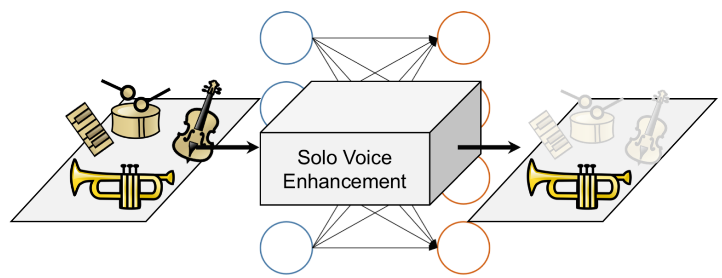
Data-Driven Solo Voice Enhancement for Jazz Music Retrieval



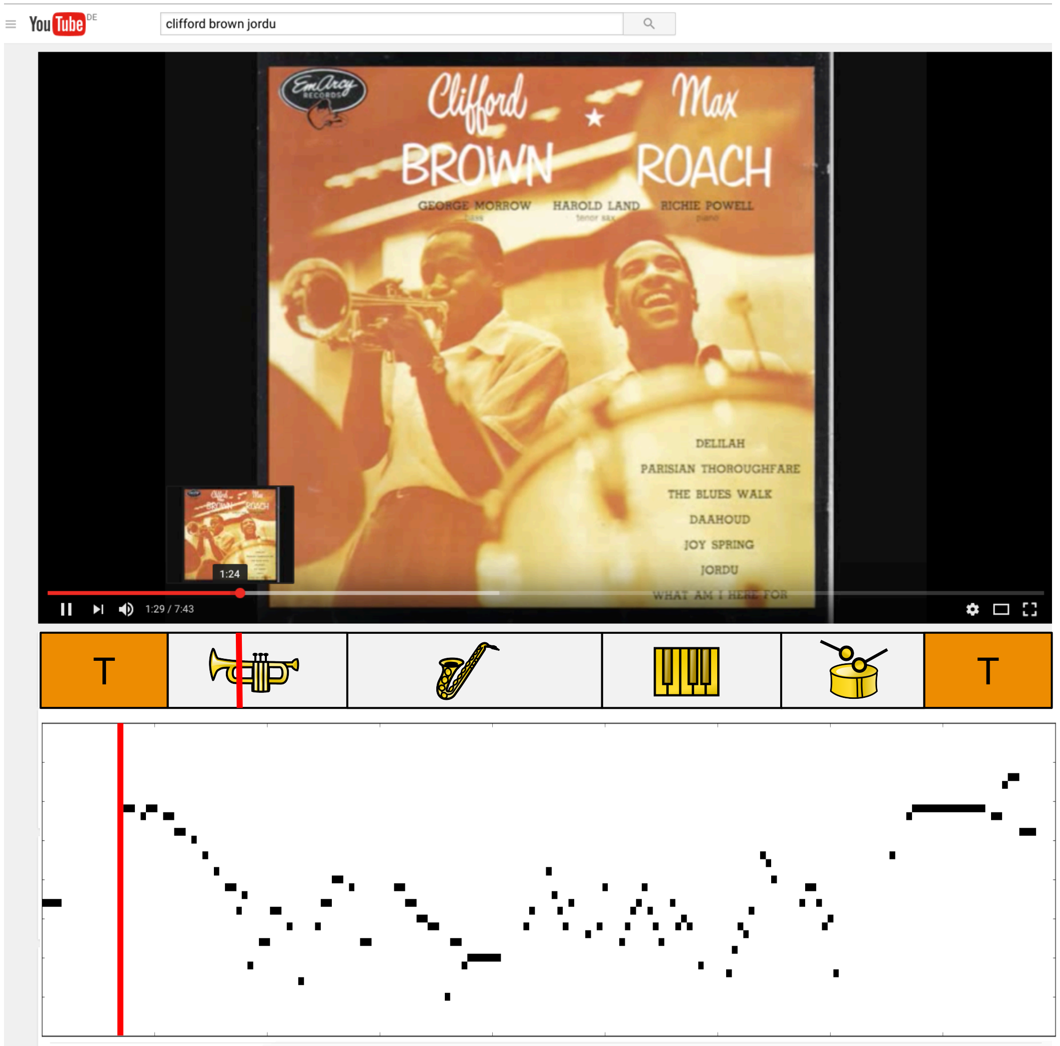
Stefan Balke[1], Christian Dittmar[1], Jakob Abeßer[2], Meinard Müller[1]

[1]International Audio Laboratories Erlangen
[2]Fraunhofer Institute for Digital Media Technology IDMT

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

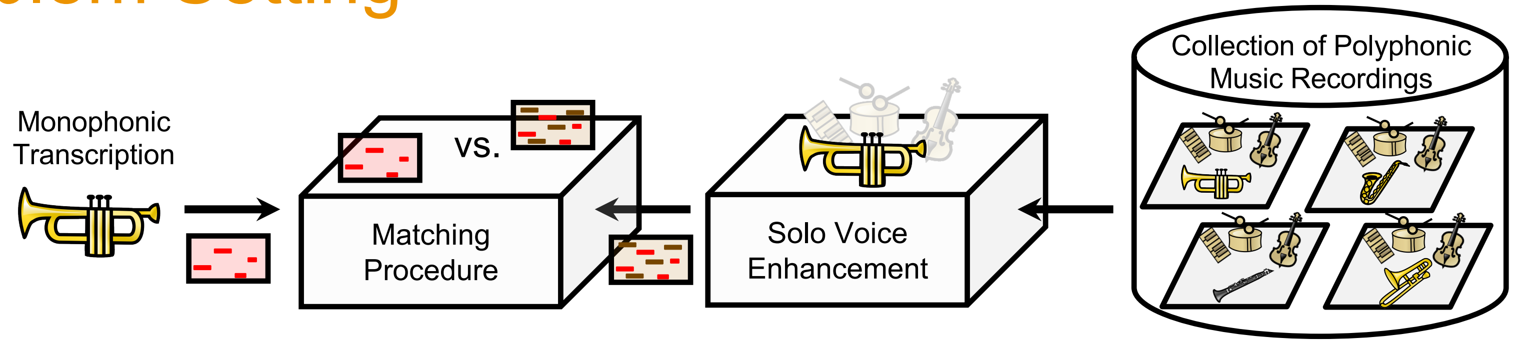
Fraunhofer IIS

Vision



clifford brown jordu
Clifford ★ Max
BROWN ROACH
GEORGE MORROW HAROLD LAND RICHIE POWELL
DELILAH
PARISIAN THOROUGHFARE
THE BLUES WALK
DAAHOUD
JOY SPRING
JORDU
WHAT AM I HERE FOR
1:24
1:29 / 7:43
T
T

Problem Setting



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.

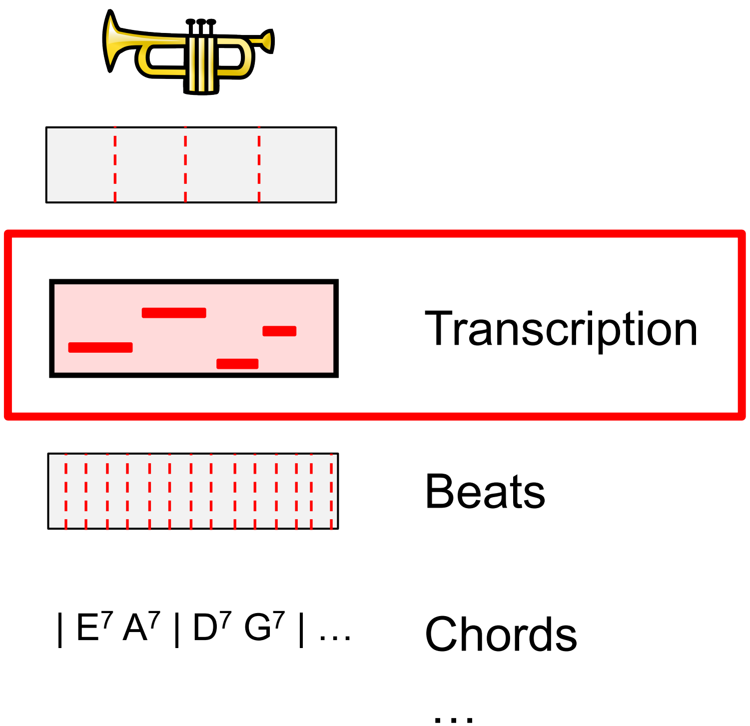
Overview

Philippe Halsman, “Louis Armstrong”

1. Background on the Data
2. DNN Architecture & Training
3. Evaluation within Retrieval Scenario

Weimar Jazz Database (WJD)

[Pfleiderer17]

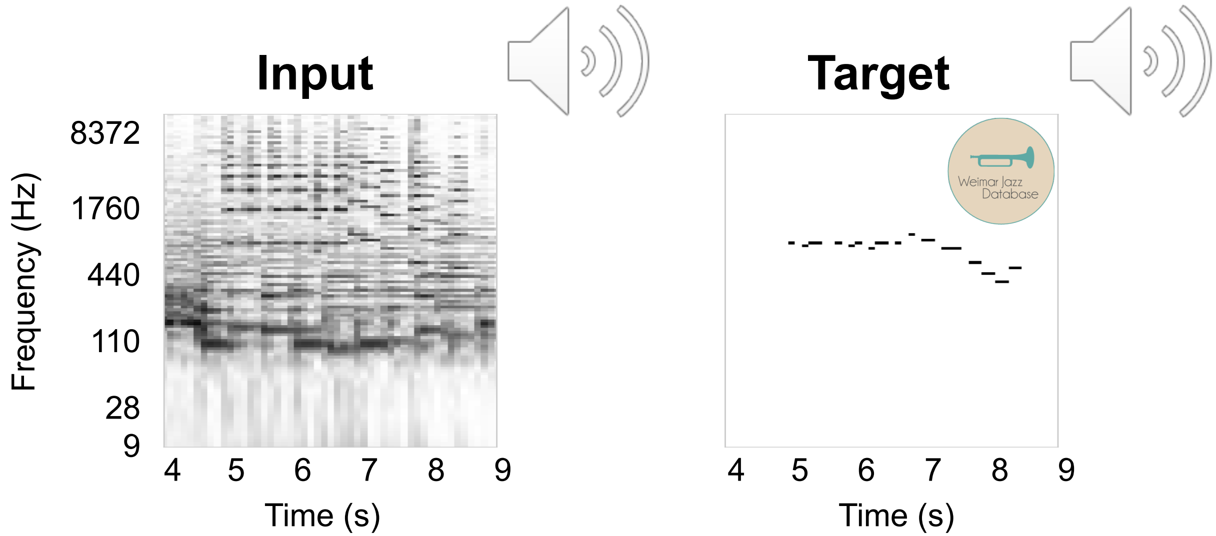


- 299 transcribed jazz solos of monophonic instruments.
- Transcriptions specify a musical pitch for physical time instances.
- 570 min. of audio recordings.

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

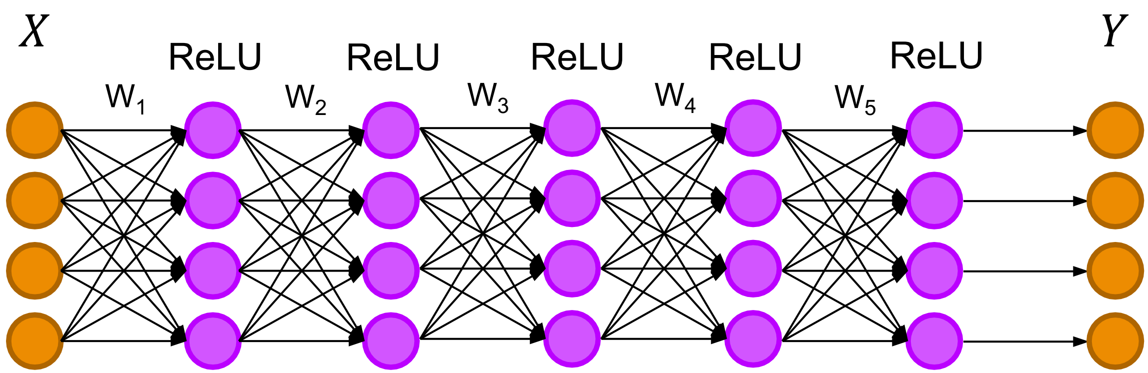
DNN Training

- **Input:** Log-freq. STFT frame (120 semitones, 10 Hz feature rate)
  - TF-representation of jazz solo recording
- **Output:** Pitch activations (120 semitones, 10 Hz feature rate)
- **Target:** TF-representation with solo instrument’s pitch activations



DNN Architecture

$X :=$ Input, $Y :=$ Output, $T :=$ Target, $L :=$ Loss



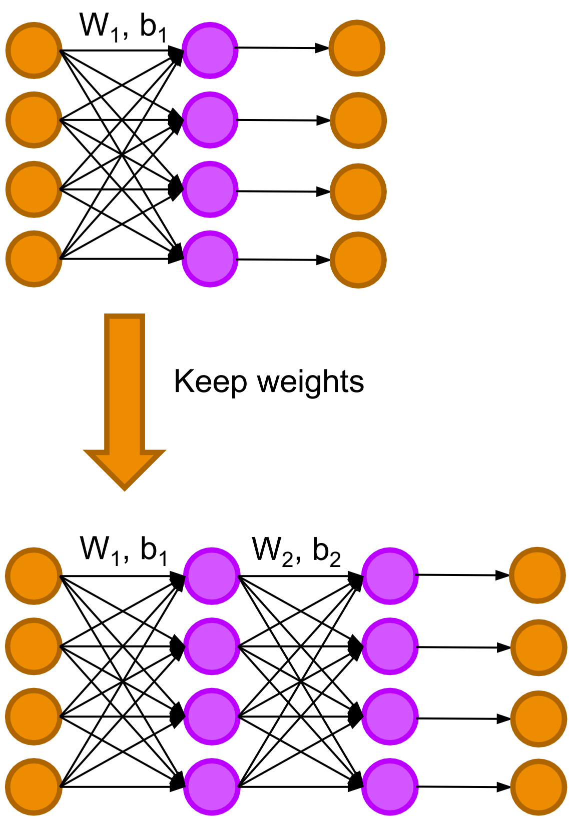
$$L = \mathrm{MSE}(X, Y)$$

Dimensions: 120 120 120 120 120 120 120

- Basic feed-forward DNN with 5 hidden layers.
- Training is applied layer-wise [Bengio06], extended in [Uhlich15].

Layer-Wise Training

[Uhlich15]



- Initialize weights (W_1) and bias (b_1) with Linear Least Squares (LLS)
- Train 600 epochs …
- Interpret output of trained network as input to the next layer

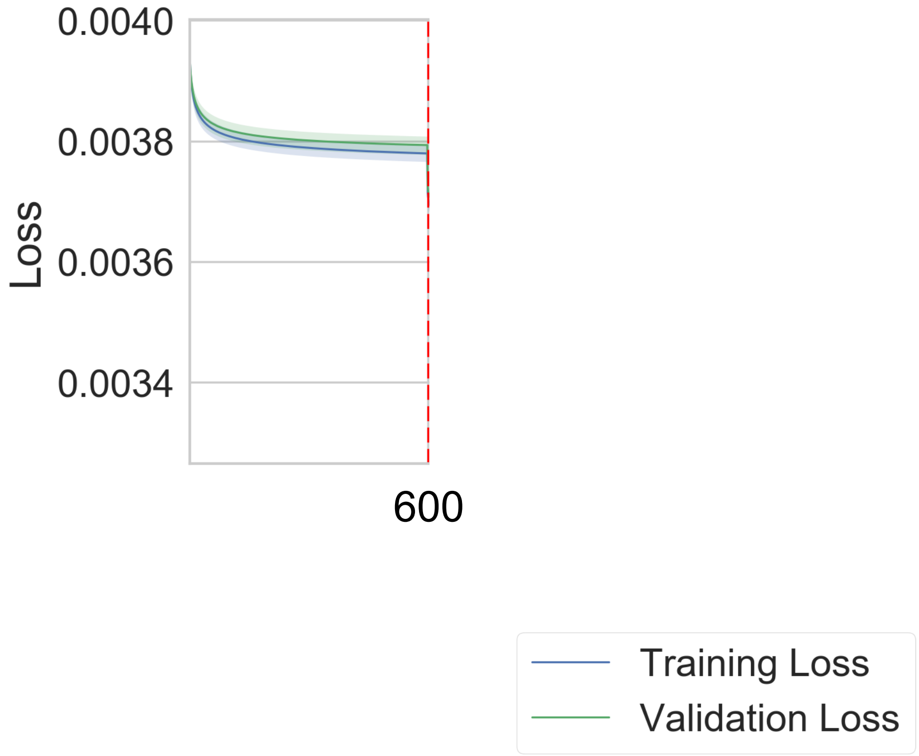
  - Append next layer
  - Initialize W_2 and b_2 with LLS
  - Train 600 epochs …

Training Details

- **Total Duration:** 570 min.
- **Active Solo Frames:** 62%
- **Split:** 10-fold cross-validation
  - Training Set: 63%, Validation Set: 27%
  - Test Set: 10%

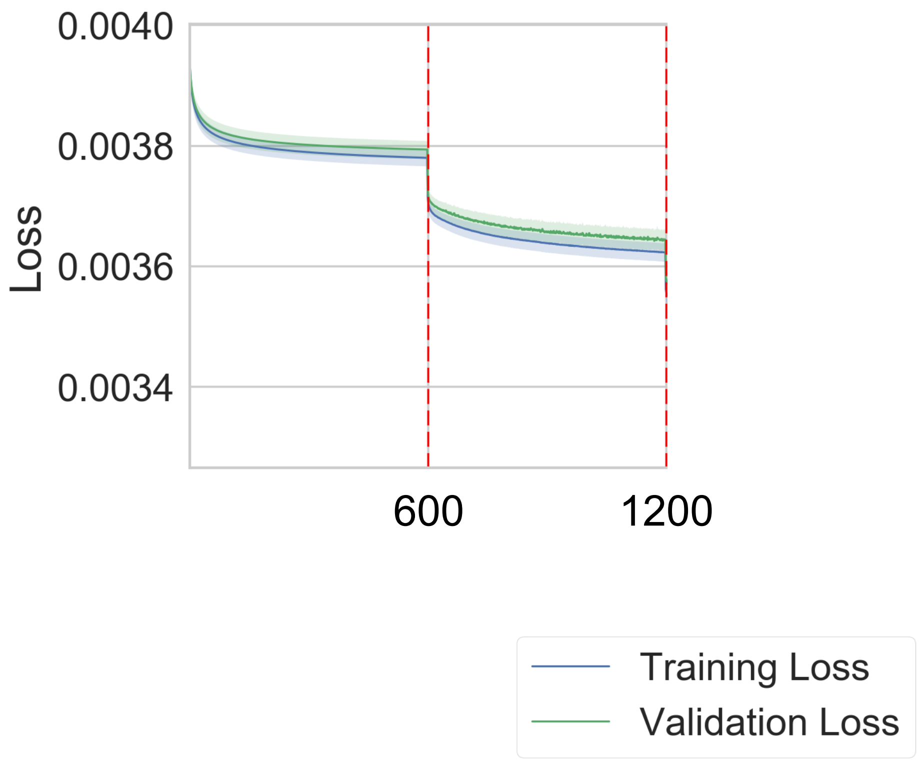
- **Loss:** Mean-Squared Error
- **Optimizer:** Stochastic Gradient Descent
  - Mini-batch size = 100 frames (10 s)
  - Learning Rate = 10^{-6}, Momentum = 0.9
  - 600 epochs per layer (3000 epochs in total)

Training Loss
Number of Hidden Layers: 1



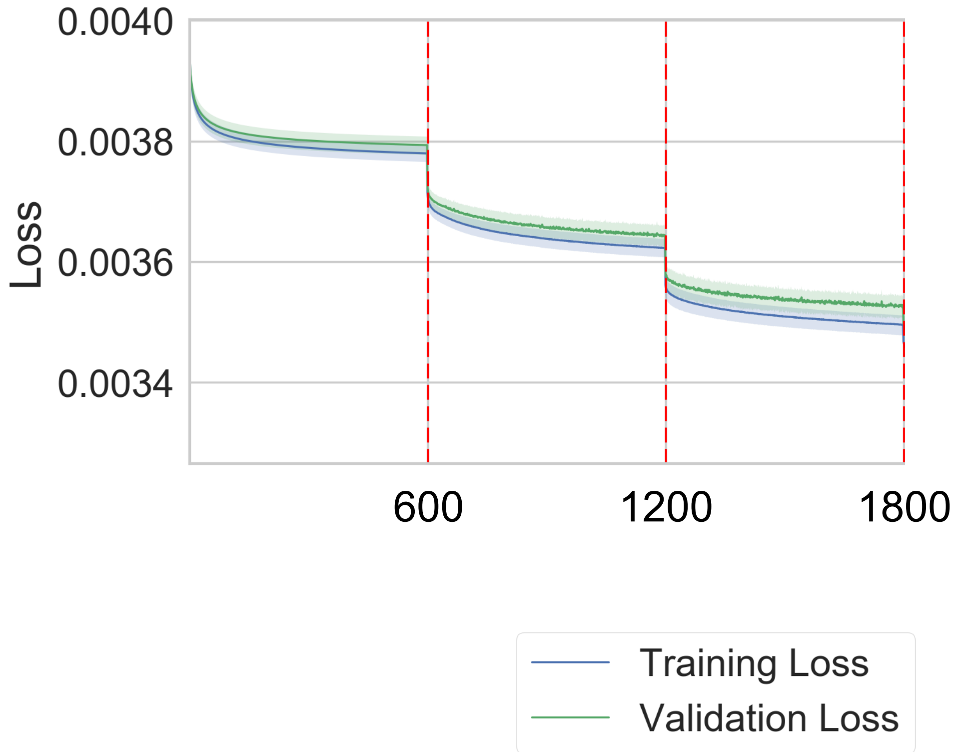
Training Loss

Number of Hidden Layers: 2



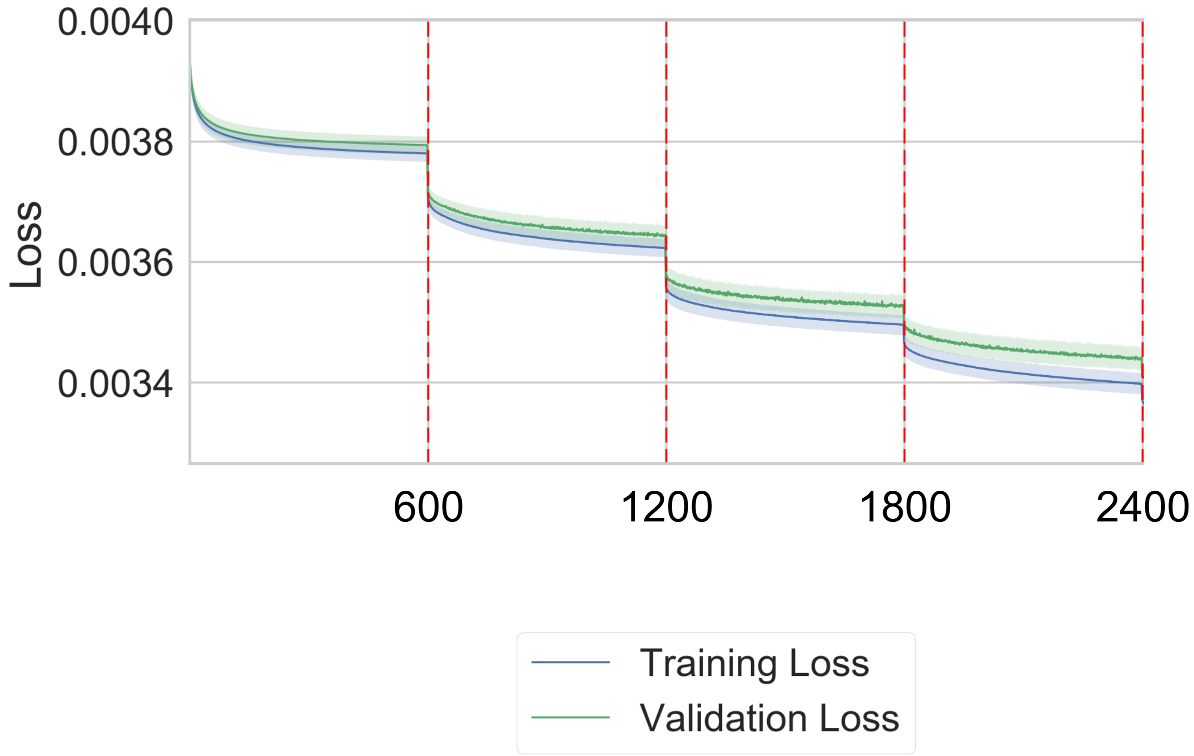
Training Loss

Number of Hidden Layers: 3



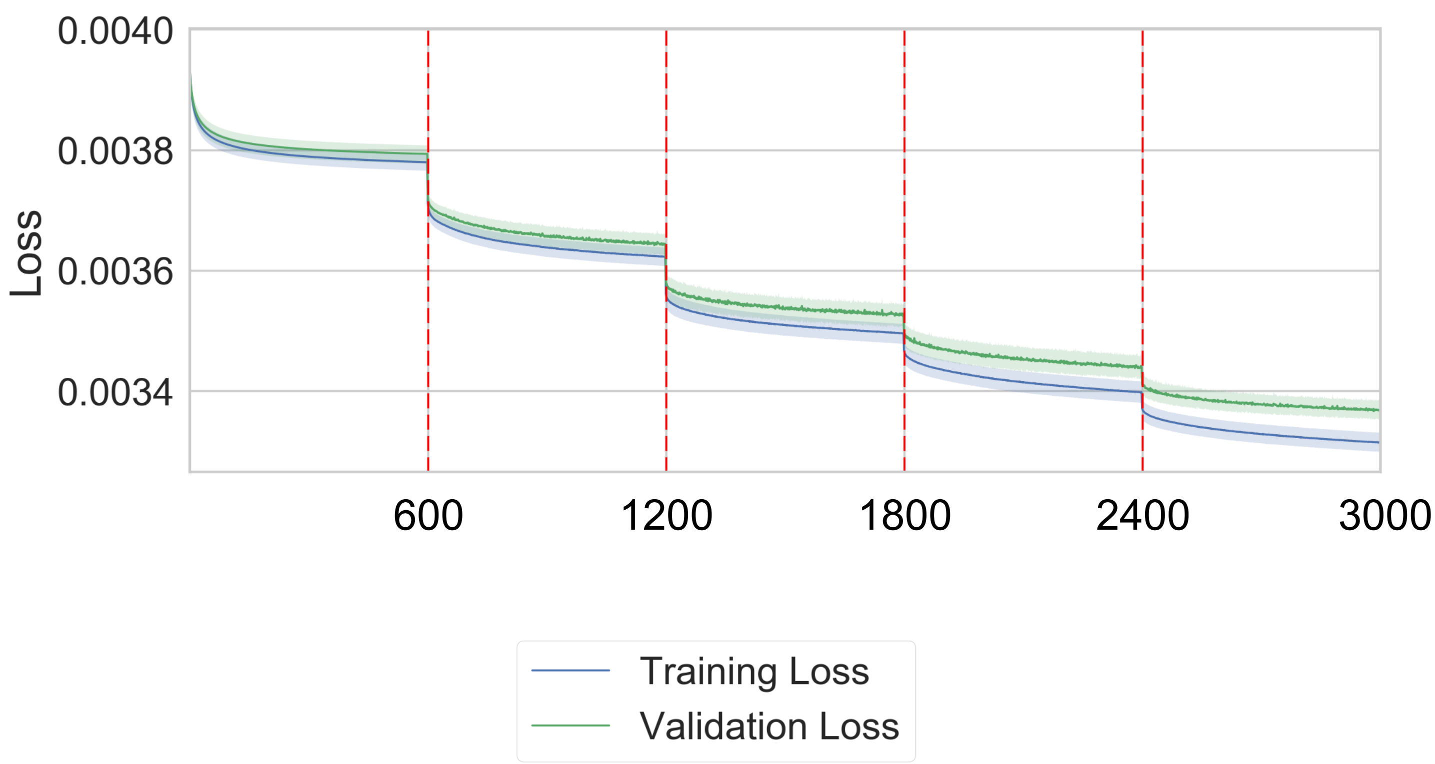
Training Loss

Number of Hidden Layers: 4

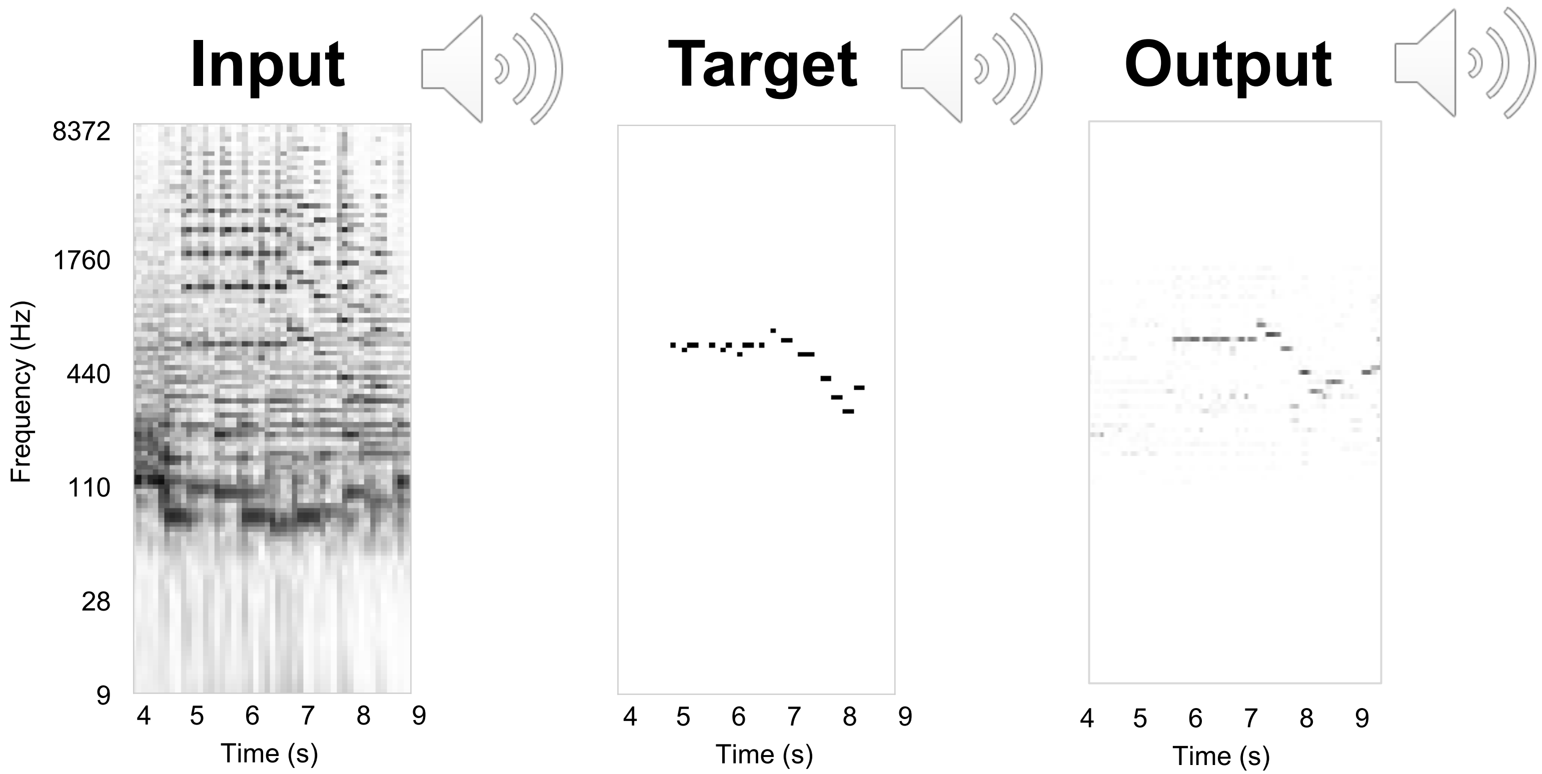


Training Loss

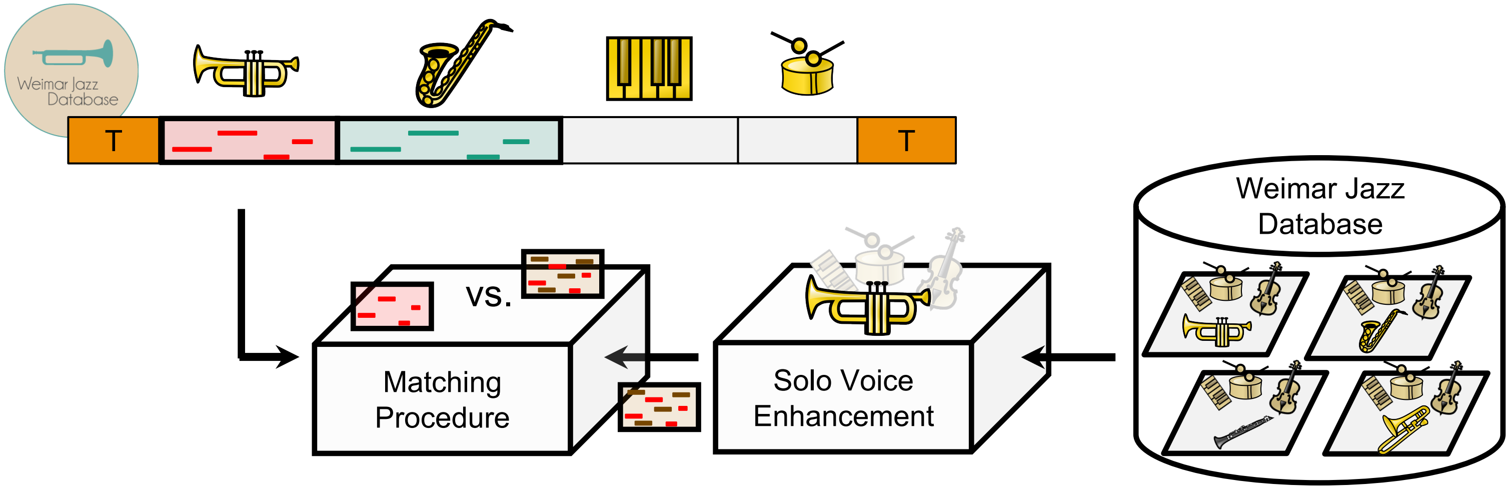
Number of Hidden Layers: 5



Qualitative Evaluation



Experiment: Jazz Music Retrieval



- 30 queries with a duration of 25 s for each fold
- 1 relevant document in the database per query
- Additional queries by shortening to [20, 15, 10, 8, 6, 5, 4, 3] s
- Evaluation measure is the mean reciprocal rank (MRR)

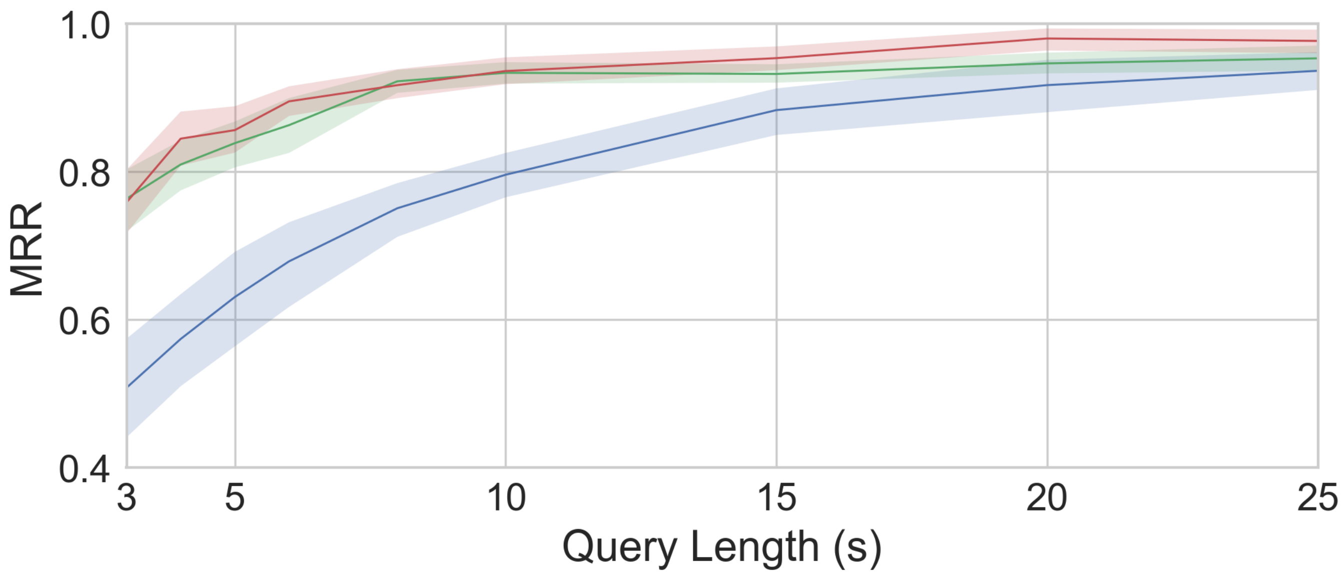
Experiment: Jazz Music Retrieval

Results

Baseline Chroma-based matching [Mueller15]

Melodia Quantized F0-trajectory [Salamon13]

DNN



Conclusions

- Data-driven approaches seem to be beneficial for solo voice enhancement.
- Data-driven and model-based approaches show similar performance in a retrieval scenario.

Future Work

- Investigate scenarios where predominance assumption is violated, e. g., walking bass transcription.
- Train instrument-specific models, e. g., implicit instrument recognition.
- Utilize DNN’s output for other tasks (e. g., F0-tracking).

Audio examples, trained models, and data:
https://www.audiolabs-erlangen.de/resources/MIR/2017-ICASSP-SoloVoiceEnhancement
stefan.balke@audiolabs-erlangen.de

feat. Masataka Goto, Mark Plumbley, and Udo Zölzer as keynote speakers.

More Details: http://www.aes.org/conferences/2017/semantic/

Fraunhofer IIS FAU FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG AUDIO LABS

References

[Salamon13] Justin Salamon, Joan Serrà, and Emilia Gómez, “Tonal representations for music retrieval: from version identification to query-by-humming,” Int. Journal of Multimedia Information Retrieval, vol. 2, no. 1, pp. 45–58, 2013.

[Rigaud16] F. Rigaud and M. Radenen, “Singing voice melody transcription using deep neural networks,” in Proc. of the Int. Society for Music Information Retrieval Conf. (ISMIR), New York City, USA, 2016, pp. 737–743.

[Bittner15] Rachel M. Bittner, Justin Salamon, Slim Essid, and Juan Pablo Bello, “Melody extraction by contour classification,” in Proc. of the Int. Society for Music Information Retrieval Conf. (ISMIR), Málaga, Spain, 2015, pp. 500–506.

[Bengio06] Yoshua Bengio, Pascal Lamblin, Dan Popovici, Hugo Larochelle, “Greedy Layer-Wise Training of Deep Networks”, in Proc. of the Annual Conference on Neural Information Processing Systems (NIPS), 2006, pp. 153–160.

[Uhlich15] Stefan Uhlich, Franck Giron, and Yuki Mitsufuji, “Deep neural network based instrument extraction from music,” in Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), April 2015, pp. 2135–2139.

[Pfleiderer17] The Jazzomat Research Project, “Database download, last accessed: 2016/02/17,” http://jazzomat.hfm-weimar.de.

[Mueller15] Meinard Müller, “Fundamentals of Music Processing”, Springer Verlag, 2015.