JAZZ SOLO INSTRUMENT CLASSIFICATION WITH CONVOLUTIONAL NEURAL NETWORKS, SOURCE SEPARATION, AND TRANSFER LEARNING

Juan S. Gómez **Jakob Abeßer** **Estefanía Cano**
Semantic Music Technologies Group, Fraunhofer IDMT, Ilmenau, Germany
{gomejn,abr,cano}@idmt.fhg.de

ABSTRACT

Predominant instrument recognition in ensemble recordings remains a challenging task, particularly if closely-related instruments such as alto and tenor saxophone need to be distinguished. In this paper, we build upon a recently-proposed instrument recognition algorithm based on a hybrid deep neural network: a combination of convolutional and fully connected layers for learning characteristic spectral-temporal patterns. We systematically evaluate harmonic/percussive and solo/accompaniment source separation algorithms as pre-processing steps to reduce the overlap among multiple instruments prior to the instrument recognition step. For the particular use-case of solo instrument recognition in jazz ensemble recordings, we further apply transfer learning techniques to fine-tune a previously trained instrument recognition model for classifying six jazz solo instruments. Our results indicate that both source separation as pre-processing step as well as transfer learning clearly improve recognition performance, especially for smaller subsets of highly similar instruments.

1. INTRODUCTION

Automatic Instrument Recognition (AIR) is a fundamental task in Music Information Retrieval (MIR) which aims at identifying all participating music instruments in a given recording. This information is valuable for a variety of tasks such as automatic music transcription, source separation, music similarity computation, and music recommendation, among others. In general, musical instruments can be categorized based on their underlying sound production mechanisms. However, various aspects of human music performance such as dynamics, intonation, or vibrato create a large timbral variety that complicate the distinction of closely-related instruments such as a violin and a cello.

As part of the ISAD (Informed Sound Activity Detection in Music Recordings) research project, we aim at improving existing methods for timbre description and instrument classification in ensemble music recordings. In particular, this paper focuses on the identification of predominant solo instruments in multitimbral music recordings, i. e., the most salient instruments in the audio mixture. This assumes that the spectral-temporal envelopes that describe the instrument's timbre are dominant in the polyphonic mixture [11]. As a particular use-case, we focus on the classification of solo instruments in jazz ensemble recordings. Here, we study the task of instrument recognition both on a class and sub-class level, e. g. between soprano, alto, and tenor saxophone. Besides the high timbral similarity between different saxophone types, a second challenge lies in the large variety of recording conditions that heavily influence the overall sound of a recording [21, 25]. A system for jazz solo instrument classification could be used for content-based metadata clean-up and enrichment of jazz archives.

As the main contributions of this paper, we systematically evaluate two state-of-the-art source separation algorithms as pre-processing steps to improve instrument recognition (see Section 3). We extend and improve upon a recently proposed hybrid neural network architecture (see Figure 1) that combines convolutional layers for automatic learning of spectral-temporal timbre features, and fully connected layers for classification [28]. We further evaluate transfer learning strategies to adapt a given neural network model to more specific classification use-cases such as jazz solo instrument classification, which require a more granular level of detail [13].

2. RELATED WORK

The majority of work towards automatic instrument recognition has focused on instrument classification of isolated note events or monophonic phrases and melodies played by single instruments. Considering classification scenarios with more than 10 instrument classes, the best-performing systems achieve recognition rates above 90%, as shown for instance in [14, 27].

In polyphonic and multitimbral music recordings, however, AIR is a more complicated problem. Traditional approaches rely on hand-crafted audio features designed to capture the most discriminative aspects of instrument timbres. Such features are based on different signal representations based on cepstrum [8–10, 29], group delay [5], or line spectral frequencies [18]. A classifier ensemble focus-

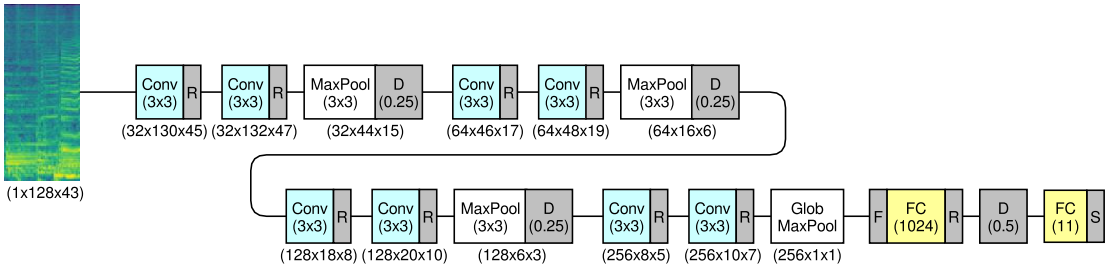

Figure 1. Reference model proposed by Han et al. [28]. Time-frequency spectrogram patches are processed by successive pairs of convolutional layers (Conv) with ReLU activation function (R), max pooling (MaxPool), and global max pooling (GlobMaxPool). Dropout (D) is applied for regularization in the feature extractor and classifier. Conv layers have increasing number of filters (32, 64, 128, and 256) and output shapes are specified for each layer.

ing on note-wise, frame-wise, and envelope-wise features was proposed in [14]. We refer the reader to [11] for an extensive overview of AIR algorithms that include hand-crafted audio features.

Novel deep learning algorithms, particularly convolutional neural networks (CNN), have been widely used for various image recognition tasks [13]. As a consequence, these methods were successfully adopted to MIR tasks such as chord recognition [17] and music transcription [1], where they significantly improved upon previous state-of-the-art results. Similarly, the first successful AIR methods based on deep learning were recently proposed and designed from the combination of convolutional layers for feature learning, and fully-connected layers for classification [24, 28]. Park et al. use a CNN to recognize instruments using single tone recordings [24]. Han et al. [28] propose a similar architecture and evaluate different late-fusion results to obtain clip-wise instrument labels. The authors aim at classifying predominant instruments in polyphonic and multitimbral recordings, and improve upon previous state-of-the-art systems by around 0.1 in f-score. Li et al. [20] propose to use end-to-end learning, considering a different network architecture. By these means, they use raw audio data as input without relying on spectral transformations such as mel spectrograms.

A variety of pre-processing strategies have been been applied MIR tasks such as singing voice detection [19] and melody line estimation [26]. Regarding the AIR task, several algorithms include a preceding source separation step. In [2], Bosch et al. evaluate two segregation methods for stereo recordings—a simple LRMS (Left/Right-Mid/Side) separation and FASST (Flexible Audio Source Separation Framework) developed by Ozerov et al. [22]. The authors report improvements of 19% in f-score using a simple panning separation, and up to 32% when the model was trained with previously separated audio, taking into account the typical artifacts produced by source separation techniques. Heittola et al. [16] propose a system that uses a source-filter model for source separation in a non-negative matrix factorization (NMF) scheme. The spectral basis functions are constrained to have harmonic spectra with smooth frequency responses. Using a Gaussian mixture model, the authors achieved a 59% recognition rate for six polyphonic notes randomly chosen from 19 different instruments.

3. PROCESSING STEPS

3.1 Baseline Instrument Recognition Framework

In this section, we briefly summarize the instrument recognition model proposed by Han et al. [28], which we use as the starting point for our experiments. As a first step, monaural audio signals are processed at a sampling rate of 22.05 kHz. A mel spectrogram with a window size of 1024, a hop size of 512, and 128 mel bands is then computed. After applying a logarithmic magnitude compression, spectral patches one second long are used as input to the deep neural network. The resulting time-frequency patches have shape $x_i \in \mathbb{R}^{128\times 43}$.

The network architecture is illustrated in Figure 1 and consists of four pairs of convolutional layers with a filter size of 3×3 and ReLU activation functions. The input of each convolution layer is zero-padded with 1×1, considered in the output shape of each layer. The number of filters in the conv layer pairs increases from 32 to 256. Max pooling over both time and frequency is performed between successive layer pairs. Dropout of 0.25 is used for regularization. An intermediate global max pooling layer and flatten layer (F) connect the feature extractor with the classifier. Finally, a fully-connected layer (FC), dropout of 0.5, and a final output layer sigmoid activation (S) with 11 classes are used. The model was trained with a learning rate of 0.001, a batch size of 128, and the Adam optimizer.

In the post-processing stage, Han et al. compare two aggregation strategies to obtain class predictions on a audio file level: first, they apply thresholds over averaged and normalized segment-wise class predictions (S1 strategy). Secondly, a sliding window of 6 segments and hop-size 3 segments is used for local aggregation prior to performing S1 strategy (S2 strategy). Refer to [28] for the identification threshold estimation. Apart from the model ensembling step (which combines different predictors), we were able to reproduce the evaluation results reported in [28], in terms of recognition performance, intermediate activation function (ReLU), and the optimal identification threshold

Method	Model Ensembling	Data set	Activation Function	Agg.	Micro Averaging			Macro Averaging			Opt. θ
					P	R	F	P	R	F	
Baseline system [28]	✓	IRMAS	**ReLU**	**S2**	**0.657**	**0.603**	**0.629**	**0.540**	**0.547**	**0.517**	**0.55**
Reproduction	-	IRMAS	ReLU	S1	0.591	0.548	0.568	0.530	0.477	0.471	0.40
			ReLU	**S2**	**0.609**	**0.544**	**0.574**	**0.501**	**0.507**	**0.475**	**0.55**
Experiment	-	MONOTIMBRAL	**LReLU**	**S1**	**0.645**	**0.678**	**0.661**	**0.685**	**0.681**	**0.657**	**0.8**
			LReLU	S2	0.619	0.695	0.655	0.657	0.690	0.649	0.7

Table 1. Performance metrics precision (P), recall (R), and F-score (F) from best results reported by [28], its reproduction with the IRMAS data set, and an experiment with the MONOTIMBRAL data set. The displayed results are the best settings obtained with respect to ReLU/LReLU activation functions, and S1/S2 aggregation strategies (see Section 3.1).

θ as shown in Table 1. Additionally, an experiment was conducted using monotimbral audio as input data to train the neural network. Following [28], we tested different intermediate activation functions (ReLU and LReLU) and both aggregation strategies. The monotimbral audio used for this experiment is further explained in Section 4.2.

3.2 Source Separation

Motivated by the previous experiment, which showed that recognition performance increases 5-10% by using monotimbral data as input, we explore the use of sound source separation as a pre-processing stage to musical instrument classification. The idea is to evaluate whether isolating the desired instrument from the mixture can improve classification performance. This section briefly describes two sound separation methods used in our experiments.

3.2.1 Phase-based Harmonic / Percussive Source Separation

The harmonic-percussive separation described in [3] works under the assumption that harmonic music instrument will exhibit stable phase contours as the ones obtained by differentiating the phase spectrogram in time. In contrast, given the broadband and transient-like characteristics of percussive instruments, this stability in phase cannot be expected. This system takes advantage of this fundamental distinction between harmonic and percussive instruments, and by calculating the expected phase change for a given frequency bin and hop size, a separation mask is created to extract harmonic components from the mix. The effects of the harmonic-percussive separation can be observed in Figure 2, where the spectrogram of the original audio mixture and of the harmonic and percussive components are displayed.

3.2.2 Pitch-Informed Solo/Accompaniment Separation

To extract solo instruments from multitimbral music, the method proposed in [4] was also used in our experiments. The system performs separation by first extracting pitch information from the solo instrument, and then closely tracking its harmonic components to create a spectral mask. To extract pitch information, the method proposed in [7] is used for main melody extraction. Pitch information is extracted by performing a pair-wise evaluation of spectral peaks, and by finding partials with well-defined frequency ratios. The pitch information extracted is then used to track the harmonic components in the separation stage, using common amplitude modulation, inharmonicity, attack length, and saliency as underlying concepts.

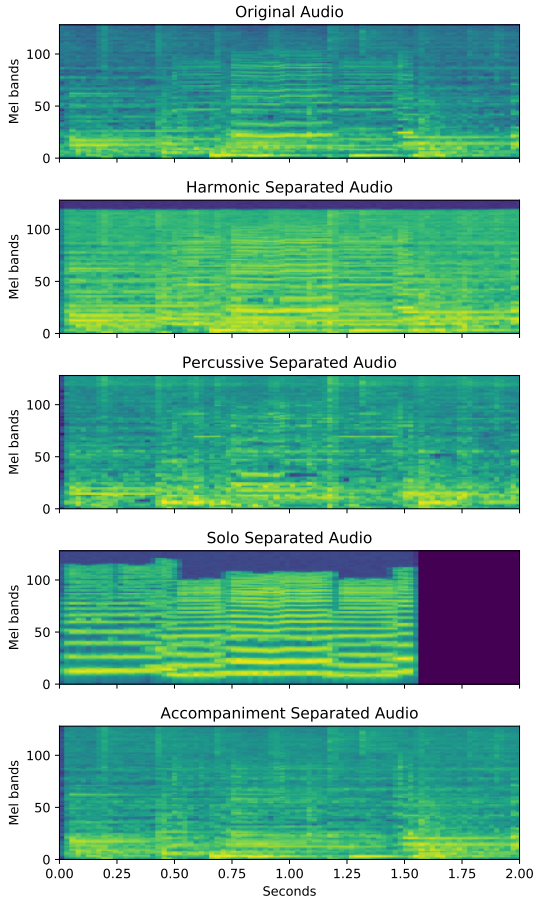

Figure 2. Mel-spectrograms of the original audio track, the harmonic/percussive components, and the solo/accompaniment components for a jazz excerpt of a saxophone solo played by John Coltrane. The audio mixture contains the solo saxophone, piano, bass and drums.

The performance of both the pitch detection and the separation stage in this system highly depend on the musical instrument to be separated: for musical instruments with clear, stable partials the separation performance can be very good. This is the case of woodwinds and string instruments such as the violin. However, for musical instru-

ments with a less stable spectral behavior such as the xylophone, or instruments with strong distortion effects such as electric guitars, separation can be noisy. The effects of the solo/accompaniment separation can be observed in Figure 2, where the spectrogram of the original audio mixture and of the solo and accompaniment components are displayed. It can be seen that starting from 1.50 seconds, the solo instrument is not detected and hence, no energy is assigned to the solo track.

3.3 Transfer Learning

For the special use-case of solo instrument recognition in jazz ensemble recordings, we aim at training a recognition model despite the small amount of available training data (see the JAZZ data set in Section 4.3). Here, transfer learning can be applied to fine-tune an existing classification model [13]. We assume that initially learnt feature representations for predominant AIR are highly relevant and therefore transferable for our use-case. Transfer learning has been successfully used in MIR for the task of sound event tagging in [6]. We refer the reader to [23] for a comprehensive overview of transfer learning in classification, regression, and clustering applications.

4. DATA SETS

4.1 IRMAS

The IRMAS data set (Instrument Recognition in Music Audio Signals) for predominant instrument recognition was first introduced by Bosch et al. in [2]. It is partitioned into separate training and test sets. The training set includes 6705 stereo audio files with a duration of 3 seconds each, extracted from more than 2000 recordings. All the recordings in the training data set are single-labeled and have a single predominant instrument. The amount of audio files per instrument is unevenly distributed and ranges from 388 to 778. The test set consists of 2874 stereo audio files with variable duration ranging from 5 to 20 seconds. These recordings are multi-labeled and cover 1-5 instrument labels per sample. The test set also shows a highly uneven instrument distribution with 62 to 1044 audio files per instrument class. As shown in Table 2, the data set contains 11 musical instruments: cello, clarinet, flute, acoustic guitar, electric guitar, organ, piano, saxophone, trumpet, violin, and singing voice. In the experiments described in Section 5.2.2, we use a subset denoted as IRMAS-Wind, which includes all recordings of the wind instruments in the IRMAS data set: flute, clarinet, saxophone, and trumpet. The motivation to create this subset is the improved performance of the solo/accompaniment separation algorithm (see section Section 3.2.1) and its timbral similarity to the JAZZ data set to apply transfer learning strategies (see Section 4.3). Following [28], training data was randomly split to training (85%) and validation (15%) to prevent overfitting by implementing early stopping. Testing data was randomly split into development testing data (50%) for optimum thresholding in post-processing, and pure testing data (50%) to obtain the final performance metrics (see Table 3).

Instrument		IRMAS		MONO.		JAZZ	
Class	**Subclass**	**#**	**h**	**#**	**h**	**#**	**h**
Cello		499	0.87				
Clarinet		567	0.71	26	0.32	31	0.53
Flute		614	1.17	29	0.42		
Acoustic Guitar		1172	3.08	30	0.38		
Electric Guitar		1702	5.00				
	Clean			28	0.43		
	Distorted			30	0.34		
Organ		1043	2.25				
Hammond Organ				30	0.44		
Piano		1716	5.40	27	0.38		
Electric Piano				29	0.31		
Saxophone		952	2.16	29	0.34		
	Soprano					30	0.53
	Alto					29	0.53
	Tenor					32	0.53
Trombone						27	0.53
Trumpet		744	1.29	29	0.35	36	0.53
Violin		791	1.56	27	0.47		
Voice		1822	5.38				
	Female			21	0.26		
	Male			20	0.26		
Double Bass				27	0.28		
Synthesizer				30	0.77		
TOTAL		11622	28.87	412	5.75	185	3.18

Table 2. Overview of the three data sets IRMAS, MONOTIMBRAL, and JAZZ, which includes various instrument classes and subclasses. Both the number of labels (#) and the total duration in hours (h) is given for each data set.

4.2 MONOTIMBRAL

The MONOTIMBRAL data set includes monotimbral (single-labeled) recordings, i. e., monophonic or polyphonic recordings without overlap of other instruments, of 15 musical instrument classes: acoustic guitar, clarinet, double bass, electric guitar clean, electric guitar distorted, electric piano, flute, hammond organ, piano, saxophone, female singing voice, male singing voice, synthesizer, trumpet, and violin. The data set contains 412 stereo audio files with variable duration from 10 to 120 seconds, manually selected from various segments of YouTube videos. The MONOTIMBRAL data set was randomly split equally into a training and test set based on an equal distribution of audio files per instrument class (see Table 3).

4.3 JAZZ

As one specific use-case, we aim at classifying among the six most popular brass and reed instruments in jazz solos: trumpet (tp), clarinet (cl), trombone (tb), alto saxophone (as), tenor saxophone (ts), and soprano saxophone (ss). While the number of instruments is smaller compared to the IRMAS and MONOTIMBRAL data sets, they have a higher timbral similarity, considering particularly the three saxophone subclasses. In order to prepare a data set, we first randomly selected solos from the Weimar Jazz Database [25] and enriched the data set with additional jazz solos. While the number of instruments is smaller compared to the IRMAS and MONOTIMBRAL data sets, the audio samples were chosen to maximize diversity of

performing artists. Moreover, examples from each class were randomly selected to have the same duration (see Table 2), achieving equal distribution of spectrogram examples across instrument classes. As with the other data sets, the JAZZ data set split randomly as the other data sets (see Table 3). Since jazz recordings cover many decades of the 20th century, the instrument recognition task is further complicated by different recording techniques.

For additional information regarding the MONOTIMBRAL and JAZZ data sets, refer to the complimentary website for this paper [12].

	Training Data Set (85/15)		Testing Data Set (50/50)	
	Train	Validation	Development	Pure
IRMAS	17094	3021	48064	48055
IRMAS-Wind	5486	970	10447	10446
Monotimbral	8676	1539	10620	10610
JAZZ	7206	1275	1678	1271

Table 3. Number of mel spectrogram examples for each data set split into Train, Validation, Development, Pure data sets.

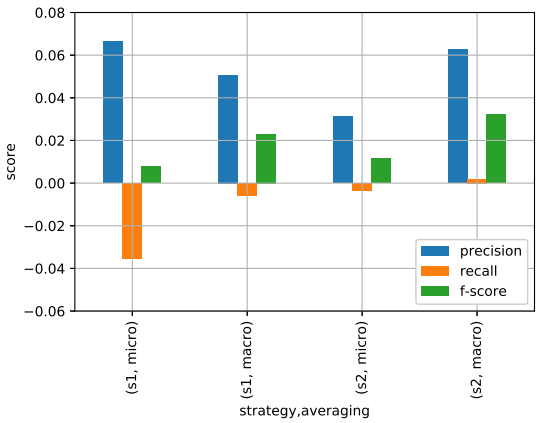

Figure 3. Comparison of the AIR system trained on the harmonic stream and the baseline model trained with the original IRMAS data set. Differences between evaluation metrics are shown for both aggregation strategies S1 and S2 (compare Section 3.1) as well as micro and macro averaging (compare Section 5.1).

5. EVALUATION

5.1 Metrics

Following [2, 11, 28], precision, recall, and f-scores were calculated for both micro and macro averages. Micro averaging gives more weight to instrument classes with higher appearance in the data distribution. Macro averaging is calculated per label, representing an overall performance of the system.

5.2 Improving Predominant Instrument Recognition using Source Separation

5.2.1 Harmonic / Percussive Separation

After processing the audio files with the harmonic/percussive separation introduced in Section 3.2.1, we first retrained the baseline model independently on the harmonic stream and percussive stream. Furthermore, we created a two-branch model that processes the harmonic and percussive stream in parallel and fuses the results in the final fully-connected layers, similar to [15]. As shown in Figure 3, using the harmonic stream marginally improved recognition results for both aggregation strategies S1 and S2 by up to 3% in f-score for the multitimbral IRMAS data set. In contrast, we did not observe an improvement for the MONOTIMBRAL data set. Using the two-branch model did not improve the performance on the IRMAS data set and worsens the performance on the MONOTIMBRAL data set.

5.2.2 Solo / Accompaniment Separation

The aim of performing this separation is to further improve the quality of the input audio to the classification system. All experiments described in this section were performed on the IRMAS-Wind and the JAZZ data sets (see Section 4), given the performance of the solo/accompaniment algorithm. Both data sets also have similar timbral characteristics, which represents our targeted scenario.

We compare AIR models trained on the original audio tracks with models trained on the solo stream obtained from the solo/accompaniment separation. As shown in Table 4, applying the solo/accompaniment separation as pre-processing step improves the AIR performance by 3.8% in macro f-score for the IRMAS-Wind data set and 13.4% for the JAZZ data set using the S1 strategy. Additionally both micro and macro averages result in similar values, given the even distribution of examples of the JAZZ data set. The results might also indicate that error propagation from transcription errors to the source separation algorithm are not critical, since the instrument recognition results are averaged over time and the approximate accuracy of the pitch detection algorithm is 80% [7].

		F-Score	
Data set	S/A Separation	Micro	Macro
IRMAS-Wind	-	0.684	0.598
IRMAS-Wind	✓	0.713	0.636
JAZZ	-	0.657	0.669
JAZZ	✓	**0.805**	**0.803**

Table 4. Performance metrics obtained by training the baseline model with the IRMAS-Wind and JAZZ data sets. Best results were obtained using aggregation strategy S1.

5.3 Combining Source Separation and Transfer Learning for Jazz Solo Instrument Recognition

For our final use-case of recognizing jazz solo instruments, we aim at combining solo/accompaniment separation and transfer learning strategies. We use the models trained on the IRMAS-Wind data set (with and with-

out solo/accompaniment separation) as starting point for the transfer learning approach. All models were trained from scratch following the original parameters from [28]. The JAZZ data set includes recordings from trombone and three saxophone subclasses: tenor, alto, and soprano. Additionally, the trumpet and the clarinet classes were already included in the IRMAS-Wind data set. One main challenge is that while the characteristics of the predominant melody instruments in the IRMAS and JAZZ data sets are similar, the background instrumentation and recording conditions are often very different. We remove the last sigmoid layers of models pre-trained with the IRMAS-Wind data set and replace them by a 6-class sigmoid layer, considering the JAZZ data set. For testing, we compare two approaches: (1) the one-pass method which re-trains the last classification layer using a learning rate of $\alpha = 0.01$ (10 times the original learning rate), while all remaining layers remain fixed, and (2) the two-pass approach where we further re-train all layers in a second training step with a smaller learning rate of $\alpha = 0.001$. Table 5 shows the classification performance on the JAZZ data set for different system configurations with the one-pass and two-pass strategies, as well as with and without the solo/accompaniment separation. The best performance was achieved by combining solo/accompaniment separation and the two-pass transfer learning strategy.

		F-score	
S/A Separation	Transfer Learning	Micro	Macro
-	One-pass	0.605	0.621
✓	One-pass	0.738	0.748
-	Two-pass	0.583	0.610
✓	Two-pass	**0.787**	**0.780**
✓	-	**0.805**	**0.803**

Table 5. Performance metrics obtained by combining solo/accompaniment separation with transfer learning on the JAZZ data set. The results obtained by training the model from scratch (without transfer learning) are also shown in the bottom row for reference. Best results were obtained using aggregation strategy S1.

It can also be observed that the transfer learning model shows a lower macro f-measure of 0.780 than the model trained from scratch with 0.803 (see bottom row of Table 5). To further understand this behavior, six additional 10 s (unseen) jazz solo excerpts [1] were analyzed. Figure 4 shows segment- and clip-wise predictions for these six solo excerpts using solo/accompaniment separation. The figure shows the results for the best transfer learning system and the model trained on the JAZZ data set from scratch [12]. A total of 20 predictions were generated per excerpt on 1 s long windows using a 50 % overlap. These results suggest that transfer learning can improve generalization of unseen data, but needs further systematic investigations on a larger testing data set.

[1] Ornette Coleman - Ramblin (as), Buddy DeFranco - Autumn Leaves (cl), John Coltrane - My Favorite Things (ss), Frank Rossolino - Moonlight in Vermont (tb), Lee Morgan - The Sidewinder (tp), Michael Brecker - African Skies (ts)

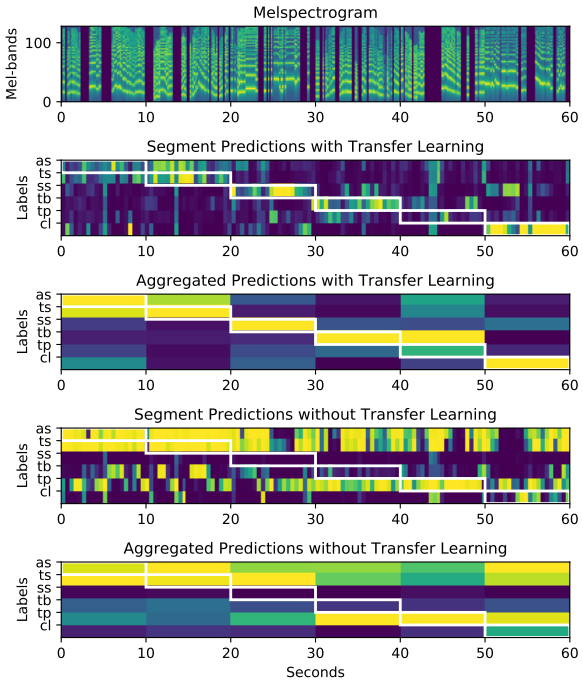

Figure 4. Mel-spectrogram of 10 second excerpts from six jazz solos covering all solo instruments (top), segment-wise and aggregated clip-wise predictions (using strategy S1) are shown below for a model trained via transfer learning (two-pass) and a model trained from scratch. Clip-wise ground truth is plotted in white rectangles [12].

6. CONCLUSION

In this paper, we investigated two methods to improve upon a system for AIR on multitimbral ensemble recordings. We first evaluated two state-of-the-art source separation methods and showed that on multitimbral audio data, analyzing the harmonic and solo streams can be beneficial compared to the mixed audio data.

For the specific use-case of jazz solo instrument classification, which involves classifying six instruments with high timbral similarity, combining solo/accompaniment source separation and transfer learning methods seems to lead to AIR models with better generalization to unseen data. This must be further investigated by increasing the size of the JAZZ data set. While source separation allows to narrow the focus on the predominant instrument, transfer learning allows to exploit useful feature representations learned from related instruments. In the future, a deep learning model capable of discriminating highly similar instruments could potentially be applied in other timbre-related recognition tasks such as performer identification [25].

7. ACKNOWLEDGEMENTS

This work has been supported by the German Research Foundation (AB 675/2-1).

8. REFERENCES

[1] Rachel M. Bittner, Brian McFee, Justin Salamon, Peter Li, and Juan P. Bello. Deep salience representations for f0 estimation in polyphonic music. In *Proceedings of the International Society of Music Information Retrieval (ISMIR)*, Suzhou, China, October 2017.

[2] Juan Bosch, Jordi Janer, Ferdinand Fuhrmann, and Perfecto Herrera. A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 559–564, Porto, Portugal, 2012.

[3] Estefanía Cano, Mark D. Plumbley, and Christian Dittmar. Phase-based harmonic/percussive separation. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1628–1632, Singapore, 2014.

[4] Estefanía Cano, Gerald Schuller, and Christian Dittmar. Pitch-informed solo and accompaniment separation towards its use in music education applications. *EURASIP Journal on Advances in Signal Processing*, 23:1–19, 2014.

[5] Aleksandr Diment, Padmanabhan Rajan, Toni Heittola, and Tuomas Virtanen. Modified group delay feature for musical instrument recognition. In *Proceedings of the International Symposium on Computer Music Multidisciplinary Research*, pages 431–438, Marseille, France, 2013.

[6] 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, USA, 2017.

[7] Karin Dressler. *Automatic transcription of the melody from polyphonic music*. PhD thesis, TU Ilmenau, Germany, Jul 2017.

[8] Zhiyao Duan, Bryan Pardo, and Laurent Daudet. A novel cepstral representation for timbre modeling of sound sources in polyphonic mixtures. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7495–7499, Florence, Italy, May 2014.

[9] Antti Eronen and Anssi Klapuri. Musical instrument recognition using cepstral coefficients and temporal features. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 753–756, Istanbul, Turkey, 2000.

[10] Slim Essid, Gael Richard, and Bertrand David. Musical instrument recognition on solo performances. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pages 1289–1292, Vienna, Austria, 2004.

[11] Ferdinand Fuhrmann. *Automatic musical instrument recognition from polyphonic music audio signals*. PhD thesis, Universitat Pompeu Fabra, 2012.

[12] Juan S. Gómez, Jakob Abeßer, and Estefanía Cano. Complementary website. `https://github.com/dfg-isad/ismir_2018_instrument_recognition`.

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

[14] Mikus Grasis, Jakob Abeßer, Christian Dittmar, and Hanna Lukashevich. A multiple-expert framework for instrument recognition. In *Proceedings of the International Symposium on Computer Music Multidisciplinary Research (CMMR)*, Marseille, France, October 2013.

[15] Thomas Grill and Jan Schlüter. Music Boundary Detection Using Neural Networks on Spectrograms and Self-Similarity Lag Matrices. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Nice, France, 2015.

[16] Toni Heittola, Anssi Klapuri, and Tuomas Virtanen. Musical instrument recognition in polyphonic audio using source-filter model for sound separation. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 327–332, Kobe, Japan, 2009.

[17] Filip Korzeniowski and Gerhard Widmer. A fully convolutional deep auditory model for musical chord recognition. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing, MLSP*, pages 1–6, Salerno, Italy, 2016.

[18] A. G. Krishna and T. V. Sreenivas. Music instrument recognition: from isolated notes to solo phrases. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 4, pages 265–268, Quebec, Canada, 2004.

[19] Simon Leglaive, Romain Hennequin, and Roland Badeau. Singing voice detection with deep recurrent neural networks. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 121–125, Brisbane, Australia, April 2015.

[20] Peter Li, Jiyuan Qian, and Tian Wang. Automatic instrument recognition in polyphonic music using convolutional neural networks. *CoRR*, abs/1511.05520, 2015.

[21] Daniel Matz, Estefanía Cano, and Jakob Abeßer. New sonorities for early jazz recordings using sound source separation and automatic mixing tools. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, pages 749–755, Malaga, Spain, 2015.

[22] Alexey Ozerov, Emmauel Vincent, and Frederic Bimbot. A general flexible framework for the handling of prior information in audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4):1118–1133, May 2012.

[23] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, Oct 2010.

[24] Taejin Park and Taejin Lee. Musical instrument sound classification with deep convolutional neural network using feature fusion approach. *CoRR*, abs/1512.07370, 2015.

[25] Martin Pfleiderer, Klaus Frieler, Jakob Abeßer, Wolf-Georg Zaddach, and Benjamin Burkhart, editors. *Inside the Jazzomat - New Perspectives for Jazz Research*. Schott Campus, 2017.

[26] H. Tachibana, T. Ono, N. Ono, and S. Sagayama. Melody line estimation in homophonic music audio signals based on temporal-variability of melodic source. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 425–428, Dallas, Texas, March 2010.

[27] Steven Tjoa and K. J. Ray Liu. Musical instrument recognition using biologically inspired filtering of temporal dictionary atoms. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 435–440, Utrecht, The Netherlands, 2010.

[28] Yoonchang Han and Jaehun Kim and Kyogu Lee. Deep convolutional neural networks for predominant instrument recognition in polyphonic music. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(1):208–221, Jan 2017.

[29] Li-Fan Yu, Li Su, and Yi-Hsuan Yang. Sparse cepstral codes and power scale for instrument identification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7460–7464, Florence, Italy, 2014.