21st Vienna Deep Learning Meetup – Hot Topics

Deep Learning Approaches for Music Information Retrieval at ISMIR conference, Sep. 2018

15 Oct 2018 by Thomas Lidy

ISMIR 2018 in a few numbers

- 5 days in Paris
- 450+ attendees, 280 unique scientific authors
- 8 tutorial sessions (including one by Tom, Alex and Sebastian)
- 100+ scientific paper presentations and posters
- 50 late-breaking demo papers
- 2 inspiring keynotes

Industry presence

DEEZER

Spotify

gracenote
A NIELSEN COMPANY

pandora

SHAZAM

NATIVE INSTRUMENTS
THE FUTURE OF SOUND

ACRCloud

smule
Let's music together

Google

Last but not least :

musimap
cognitive technologies

Tutorials

When	Sunday 2018-09-23 from 09:00 to 12:30 and from 14:00 to 17:30
What	Tutorials T1, T2, T3, T4, T5, T6, T7, T8
Where	Télécom ParisTech How to get there ?

There will be 4 parallel tutorials in the morning, and another 4 parallel tutorials in the afternoon.

Morning Tutorials

- T1 (Room B310) Open Source and Reproducible MIR Research (Speakers: Brian McFee, Thor Kell)
- T2 (Room Grenat) Computational Approaches for Analysis of Non-Western Music Traditions (Speakers: Xavier Serra, Martin Clayton, Barış Bozkurt)
- T3 (Room B312) Statistical Analysis of Results in Music Information Retrieval: Why and How (Speakers: Julian Urbano and Arthur Flexer)
- T4 (Room: Emeraude) Music Separation with DNNs: Making It Work (Speakers: Antoine Liutkus, Fabian-Robert Stöter)

Afternoon Tutorials

- T5 (Room Estaunié) Deep Learning for MIR (Speakers: Alexander Schindler, Thomas Lidy, Sebastian Böck)
- T6 (Room Grenat) Fundamental Frequency Estimation in Music (Speakers: Rachel Bittner, Alain de Chevigne, Johanna Devaney)
- T7 (Room B312) Optical Music Recognition for Dummies (Speakers: Jorge Calvo-Zaragoza, Jan Hajič jr., Alexander Pacha, Ichiro Fujinaga)
- T8 (Room B310) Overview and New Challenges of Music Recommendation Research in 2018 (Speakers: Markus Schedl, Peter Knees, Fabien Gouyon)

available on Github!

DL Tutorial (Sunday)

[redacted] SINGING VOICE DETECTION

Jan Schlüter Bernhard Lehner

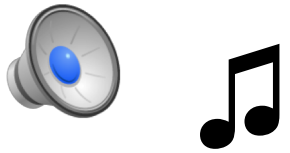
Austrian Research Institute for Artificial Intelligence
OFAI

JOHANNES KEPLER UNIVERSITY LINZ
JKU

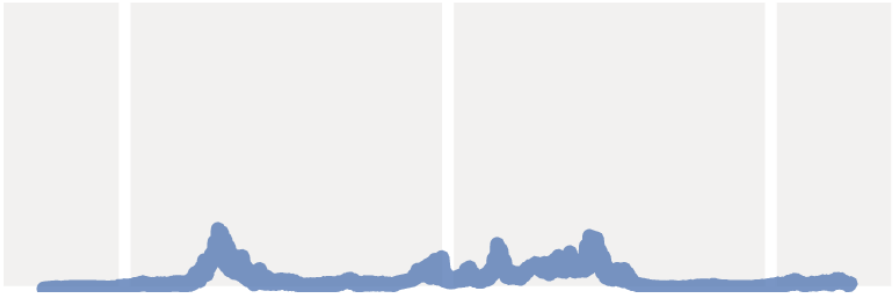
2018-09-25
ISMIR
Paris

github.com/f0k/ismir2018

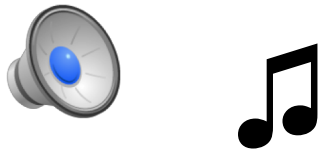
You will hear an example.
Raise your hand if you hear singing voice.



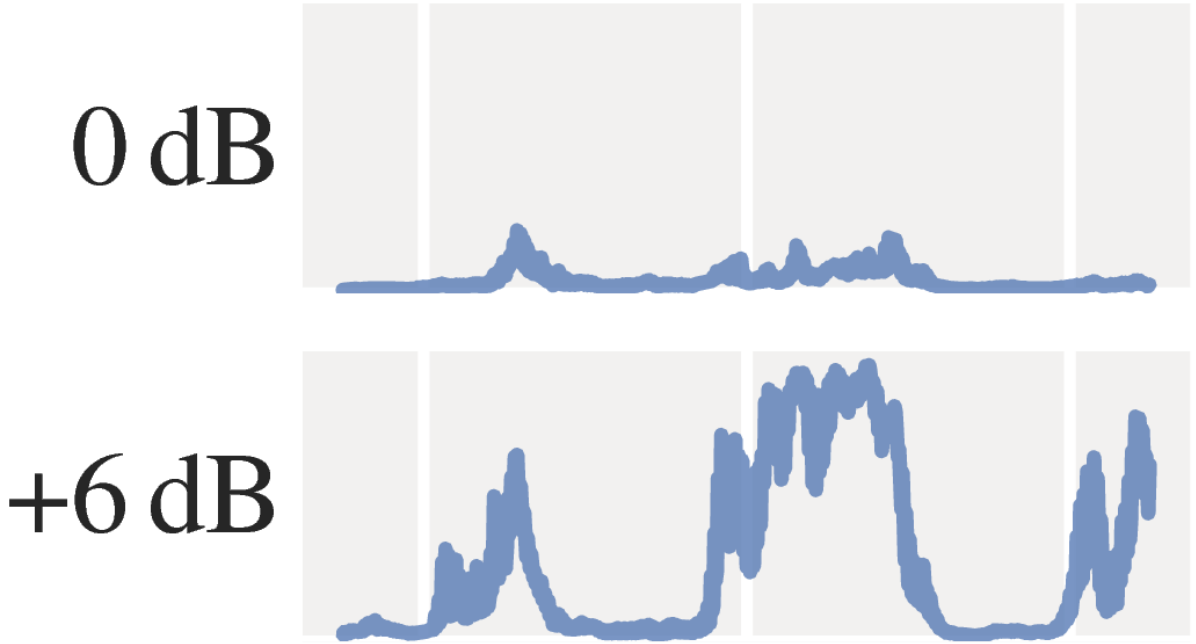
- These are the predictions of a state-of-the art neural network. No voice throughout. Good.



Let us turn up the volume a bit.
Raise your hand if you hear singing voice.



The network detects singing voice. It learned that everything loud is voice. How do we fix it?



ZERO-MEAN CONVOLUTIONS FOR LEVEL-INVARIANT SINGING VOICE DETECTION

Jan Schlüter

Bernhard Lehner

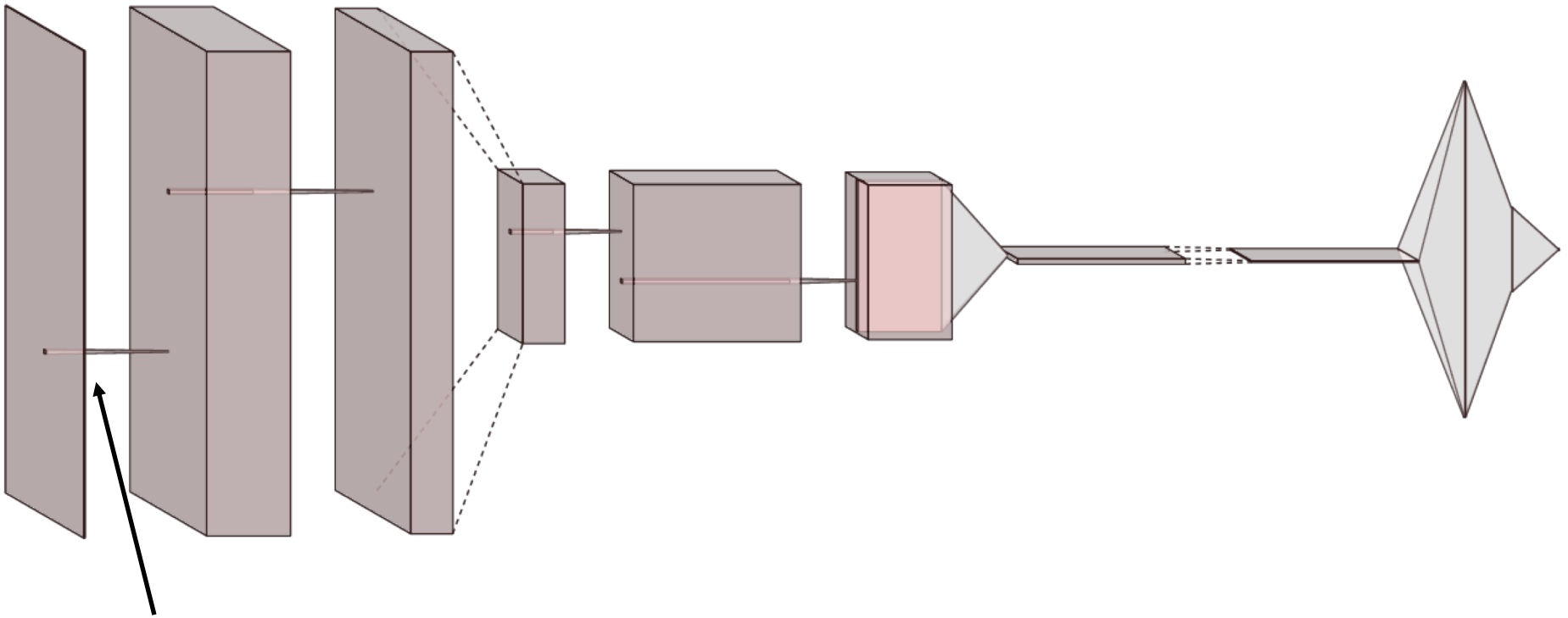
Austrian Research Institute for Artificial Intelligence

JOHANNES KEPLER UNIVERSITY LINZ

2018-09-25
ISMIR
Paris

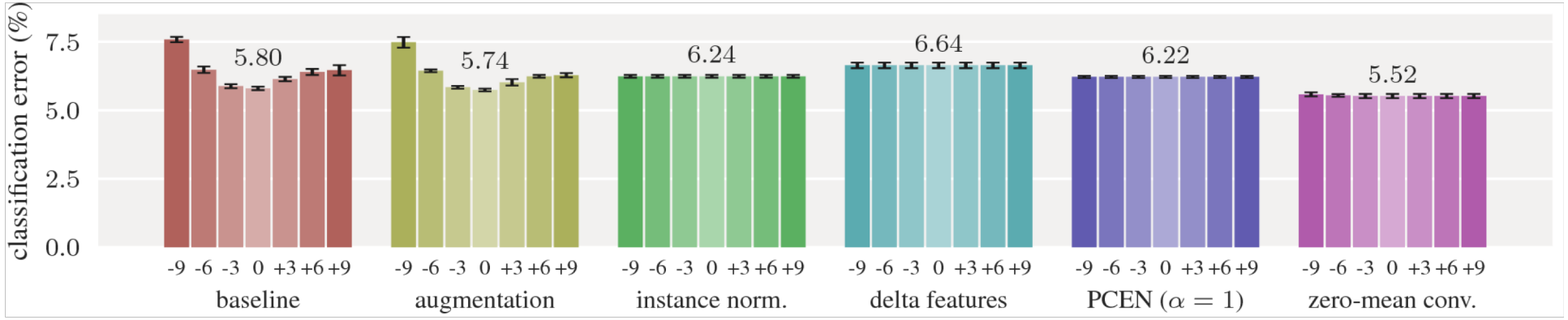
github.com/f0k/ismir2018

Input: logarithmic-magnitude spectrogram
→ logarithm turns input scale into input shift:
log(a x) = log(a) + log(x)



First layer: zero-mean convolutions
→ constrain each filter to sum up to zero
→ removes input shift

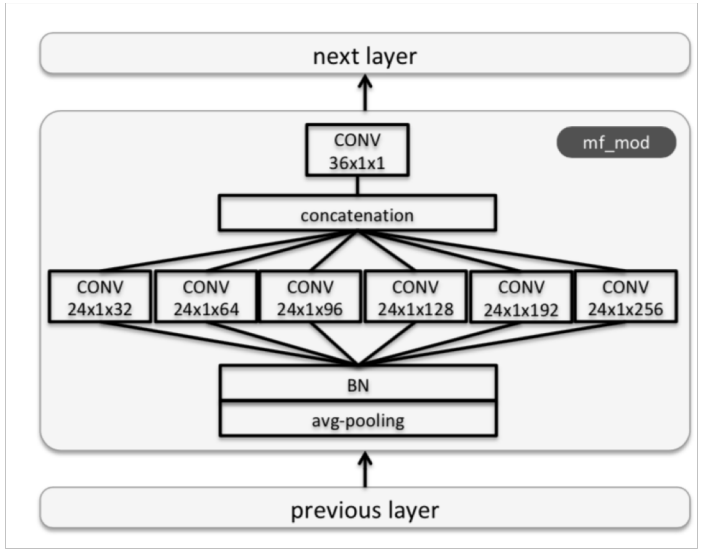
Other attempts (e.g., augmentation, input normalization): Either still depend on input scale, or worse classification error.

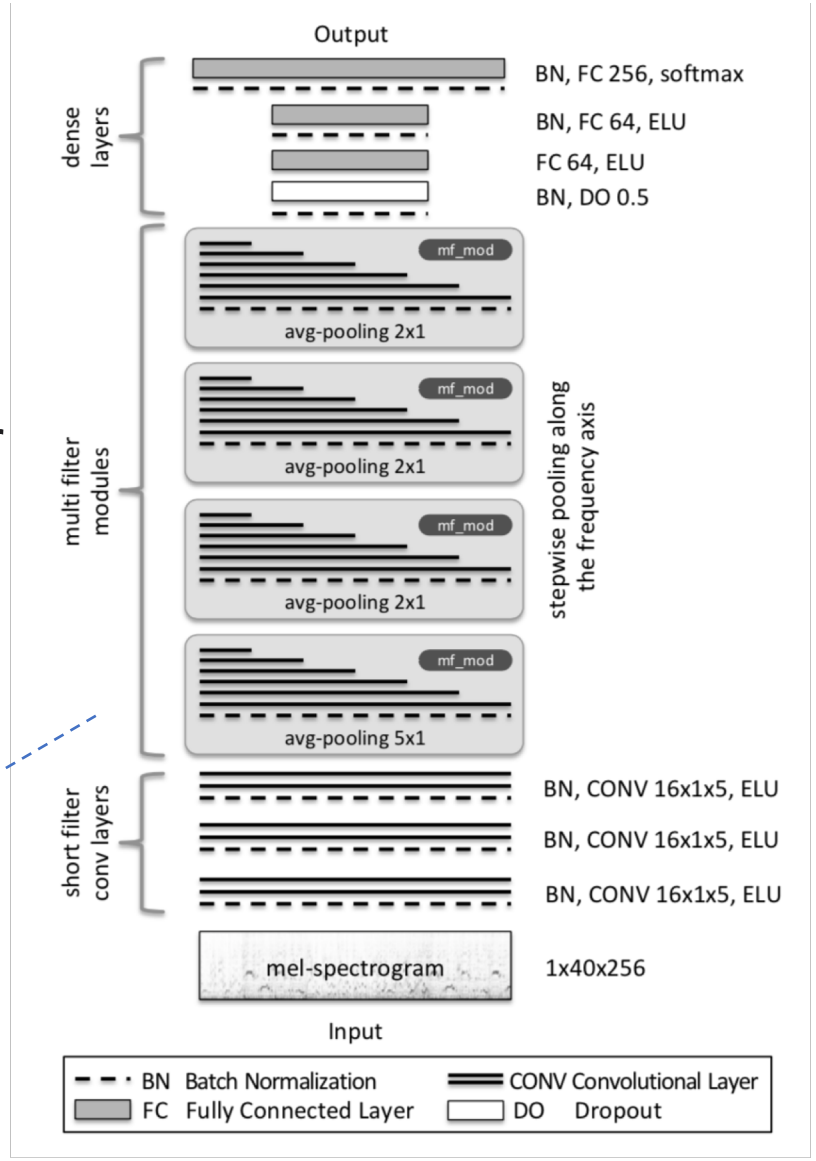


Zero-mean Convolution: Level-invariant, slightly better than baseline.
→ Could be useful for other audio tasks!

Tempo (Bpm) Estimation

- A single-step approach to musical tempo estimation using a CNN *(Hendrik Schreiber, Audiolabs)*
- proposed a novel deep learning approach for one-step BPM detection
- very interesting DL model architecture





Key Detection (“genre agnostic”)

- (B-17) Genre-Agnostic Key Classification With Convolutional Neural Networks
- Filip Korzeniowski and Gerhard Widmer
- Key classification without genre profiles needed anymore

(a) KeyNet Architecture

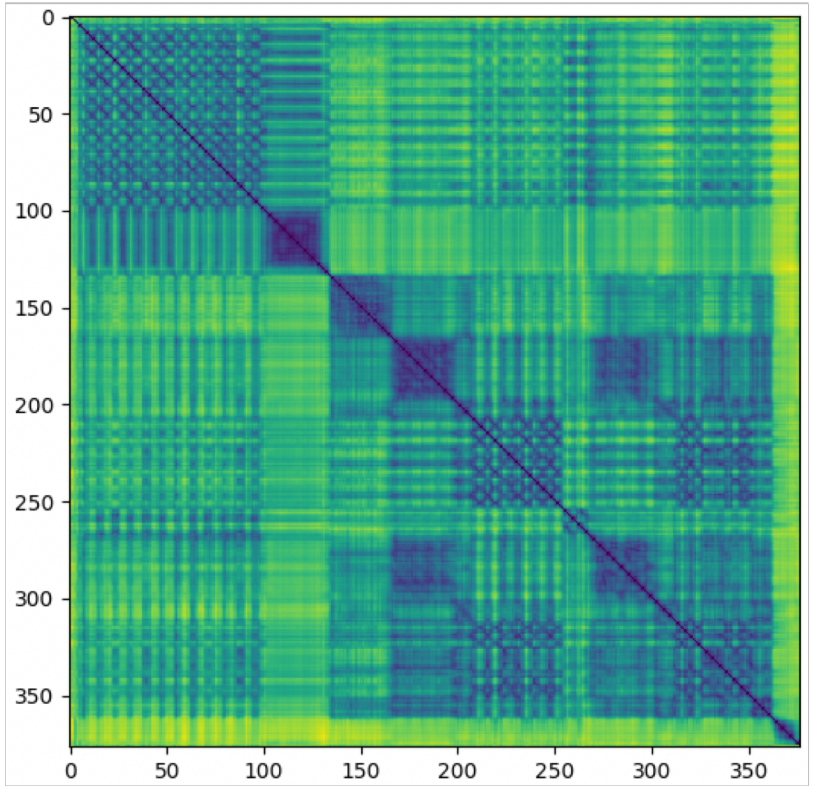
Layer Type	FMaps	Params
Input		
Conv-ELU	N_f	5×5
Conv-ELU	N_f	5×5
Conv-ELU	N_f	5×5
Conv-ELU	N_f	5×5
Conv-ELU	N_f	5×5
Dense-ELU		$2 \cdot N_f$
Pool-Time Avg.		
Dense-Softmax		24

(b) AllConv Architecture

Layer Type	FMaps	Params
Input		
Conv-ELU	N_f	5×5
Conv-ELU	N_f	3×3
Pool-Max		2×2
Conv-ELU	$2N_f$	3×3
Conv-ELU	$2N_f$	3×3
Pool-Max		2×2
Conv-ELU	$4N_f$	3×3
Conv-ELU	$4N_f$	3×3
Pool-Max		2×2
Conv-ELU	$8N_f$	3×3
Conv-ELU	$8N_f$	3×3
Conv-ELU	24	1×1
Pool-Global Avg.		
Softmax		

Music Segmentation

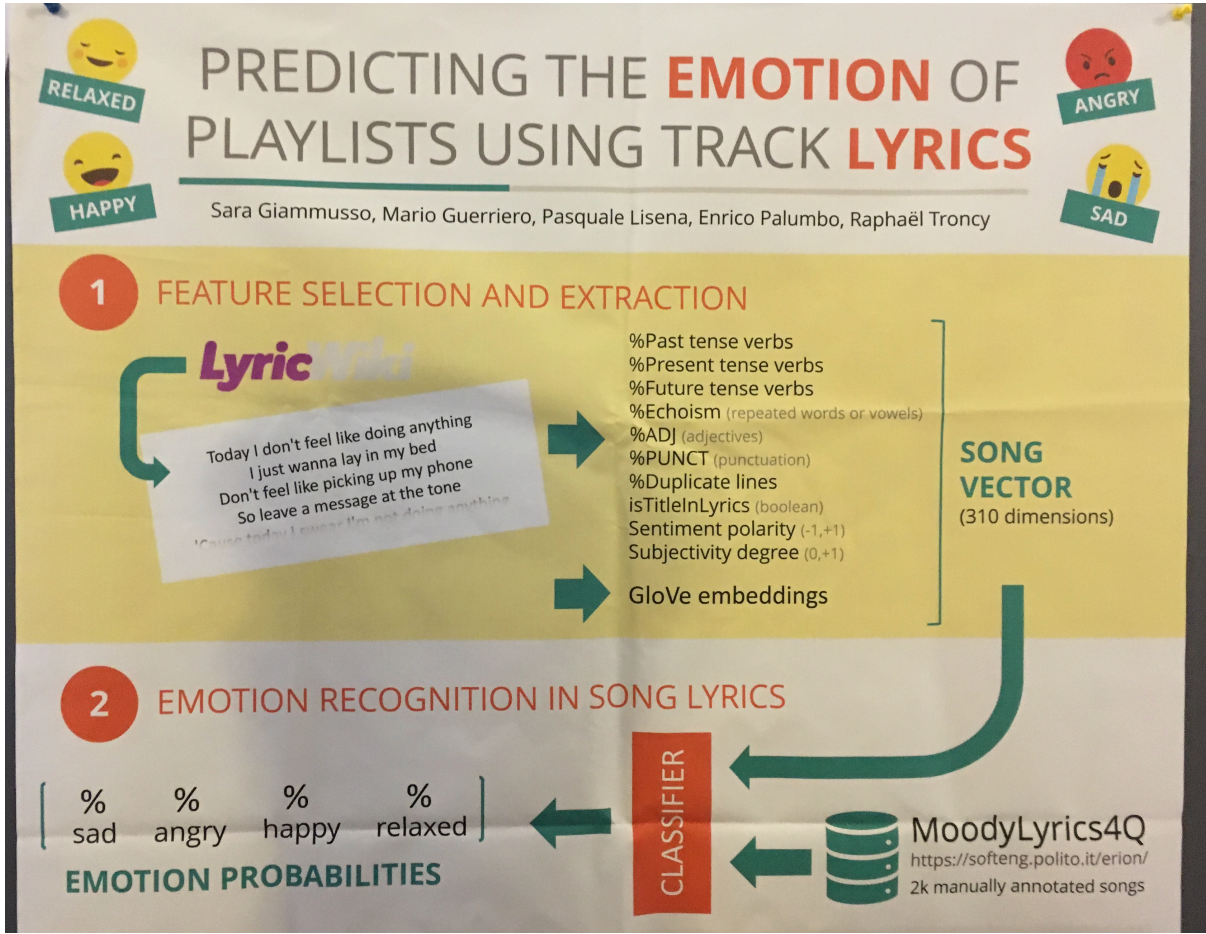
- (Demo 48) Unsupervised Deep Feature Learning For Music Segmentation
  Matthew McCallum, Gracenote
- Unsupervised (!) Triplet Networks
- means they have no labelled training data!
  (take positive example from middle of song,
  negative from beginning and end)
- works to learn a ***sufficiently good*** feature representation
- then traditional auto regression for self-similarity



Self similarity matrix of The Beatles, “Birth-day”, computed on Euclidean distance between deep embeddings of beat synchronized CQT features.

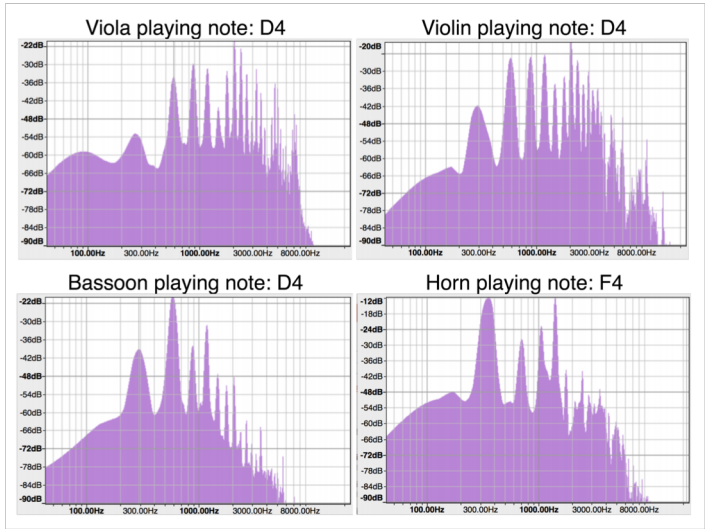
Emotion Detection from Lyrics

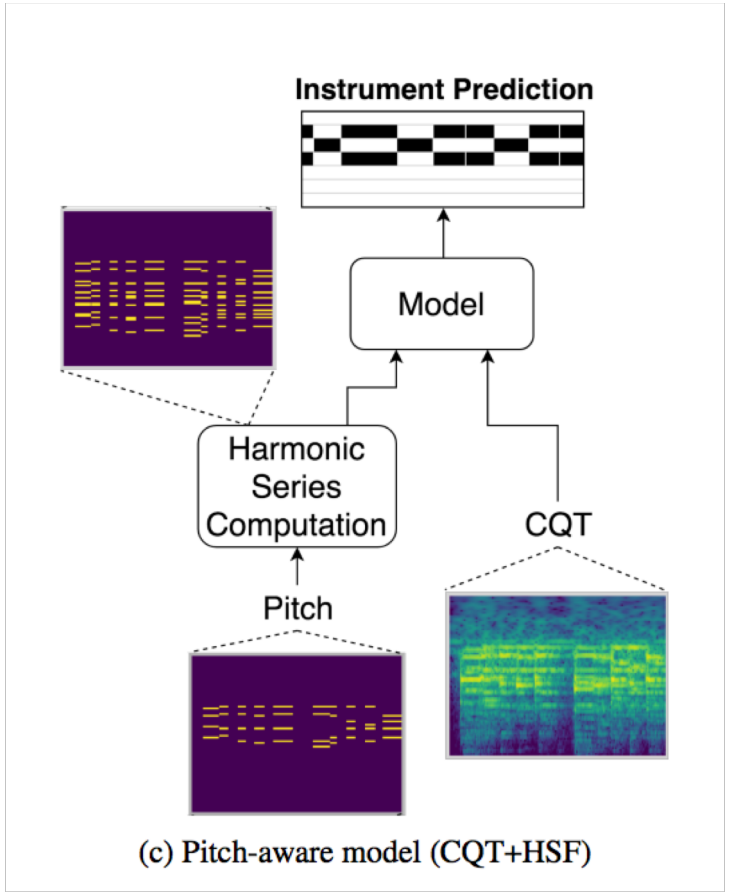
- [PREDICTING THE EMOTION OF PLAYLISTS USING TRACK LYRICS]
- playlists downloaded from Spotify with emotional word in title
- analysis only based on **lyrics**
- emotion average with NN (works better than other ML)
- 4 moods: sad, angry, happy, relaxed
- demo: data.doremus.org/emotion



Instrument Recognition (1)

- Frame-level Instrument Recognition by Timbre and Pitch (Yun-Ning Hung , Citi Taiwan)
- Comparison of 3 DL models for instrument recognition
- Automatic extraction of melody information
- Pitch (melody) information + acoustic features lead to better results
- 89,6 % average accuracy on the detection of 7 instruments.

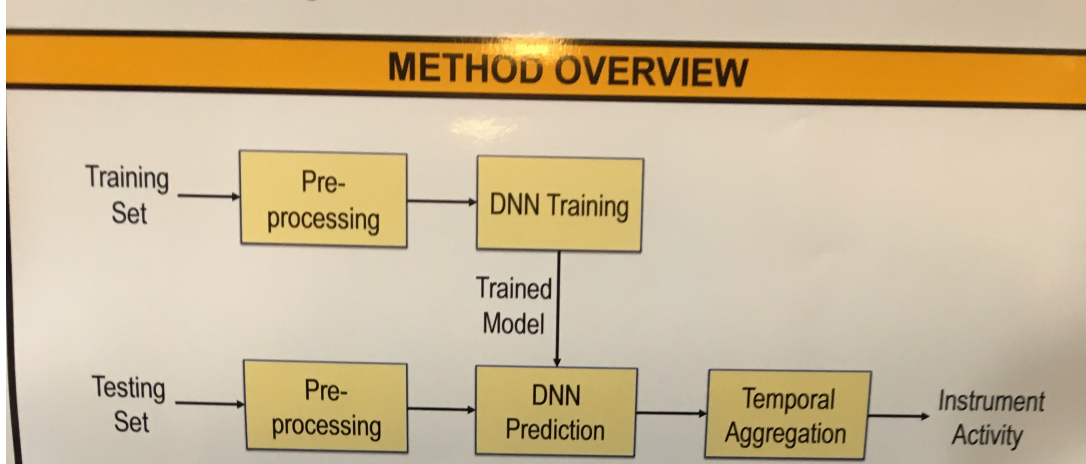




(c) Pitch-aware model (CQT+HSF)

Instrument Recognition (2)

- (E-5) Instrument Activity Detection in Polyphonic Music using Deep Neural Networks – GRACENOTE
  - 18 instruments
    800k examples
    (very inbalanced: 720k drumset, low classes have only 10k)
  - compared FNN, CNN, CRNN (last two are better)
  - input: 1 second log Mel spectrograms (96x85 frames)



DATASET

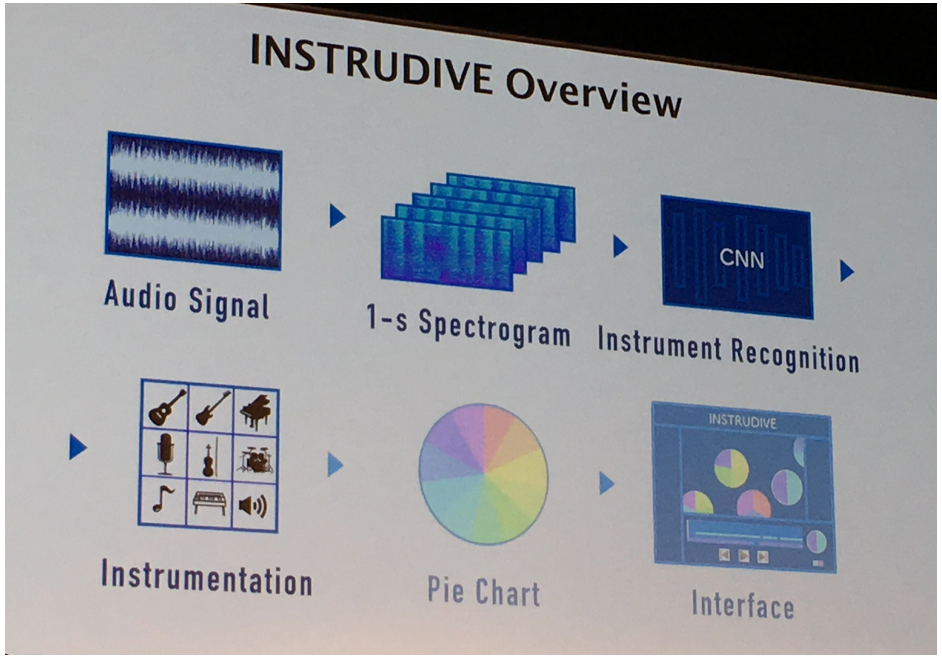
- **MedleyDB** and **Mixing Secrets** multi-track datasets are combined
- Automatic instrument activity annotation available for all stems in the mix
- We choose **18** most frequently occurring instruments

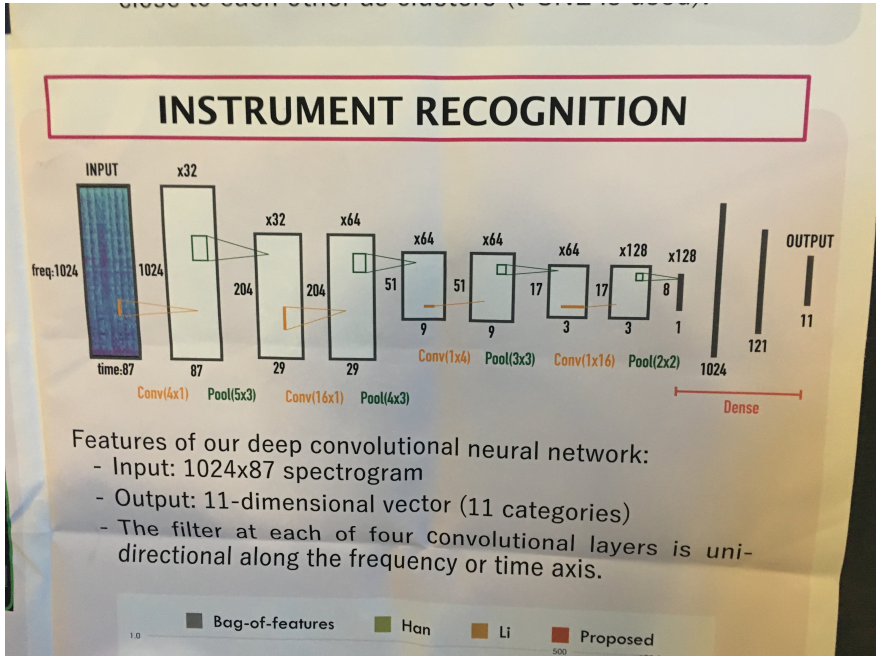
Dataset Distribution: Bold indicates AUC > 0.8 after evaluation

		Training		Testing	
Instrument	Abbr	Tracks	Instances	Tracks	Instances
drum set	**dru**	300	720036	79	15957
electric bass	**bgtr**	253	620592	62	13344
male singer	**ms**	200	351384	62	10038
distorted electric guitar	**dgtr**	171	396204	40	7522
clean electric guitar	cgtr	119	225456	34	5875
synthesizer	**syn**	118	295524	33	5712
acoustic guitar	**agtr**	91	230556	25	5241
piano	pf	89	187536	24	4063
vocalists	vox	84	154596	12	1895
female singer	fs	79	149232	23	3733
string section	str	24	39444	10	1278
electric piano	epf	24	52680	14	2075
electronic organ	eorg	22	39516	11	2117
double bass	db	21	40116	9	1786
cello	**vc**	13	22176	9	1623
violin	**vn**	10	28452	15	2385
tabla	**tab**	9	41640	3	806
flute	**fl**	7	9972	7	1171

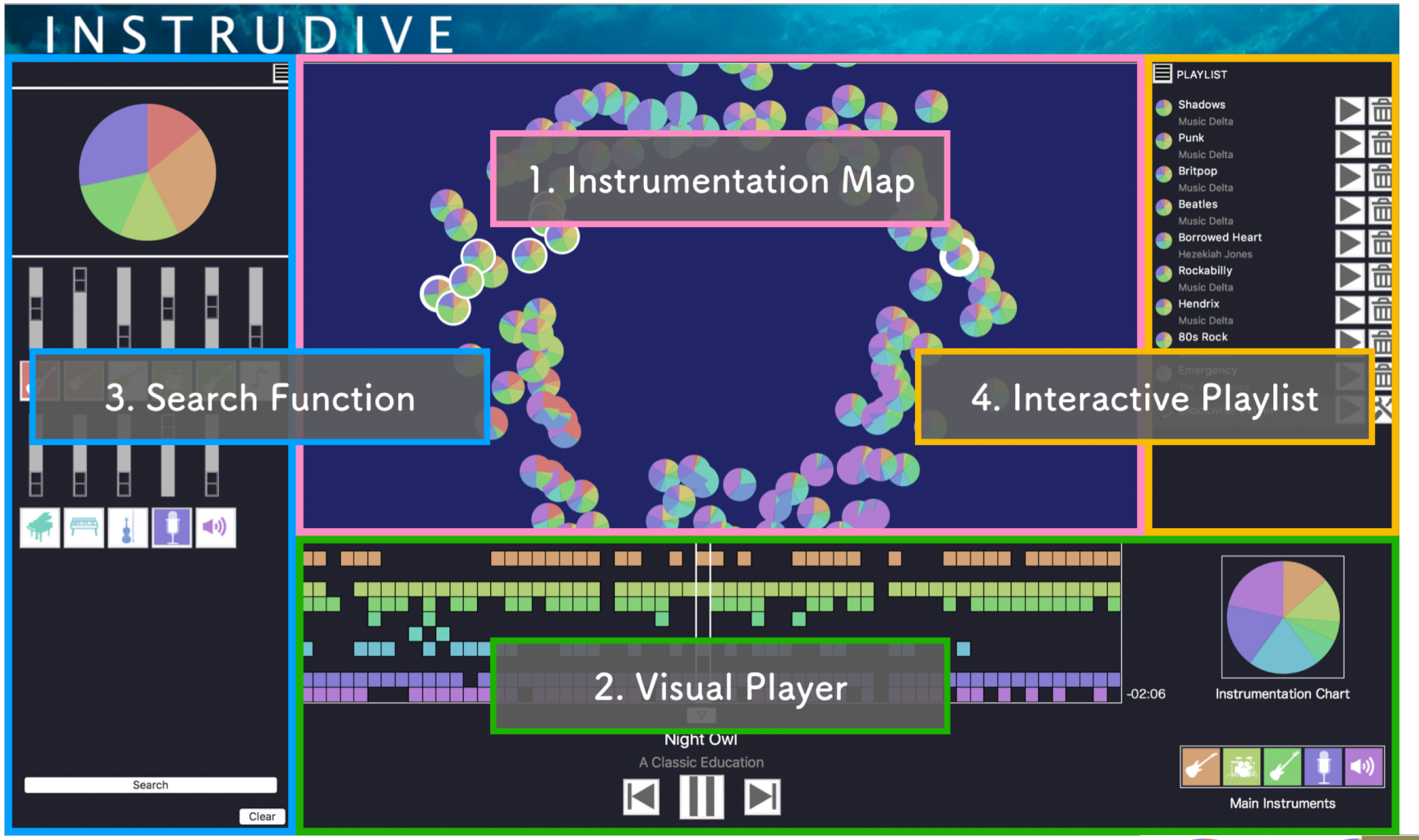
Instrument Recognition (3)

- (E-4) Instrudive: A Music Visualization System Based on Automatically Recognized Instrumentation
  - AIST Japan - Masataka Goto
    - use CNNs with 1 second input (1024 freq bands) for better instrument detection

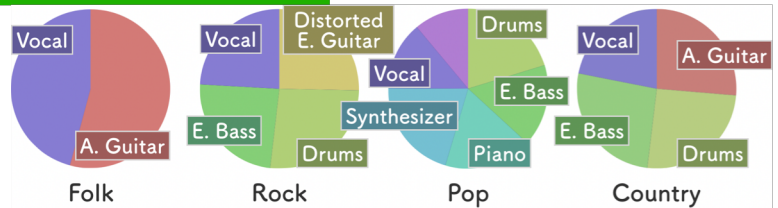




Instrument Recognition (3)

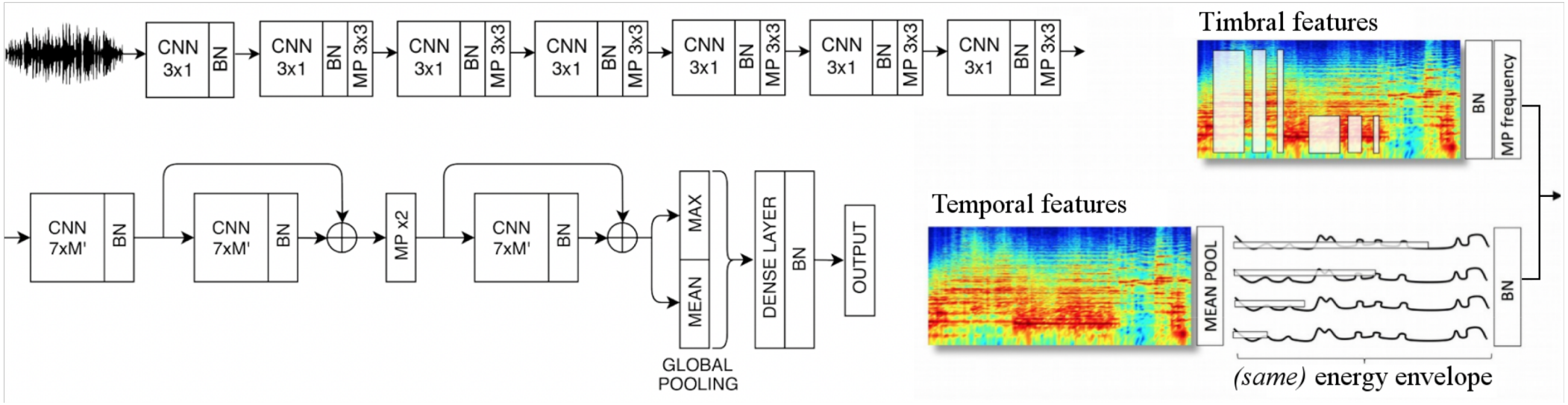


Player for Instrumentation based Playlists



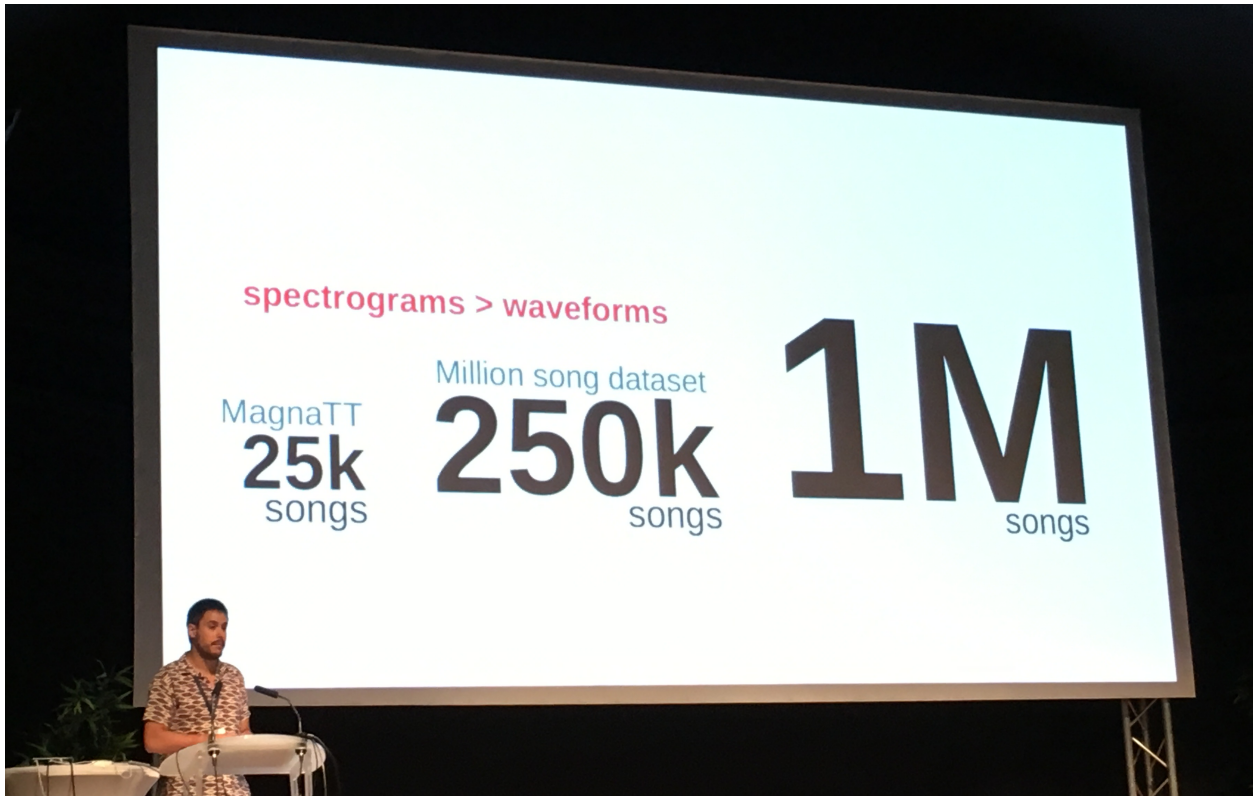
Music Tagging “at scale” (UPF-MTG + Pandora)

- (E-14) End-to-end Learning for Music Audio Tagging at Scale
  Jordi Pons, Oriol Nieto, Matthew Prockup, Erik M. Schmidt, Andreas F. Ehmann and Xavier Serra
- compare the parallel CNN approach based on spectrograms processed by rhythm and timbre separately (created by Jordi and Tom in 2016) with a new end to end learning CNN that **learns directly from waveforms**



Music Tagging “at scale” (UPF-MTG + Pandora)

- Conclusion: **waveform input performs better only** when you have **at last 1 million songs to train (private Pandora dataset)**
  (the increase in ROC AUC was very small though: less than 1%)
- interesting is the improved parallel CNN approach (Pons 2017) based on Jordi’s+Tom’s original work: he uses now **6 (!) parallel networks**



Other Topics

- Singing Voice Detection (vs. Instrumental)
- Cover Song Identification
- Melodic Similarity
- Audio Source Separation
- Audio-to-Score Alignment / Score Following (Best paper winner)
- …

(… a lot of approaches use Deep Learning…)

Resources

- All ISMIR 2018 paper PDFs available here : http://ismir2018.ircam.fr/pages/events-main-program.html