Data-Driven Solo Voice Enhancement for Jazz Music Retrieval

Stefan Balke\textsuperscript{1}, Christian Dittmar\textsuperscript{1}, Jakob Abeßer\textsuperscript{2}, Meinard Müller\textsuperscript{1}

\textsuperscript{1}International Audio Laboratories Erlangen
\textsuperscript{2}Fraunhofer Institute for Digital Media Technology IDMT
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

1. Background on the Data
2. DNN Architecture & Training
3. Evaluation within Retrieval Scenario
Weimar Jazz Database (WJD)

- 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

---

© AudioLabs, 2017

Data-Driven Solo Voice Enhancement for Jazz Music Retrieval

Balke et al.
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

\[ X := \text{Input}, \ Y := \text{Output}, \ T := \text{Target}, \ L := \text{Loss} \]

\[ L = \text{MSE}(X, Y) \]

Dimensions: 120 120 120 120 120 120 120 120

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

- 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

Keep weights

- 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

![Graph showing training and validation loss over 600 iterations.](image)
Training Loss
Number of Hidden Layers: 2
Training Loss

Number of Hidden Layers: 3

- Training Loss
- Validation Loss

Number of Iterations:
- 600
- 1200
- 1800
Training Loss
Number of Hidden Layers: 4

![Graph showing Training Loss and Validation Loss over epochs with a maximum of 2400 epochs. The graph indicates a downward trend in loss values with increasing epochs.]
**Training Loss**

Number of Hidden Layers: 5
Qualitative Evaluation

**Input**

**Target**

**Output**
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**

![Graph showing MRR vs Query Length (s)]
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/
References


