IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 April 2023, accepted 25 April 2023, date of publication 27 April 2023, date of current version 5 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3271146

RESEARCH ARTICLE

Improved Separation of Polyphonic Chamber Music Signals by Integrating Instrument Activity Labels

MARKUS SCHWABE AND MICHAEL HEIZMANN
Institute of Industrial Information Technology (IIIT), Karlsruhe Institute of Technology (KIT), 76187 Karlsruhe, Germany
Corresponding author: Markus Schwabe (markus.schwabe@kit.edu)

This work was supported by the Karlsruhe Institute of Technology (KIT).

ABSTRACT The separation of music signals is a very challenging task, especially in case of polyphonic chamber music signals because of the similar frequency ranges and sound characteristics of the different instruments to separate. In this work, a joint separation approach in the time domain with a U-Net architecture is extended to incorporate additional time-dependent instrument activity information for improved instrument track extractions. Different stages are investigated to integrate the additional information, but an input before the deepest encoder block achieves best separation results as well as highest robustness against randomly wrong labels. This approach outperforms a label integration by multiplication and the input of a static instrument label. Targeted data augmentation by incoherent mixtures is used for a trio example of violin, trumpet, and flute to improve separation results. Moreover, an alternative separation approach with one independent separation model for each instrument is investigated, which enables a more flexible architecture. In this case, an input after the deepest encoder block achieves best separation results, but the robustness is slightly reduced compared to the joint model. The improvements by additional information on active instruments are verified by using real instrument activity predictions for both the joint and the independent separation approaches.

INDEX TERMS Music source separation, polyphonic chamber music, active instruments, end-to-end deep learning.

I. INTRODUCTION

Music Source Separation (MSS) is one of the main parts of Music Information Retrieval (MIR). Since its aim is to extract isolated tracks for the sources recorded together, MSS can either be used in specific scenarios like singing voice [1] or lead and accompaniment separation [2], or it can be applied as a preprocessing tool for many subsequent MIR tasks like music transcription [3] or audio remixing [4].

Although those applications are very useful for various music sources, especially chamber music instruments, most multi-instrument source separation approaches concentrate on the separation of the four target sources 'vocals', 'bass', 'drums', and 'other', like in the commonly used MUSDB18 dataset [5]. One reason is the lack of large datasets with separated music track recordings of many different instruments. Generating synthesized music tracks [6] is an option to get large amounts of data, but their variability is very limited and therefore the dataset is unsuitable for learning approaches. An example of a small dataset with real instrument track recordings is URMP [7], in which the line-up changes for the different music pieces, therefore the source number and source types to separate vary. Additionally, the separation of chamber music instruments is challenging due to their similar frequency ranges and sound characteristics.

Generally, MSS has benefited heavily from the introduction of deep learning techniques. They clearly outperformed former signal processing approaches by means of convolutional structures in neural networks [8], denoising autoencoders [9], and variational autoencoders [10]. Further improvements could be achieved through data augmentation and a blending of different neural networks [11]. Nowadays,

The associate editor coordinating the review of this manuscript and approving it for publication was Tao Huang.

deep learning approaches are used as standard technique in all kinds of source separation fields. All recent MSS approaches can be divided into two basic concepts: One concept is the separation based on time-frequency representations (mostly the short-time Fourier transform), which is used for example in Open-Unmix [12], Spleeter [13], and D3Net [14]. To obtain the final separated sources, masks are calculated by means of the predicted matrices for each source and then multiplied with the time-frequency representation of the mixture. Those masks usually comprise only real values between 0 and 1, but the separation results can be improved by complex masks [15]. The second concept is the end-to-end separation in the time domain, which does not require any preprocessing and predicts all track signals directly. As this approach can estimate signal amplitudes and phases directly, the potential performance is much higher than with mask multiplication, but in reality systems like Demucs [16] or Wave-U-Net [17] reach comparable results to those of the first concept. Latest systems like KUIELab-MDX-Net [18] and Hybrid Demucs [19] proposed the fusion of two parallel networks of both domains, which improves MSS performance but results in large architectures.

In order to further improve the separation performance, additional information about the music pieces has been used. Miron et al. proposed a score-informed system trained on four synthetic classical instruments [20], but synthetic instruments are easy to separate because their timbre is well-defined and has little variability during playing. During an optimization-based approach, synthesized music signals can help to improve separation results by means of minimizing the error between the resynthesized and the input mixture [21]. But the synthesizer requires an exact audio-synchronized musical score, which represents a strong limit. This drawback can be avoided by fusing music separation and transcription like in [22], where separated instrument tracks and the music score are predicted jointly in a multitask approach. Multitask learning for MSS and instrument activation, i.e. time-dependent detection of playing instruments, is proposed in [23]. Both multitask approaches show that each task can benefit from the fusion, but they need large training datasets because of the large models.

Another possibility for integrating additional information is conditioned source separation, in which a label vector, describing whether the respective instrument is playing or not in the given music piece, is fed into the MSS network. This one-hot-encoded vector can be integrated directly in the neural network by multiplication [24] or by means of a second control network and special layers in the separation network, e.g. feature-wise linear modulation (FiLM) layers [25] or latent source attentive frequency transformation (LaSAFT) [26]. With the control network, the condition enables the system to separate different instruments by the same architecture with only one output track. Beside a label vector, other input data like an audio sample of one undisturbed instrument [27] or a video stream [28] can be given to the control network. Moreover, a conditioning can even select the task of a unified model that is able to do music source separation, transcription, or synthesis for unseen instruments [29]. Core of this approach is a pitch-timbre disentanglement module, which is based on a common encoder-decoder architecture like in many source separation approaches presented above. The conditioning on unseen instruments is enabled by a query-by-example subnet that encodes the query spectrogram of a given instrument sound example, which defines the desired instrument type, and feeds the result into the latent space.

Most conditioning approaches only include static additional information like timbre or line-up, but some applications benefit more from time-dependent information. Meseguer-Brocal and Peeters improved the singing voice extraction with known lyrics by conditioning a separation model with time-aligned phonemes [30]. The time-dependent phoneme matrix is generated from the respective lyrics. For instrument recordings, prior knowledge about time-dependent instrument activity can be introduced by a temporal segmentation of the mixture [31]. This segmentation has to be done by a user and enables a source separation model adaption on the given music piece.

In this work, the separation of monaural chamber music signals from small ensembles with different instruments is improved by the integration of additional instrument labels. Existing conditioned approaches consider only static instrument information or need user input, which is not practical for automated source separation. Since time-dependent instrument activity contains useful information for MSS, especially in case of chamber music with many instruments and different line-ups, it is integrated here as additional input. The main contributions of this work are:

- A time-dependent conditioning approach for music instrument separation of chamber music is introduced.
- Two separation model approaches are investigated, one joint model that predicts all source signals at the same time and a more flexible approach that uses independent source-targeted models for each instrument.
- The integration of time-dependent instrument activity labels by concatenation is investigated for all encoder input steps. Furthermore, the robustness against random label errors is analyzed by means of toggled labels.
- Real instrument activity predictions are analyzed to verify the results with simulated label errors.

An end-to-end MSS approach is developed whose architecture is presented in section II. Despite the lack of large chamber music datasets for MSS, sufficient data can be generated for predefined ensembles, which is explained in section III. All experiments and separation results for different integration of additional input data are analyzed and discussed in section IV-A for the joint separation models and in section IV-B for the independent separation approach. The influence of real instrument activity predictions is investigated in section IV-C. They are necessary if the time-dependent instrument activity is not available in advance.

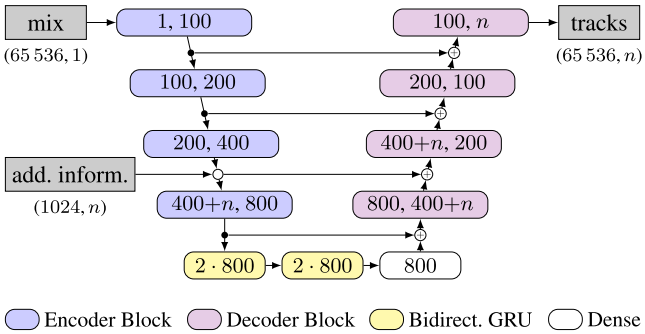

FIGURE 1. Schematic model structure with channel numbers of each layer for additional information integration before encoder 4.

II. MODEL ARCHITECTURE

For the core MSS model of this work, an architecture for the separation in the time domain is chosen due to the improved potential performance compared to approaches based on time-frequency representations, as stated in the introduction. Moreover, the additional time-dependent information about active instruments can be included more easily in an arbitrary fine-grained resolution. An example of the whole model architecture with additional label integration before encoder block 4 is given in Fig. 1.

A. SOURCE SEPARATION MODEL

The separation model architecture is inspired by Demucs [16], which achieved one of the best separation performances for the MUSDB18 dataset with its target sources 'vocals', 'bass', 'drums', and 'other'. It consists of a U-Net structure with several encoder and decoder blocks and two bidirectional LSTMs in the bottleneck of the network. Several modifications have been made to adapt Demucs on the case of small chamber music ensembles with different instruments of partially similar frequency range.

First, the number of output tracks is set before the training process to the desired instrument number n that should be separated. Preliminary experiments showed that four encoder and decoder blocks performed comparably to six encoder and decoder blocks like in Demucs, therefore the smaller architecture with four blocks each is used. Since separation models in the time domain usually have lots of parameters, model size reduction is very important regarding the applicability. All encoder and decoder blocks are identical to those in Demucs, consisting of two 1D convolutions and the activation functions ReLU and GLU. The channel number of each encoder and decoder block is set equal to 100 in the first stage and then doubled in all subsequent stages to enable the incorporation of a sufficiently large number of specialized features for the separation. In the bottom layer, two bidirectional gated recurrent units (GRUs) replace the two LSTMs of the Demucs architecture, because GRUs have less parameters than LSTMs and perform comparably in most cases, also in MSS [32]. This model architecture consists of about 41 million parameters, only slightly dependent on n.

The input mixture signal is cut into segments of 65 536 samples due to GPU memory constraints and then fed in the first encoder block. Input segments with 8 times more samples were investigated in preliminary experiments because of the longer input in the original Demucs architecture, but they did not lead to improvements, so the smaller length is used. We consider the basic case of monaural mixture and track signals in this work, to which signals with any higher channel number can be reduced. Due to the kernel stride of 4 of each encoder block's first convolution layer, the input length is reduced by the factor 4 in each encoder stage.

Beside a joint separation model for all instruments to separate, using independent separation models for each instrument allows for more flexibility, according to the playing instruments. Moreover, later model enhancements with further instruments can be easily implemented by integrating additional independent separation models. Therefore that separation architecture is investigated in this work, too. Each independent model has the same structure depicted in Fig. 1 with $n = 1$, but the channel numbers per encoder and decoder block are halved compared to the case of joint separation because each model has only to extract one instrument track and n independent models need more resources. Each instrument model has about 11.5 million parameters. The separation prediction needs about 40 ms per model for one input sequence, whereas the whole calculation time of the joint approach is about 65 ms for one sequence.

B. INTEGRATION OF ADDITIONAL INFORMATION

Additional time-dependent instrument activity information is included by concatenating one encoder block input with the activity labels of all considered instruments. Thus, this information comprises the activity labels of n instruments in case of joint separation models and only the particular instrument in case of independent models. Due to the concatenation at one encoder stage, the MSS network can already profit from the additional information during the data compression in the encoder. Moreover, the concatenation enables the model to extract only the relevant information from the activity labels and to be robust to the case of wrong labels, which is not assured in case of a multiplicative incorporation like in [24]. A multiplication or an addition of the labels, like e.g. in positional encoding, would distort the data and is therefore not suitable for source separation. As no additional layers like FiLM or LaSAFT are needed, the model size is nearly the same as without additional information.

Beside the incorporation before encoder 1, which means the concatenation of the input mixture signal and the instrument activity labels, the additional information can be included at later encoder stages as well. Then the time resolution of the labels has to be reduced by the factor 4^m to fit the output length of the preceding encoder block m, e.g. factor 64 in Fig. 1. If the concatenation is not done in stage 1, the resulting encoder input is also transmitted to decoder block m via the respective skip connection and added to the output of the preceding decoder block.

III. DATASET

Existing datasets with chamber music and corresponding instruments are often not suited to the MSS problem because the tracks of the different instruments are not available or every instrument recording comprises only few notes. But the goal of polyphonic MSS is to separate different melodies sounding together that are played by different instruments. Such a polyphonic scenario was considered in the University of Rochester Musical Performance (URMP) dataset [7]. It consists of 44 classical music pieces of small ensembles (11 duets, 12 trios, 14 quartets, and 7 quintets) with a total duration of about 80 min and a sampling rate of 48 kHz. The ensembles have various line-ups from 13 instruments including strings, woodwinds, and brass. In addition, note-level pitches of all separated instrument tracks are given with onsets and note duration.

Some ensemble line-ups contain more than one player of the same instrument type. But we aim to separate tracks of different instruments in this work, so all tracks of the same instrument and the same music piece are mixed to one output track of the respective instrument. Consequently, the music pieces of this modified URMP dataset reduce to 4 solos, 12 duets, 20 trios, and 8 quartets. Those recordings are divided into 36 train, 5 validation (No. 5, 12, 17, 24, and 40), and 3 test pieces (No. 8, 18, 41), all consisting of at least one duet, trio, and quartet. In order to enlarge this relatively small dataset, single instrument recordings taken from different music pieces of the original training dataset are remixed. This data augmentation strategy enables the creation of a large quantity of incoherent mixtures by combining various instruments, but decreases the percentage of training samples with harmonic relation between partials. To ensure the harmonic separation ability of the trained network, several coherent mixtures are necessary. Thus, this data augmentation can only be used to a limited extent.

For the separation of a known ensemble line-up, targeted data augmentation can focus on training samples with only the relevant instrument combination, which reduces the augmentation dataset size drastically. That is done in this work, creating 50 additional music pieces by remixing combinations of the predefined example trio of violin, trumpet, and flute before the training process of all models presented. All instrument tracks are shifted in time by a random time offset and are multiplied by a random loudness factor between 0.7 and 1.3 during the mixing process to ensure a large mixture variety. The additional music pieces are divided into 45 for training and 5 for validation, which results in 81 training and 10 validation music pieces for the augmented dataset.

During training, the mixture and track signals are cut in segments of about 1.365 s due to the model input length of 65 536 samples and 48 kHz sampling rate. For this segmentation, a random time offset in the range $[0, 65535]$ is used for every music piece in each training epoch to avoid overfitting. All non-overlapping segments of the music piece are extracted based on this time offset. Downsampling could enlarge the time span of the input segments but could also cause a loss of information, therefore it is not considered.

Time-dependent instrument activity labels are generated by means of the note-level pitches of the instrument tracks. The labels of each instrument represent a binary vector with length according to the input stage (described in section II-B), containing '1' in all time samples where a note is active for the respective instrument and '0' in all the others. Those binary label vectors of all n sources are fed directly to the respective independent separation models or they are concatenated and given to the joint separation model as a binary time-dependent matrix. As time-dependent instrument activity labels are rarely known in real applications, they can be predicted by a preceding time-dependent instrument detection approach like [33], causing some errors in the additional information. In order to investigate the integration of such imperfectly predicted labels, a defined percentage of the input instrument label vector values is toggled. Thereby, the indices of the toggled values are chosen randomly.

IV. EXPERIMENTS

Several experiments are conducted to analyze the integration of time-dependent information about active instruments in MSS. First, the influence of correct and randomly toggled ground truth labels is investigated for joint as well as independent separation model architectures. Thereby, both the generic case with $n = 13$ instruments and the targeted case of the predefined example trio are analyzed. Then the separation results are validated by real instrument activity predictions as additional input. The separation models are implemented in TensorFlow and trained using Adam optimizer [34] with MSE loss, a learning rate of 3×10^{-4}, batch size 32, and maximum 500 epochs. Early stopping is enabled in case of no validation improvement during 50 epochs. At the end, the model with the best performance on the validation dataset is taken as the resulting separation system. For all random functions, the same seed (0) is defined at the beginning to reduce performance variances by different initialization as well as to make the results more comparable.

Preliminary experiments showed that targeted data augmentation improves not only the separation of the predefined instruments, but in the vast majority of cases all instrument results. Therefore the augmented dataset is taken during training. Performance is evaluated by means of the scale-invariant signal-to-distortion ratio (SI-SDR) [35], which is a robust quality metric for separated sources.

A. JOINT SEPARATION MODELS

The modified URMP dataset comprises 13 instruments, therefore joint models predicting $n = 13$ instrument tracks are investigated first. For the integration of the time-dependent instrument activity labels, all possible input stages in the encoder are analyzed, whereby the input stage is named by the subsequent encoder block. As the three music pieces of the test dataset comprise only six playing instruments, Table 1 lists the averaged separation results for those six

TABLE 1. SI-SDR results in dB for test dataset instruments and different stages of additional information integration in a joint separation model.

Additional Information	Instruments						Average	Data Augm. Improvement
	Bassoon	*Flute*	*Oboe*	*Saxophone*	*Trumpet*	*Violin*		
-	−7.60	3.06	−0.99	−6.62	−0.28	−1.07	−2.25	+3.15
Encoder 1	**1.31**	1.89	−0.75	−5.86	**3.95**	**1.62**	0.36	+2.74
Encoder 2	0.26	0.82	−2.28	−3.80	3.80	−1.51	−0.45	+8.17
Encoder 3	−2.77	3.47	−3.04	−3.80	1.60	−1.68	−1.04	+2.45
Encoder 4	0.86	**4.74**	**−0.42**	**−3.53**	1.36	0.65	**0.61**	+1.98
GRU	0.62	4.01	−1.38	−3.80	1.00	−0.16	0.05	+4.04
Mixture (no sep.)	−3.51	1.29	−11.29	−6.17	−4.12	−4.98	−4.80	-
Open-Unmix [12]	−7.52	1.92	−0.45	−30.92	0.08	**2.53**	−5.73	-
Label multiplication	−7.60	3.17	−0.96	−6.59	−0.18	−1.12	−2.22	+3.13
Const. label (Enc. 4)	0.94	3.73	**−0.39**	−4.33	0.94	−0.42	0.08	+1.93

instruments. The estimated signals of the other seven instruments consist mostly silence and are therefore not evaluated in detail here. Beside the separation results with additional information at the input of all encoder blocks and the first GRU block, the results for the predicted tracks of a model without additional instrument activity labels (-) are given as a reference. Further references are the mixture signal without any applied separation algorithm as 'worst case', the literature approach Open-Unmix [12] with time-frequency matrix input and independent instrument models trained with our augmented dataset, and the multiplication of the instrument activity labels with the respective track estimation of the separation model without additional information. Their separation results are added in Table 1. To investigate the improvement due to time-dependency of the activity information, constant labels for each instrument are fed in the separation model at the 4th encoder block, which represents the best stage for the time-dependent case. This third reference is also added in the result table.

As given in Table 1, the integration of time-dependent instrument activity labels improves the average separation results for all input stages by at least 1.21 dB. The best separation results are achieved by an integration of the additional information before the last encoder block (encoder 4). Thus, the instrument activity labels seem to be very useful in combination with 'higher' features with a large receptive field like in the latent representation. But a subsequent encoder block is advantageous to connect the additional information with features of the preceding encoder, in contrast to the integration directly before GRU. Although the label input before encoder block 4 has a lower time resolution of about 1.33 ms, it is still time-dependent over 1024 values.

An integration of time-dependent instrument activity by multiplication with the estimated instrument tracks improves the separation results of its basis, the separation model without additional information, only marginally. The multiplication suppresses artifacts and other instrument sounds during the time in which the target sources are not played, so the basic separation model seems to suppress the majority of those sounds by itself. Consequently, the concatenation of the additional information in the model increases its separation ability and performance. Time-dependent instrument activity labels lead to a further performance improvement, because

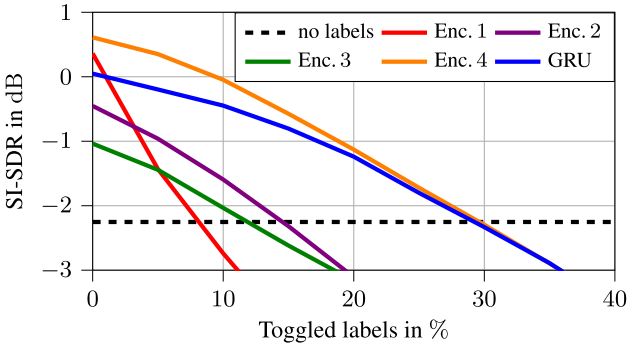

FIGURE 2. SI-SDR results for randomly toggled labels of defined percentage and joint separation models with 13 instruments.

the separation results with input stage before encoder block 4 and a constant instrument label, like e.g. in the literature approach of Slizovskaia [24], are 0.53 dB lower than the results for the time-dependent case. For the other input stages, the improvement due to time-dependency compared to the constant label case is even higher.

The usage of the augmented dataset improves the separation results by at least 1.98 dB, even in the investigated case of models with 13 instruments. Especially for the input stage before encoder 2, data augmentation is necessary for acceptable separation results, because the SI-SDR of the original dataset is with −8.62 dB far below the mixture case without any separation. Obviously, that model cannot extract the relevant features to separate the instrument signals. One reason is the direct transmission of the additional information to the last decoder block through the highest skip connection (compare Fig. 1). From the combination of the resulting 113 decoder channels and the small amount of data, the last decoder block and therefore the whole model learns wrong connections during training. These connections lead to very low SI-SDR values for bassoon and trumpet. In the case of encoder 2 with data augmentation, the performance is comparable to the other stages. Therefore data augmentation should be used for a better separation with small datasets.

Toggled labels mimic wrong input labels which influence the separation results. The specific performance reduction for different percentages of randomly toggled labels (only during testing) is visualized in Fig. 2. Input stages near

TABLE 2. SI-SDR results in dB for test dataset and defined trio with different stages of additional information integration in a joint separation model.

Additional Inform.	Instruments *Flute*	*Trumpet*	*Violin*	Average	Augm. Impr.
-	2.60	−0.26	−0.07	0.76	+3.97
Encoder 1	**5.06**	0.77	1.69	2.51	+5.36
Encoder 2	4.00	1.20	1.40	2.20	+5.19
Encoder 3	4.78	1.23	2.00	2.67	+4.86
Encoder 4	4.96	**1.46**	**2.38**	**2.93**	+4.42
GRU	4.80	0.98	2.35	2.71	+4.95
Label mult.	2.70	−0.12	−0.12	0.82	+3.91
Const. (E. 4)	4.14	1.16	1.41	2.24	+4.29

TABLE 3. SI-SDR results in dB for test dataset and independent models for each trio instrument with different stages of additional information integration.

Additional Inform.	Instruments *Flute*	*Trumpet*	*Violin*	Average	Augm. Impr.
-	1.99	3.38	−1.00	1.45	+5.27
Encoder 1	0.46	1.58	−2.36	−0.11	+2.45
Encoder 2	1.63	1.22	0.10	0.98	+3.61
Encoder 3	2.42	4.03	0.42	2.29	+3.96
Encoder 4	1.90	4.21	**0.69**	2.26	+2.91
GRU	**3.86**	**4.38**	0.68	**2.97**	+3.71
Op.Unm. [12]	1.92	0.08	**2.53**	1.51	-
Label mult.	2.12	3.46	−1.05	1.51	+5.22
Const. (GRU)	2.55	3.70	0.70	2.31	+3.50

the latent representation, like encoder 4 or GRU, are more robust against wrong input labels than input stages in the first encoder blocks, whose separation results decrease very fast for high percentages of toggled labels. That strengthens the explanation of better separation results by combining additional information with higher or more abstract features. In our case, the best additional information integration stage before encoder 4 improves the separation performance up to a toggled labels rate of 30 %.

Further separation improvements can be achieved in case of a known ensemble line-up by targeting the model output. Exemplarily, the predefined trio of violin, trumpet, and flute is investigated here, so $n = 3$ instrument tracks have to be separated. The averaged separation results for this trio model are listed in Table 2. Moreover, the previously discussed references of label multiplication and constant instrument label input before encoder block 4 are added to the result table. Similar to the separation of 13 instruments, additional information integration improves MSS results for all input stages and the best input stage is before encoder block 4.

Targeted data augmentation leads to an improvement of at least 4.42 dB compared to the original URMP dataset. This is higher than the improvement of the trio separation without additional information of about 4 dB. Thus, the joint separation models can benefit more from the data augmentation of music pieces and their corresponding labels than only from the augmented music pieces. As for the separation model with 13 instruments, the multiplication of the instrument labels with their respective track estimations lead to only marginally improvements. The usage of time-dependent labels improves the separation results by 0.69 dB for the trio model and input stage before encoder block 4. In case of the other input stages, the separation improvements are similar.

B. INDEPENDENT INSTRUMENT MODELS

An alternative to the joint separation model analyzed in section IV-A is the separation system approach with n independent instrument extraction models. The main advantage of this approach is the flexibility of the line-up to be separated. For example, the separation system for the predefined trio of violin, trumpet, and flute, to which the data augmentation is tailored in this work, can easily be created by choosing only the three particular independent models. If a separation system with independent models of all 13 instruments is used, the separation results of the trio instruments remain the same as in the trio case, in contrast to the joint separation approach. The resulting SI-SDR scores for the trio instruments and different input stages of the additional time-dependent instrument activity information are presented in Table 3. Additionally, the results for the literature approach Open-Unmix [12] with time-frequency representation input, a multiplication of time-dependent labels and the estimated instrument tracks from the separation system without additional information, and the results for a static label concatenation before the first GRU block are given as references.

According to Table 3, the separation performance benefits from the time-dependent instrument activity integration at input stages before encoder block 3, encoder block 4, and the first GRU block. Those stages contain complex features with a larger receptive field than earlier stages. These features have also proven to be beneficial in the joint separation approach. Early input stages before encoder block 1 or 2 lead to worse results than without instrument activity (-), especially for the trumpet. That could indicate overfitting to the instrument activity labels in those early stages. Consequently, they are not suitable for an additional information integration. Without data augmentation, the separation results for all input stages are better than those of the approach without additional information. That strengthens the positive effect of instrument activity integration for MSS with small datasets. In case of augmented trio data, all separation performances are improved, therefore data augmentation is useful during training of the independent models.

Similar to the joint separation approach, a multiplication of time-dependent instrument activity labels with its respective estimated tracks leads only to a marginal improvement. The separation by independent models suppresses other instruments and artifacts during pauses of the target source successfully by itself. If the additional label input comprises only a constant label about the instrument presence in the whole music piece, the separation performance is decreased by 0.66 dB in case of the input stage before the first GRU block. For the other input stages, a similar increase is detected using time-dependent additional information. Using time-frequency input matrices like in Open-Unmix, the violin can

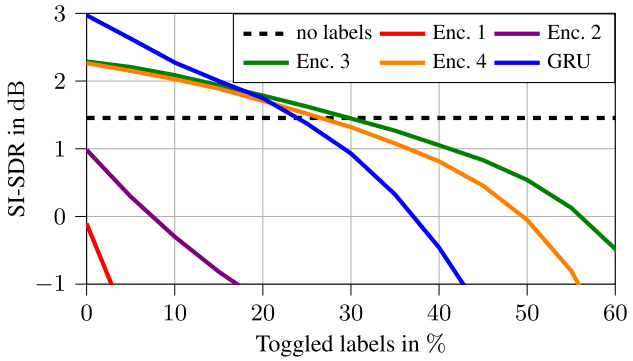

FIGURE 3. SI-SDR results for randomly toggled labels of defined percentage and 3 independent separation models of the defined trio.

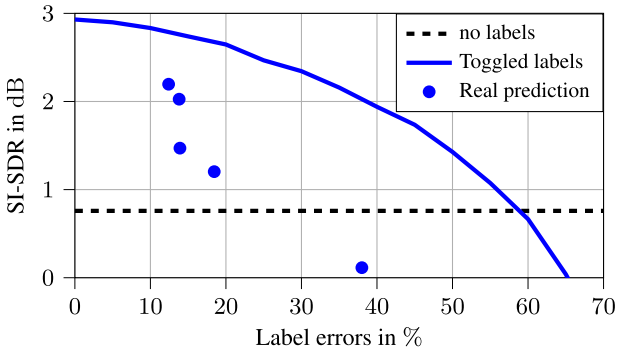

FIGURE 4. SI-SDR results of the joint trio model with input before encoder block 4 for toggled labels and real instrument predictions.

be separated very well, whereas the trumpet separation is very difficult. Since the violin is the predominant instrument in most music pieces, a separation based on time-frequency representations could focus more on the predominant instrument than an appoach in the time domain like ours. However, the average separation result of Open-Unmix is comparable to our separation model without additional label input.

Although the average results for the best separation systems in Table 2 and Table 3 are nearly the same, the SI-SDR scores of the three instruments are very different. The separation of the flute and the violin track seem to include activity information of other instruments because the results are much better for almost all joint separation models than the respective independent ones. In contrast, the SI-SDR values for the trumpet are drastically decreased using additional inputs of other instruments, probably because the independent trumpet model can focus more on the special instrument characteristics. But for rarely present instruments in the training dataset, the activity information about other instruments should in general be useful.

In order to analyze the robustness of the separation systems with independent instrument models, randomly toggled activity labels are fed in the systems. The resulting SI-SDR values for the defined trio are illustrated in Fig. 3 as a function of the percentage of toggled labels during testing. An input of additional information before encoder 1 and 2 remains unsuitable due to bad separation results as well as low robustness against label errors. The input stages before encoder block 3 and encoder block 4 show highest robustness because of the slow SI-SDR score decrease for high percentages of toggled labels. However, the input stage before the first GRU block offers the best separation performance until approximately 20 % of random label errors. It achieves an improved separation performance by additional label integration until approximately 23 % of toggled labels.

C. REAL PREDICTIONS AS ADDITIONAL INFORMATION

In real MSS applications, time-dependent instrument activity is no available information and has to be estimated. Exemplarily, the impact of such real time-dependent label predictions by a pretrained ResNet approach [33] is investigated here. The predictions with a time resolution of about 100 ms are only used during testing in order to use exact labels during supervised training. Depending on the chosen detection thresholds for the predictions, which define the minimum estimation values to detect an instrument as active, different label accuracies can be achieved. All thresholds are selected individually for each instrument. During the experiments, different detection thresholds are investigated. The corresponding separation results (average of test dataset) are illustrated in Fig. 4 for the joint trio model and different label accuracies of real time-dependent instrument activity predictions. As a reference, the separation results for the randomly toggled ground truth inputs are given.

The separation results with real predictions are not as robust as those with randomly toggled labels because real predictions contain coherent sections of false labels. These sections of many time steps result from musical notes not taken into account, unknown or unusual instrument characteristics, unsuitable thresholds, or jittered audio signals. Single label errors due to toggling can be handled relatively easy by means of the surroundings or other instrument activity labels, but larger sections of false labels influence the separation performance drastically. Nevertheless, the separation approach with real time-dependent instrument activity predictions still outperforms the approach without additional information if the instrument detection accuracy is higher than 80 %. With the best instrument detection accuracy of about 87.59 %, an average trio separation performance of 2.20 dB is achieved.

For the separation based on independent instrument models, the influence of label errors on the separation performance can be analyzed isolated for each instrument. The separation results are only affected by the label errors of this instrument. In Fig. 4, the results of the independent flute and violin models are visualized exemplarily for different accuracies of real instrument activity predictions as well as simulated estimation errors by toggled labels. Using the best detection thresholds, the SI-SDR scores of the separated instrument signals are 2.45 dB for flute and 0.06 dB in case of

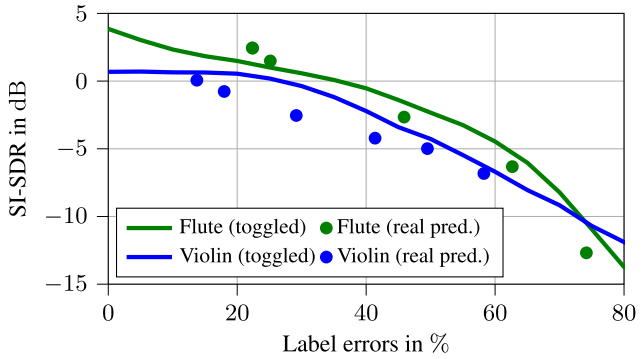

FIGURE 5. SI-SDR results of independent violin and flute model with input before the first GRU block for toggled labels and real instrument predictions.

violin, which are still better than the results without additional information input in Table 3.

In contrast to the joint separation case in Fig. 4, the data points of real label predictions validate the relation between toggled labels and separation performance of flute and violin. The most important reason for this successful validation is the focus on one instrument, which neglects any influence of other instrument labels on the separation results. Between approximately 15 % and 50 % of label errors, the separation performance of violin is a little lower for real predictions than the results for toggled labels. But in general, the separation results for real instrument activity predictions match the curves for toggled labels. Consequently, the robustness of separation systems with independent instrument models can be analyzed by toggled labels to simulate time-dependent instrument activity prediction errors.

V. CONCLUSION

An MSS system for polyphonic chamber music signals is presented in this work. The integration of additional time-dependent instrument activity information improves the separation performance of the basic joint time domain approach. Although it is possible to integrate the additional information in different stages of the U-Net architecture, the integration before the deepest encoder block ensures best separation and highest robustness against randomly toggled instrument labels. Furthermore, an alternative separation architecture with independent instrument models is investigated. This approach enables a more flexible separation line-up and shows best separation results with an integration of additional information before the first GRU block. The separation improvement by additional information is verified for real instrument predictions and both separation architectures.

In future work, the generalization ability of the presented models has to be analyzed by signals with other recording conditions than in the dataset used in our experiments. Moreover, the instrument activity labels can be extended to a larger range of values. Several active instruments of the same instrument category can be labeled by a value higher than 1, like e.g. the first and second violin, which would improve the additional information given to the separation model.

REFERENCES

[1] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep U-Net convolutional networks," in *Proc. Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, Oct. 2017, pp. 745–751.

[2] Z. Rafii, A. Liutkus, F.-R. Stoter, S. I. Mimilakis, D. FitzGerald, and B. Pardo, "An overview of lead and accompaniment separation in music," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 8, pp. 1307–1335, Aug. 2018.

[3] C. Hawthorne, E. Elsen, J. Song, A. Roberts, I. Simon, C. Raffel, J. Engel, S. Oore, and D. Eck, "Onsets and frames: Dual-objective piano transcription," in *Proc. Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, Sep. 2018, pp. 50–57.

[4] L. V. Veire and T. De Bie, "From raw audio to a seamless mix: Creating an automated DJ system for drum and bass," *EURASIP J. Audio, Speech, Music Process.*, vol. 2018, no. 1, pp. 1–21, Dec. 2018.

[5] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "MUSDB18—A corpus for music separation," (1.0.0) [Data set], Zenodo, Dec. 2017. [Online]. Available: https://zenodo.org/record/1117372, doi: 10.5281/zenodo.1117372.

[6] M. Miron, J. Janer, and E. Gómez, "Generating data to train convolutional neural networks for low latency classical music source separation," in *Proc. 14th Sound Music Comput. Conf.*, Jul. 2017, pp. 227–233.

[7] B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma, "Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications," *IEEE Trans. Multimedia*, vol. 21, no. 2, pp. 522–535, Feb. 2019.

[8] P. Chandna, M. Miron, J. Janer, and E. Gómez, "Monoaural audio source separation using deep convolutional neural networks," in *Latent Variable Analysis and Signal Separation*, P. Tichavsky, M. Babaie-Zadeh, O. J. J. Michel, and N. Thirion-Moreau, Eds. Cham, Switzerland: Springer, 2017, pp. 258–266.

[9] E. M. Grais and M. D. Plumbley, "Single channel audio source separation using convolutional denoising autoencoders," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Nov. 2017, pp. 1265–1269.

[10] L. Pandey, A. Kumar, and V. Namboodiri, "Monoaural audio source separation using variational autoencoders," in *Proc. Interspeech*, Sep. 2018, pp. 3489–3493.

[11] S. Uhlich, M. Porcu, F. Giron, M. Enenkl, T. Kemp, N. Takahashi, and Y. Mitsufuji, "Improving music source separation based on deep neural networks through data augmentation and network blending," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 261–265.

[12] F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, "Open-unmix—A reference implementation for music source separation," *J. Open Source Softw.*, vol. 4, no. 41, p. 1667, Sep. 2019.

[13] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, "Spleeter: A fast and efficient music source separation tool with pre-trained models," *J. Open Source Softw.*, vol. 5, no. 50, p. 2154, Jun. 2020.

[14] N. Takahashi and Y. Mitsufuji, "D3Net: Densely connected multidilated DenseNet for music source separation," 2020, *arXiv:2010.01733*.

[15] Q. Kong, Y. Cao, H. Liu, K. Choi, and Y. Wang, "Decoupling magnitude and phase estimation with deep ResUNet for music source separation," in *Proc. Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, Nov. 2021, pp. 342–349.

[16] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Music source separation in the waveform domain," 2019, *arXiv:1911.13254*.

[17] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A multi-scale neural network for end-to-end audio source separation," in *Proc. Int. Soc. Music Inf. Retr. (ISMIR) Conf.*, Sep. 2018, pp. 334–340.

[18] M. Kim, W. Choi, J. Chung, D. Lee, and S. Jung, "KUIELab-MDX-Net: A two-stream neural network for music demixing," in *Proc. ISMIR Workshop Music Source Separat.*, Nov. 2021, pp. 1–7.

[19] A. Défossez, "Hybrid spectrogram and waveform source separation," in *Proc. ISMIR Workshop Music Source Separat.*, Nov. 2021, pp. 1–11.

[20] M. Miron, J. Janer, and E. Gómez, "Monaural score-informed source separation for classical music using convolutional neural networks," in *Proc. Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, Oct. 2017, pp. 55–62.

[21] M. Kawamura, T. Nakamura, D. Kitamura, H. Saruwatari, Y. Takahashi, and K. Kondo, "Differentiable digital signal processing mixture model for synthesis parameter extraction from mixture of harmonic sounds," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 941–945.

[22] E. Manilow, P. Seetharaman, and B. Pardo, "Simultaneous separation and transcription of mixtures with multiple polyphonic and percussive instruments," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 771–775.
[23] Y.-N. Hung and A. Lerch, "Multitask learning for instrument activation aware music source separation," in *Proc. Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, Oct. 2020, pp. 748–755.
[24] O. Slizovskaia, L. Kim, G. Haro, and E. Gomez, "End-to-end sound source separation conditioned on instrument labels," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 306–310.
[25] G. Meseguer-Brocal and G. Peeters, "Conditioned-U-Net: Introducing a control mechanism in the U-Net for multiple source separations," in *Proc. Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, Nov. 2019, pp. 159–165.
[26] W. Choi, M. Kim, J. Chung, and S. Jung, "LaSAFT: Latent source attentive frequency transformation for conditioned source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 171–175.
[27] E. Manilow, G. Wichern, and J. L. Roux, "Hierarchical musical instrument separation," in *Proc. Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, Oct. 2020, pp. 376–383.
[28] O. Slizovskaia, G. Haro, and E. Gomez, "Conditioned source separation for musical instrument performances," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 2083–2095, 2021.
[29] L. Lin, Q. Kong, J. Jiang, and G. Xia, "A unified model for zero-shot music source separation, transcription and synthesis," in *Proc. Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, Nov. 2021, pp. 381–388.
[30] G. Meseguer-Brocal and G. Peeters, "Content based singing voice source separation via strong conditioning using aligned phonemes," in *Proc. Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, Oct. 2020, pp. 819–827.
[31] G. Cantisani, A. Ozerov, S. Essid, and G. Richard, "User-guided one-shot deep model adaptation for music source separation," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, Oct. 2021, pp. 111–115.
[32] J.-Y. Liu and Y.-H. Yang, "Dilated convolution with dilated GRU for music source separation," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, Jul. 2019, pp. 4718–4724.
[33] M. Schwabe and M. Heizmann, "Influence of input data representations for time-dependent instrument recognition," *Tm Technisches Messen*, vol. 88, no. 5, pp. 274–281, Feb. 2021.
[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[35] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR—Half-baked or well done?" in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 626–630.

MARKUS SCHWABE received the B.Sc. and M.Sc. degrees in electrical engineering and information technology from the Karlsruhe Institute of Technology (KIT), Germany, in 2014 and 2016, respectively. He is currently a Research Associate with the Institute of Industrial Information Technology (IIIT), KIT. His research interests include signal and audio processing, music signal analysis, and machine learning.

MICHAEL HEIZMANN received the M.Sc. degree in mechanical engineering and the Ph.D. degree in automated visual inspection from the University of Karlsruhe, Germany, in 1998 and 2004, respectively. From 2004 to 2009, he was a Postdoctoral Research Assistant with the Fraunhofer IOSB, Karlsruhe, Germany, where he was the Head of the Department Systems for Measurement, Control and Diagnosis, from 2009 to 2016. From 2014 to 2016, he was a Professor in mechatronic systems with the Karlsruhe University of Applied Sciences. Since 2016, he has been a Full Professor in mechatronic measurement systems and the Director of the Institute of Industrial Information Technology, Karlsruhe Institute of Technology. His research interests include machine vision, image processing, image and information fusion, measurement technology, machine learning, artificial intelligence, and their applications.