Processes*, IEEE International Workshop on Machine Learning for Signal Processing, 2012.
- S. G. Karadogan, J. Larsen, *Combining Semantic and Acoustic Features for Valence and Arousal Recognition in Speech*, Cognitive Information Processing CIP2012, IEEE Press, 2012
- Bjørn Sand Jensen, Integration of top-down and bottom-up information for audio organization and retrieval, PhD thesis, Kgs. Lyngby, Technical University of Denmark, 2012. 197 p. (IMM-PhD-2012; No. 291).
- Seliz Karadogan, Towards Cognizant Hearing Aids: Modeling of Content, Affect and Attention. PhD Thesis, Technical University of Denmark, 2012. 142 p. (IMM-PhD-2012; No. 275).

Research contributions 2011

- Bjørn Sand Jensen, Jens Brehm Nielsen, and Jan Larsen. *Efficient Preference Learning with Pairwise Continuous Observations and Gaussian Processes*, IEEE International Workshop on Machine Learning for Signal Processing, 2011.
- S. G. Karadogan, L. Marchegiani, J. Larsen, L. K. Hansen, *Top-Down Attention with Features Missing at Random*, International Workshop on Machine Learning for Signal Processing, IEEE Press, 2011
- J. B. Nielsen, B. S. Jensen, J. Larsen, *On Sparse Multi-Task Gaussian Process Priors for Music Preference Learning*, NIPS 2011 Workshop on Choice Models and Preference Learning, 2011
- L. Marchegiani, S. G. Karadogan, T. Andersen, J. Larsen, L. K. Hansen, *The Role of Top-Down Attention in the Cocktail Party: Revisiting Cherry's Experiment after Sixty Years*, The tenth International Conference on Machine Learning and Applications (ICMLA'11), 2011

Goal is to efficiently and robustly to elicit, model and predict top-down aspects such as affective, perceptual and other cognitive aspects

Modelling cognitive aspects

Affection

- **Preference elicitation** refers to the problem of developing a decision support system capable of **generating recommendations to a user, thus assisting him in decision making**. It is important for such a system to model user's preferences accurately, find hidden preferences and avoid redundancy. This problem is sometimes studied as a **computational learning theory** problem (ref. Wikipedia)
- Affect refers to the experience of feeling or emotion

Modelling cognitive aspects

Perception

Perception is the organization, identification, and interpretation of sensory information in order to represent and understand the environment. All perception involves signals in the nervous system, which in turn result from physical stimulation of the sense organs. Perception is not the passive receipt of these signals, but can be shaped by learning, memory, and expectation. Perception involves these "top-down" effects as well as the "bottom-up" process of processing sensory input (ref. Wikipedia)

Use cases

- Identify the best audio system among a fixed set of systems
- Audio system feature sensitivity/importance
- Evaluation and comparison of system performance

- Predict the best *unknown* audio system from a set of evaluated audio systems
- Identify best tuning of a single audio system
- Iterative system development on a budget
- Personalization of audio systems

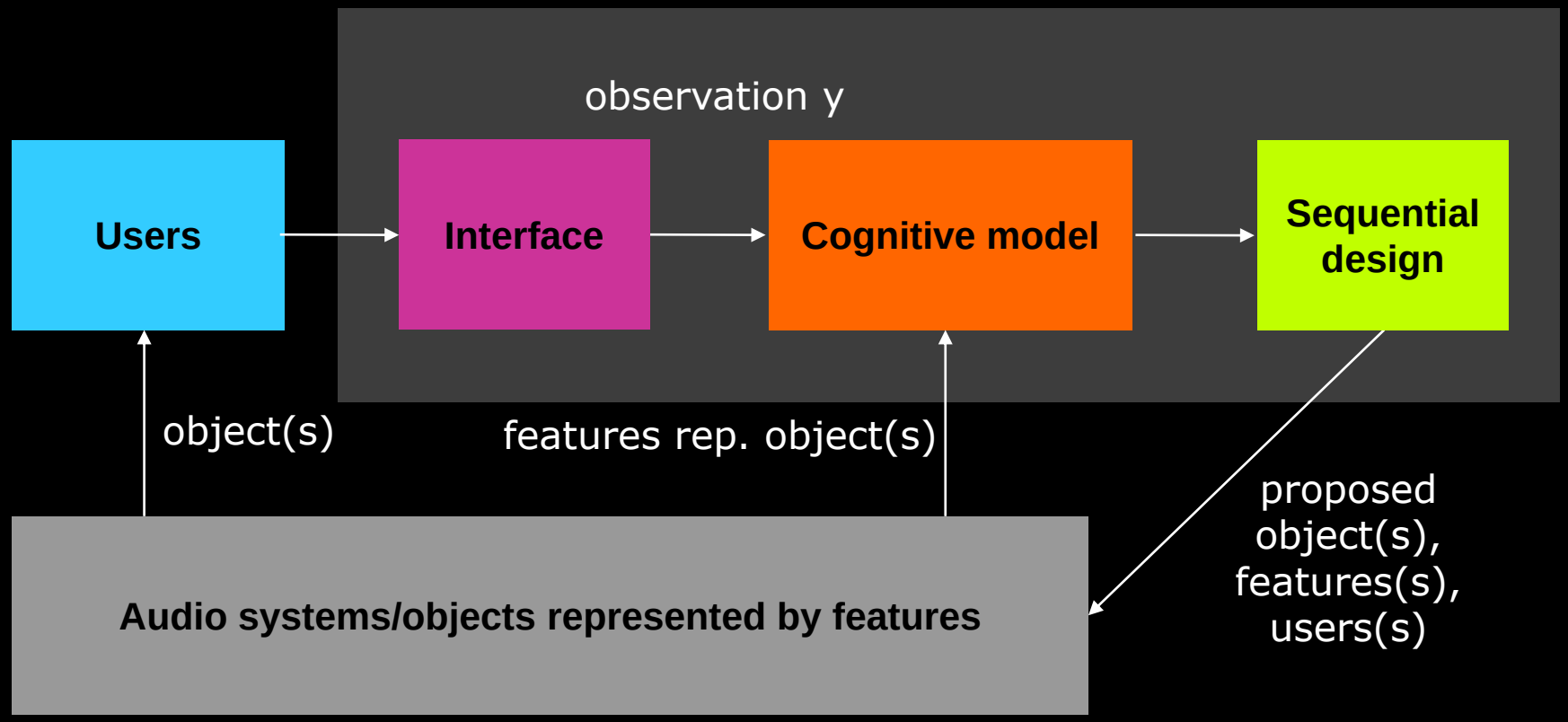
Optimization

Interactive development

Performance evaluation

Individualization

Framework


observation y
Users
Interface
Cognitive model
Sequential design
object(s)
features rep. object(s)
proposed object(s), features(s), users(s)
Audio systems/objects represented by features

Observations

Absolute
Relative
Continuous
Discrete (nominal/ordinal)
Multi vs. single-label

Multiple objects

Ranking
k-AFC
Triangle (odd out)

Noise models

user consistency

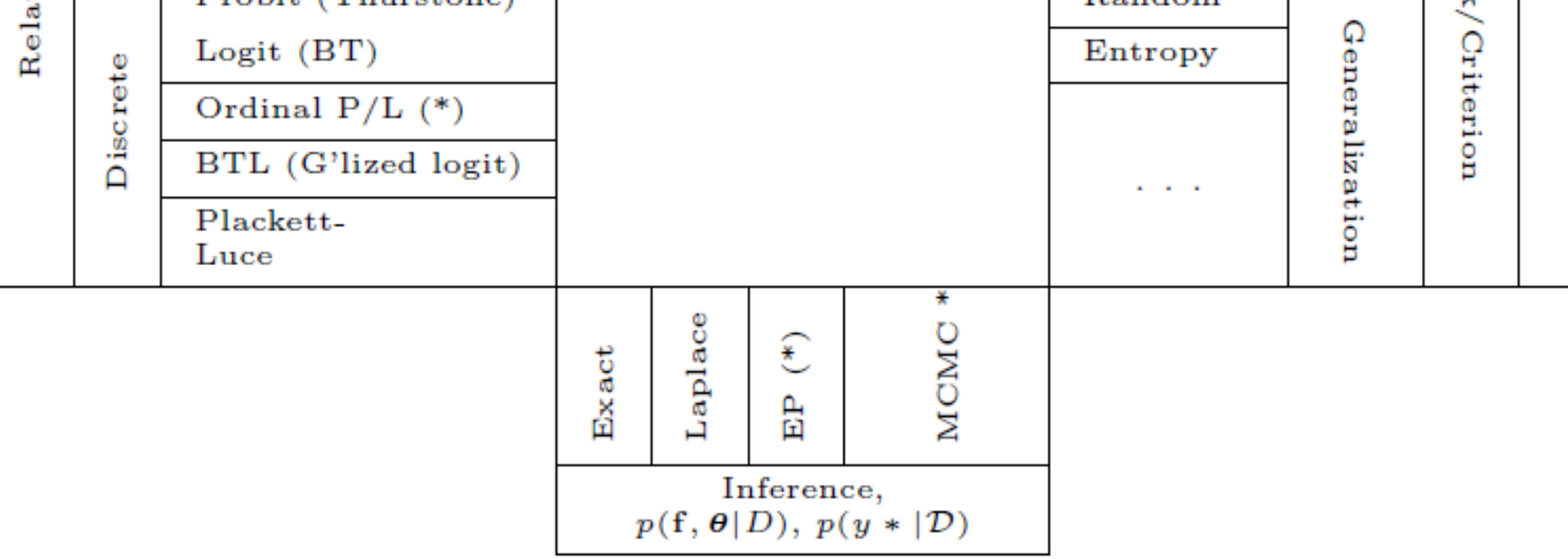
User modeling

individual approach
pooled approach
hierarchical approach based on:
  user features and/or user observations

Observations, $p(y\|f)$				HB*
	Absolute	Continuous	Normal **	
			Student-t **	
			Warped	
			Beta	
			Truncated G.	
		Discrete	Probit/Logit	
			G'lized P/L *	
			Ordinal P/L *	
	Relative	Continious	Warped (*)	
			Beta	
			Truncated G. (*)	
		Discrete	Probit (Thurstone)	
			Logit (BT)	
			Ordinal P/L (*)	
			BTL (G'lized logit)	
			Plackett-Luce	
				Exact

Bayesian nonlinear kernel modelling

			$p(\mathbf{f}\|\boldsymbol{\theta})$							
			Covarince			Induced Sparsity				
			HB* / MTK	ARD/MKL	PPK / SSK	Pseudo input	FITC/PITC (*)			
Absolute	Continuous	Normal **						Random *	*Iterative*	
		Student-t **						IVM *	*Acive Set*	
		Warped						. . .	*Methods*	
		Beta						Approx. *	Plan	I: Co
		Truncated G.						Exact *		
								VOI		



Exact and approximate Inference (learning)

Sequential design of objects, users or inputs

Random *	*Iterative*			Sequential Design
IVM *	*Acive Set*			
. . .	*Methods*			
Approx. *	Plan	I: Computation	Active Learning	
Exact *				
VOI	Greedy			
EVOI				
G(E)VOI				
CWS				
PoI	Optimize	II: Task/Criterion		
EI				
UCB				
THOMP				
Random	Generalization			
Entropy				
. . .				

Fixed design:

m observations

Sequential design:

αm observations

Modeling cognitive aspects

Is it possible to model the users representation of expressed emotion using pairwise comparisons?

Which scaling method should we use?

Is it possible to design a personalized audio system from user's preference of audio clips?

Is it possible to model, interpret and predict individual music preference based on low-level audio features and pairwise comparisons?

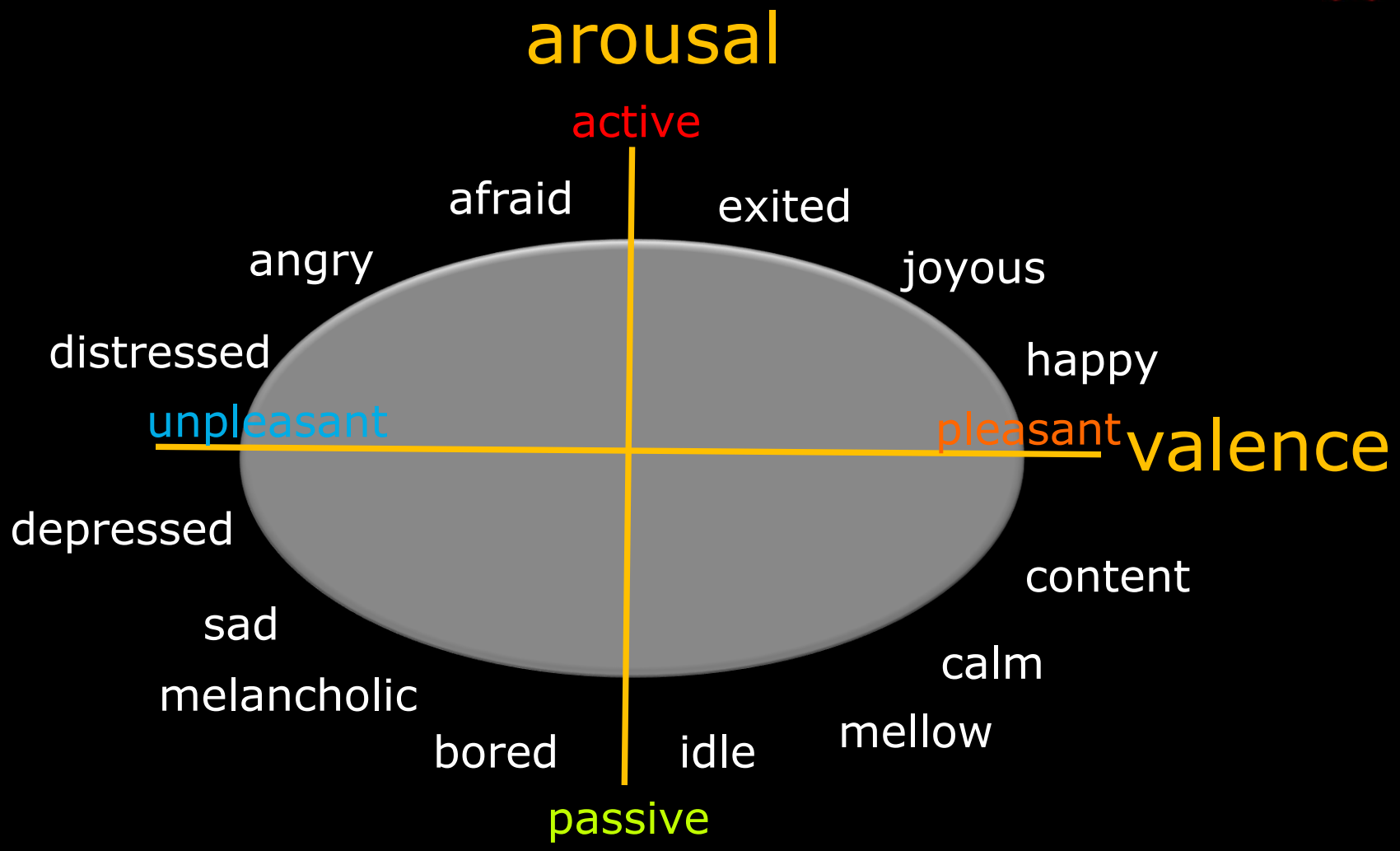
Expressed emotions in music

- Jens Madsen, Bjørn Sand Jensen, Jan Larsen and Jens Brehm Nielsen. *Towards Predicting Expressed Emotion in Music from Pairwise Comparisons*, 9th Sound and Music Computing Conference, 2012.
- Jens Madsen, Jens Brehm Nielsen, Bjørn Sand Jensen and Jan Larsen. *Modeling Expressed Emotions in Music using Pairwise Comparisons*. 9th International Symposium on Computer Music Modeling and Retrieval (CMMR) 2012.
- Madsen, J., Jensen, B.S., Larsen, J., Predictive modeling of expressed emotions in music using pairwise comparisons. M. Aramaki et al. (Eds.): CMMR 2012, LNCS 7900, pp. 253–277, 2013. Springer-Verlag Berlin Heidelberg 2013.

Is it possible to model the users representation of expressed emotion using pairwise comparisons?

Which scaling method should we use?

Emotional spaces



J. A. Russel: "A Circumplex Model of Affect," *Journal of Personality and Social Psychology*, 39(6):1161, 1980

J. A. Russel, M. Lewicka, and T. Niit, "A Cross-Cultural Study of a Circumplex Model of Affect," *Journal of Personality and Social Psychology*, vol. 57, pp. 848-856, 1989

Experimental setup

- **20 excerpts** of **15 second** length were chosen to be evenly distributed in the AV space using a linear regression model and subjective evaluation.
- **8 participants** each evaluated all **190 unique pairwise comparisons**.
- **Question to participants:** Which sound clip was the most

(Arousal) ***excited, active, awake?*** and (Valence) ***positive, glad, happy?***

Audio representation

- 30 dimensions of Mel-frequency cepstral coefficients (MFCC).
- Spectral- flux, roll-off, slope and variation (SSD).
- Zero crossing rate and statistical shape descriptors (TSS).

Features extracted by YAAFE (Yet-Another-Audio-Feature-Extraction) Toolbox

Performance predicting arousal using different audio features

Training size	5%	7%	10%	20%	40%	60%	80%	100%
MFCC	0.3402	0.2860	0.2455	0.2243	0.2092	0.2030	0.1990	0.1949
Envelope	0.4110*	0.4032	0.3911	0.3745	0.3183	0.2847	0.2780	0.2761
Chroma	0.3598	0.3460	0.3227	0.2832	0.2510	0.2403	0.2360	0.2346
CENS	0.3942	0.3735	0.3422	0.2994	0.2760	0.2676	0.2640	0.2621
CRP	0.4475	0.4336	0.4115	0.3581	0.2997	0.2790	0.2735	0.2729
Sonogram	0.3325	0.2824	0.2476	0.2244	0.2118	0.2061	0.2033	0.2026
Pulse clarity	0.4620	0.4129	0.3698	0.3281	0.2964	0.2831	0.2767*	0.2725
Loudness	0.3261	0.2708	**0.2334**	**0.2118**	**0.1996**	**0.1944**	**0.1907**	**0.1862**
Spec. disc.	0.2909	0.2684	0.2476	0.2261	0.2033	0.1948	0.1931	0.1951
Spec. disc. 2	0.3566	0.3223	0.2928	0.2593	0.2313	0.2212	0.2172	0.2138
Key	0.5078	0.4557	0.4059	0.3450	0.3073*	0.2959	0.2926	0.2953
Tempo	0.4416	0.4286	0.4159	0.3804	0.3270	0.3043	0.2953	0.2955
Fluctuations	0.4750	0.4247	0.3688	0.3117	0.2835	0.2731	0.2672	0.2644*
Pitch	0.3173	0.2950	0.2668	0.2453	0.2301	0.2254	0.2230	0.2202
Roughness	**0.2541**	**0.2444**	0.2367	0.2304	0.2236	0.2190	0.2168	0.2170
Spectral crest	0.4645	0.4165	0.3717	0.3285	0.2979	0.2866*	0.2828	0.2838
Echo. timbre	0.3726	0.3203	0.2797	0.2524	0.2366	0.2292	0.2258	0.2219
Echo. pitch	0.3776	0.3264	0.2822	0.2492	0.2249	0.2151	0.2089	0.2059
$Base_{low}$	0.4122	0.3954	0.3956	0.3517	0.3087	0.2879	0.2768	0.2702

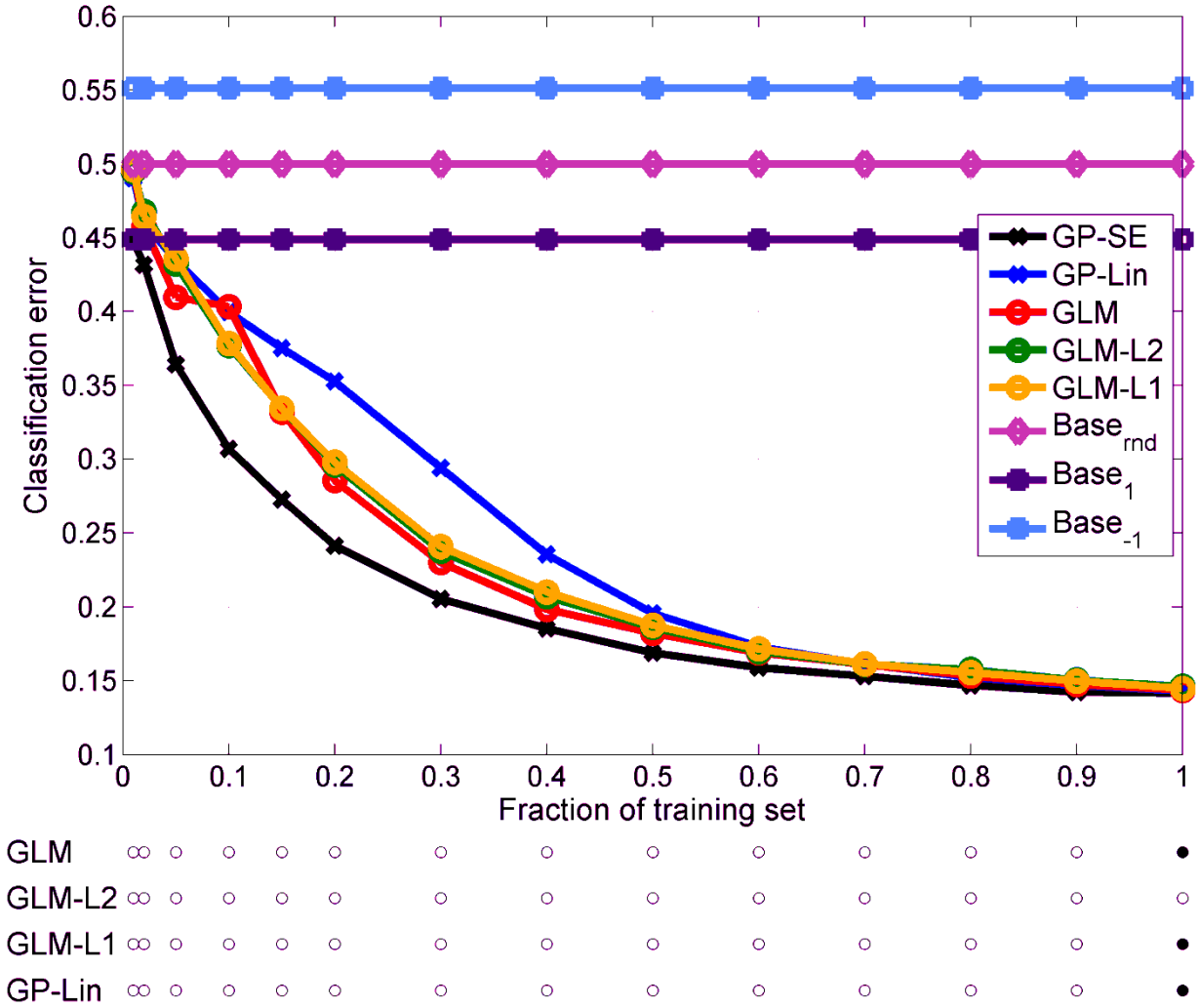
Table 4.2. Arousal: Classification error learning curves as an average of 50 repetitions and 13 individual user models, using only the mean of the features. McNemar test between all points on the learning curve and $Base_{low}$ resulted in $p < 0.05$ for all models except results marked with *, with a sample size of 12.350

Performance predicting valence using different audio features

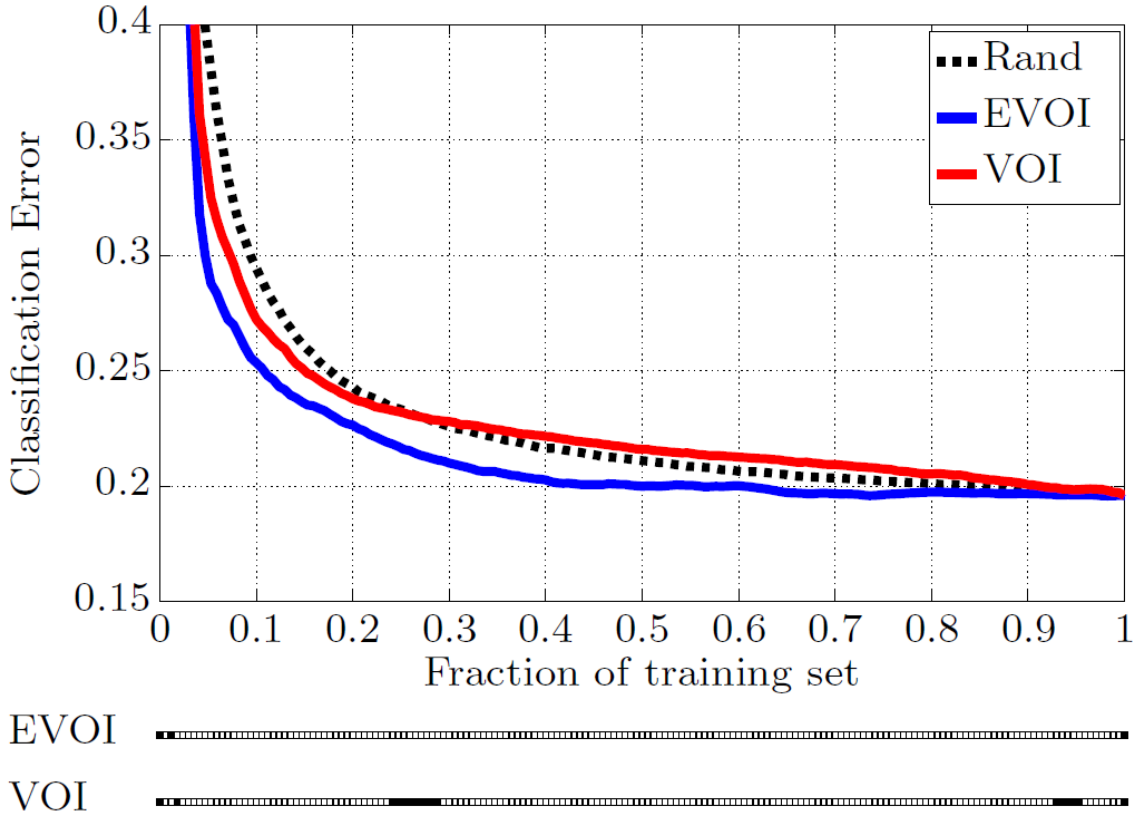
Training size	5%	7%	10%	20%	40%	60%	80%	100%
MFCC	0.4904	0.4354	0.3726	0.3143	0.2856	0.2770	0.2719	0.2650
Envelope	**0.3733**	**0.3545**	0.3336	0.3104	0.2920	0.2842	0.2810	0.2755
Chroma	0.4114*	0.3966*	0.3740	0.3262	0.2862	0.2748	0.2695	0.2658
CENS	0.4353	0.4139	0.3881	0.3471	0.3065	0.2948	0.2901*	0.2824
CRP	0.4466	0.4310	0.4111	0.3656	0.3066	0.2925	0.2876	0.2826
Sonogram	0.4954	0.4360	0.3749	0.3163	0.2884	0.2787	0.2747	0.2704
Pulse clarity	0.4866	0.4357	0.3856	0.3336	0.3026	0.2930	0.2879	0.2810
Loudness	0.4898	0.4310	0.3684	0.3117	0.2854	0.2768	0.2712	0.2664
Spec. disc.	0.4443	0.4151	0.3753	0.3263	0.2939	0.2857	0.2827	0.2794
Spec. disc. 2	0.4516	0.4084	0.3668	0.3209	0.2916	0.2830	0.2781	0.2751
Key	0.5303	0.4752	0.4104	0.3370	0.2998	0.2918	0.2879	0.2830*
Tempo	0.4440	0.4244	0.3956	0.3559*	0.3158	0.2985	0.2933	0.2883
Fluctuations	0.4015	0.3584	**0.3141**	**0.2730**	**0.2507**	**0.2433**	**0.2386**	**0.2340**
Pitch	0.4022	0.3844	0.3602	0.3204	0.2926	0.2831	0.2786	0.2737
Roughness	0.4078	0.3974	0.3783	0.3313	0.2832	0.2695	0.2660	0.2605
Spec. crest	0.4829	0.4289	0.3764	0.3227	0.2994	0.2942	0.2933	0.2923
Echo. timbre	0.4859	0.4297	0.3692	0.3127	0.2859	0.2767	0.2732	0.2672
Echo. pitch	0.5244	0.4643	0.3991*	0.3275	0.2942	0.2841	0.2790	0.2743
$Base_{low}$	0.4096	0.3951	0.3987	0.3552	0.3184	0.2969	0.2893	0.2850

Table 4.1. Valence: Classification error learning curves as an average of 50 repetitions and 13 individual user models, using both mean and standard deviation of the features. McNemar test between all points on the learning curve and $Base_{low}$ resulted in $p < 0.05$ for all models except results marked with *, with a sample size of 12.350

Learning curve modeling valence shows nonlinear modeling is best



How many pairwise comparisons do we need to model emotions?



Using active learning

15% for valence

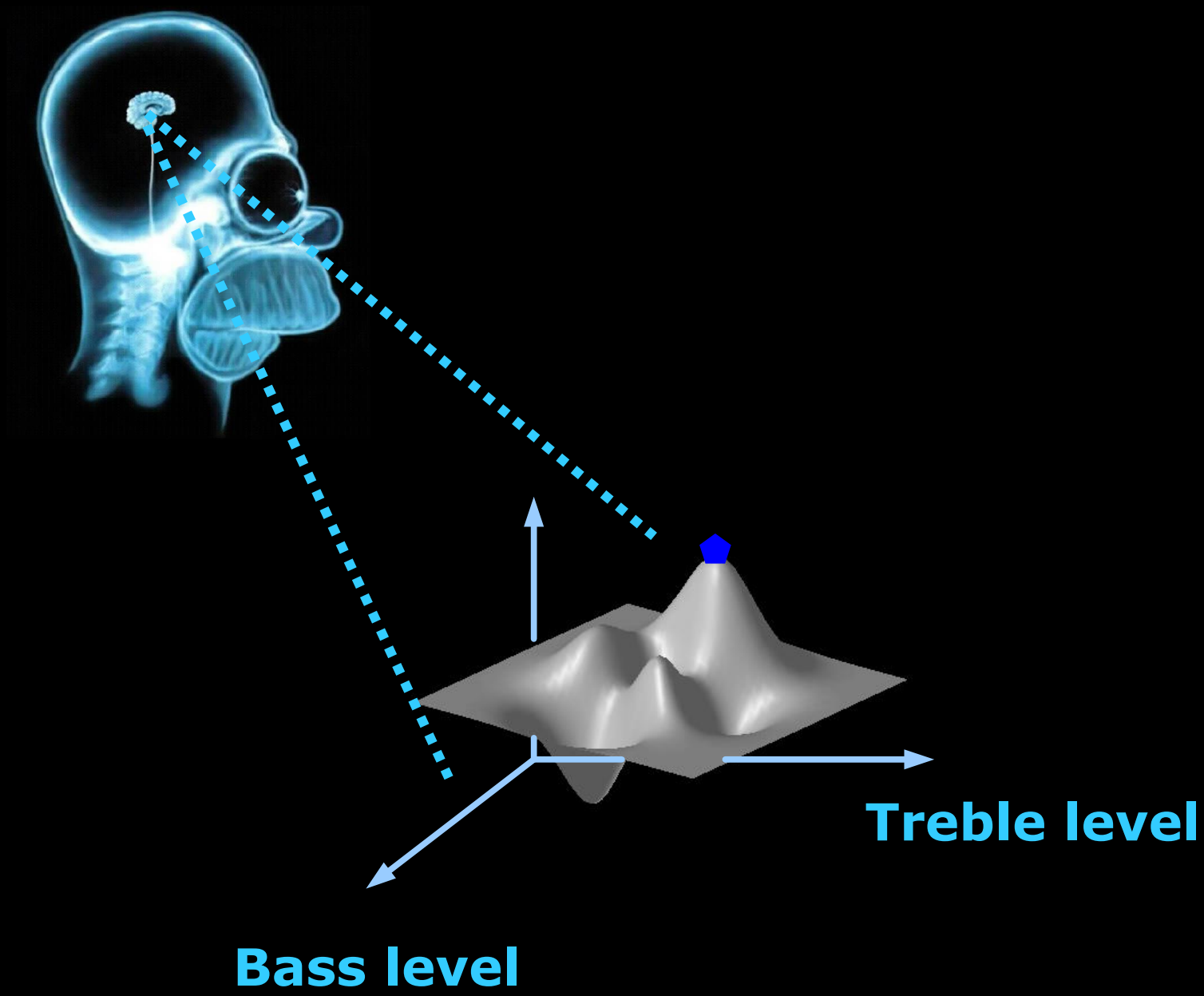
9% for arousal

Main conclusion on eliciting emotions

- Models produce similar results using a learning curve
- Models produce different rankings specially when using a fraction of comparisons
- Large individual differences between the ranking of music expressed in music on dimensions of Valence and Arousal
- Promising error rates for both arousal and valence using as little as 30% of the training set corresponding to 2.5 comparisons per excerpt.
- Pairwise comparisons (2AFC) can scale when using active learning.

Personalized Audio Systems – a Bayesian Approach

- Jens Brehm Nielsen, Bjørn Sand Jensen,Toke Jansen Hansen, Jan Larsen, *AES Convention 135, New York, 17-20 October 2013.*
- Jens Brehm Nielsen, Jakob Nielsen, Jan Larsen: Perception based Personalization of Hearing Aids using Gaussian Process and Active Learning, *in preparation for IEEE Trans. ASLP, 2013.*


Treble level
Bass level

Personalizing an audio system

Machine Learning

(1) A setting is selected in a clever way based on the model of the user's *internal representation*

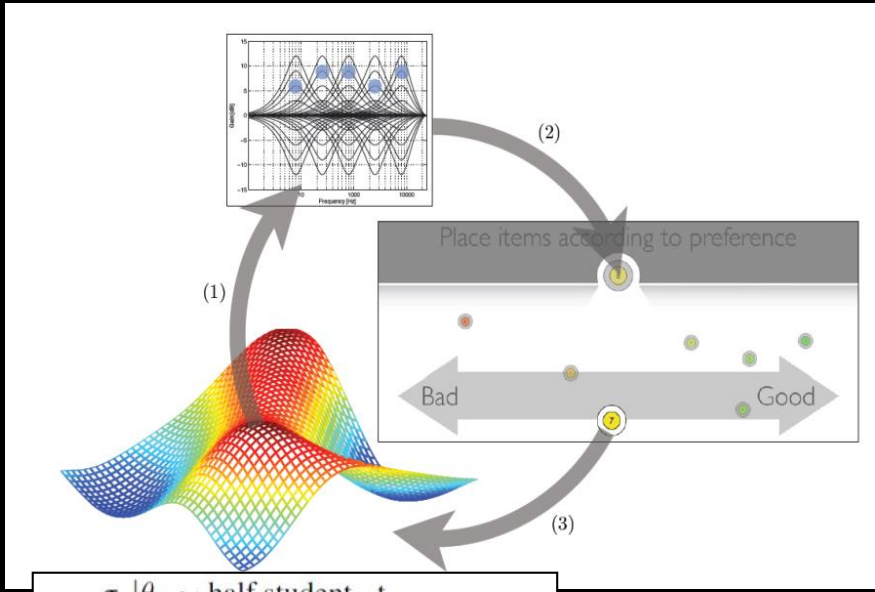
- which is a function, $f(x)$, (modeled by the Gaussian process) over device parameters, x.

DSP

(2) The new setting is *presented* to the user by processing the audio accordingly (standard DSP).

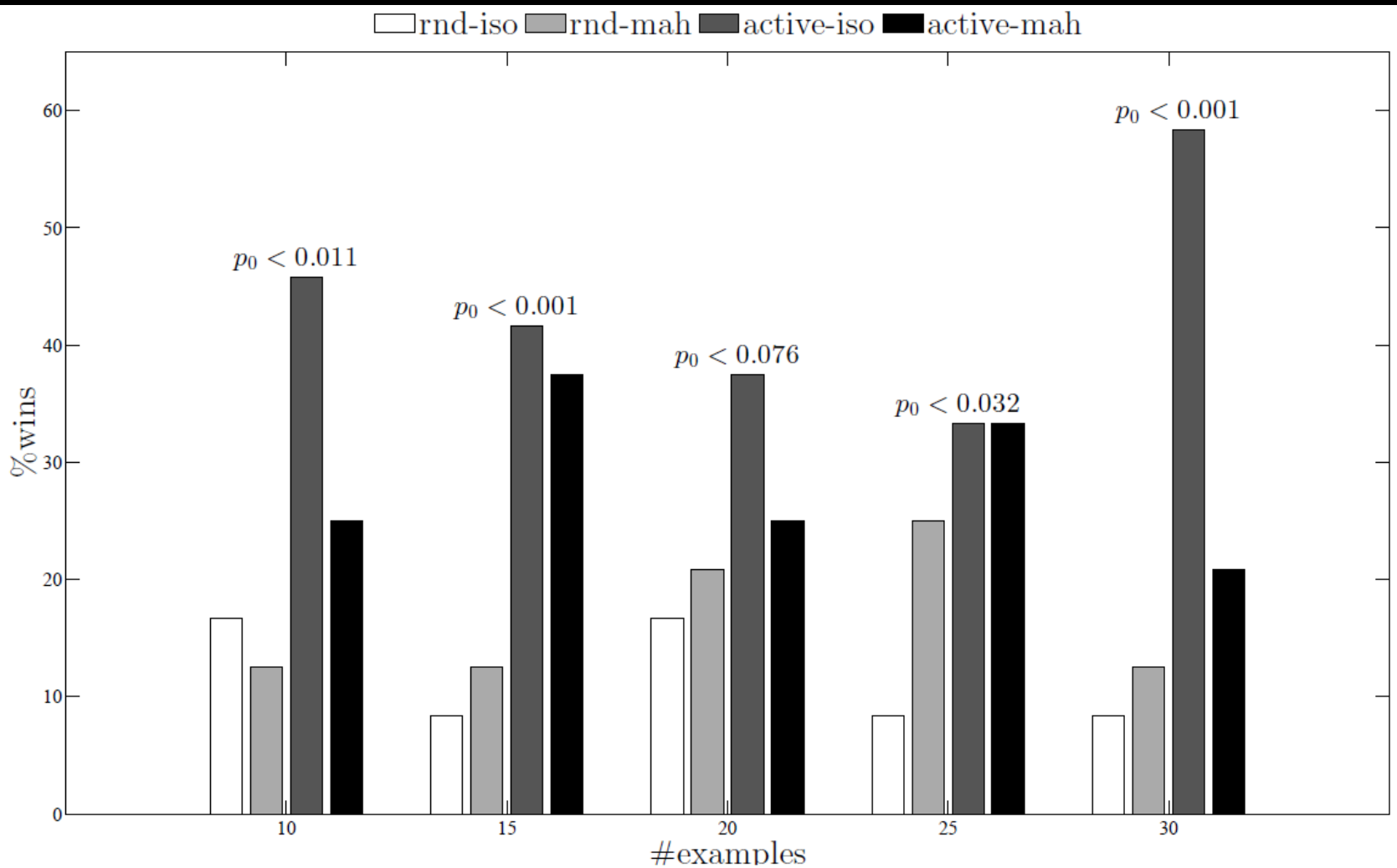
HCI

(3) The users listens to a stimuli and indicates his/her preferences in a simple interfaces with anchors



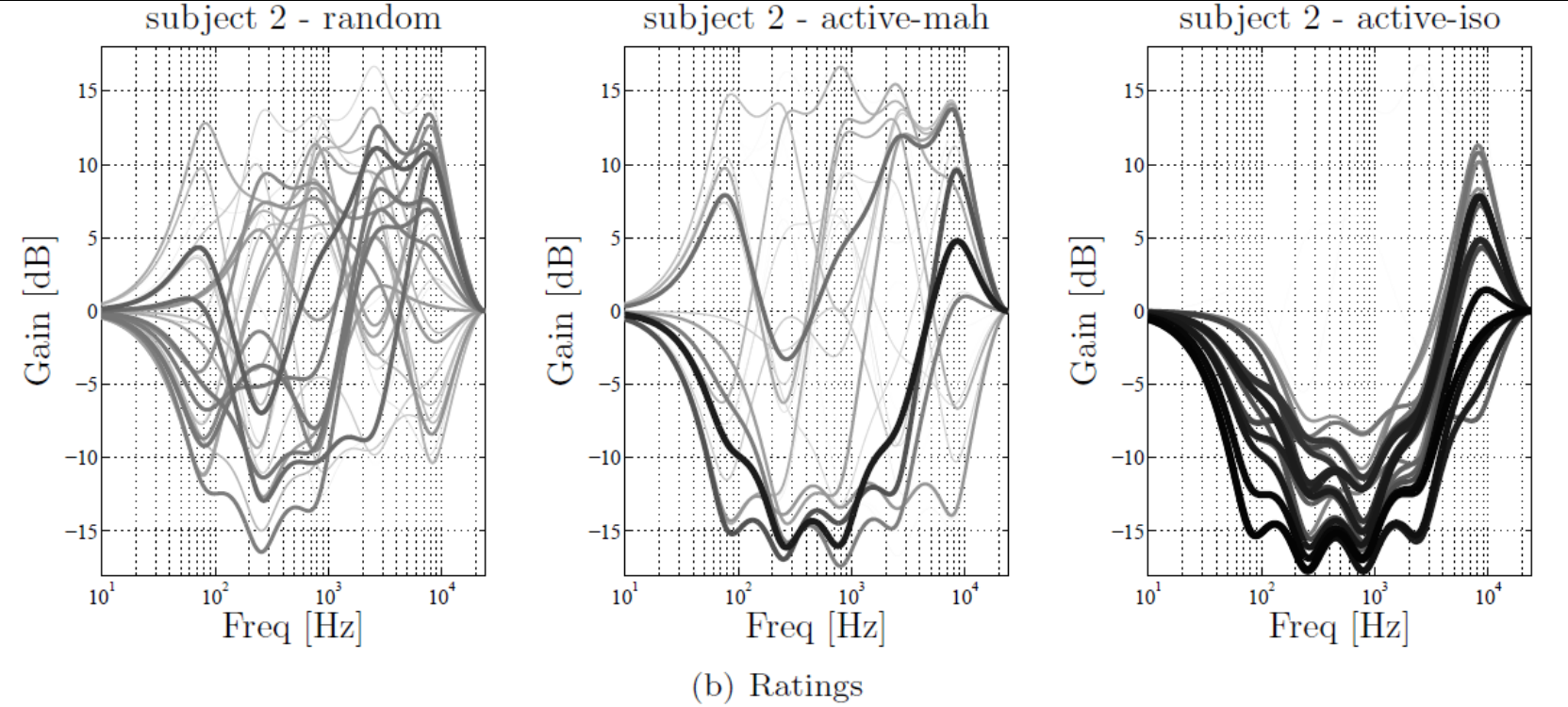
$$\sigma_s|\theta_s \sim \text{half student - t}$$
$$\sigma_\ell|\theta_\ell \sim \text{half student - t}$$
$$f_i|\sigma_s,\sigma_\ell \sim \mathcal{GP}\left(m\left(\mathbf{x}_i\right), \mathrm{k}\left(\mathbf{x}_i,\cdot\right)_{\sigma_s,\sigma_\ell}\right)$$
$$z_i|f_i \sim \mathcal{N}\left(f_i,\sigma_i\right)$$
$$y_i = \Phi^{-1}\left(z_i\right),$$

Results



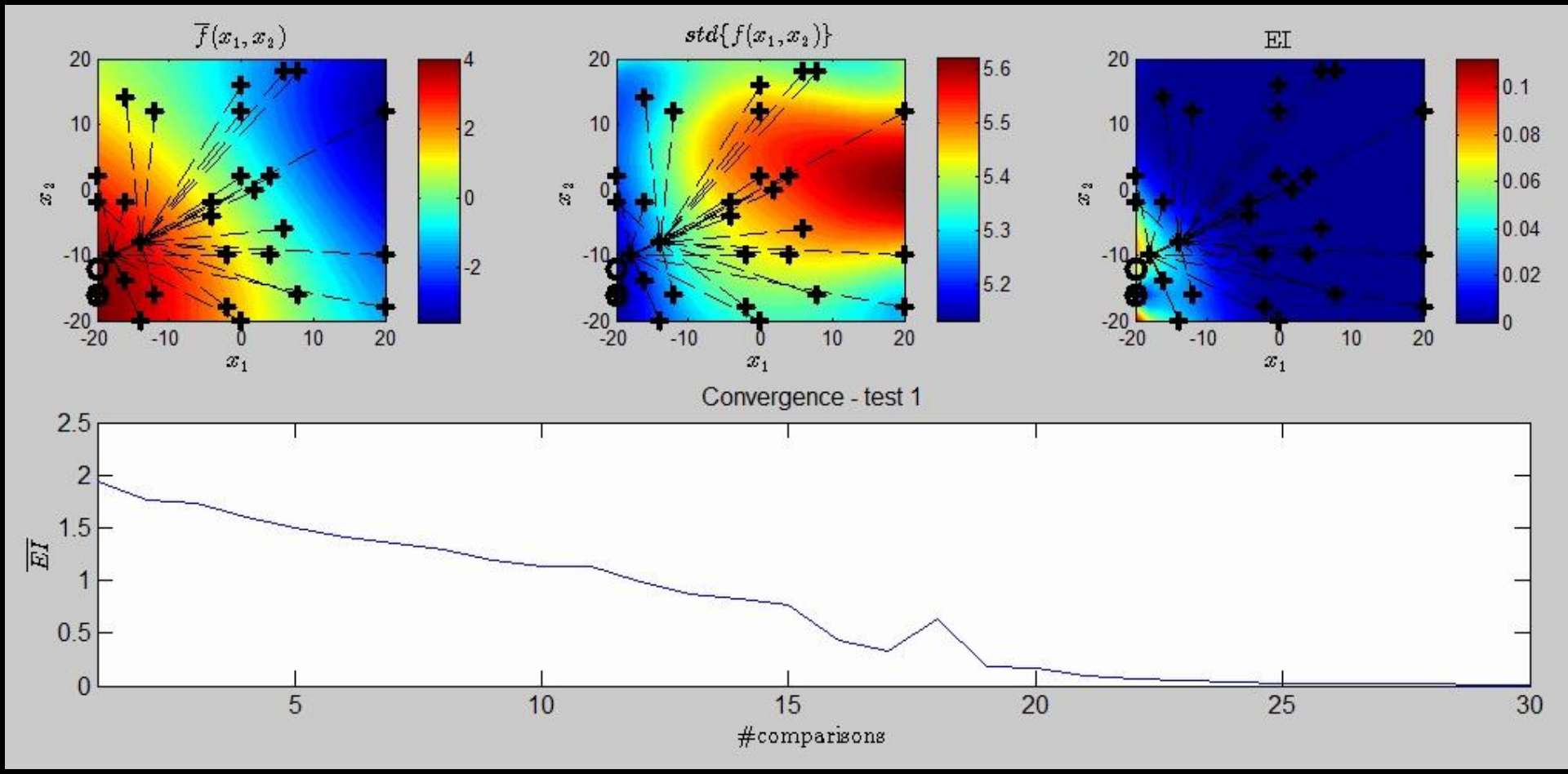
(a) Learning Curve

Some Results

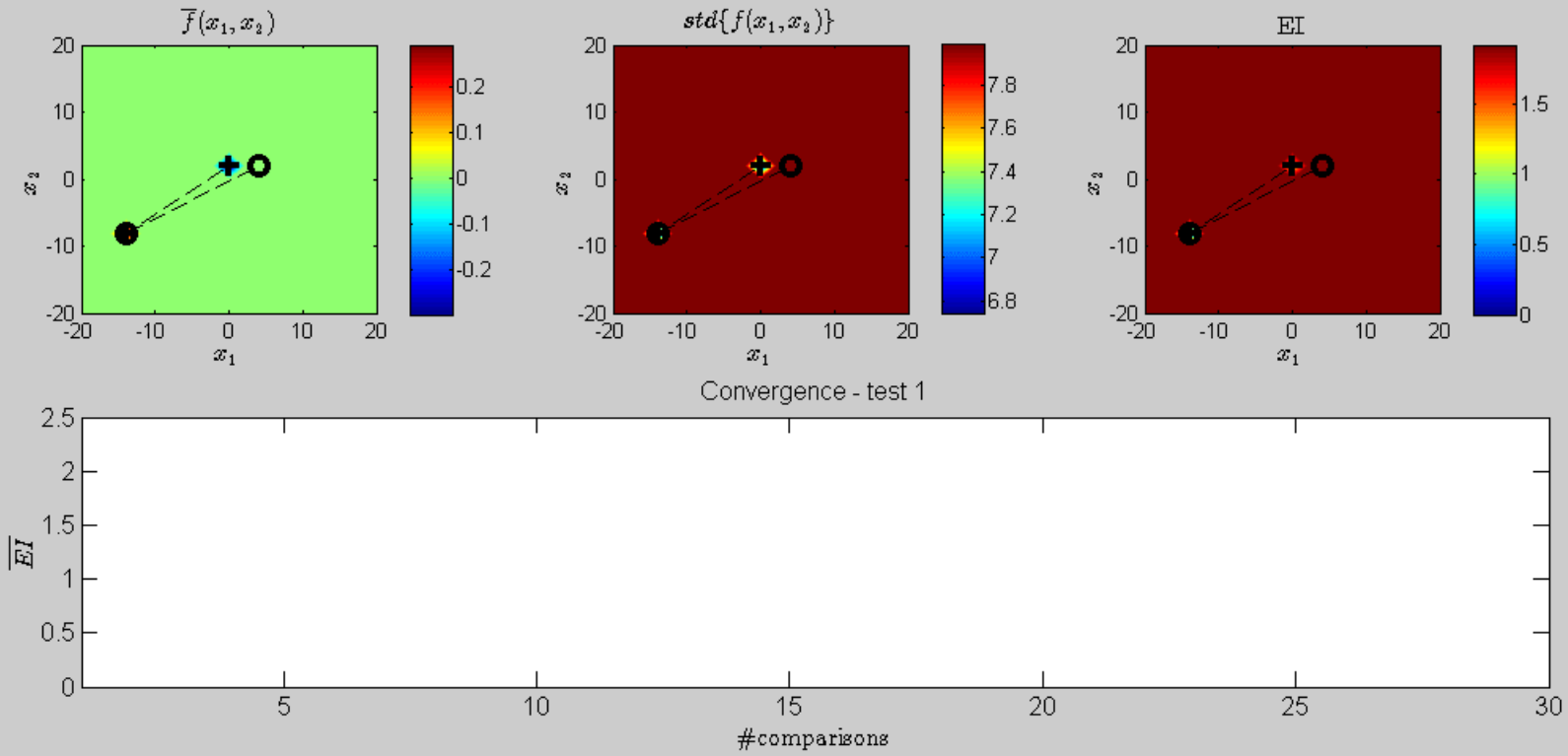


(b) Ratings

Pairwise (2AFC) personalization of HA



Active learning process



METADATA PREDICTION


WINAMP
File
Play
Options
View
Help
Detach Visualizer
Alternative Rock
Blues
Classical
Country
Dance
Folk
Jazz
Opera & Vocal
Pop
R&B
Rap & Hip-Hop
Rock
Prev
Next
Random
0:39
STEREO
MUDDY WATERS - GOT MY MOJO WORKING (3:06)

AUDIO SOURCE SEPARATION

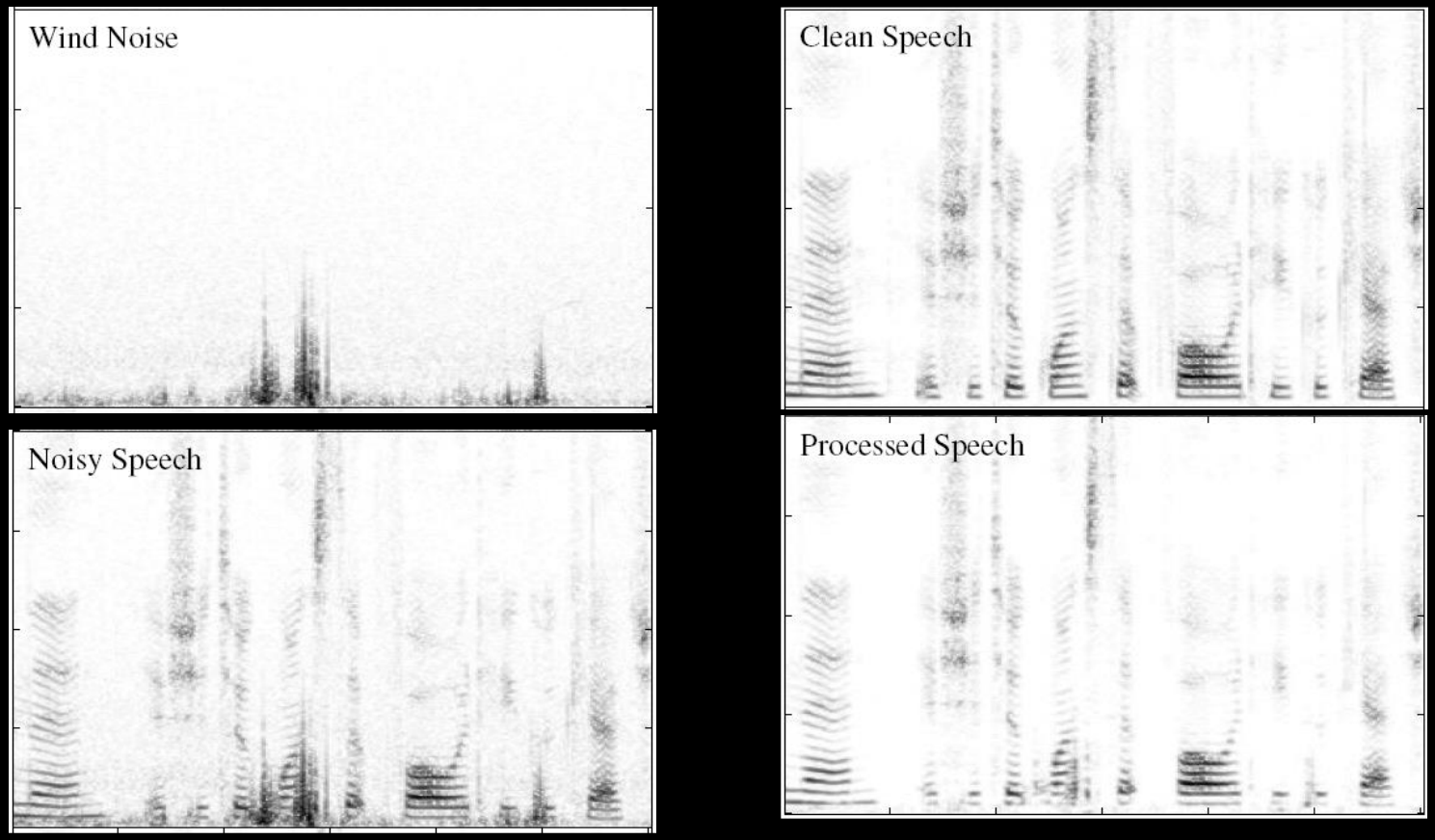
Audio separation

- 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

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

Wind noise reduction



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

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)		