IMPROVING BASS SALIENCY ESTIMATION USING LABEL PROPAGATION AND TRANSFER LEARNING

Jakob Abeßer[1] **Stefan Balke**[2] **Meinard Müller**[2]
[1] Semantic Music Technologies Group, Fraunhofer IDMT, Ilmenau, Germany
[2] International Audio Laboratories Erlangen, Germany
jakob.abesser@idmt.fraunhofer.de

ABSTRACT

In this paper, we consider two methods to improve an algorithm for bass saliency estimation in jazz ensemble recordings which are based on deep neural networks. First, we apply label propagation to increase the amount of training data by transferring pitch labels from our labeled dataset to unlabeled audio recordings using a spectral similarity measure. Second, we study in several transfer learning experiments, whether isolated note recordings can be beneficial for pre-training a model which is later fine-tuned on ensemble recordings. Our results indicate that both strategies can improve the performance on bass saliency estimation by up to five percent in accuracy.

1. INTRODUCTION

Recent developments in the field of machine learning, in particular deep learning, stimulated a significant performance boost in various Music Information Retrieval (MIR) tasks [7] such as audio tagging [23], audio source separation [27], and automatic music transcription (AMT) [15]. One major challenge in training deep neural networks (DNNs) that generalize well to unseen data lies in the large amount of required labeled training data, which is often not available.

In this context, semi-supervised learning strategies can help to solve this data problem. A first approach is to apply *transfer learning*, i. e., training a network on a related classification task and fine-tune the model parameters for the target task with the (usually smaller) amount of training data at hand [10, 18]. Both training steps are fully supervised and therefore require labeled datasets. A second approach is *label propagation*, where labels from labeled feature vectors are propagated to unlabeled feature vectors if some pre-defined similarity measure exceeds a particular threshold. Label propagation can help to significantly increase the amount of available training data.

In this paper, we focus on the task of estimating the pitch salience of the bass instrument in jazz ensemble recordings. In general, pitch saliency refers to a likelihood measure of an instrument playing certain pitch frequencies at given times. Figure 1 illustrates the DNN-based approach that we use. Given a time-frequency representation of an audio recording of a jazz ensemble, the goal is to estimate a bass salience representation in a frame-wise fashion. As outlined in [1], these frame-wise estimates of the bass saliency can then be aggregated using beat annotations to obtain a beat-wise pitch representation, which is a musically meaningful approximation of the commonly played walking bass lines in jazz music.

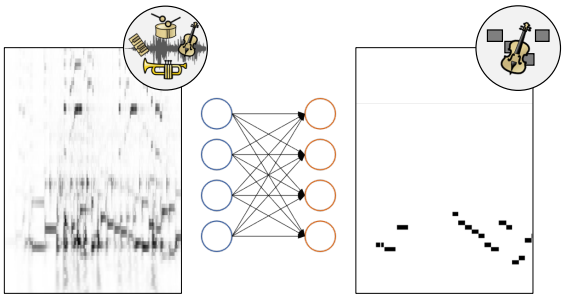

Figure 1. Flowchart summarizing the main idea of training a deep neural network to learn a mapping function from a constant-Q spectrogram of a jazz ensemble recording (left) towards a bass pitch saliency representation (right) using a deep neural network.

As the main contributions of this paper, we investigate transfer learning and label propagation strategies for improving fully-connected deep neural networks for the task of bass saliency estimation, as shown in Figure 2. Both techniques aim to compensate the lack of available labeled data for the task of bass salience estimation. For label propagation, the core idea is to enrich an unlabeled dataset with labels from a labeled dataset. For transfer learning, we investigate whether training models on music data of lower timbral complexity (e. g., isolated instrument tones) is beneficial for transferring them to complex mixture recordings.

The remainder of this paper is structured as follows. In Section 2, we review related work. In Section 3, we introduce the underlying datasets used throughout our experiments and propose additional data augmentation steps. Section 4 introduces the feature extraction approach, DNN architecture, and the evaluation methodology. In Section 5,

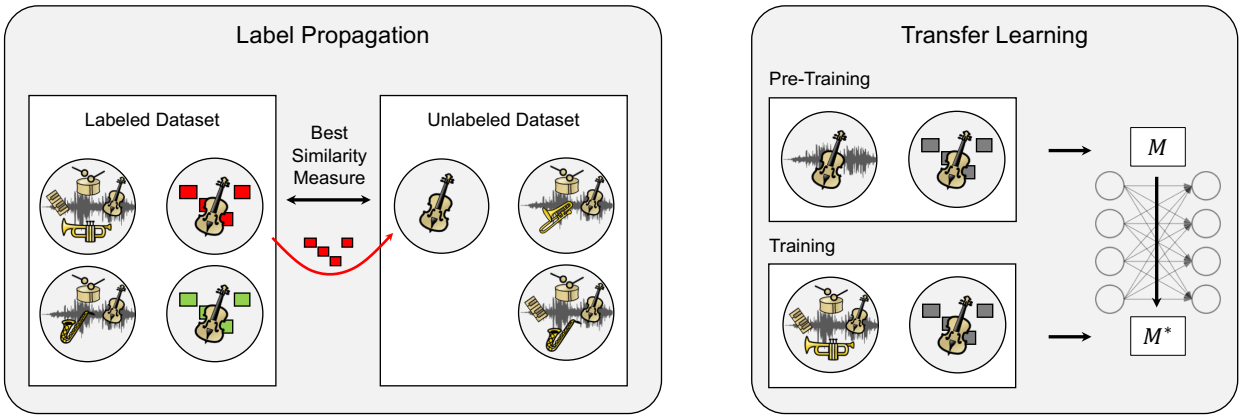

Figure 2. Illustration of label propagation and transfer learning. In label propagation (left), an unlabeled audio dataset is enriched with frame-wise pitch labels from a labeled dataset. In transfer learning (right), a DNN is first pre-trained on a dataset with lower complexity (isolated bass recordings) and then trained further on jazz ensemble recordings.

we present experiments towards hyperparameter optimization (Section 5.1), label propagation (Section 5.2), as well as transfer learning (Section 5.3). Finally, Section 6 concludes our work and gives perspectives towards future work.

2. RELATED WORK

Salience representations are a common intermediate representation in many automatic music transcription (AMT) systems prior to the formation of note events. Most previous approaches for bass saliency estimation rely on handcrafted algorithms rather than on automatically learnt mappings. For instance, Goto derives pitch saliency values from predominant peaks in a spectral representation based on instantaneous frequency values [13]. Ryynänen and Klapuri compute a saliency measure for a given pitch from a weighted sum over the spectral magnitude values at its harmonic frequencies [24]. Salamon et al. apply harmonic summation based on a logarithmic frequency representation combined with instantaneous frequency estimation methods [25].

In [1], the mapping from a constant-Q spectrogram to a bass saliency function is automatically learnt using fully-connected deep neural networks. The authors also investigated a semi-supervised learning step where parts of predicted pitch saliency estimations on unlabeled audio data were added to the training data based on a sparsity criterion. The modeling strategy was inspired by Balke et al. [2], who used a similar approach to estimate a saliency representation of the predominant melody instrument in jazz music recordings. Bittner et al. [4] proposed a fully convolutional neural network (CNN) to extract a saliency representations from different constant-Q transforms used as input for both multiple fundamental frequency estimation and melody tracking.

Models with state-of-the-art performance in related disciplines such as image processing (mostly CNN-based models) are rarely trained from scratch due to the large amount of required training data. Instead, only the last layers of existing "general purpose" classification models (such as the ImageNet model [9]) are fine-tuned for related classification tasks using smaller amounts of training data [21]. Similarly, in the field of MIR, Choi et al. [8] used a pre-trained CNN-based feature extractor trained on music tagging data for related music classification and regression tasks. However, for the task of AMT, no such general-purpose model was established so far.

Dataset	Usage	Labels	# Feature Vectors	Duration [h]
ISO$^+$	Training	✓	448,626	5.79
WJD$^+$	Training	✓	305,507	3.94
WJD$^-$	Training	-	500,000	6.45
WJD$^+$-TEST	Test	✓	8,318	0.1

Table 1. Summary of the datasets. The number of feature vectors after data augmentation and voiced frame selection as well as the corresponding duration in hours is given in the last two columns. For the WJD$^-$ dataset, 500,000 feature vectors were randomly selected due to memory limitations on the hardware in use for the experiments.

3. DATASETS

The spectral characteristics of the targeted upright bass tones are affected by different factors of variation such as the pitch, the loudness, as well as the overlap with tones from simultaneously playing instruments. In our considered datasets, we use different sets of upright bass recordings that try to address these variations. All considered audio files used in this paper include an acoustic upright bass played with the plucked (pizzicato) plucking style—as opposed to using a bow—as this is the common playing style for jazz bass players. Table 1 gives an overview of the datasets used, which we discuss in the following.

3.1 Isolated Upright Bass Recordings (ISO$^+$)

The ISO$^+$ dataset is a collection of isolated chromatic note recordings. The recordings stem from various commercial and non-commercial upright bass sample datasets: Adam

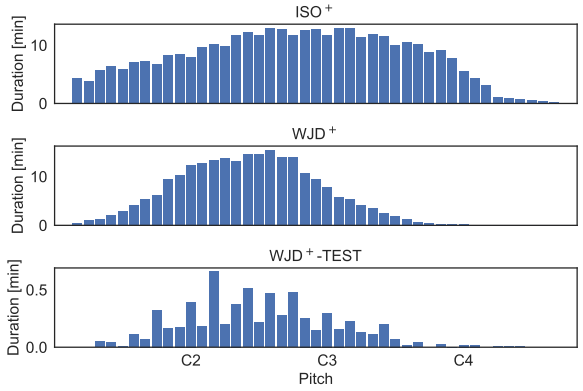

Figure 3. Pitch histogram over labeled datasets ISO$^+$, WJD$^+$, and WJD$^+$-TEST after data augmentation. Total duration of all notes in minutes is shown for each pitch.

Monroe's Upright Bass Sample Library , Meatbass , Trillian , Steinberg Halion Symphonic Orchestra, and SWAM Double Bass. Furthermore, we collected recordings from the Real World Computing Music Database (RWC) [14], the McGill University Master Samples [11] and the Iowa Classical Instrument Samples [1] .

3.2 Jazz Ensemble Recordings (WJD)

The Weimar Jazz Database (WJD) contains 456 manually transcribed solos from famous jazz recordings [22]. For a subset of 40 of these recordings, excerpts of walking bass lines using the Sonic Visualiser software [6]. All of the selected recordings are typical jazz ensembles that consist of upright bass, drums, piano, as well as melody instruments such as trumpet or saxophone. We use 30 of these annotated recordings for training (WJD$^+$) and 10 for testing (WJD$^+$-TEST). The remaining 416 recordings from the WJD are denoted by WJD$^-$. These recordings come without bass pitch annotations and will be used in the label propagation experiment detailed in Section 5.2.

3.3 Data Augmentation

On the datasets that we use for supervised training (ISO$^+$ and WJD$^+$), we generated 15 augmented versions from each original audio file by combining three time-stretching settings (stretch factors 0.9, 1, and 1.1) and five pitch-shifting settings (shifts between -2 and +2 semitones) using the software package sox [2] .

For all labeled datasets, we discard all non-voiced frames. Furthermore, we only keep the spectral frames from the first 75 % of the note duration as especially higher harmonics from upright bass tones decay much faster than the fundamental frequency contours. In order to make the final results comparable to [1], data augmentation is not applied to the test set WJD$^+$-TEST (compare Section 3).

Figure 3 illustrates the pitch distribution over the three labeled datasets after applying data augmentation. While the WJD$^+$ and WJD$^+$-TEST datasets similarly include notes up to C4, the isolated tones in ISO$^+$ cover a wider pitch range and distribute among the pitches in a more balanced fashion.

[1] http://theremin.music.uiowa.edu/MIS.html

[2] http://sox.sourceforge.net

Hyperparameter	Search Space	Importance
Magnitude scaling	{**linear**, logarithmic}	0.015
# hidden layers	$n \in \{\mathbf{3}, 4, 5, 6\}$	0.020
Hidden layer size	$H = 2^h, h \in \{7, \mathbf{8}, 9, 10\}$	0.040
Learning rate	$\alpha = 10^r, r \in [-3, -6]$, (**-4.27**)	**0.485**
Batch normalization	{**no**, yes}	0.017
ℓ_2 weight regularization	$\lambda \in \{0, 10^{-4}, \mathbf{10^{-3}}, 10^{-2}\}$	0.038
Dropout ratio	$d \in [0, 0.5]$, (**0.06**)	**0.385**

Table 2. Search space for hyperparameter optimization. Optimal parameter set for $a_{\text{val, opt}} = 0.62$ is given in bold font. Feature importance values in a random forest regression model are shown in the last column.

4. METHODOLOGY

4.1 Feature Extraction

Audio files are resampled to 22.05 kHz before constant-Q magnitude spectrograms are computed with a hopsize of 1024 samples (46.4 ms) and a frequency resolution of 12 bins per octave between 34.65 Hz (MIDI pitch 25, note D$^\flat$1) and 1567 Hz (MIDI pitch 91, note G6) using the librosa Python library [19]. Hence, the input vectors have the dimensionality of 67. In contrast to [1], we extend the frequency range by a small margin in the low frequency range and by two octaves in the upper frequency range in order to incorporate overtone frequencies of higher bass notes. In Section 5.1 we evaluate, to which extent a logarithmic compression of the magnitude spectrogram is beneficial for bass saliency.

4.2 Deep Neural Network Architecture

Throughout this paper, we use a fully-connected network architecture for the given task of bass saliency estimation, see Table 2 for an overview of parameters. The model is a cascade of n hidden layers of size H with optional intermediate layers for batch normalization (prior to the ReLU activation function) [16] and dropout (dropout ratio d) [26]. In contrast to [1], we do not use frame stacking here as we aim to directly compare local feature vectors later in the label propagation step described in Section 5.2. The model instead processes individual spectrogram frames as input and predicts the corresponding pitch saliency vector. Furthermore, all hidden layers have an optional ℓ_2 weight regularization (with regularization parameter λ) [12]. For each model training, we use 500 training epochs, a batch size of 250, early stopping with a patience of 25 epochs based on the validation accuracy, and the categorical cross-entropy as loss function. The keras [3] Python library is used for all experiments in this paper.

Score annotations are converted into frame-wise binary pitch activities, which are used as targets for the model training. In the annotated datasets used in this paper, bass lines are strictly monophonic. In the final layer, we use a

[3] https://www.keras.io

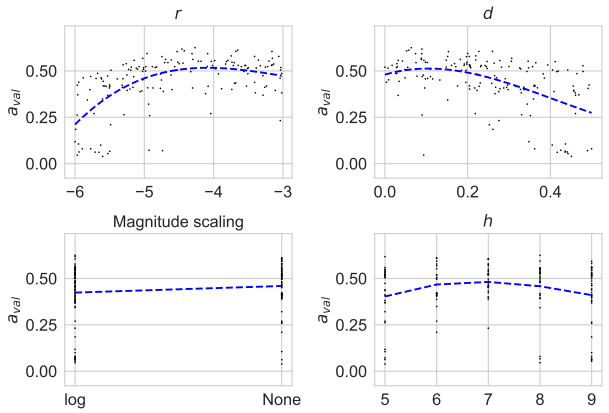

Figure 4. Validation accuracy a_{val} and hyperparameter values for learning rate exponent r, dropout ratio d, magnitude scaling, and layer size exponent h over all random parameter configurations (black dots). Cubic regression lines (blue dashed lines) show trends in the data. Optimal values for the other hyperparameters are given in Table 2.

sigmoid instead of a softmax activation function to be able to model the activity of all pitches independently. This also allows us to model polyphonic parts or rests within bass lines. However, in order to compare our results with [1], we only focus on bass saliency estimation from voiced frames in this paper and leave bass voicing detection open for future work.

As pitch range for the targets, we use $[26, 69]$ (notes D1 to G4) whereas in [1], a slightly smaller pitch range $[28, 67]$ (notes E1 to F4) was used. The dimensionality of the target vectors is 44.

4.3 Evaluation

We derive pitch estimates by looking at the highest output value of the final sigmoid layer. For the evaluation, we use the standard evaluation measures *Raw Pitch Accuracy* (denoted as a) and *Raw Chroma Accuracy* (denoted as a_{12}) as used in the MIREX *Audio Melody Extraction* task. For the definition of these measures, we refer to [20]. During training, we randomly split the training dataset(s) into training and validation dataset based on a 80:20 split. Accuracy values a_{train}, a_{val}, and a_{test} are computed on the training, validation, and test set, respectively.

5. EXPERIMENTS

5.1 Hyperparameter Optimization

A systematic grid search of possible hyperparameter combinations is not feasible for deep neural networks due to the high computational costs. In our approach, we train 160 models with different combinations of hyperparameters. These combinations are randomly sampled from the hyperparameters given in Table 2. The best hyperparameter combination is then retrieved by testing the model performance on the validation set (a_{val}). Figure 4 shows the validation set accuracy for the different hyperparameter

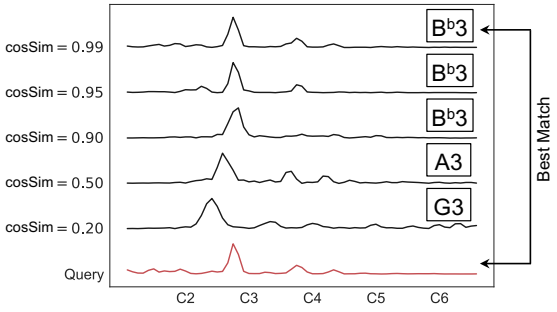

Figure 5. Label propagation example: given a query constant-Q spectrogram frame with unknown pitch (bottom, red), candidates with different cosSim similarity values are shown (above, black). The pitch label B♭3 will be transferred from the most similar candidate shown on top ($s = 0.99$).

configurations. To get an intuition about the influence of the different hyperparameters, we follow an approach previously presented in [17]. In that approach, a random forest regression model [5] is fitted to a_{val} over all parameter configurations. From the random forest regression model, we can obtain the relative importance of all hyperparameters, see Table 2 (third column) for the results.

As previously found in [17], the learning rate exponent r is by far the most important hyperparameter (0.485) with optimal values around 10^{-4}, as shown in Figure 4. Interestingly, as indicated in Table 2, the dropout ratio d also has a high relative importance (0.385) and an optimal value only slightly above zero.

5.2 Label Propagation

A first approach to enrich the available amount of training data is to use label propagation. We derive pitch labels for feature vectors in the unlabeled WJD$^-$ dataset by transferring labels from their most similar counterparts in WJD$^+$ dataset. To this end, we compute a similarity score s_i for the i-th feature vector in the WJD$^-$ database $x_i^{\text{WJD}^-} \in \mathbb{R}^{67}$ by maximizing its cosine similarity (cosSim) towards all feature vectors in the WJD$^+$ database as

$$s_i = \max_k \text{cosSim}(x_i^{\text{WJD}^-}, x_k^{\text{WJD}^+}). \quad (1)$$

An example is shown in Figure 5. Given a query spectrogram frames (bottom), we show five example spectrogram frames with different similarity values. The most similar frame ($\text{cosSim} = 0.99$) shows an almost identical overtone structure, which motivates the transfer of its pitch label.

As shown in Figure 6, most feature vectors in the unlabeled WJD$^-$ dataset have very similar counterparts in the WJD$^+$ dataset, which is somewhat intuitive as both datasets originate from the Weimar Jazz Database (WJD). We derive three similarity thresholds $\tau_{25} = 0.914$, $\tau_{50} = 0.940$, and $\tau_{75} = 0.96$ from the 25^{th}, 50^{th}, and 75^{th} percentile of the distribution over s. The label-enriched WJD$^-$ dataset is denoted as $(\text{WJD}^-)^+$ in the following.

We use the best model architecture obtained via hyperparameter optimization (see Section 5.1) and train

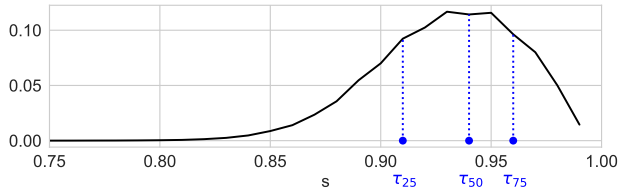

Figure 6. Histogram over best-match cosine-similarity values for mapping feature vectors from the unlabeled dataset WJD$^-$ to the labeled dataset WJD$^+$. Similarity thresholds τ_{25}, τ_{50}, and τ_{75} are derived from the respective percentiles of the distribution over s.

models from different training sets. For that purpose, we combine the full WJD$^+$ dataset with feature vectors of the (WJD$^-$)$^+$ dataset based on the criterion $\tau_- \leq s_i \leq \tau_+$. We test different pairs (τ_-, τ_+) using combinations of the percentile-based thresholds τ_{25}, τ_{50}, and τ_{75} as well as $\tau_0 = 0$ and $\tau_{100} = 1$ as shown in the lower subplot of Figure 7.

Feature vectors with lower similarity scores more likely introduce label noise to the mixed training dataset. Since the WJD$^+$ dataset only contains voiced frames, even unvoiced frames in the WJD$^-$ will be be mapped to voiced frames. This is a drawback due to the given dataset configurations. Another possible reason for low similarity scores are notes played by other instruments in the ensemble such as the piano or the soloist. However, voiced frames from the WJD$^-$ database with a lower similarity can provide novel information for the classification task, which can help to improve the existing model. At the same time, feature vectors with high similarity scores can be redundant without providing much novel information for the pitch saliency estimation task. In contrast to [1], label propagation is performed based on feature vector similarity and not based on predictions of existing models.

From the results shown in Figure 7, we make the following observations for all configurations. First, we observe that difference between a_{train} and a_{val} (overfitting) remains almost constant across different configurations. Also, the raw chroma accuracy $a_{\text{test,12}}$ is consistently about 0.07 higher than the raw pitch accuracy a_{test}, which indicates that octave errors make up only a small fraction of the remaining pitch estimation errors.

Using the WJD$^+$ dataset alone or combined with the most similar feature vectors in WJD$^-$ (configurations 0:0 and 75:100), we observe that the training and validation accuracies are clearly higher than the test accuracy. The reason is that these configurations correspond to the best parameter settings found in the hyperparameter optimization step (compare Section 5.1), where maximizing a_{val} was the main objective. Due to their small size, the data distribution in the WJD$^+$ and WJD$^+$-TEST datasets presumably is only similar to a certain degree although both are taken from the Weimar Jazz Database.

In contrast, by adding feature vectors from the WJD$^-$ dataset with lower similarity values and higher novelty (configurations 0:25, 0:50, and 0:75), the modeling task

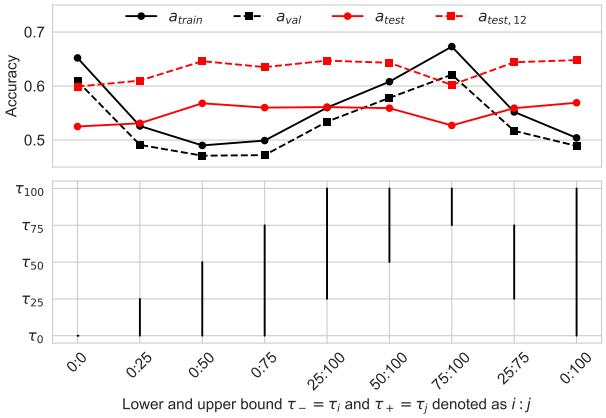

Figure 7. Label propagation results for different dataset configurations (see Section 5.2). Training accuracy a_{train}, validation set accuracy a_{val}, test set accuracy a_{test}, and chroma pitch accuracy $a_{\text{test,12}}$ are shown.

becomes harder and the training and validation accuracies decrease. Interestingly, the models' ability to generalize to the test set improves and a_{test} increases. The relatively high difference between validation and test accuracy of up to 0.09 indicates that the small test set size needs to be increased in future work, as both, test and validation set, should come from the same distribution.

For the configurations 0:50 and 0:100, we observe the highest test accuracy of around $a_{\text{test}} = 0.57$. This result is notable as by using label propagation, we are able to train a model which achieves a performance comparable to the highest test accuracy reported in [1] without requiring additional temporal context information using frame stacking. Therefore, label propagation seems a promising approach to improve the model performance.

5.3 Transfer Learning

State-of-the-art music transcription algorithms based on spectral decomposition algorithms such as Non-Negative Matrix Factorization (NMF) are commonly initialized with isolated instrument tones, e. g., for learning spectral note templates [3]. We aim to investigate to which extent a similar strategy can be used to improve neural networks for pitch saliency estimation tasks. As an alternative, we want to find out if it is instead better to train the networks solely on more complex instrumental mixtures (ensemble recordings, see Section 3.2) as these are more similar to the final test data.

We compare three training scenarios in our experiment. First, we train the model solely using the isolated bass tones (ISO$^+$ dataset) to evaluate the generalization potential of the trained model towards mixture signals in the test set. Secondly, we apply transfer learning, i. e., we pre-train an initial model for bass saliency estimation using isolated bass tracks (ISO$^+$ dataset) and then fine-tune the model in a second training step on the WJD$^+$ dataset. In a third scenario, we mix and shuffle the WJD$^+$ and ISO$^+$ datasets and perform a single training step. The model trained only

on the WJD$^+$ dataset serves as baseline for comparison. We train for 750 epochs for both training steps but apply early stopping as detailed in Section 4.2 if possible.

The results are shown in Table 3. Accuracy values are computed on a macro-level by averaging across all spectrogram frames of the test set files. We observe that a pitch saliency model, which is only trained on isolated tones of the target instrument, is not capable to generalize to more complex mixtures as it performs poorly on the test set ($a_{\text{test}} = 0.095$). Combining pre-training on the isolated note database with fine-tuning on the mixture dataset improves the performance by around four percent on the test set accuracy ($a_{\text{test}} = 0.542$) compared to a baseline model, which is only trained on the mixture recordings. The best configuration improves on the test set accuracy by 6 percent ($a_{\text{test}} = 0.561$) compared to the baseline model. It does not involve a pre-training step but uses a mix of isolated and mixed recordings (ISO$^+$ and WJD$^+$) for training instead.

The results of the transfer learning experiments suggest that combining training data with different levels of complexity, i. e., different amount of instrumental overlap, can be useful to improve DNN-based models for pitch saliency estimation in ensemble recordings. By using a mixture of both isolated and mixed recordings in one training step, it appears as if the neural network learns best to "focus" on the targeted instrument. Future work could address a different order in the training process, i. e., first training on the mixture tracks and then fine-tuning the model on the isolated note recordings.

6. CONCLUSION

We investigated strategies for label propagation and transfer learning in order to improve bass saliency estimation using deep neural networks. We could show that unlabeled feature vectors from datasets with a similar spectral distribution as the target scenario can be mapped towards labeled datasets to derive pitch labels. By combining labeled datasets and unlabeled datasets through label propagation, we were able to improve the model's accuracy by around six percent compared to a baseline model. Similarly, we could show that by combining isolated note recordings of the targeted instrument with mixture recordings as training set, we gain around five percent in accuracy. This joint training slightly outperformed our considered transfer learning strategy with two successive training steps. Future work could deal with strategies on how to combine frame-stacking (compare [1]), label propagation, and transfer learning.

For a systematic bias–variance analysis of the given modeling task, it remains challenging to define human level performance as we only focus on frame-wise pitch estimation here. Human experts, i. e., musicians or musicologists, are capable to generate near-perfect note-wise transcriptions. This related task, however, involves listening to longer audio excerpts and allows to include additional cues from the metric structure, tone duration, and local harmony.

Pre-Training	**Training**	a_{train}	a_{val}	a_{test}	$a_{\text{test,12}}$
-	WJD$^+$ (baseline)	0.665	0.614	0.508	0.589
-	ISO$^+$	0.514	0.538	0.095	0.234
ISO$^+$	WJD$^+$	0.652	0.603	0.542	0.614
-	ISO$^+$ & WJD$^+$	0.507	0.531	**0.561**	**0.655**

Table 3. Performance comparison with and without transfer learning. All experiments were evaluated using the WJD$^+$-TEST dataset.

7. ACKNOWLEDGEMENTS

This work has been supported by the German Research Foundation (MU 2686/11-1, AB 675/2-1). The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institut für Integrierte Schaltungen IIS.

8. REFERENCES

[1] 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 202–209, Erlangen, Germany, 2017.

[2] 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.

[3] Emmanouil Benetos and Tillman Weyde. An efficient temporally-constrained probabilistic model for multiple-instrument music transcription. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 701–707, Málaga, Spain, 2015.

[4] 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)*, pages 63–70, Suzhou, China, 2017.

[5] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[6] C. Cannam, C. Landone, and M. Sandler. Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files. In *Proceedings of the ACM Multimedia 2010 International Conference*, pages 1467–1468, Firenze, Italy, October 2010.

[7] Keunwoo Choi, György Fazekas, Kyunghyun Cho, and Mark Sandler. A Tutorial on Deep Learning for Music Information Retrieval. *ArXiv e-prints*, September 2017.

[8] Keunwoo Choi, György Fazekas, Mark B. Sandler, and Kyunghyun Cho. Transfer learning for music classification and regression tasks. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 141–149, Suzhou, China, 2017.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, Miami, FL, USA, 2009.

[10] Aleksandr Diment and Tuomas Virtanen. Transfer learning of weakly labelled audio. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 6–10, New Paltz, NY, USA, 2017.

[11] Tuomas Eerola and Rafael Ferror. Instrument library (MUMS) revised. *Music Perception*, 25(3):253–255, 2008.

[12] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[13] Masataka Goto. A real-time music-scene-description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Communication (ISCA Journal)*, 43(4):311–329, 2004.

[14] 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.

[15] Curtis Hawthorne, Erich Elsen, Jialin Song, Adam Roberts, Ian Simon, Colin Raffel, Jesse Engel, Sageev Oore, and Douglas Eck. Onsets and Frames: Dual-Objective Piano Transcription. *ArXiv e-prints*, October 2017.

[16] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the International Conference on International Conference on Machine Learning (ICML)*, pages 448–456, 2015.

[17] Rainer Kelz, Matthias Dorfer, Filip Korzeniowski, Sebastian Böck, 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, USA, 2016.

[18] Anurag Kumar, Maksim Khadkevich, and Christian Fügen. Knowledge Transfer from Weakly Labeled Audio using Convolutional Neural Network for Sound Events and Scenes. *ArXiv e-prints*, November 2017.

[19] Brian McFee, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the Scientific Computing with Python conference (Scipy)*, Austin, Texas, 2015.

[20] MIREX. Audio melody extraction task. Website `http://www.music-ir.org/mirex/wiki/2016:Audio_Melody_Extraction`, last accessed 03/29/2018, 2016.

[21] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[22] Martin Pfleiderer, Klaus Frieler, Jakob Jakob Abeßer, Wolf-Georg Zaddach, and Benjamin Burkhart. *Inside the Jazzomat. New perspectives for jazz research*. Schott Campus, Mainz, Germany, 2017.

[23] Jordi Pons, Oriol Nieto, Matthew Prockup, Erik M. Schmidt, Andreas F. Ehmann, and Xavier Serra. End-to-end learning for music audio tagging at scale. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, pages 1–5, Long Beach, CA, USA, 2015.

[24] Matti Ryynänen and Anssi P. Klapuri. Automatic transcription of melody, bass line, and chords in polyphonic music. *Computer Music Journal*, 32(3):72–86, 2008.

[25] Justin Salamon and Emilia Gómez. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6):1759–1770, 2012.

[26] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

[27] 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.