INTERNATIONAL AUDIO LABORATORIES ERLANGEN

AUDIO LABS

Tutorial
Automatisierte Methoden der Musikverarbeitung
47. Jahrestagung der Gesellschaft für Informatik

Deep Neural Networks in MIR

Meinard Müller, Christof Weiss, Stefan Balke

International Audio Laboratories Erlangen

{meinard.mueller, christof.weiss, stefan.balke}@audiolabs-erlangen.de

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

Fraunhofer IIS

Motivation

- DNNs are very powerful methods
- Define the state of the art in different domains
- Lots of decisions involved when designing a DNN
  - Input representation, input preprocessing
  - #layers, #neurons, layer type, dropout, regularizers, cost function
  - Initialization, mini-batch size, #epochs, early stopping (patience)
  - Optimizer, learning rate…

Neural Networks

Black Box

Input

$X \in \mathbb{R}^N$

Animal Images

Speech

Music

Output

$Y \in \mathbb{R}^M$

{Cats, Dogs}

Text

Genre, era, chords…

$$f : \mathbb{R}^N \rightarrow \mathbb{R}^M$$

Neural Networks
Black Box

Input

$X \in \mathbb{R}^N$

Animal Images

Speech

Music

Output

$Y \in \mathbb{R}^M$

{Cats, Dogs}

Text

Genre, era, chords…

$$f : \mathbb{R}^N \rightarrow \mathbb{R}^M$$

Neural Networks
Black Box

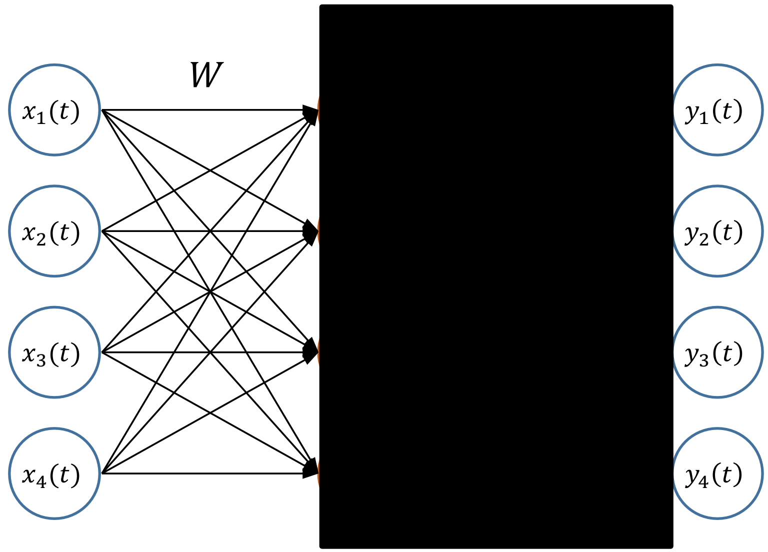
Input

$X \in \mathbb{R}^N$

Animal Images

Speech

Music



Output

$Y \in \mathbb{R}^M$

{Cats, Dogs}

Text

Genre, era, chords…

$$f : \mathbb{R}^N \rightarrow \mathbb{R}^M$$

Neural Networks

Black Box

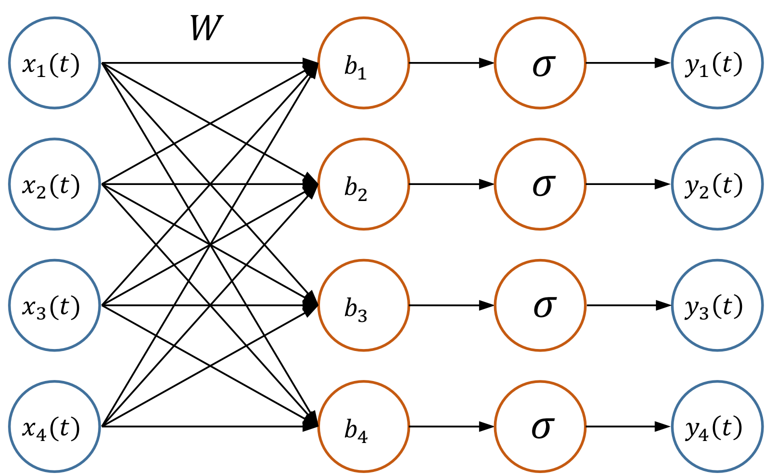
Input

$X \in \mathbb{R}^N$

Animal Images

Speech

Music



Output

$Y \in \mathbb{R}^M$

{Cats, Dogs}

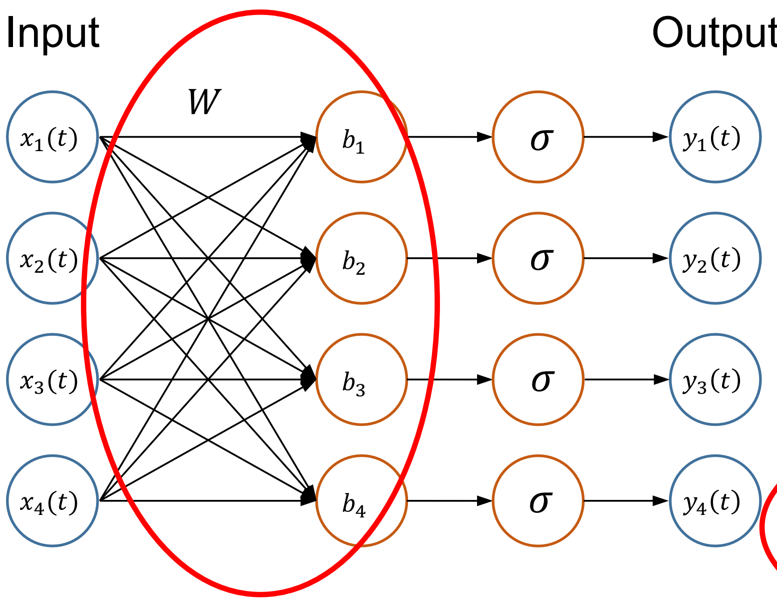
Text

Genre, era, chords…

$$f : \mathbb{R}^N \rightarrow \mathbb{R}^M$$

Neural Network
Intuition

- NN is a non-linear mapping from input- to output-space
- Free parameters are trained with examples (supervised)



Definition: $f : \mathbb{R}^N \to \mathbb{R}^M$

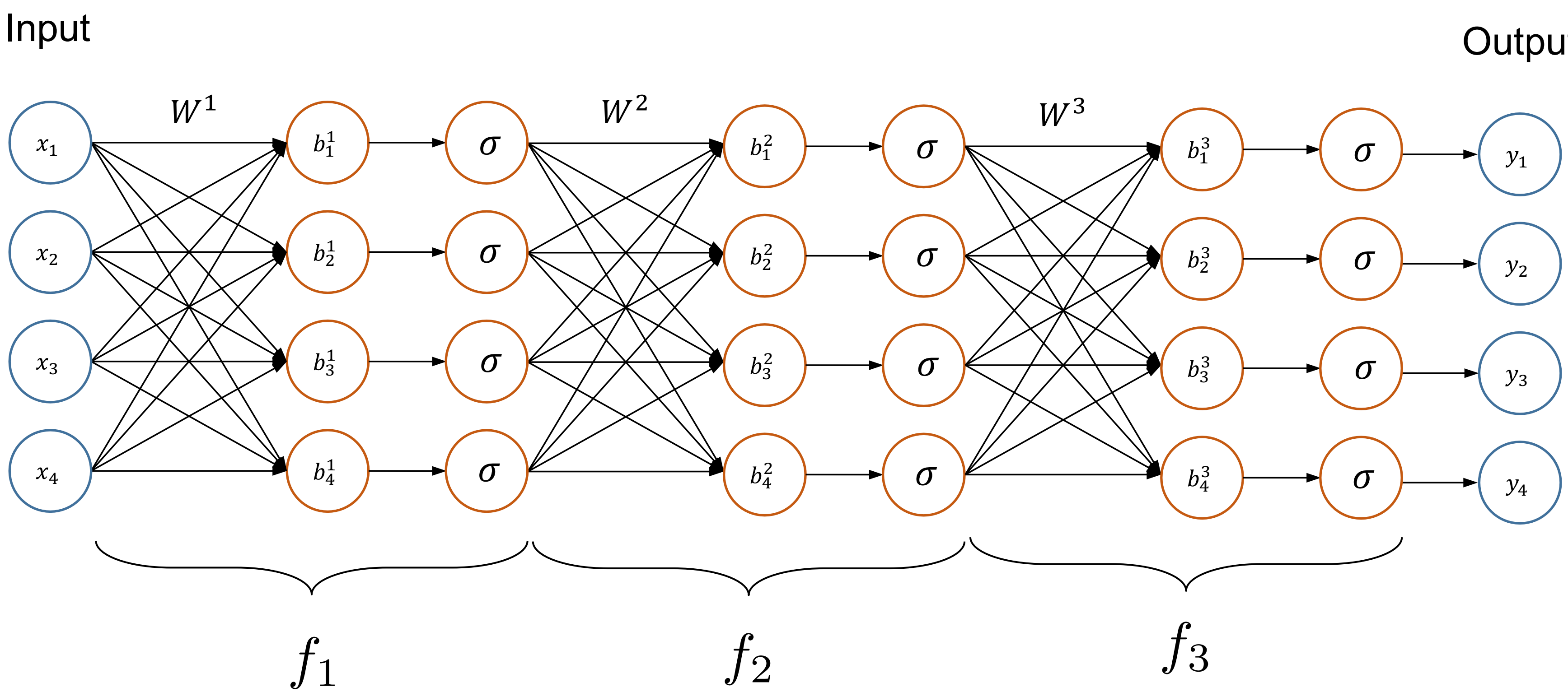
Mapping: $f(x) = \sigma(W^T x + b),$

Nonlinearity: $\sigma : \mathbb{R} \to \mathbb{R}$

Weights: $W \in \mathbb{R}^{N \times M}$

Bias: $b \in \mathbb{R}^M$

Deep Neural Network
Going Deep

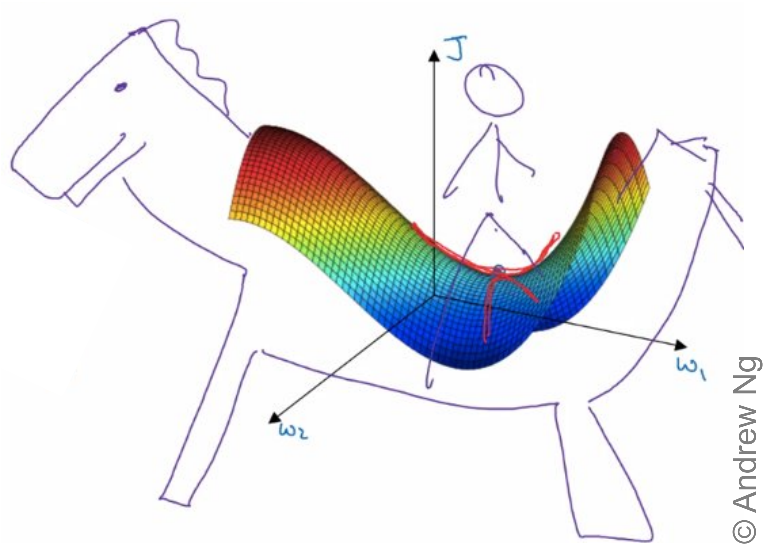


$$f(x) = (f_3 \circ f_2 \circ f_1)(x)$$

Deep Neural Networks
Training

- Collect labeled dataset (e.g., images with cats and dogs)
- Define a quality measure: Loss function
- Task: Find minimum of loss function (not trivial)

➡ Gradient Descent



Deep Neural Networks
Gradient Descent

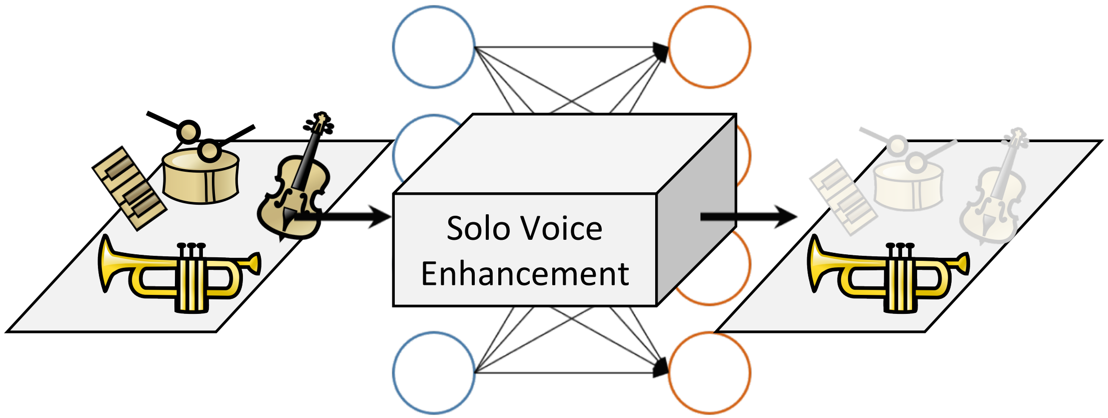
- **Idea:** Find the minimum of a function in an iterative way by following the direction of steepest descent of the gradient

- Initialize all free parameters randomly
- Repeat until convergence:
  - Let the DNN perform predictions on the dataset
  - Measure the quality of the predictions w. r. t. the loss function
  - Update the free parameters based on the prediction quality

- Common extension: Stochastic Gradient Descent

Overview

1. Feature Learning
2. Beat and Rhythm Analysis
3. Music Structure Analysis
4. Literature Overview



Feature Learning

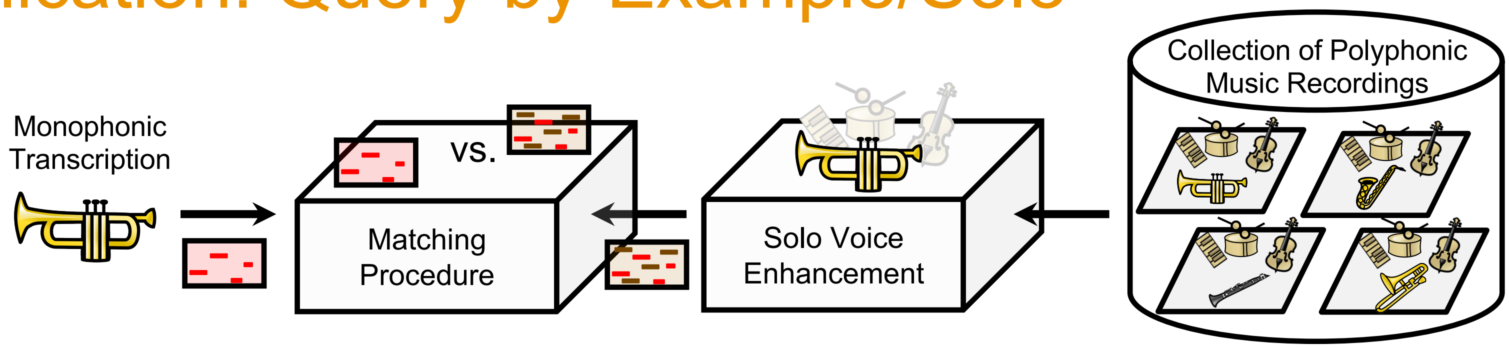
Feature Learning

…where it all began

- Core task for DNNs:
  Learn a representation from the data to solve a problem.
- Task is very hard to define!
  Often evaluated in tagging, chord recognition, or retrieval application.

Task	Year	Authors	Ref.	Type	Input	Pre-proc.
FL	2013	Schmidt and Kim	[67]	DBN	HC	—
FL	2010	Hamel and Eck	[30]	DBN	LinS	—
FL	2017	Dai et al.	[15]	CNN	Raw	—
FL	2012	Hamel et al.	[33]	FNN	LogMelS	PCA
FL	2016	Korzeniowski and Widmer	[43]	FNN	LogLogS	—
FL	2017	Balke et al.	[2]	FNN	LogS	—
FL	2011	Hamel et al.	[32]	FNN	MelS	PCA
FL	2014	Dieleman and Schrauwen	[17]	CNN	Raw	—

Application: Query-by-Example/Solo



Retrieval Scenario

Given a monophonic transcription of a jazz solo as query, find the corresponding document in a collection of polyphonic music recordings.

Solo Voice Enhancement

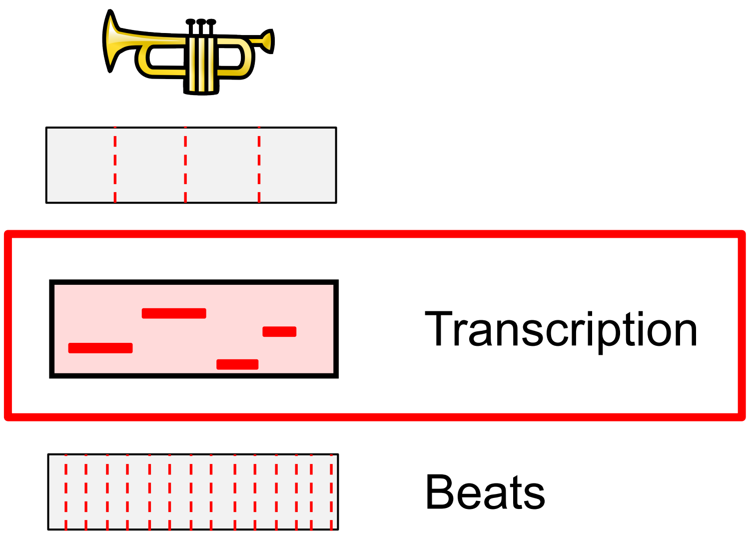
1. Model-based Approach [Salamon13]
2. Data-Driven Approach [Rigaud16, Bittner15]

Our Data-Driven Approach

Use a **DNN** to learn the mapping from a “polyphonic” TF representation to a “monophonic” TF representation.

Weimar Jazz Database (WJD)

[Pfleiderer17]



| E^7 A^7 | D^7 G^7 | … Chords

…

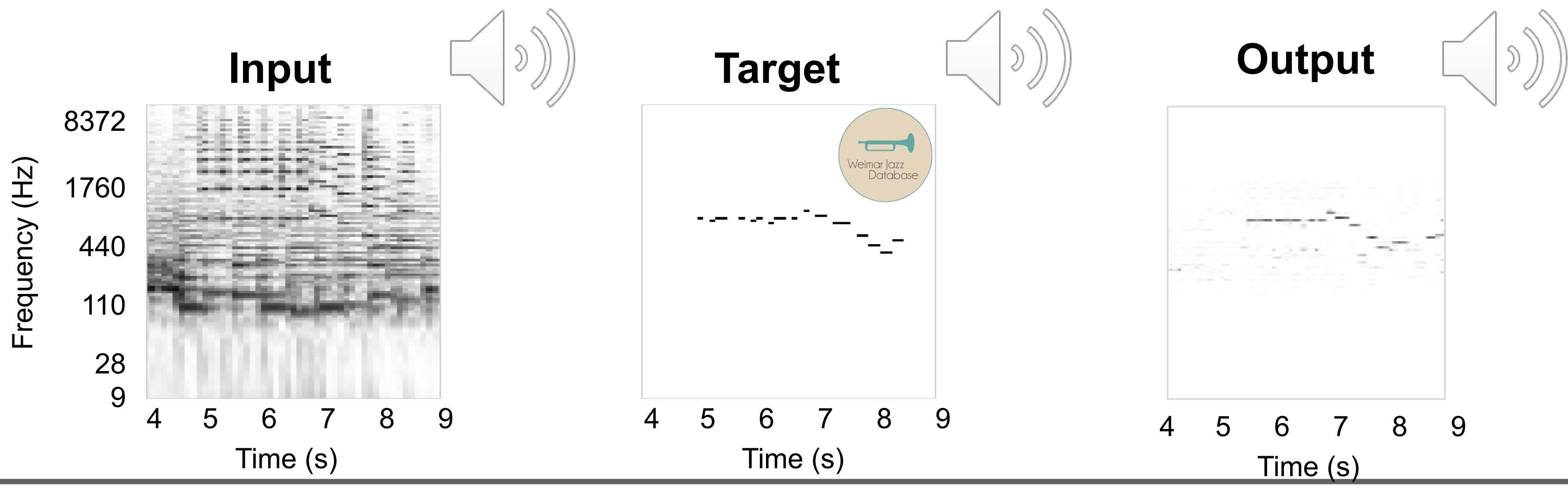
- 456 transcribed jazz solos of monophonic instruments.
- Transcriptions specify a musical pitch for physical time instances.
- 810 min. of audio recordings.

Thanks to the Jazzomat research team: M. Pfleiderer, K. Frieler, J. Abeßer, W.-G. Zaddach

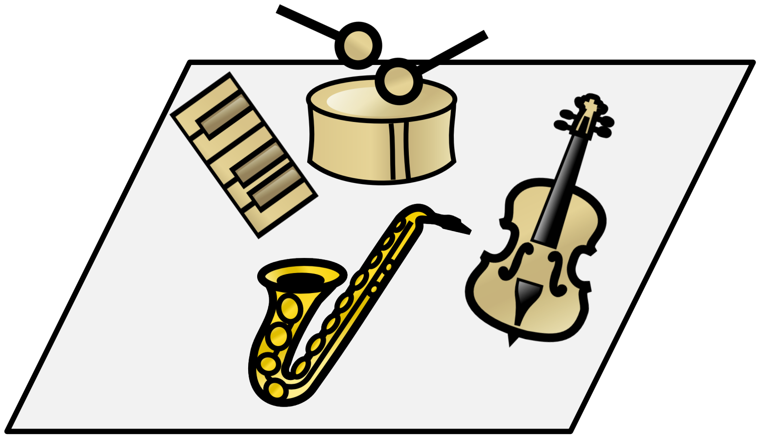
DNN Training

Stefan Balke, Christian Dittmar, Jakob Abeßer, Meinard Müller, ICASSP 17

- **Input:** Log-freq. Spectrogram (120 semitones, 10 Hz feature rate)
- **Target:** Solo instrument's pitch activations
- **Output:** Pitch activations (120 semitones, 10 Hz feature rate)
- **Architecture:** FNN, 5 hidden layers, ReLU, Loss: MSE, layer-wise training
- **Demo:** https://www.audiolabs-erlangen.de/resources/MIR/2017-ICASSP-SoloVoiceEnhancement



Walking Bass Line Extraction



- Harmonic analysis
  - Composition (lead sheet) vs. actual performance
  - Polyphonic transcription from ensemble recordings is challenging
  - Walking bass line can provide first clues about local harmonic changes
- Features for style & performer classification

What is a Walking Bass Line?

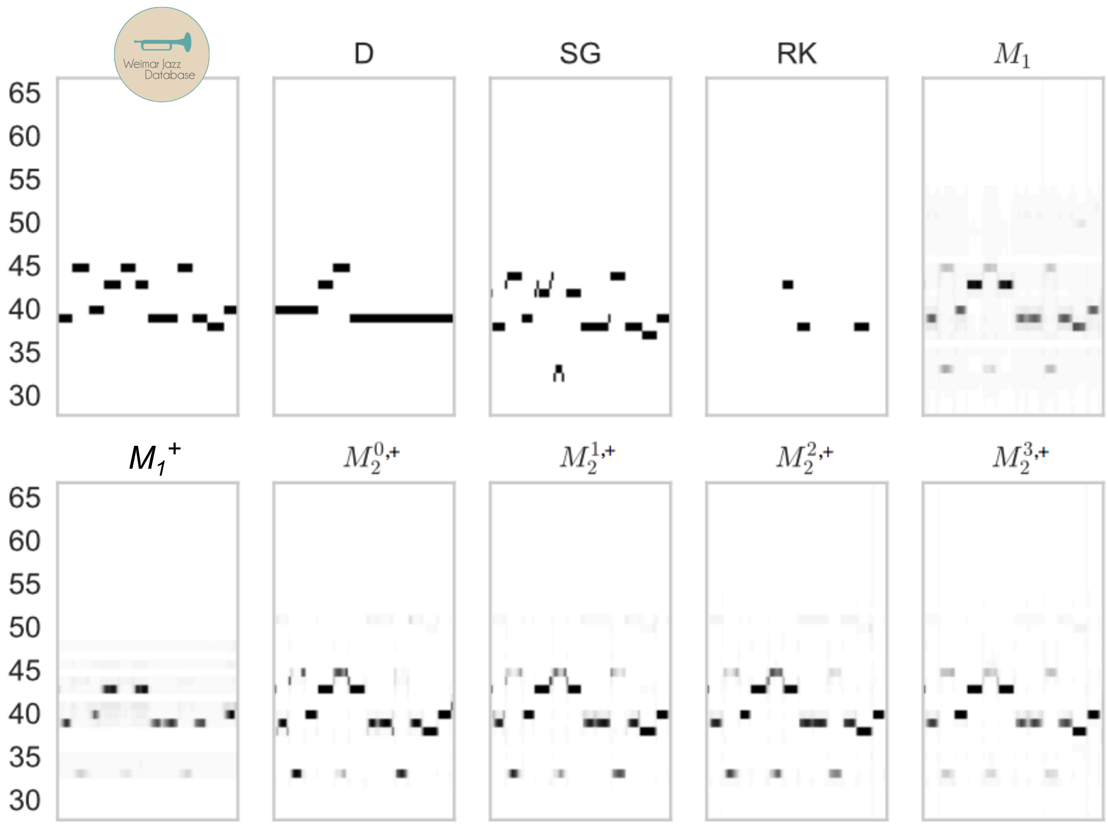
- **Example:** Miles Davis: So What (Paul Chambers: b)

- Our assumptions for this work:
- Quarter notes (mostly chord tones)
- Representation: beat-wise pitch values

Example

D - Dittmar et al.
SG - Salamon et al.
RK - Ryynänen & Klapuri

- Chet Baker: "Let's Get Lost" (0:04 – 0:09)
- **Demo:** https://www.audiolabs-erlangen.de/resources/MIR/2017-AES-WalkingBassTranscription



Initial model

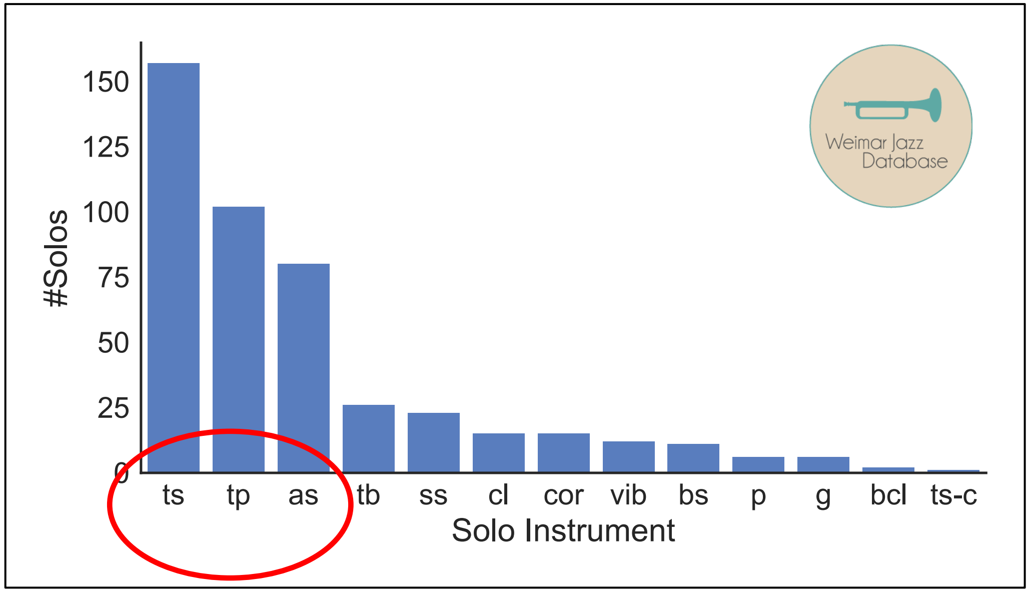
M_1 - without data aug.
M_1^+ - with data aug.

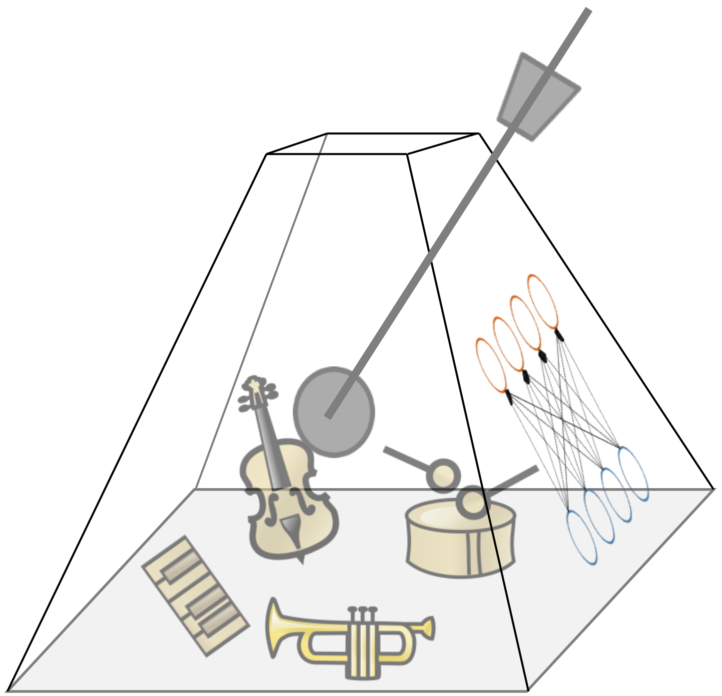
Semi-supervised learning

$M_2^{0,+}$ - τ^0
$M_2^{1,+}$ - τ^1
$M_2^{2,+}$ - τ^2
$M_2^{3,+}$ - τ^3

Feature Learning

- Less domain knowledge needed to learn working features.
- Know your task/data.
  Accuracy is not everything!





Beat and Rhythm Analysis

Beat and Rhythm Analysis

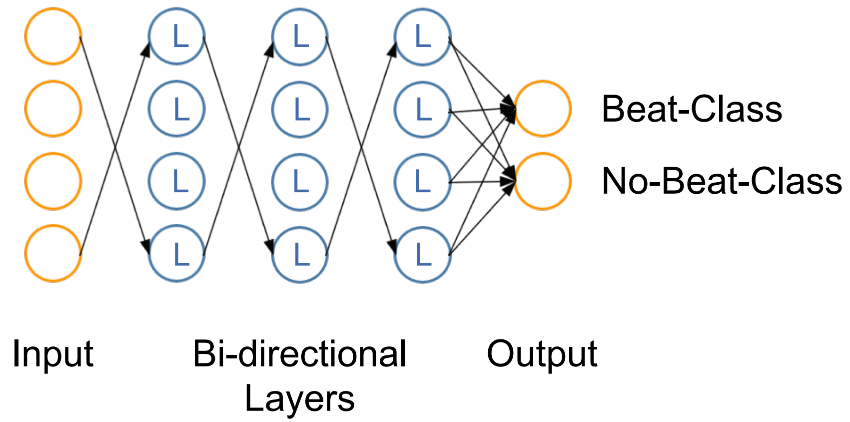
Task	Year	Authors	Ref.	Type	Input	Pre-proc.
BRA	2010	Eyben et al.	[25]	RNN-BLSTM	LogMelS	DERIV
BRA	2011	Böck and Schedl	[5]	RNN-BLSTM	LogMelS	DERIV
BRA	2012	Battenberg and Wessel	[3]	DBN	—	—
BRA	2014	Böck et al.	[7]	RNN-BLSTM	LogS	—
BRA	2016	Böck et al.	[9]	RNN-BLSTM	LogS	DERIV
BRA	2016	Elowsson	[23]	FNN	HC	—
BRA	2016	Holzapfel and Grill	[35]	CNN	LogLogS	STDF
BRA	2016	Krebs et al.	[46]	RNN-BGRU	HC	—
BRA	2016	Durand and Essid	[21]	CNN	HC	—
BRA	2017	Durand et al.	[22]	CNN	HC	—
BRA	2015	Böck et al.	[8]	RNN-BLSTM	LogMelS	DERIV

- **Beat Tracking:**
  Find the pulse in the music which you would tap/clap to.

Beat and Rhythm Analysis

Sebastian Böck, Florian Krebs, and Gerhard Widmer, DAFx 2011

- **Input:** 3 LogMel spectrograms (varying win-length) + derivatives
- **Target:** Beat annotations
- **Output:** Beat activation function $\in [0, 1]$
- **Post-processing:** Peak picking on beat activation function
- **Architecture:** RNN, 3 bidirectional layers, 25 LSTM per layer/direction



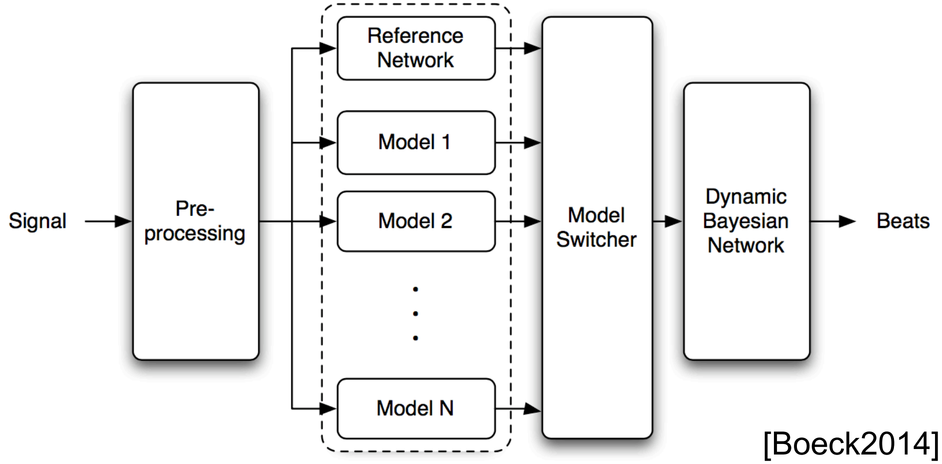
Beat Tracking
Examples

	Borodin String Quartet 2, III. 65 bpm	Carlos Gardel Por una Cabeza 114 bpm	Sidney Bechet Summertime 87 bpm	Wynton Marsalis Caravan 195 bpm	Wynton Marsalis Cherokee 327 bpm
Original					
Ellis (librosa) Init = 120 bpm					
Böck2015 (madmom)					

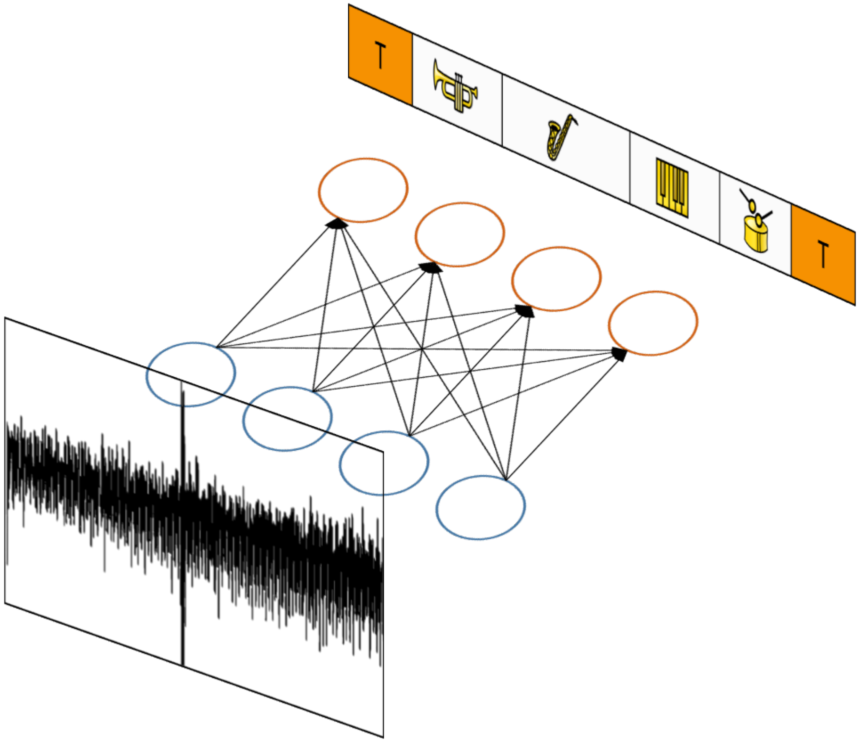
Beat Tracking

- DNN-based methods need less task-specific initialization (e.g., tempo).
- Closer to a “universal” onset detector.

- Task-specific knowledge is introduced as post-processing step:



[Boeck2014]

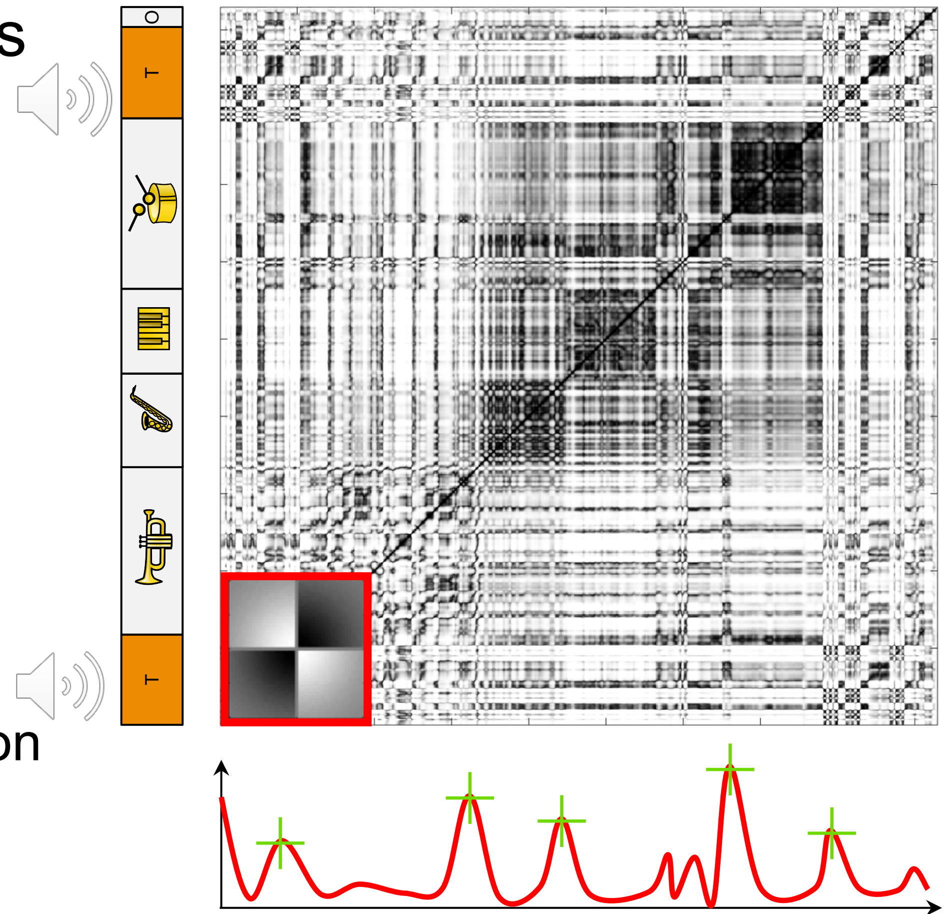


Music Structure Analysis

Music Structure Analysis

Task	Year	Authors	Ref.	Type	Input	Pre-proc.
MSA	2017	Cohen-Hadria and Peeters	[14]	CNN	LogMelS, SSM	—
MSA	2014	Ullrich et al.	[75]	CNN	LogMelS	—
MSA	2015	Grill and Schlüter	[28]	CNN	LogMelS	—
MSA	2015	Grill and Schlüter	[29]	CNN	LogMelS	HPSS

- Find boundaries/repetitions in music
- Classic approaches:
  - Repetition-based
  - Homogeneity-based
  - Novelty-based
- Main challenges:
  - What is structure?
  - Model assumptions based on musical rules (e.g., sonata).

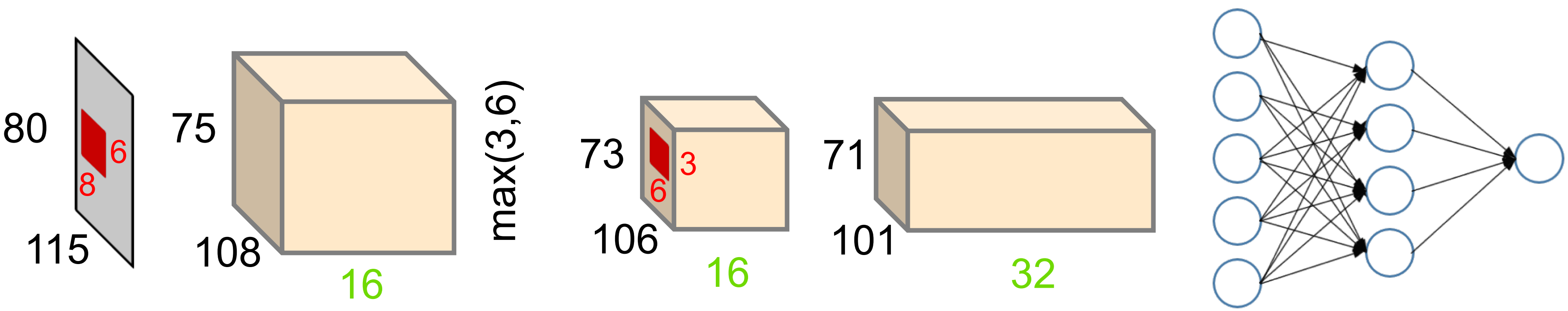


[Foote]

Music Structure Analysis

Karen Ullrich, Jan Schlüter, and Thomas Grill, ISMIR 2014

- **Input:** LogMel spectrogram
- **Target:** Boundary annotations
- **Output:** Novelty function ∈ [0, 1]
- **Post-processing:** Peak picking on novelty function



71 * 101 * 32 = 229’472 128 1

8 * 6 * 16
= 768

6 * 3 * 16 * 32
= 9216

229’472 * 128
= 29’372’416

128 * 1
= 128

* ignoring bias

Music Structure Analysis

Results

Tolerance	SALAMI 1.3 Ullrich et al. (2014)

0.5 s:

Algorithm	F-measure	Precision	Recall
Upper bound (est.)	0.68		
16s_std_1.5s	**0.4646**	0.5553	0.4583
MP2 (2013)	0.3280	0.3001	0.4108
MP1 (2013)	0.3149	0.3043	0.3605
OYZS1 (2012)	0.2899	0.4561	0.2583

3.0 s:

Algorithm	F-measure	Precision	Recall
Upper bound (est.)	0.76		
32s_low_6s	**0.6164**	0.5944	0.7059
16s_std_1.5s	0.5726	0.5648	0.6675
MP2 (2013)	0.5213	0.4793	0.6443
MP1 (2013)	0.5188	0.5040	0.5849

SALAMI 2.0

Grill et al. (2015)

Algorithm	**F_1**	**$F_{.58}$**	**Rec.**	**Prec.**
Upper bound (est.)	.74	.74		
All features, multi+fine ann.	**.508**	.529	.502	.572
MLS+SSLM-near, multi+fine	.496	.506	.509	.536
MLS+SSLM-near, single ann.	.469	.466	.504	.475
SUG1 (2014)	.422	.442	.422	.490
MP2 (2013)	.294	.280	.362	.271
MP1 (2013)	.276	.270	.311	.269
NB1 (2014)	.270	.246	.374	.229
KSP2 (2012)	.263	.231	.422	.209
Baseline (est.)	.15	.21		

- Added features (SSLM)
- Trained on 2 levels of annotations
- SUG1 is similar to [Ullrich2014]

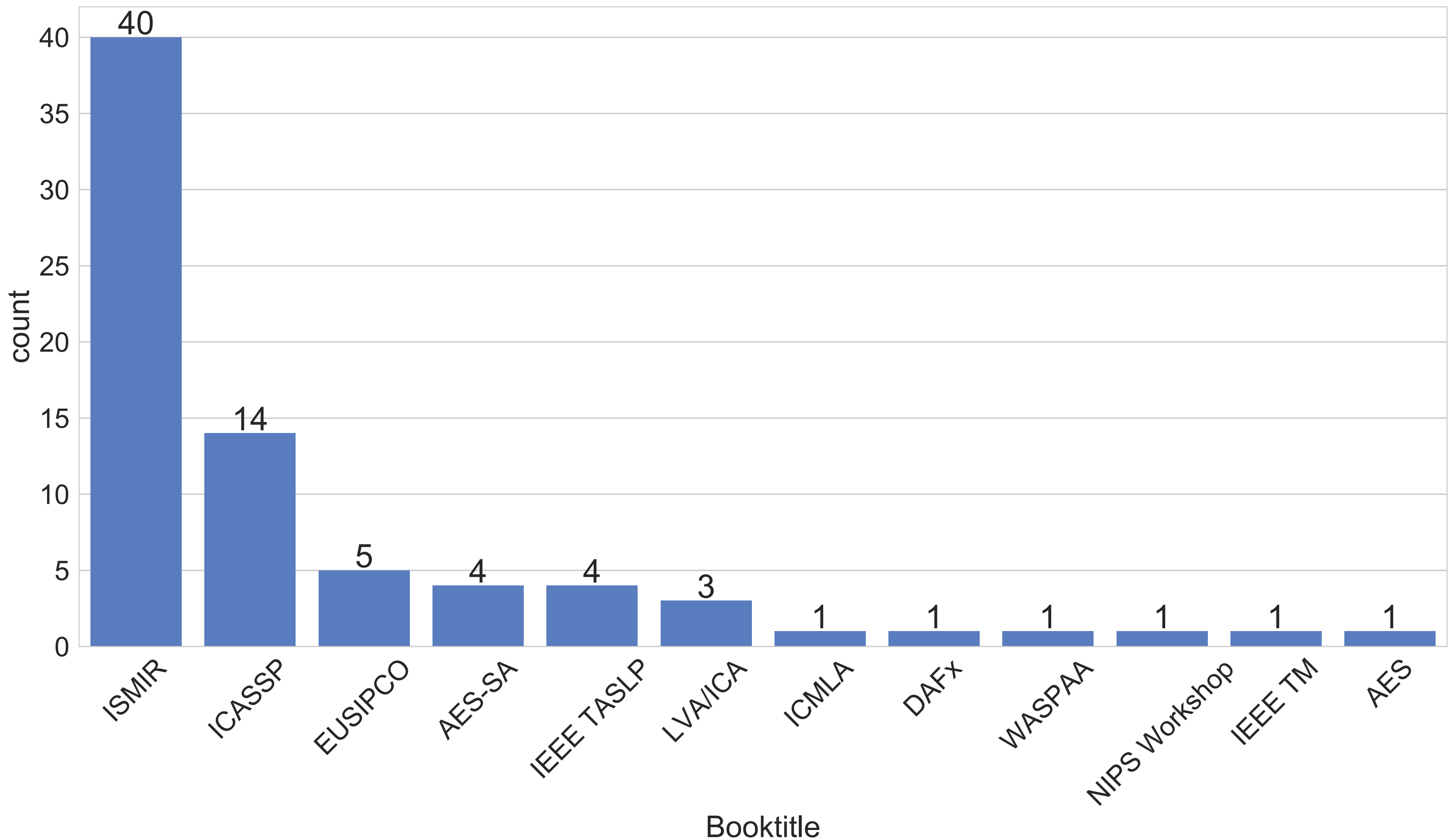
Music Structure Analysis

Task	Year	Authors	Ref.	Type	Input	Pre-proc.
MSA	2017	Cohen-Hadria and Peeters	[14]	CNN	LogMelS, SSM	—
MSA	2014	Ullrich et al.	[75]	CNN	LogMelS	—
MSA	2015	Grill and Schlüter	[28]	CNN	LogMelS	—
MSA	2015	Grill and Schlüter	[29]	CNN	LogMelS	HPSS

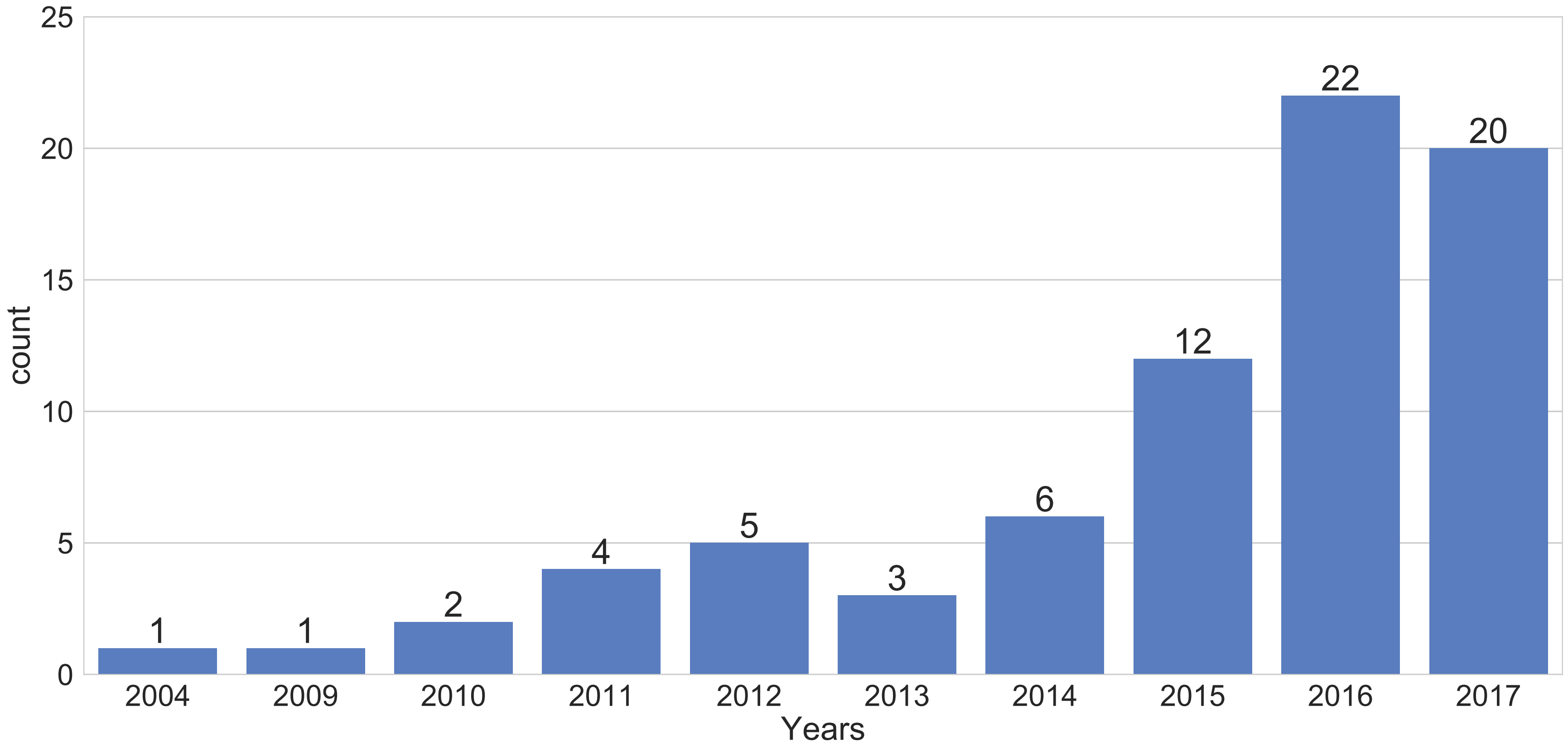
- Re-implementation by *Cohen-Hadria and Peeters* did not reach reported results.
- Possible reasons:
  - Data identical?
  - Different kind of convolution? What was the stride?
  - Didn’t ask?
  - Availability of pre-trained model would be awesome!

Literature Overview

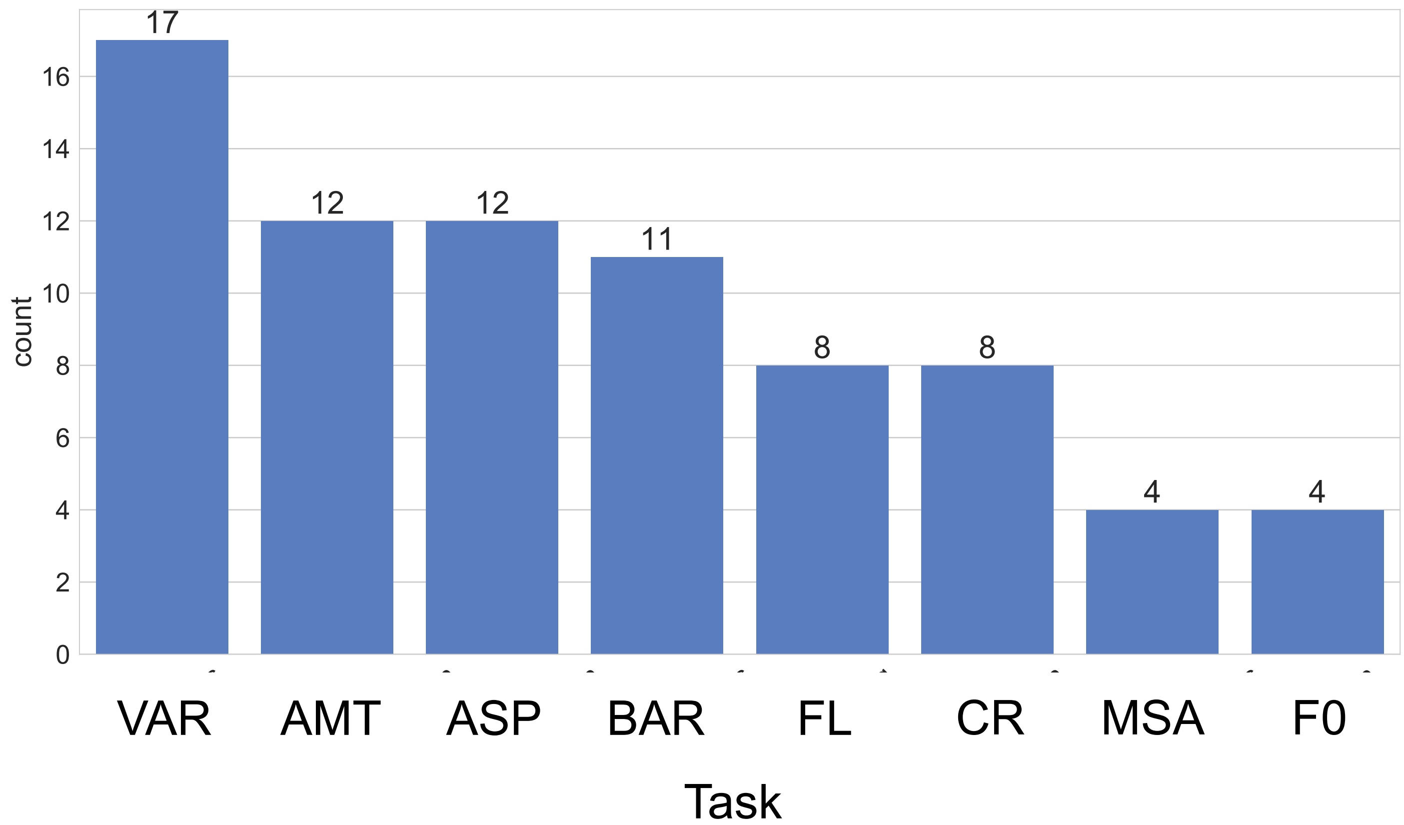
Publications by Conference



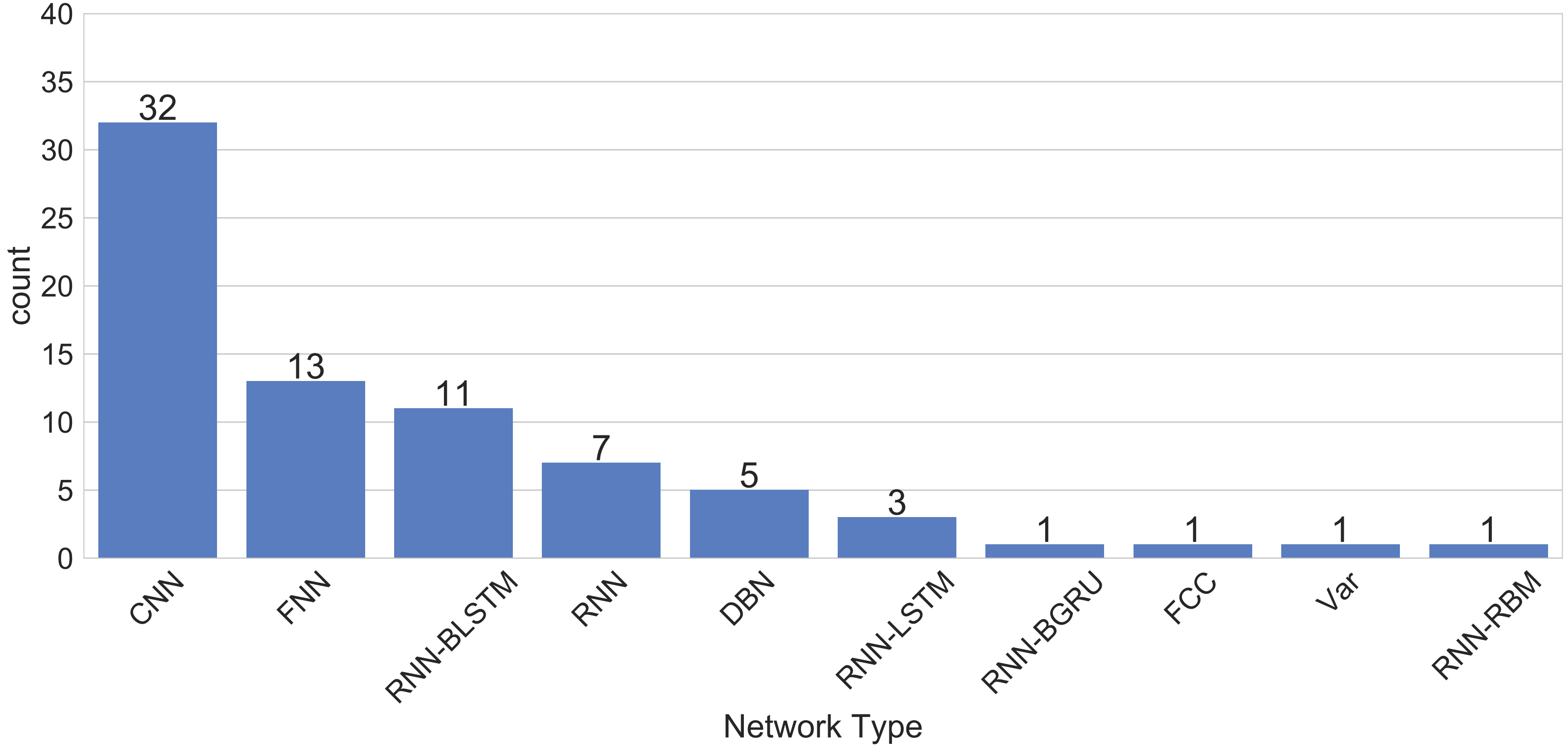
Publications by Year



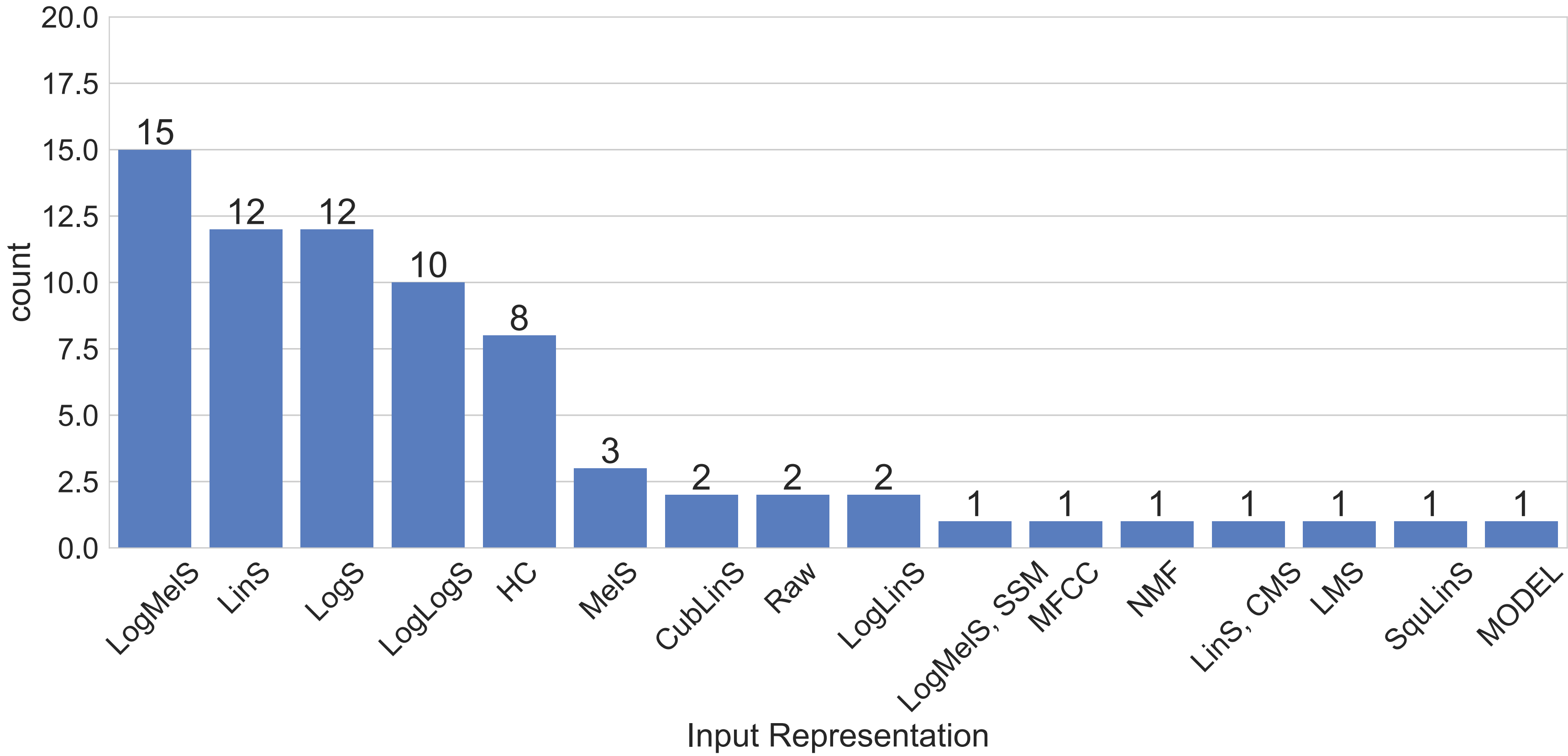
Publications by Task



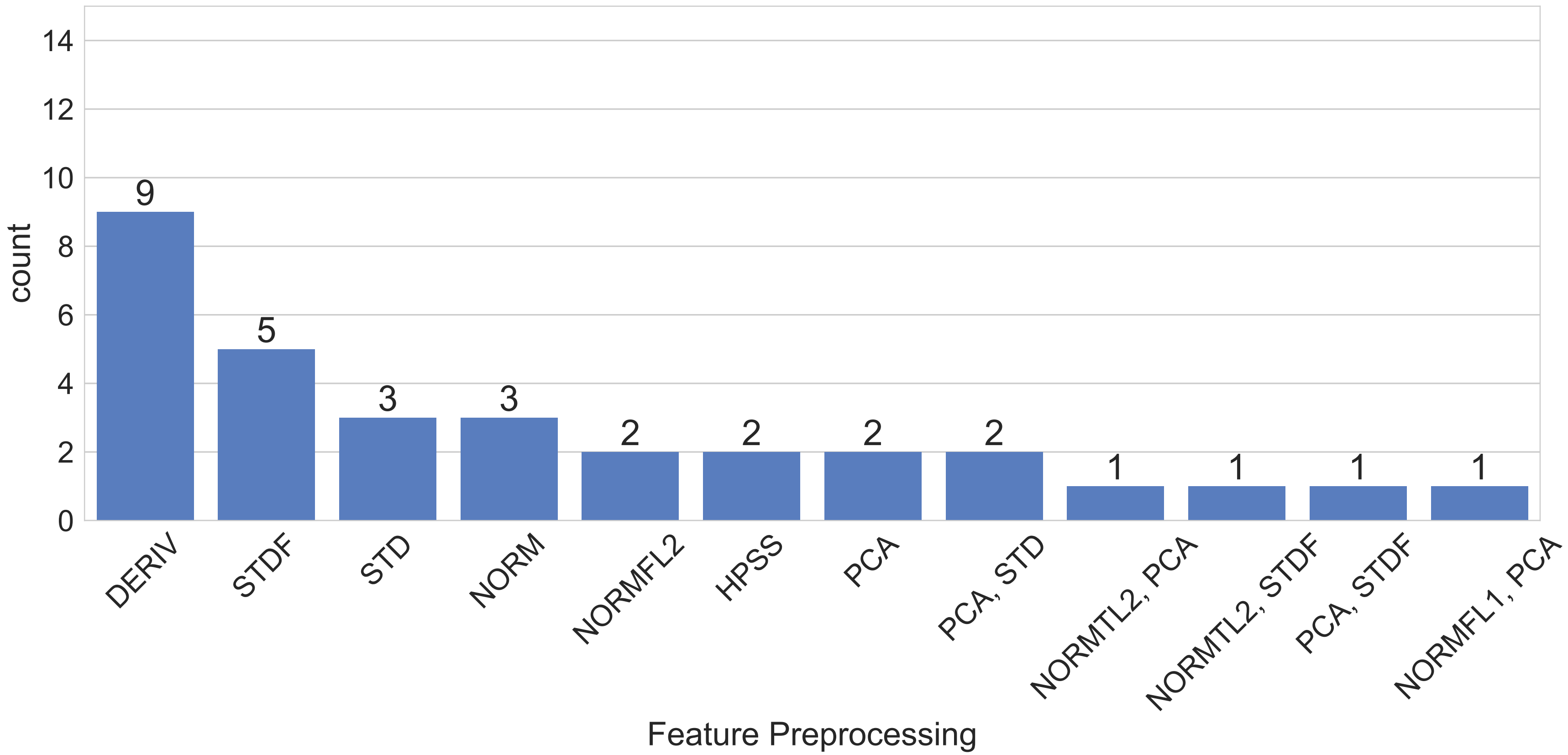
Publications by Network



Input Representations



Feature Preprocessing



Technical Background
Overview

- DNN problems are tensor problems
- Lots of different open source frameworks available
  - Theano (University of Montreal)
  - tensorflow (Google)
  - PyTorch (Facebook)
- Support training DNNs on GPUs (NVIDIA GPUs are currently leading)
- Python is mainly used in this research area

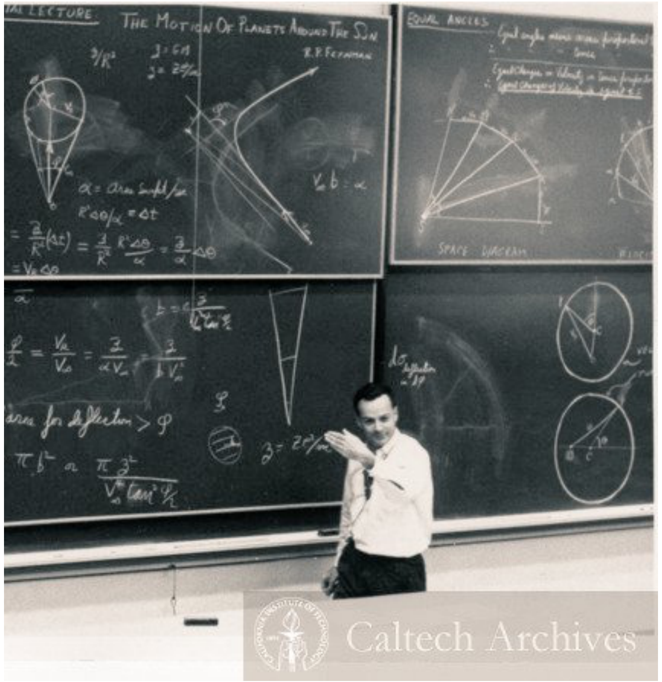
Technical Background
Python Starter-Kit

- **NumPy** Basics for matrices and tensors
- **Pandas** General operations on any data
- **Matplotlib** plotting your data

- **Librosa** General Audio library (STFT, Chroma, etc.)
- **Scikit-learn** For all kinds of machine learning models
- **Keras** High-Level wrapper for neural networks
- **Pescador** Data streaming
- **mir_eval** Common evaluation metrics used in MIR

Deep Neural Networks in MIR

- Online Lectures:
  - Andrew Ng: Machine Learning
    (Coursera class, more a general introduction to machine learning)
  - Google: Deep Learning
    (Udacity class, hands on with tensorflow)
  - CS231n: Convolutional Neural Networks for Visual Recognition
    (Stanford class, available via YouTube)
- Goodfellow, Bengio, Courville: Deep Learning Book.
- Other MIR resources:
  - Jordi Pons: http://jordipons.me/wiki/index.php/MIRDL
  - Keunwoo Choi: https://arxiv.org/abs/1709.04396
  - Yann Bayle: https://github.com/ybayle/awesome-deep-learning-music
  - Jan Schlüter: http://www.univie.ac.at/nuhag-php/program/talks_details.php?nl=Y&id=3358



"…if you're doing an experiment, you should **report everything that you think might make it invalid—not only what you think is right about it**: other causes that could possibly explain your results; and things you thought of that you've eliminated by some other experiment, and how they worked—to make sure the other fellow can tell they have been eliminated."

Richard Feynman, Surely You're Joking, Mr. Feynman!: Adventures of a Curious Character

Bibliography

[1] Jakob Abeßer, Klaus Frieler, Wolf-Georg Zaddach, and Martin Pfleiderer. Introducing the Jazzomat project - jazz solo analysis using Music Information Retrieval methods. In Proceedings of the International Symposium on Sound, Music, and Motion (CMMR), pages 653–661, Marseille, France, 2013.

[2] Jakob Abeßer, Stefan Balke, Klaus Frieler, Martin Pfleiderer, and Meinard Müller. Deep learning for jazz walking bass transcription. In Proceedings of the AES International Conference on Semantic Audio, pages 210–217, Erlangen, Germany, 2017.

[3] Stefan Balke, Christian Dittmar, Jakob Abeßer, and Meinard Müller. Data-driven solo voice enhancement for jazz music retrieval. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 196–200, New Orleans, USA, 2017.

[4] Eric Battenberg and David Wessel. Analyzing drum patterns using conditional deep belief networks. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 37–42, Porto, Portugal, 2012.

[5] Rachel M. Bittner, Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam, and Juan Pablo Bello. MedleyDB: A multitrack dataset for annotation-intensive MIR research. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 155–160, Taipei, Taiwan, 2014.

[6] Rachel M. Bittner, Brian McFee, Justin Salamon, Peter Li, and Juan P. Bello. Deep salience representations for F0 tracking in polyphonic music. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), Suzhou, China, 2017.

[7] Sebastian Böck and Markus Schedl. Enhanced beat tracking with context-aware neural networks. In Proceedings of the International Conference on Digital Audio Effects (DAFx), pages 135–139, Paris, France, 2011.

[8] Sebastian Böck and Markus Schedl. Polyphonic piano note transcription with recurrent neural networks. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 121–124, Kyoto, Japan, 2012.

[9] Sebastian Böck , Florian Krebs, and Gerhard Widmer. A multi-model approach to beat tracking considering heterogeneous music styles. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 603–608, Taipei, Taiwan, 2014.

[10] Sebastian Böck , Florian Krebs, and Gerhard Widmer. Accurate tempo estimation based on recurrent neural networks and resonating comb filters. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 625–631, Málaga, Spain, 2015.

[11] Sebastian Böck, Florian Krebs, and Gerhard Widmer. Joint beat and downbeat tracking with recurrent neural networks. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 255–261, New York City, United States, 2016.

[12] Nicolas Boulanger-Lewandowski, Yoshua Bengio, and Pascal Vincent. Audio chord recognition with recurrent neural networks. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 335–340, Curitiba, Brazil, 2013.

[13] Nicolas Boulanger-Lewandowski, Yoshua Bengio, and Pascal Vincent. High-dimensional sequence transduction. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 3178–3182, Vancouver, Canada, 2013.

[14] Pritish Chandna, Marius Miron, Jordi Janer, and Emilia Gómez. Monoaural audio source separation using deep convolutional neural networks. In Proceedings of the International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA), pages 258–266, Grenoble, France, 2017.

[15] Keunwoo Choi, Gyo ̈rgy Fazekas, and Mark B. Sandler. Automatic tagging using deep convolutional neural networks. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 805–811, New York City, United States, 2016.

[16] Alice Cohen-Hadria and Geoffroy Peeters. Music structure boundaries estimation using multiple self-similarity matrices as input depth of convolutional neural networks. In Proceedings of the AES International Conference on Semantic Audio, pages 202–209, Erlangen, Germany, 2017.

[17] Wei Dai, Chia Dai, Shuhui Qu, Juncheng Li, and Samarjit Das. Very deep convolutional neural networks for raw waveforms. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 421–425, New Orleans, USA, 2017.

[18] Jun-qi Deng and Yu-Kwong Kwok. A hybrid gaussian-hmm-deep learning approach for automatic chord estimation with very large vocabulary. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 812–818, New York City, United States, 2016.

[19] Sander Dieleman and Benjamin Schrauwen. End-to-end learning for music audio. In IEEE Interna- tional Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6964–6968, Florence, Italy, 2014.

[20] Sander Dieleman, Philemon Brakel, and Benjamin Schrauwen. Audio-based music classification with a pretrained convolutional network. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 669–674, Miami, Florida, 2011.

[21] Matthias Dorfer, Andreas Arzt, and Gerhard Widmer. Towards score following in sheet music images. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 789–795, New York, USA, 2016.

[22] Matthias Dorfer, Andreas Arzt, and Gerhard Widmer. Learning audio-sheet music correspondences for score identification and offline alignment. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), Suzhou, China, 2017.

[23] Simon Durand and Slim Essid. Downbeat detection with conditional random fields and deep learned features. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 386–392, New York City, United States, 2016.

[24] Simon Durand, Juan P. Bello, Bertrand David, and Gaël Richard. Robust downbeat tracking using an ensemble of convolutional networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(1):76–89, 2017.

[25] Anders Elowsson. Beat tracking with a cepstroid invariant neural network. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 351–357, New York City, United States, 2016.

[26] Valentin Emiya, Roland Badeau, and Bertrand David. Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. IEEE Transactions on Audio, Speech, and Language Processing, 18(6):1643–1654, 2010.

[27] Sebastian Ewert and Mark B. Sandler. An augmented lagrangian method for piano transcription using equal loudness thresholding and LSTM-based decoding. In Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, New York, USA, 2017.

[28] Florian Eyben, Sebastian Böck, Björn Schuller, and Alex Graves. Universal onset detection with bidirectional long-short term memory neural networks. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 589–594, Utrecht, The Netherlands, 2010.

[29] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 776–780, 2017.

[30] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. RWC music database: Popular, classical and jazz music databases. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), Paris, France, 2002.

[31] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. RWC music database: Music genre database and musical instrument sound database. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 229–230, Baltimore, Maryland, USA, 2003.

[32] Emad M. Grais, Gerard Roma, Andrew J. R. Simpson, and Mark D. Plumbley. Single-channel audio source separation using deep neural network ensembles. In Proceedings of the Audio Engineering Society (AES) Convention, Paris, France, May 2016.

[33] Thomas Grill and Jan Schlüter. Music boundary detection using neural networks on spectrograms and self-similarity lag matrices. In Proceedings of the European Signal Processing Conference (EUSIPCO), pages 1296–1300, Nice, France, 2015.

[34] Thomas Grill and Jan Schlüter. Music boundary detection using neural networks on combined features and two-level annotations. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 531–537, Màlaga, Spain, 2015.

[35] Philippe Hamel and Douglas Eck. Learning features from music audio with deep belief networks. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 339–344, Utrecht, The Netherlands, 2010.

[36] Philippe Hamel, Sean Wood, and Douglas Eck. Automatic identification of instrument classes in polyphonic and poly-instrument audio. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 399–404, Kobe, Japan, 2009.

[37] Philippe Hamel, Simon Lemieux, Yoshua Bengio, and Douglas Eck. Temporal pooling and multiscale learning for automatic annotation and ranking of music audio. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 729–734, Miami, Florida, 2011.

[38] Philippe Hamel, Yoshua Bengio, and Douglas Eck. Building musically-relevant audio features through multiple timescale representations. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 553–558, Porto, Portugal, 2012.

[39] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. CNN architectures for large-scale audio classification. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 131–135, 2017.

[40] Andre Holzapfel and Thomas Grill. Bayesian meter tracking on learned signal representations. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 262–268, New York City, United States, 2016.

[41] André Holzapfel, Matthew E. P. Davies, Jos ́e R. Zapata, Joa ̃o Lobato Oliveira, and Fabien Gouyon. Selective sampling for beat tracking evaluation. IEEE Transactions on Audio, Speech, and Language Processing, 20(9):2539–2548, 2012. doi: 10.1109/TASL.2012.2205244. URL http://dx.doi.org/10. 1109/TASL.2012.2205244.

[42] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis. Singing-voice separation from monaural recordings using deep recurrent neural networks. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 477–482, Taipei, Taiwan, 2014.

[43] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis. Joint optimization of masks and deep recurrent neural networks for monaural source separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(12):2136–2147, 2015.

[44] Eric J. Humphrey and Juan P. Bello. Rethinking automatic chord recognition with convolutional neural networks. In Proceedings of the IEEE International Conference on Machine Learning and Applications (ICMLA), pages 357–362, Boca Raton, USA, 2012.

[45] Eric J. Humphrey, Taemin Cho, and Juan P. Bello. Learning a robust tonnetz-space transform for automatic chord recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 453–456, Kyoto, Japan, 2012.

[46] Il-Young Jeong and Kyogu Lee. Learning temporal features using a deep neural network and its application to music genre classification. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 434–440, New York City, United States, 2016.

[47] Rainer Kelz and Gerhard Widmer. An experimental analysis of the entanglement problem in neural- network-based music transcription systems. In Proceedings of the AES International Conference on Semantic Audio, pages 194–201, Erlangen, Germany, 2017.

[48] Rainer Kelz, Matthias Dorfer, Filip Korzeniowski, Sebastian B ̈ock, Andreas Arzt, and Gerhard Widmer. On the potential of simple framewise approaches to piano transcription. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 475–481, New York City, United States, 2016.

[49] Filip Korzeniowski and Gerhard Widmer. Feature learning for chord recognition: The deep chroma extractor. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 37–43, New York City, United States, 2016.

[50] Filip Korzeniowski and Gerhard Widmer. End-to-end musical key estimation using a convolutional neural network. In Proceedings of the European Signal Processing Conference (EUSIPCO), Kos Island, Greece, 2017.

[51] Filip Korzeniowski and Gerhard Widmer. On the futility of learning complex frame-level language models for chord recognition. In Proceedings of the AES International Conference on Semantic Audio, pages 179–185, Erlangen, Germany, 2017.

[52] Florian Krebs, Sebastian Böck, Matthias Dorfer, and Gerhard Widmer. Downbeat tracking using beat synchronous features with recurrent neural networks. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 129–135, New York City, United States, 2016.

[53] Sangeun Kum, Changheun Oh, and Juhan Nam. Melody extraction on vocal segments using multi- column deep neural networks. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 819–825, New York City, United States, 2016.

[54] Simon Leglaive, Romain Hennequin, and Roland Badeau. Deep neural network based instrument extraction from music. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 121–125, Brisbane, Australia, 2015.

[55] Bernhard Lehner, Gerhard Widmer, and Sebastian B ̈ock. A low-latency, real-time-capable singing voice detection method with lstm recurrent neural networks. In Proceedings of the European Signal Processing Conference (EUSIPCO), pages 21–25, Nice, France, 2015.

[56] Antoine Liutkus, Fabian-Robert Stöter, Zafar Rafii, Daichi Kitamura, Bertrand Rivet, Nobutaka Ito, Nobutaka Ono, and Julie Fontecave. The 2016 signal separation evaluation campaign. In Proceedings of the International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA), pages 323–332, Grenoble, France, 2017.

[58] Yi Luo, Zhuo Chen, John R. Hershey, Jonathan Le Roux, and Nima Mesgarani. Deep clustering and conventional networks for music separation: Stronger together. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 61–65, New Orleans, USA, 2017.

[59] Matija Marolt. A connectionist approach to automatic transcription of polyphonic piano music. IEEE/ACM Transactions on Multimedia, 6(3):439–449, 2004.

[60] Marius Miron, Jordi Janer, and Emilia Gómez. Monaural score-informed source separation for classical music using convolutional neural networks. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), Suzhou, China, 2017.

[61] Juhan Nam, Jiquan Ngiam, Honglak Lee, and Malcolm Slaney. A classification-based polyphonic piano transcription approach using learned feature representations. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 175–180, Miami, Florida, 2011.

[62] Aditya Arie Nugraha, Antoine Liutkus, and Emmanuel Vincent. Multichannel music separation with deep neural networks. In Proceedings of the European Signal Processing Conference (EUSIPCO), pages 1748–1752, Budapest, Hungary, 2016.

[63] Aditya Arie Nugraha, Antoine Liutkus, and Emmanuel Vincent. Multichannel audio source separation with deep neural networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24 (9):1652–1664, 2016.

[64] Hyunsin Park and Chang D. Yoo. Melody extraction and detection through LSTM-RNN with harmonic sum loss. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 2766–2770, New Orleans, USA, 2017.

[65] Graham E. Poliner and Daniel P.W. Ellis. A discriminative model for polyphonic piano transcription. EURASIP Journal on Advances in Signal Processing, 2007(1), 2007.

[66| Jordi Pons and Xavier Serra. Designing efficient architectures for modeling temporal features with convolutional neural networks. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 2472–2476, 2017.

[67] Jordi Pons, Olga Slizovskaia, Rong Gong, Emilia Gómez, and Xavier Serra. Timbre analysis of music audio signals with convolutional neural networks. In Proceedings of the European Signal Processing Conference (EUSIPCO), Kos Island, Greece, 2017.

[68| Colin Raffel and Dan P. W. Ellis. Pruning subsequence search with attention-based embedding. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 554–558, Shanghai, China, 2016.

[69] Colin Raffel and Daniel P. W. Ellis. Accelerating multimodal sequence retrieval with convolutional networks. In Proceedings of the NIPS Multimodal Machine Learning Workshop, Montréal, Canada, 2015.

[70] Francois Rigaud and Mathieu Radenen. Singing voice melody transcription using deep neural networks. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 737–743, New York City, United States, 2016.

[71] Jan Schlüter. Learning to pinpoint singing voice from weakly labeled examples. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 44–50, New York City, United States, 2016.

[72] Jan Schlüter and Thomas Grill. Exploring data augmentation for improved singing voice detection with neural networks. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 121–126, Màlaga, Spain, 2015.

[73] Erik M. Schmidt and Youngmoo Kim. Learning rhythm and melody features with deep belief networks. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 21–26, Curitiba, Brazil, 2013.

[74] Siddharth Sigtia, Emmanouil Benetos, Srikanth Cherla, Tillman Weyde, Artur S. d'Avila Garcez, and Simon Dixon. An rnn-based music language model for improving automatic music transcription. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 53–58, Taipei, Taiwan, 2014.

[75] Siddharth Sigtia, Nicolas Boulanger-Lewandowski, and Simon Dixon. Audio chord recognition with a hybrid recurrent neural network. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 127–133, Màlaga, Spain, 2015.

[76] Siddharth Sigtia, Emmanouil Benetos, and Simon Dixon. An end-to-end neural network for polyphonic piano music transcription. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24 (5):927–939, 2016.

[77] Andrew J. R. Simpson, Gerard Roma, and Mark D. Plumbley. Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network. In Proceedings of the International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA), pages 429–436, Liberec, Czech Republic, 2015.

[78] Jordan Bennett Louis Smith, John Ashley Burgoyne, Ichiro Fujinaga, David De Roure, and J. Stephen Downie. Design and creation of a large-scale database of structural annotations. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 555–560, Miami, Florida, USA, 2011.

[79] Carl Southall, Ryan Stables, and Jason Hockman. Automatic drum transcription using bi-directional recurrent neural networks. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 591–597, New York City, United States, 2016.

[80] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing, 10(5):293–302, 2002.

[81] Stefan Uhlich, Franck Giron, and Yuki Mitsufuji. Deep neural network based instrument extraction from music. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 2135–2139, Brisbane, Australia, 2015.

[82] Stefan Uhlich, Marcello Porcu, Franck Giron, Michael Enenkl, Thomas Kemp, Naoya Takahashi, and Yuki Mitsufuji. Improving music source separation based on deep neural networks through data augmentation and network blending. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 261–265, New Orleans, USA, 2017.

[83] Karen Ullrich, Jan Schlüter, and Thomas Grill. Boundary detection in music structure analysis using convolutional neural networks. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 417–422, Taipei, Taiwan, 2014.

[84] Aäron van den Oord, Sander Dieleman, and Benjamin Schrauwen. Transfer learning by supervised pre-training for audio-based music classification. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 29–34, Taipei, Taiwan, 2014.

[85] Richard Vogl, Matthias Dorfer, and Peter Knees. Recurrent neural networks for drum transcription. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 730–736, New York City, United States, 2016.

[86] Xinquan Zhou and Alexander Lerch. Chord detection using deep learning. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 52–58, Màlaga, Spain, 2015.