Transfer Learning for Violinist Identification

Yudong Zhao, György Fazekas, Mark Sandler
Centre for Digital Music, Queen Mary University of London, UK
Email: {yudong.zhao; g.fazekas; mark.sandler}@qmul.ac.uk

***Abstract*—Music performer identification is important for music recommendation, music expression analysis and playlist generation. In previous research, audio feature learning methods were commonly used for both singer identification (SID) and instrument player identification (IPID) with good results. In the current deep learning era, SID results are greatly improved using neural networks, however, instrument player identification is rarely investigated in recent works primarily due to the shortage of open-access datasets. To solve this problem, we construct a concerto violin dataset as well as a solo dataset, and present a transfer learning approach for violinist identification from pre-trained music auto-tagging neural networks and singer identification models. We then transfer pre-trained weights and fine-tune the models using violin datasets and finally obtain violinist identification results. We compare our system with a number of state-of-the-art methods and show that our model outperforms them using both of our datasets.**

***Index Terms*—violinist identification, transfer learning**

I. Introduction

Expressive performance of music is mostly determined by musical structures and surfaces, as well as performer's individual interpretations. For a specific music piece, the diversity of expression mainly depends on performers' characteristic playing styles. Therefore, modelling performers' individual styles and identifying performers are important for music education, music expression analysis and music synthesis. Furthermore, performer identification systems can be used in commercial application scenarios, such as recommending similar performers to users, managing large amounts of unlabelled songs, and generating playlists in different styles.

In previous research, musical performer identification was mainly considered in two areas: singer identification (SID) and instrument player identification (IPID). As for singer identification, there are several studies using audio feature learning [26] and deep neural networks (DNN) [10], [18], [27]. The works using DNNs reported outstanding performance in this task (with 0.99 f1-score in [27] based on the artist20 dataset [7]). There are also some existing works focussing on instrument player identification, such as pianist identification [22] and violinist identification [21]. Although many attempts have been reported in instrument player identification, the approaches and results are often incremental, and the IPID problem remains challenging and not fully solved.

There are several possible reasons for this. First, unlike singing voice produced by singers' vocal cords directly, instrument players present their characteristic styles via musical instruments. It means that music performers' individual styles not only depend the performers themselves, but also on the instrument, which makes the analysis and detection more challenging. Secondly, although data driven methods, such as DNN, are very powerful in other areas, there are not many published large datasets for training DNNs for IPID. Existing music datasets like MagnaTagATune (MTAT) [16], the Million Song Dataset (MSD) [1], and Jamendo [2] are mostly designed for music tagging or classification. Performer information is not contained in the data. Moreover, training on small-size data tends to overfit to the training set, which typically leads to poor generalization of performance.

In order to solve the problem of insufficient datasets for training DNNs, the idea of transfer learning has been increasingly applied in recent years. Small datasets can be used to train neural networks by transferring pre-trained weights, and achieve reasonable performance in several MIR tasks [5], [17]. Cramer [6] also found that pre-training a model on a large amount of data resulted in models that could be fine-tuned to downstream tasks with little data. Although there are only few open access datasets for IPID, datasets and pre-trained DNNs for music tagging and singer identification are widely published in the MIR community. Music auto-tagging [4], [25] can be considered a combination of multiple tasks such as genre classification, emotion recognition and instrument identification, which contributes to learning the relation between tags and audio content. Additionally, research concerning singer identification proposed methods for recognising individual styles of different singers from audio recordings, which is similar to the IPID. Since transfer learning can be used for different music classification and regression tasks [5], we hypothesise that effective trained-models based on music tagging and singer identification are helpful for identifying instrument players.

In this paper, we propose a case study for violinist identification using transfer learning, which is based on pre-trained music tagging and singer identification neural networks. To best of our knowledge, this is the first work to identify violinists using transfer learning from DNNs, which are pretrained for other MIR tasks. We first construct two violin datasets from solo musical scale recordings played by 22 performers and commercial concerto recordings performed by 9 master players separately. Details of these datasets are introduced in Section II. In the next step, we choose six neural networks for music tagging and three neural networks for singer identification, then train them using a corresponding dataset and obtain pre-trained weights. We further retrain the selected models on our two datasets separately, and use pre-trained weights during initialization. This transfer learning

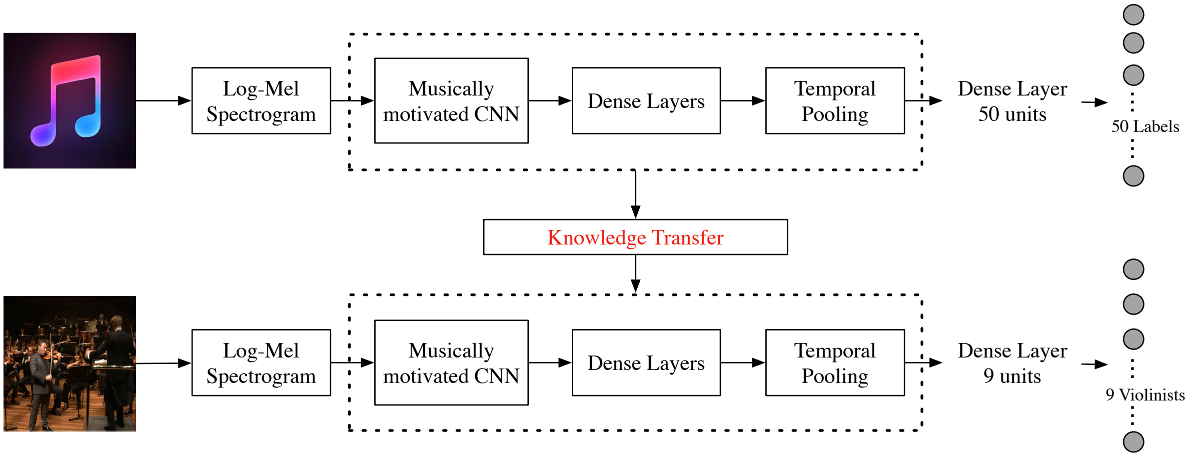

Fig. 1. Transfer learning process using pre-trained Musicnn model on concerto dataset.

process is proposed in Section III. Finally, the results obtained from different pre-trained models and datasets are compared and discussed in Section IV.

II. Datasets

A. Solo violin dataset

To assess the performance of our proposed method on solo music, and model the characteristic styles of professional players, we first label a solo musical scale violin dataset. During the European Bilbao project[1] [8], thirteen new (white) violins were designed and built and then evaluated within a free categorisation task by 22 professional violinists. At the end of the task, they were invited to play a scale on each of these violins. The recordings were all made under the same conditions keeping the position of the player and the microphone constant, in a large rehearsal room at the Bilbao conservatory. We selected a group of 22 players for our dataset, which thus consists of 22 × 13 musical scales in total. Each scale contains around 37 notes.

B. Concerto dataset

To investigate expressive performances of master players, and test our proposed method in more complicated scenarios, we created a dataset of violin concerto pieces. We selected five concertos including Beethoven Op.61, Brahms Op.77, Mendelssohn Op.64, Sibelius Op.47 and Tchaikovsky Op.35. These pieces have all been performed by nine violinists: Jascha Heifetz, Anne Sophie Mutter, David Oistrakh, Itzhak Perlman, Pinchas Zukerman, Isaac Stern, Salvatore Accardo, Yehudi Menuhin and Maxim Vengerov, who are all leading master violinists. We excluded performances in the 'prelude' or 'interlude', sections that are performed by orchestra alone. Additionally, there are segments where the accompaniment is too loud to hear the violin performance. To minimise the impact of accompaniment, we manually remove the parts of the music without violin or where the violin cannot be heard clearly. The remaining music pieces constitute our concerto dataset, which contains approximately two hours of audio for each performer.

III. Methodology

In this section, our proposed transfer learning approach is described. We first train seven music tagging models using three datasets: MSD, MTAT and Jamendo (see Section I for introductions). Next, we select three singer identification neural networks and train them using the artist20 dataset. These neural networks are trained as source task, details of pretrained models are introduced in Section III-A. Then, the networks with trained weights are used as a feature extractor for violinist identification. We then modify the model architecture and fine-tune the models on the datasets mentioned in Section II. The details of target task are presented in Section III-B.

A. Source tasks

1) Music auto tagging: We select seven music tagging models as source task, including a fully convolutional network (FCN) [3], short-chunk CNN with Residual connections [25], Sample-level CNN [14], Musicnn [19], Harmonic CNN [23], Convolutional Recurrent Neural Network (CRNN) [4], and self-attention-based CRNN (self-attention-CRNN) [24].

The FCN consists of 4 convolutional layers and 4 max-pooling layers. It takes a log-amplitude mel-spectrogram as input and predicts a 50 dimensional tag vector [3]. Similarly, another FCN with 7-layer CNN with a fully-connected layer and its extension with residual connections is validated in [25], which shows outstanding performance. Sample-level CNN [14] is an end-to-end model that takes raw audio waveforms as its inputs. It consists of ten 1D convolutional layers with 1 × 3 filters and 1 × 3 max-poolings, and simpler and deeper than Mel spectrogram-based approaches [25]. Since a variation of Sample-level CNN with squeeze-and-excitation (SE) [11] blocks performs better than the original one, we use this model in our paper. Musicnn [19] is different from previously proposed models although it also uses Mel spectrograms

[1] https://www.bele.es/en/bilbao-project-introduction

TABLE I
THE DETAILS OF SOURCE TASKS IN OUR TRANSFER LEARNING EXPERIMENT.

Task	Dataset	Models	Input Length	Input Feature	Classes
Music Tagging	MSD/MTAT/Jamendo	FCN	29.1s	Mel-spectrogram	50
		Musicnn	3s	Mel-spectrogram	50
		Harmonic CNN	5s	Stacked harmonic tensor	50
		Sample-level CNN	3.69s	Raw Waveform	50
		Short-chunk CNN	3.69s	Mel-spectrogram	50
		CRNN	29.1s	Mel-spectrogram	50
		CRNN-self-attention	15s	Mel-spectrogram	50
Singer Identification	Artist20	CRNNM	5s	Mel-spectrogram & Melody contour	20
		CRNN-attention	5s	Mel-spectrogram	20
		CRNN-attention-KNN	5s	Mel-spectrogram	20

as input. It is designed to rely on music domain knowledge. Harmonic CNN [23] takes advantage of trainable band-pass filters and harmonically stacked time-frequency representation inputs. The number of frequency bands is set to 128 and the number of harmonics is six.

Due to CRNN being widely used for music auto tagging [4], we consider CRNN and self-attention-CRNN as source task models as well. CRNN is a combination of CNNs and RNNs, where the CNN front-end extracts local features and the RNN back-end summarises them temporally. The architecture of self-attention-CRNN is similar to CRNN, The only difference is that the self-attention mechanism is used instead of the RNNs as a temporal summarisation back-end [24]. The inputs of these two CRNN based models are Mel spectrograms.

We train these models using the MSD, MTAT and Jamendo datasets separately. All Mel-spectrogram based approaches use 512-point FFT with a 50% overlap, and number of frequency bins are all set as 128. For pre-training and the input length of audio for each model, see Table I and the original paper [25]. While training the models, we used a optimization method that combine scheduled ADAM [15] and stochastic gradient descent (SGD) [12], which is also proposed in [24].

2) Singer Identification: Since the CRNN based models have recently been used for singer identification and present good results [10], [18], [27], we select three CRNN based models trained for SID as source tasks.

The first model is CRNNM [10] which extends the original CRNN model [18] for SID. The input features of CRNNM are mel-spectrogram and melody contour, with the melody contour extracted using CREPE [13]. Another two models, Attention-CRNN and Attention-CRNN-KNN are proposed in [27], which perform best in existing SID works using the artist20 dataset.

We therefore train CRNNM, Attention-CRNN and Attention-CRNN-KNN using the artist20 dataset as source tasks, following exactly the same training setup (i.e., data split method, filters numbers, kernel sizes,optimizer, learning rate, activation functions, loss function, etc) described in the original works. To make the results comparable, we split the dataset at the album-level, and the input length of audio for each model is set as 5s.

B. Target tasks

After training all models in the source task on source datasets, the pre-trained networks are used as feature extractors and transferred to the target datasets introduced in Section II. In order to adapt the violin dataset, we change the final dense layer of each pre-trained model, which outputs probabilities of violinists instead of original labels or singers. We retrain these models separately using weights from each pre-trained model during initialisation, and select the best model based on validation loss. Finally, the violinist classification results obtained for each model are compared.

The transfer learning process using a pre-trained Musicnn model to identify master violinists on concerto dataset is shown in Figure 1 as an example. We transfer the learned knowledge from pre-trained music tagging network, then modify the output layer and fine-tune the model using concerto data to obtain violinist identification result.

IV. EXPERIMENTS AND RESULTS

A. Dataset preparation

To adapt the requirement of different input length for each model, we segment the audio for each performer in lengths of 29.1s, 15s, 5s, 3.69s and 3s separately without overlaps after re-sampling the audio using $F_s = 16000$Hz. For all audio segments of each performer, we randomly shuffle them into training, validation and test sets using a ratio of 6:2:2.

B. Estimation metric and baseline method

To evaluate and compare the performances of the proposed models for violinist identification, *accuracy* is used as an evaluation metric. To validate the effectiveness of our proposed method, we consider "violinist identification methods using timbre feature distribution [29] and onset time deviation [28]" as baseline methods. The former method had an accuracy of 0.94 using the solo dataset and 0.35 using the concerto dataset; the latter method had an accuracy of 0.741 using the concerto dataset. Similar statistical approaches to IPID have also been validated in the context of piano music [20].

TABLE II
VIOLINIST IDENTIFICATION RESULTS USING TWO DATASETS.

Models	Concerto dataset					Solo dataset				
	Scratch	MTAT	MSD	Jamendo	Artist20	Scratch	MTAT	MSD	Jamendo	Artist20
FCN	0.950	0.958	0.969	0.981	–	0.850	0.855	0.873	0.874	–
Musicnn	0.905	0.913	0.932	0.907	–	0.960	0.971	0.980	0.962	–
Harmonic CNN	0.955	0.964	0.973	0.962	–	0.981	**0.988**	**0.988**	0.984	–
Sample-level CNN	0.908	0.934	0.956	0.924	–	0.786	0.913	0.953	0.925	–
Short-chunk CNN	0.976	0.978	0.978	**0.991**	–	0.953	0.965	0.972	0.981	–
CRNN	0.548	0.927	0.789	0.625	–	0.513	0.648	0.564	0.611	–
CRNN-self-attention	0.937	0.977	0.942	0.976	–	0.978	0.926	0.811	0.884	–
CRNNM	0.479	–	–	–	0.492	0.546	–	–	–	0.539
CRNN-attention	0.755	–	–	–	0.809	0.825	–	–	–	0.830
CRNN-attention-KNN	0.745	–	–	–	0.776	0.793	–	–	–	0.822

C. Results

Table II summarises the results obtained by our proposed method, with the test accuracy based on the concerto dataset on the left hand side and the results based on the solo dataset on the right hand side. To compare the differences in results with and without transfer learning, we first show the results trained from scratch (using random initialisation) for each model, which corresponds to the "Scratch" column in Table II. The evaluation of violinist identification based on different source datasets and pre-trained models are then shown separately.

It can be seen in the table that the transferred knowledge is very useful for improving violinist identification performance. Short-chunk CNN, and Harmonic CNN showed the best results for both target datasets, no matter which source datasets are used for pre-training. The best accuracy on concerto dataset is obtained by the Short-chunk CNN pre-trained on the Jamendo dataset, which is 0.991; the best solo violinists identification performance is 0.988, which is obtained by Harmonic CNN pre-trained on the MSD dataset.

For the CRNN models, self-attention mechanisms can improve the performance no matter which pre-trained model are used. However, the results obtained from pre-trained SID networks are generally inferior. One possible reason is that the characteristic features of singers are not directly transferable to identify the results obtained from pre-trained SID networks.

V. Discussion and conclusion

In this paper, we first constructed two violin datasets using concerto collections and solo recordings separately. Pre-trained models were then obtained using the source tasks of music auto-tagging and singer identification. Results show that pre-trained models can be successfully adapted to the target task, and outperforms the current baseline methods to achieve high performance on both datasets.

In general, the pre-trained music tagging networks perform better. We suspect that this is because the former was pre-trained on datasets that contain a broader set of musical styles and types of music in general, and the networks were designed to facilitate the output a set of 50 broad music labels belonging to different categories. Therefore those models learned more detailed musical features, which may include feature spaces suitable for characterising violinist's style. In contrast, the source dataset of latter task (artist20) contains human voices, and the corresponding models were designed to find stylistic features of vocal performances, which is somewhat different from our target task.

Among the results of the pre-trained music tagging networks, models trained on shorter music clips (short block CNN, sample-level CNN, harmonic CNN and Musicnn) outperformed models trained on longer music clips (FCN, CRNN). Intuitively, when the models are trained on short examples, there are a larger number of examples during the training process and it is very likely that the performer's style can be identified within a few seconds, which brings good performance to these models.

In the future, we will investigate the proposed approaches using larger datasets to obtain more robust results. We may also apply source separation to isolate the violin performance, assuming the input feature would reflect the performers' individual style better. Regarding improved explainability of differences in model performance, we may perform an ablation study or use a teacher-student learning framework [9] to transfer knowledge from networks trained on different tasks.

References

[1] Thierry Bertin-Mahieux, Daniel Ellis, Brian Whitman, and Paul Lamere. The million song dataset. 2011.

[2] Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. The mtg-jamendo dataset for automatic music tagging. 2019.

[3] Keunwoo Choi, György Fazekas, and Mark Sandler. Automatic tagging using deep convolutional neural networks. *Proc. of the 17th International Society for Music Information Retrieval (ISMIR) conference, August 7-11., New York, USA*, 2016.

[4] Keunwoo Choi, György Fazekas, Mark Sandler, and Kyunghyun Cho. Convolutional recurrent neural networks for music classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2392–2396. IEEE, 2017.

[5] Keunwoo Choi, György Fazekas, Mark Sandler, and Kyunghyun Cho. Transfer learning for music classification and regression tasks. *Proc. 18th International Society for Music Information Retrieval Conference (ISMIR), Oct. 23-27, Suzhou, China*, 2017.

[6] Jason Cramer, Ho-Hsiang Wu, Justin Salamon, and Juan Pablo Bello. Look, listen, and learn more: Design choices for deep audio embeddings. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3852–3856. IEEE, 2019.

[7] Daniel PW Ellis. Classifying music audio with timbral and chroma features. 2007.

[8] Claudia Fritz, George Stoppani, Igartua Unai, Roberto Jardón Rico, Arroitajauregi Ander, and Luis Artola. The bilbao project: How violin makers match backs and tops to produce particular sorts of violins. In *International Symposium on Musical Acoustics*, 2019.

[9] Geoffrey Hinton, Orio Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning Workshop*, 2015.

[10] Tsung-Han Hsieh, Kai-Hsiang Cheng, Zhe-Cheng Fan, Yu-Ching Yang, and Yi-Hsuan Yang. Addressing the confounds of accompaniments in singer identification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2020.

[11] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[12] Nitish Shirish Keskar and Richard Socher. Improving generalization performance by switching from adam to sgd. *arXiv:1712.07628*, 2017.

[13] Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello. Crepe: A convolutional representation for pitch estimation. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 161–165. IEEE, 2018.

[14] Taejun Kim, Jongpil Lee, and Juhan Nam. Sample-level cnn architectures for music auto-tagging using raw waveforms. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 366–370. IEEE, 2018.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[16] Edith Law, Kris West, Michael I Mandel, Mert Bay, and J Stephen Downie. Evaluation of algorithms using games: The case of music tagging. In *ISMIR*, pages 387–392. Citeseer, 2009.

[17] Beici Liang, Fazekas G., and Sandler M. Transfer learning for piano sustain-pedal detection. In *Proc. International Joint Conf. on Neural Networks (IJCNN), July 14-19, Budapest, Hungary*. IEEE, 2019.

[18] Zain Nasrullah and Yue Zhao. Music artist classification with convolutional recurrent neural networks. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

[19] Jordi Pons and Xavier Serra. musicnn: Pre-trained convolutional neural networks for music audio tagging. *arXiv:1909.06654*, 2019.

[20] Syed Rifat Mahmud Rafee, G. Fazekas, and G. A. Wiggins. Performer identification from symbolic representation of music using statistical models. In *International Computer Music Conference (ICMC), Santiago, Chile, July 25-31*, 2021.

[21] Rafael Ramirez, Esteban Maestre, Alfonso Perez, and Xavier Serra. Automatic performer identification in celtic violin audio recordings. *Journal of New Music Research*, 40(2):165–174, 2011.

[22] Efstathios Stamatatos and Gerhard Widmer. Automatic identification of music performers with learning ensembles. *Artificial Intelligence*, (1):37–56, 2005.

[23] Minz Won, Sanghyuk Chun, Oriol Nieto, and Xavier Serrc. Data-driven harmonic filters for audio representation learning. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 536–540. IEEE, 2020.

[24] Minz Won, Sanghyuk Chun, and Xavier Serra. Toward interpretable music tagging with self-attention. *arXiv:1906.04972*, 2019.

[25] Minz Won, Andres Ferraro, Dmitry Bogdanov, and Xavier Serra. Evaluation of cnn-based automatic music tagging models. *arXiv arXiv:2006.00751*, 2020.

[26] Tong Zhang. Automatic singer identification. In *2003 International Conference on Multimedia and Expo. ICME'03. Proceedings (Cat. No. 03TH8698)*, volume 1, pages I–33. IEEE, 2003.

[27] Xulong Zhang, Jiale Qian, Yi Yu, Yifu Sun, and Wei Li. Singer identification using deep timbre feature learning with knn-net. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3380–3384. IEEE, 2021.

[28] Yudong Zhao, György Fazekas, and Mark Sandler. Identifying master violinists using note-level audio features. In *17th Sound and Music Computing Conference*, 2020.

[29] Yudong Zhao, Changhong Wang, György Fazekas, Emmanouil Benetos, and Mark Sandler. Violinist identification based on vibrato features. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 381–385. IEEE, 2021.