MULTITASK LEARNING FOR INSTRUMENT ACTIVATION AWARE MUSIC SOURCE SEPARATION

Yun-Ning Hung
Center for Music Technology
Georgia Institute of Technology
`amyhung@gatech.edu`

Alexander Lerch
Center for Music Technology
Georgia Institute of Technology
`alexander.lerch@gatech.edu`

ABSTRACT

Music source separation is a core task in music information retrieval which has seen a dramatic improvement in the past years. Nevertheless, most of the existing systems focus exclusively on the problem of source separation itself and ignore the utilization of other —possibly related— MIR tasks which could lead to additional quality gains. In this work, we propose a novel multitask structure to investigate using instrument activation information to improve source separation performance. Furthermore, we investigate our system on six independent instruments, a more realistic scenario than the three instruments included in the widely-used MUSDB dataset, by leveraging a combination of the MedleyDB and Mixing Secrets datasets. The results show that our proposed multitask model outperforms the baseline Open-Unmix model on the mixture of Mixing Secrets and MedleyDB dataset while maintaining comparable performance on the MUSDB dataset.

1. INTRODUCTION

Music source separation has long been an important task for Music Information Retrieval (MIR) with numerous practical applications. By isolating the sound of individual instruments from a mixture of music, source separation systems can be used, for example, as a pre-processing tool for music transcription [22] or for audio remixing [37]. They also enable special applications such as the automatic generation of karaoke tracks by separating vocals from the accompaniment, stereo-to-surround upmixing, and instrument-wise equalization [1, 24].

Most of the current source separation systems use deep learning approaches to estimate a spectral mask for each independent instrument, then apply the mask to the mixture audio for separation. Although the utilization of deep learning has improved source separation performance dramatically, one problem of this approach is the limited amount of training data for the prevalent supervised learning approaches. More specifically, the datasets need to comprise

of the separated tracks of each instrument, which renders most easily accessible music data useless as it is already mixed. Multiple open-source datasets attempt to address this issue [5, 11, 17]. MUSDB [23] is nowadays the most frequently used dataset for the training and evaluation of source separation systems. In addition to the limited size, the main shortcoming of MUSDB is the limited number of instrument tracks: 'Bass,' 'Drums,' 'Vocals,' and 'Other.' Other datasets show other drawbacks, for instance, the iKala and MIR-1K datasets only contain short clips of music instead of complete songs [5, 11].

In addition to the data challenge, one potential issue with most existing music source separation systems is that they exclusively focus on the source separation task itself. Harnessing the information of other MIR tasks by incorporating them into source separation, however, has not been explored in-depth. For example, Instrument Activation Detection (IAD) can help determine which time frame contains the target instrument, while pitch detection can help determine which frequency bins are more likely to contain a harmonic series [13, 19]. This kind of multitask learning approach has been reported to be efficient for multiple other MIR tasks. Böck et al. achieve state-of-the-art performance for both tempo estimation and beat tracking by learning these two tasks at the same time [4]. Bittner et al. show that by estimating multi-f0, melody, bass line, and vocals at the same time, the system outperforms its single-task counterparts on all four tasks [2]. Similar results have been reported for simultaneously estimating score, instrument activation, and multi-f0 [12]. However, only recently was multitask learning successfully applied to source separation by combining it with pitch estimation [28].

In this paper, we propose a novel multitask learning structure to explore the combination of IAD and music source separation. By training for both tasks in an end-to-end manner, the estimated instrument labels can be used during inference as a weight for each time frame. The goal is both to suppress the frames not containing the target instrument and to correct a potentially incorrectly estimated mask. To increase the size of the available training data, we leverage two open-source large-scale multi-track datasets (MedleyDB [3] and Mixing Secrets [8]) in addition to MUSDB to evaluate on a larger variety of separable instruments. We refer to the combination of these two datasets as the *MM dataset*.

In summary, the main contributions of this work are

- the systematic investigation of the first multi-task source separation deep learning model that incorporates source separation with IAD in the spectral domain,

- the application of the IAD predictions during inference, and

- the presentation of the first open-source model that separates up to six instruments instead of the four tracks (3 independent instruments) featured by MUSDB.

2. RELATED WORK

State-of-the-art systems for music source separation are all based on deep learning due to proven superior performance. Uhlich et al. presented one of the pioneering works using a Deep Neural Network (DNN) architecture for music source separation [35], and Nugraha et al. used a DNN architecture and fully-connected layers for multichannel music source separation [21]. In the following years, more deep learning related systems were introduced. For example, Takahashi and Mitsufuji used recurrent neural networks to deal with temporal information [36], while others proposed the U-net structure for multiple separation tasks [14]. The U-net structure had been previously found useful for image segmentation [26] and treats the decomposition of a musical audio signal into the target and accompanying instrument tracks analogous to image-to-image translation. Takahashi et al. presented a dense LSTM that achieved the highest score in the SiSEC2018 [31] competition [34]. To preserve high resolution information, Liu and Yang introduced dilated 1-D convolution and a GRU network to replace pooling [16]. Different from other approaches using spectrograms as the input representation, Défossez et al. experimented on time-domain waveform source separation and showed that results comparable to spectrogram-based source separation systems are achievable [6]. "Spleeter," based on a U-net model structure, is currently regarded as one of the most powerful source separation systems [10]. It should be noted that —although the pre-trained model is freely available— Spleeter is trained on a proprietary, publicly unavailable dataset.

Stöter et al.'s "Open-Unmix" is frequently used as modern benchmark system on the MUSDB dataset [23]. It is a well documented open-source music source separation system with a recurrent architecture that achieves good separation results [32].

Most of the methods mentioned above are trained and evaluated on the open-source dataset used in SiSEC2018 competition [31]: MUSDB [23]. As mentioned above, one of the main problems of the MUSDB dataset is that it has only a limited amount of songs and instrument categories: it only includes three separable independent instruments. To include more separable instruments, Miron et al. proposed a score-informed system able to separate four classical instruments by training it on synthetic renditions [19]. However, their system is limited to classical music and requires the musical score for separation. While Spleeter is able to separate four independent instruments and Uhlich et al.
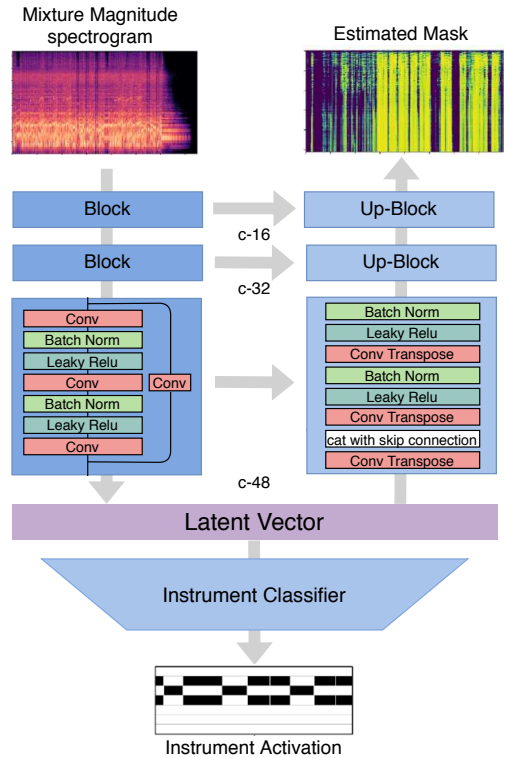

Figure 1. Multitask model structure for our proposed source separation system. Block is the residual block composed of convolutional layers while Up-Block is the residual block composed of transposed convolutional layers. "c" is the number of features.

constructed a dataset which contains nine separable instruments [35], both their datasets are not publicly available.

To explore the possibility of separating unseen instruments, Seetharaman et al. proposed to use instrument class labels as a condition to cluster time-frequency bins for different instruments into the embedding space [28]. Their work showed that the system can also separate unseen instruments during testing by sampling from the learned embedding space. Lee et al. proposed to use audio queries for music separation [15]. The learned feature vector from the audio query acts as the condition to inform the separation of the target source. Manilow et al. introduced a multitask learning structure for source separation, instrument classification, and music transcription [18]. They showed that by jointly learning these three tasks, the source separation quality increased. Furthermore, the network seems to generalize better to unseen instruments. Other works such as [29, 33] combine instrument activity with source separation, however, they utilized either the predicted activity or ground truth as an input condition instead of learning in an end-to-end manner.

3. METHOD

We propose a U-net-based [27] multitask structure to incorporate instrument recognition with music source separation, which we refer to as Instrument Aware Source Separation (IASS) system. An overview of the model is shown in
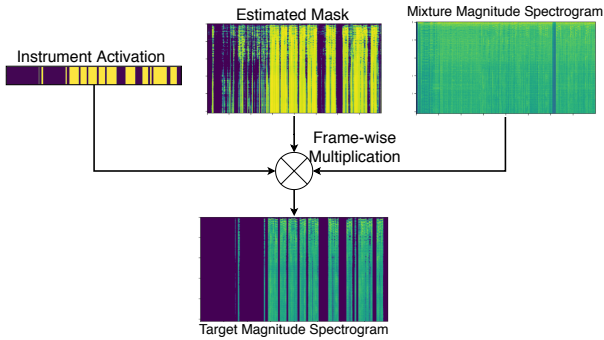
Figure 2. Using instrument activation as a weight to filter the estimated mask, which will be used to multiply with the mixture of the magnitude spectrogram.

Figure 1. Although the multitask approach shows similarities with previous approaches (compare [18]), we design our model with a different goal: instead of just learning a joint representation using the multitask structure, our model uses estimated labels from multitask learning during inference to improve source separation estimation.

3.1 Model structure

The U-net structure has been found useful for image decomposition [26], a task with general similarities to source separation. The skip connections of U-net enable the model to learn from both high-level and low-level features leading to its success for music source separation [10, 14, 30].

Our model differs from previous U-net-based source separation systems by using a residual block instead of a CNN in each layer. The residual block allows the information from the current layer to be fed into a layer 2 hops away and deepens the structure. Each encoder and decoder contains three blocks with each block containing three convolutional or transposed convolutional layers, respectively, two batch normalization layers, and two leaky ReLU layers. The multitask objective is achieved by attaching a CNN classifier to the latent vector. This classifier predicts the instrument activity and has four transposed convolutional layers and three batch normalization layers in between. The last convolutional and transposed convolutional layers in each block feature a (3,1) filter size for up-sampling or down-sampling, while the others have a (3,3) filter size. During training, we use a Mean Square Error (MSE) loss for source separation and a Binary Cross-Entropy (BCE) loss for the prediction of the instrument activity. A hyperparameter α is manually tuned to balance these two loss functions:

$$L = L_{\mathrm{MSE}} + \alpha L_{\mathrm{BCE}}. \tag{1}$$

After successful training, the predicted instrument activity is used as a binary weight to multiply with the magnitude spectrogram along the time dimension. By doing so, the instrument labels are able to suppress the frames not containing any target instrument as shown in Figure 2. However, this binary instrument mask has two potential problems. First, false negatives of the predicted labels
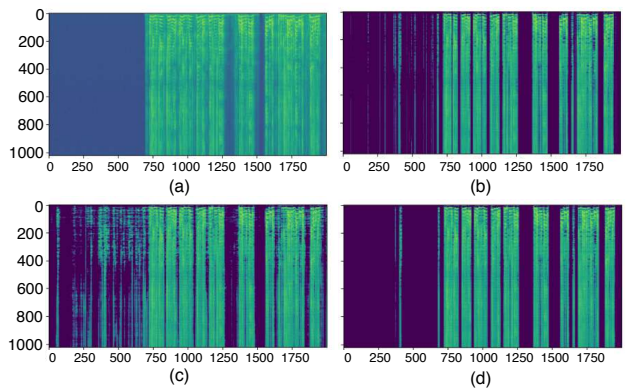

Figure 3. First 2000 frames of the separated vocal track from the song *Angels In Amplifiers — I'm Alright* from the MUSDB-HQ dataset, visualizing different post-processing methods for applying instrument activity as a weight on the predicted magnitude spectrogram: (a) ground truth spectrogram, (b) predicted spectrogram without post-processing, (c) predicted spectrogram with raw predicted instrument labels as a weight, (d) predicted spectrogram with smoothed predicted instrument labels.

might mistakenly suppress wanted components in the spectrogram. Figure 3 (b) exemplifies that around frames 800 and 1000 with gaps caused by false negative prediction within a continuous sound. Even if the gaps have only a the length of a few milliseconds, it will have negative impact on the perceived quality. Second, the binary mask might cause repeated abrupt switching between silence and sound, which might lead to artifacts such as musical noise further decreasing the perceived quality. To address these problems, a median filter is applied to smooth the predicted instrument activities. The influence is discussed in Sect. 4.

An implementation of our system is publicly available online. [1]

3.2 Data representation

We extract magnitude spectrograms with a window length and hop size of 4096 and 1024 samples, respectively, at a sample rate of 44100 Hz for the input of our source separation model. The same magnitude spectrogram is extracted for the target audio reference. The instrument activity ground truth is at the frame level, meaning there is a binary label for each instrument to show whether the instrument is active or not in each time frame. Instrument labels have the same time resolution as the input spectrogram. We use the original activation probability computed by both datasets [3, 8], and binarize the activation with a threshold of 0.5 as suggested.

4. EXPERIMENT

To show the efficiency of our proposed model, we first compare our model with the baseline Open-Unmix model [32] on the MUSDB-HQ dataset. Note that we choose

---

[1] https://biboamy.github.io/Source_Separation_Inst

Method	Vocals	Bass	Drums	Other
IASS 3 blocks	6.46	4.18	5.56	4.19
IASS 4 blocks	6.51	4.25	5.15	4.38

Table 1. SDR score for IASS source separation performance with 3 and 4 residual blocks.

	Method	SDR	SIR	SAR	ISR
Vocals	Open-Unmix	6.11	13.21	6.75	12.43
	IASS	**6.46**	**14.70**	**6.98**	**14.30**
Bass	Open-Unmix	**4.48**	**8.23**	**5.40**	**10.29**
	IASS	4.18	7.30	4.52	6.85
Drums	Open-Unmix	5.02	10.17	6.05	10.55
	IASS	**5.56**	**10.74**	**6.86**	**10.92**
Other	Open-Unmix	**4.23**	**9.90**	3.88	7.34
	IASS	4.19	8.78	**4.70**	**9.32**

Table 2. BSS metrics for Open-Unmix and IASS on the MUSDB-HQ dataset.

MUSDB-HQ instead of MUSDB because we want to obtain a high-quality separation system without potential coding artifacts. The audio of MUSDB is encoded in a lossy format while MUSDB-HQ provides the raw audio data. Other than that, there is no difference between the two datasets. However, since MUSDB-HQ is a newly released dataset it complicates comparing our results to other approaches directly as most of the previous systems have not been evaluated on the MUSDB-HQ dataset. Both the baseline and the proposed method are then evaluated on the MM dataset with six different instruments. For each source, a separate model is trained for both the baseline and our proposed method. Finally, an ablation study is conducted to investigate the impact of instrument labels and median filter on the source separation results. We train our models with the Adam optimizer with a learning rate of 0.001 and apply early stopping if the validation loss does not change for 100 epochs.

4.1 Dataset

Two open-source datasets, MUSDB-HQ [25] and the combination of Mixing Secrets [8] and MedleyDB [3] dataset (MM dataset) are used for the experiments. MUSDB is the most widely used dataset for music source separation and contains four separated tracks: 'Bass,' 'Drums,' 'Vocals,' and 'Others.' The dataset has 150 full-length stereo music tracks. We use the data split proposed by Stöter [32]: 86/14/50 songs for training, validation, and testing, respectively. Since data augmentation has been proven to be helpful [34], the following data augmentation is applied during the training. First, one track is randomly selected from each source and multiplied with a random gain factor ranging from 0.25 to 1.25. The starting point of each track is randomly chosen for chunking into a clip with length of 6 s. Finally, the chunked audio clips from each source are remixed for training. Since the original MUSDB-HQ dataset does not include instrument activation labels, we apply the energy tracking method proposed for MedleyDB [3] with a threshold of 0.5 to obtain the frame-level binary instrument activity labels.

The MM dataset contains 585 pieces of songs (330 from MedleyDB and 258 from Mixing Secrets) with more than 100 instruments and their individual tracks. We use the training and testing split proposed by Gururani et al. [9] for training (488 songs) and evaluating (100 songs) our system. The most frequently occurring 6 instruments are picked as target instruments: 'Bass,' 'Drums,' 'Vocals,' 'Electrical Guitar,' 'Acoustic Guitar,' and 'Piano.' One of the problems with this dataset is that not all of the songs provide parameters on how to remix the individual tracks into the mixture. Therefore, the volume of each track is adjusted

to the same loudness (RMS) during training before applying the random gain as detailed above. In addition, each track is downmixed to a single channel. Furthermore, the data augmentation technique introduced above is applied to generate a large number of training samples from the MM dataset. We construct two separate groups of songs. One contains all the tracks including the target instrument, while another contains the tracks without the target instrument ("accompaniments"). There are total of 128 target tracks for 'Acoustic Guitar,' 189 for 'Piano,' 325 for 'Electrical Guitar,' 374 for 'Vocals,' 468 for 'Drums,' and 458 for 'Bass.' During training, we randomly select 1 to 5 tracks in the accompaniment pool to mix with the target instrument. By doing so, we can generate various combinations of training "songs." The random chunking approach applied to MUSDB-HQ is also applied on MM dataset. The testing set also balances the loudness of each track. We filter out the songs from the testing set which do not contain any of the 6 target instruments, resulting in 20 songs with 'Piano,' 23 songs with 'Acoustic Guitar,' 54 songs with 'Electrical Guitar,' 71 songs with 'Vocals,' 76 songs with 'Bass,' and 81 songs with 'Drums.'

4.2 Evaluation

To reconstruct the waveforms from the resulting magnitude spectrograms, we multiply the magnitude spectrogram with the phase of the original complex spectrograms and apply the inverse short-time Fourier transform on the complex spectrogram. We do not use any post-processing such as Wiener filtering here to focus on the raw result without potentially confounding quality gains in post-processing. Therefore, the Open-Unmix post-processing is disabled for the evaluation.

The quality of the source separation is evaluated with the four most frequently used objective metrics: source to distortion ratio (SDR), source to interference ratio (SIR), source to artifact ratio (SAR), and Image to Spatial distortion Ratio (ISR) [7]. We use the *museval* package for calculating the evaluation metrics [31].

4.2.1 Source separation on MUSDB-HQ

To allow for comparison with other systems trained on the MUSDB, the first experiment reports the result on MUSDB-HQ.

	Open-Unmix	IASS	IBM	input-SDR
Vocals	3.68	4.78	6.49	-6.24
Elecgtr	1.55	1.77	4.56	-5.90
Acgtr	0.95	1.29	3.38	-6.65
Piano	1.08	1.91	3.63	-6.31
Bass	4.04	5.26	5.34	-5.77
Drums	4.45	4.89	6.23	-6.05

Table 3. SDR score for Open-Unmix, IASS, an ideal binary mask and input-SDR.

In a first preliminary experiment we investigate whether adding more residual blocks influences the performance of the proposed method. We report the SDR score for the four sources of the MUSDB-HQ dataset in Table 1. The performance with three residual blocks and with four residual blocks is comparable on all instruments. For training efficiency, we use three residual blocks in the following experiment.

Table 2 shows the results of our proposed model compared to Open-Unmix. Our model outperforms the Open-Unmix model on 'Vocals' and 'Drums', performs equally on 'Other', and slightly worse on 'Bass'. This might be because 'Bass' most likely to appear throughout the songs. As a result, the improvement of using the instrument activation weight is limited. The imbalanced activity might also impact the instrument classifier.

4.2.2 Source separation on MM

The results for the MM dataset are summarized in Table 3. We re-trained the Open-Unmix model on the MM dataset by using the default training setting provided with their code. The results for the Ideal Binary Mask (IBM) (source code [31]) represent the best case scenario. The worst case scenario is represented by the results for input-SDR, which is the SDR score when using mixture as the input. Compared to MUSDB-HQ, the MM dataset has a larger amount of training data. It can be observed from Table 3 that our proposed model generally achieves better source separation performance on six instruments. We can also observe a trend that both models have higher scores on 'Drums,' 'Bass,' and 'Vocals' than on 'Electrical Guitar,' 'Piano,' and 'Acoustic Guitar.' This might be attributed to the fact that 'Guitar,' 'Piano,' and 'Acoustic Guitar' have fewer training samples (cf. Sect. 4.1). Another possible reason is that the more complicated spectral structure of polyphonic instruments such as 'Guitar' and 'Piano' make the separation task more challenging.

4.2.3 Instrument activity detection

While our system's source separation performance was the primary concern, the accuracy of the instrument predictions is also of interest. Our classifier output is compared to the model proposed by Gururani et al. [9], which was trained and evaluated on the same MM dataset. Note that this comparison is still not completely valid as their system uses multi-label prediction while our model is single-label.
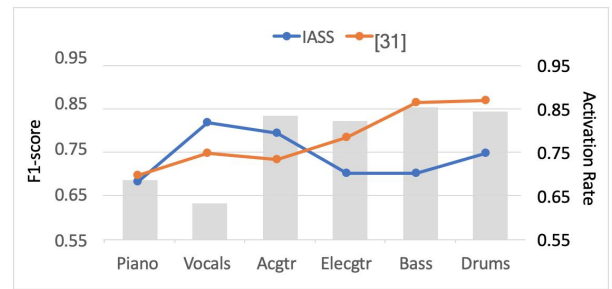

Figure 4. IAD result for both Gururani et al.'s method (orange) and our IASS (blue) with label aggregation. Indicated in gray is the activation rate (percentage of the training frames containing positive activity labels).

Still, it can provide some insights into how well our system predicts the instrument activity. As Gururani et al.'s system predicts instrument labels with a time resolution of 1 s, the output resolution of our prediction has to be reduced. For each second, all the estimated activations are aggregated by calculating their median. Furthermore, instrument subcategories from their work are combined. For example, female and male singers are combined into 'Vocals,' electrical bass and double bass are combined into 'Bass,' electrical and acoustic piano are combined into 'Piano' and clean and distorted electrical guitar are combined into 'Electrical Guitar.' We report the AUC score in Figure 4.

We can make the following observations. First, 'Piano,' 'Electrical Guitar,' and 'Bass' tend to have lower detection rates. This might be because all these instrument categories include both acoustic and electric instruments which the model might easily confuse with the background music. This might also influence the source separation performance. The result can explain that in Table 4, 'Vocals' have the highest increase in the average score when applying instrument labels since 'Vocals' has better instrument detection accuracy. In contrast, 'Bass' has a lower increase since it has poorer instrument detection results. Second, from Figure 4 we can observe that 'Vocals' and 'Piano' have a lower activation rate, which means the model has fewer sound samples containing 'Vocals' and 'Piano' during training. This aligns with the highest SIR score increase on 'Vocals' and 'Piano' in Table 4 when instrument activation is added, since instrument activation can help suppress the interference at the non-active frames. This also shows the potential of our model to be used on instruments which only appear in the song infrequently.

4.3 Ablation study

In this experiment, we investigate the impact of the instrument labels on our model. First, the IASS model is trained and evaluated without using instrument labels, i.e., as a standard U-net without instrument classifier on the latent vector. The model will only be updated by the MSE between the ground-truth magnitude spectrogram and predicted spectrogram ($\alpha = 0$). For testing, all instrument "predictions" are set to 1. Second, the IASS model is trained

	Train	Test	SDR	SIR	SAR	Avg
Vocals	✗	✗	4.26	8.58	4.48	5.77
	✓	✗	3.94	8.48	4.69	5.70
	✓	✓	**4.78**	**11.62**	**5.31**	**7.24**
Elecgtr	✗	✗	1.75	0.61	**4.94**	**2.46**
	✓	✗	**1.82**	1.27	4.29	2.43
	✓	✓	1.77	**1.64**	4.48	2.63
Acgtr	✗	✗	1.11	0.75	2.42	1.43
	✓	✗	1.15	0.48	**2.52**	1.38
	✓	✓	**1.29**	**1.80**	2.45	**1.85**
Piano	✗	✗	1.55	2.97	2.13	2.31
	✓	✗	1.70	3.16	2.06	2.22
	✓	✓	**1.91**	**4.17**	2.15	**2.74**
Bass	✗	✗	4.10	**8.34**	4.74	5.72
	✓	✗	4.12	7.82	5.10	5.68
	✓	✓	**4.34**	8.13	5.10	**5.85**
Drums	✗	✗	4.50	9.54	5.15	6.40
	✓	✗	4.38	9.87	4.95	6.40
	✓	✓	**4.89**	**10.72**	**5.26**	**6.96**

Table 4. Ablation study for IASS source separation performance training and evaluating with (✓) or without (✗) instrument labels.

with instrument labels but evaluated without instrument labels. This is a traditional multitask scenario: the model will be trained with both the MSE and the BCE losses in Eq. (1). However, during evaluation, the output magnitude spectrogram is not weighted by the instrument activity (predictions equal 1). Third, we include the IASS results from Table 3 —computed with both losses and using the instrument predictions as mask weights— for convenience.

The results are shown in Table 4. It can be observed that using instrument labels as a weight generally leads to a better performance than without using instrument labels. The result also somewhat unexpectedly shows that training with instrument detection loss influences source separation performance, as the average quality score is often lower when training with the multitask loss. One possible reason for this is that adding the IAD sub-task forces more information to be passed to the bottom layers, where the resolution is compressed. We argue, however, that the multitask learning structure does bring an extra benefit to the system: using the instrument activity predictions as a weight leads to better separation quality. Figure 3 visualizes the effect on one of the songs from the MUSDB-HQ dataset. This song does not have any vocals before 16 s, which is around time frame 700. Subfigure (a) shows the ground truth magnitude spectrogram before applying instrument labels while (c) and (d) show the predicted spectrograms after applying the instrument activations or the smoothed instrument activations, respectively. Both the false positive predictions in the beginning before time frame 700 as well as the false negative predictions around frames 800 and 1000 have been repaired by using smoothed activations. This result is consistent with the results in Table 4 where SIR has the highest increase: interferences are more successfully suppressed.

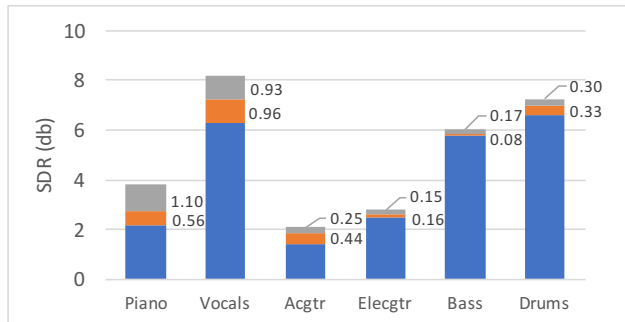Furthermore, we investigate the influence of median

Figure 5. Ablation study for IASS source separation performance with or without median filter on instrument labels. The orange bar shows the increase of the average score (average of SDR, SIR, and SAR) after applying median filtering. Gray shows the average score increase when using instrument ground truth labels instead of estimated labels.

filtering the predicted instrument activity on our results. Figure 5 shows the performance of our proposed source separation model with and without applying the median filter on the predicted instrument activities. Using the median filter generally increases performance across all instruments as it eliminates spurious prediction errors.

Finally we are using the oracle ground truth labels instead of the estimated labels as the weight. As we can observe from Figure 5, using the ground truth labels brings an average score increase in all instruments, especially for vocals and piano. This can be seen as the upper-bound best case scenario of our instrument-activity-weighted model and emphasizes the potential for improvement when combining instrument prediction with source separation.

5. CONCLUSION

In this paper, we proposed a novel multitask structure combining instrument activation detection with multi-instrument source separation. We utilize a large dataset to evaluate on various instruments and show that our model achieves equal or better separation quality than the baseline Open-Unmix model. The ablation study also shows that using instrument activation as a weight is able to correct the false estimation from the source separation task and improve source separation performance. In summary, the main contributions of this work are the proposal of a multitask learning structure combining IAD with source separation, and insights into using new open-source datasets to increase the number of separable instrument categories.

We have identified several directions for future extensions of this model. First, we plan to increase the number of target instruments by combining synthesized data with the MM dataset especially for underrepresented instrument classes. Second, we plan to incorporate other tasks, such as multi-pitch estimation, into our current multi-task structure [20]. Third, we will explore using multi-label instrument detection to separate multiple instruments at the same time. Lastly, we will explore post-processing methods such as Wiener filter to improve our system's quality.

6. ACKNOWLEDGMENT

7. REFERENCES

[1] Tülay Adali, Christian Jutten, Arie Yeredor, Andrzej Cichocki, and Eric Moreau. From raw audio to a seamless mix: creating an automated DJ system for Drum and Bass. *IEEE Signal Processing Magazine*, pages 16–17, 2014.

[2] Rachel M. Bittner, Brian McFee, and Juan Pablo Bello. Multitask learning for fundamental frequency estimation in music. *arXiv preprint arXiv:1809.00381*, 2018.

[3] Rachel M. Bittner, Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam, and Juan Pablo Bello. Medleydb: A multitrack dataset for annotation-intensive mir research. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 155–160, 2014.

[4] Sebastian Böck, Matthew E. P. Davies, and Peter Knees. Multi-task learning of tempo and beat: Learning one to improve the other. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 569–576, 2019.

[5] Tak-Shing Chan, Tzu-Chun Yeh, Zhe-Cheng Fan, Hung-Wei Chen, Li Su, Yi-Hsuan Yang, and Jyh-Shing Roger Jang. Vocal activity informed singing voice separation with the ikala dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 718–722, 2015.

[6] Alexandre Défossez, Nicolas Usunier, Léon Bottou, and Francis Bach. Music source separation in the waveform domain. *arXiv preprint arXiv:1911.13254*, 2019.

[7] Cédric Févotte, Rémi Gribonval, and Emmanuel Vincent. Bss_eval toolbox user guide – revision 2.0. In *IRISA Technical Report 1706, Rennes, France*, 2005.

[8] Siddharth Gururani and Alexander Lerch. Mixing secrets: a multi-track dataset for instrument recognition in polyphonic music. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2017. Late-breaking paper.

[9] Siddharth Gururani, Cameron Summers, and Alexander Lerch. Instrument activity detection in polyphonic music using deep neural networks. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 569–576, 2018.

[10] Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam. Spleeter: A fast and state-of-the art music source separation tool with pre-trained models. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2018. Late-breaking paper.

[11] Chao-Ling Hsu and Jyh-Shing Roger Jang. On the improvement of singing voice separation for monaural recordings using the mir-1k dataset. *IEEE Transactions on Audio, Speech, and Language Processing*, 18:310–319, 2010.

[12] Yun-Ning Hung, Yi-An Chen, and Yi-Hsuan Yang. Multitask learning for frame-level instrument recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 381–385, 2019.

[13] Yun-Ning Hung and Yi-Hsuan Yang. Frame-level instrument recognition by timbre and pitch. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 569–576, 2018.

[14] Andreas Jansson, Eric Humphrey, Nicola Montecchio, Rachel Bittner, Aparna Kumar, and Tillman Weyde. Singing voice separation with deep U-Net convolutional networks. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, page 323–332, 2017.

[15] Jie Hwan Lee, Hyeong-Seok Choi, and Kyogu Lee. Audio query-based music source separation. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 878–885, 2019.

[16] Jen-Yu Liu and Yi-Hsuan Yang. Dilated convolution with dilated gru for music source separation. In *International Joint Conferences on Artificial Intelligence*, pages 4718–4724, 2019.

[17] Antoine Liutkus, Fabian-Robert Stöter, Zafar Rafii, Daichi Kitamura, Bertrand Rivet, Nobutaka Ito, Nobutaka Ono, and Julie Fontecave. The 2016 signal separation evaluation campaign. In *International conference on latent variable analysis and signal separation*, pages 323–332, 2017.

[18] Ethan Manilow, Prem Seetharaman, and Bryan Pardo. Simultaneous separation and transcription of mixtures with multiple polyphonic and percussive instruments. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 771–775, 2020.

[19] Marius Miron, Jordi Janer, and Emilia Gómez. Monaural score-informed source separation for classical music using convolutional neural networks. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 569–576, 2017.

[20] Tomoyasu Nakano, Kazuyoshi Yoshii, Yiming Wu, Ryo Nishikimi, Kin Wah Edward Lin, and Masataka Goto. Joint singing pitch estimation and voice separation based on a neural harmonic structure renderer. *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 160–164, 2019.

[21] Aditya Arie Nugraha, Antoine Liutkus, and Emmanuel Vincent. Multichannel music separation with deep neural networks. In *European Signal Processing Conference (EUSIPCO)*, pages 1748–1752, 2016.

[22] Jouni Paulus and Tuomas Virtanen. Drum transcription with non-negative spectrogram factorisation. In *European Signal Processing Conference*, pages 1–4, 2005.

[23] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. The MUSDB18 corpus for music separation, December 2017. [Online] https://doi.org/10.5281/zenodo.1117372.

[24] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stoter, Stylianos Ioannis Mimilakis, Derry FitzGerald, and Bryan Pardo. An overview of lead and accompaniment separation in music. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 26(8):1307–1335, 2018.

[25] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. MUSDB18-HQ - an uncompressed version of MUSDB18, August 2019.

[26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, 2015.

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241, 2015.

[28] Prem Seetharaman, Gordon Wichern, Shrikant Venkataramani, and Jonathan Le Roux. Class-conditional embeddings for music source separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 301–305, 2019.

[29] Olga Slizovskaia, Leo Kim, Gloria Haro, and Emilia Gomez. End-to-end sound source separation conditioned on instrument labels. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 306–310, 2019.

[30] Daniel Stoller, Sebastian Ewert, and Simon Dixon. Wave-u-net: A multi-scale neural network for end-to-end audio source separation. *arXiv preprint arXiv:1806.03185*, 2018.

[31] Fabian-Robert Stöter, Antoine Liutkus, and Nobutaka Ito. The 2018 signal separation evaluation campaign. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 293–305, 2018.

[32] Fabian-Robert Stöter, Stefan Uhlich, Antoine Liutkus, and Yuki Mitsufuji. Open-unmix-a reference implementation for music source separation. *Journal of Open Source Software*, 2019.

[33] Rupak Swaminathan and Alexander Lerch. Improving singing voice separation using attribute-aware deep network. *International Workshop on Multilayer Music Representation and Processing (MMRP)*, pages 60–65, 2019.

[34] Naoya Takahashi, Nabarun Goswami, and Yuki Mitsufuji. Mmdenselstm: An efficient combination of convolutional and recurrent neural networks for audio source separation. In *International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 106–110, 2018.

[35] Stefan Uhlich, Franck Giron, and Yuki Mitsufuji. Deep neural network based instrument extraction from music. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2135–2139, 2015.

[36] Stefan Uhlich, Marcello Porcu, Franck Giron, Michael Enenkl, Thomas Kemp, Naoya Takahashi, and Yuki Mitsufuji. Improving music source separation based on deep neural networks through data augmentation and network blending. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 261–265, 2017.

[37] Len Vande Veire and Tijl De Bie. From raw audio to a seamless mix: creating an automated dj system for drum and bass. *EURASIP Journal on Audio, Speech, and Music Processing*, 2018(1):13, 2018.