MIRages: an account of music audio extractors, semantic description and context-awareness, in the three ages of MIR

Perfecto Herrera Boyer

Music Technology Group, DTIC, UPF

PhD Thesis defence

Directors:
Xavier Serra & Emilia Gómez
Committee:
Geoffroy Peeters (Télécom ParisTech),
Sergi Jordà (UPF), Josep Lluís Arcos (IIIA)

December 12th, 2018, Barcelona, Spain

Outline

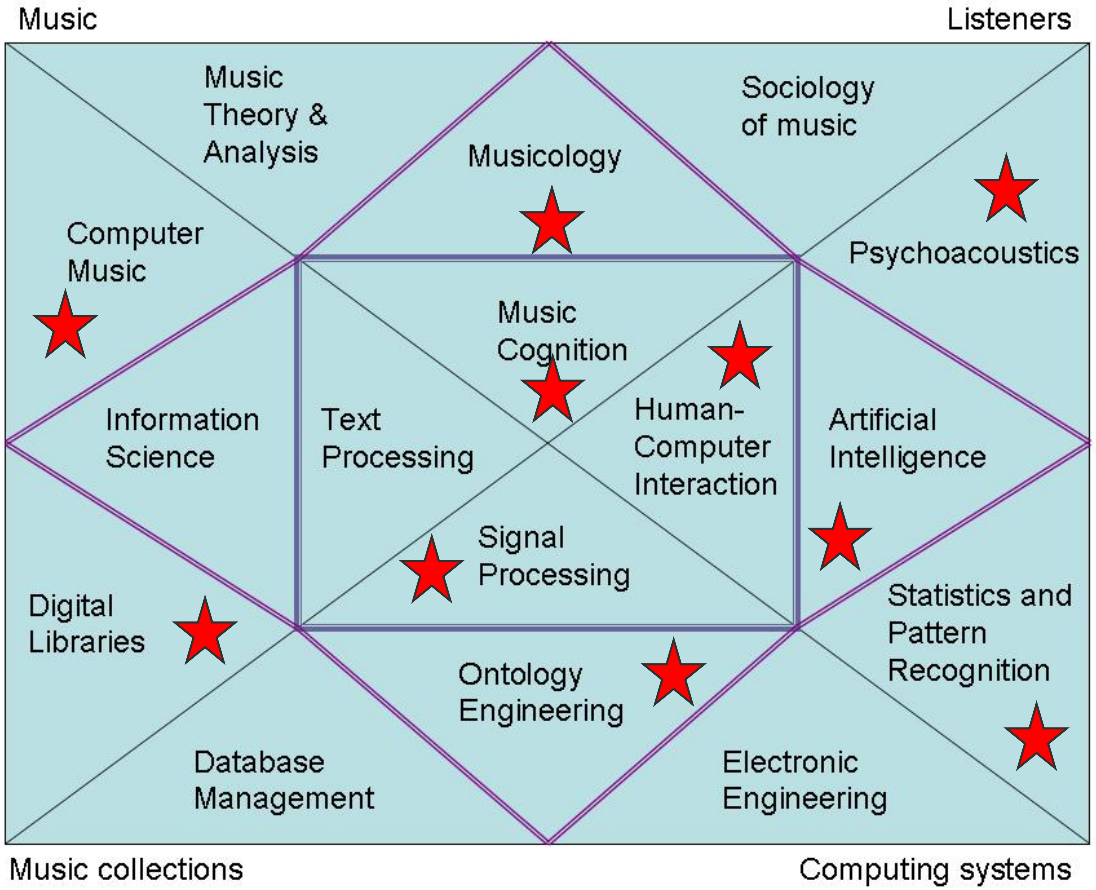
- Motivation and context of the thesis
- The age of extractors
- The age of semantic descriptors
- The age of context-aware systems
- The age of creative systems?
- Concluding thoughts

Frank Nack, "The Future in Digital Media Computing is Meta“, IEEE MultiMedia, pp. 10-13, 2004

Motivation and context

- Atypical dissertation
- Perspective gained after 20 years in the field and involvement in many MTG projects
- Report on a personal "way of thinking/doing"
- Compilation of journal articles (with a couple of "special" conference papers)
- Articles selected combining relevance, impact, personal contribution, breadth of journals, and fit to narrative purposes (among 33 journal articles, 150 conference papers)
- Essential role of collaborators (>80!)

Personal statements



Music

Listeners

Music Theory & Analysis

Musicology

Sociology of music

Computer Music

Psychoacoustics

Music Cognition

Information Science

Text Processing

Human-Computer Interaction

Artificial Intelligence

Signal Processing

Digital Libraries

Statistics and Pattern Recognition

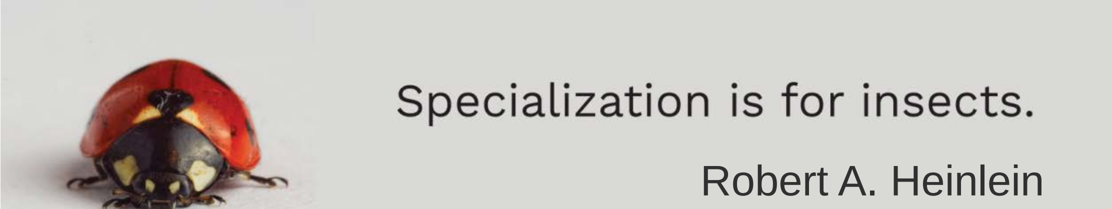
Ontology Engineering

Database Management

Electronic Engineering

Music collections

Computing systems



Specialization is for insects.

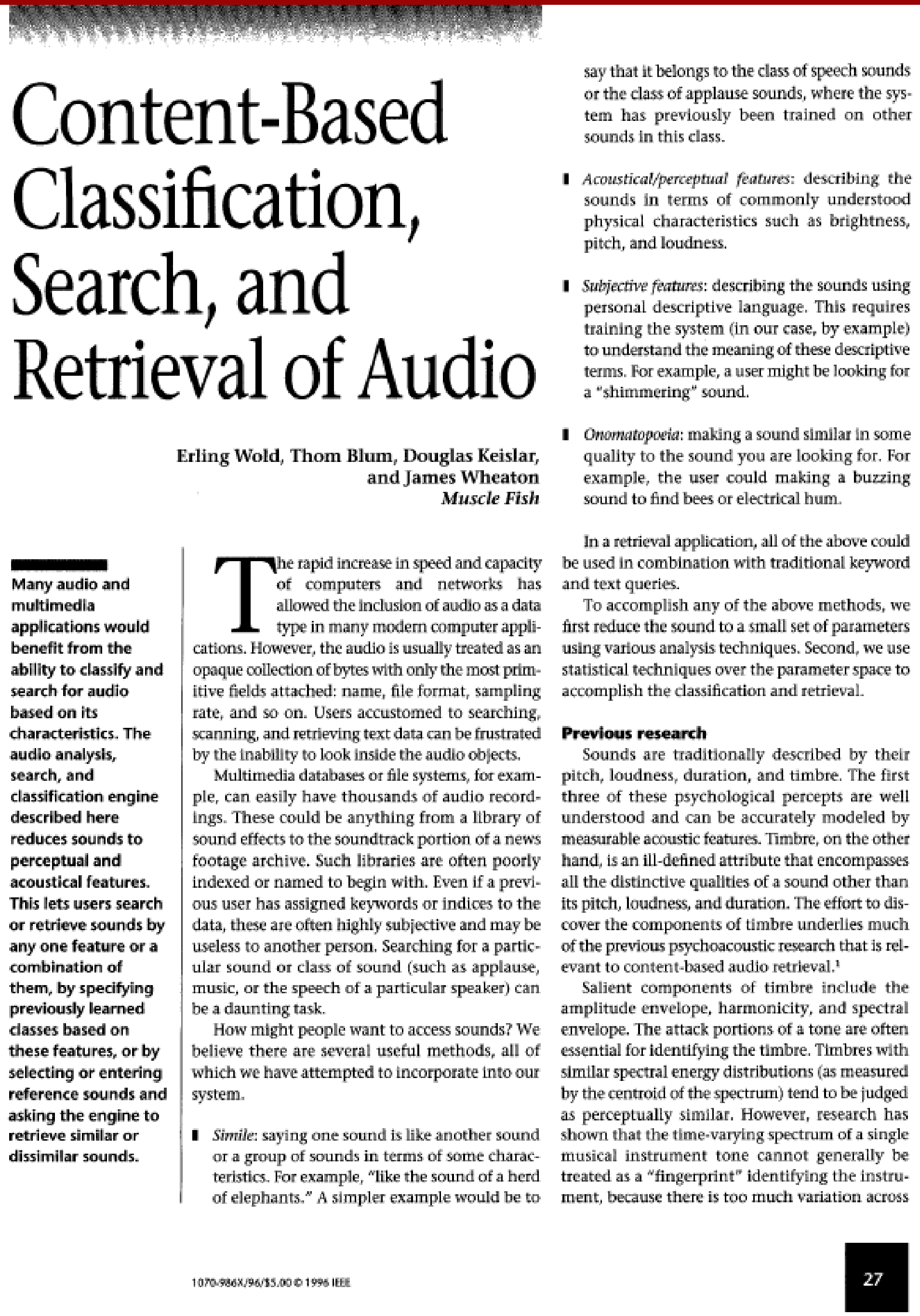
Robert A. Heinlein

Once
upon
a
time...

… when everything was

IEEE MultiMedia

1070-986X/96/$5.00 © 1996 IEEE



Content-Based Classification, Search, and Retrieval of Audio

Erling Wold, Thom Blum, Douglas Keislar, and James Wheaton
Muscle Fish

Many audio and multimedia applications would benefit from the ability to classify and search for audio based on its characteristics. The audio analysis, search, and classification engine described here reduces sounds to perceptual and acoustical features. This lets users search or retrieve sounds by any one feature or a combination of them, by specifying previously learned classes based on these features, or by selecting or entering reference sounds and asking the engine to retrieve similar or dissimilar sounds.

The rapid increase in speed and capacity of computers and networks has allowed the inclusion of audio as a data type in many modern computer applications. However, the audio is usually treated as an opaque collection of bytes with only the most primitive fields attached: name, file format, sampling rate, and so on. Users accustomed to searching, scanning, and retrieving text data can be frustrated by the inability to look inside the audio objects.

Multimedia databases or file systems, for example, can easily have thousands of audio recordings. These could be anything from a library of sound effects to the soundtrack portion of a news footage archive. Such libraries are often poorly indexed or named to begin with. Even if a previous user has assigned keywords or indices to the data, these are often highly subjective and may be useless to another person. Searching for a particular sound or class of sound (such as applause, music, or the speech of a particular speaker) can be a daunting task.

How might people want to access sounds? We believe there are several useful methods, all of which we have attempted to incorporate into our system.

- *Simile*: saying one sound is like another sound or a group of sounds in terms of some characteristics. For example, "like the sound of a herd of elephants." A simpler example would be to say that it belongs to the class of speech sounds or the class of applause sounds, where the system has previously been trained on other sounds in this class.
- *Acoustical/perceptual features*: describing the sounds in terms of commonly understood physical characteristics such as brightness, pitch, and loudness.
- *Subjective features*: describing the sounds using personal descriptive language. This requires training the system (in our case, by example) to understand the meaning of these descriptive terms. For example, a user might be looking for a "shimmering" sound.
- *Onomatopoeia*: making a sound similar in some quality to the sound you are looking for. For example, the user could making a buzzing sound to find bees or electrical hum.

In a retrieval application, all of the above could be used in combination with traditional keyword and text queries.

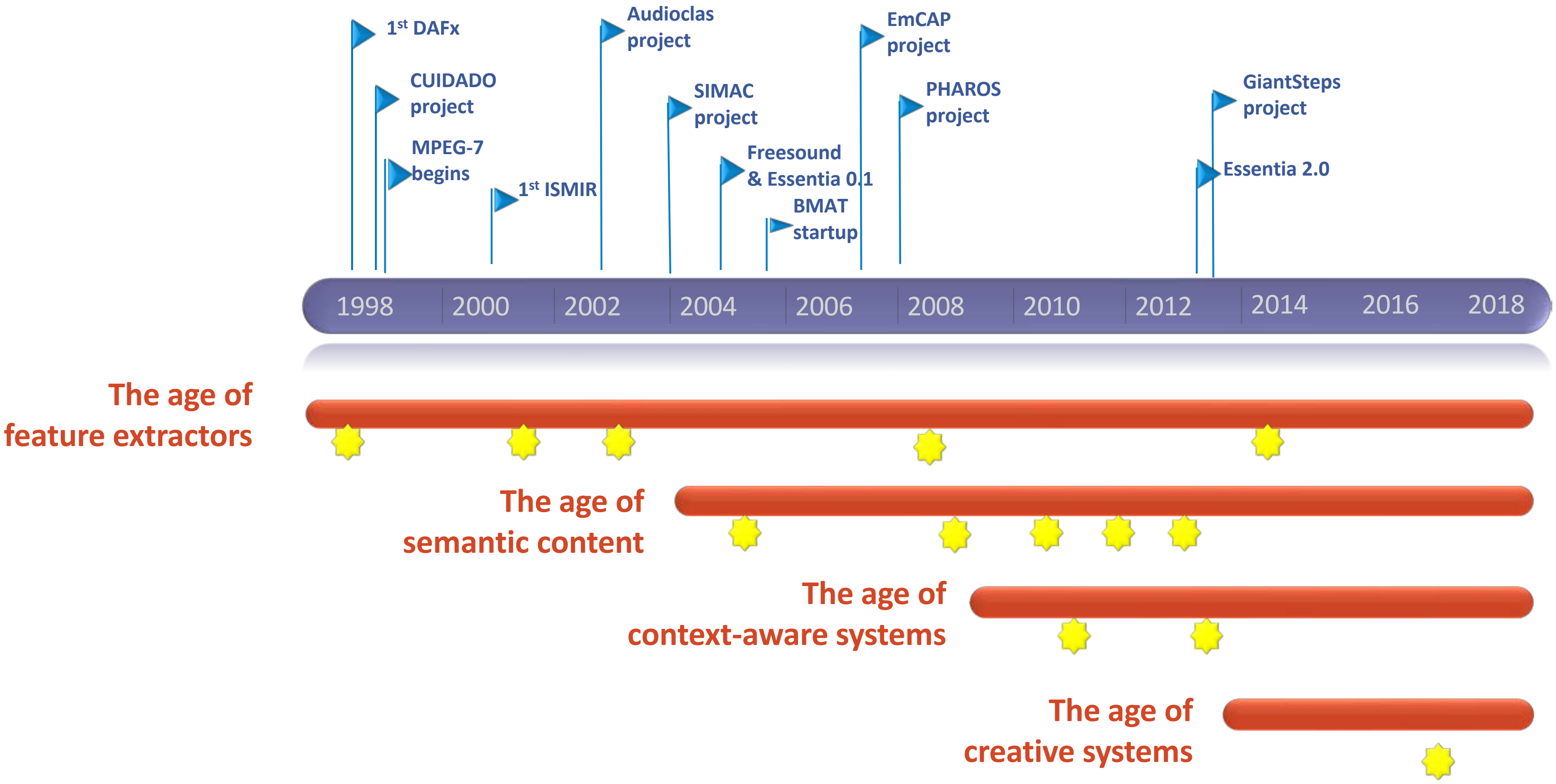
To accomplish any of the above methods, we first reduce the sound to a small set of parameters using various analysis techniques. Second, we use statistical techniques over the parameter space to accomplish the classification and retrieval.

Previous research

Sounds are traditionally described by their pitch, loudness, duration, and timbre. The first three of these psychological percepts are well understood and can be accurately modeled by measurable acoustic features. Timbre, on the other hand, is an ill-defined attribute that encompasses all the distinctive qualities of a sound other than its pitch, loudness, and duration. The effort to discover the components of timbre underlies much of the previous psychoacoustic research that is relevant to content-based audio retrieval.[1]

Salient components of timbre include the amplitude envelope, harmonicity, and spectral envelope. The attack portions of a tone are often essential for identifying the timbre. Timbres with similar spectral energy distributions (as measured by the centroid of the spectrum) tend to be judged as perceptually similar. However, research has shown that the time-varying spectrum of a single musical instrument tone cannot generally be treated as a "fingerprint" identifying the instrument, because there is too much variation across

1070-986X/96/$5.00 © 1996 IEEE



1st DAFx

CUIDADO project

MPEG-7 begins

1st ISMIR

Audioclas project

SIMAC project

Freesound & Essentia 0.1

BMAT startup

EmCAP project

PHAROS project

GiantSteps project

Essentia 2.0

1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018

The age of feature extractors

The age of semantic content

The age of context-aware systems

The age of creative systems

1. The age of feature extractors

“It’s more fun to compute (x2)”

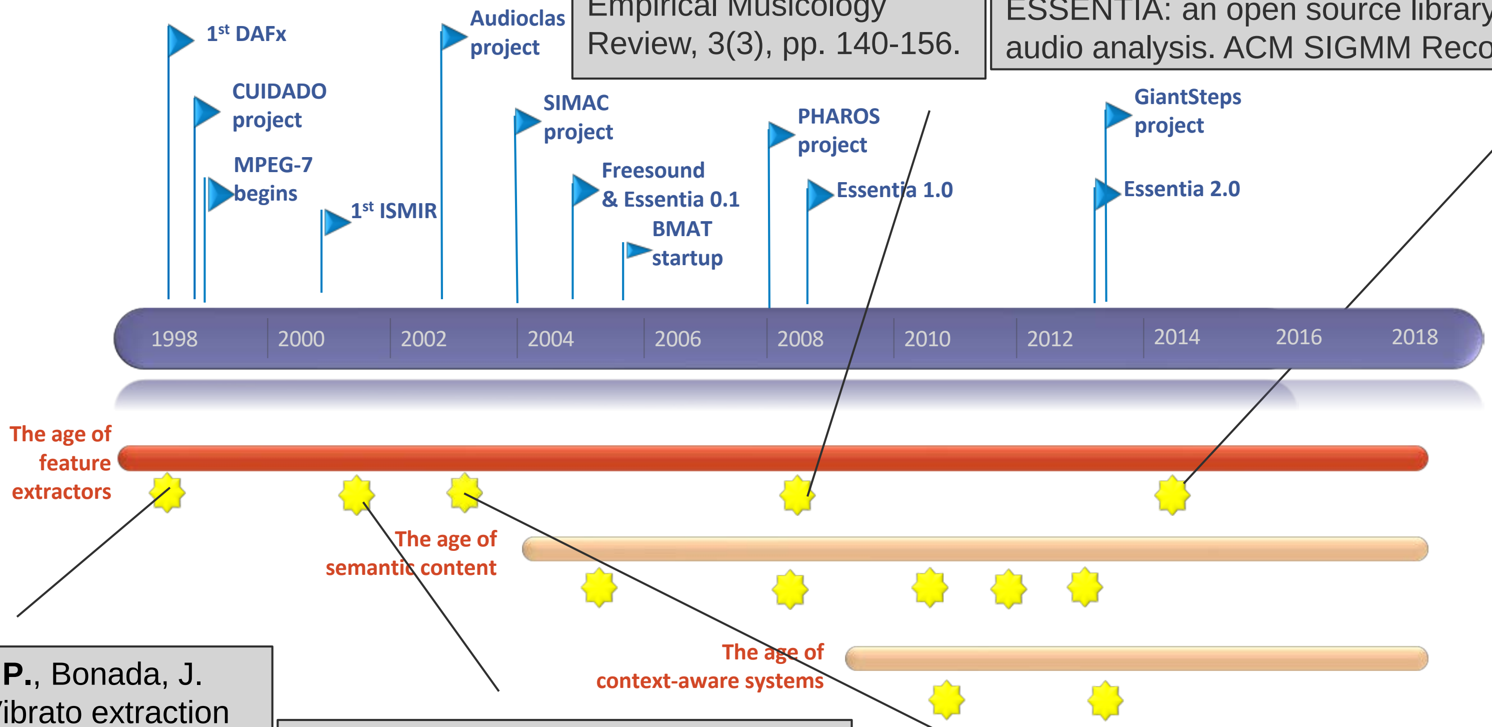
Ralf Hütter / Florian Schneider-Esleben / Karl Bartos

The age of feature extractors

- *Understanding without separation*
- Involvement in MPEG-7 (1998-2000): multimedia content description
- First ISMIR (2000)
- CUIDAD and CUIDADO EU projects (1999-2003):
  - Our first descriptors
  - Tools for metadata generation in parallel with the generation of content in music production
  - Search in instrument sounds databases

Gómez, E. & **Herrera, P.** (2008). Comparative Analysis of Music Recordings from Western and Non-Western traditions by Automatic Tonal Feature Extraction. Empirical Musicology Review, 3(3), pp. 140-156.

Bogdanov, D., Wack, N., Gómez, E., Gulati S., **Herrera, P.**, Mayor, O., Roma, G., Salamon, J., Zapata, J. & Serra, X. (2014). ESSENTIA: an open source library for audio analysis. ACM SIGMM Records. 6(1).



Herrera, P., Bonada, J. (1998). Vibrato extraction and parameterization in the spectral modeling synthesis framework. Proceedings of the Digital Audio Effects Workshop (DAFX98), Barcelona, Spain, 99-103.

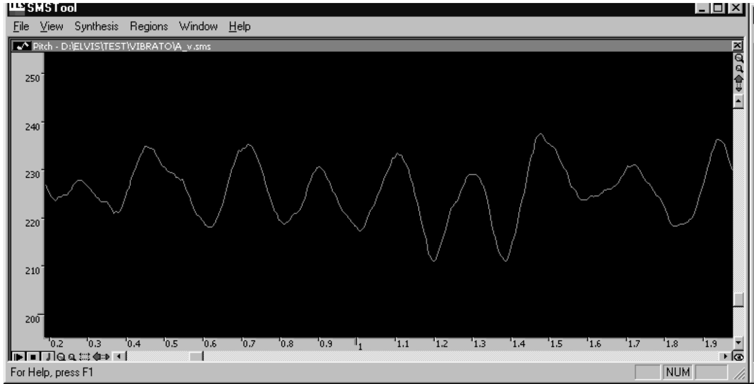
Herrera, P., Yeterian, A., Gouyon, F. (2002). Automatic classification of drum sounds: a comparison of feature selection methods and classification techniques. In C. Anagnostopoulou et al. (Eds), "Music and Artificial Intelligence". Lecture Notes in Computer Science V. 2445. Berlin: Springer-Verlag.

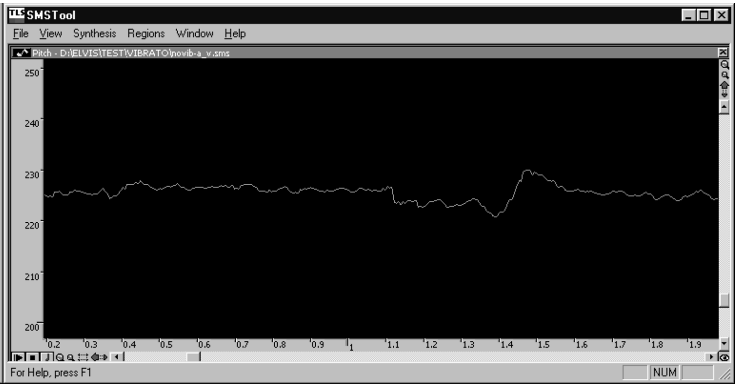
Herrera, P., Peeters, G., Dubnov, S. (2003). Automatic Classification of Musical Instrument Sounds. Journal of New Music Research. 32(1), pp. 3-21.

Vibrato

Herrera, P., Bonada, J. (1998). Vibrato extraction and parameterization in the spectral modeling synthesis framework. Proceedings of the Digital Audio Effects Workshop (DAFX98) 99-103. (paper cited 74 times)

- Analysis of monophonic audio
- Vibrato as a property of F0
- FFT of short-excerpts of F0 trajectories yielded rate and magnitude
- NO systematic EVALUATION (which was normal at that time) !!!





Timbre features

Herrera, P., Yeterian, A., Gouyon, F. (2002). Automatic classification of drum sounds: a comparison of feature selection methods and classification techniques. In C. Anagnostopoulou et al. (Eds), "Music and Artificial Intelligence". Lecture Notes in Computer Science V. 2445. Berlin: Springer-Verlag. (Series IF: 0.8; Q2 in Computer Science journals; paper cited 148 times)

- Context: MPEG-7 features "chase", validation and application
- One of the early ML papers in the MTG
- First paper on generic automatic detection of drum sounds
- Focus on feature selection and classification models
- Hierarchical models (classifiers for individual instruments and for families –membranes vs plates)

Timbre features

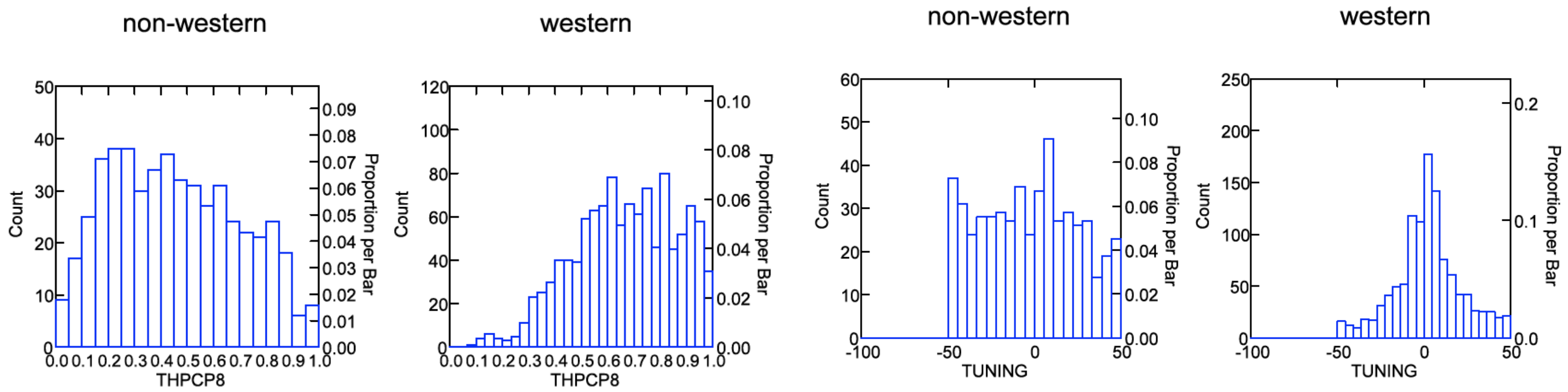
Herrera, P., Peeters, G., Dubnov, S. (2003). "Automatic Classification of Musical Instrument Sounds". Journal of New Music Research. 32(1), pp. 3-21. (Journal h-index: 22; Journal IF 2016: 1.122; Q1 in music-related journals; paper cited 231 times)

- My most cited paper until June 2016!
- Review paper derived from ISMIR 2000 paper
- No empirical research included, value of tutorial-like texts
- One of the earliest papers remarking the potential of SVM

Tonal features

Gómez, E., **Herrera, P.** (2008). "Comparative Analysis of Music Recordings from Western and Non-Western traditions by Automatic Tonal Feature Extraction". Empirical Musicology Review, 3(3), pp. 140-156. (paper cited 33 times)

- Tonal features (HPCP bins, equal-temprered deviation, non-tempered energy ratio, diatonic strength, dissonance) used to tell apart music from different cultures
- Use of statistical distribution comparisons
- Early piece of literature dealing with (rough and naïve) characterization of musical cultures



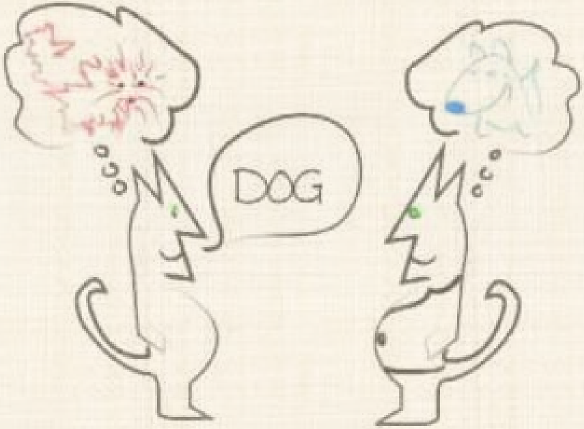
Feature extraction library

Bogdanov, D., Wack, N., Gómez, E., Gulati S., **Herrera, P**., Mayor, O., Roma, G., Salamon, J., Zapata, J. & Serra, X. (2014). ESSENTIA: an open source library for audio analysis. ACM SIGMM Records. 6(1). (Winner of ACM MM 2013 Open Source competition; 5 citations, but a longer report of Essentia (Bogdanov et al., 2013a), not from a journal, has been cited 204 times)

- Cross-platform open library for audio and music features
- Result of 10+ years of studying/using features
- Includes timbre, loudness, pitch, rhythm, tonal and morphological descriptors + statistical moments
- Includes Python bindings and vamp plugins for easy extension/integration/prototyping

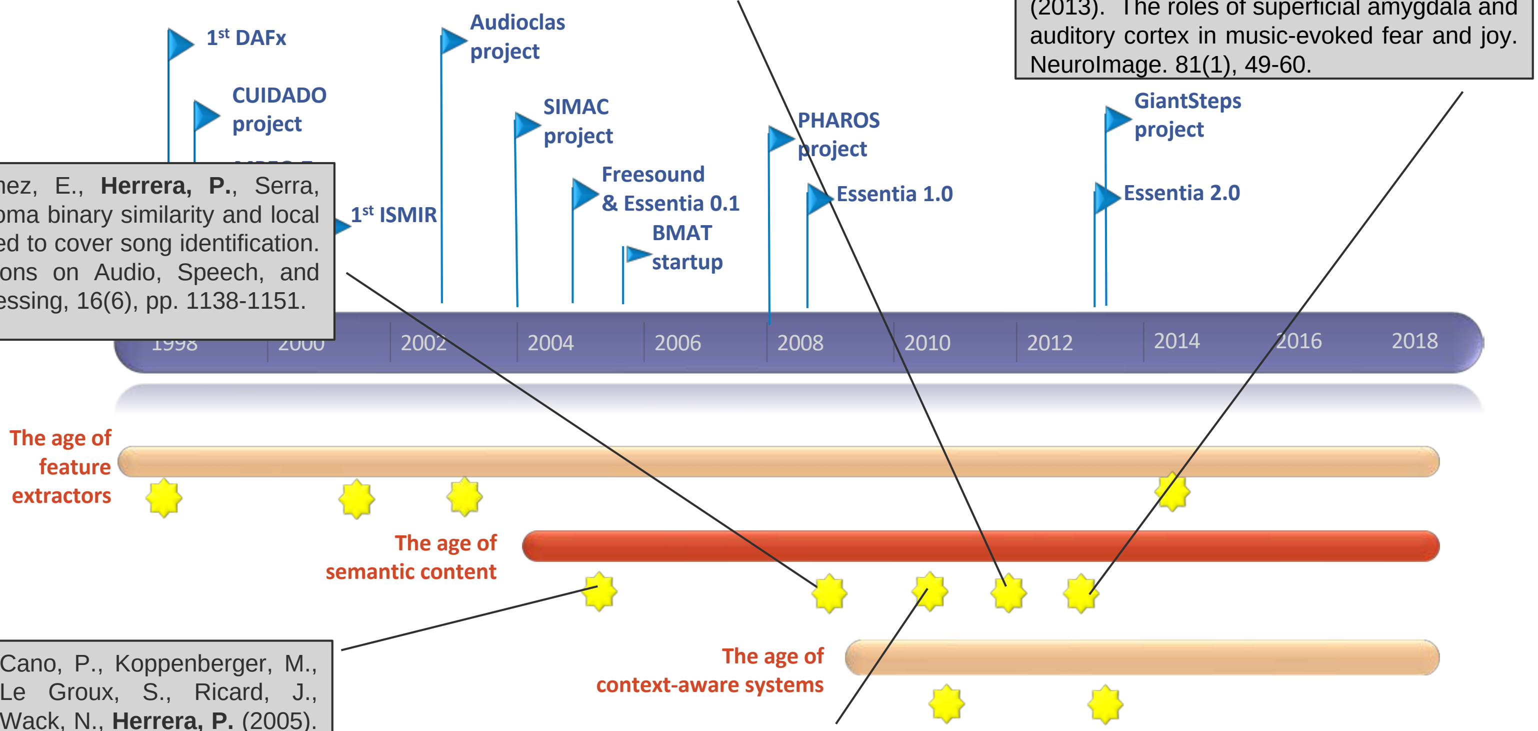
2. The age of semantic content

There are two kinds of sounds of rain: the sounds of raindrops upon the leaves of wu'tung and lotus, and the sounds of rain water coming down from the eaves into bamboo pails.
Lin Yutang, The importance of living (1937), p. 322.



The age of semantic content

- The *semantic gap:* connecting audio features and human concepts by means of models
- Semantic features (similarity, structure, mood, tonality, version, complexity, genre, energeticness, danceability, other *tags*...)
- Role of annotated collections
- SIMAC project: Semantic Interaction with Music Audio Contents (2004-2006):
  - Our first MTG-led EU project
  - Annotation, Collection Navigation, Personal tagger, Music Recommender
- AudioClas (2003-2005)
  - Essentia v0 (2005)
- BMAT (2005), first UPF start-up
- Freesound (2005 →)



Bogdanov, D., Serrà J., Wack N., **Herrera P.**, & Serra X. (2011). Unifying Low-level and High-level Music Similarity Measures. IEEE Transactions on Multimedia. 4, 687-701.

Koelsch, S., Skouras S., Fritz T., **Herrera P.**, Bonhage C., Küssner M. B., et al. (2013). The roles of superficial amygdala and auditory cortex in music-evoked fear and joy. NeuroImage. 81(1), 49-60.

Serrà, J., Gómez, E., **Herrera, P.**, Serra, X. (2008). Chroma binary similarity and local alignment applied to cover song identification. IEEE Transactions on Audio, Speech, and Language Processing, 16(6), pp. 1138-1151.

Cano, P., Koppenberger, M., Le Groux, S., Ricard, J., Wack, N., **Herrera, P.** (2005). "Nearest-neighbor sound annotation with a Wordnet taxonomy". Journal of Intelligent Information Systems, 24 (2), pp. 99-111.

Laurier, C., Meyers, O., Serrà, J., Blech, M., **Herrera, P.**, Serra, X. (2010). Indexing Music by Mood: Design and Integration of an Automatic Content-based Annotator. Multimedia Tools and Applications. 48(1), 161-184.

Similarity

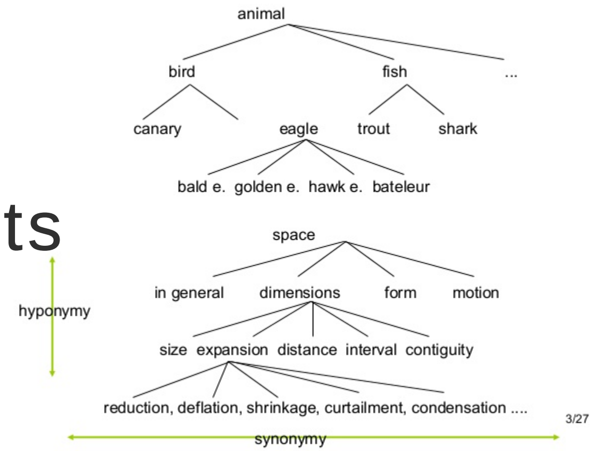
Bogdanov, D., Serrà J., Wack N., **Herrera P.**, & Serra X. (2011). "Unifying Low-level and High-level Music Similarity Measures". IEEE Transactions on Multimedia. 4, 687-701. (Journal h-index: 101; Journal IF 2016: 3.509; Q1 in Computer Science Applications journals; 72 citations)

- Development and evaluation of several polyphonic music similarity distances (with different abstraction levels)
- Exploration of similarity through classification
- Best results with a hybrid euclidean distance combining timbral, temporal, tonal and semantic descriptors (LLD+HLD)
- Among top systems in MIREX 2009 and 2010

Tags

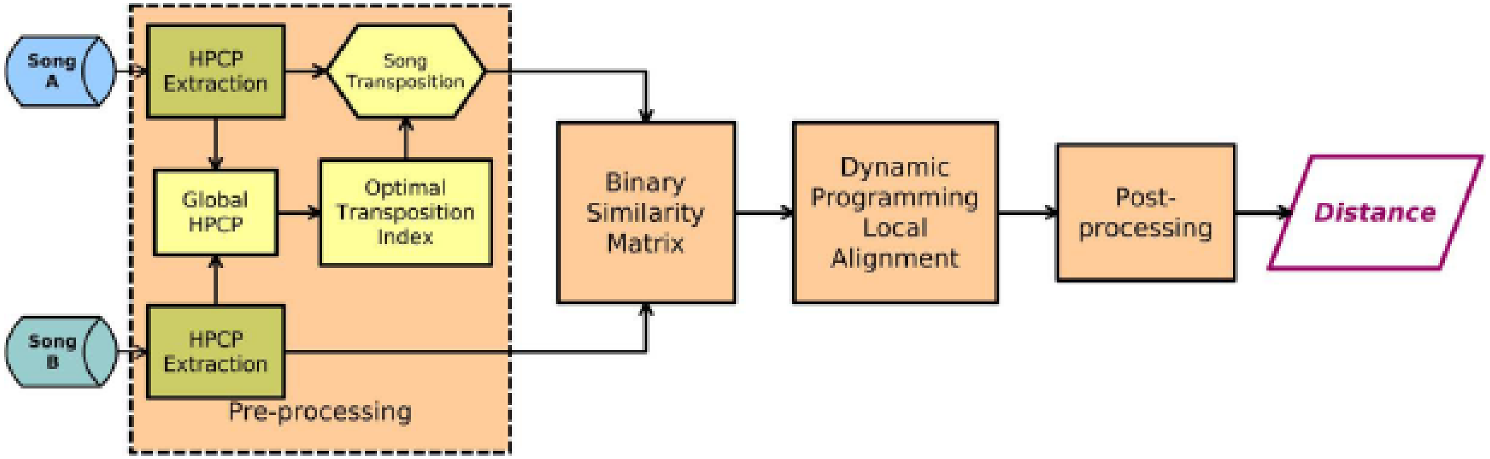
Cano, P., Koppenberger, M., Le Groux, S., Ricard, J., Wack, N., **Herrera, P.** (2005). "Nearest-neighbor sound annotation with a Wordnet taxonomy". Journal of Intelligent Information Systems, 24 (2), pp. 99-111. (Journal h-index: 47; Journal IF 2016; 1.107; Q2 in Information Systems journals; 20 citations)

- How to classify/multi-tag thousands of categories?
- Wordnet as the backbone of taxonomical knowledge and inference
- First use of Wordnet in MIR
- 30% accuracy for 1600 concepts and over 50000 instances
- Features robust to transcoding
- Semantics as network of concepts



Covers

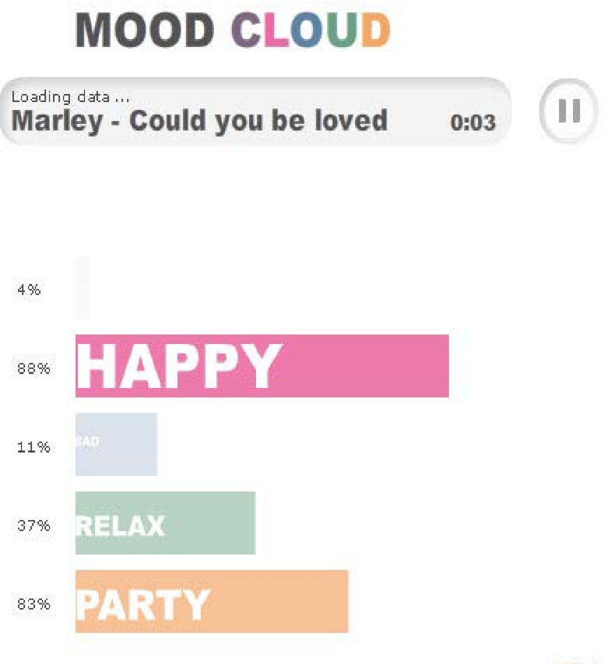
Serrà, J., Gómez, E., **Herrera, P.**, Serra, X. (2008). "Chroma binary similarity and local alignment applied to cover song identification". IEEE Transactions on Audio, Speech, and Language Processing, 16(6), pp. 1138-1151. (Journal h-index: 91; Journal IF 2016: 2.491; Q1 in Acoustics and Ultrasonics journals; 245 citations)



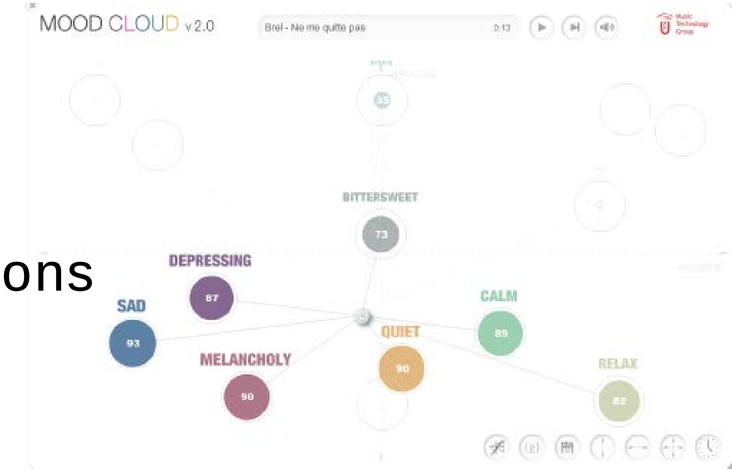
- Tonal and tempo invariance required to match tracks
- 1st systematic evaluation of factors influencing cover identification
- Best system in MIREX 2008 and 2009
- “Understanding music understanding” pays for improving technologies

Mood

Laurier, C., Meyers, O., Serrà, J., Blech, M., **Herrera, P.**, Serra, X. (2010). "Indexing Music by Mood: Design and Integration of an Automatic Content-based Annotator". Multimedia Tools and Applications. 48(1), 161-184. (Journal h-index: 45; Journal IF 2016: 1.541; Q2 in Computer Networks and Communications journals; 40 citations)



- Modeling happy, sad, angry, relaxed and "NOT-" categories
- Annotations from social networks+expert supervision/filtering
- Importance of spectral complexity, dissonance and mode
- SVM-based Multimedia mood annotator
- Web-based "original" prototype
- Very good results in several MIREX editions
- Moodcloud prototypes



Mood

Koelsch, S., Skouras S., Fritz T., **Herrera P.**, Bonhage C., Küssner M. B., et al. (2013). The roles of superficial amygdala and auditory cortex in music-evoked fear and joy. NeuroImage. 81(1), 49-60. (Journal h-index: 307; Journal IF 2017: **5.426**; Q1 in Cognitive Neuroscience journals; 79 citations)

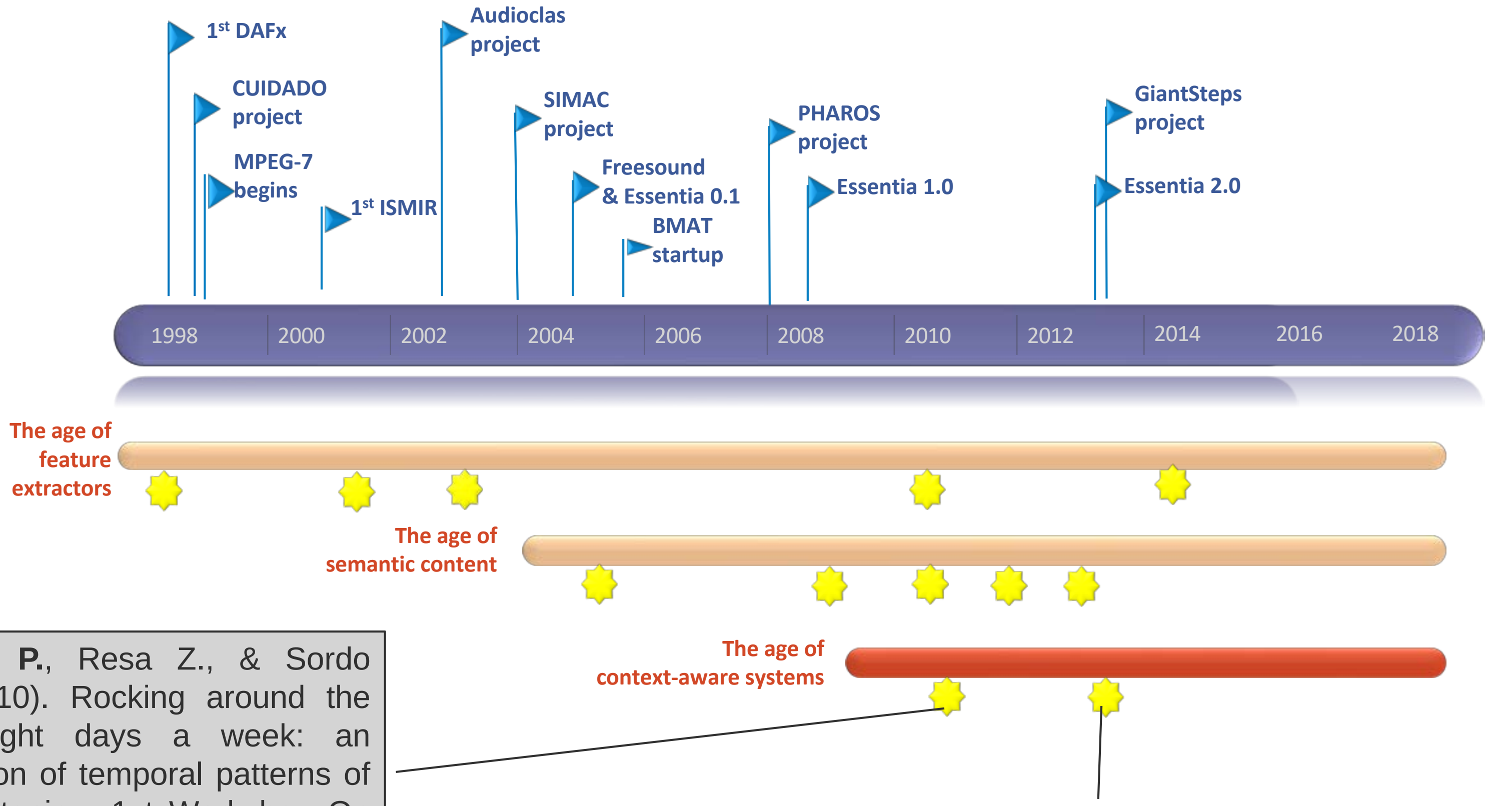
- Use of descriptors to confirm stimuli selection for studies on the neural bases of musical emotions
- Use of descriptors to specify acoustical differences between stimuli
- Unexpected connections between visual imagery and emotional music (especially fear-evoking) (mediated by the amygdala)
- The auditory cortex as a central hub of an extended affective-attentional network

3 . The age of context-aware systems

*My cow is not pretty,
but it's pretty to me*
—David Lynch

The age of context-aware systems

- Any information that can be used to characterize the situation of users, content and applications
  - Listener context (time, space, activity, preference, usage history, biography…)
  - Audio content context (linked media, within-track, between-tracks, styles, history, geography…)
- The age of music recommenders
- No targetted project
- Embedded research (somehow) in
  - PHAROS (2007-2009)
  - EmCAP (2006-2008)

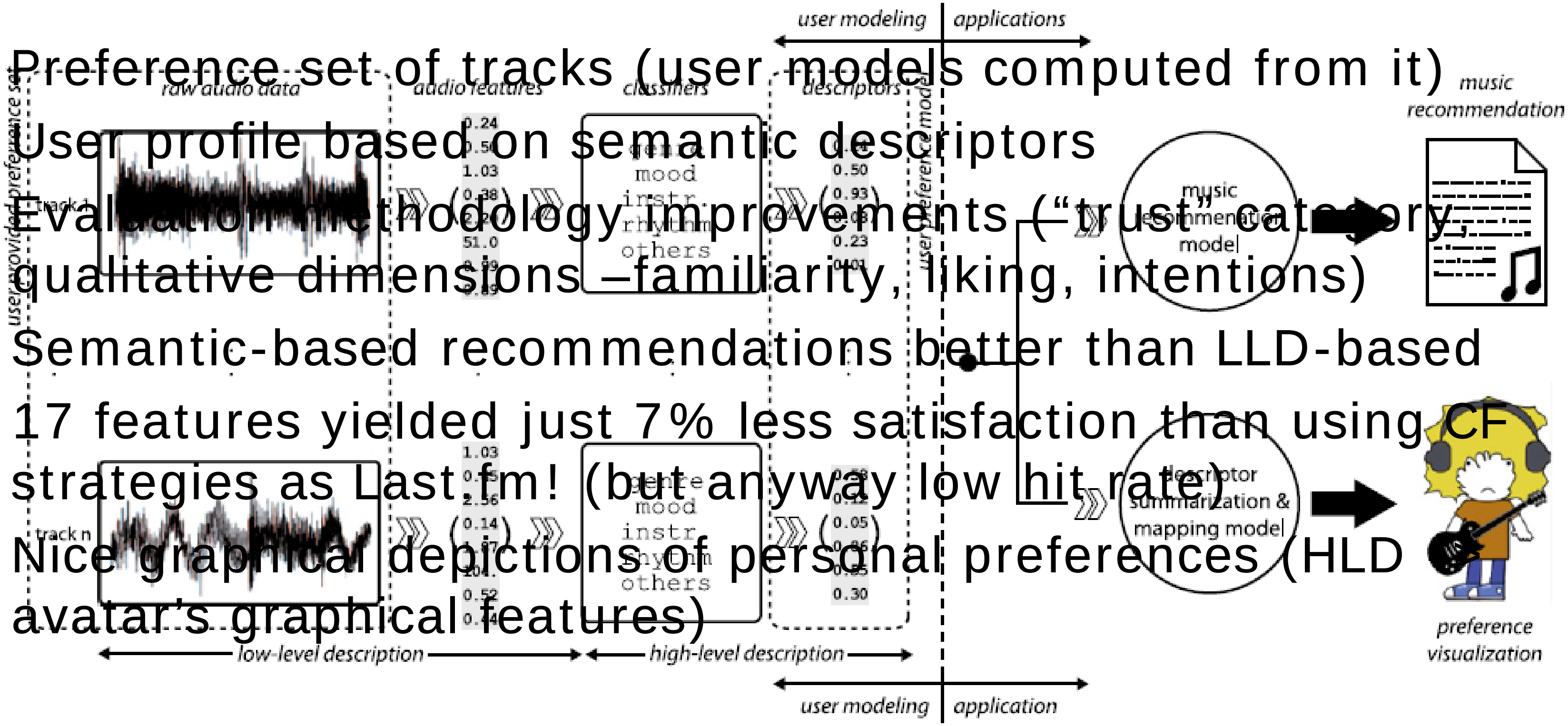


Herrera, P., Resa Z., & Sordo M. (2010). Rocking around the clock eight days a week: an exploration of temporal patterns of music listening. 1st Workshop On Music Recommendation And Discovery (WOMRAD), ACM RecSys, 2010, Barcelona, Spain.

Bogdanov, D., Haro, M., Fuhrmann, F., Xambó, A., Gómez, E. & **Herrera, P.** (2013) Semantic content-based music recommendation and visualization based on user preference examples. Information Processing and Management, 49(1), 13-33.

User profiles

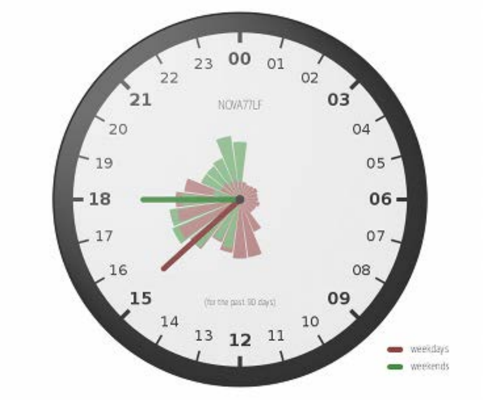
Bogdanov, D., Haro, M., Fuhrmann, F., Xambó, A., Gómez, E. & **Herrera, P.** (2013) "Semantic content-based music recommendation and visualization based on user preference examples". Information Processing and Management, 49(1), 13-33. (Journal h-index: 84; Journal IF 2017: 3.444, Q1 in Information Processing journals; 71 citations)



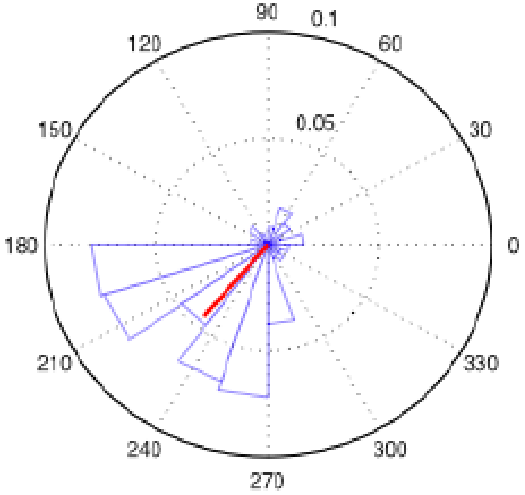
- Preference set of tracks (user models computed from it)
- User profile based on semantic descriptors
- Evaluation methodology improvements (“trust” category; qualitative dimensions –familiarity, liking, intentions)
- Semantic-based recommendations better than LLD-based
- 17 features yielded just 7% less satisfaction than using CF strategies as Last.fm! (but anyway low hit rate)
- Nice graphical depictions of personal preferences (HLD avatar’s graphical features)

Time

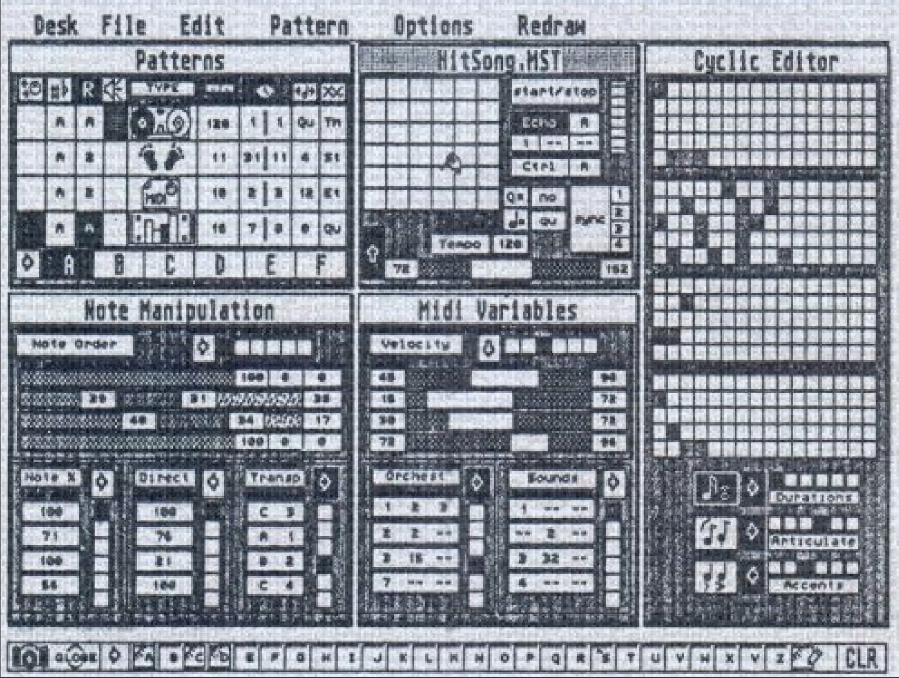
Herrera, P., Resa Z., & Sordo M. (2010). Rocking around the clock eight days a week: an exploration of temporal patterns of music listening. 1st Workshop On Music Recommendation And Discovery (WOMRAD), ACM RecSys, 2010, Barcelona, Spain. (27 citations, WIRED magazine short note, last.fm idea adoption)



- AFAIK, first paper on this subject (others have been following since then)
- First MIR paper showing the possibilities of circular statistics
- Listening genre/artist choices dependent on day and time
- Some listeners more influenced than others
- Further research by other people made this topic evolve



4. The age of creative systems?



In the future, you won't buy artists' works; you'll buy software that makes original pieces of "their" works, or that recreates their way of looking at things.

Brian Eno,

Wired 3.05, May 1995, p. 150

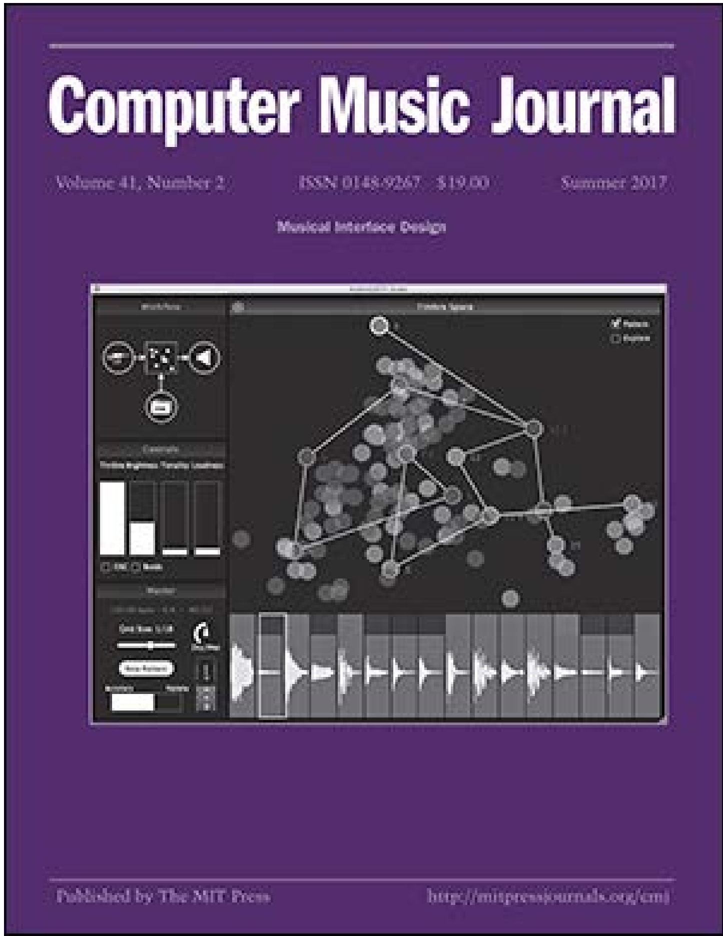
- Creation =? Features + Meaning + Context
- Creation =? Description + Modelling + Interaction

- "Creative MIR" (late breaking session, ISMIR 2013)
- MIRES roadmap (2013):
  - Content-based sound processing
  - Computer-aided composition
  - Databases for music and sound production
  - Content and context-aware Djing and improvisation
- GiantSteps project (2013-2015)
- Creative systems to enhance music creativity (not for the sake of showing creativity)
- Evaluation issues

MIR and music creation

Nuanáin, C. Ó., **Herrera P.**, & Jordà S. (2017). Rhythmic Concatenative Synthesis for Electronic Music: Techniques, Implementation, and Evaluation. Computer Music Journal. 41(2), 21-37 (Journal h-index: 35; Journal IF 2016: 0.405; Q1 in Music Journals; 0 citations; a shorter version was selected best paper in NIME 2016)

- RhythmCAT, a user-friendly plug-in for generating rhythmic loops that model the timbre and rhythm of an initial target
- Up-to-date state of the art
- 2D interactive timbre space to modulate, in real-time, the concatenation sequence
- 3-tiered evaluation: system, performer, listener

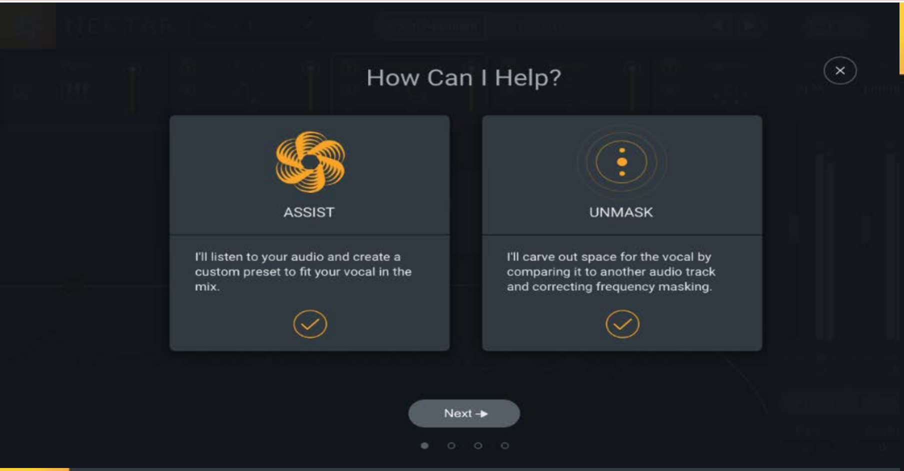


Computer Music Journal

Volume 41, Number 2 ISSN 0148-9267 $19.00 Summer 2017

Musical Interface Design

Published by The MIT Press http://mitpressjournals.org/cmj



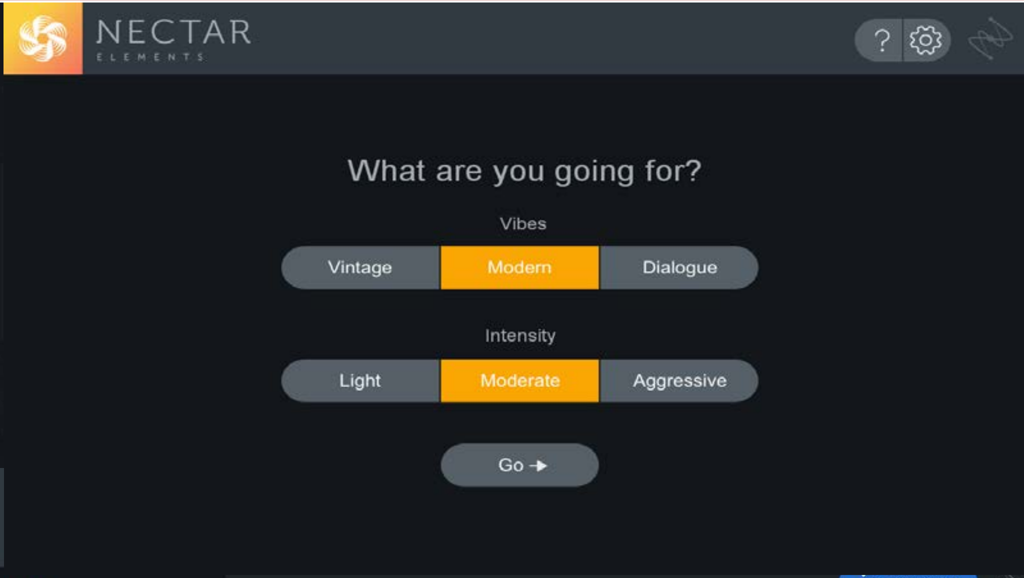
How Can I Help?

ASSIST

I'll listen to your audio and create a custom preset to fit your vocal in the mix.

UNMASK

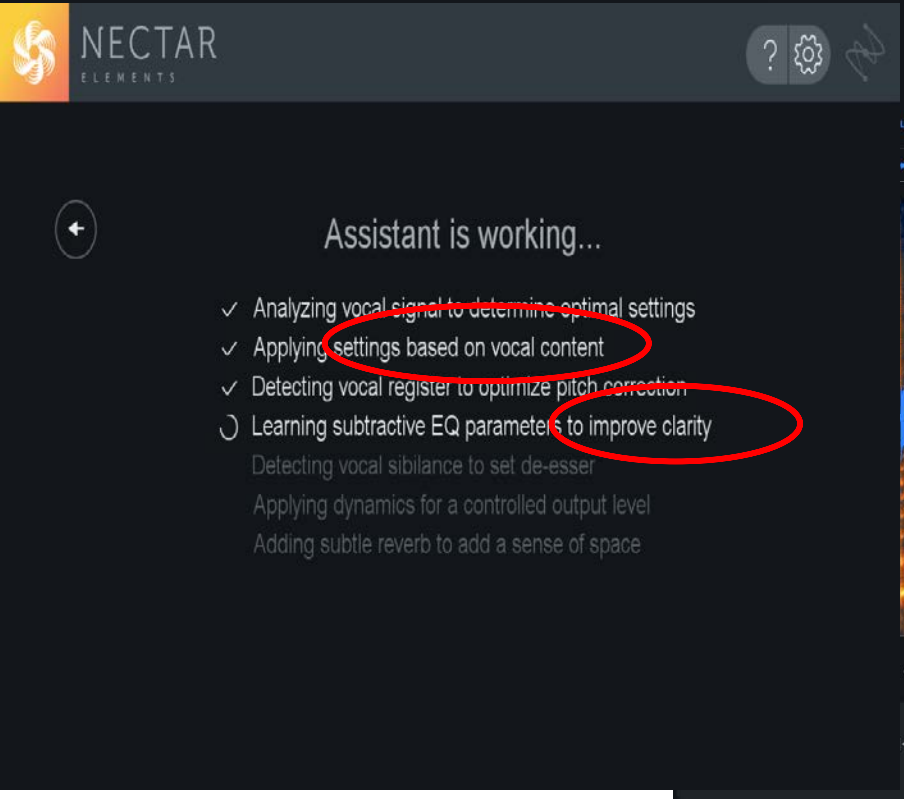
I'll carve out space for the vocal by comparing it to another audio track and correcting frequency masking.



What are you going for?

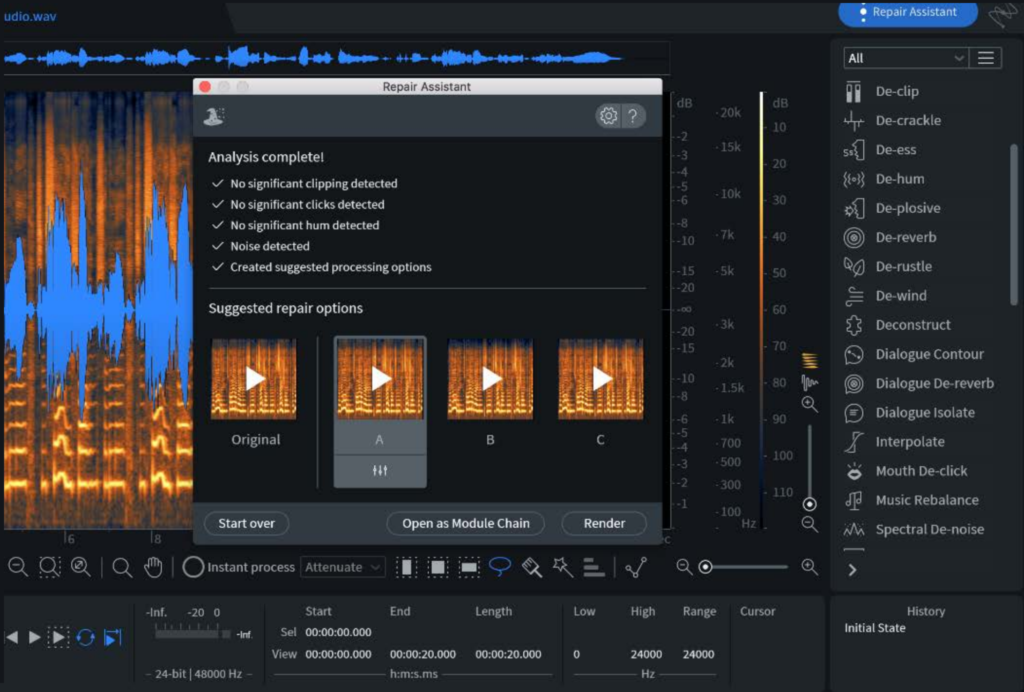
Vibes: Vintage | **Modern** | Dialogue

Intensity: Light | **Moderate** | Aggressive



Assistant is working...

- ✓ Analyzing vocal signal to determine optimal settings
- ✓ Applying settings based on vocal content
- ✓ Detecting vocal register to optimize pitch correction
- Learning subtractive EQ parameters to improve clarity
- Detecting vocal sibilance to set de-esser
- Applying dynamics for a controlled output level
- Adding subtle reverb to add a sense of space



Repair Assistant

Analysis complete!

- ✓ No significant clipping detected
- ✓ No significant clicks detected
- ✓ No significant hum detected
- ✓ Noise detected
- ✓ Created suggested processing options

Suggested repair options

"time has arrived for a paradigm shift towards doing use-inspired basic research where the focus on 'information' shifts towards 'interaction”

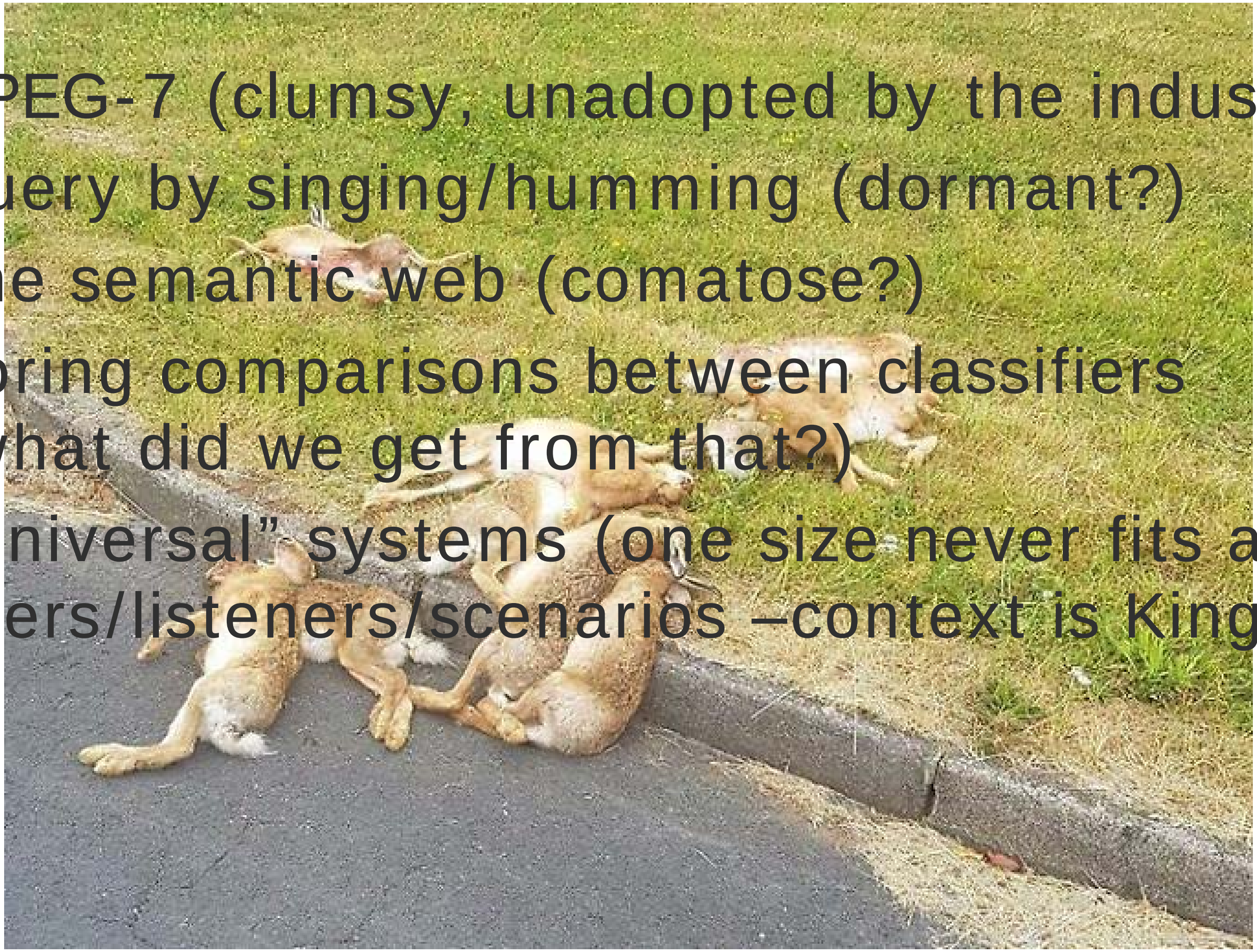
MIIR?

Concluding thoughts

Bumps found on the road

- Western-centric views (though improving)
- Poor methodology (though improving)
- Lack of replicability (though improving)
- Poor understanding of music understanding (though improving) (e.g., *bag of frames*)
- The tyranny of big numbers (sometimes a few cases give you a better insight)
- Banalization of music experiencing (emotions are not tags)
- Technology neutrality assumption (though...)
- MIR as pure engineering (is this just an optimization game?)

Corpses left on the road



- MPEG-7 (clumsy, unadopted by the industry)
- Query by singing/humming (dormant?)
- The semantic web (comatose?)
- Boring comparisons between classifiers (what did we get from that?)
- “Universal” systems (one size never fits all users/listeners/scenarios –context is King!)

Concluding thoughts

- A mature discipline has been developed along 3 or 4 different “ages”
- Specific problems, techniques and communication channels are set and clear
- Performance improved in all the addressed problems
- Still challenging open issues (e.g., similarity -still poorly understood, better engineered)
- Do we better understand music and music experiencing? (prediction=?understanding)
- Lack of theoretical models (of interactions, of users, of learning, of operations on information...)

THANKS FOR
LISTENING!
IT'S
TIME FOR
QUESTIONS!