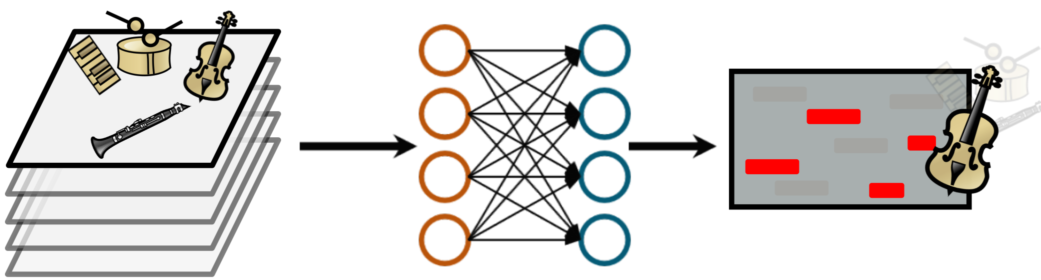
Deep Learning for Jazz Walking Bass Transcription

Jakob Abeßer[1,3], Stefan Balke[2], Klaus Frieler[3]
Martin Pfleiderer[3], Meinard Müller[2]



[1] Semantic Music Technologies Group, Fraunhofer IDMT, Ilmenau, Germany
[2] International Audio Laboratories Erlangen, Germany
[3] Jazzomat Research Project, University of Music Franz Liszt, Weimar, Germany

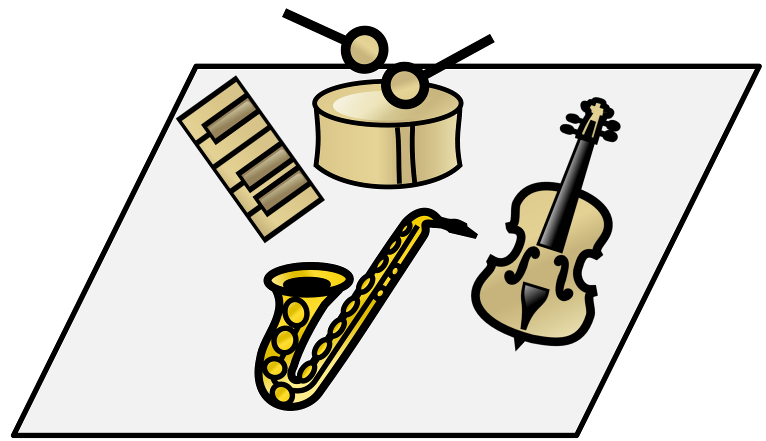
Motivation – What is a Walking Bass Line?

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

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

Fraunhofer IDMT

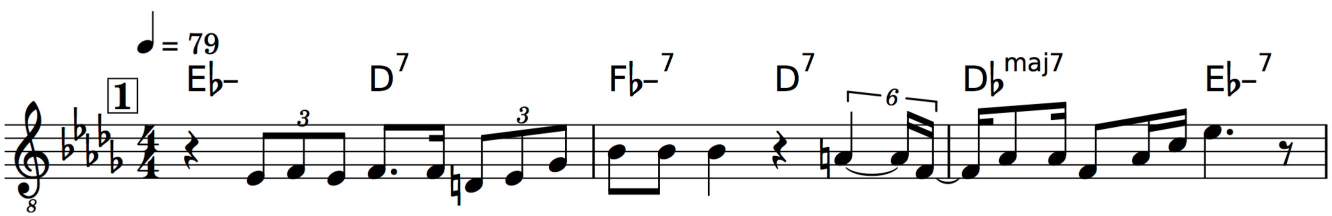
Motivation – How is this useful?



- **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**

Problem Setting

- **Challenges**
  - Bass is typically **not salient**
    - Overlap with drums (bass drum) and piano (lower register)
  - High **variability of recording quality** and playing styles
    - Example: Lester Young – Body and Soul



- **Goals**
  - Train a DNN to extract bass pitch saliency representation
  - Postprocessing: manual beat-annotations for beat-wise bass pitch

Outline

- Dataset
- Approach
  - Bass Saliency Mapping
  - Semi-Supervised Model Training
  - Beat-Informed Late Fusion
- Evaluation
- Results
- Summary & Outlook

Dataset

- **Weimar Jazz Database (WJD)** [1]
  - **456** high-quality jazz solo transcriptions
  - Annotations: solo melody, beats, chords, segments (phrase, chorus ...)
  - **41** files with **bass annotations**
- Data augmentation(+): Pitch-shifting +/- 1 semitone (sox library [2])

Dataset	Usage	Ann.	# Files	# Notes	Duration [h]
D_1	Training	✓	31	3899	0.43
D_1^+	Training	✓	93	11697	1.30
D_2	Training	-	237	-	7.16
D_2^+	Training	-	711	-	21.49
D_3	Test	✓	10	1101	0.12

Bass-Saliency Mapping

- **Data-driven** approach
  - Use **Deep Neural Network (DNN)** to learn **mapping** from magnitude spectrogram to bass saliency representation
- Spectral Analysis
  - Resampling to 22.05 kHz
  - **Constant-Q magnitude spectrogram** (librosa [3])
  - Pitch range 28 (41.2 Hz) – 67 (392 Hz)
- **Multilabel classification**
  - Input dimensions: 40 * $N_{ContextFrames}$
  - Output dimensions: 40
  - Learning Target: Bass pitch annotations from the WJD

Fraunhofer IDMT

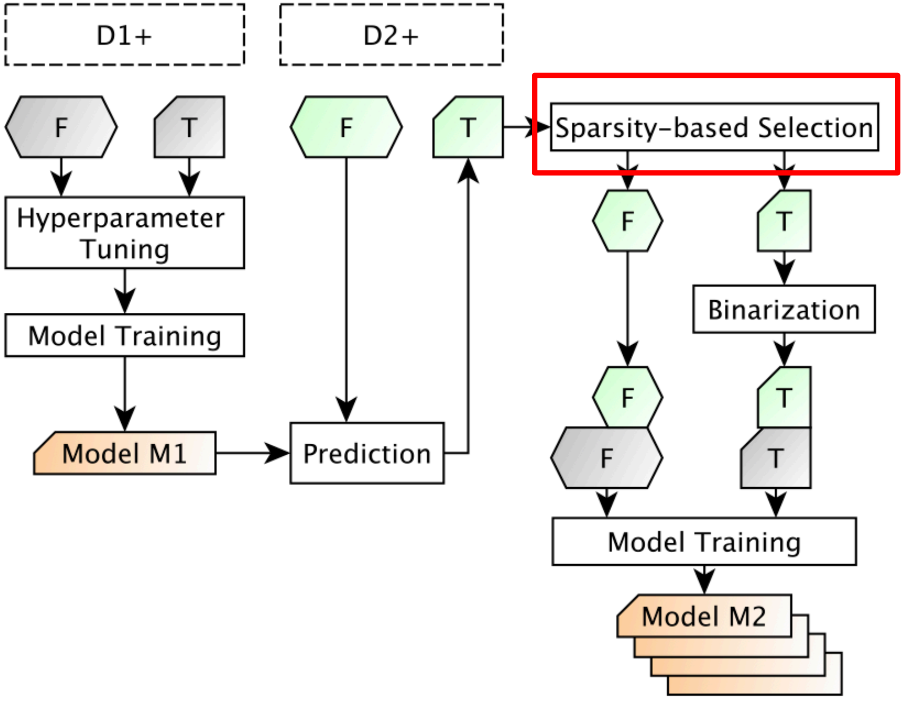
DNN Hyperparameter

- **Layer-wise training** [4, 5]
  - Least-squares estimate for weight & bias initialization
  - (3-5) fully connected layers, MSE loss
- **Frame-stacking** (3-5 context frames) & feature normalization
- Activation functions: ReLU, Sigmoid (final layer)
- Dropout & L_2 weight regularization
- Adadelta optimizer
  - Mini-batch size = 500
  - 500 epochs / layer
  - learning rate = 1

Hyperparameter	Values
# Hidden layers	3, **4**, 5
# Context frames	1, 3, **5**
Dropout (%)	0, **25**, 50
L_2 weight regularization	**disabled**, 10^{-3}

Semi-Supervised Training

- **Goal:** Select prediction on unseen data as additional training data

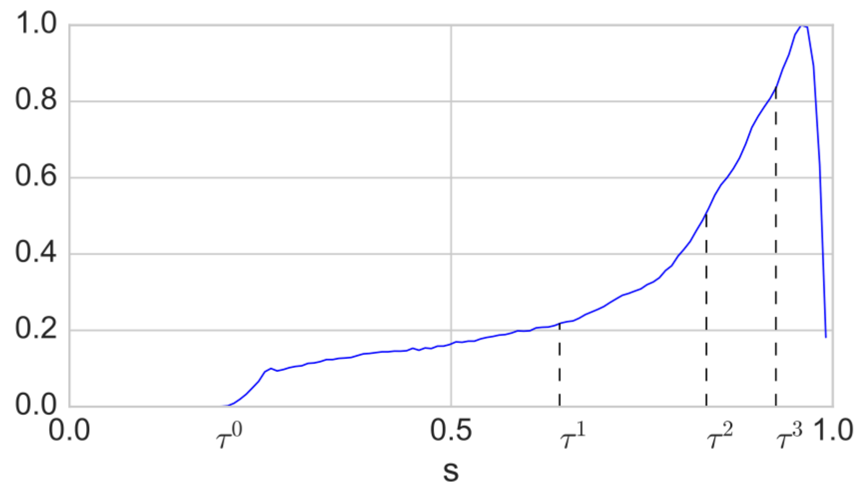


F: Features
T: Targets

Semi-Supervised Training
Sparsity-based Selection

- Train model on labelled dataset D_1^+ (3899 notes)
- Predictions on unlabelled dataset D_2^+ (11697 notes)
- Select additional training data via sparsity greater than threshold τ

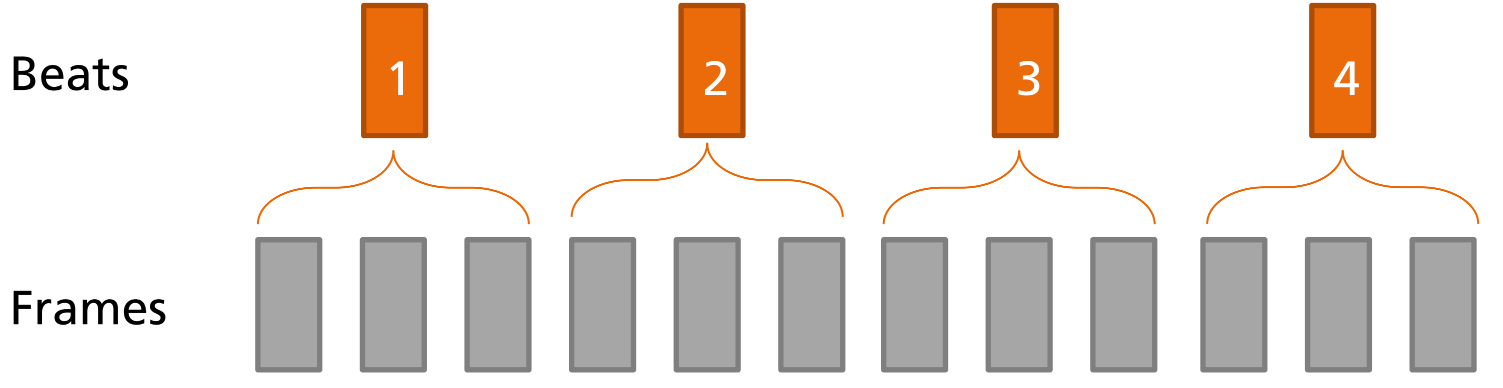
$$s(T) = \frac{\sqrt{N} - (\sum_{i=1}^{N} T_i)/(\sqrt{\sum_{i=1}^{N} T_i^2})}{\sqrt{N} - 1}$$



- Re-training

Beat-Informed Late Fusion

- Use manual beat-annotations from the Weimar Jazz Database
- Find most salient pitch per beat



Evaluation

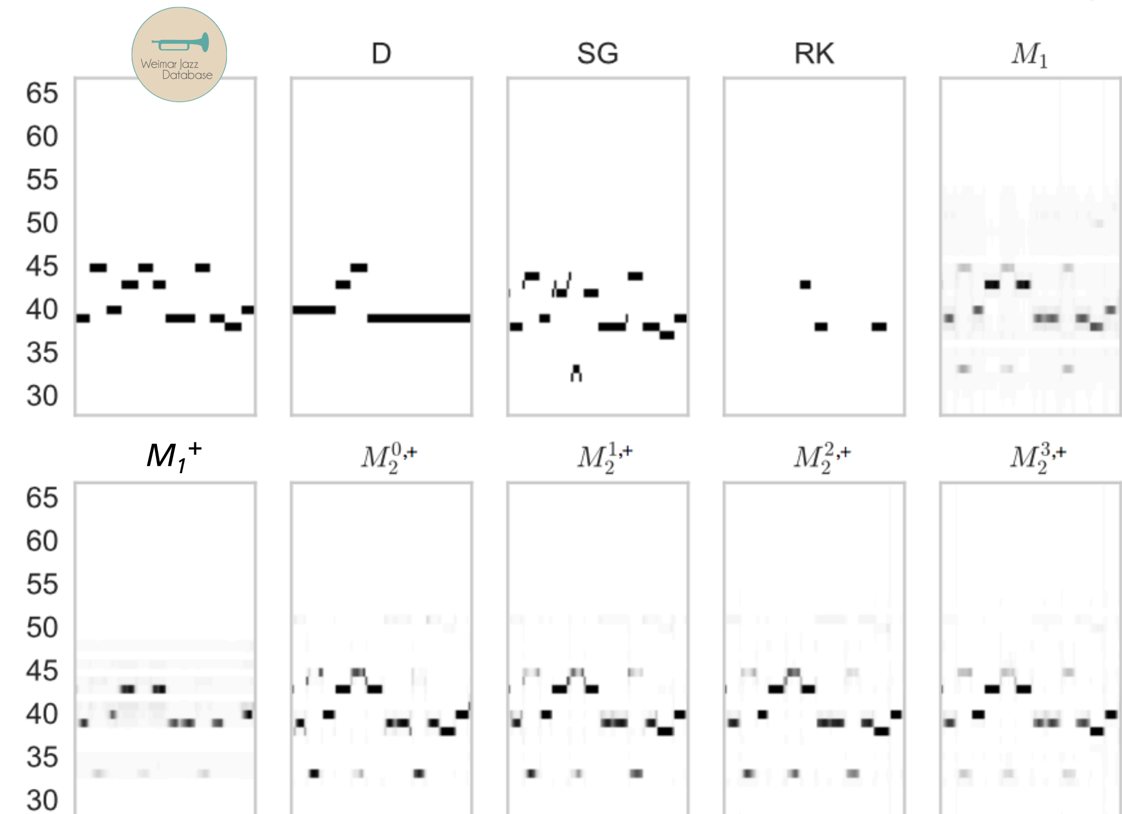
- Use manual beat-annotations from the Weimar Jazz Database
- Compare against state-of-the-art bass transcription algorithms
  - **D:** Dittmar, Dressler, and Rosenbauer [8]
  - **SG:** Salamon, Serrà, and Gómez [9]
  - **RK:** Ryynänen and Klapuri [7]

Dataset	Usage	Ann.	# Files	# Notes	Duration [h]
D_1	Training	✓	31	3899	0.43
D_1^+	Training	✓	93	11697	1.30
D_2	Training	-	237	-	7.16
D_2^+	Training	-	711	-	21.49
D_3	Test	✓	10	1101	0.12

Example

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

- Chet Baker: “Let’s Get Lost” (0:04 – 0:09)



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

Results

Alg.	Frame-wise		Beat-wise		Sparseness
	A	A_{PC}	A	A_{PC}	s
SG	0.28 (0.14)	0.39 (0.15)	0.68 (0.22)	0.75 (0.21)	-
RK	0.12 (0.13)	0.18 (0.14)	0.60 (0.27)	0.64 (0.26)	-
D	0.37 (0.20)	0.41 (0.19)	0.72 (0.16)	0.75 (0.15)	-
M_1	0.31 (0.09)	0.43 (0.10)	0.71 (0.17)	0.78 (0.14)	0.684 (0.035)
M_1^+	**0.57** (0.13)	0.70 (0.11)	**0.83** (0.13)	0.89 (0.11)	0.761 (0.018)
$M_2^{0,+}$	0.54 (0.12)	0.68 (0.11)	0.81 (0.14)	0.88 (0.12)	**0.954** (0.010)
$M_2^{1,+}$	0.54 (0.13)	0.70 (0.11)	0.81 (0.14)	0.89 (0.11)	0.935 (0.015)
$M_2^{2,+}$	0.55 (0.12)	0.71 (0.11)	0.82 (0.14)	0.89 (0.12)	0.922 (0.019)
$M_2^{3,+}$	0.56 (0.12)	0.70 (0.11)	0.82 (0.14)	0.88 (0.12)	0.862 (0.030)

Summary

- Data-driven approach seems to enhance non-salient instruments.
- Beneficial
  - Data augmentation & dataset enlargement
  - Frame stacking (stable bass pitches)
  - Beat-informed late fusion
- Semi-supervised training did not improve accuracy but made bass-saliency maps sparser
- Model is limited to training set‘s pitch range

Acknowledgements

- Thanks to the authors for sharing bass transcription algorithms / results.
- Thanks to the students for transcribing the bass lines
- German Research Foundation
  - DFG-PF 669/7-2 : Jazzomat Research Project (2012-2017)
  - DFG MU 2686/6-1: Stefan Balke and Meinard Müller

References

[1] Weimar Jazz Database: http://jazzomat.hfm-weimar.de
[2] sox http://sox.soundforge.net
[3] McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., and Nieto, O., "librosa: Audio and Music Signal Analysis in Python," in Proc. of the Scientific Computing with Python conference (Scipy), Austin, Texas, 2015.
[4] Uhlich, S., Giron, F., and Mitsufuji, Y., "Deep neural network based instrument extraction from music," in Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 2135–2139, Brisbane, Australia, 2015.
[5] Balke, S., Dittmar, C., Abeßer, J., and Müller, M., "Data-Driven Solo Voice Enhancement for Jazz Music Retrieval," in Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, USA, 2017.
[6] Hoyer, P.O., "Non-negative Matrix Factorization with Sparseness Constraints," J. of Machine Learning Research, 5, pp. 1457–1469, 2004.

[7] Ryynänen, M.P. and Klapuri, A., "Automatic transcription of melody, bass line, and chords in polyphonic music," Computer Music J., 32, pp. 72–86, 2008

[8] Dittmar, C., Dressler, K., and Rosenbauer, K., "A Toolbox for Automatic Transcription of Polyphonic Music," Proc. of the Audio Mostly Conf., pp. 58–65, 2007.

[9] Salamon, J., Serrà, J., and Gómez, E., "Tonal Representations for Music Retrieval: From Version Identification to Query-by-Humming," Int. J. of Multimedia Inf. Retrieval, 2, pp. 45–58, 2013

Fraunhofer IDMT

Thank You!

- Jazzomat Research Project & Weimar Jazz Database
  - http://jazzomat.hfm-weimar.de/
- Python code and trained model available
  - https://github.com/jakobabesser/walking_bass_transcription_dnn
- Additional online demos
  - https://www.audiolabs-erlangen.de/resources/MIR/2017-AES-WalkingBassTranscription