Multi-channel U-Net for Music Source Separation

Venkatesh S. Kadandale*, Juan F. Montesinos*, Gloria Haro*, Emilia Gómez*†
* Department of Information and Communications Technologies, Universitat Pompeu Fabra, Barcelona, Spain
† Joint Research Centre, European Commission, Seville, Spain
{venkatesh.kadandale, juanfelipe.montesinos, gloria.haro, emilia.gomez}@upf.edu

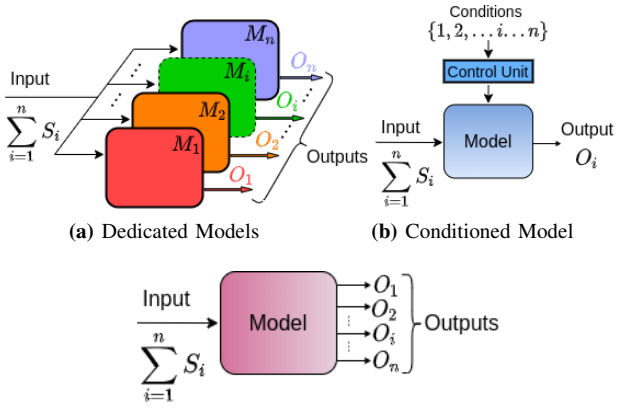
***Abstract*—A fairly straightforward approach for music source separation is to train independent models, wherein each model is dedicated for estimating only a specific source. Training a single model to estimate multiple sources generally does not perform as well as the independent dedicated models. However, Conditioned U-Net (C-U-Net) uses a control mechanism to train a single model for multi-source separation and attempts to achieve a performance comparable to that of the dedicated models. We propose a multi-channel U-Net (M-U-Net) trained using a weighted multi-task loss as an alternative to the C-U-Net. We investigate two weighting strategies for our multi-task loss: 1) Dynamic Weighted Average (DWA), and 2) Energy Based Weighting (EBW). DWA determines the weights by tracking the rate of change of loss of each task during training. EBW aims to neutralize the effect of the training bias arising from the difference in energy levels of each of the sources in a mixture. Our methods provide two-fold advantages compared to the C-U-Net: 1) Fewer effective training iterations per epoch with no conditioning, and 2) Fewer trainable network parameters (no control parameters). Our methods achieve performance comparable to that of C-U-Net and the dedicated U-Nets at a much lower training cost.**

***Index Terms*—source separation, multi-task loss, supervised, deep learning, weighted loss**

I. Introduction

Music source separation is the automatic estimation of the individual isolated sources that make up the audio mixture. It has been one of the most popular research problems in the music information retrieval community. Since most of the music audio present in the world exists in the form of mixtures, there are several applications of a system capable of music source separation – e.g. automatic creation of karaoke, music transcription, music unmixing and remixing, music production and assistance in music education.

We are interested in training a system discriminatively to estimate the sources present in the audio mixture. The deep neural networks (DNN) have been extensively used for this purpose. The existing methods mostly use DNN with either the spectrogram as the input signal representation [1], [2], [3] or directly the time-domain representation [4], [5] to train such a system. The spectrograms are compact representations of time-domain waveforms. The networks operating directly on the time-domain waveforms require larger convolution kernels than those operating on the spectrograms because of the higher time resolution in the time-domain waveforms. Hence, the number of trainable network parameters in the waveform based models are generally higher than that of spectrogram based models. In this way, the spectrogram based models have lesser training cost than the waveform based models. In most of the spectrogram based methods, the networks are trained to estimate masks like binary masks or ratio masks. These masks are then multiplied with the magnitude spectrogram of the mixture to obtain the estimates of the corresponding sources.



(a) Dedicated Models **(b)** Conditioned Model

(c) Multi-task Model

Fig. 1: Typical models for music source separation.

Convolutional neural networks (CNN) [2], [6] and long short term memory (LSTM) [7], [8] networks are the popular choices for DNN model architectures adapted for music source separation. Some of the latest top performing music source separation models are Open-Unmix [8], MMDenseLSTM [9], Demucs [5] and Meta TasNet [10]. While Open-Unmix and MMDenseLSTM models comprise of LSTMs and operate on spectrogram input, the other two methods operate directly on the time-domain waveform. It is not possible to single out any one of these models as the best model because they differ in number of trainable parameters, training time and performance metrics with respect to each of the sources. Li et al. propose Sams-Net [11] which uses attention mechanism along with CNN layers and achieves larger receptive field than CNNs and LSTMs. Considering the low number of trainable parameters

This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 713673. V. S. K. has received financial support through la Caixa Foundation (ID 100010434), fellowship code: LCF/BQ/DI18/11660064. Additional funding comes from the MICINN/FEDER UE project with reference PGC2018-098625-B-I00, H2020-MSCA-RISE-2017 project with reference 777826 NoMADS, Spanish Ministry of Economy and Competitiveness under the Mara de Maeztu Units of Excellence Program (MDM-2015-0502) and the Social European Funds. We also thank Nvidia for the donation of GPUs.

in the network, Sams-Net performs remarkably better than most of the other music source separation methods. Among the CNN based methods operating on spectrogram input, the U-Net [12] based methods [13], [14], [15] have been popular owing to their simplicity and ease of training.

In this work, we train a single multi-channel U-Net (M-U-Net) for multi-instrument source separation using a weighted multi-task loss function. We investigate the source separation task in two settings: 1) singing voice separation (two sources), and 2) multi-instrument source separation (four sources). The number of final output channels of our M-U-Net corresponds to the total number of sources in the chosen setting. Each loss term in our multi-task loss function corresponds to the loss on the respective source estimates. We explore Dynamic Weighted Average (DWA) [16] and Energy Based Weighting (EBW) strategies to determine the weights for our multi-task loss function. We compare the performance of our M-U-Net trained with multi-task loss to that of dedicated U-Nets and the C-U-Net. Then, we investigate the effect of training with the silent-source samples[1] on the performance. We also study the effect of the choice of loss term definition on the source separation performance.

Our main contributions are:

- to propose M-U-Net as a computationally cheaper alternative (in terms of the number of training iterations and trainable parameters) for multi-instrument source separation to C-U-Net and the dedicated U-Nets.
- to propose a novel weighting strategy, EBW, for the multi-task loss function, based on the energy distribution in the ground truth sources.
- to show that training a model by discarding the data samples containing silent sources could reduce the overall number of training iterations and yet perform as good as the model trained with all the training data samples.
- to emphasize the importance of choosing appropriate signal representation for computing the loss term.

The rest of the paper is organized as follows. We review the related works in the context of our work in Section II. In Section III, we describe our source separation methodology. In Section IV, we explain our experimental setup and the experiments in detail along with the ablation studies. The final section is reserved for conclusion. The source code, along with the pre-trained model weights, audio examples and more elaborate tabulation of results are made available on .

II. Related Work

In this section, we mainly focus on the U-Net based source separation methods. For the sake of clear comparison, we restrict the comparison of performance of our method with these methods only. The objective of our work is not to achieve the best performance in source separation among all other approaches, but to highlight how a multi-task model could achieve performance comparable to that of a group of isolated single-task models at a much lesser training cost. Hence, we choose to work with a simple U-Net based model operating on spectrogram input. In our context, multi-task refers to group of parallellized single tasks where each single task corresponds to estimating a specific source, as in [17].

Jansson et al. [13] proposed using a pair of independently trained U-Nets (type (a) system in Fig. 1) for the singing voice separation. Meseguer-Brocal and Peeters [17] pointed out that the implementations of such source-specific models get computationally expensive when there is a larger number of sources to be estimated. They proposed Conditioned U-Net (C-U-Net) as a cheaper alternative to the system of dedicated U-Nets, achieving comparable performance to that of the latter despite being a single model. The C-U-Net introduces control parameters through Feature-wise Linear Modulation (FiLM) layers in the encoder part of U-Net which adapts the model to estimate the source of the desired choice. C-U-Net corresponds to the type (b) system in Fig. 1. Though it is an interesting development over the work of [13], we notice that training a C-U-Net could also get expensive as we scale up the number of sources to be estimated. This is so because a data sample needs to be passed through C-U-Net multiple times in every training epoch – each time with a source specific conditioning. In this way, for K sources, the effective number of C-U-Net training iterations will be at least K times that of the number of training iterations for a single U-Net for every epoch. There could be even higher number of training iterations for an epoch if the C-U-Net is also conditioned on more than one instrument at a time as shown in [17]. Also, the addition of control parameters further adds to the training cost. For these reasons, we investigate the possibility of using a single multi-channel U-Net (M-U-Net) (a type (c) system from Fig. 1) which neither requires training over a data sample multiple times in an epoch nor does it involve any additional trainable parameters. Another motivation for using the multi-task model comes from the fact that it could potentially perform even better than the dedicated models by learning from the extra mutual information across the tasks and sharing inductive bias as pointed out by Caruana [18]. Oh et al. [19] propose a multi-channel U-Net for music source separation by adjusting the number of output channels to match the number of sources to be estimated. They also propose a weighting scheme for the multi-task loss function to balance the effect of unequal volume levels of different sources. Their network estimates the magnitude spectrograms directly as the output. In our work, we explore several weighting schemes for optimizing the multi-task loss function and compare the performance of our approach with that of dedicated U-Nets, C-U-Net and Oh et al. For the sake of fair comparison, we adapt implementation of Oh et al. and C-U-Net to our experimental setup.

III. Proposed Method

We train a Multi-channel U-Net (M-U-Net) that generates multiple outputs, one per source in the mixture. Having multiple outputs gives rise to multiple task-specific loss terms

[1]Data samples containing at least one silent source.

and hence the following multi-task loss function:

$$\mathcal{L} = \sum_{i=1}^{K} w_i L_i, \tag{1}$$

where L_i is the loss term corresponding to the i-th source, w_i is its corresponding weight, K is the number of sources and $\mathcal{L}$ is the overall scalar-valued loss. The input to our M-U-Net is the log-magnitude spectrogram of an audio mixture data sample. We train the U-Net to generate soft masks $\hat{M}_i$ as the outputs.

We explore two different definitions for the individual loss terms L_i:

a) Direct Loss: In this case, firstly, we determine the Ideal Amplitude Masks (IAM [20]) M_i for the ground truth source magnitude spectrograms S_i for each time-frequency bin (n, m) as:

$$M_i(n,m) = \min\left\{\frac{S_i(n,m)}{S_{mix}(n,m)}, 10\right\}, \tag{2}$$

where $S_{mix} \in \mathbb{R}_{+}^{F\times T}$ is the mixture magnitude spectrogram. We clip the values exceeding 10 for numerical stability in training. We then find the mean absolute value error (L1 loss) directly between the original source IAM masks M_i and their respective estimated masks $\hat{M}_i$:

$$L_i = \sum_{n=1}^{T}\sum_{m=1}^{F} |M_i(n,m) - \hat{M}_i(n,m)| \tag{3}$$

b) Indirect Loss: In this case, we find the mean absolute error between the original source spectrograms and the estimated spectrograms as shown in (4). It is 'indirect' in the sense that the U-Net outputs the masks but the loss term is defined on the spectrogram representations rather than the masks. This kind of loss term definition has been used in source separation works like [21], [17]. Michelsanti et al. [22] showed that such an indirect loss performs better than the direct loss term for the speech enhancement task.

$$L_i = \sum_{n=1}^{T}\sum_{m=1}^{F} |S_i(n,m) - \hat{M}_i(n,m)\, S_{mix}(n,m)| \tag{4}$$

A. Loss Weighting Strategies

Now, we shift the focus on determining the weights w_i for each loss term L_i in (1). The ranges of loss values vary from one task to another. This results in competing tasks which could eventually make the training imbalanced. Training a multi-task model with imbalanced loss contributions might eventually bias the model in favor of the task with the highest individual loss, undermining the other tasks. Since all the tasks are of equal importance to us and the ranges of their individual loss terms differ, we cannot treat the loss terms equally. We need to assign weights to these individual loss terms indicative of their relative importance with respect to each other. Finding the right set of weights helps to counter the imbalance caused by the competing tasks during training and helps the multi-task system learn better. For determining the weights of losses in our multi-task loss function, we explore mainly the Dynamic Weight Average (DWA) and the Energy Based Weighting (EWB) strategies in this paper:

1) Dynamic Weight Average (DWA): Liu et al. [16] proposed the Dynamic Weight Average method for continuously adapting the weights of losses in a multi-task loss function during training. In this method, the weights are distributed such that a loss term decreasing at a higher rate is assigned a lower weight than the loss which does not decrease much. In this way, the model learns to focus more on difficult tasks rather than selectively learning easier tasks. The weight w_i for the i-th task is determined as:

$$w_i(t) := \frac{K\exp\left(\gamma_i(t-1)/T\right)}{\sum_j \exp\left(\gamma_j(t-1)/T\right)}, \gamma_i(t-1) = \frac{L_i(t-1)}{L_i(t-2)}, \tag{5}$$

where γ_i indicates the relative descending rate of the loss term L_i, t is the iteration index, and T corresponds to the temperature which controls the softness of the task weighting. More the value of T, more even the distribution of weights across all the tasks.

In our work, we use DWA with $T = 2$, the loss term $L_i(t)$ being the average loss across the iterations in an epoch for the i-th task. Like in [16], we also set $\gamma_i(t) = 1$ for $t = 1, 2$ to avoid improper initialization.

2) Energy Based Weighting (EBW): We determine the energy of a target source by summing the square value of each time-frequency bin in a magnitude spectrogram of a target source for a sample, normalize it by dividing by the number of time-frequency bins and then averaging across all the samples for the specific source. We notice that the energy distribution across the target sources is non-uniform (see Fig. 2). In the singing voice separation setting, the average energy of accompaniment is more than that of the vocals. In the multi-instrument source separation setting, the bass has relatively higher average energy than the other sources. We hypothesize that the uneven energy distribution could be a reason why the multi-task model preferentially learns certain tasks more than the others. When we trained our multi-task model with unit weighted loss, the estimates of sources with higher average energy were better than that of lower energy sources. Hence, we propose a weighting strategy based on energy distribution in the target representations such that the model does not become biased to the source with high-energy.

We explore the following energy-based weighting settings in this work:

a) EBW_P1: In this setting, we use the average energy content in i-th source, E_i, across all the samples (see (6)). Note that the weights are constant throughout the training for this setting.

$$w_i = \max_{j\in\{1,\ldots,K\}} \mathrm{E}_j \,/\, \mathrm{E}_i \tag{6}$$

This way, $w_i \geq 1$ for all tasks, being $w_i = 1$ for the task associated to the source with highest energy and $w_i > 1$ for the rest, and, in particular, the lower the energy of a specific source the higher its corresponding weight w_i, thus keeping the balance among tasks.

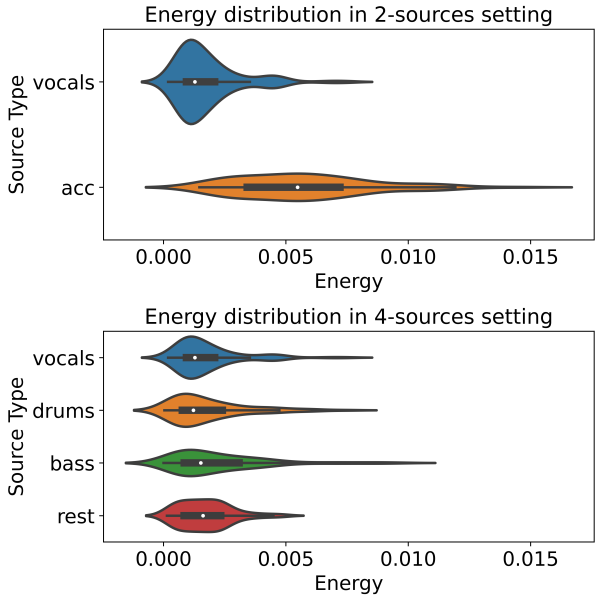


Fig. 2: Energy distribution across the sources.

b) EBW_InstP1: In this setting, we use the average energy content in i-th source, E_i, across all the samples in a batch at iteration t as shown in (7). Note that the weights change during the training for this setting.

$$w_i(t) = \max_{j \in \{1,\ldots,K\}} \mathrm{E}_j(t) \,/\, \mathrm{E}_i(t) \tag{7}$$

c) EBW_P2: This setting is very similar to that of EBW_P1 except for the fact that there is a power of 2 while determining the weights thus strengthening the relative importance between the task related to the highest-energy source and the rest of sources:

$$w_i = \max_{j \in \{1,\ldots,K\}} \mathrm{E}_j^2 \,/\, \mathrm{E}_i^2 \tag{8}$$

Note that these weights are constant throughout the training for this setting as in EBW_P1.

d) Oh et al. [19]: Finally, we also experiment with the weighting scheme proposed by Oh et al. [19]. In this setting, the weights are determined by solving the following pair of equations:

$$w_1\mathrm{E}_1 = w_2\mathrm{E}_2 = \ldots = w_i\mathrm{E}_i = \ldots = w_K\mathrm{E}_K \tag{9}$$

$$\sum_{i=1}^{K} w_i = 1 \tag{10}$$

IV. Experiments

In this section, we explore the effect of weighting strategies discussed in the previous section in training a multi-task model for source separation and compare their performance to that of a system of dedicated models and a conditioned multi-task model. We also perform ablation studies concerning the effect of choice of loss term and the effect of silent-source samples.

TABLE I: Training Cost

Model	Dedicated U-Nets	C-U-Net	M-U-Net
# params (approx.)	124M $\times K$	162M	**124M**
# training iterations	N each	N $\times K$	**N**

(N = number of training samples, M = million)

A. Dataset

We use the Musdb18 [23] dataset for this work. It contains 150 full-length stereo (two channels) audio tracks along with the isolated constituent sources. The ground truth sources are available in two settings: i) 2 sources (vocals and accompaniment), and ii) 4 sources (vocals, drums, bass and rest). The dataset comes with a pre-defined split of 100 tracks for training and 50 for testing. We convert them to mono (single channel), downsample the audio to 10880Hz (as in [21]) and split them into 6s long chunks without any overlap. We then apply the Short-time Fourier Transform (STFT) on these chunks using a 'Hanning' window of size of 1022 and hop-size of 256. This results in spectrograms of size 512×256. We resample these spectrograms to 256×256. These preprocessing steps are similar to that of [21]. We move 5% of the spectrogram samples from the training set to form our validation set. From this new training set, we filter out the silent-source samples.

B. Network Architecture

All the models in this work are based on a basic U-Net [12] model comprising of filters of sizes {32, 64, 128, 256, 512, 1024, 2048}. We have 6 down-convolution blocks, a transition block and 6 upconvolution blocks along with the skip connections in-between them. Throughout the experiments, we train the models using the Stochastic Gradient Descent (SGD) optimizer with a learning rate to 0.01 (unless otherwise mentioned) and a dropout of 0.1. The input to all our models is the log-magnitude spectrogram of audio mixture data sample of dimensions 256×256.

In case of the dedicated U-Nets, we use a U-Net with single channel output since it estimates only a single source at a time. For the C-U-Net model, we adapt the implementation of C-U-Net provided by [17] to make it consistent with our U-Net architecture for a fairer comparison. In C-U-Net too, there is a single channel output as it estimates only one source at a time. With regard to Oh et al., we only test the performance of their weighting scheme in our experiment setup along with the other weighting schemes that we propose in this paper. For the sake of a fair comparison, we adapt the Oh et al. method to estimate masks instead of magnitude spectograms as in their original work. In our M-U-Net, the number of output channels correspond to the total number of sources to be estimated, K. The training cost for each of these models is reported in Table I. Note that our M-U-Net has the least number of trainable parameters as well as the training iterations as compared to the others. More the training iterations longer the training time.

TABLE II: Results of Singing Voice Separation in SDR (median in parenthesis)

Model	Vocals	Accompaniment	Overall
Dedicated U-Nets (x2)	5.09 ± 4.31 (5.61)	12.95 ± 3.18 (12.53)	9.02 ± 5.46 (9.64)
C-U-Net	4.42 ± 4.98 (5.17)	12.21 ± 2.58 (12.16)	8.31 ± 5.56 (9.26)
UW	5.06 ± 4.93 (5.75)	12.98 ± 3.14 (12.48)	9.02 ± 5.72 (9.74)
DWA	5.20 ± 4.50 (5.67)	12.96 ± 3.11 (12.44)	9.08 ± 5.48 (9.61)
EBW P1	5.12 ± 4.78 **(5.89)**	**13.06** ± 2.91 **(12.88)**	9.09 ± 5.60 (9.77)
EBW InstP1	**5.28** ± 4.60 (5.79)	13.04 ± 3.02 (12.69)	**9.16** ± 5.50 **(9.79)**
Oh et al. [19]	5.18 ± 4.17 (5.67)	13.00 ± 3.03 (12.63)	9.09 ± 5.35 (9.78)
EBW P2*	5.07 ± 4.56 (5.63)	12.89 ± 2.95 (12.39)	8.98 ± 5.48 (9.66)

* trained with learning rate 0.001 instead of 0.01

TABLE III: Results of Multi-instrument Source Separation in SDR (median values)

Model	Vocals	Drums	Bass	Rest	Overall
Dedicated U-Nets (x4)	**5.77**	4.60	**3.19**	2.23	3.61
C-U-Net	5.26	4.30	2.97	1.69	3.37
UW	5.46	4.72	2.81	2.49	3.58
DWA	5.24	**4.92**	2.88	2.45	3.61
EBW P1	5.41	4.77	2.94	**2.64**	3.65
EBW InstP1	5.46	4.85	2.86	2.58	3.52
Oh et al. [19]	5.29	4.86	2.85	2.55	3.60
EBW P2	5.44	4.89	2.99	2.58	**3.66**

TABLE IV: Results of Ablation Studies (median in parenthesis)

Model	Overall Performance Metrics		
	SDR	SIR	SAR
EBW_P1*	**3.46** ± 4.15 **(3.65)**	7.97 ± 5.04 (7.93)	**6.83** ± 3.18 **(6.77)**
EBW_P1 with Direct Loss (3)	3.31 ± 3.96 (3.46)	7.98 ± 4.97 (8.18)	6.64 ± 3.12 (6.67)
EBW_P1* without filtering	3.44 ± 4.28 (3.59)	**8.33** ± 5.10 **(8.24)**	6.61 ± 3.46 (6.72)

* trained using Indirect Loss (4)

C. Evaluation Metrics

We choose to evaluate the following metrics [24] : Source-to-Distortion Ratio (SDR), Source-to-Interference Ratio (SIR) and Source-to-Artifact Ratio (SAR) typically used for evaluating music source separation performance. We use the mir_eval toolbox [25] to get these metrics. Note that all our models estimate the soft masks and we obtain the magnitude spectrogram estimates of each source by multiplying these masks with the magnitude spectrogram of the mixture. We combine the phase of the mixture spectrogram along with the magnitude spectrograms of the estimated sources and apply inverse-STFT transform to obtain the waveforms. The metrics are evaluated on the waveforms of the estimated sources with respect to the appropriately downsampled ground truth audio waveforms. Among these three metrics, SDR is more indicative of the source separation quality as a global performance measure [24] and some works (e.g. [4]) report only this metric. We have not published the SAR and SIR performance metrics in this paper for the sake of ease of readability. A more detailed tabulation of the results along with these metrics, is made available on the project webpage.

D. Multi-task Experiments

We aim to show that our M-U-Net can perform as good as the dedicated U-Nets for both singing voice separation (2 sources) and multi-instrument source separation (4 sources) with fewer trainable parameters. We conduct experiments with the M-U-Net exploring the weighting strategies: {DWA, EBW_P1, EBW_InstP1, EBW_P2 and Oh et al. [19]} which have been discussed earlier. We compare the performance of these M-U-Nets with the dedicated U-Nets and the C-U-Net. To notice the effectiveness of the weighting strategies, we also train an M-U-Net with unit weights (UW) and compare the performance with the models trained with our weighting strategies. Throughout the experiments, unless otherwise mentioned, we use the indirect loss function definition (4). Table II and Table III, respectively, report the results for singing voice separation and multi-instrument source separation.

From these tables, we notice that the source separation performance gets worse as the sources increase from 2 to 4, across all methods. In general, we find the performance of M-U-Nets trained using our weighting strategies comparable to that of Oh et al. method, C-U-Net and the dedicated U-Nets. Especially for the 2 source setting, the energy based methods and DWA method not only perform better than the C-U-Net and the Dedicated U-Nets, but also outperform the naive unit weighting (UW) based model, indicating the usefulness of our weighting strategies. As seen in Fig. 2, the average energy value between the sources differs a lot more in the 2 source setting than in the 4 source setting. Hence, the EBW methods which incorporate the signal energy information to weight the loss terms, perform better than the energy agnostic DWA method in the 2 source setting rather than in the 4 source setting. Looking at the *Overall* SDR metrics (last column in the tables), we notice that our M-U-Net performs better than the Dedicated U-Nets, C-U-Net and Oh et al. method in both 2-source and 4-source settings at much lesser training cost compared to C-U-Net and Dedicated U-Nets.

E. Ablation Studies

Now, we present some additional experiments to evaluate which loss definition, among direct loss and indirect loss, performs better. We also analyze the effect of including silent-source samples in the training set. For these additional experiments, we consider the setting EBW_P1 on 4 sources as a reference. Table IV reports the performance metrics pertaining to these experiments.

From Table IV, we notice that the overall performance drops in both the ablation studies. Despite the slight improvement in the SIR metric on using direct loss (3), based on the SDR and SAR metrics, we recommend using indirect loss (4) definition which is congruent with the findings for speech enhancement [22]. We also infer that training with silent-source samples does not contribute much to the overall performance and we recommend discarding them.

V. Conclusion

We presented a multi-channel U-Net as a cheaper alternative (in terms of the number of training iterations and number

of trainable parameters) to the Conditioned U-Net and the system of dedicated U-Nets for music source separation. Such an approach could be potentially extended to models other than the U-Net and perhaps also for other kinds of tasks. We also presented a novel weighting strategy, EBW, for training the multi-task loss function based on the energy of the signal representations. We showed how the EBW method is effective when the average energy values across the sources to be estimated are very different. We believe there are other ways of distilling the energy information into the weighting strategy and leave it for future work. We also showed that discarding silent-source samples during training saves on the training cost without much compromise in performance. We also showed that M-U-Net performs better when trained with an indirect loss term rather than the direct loss on the masks.

ACKNOWLEDGMENT

We thank Daniel Michelsanti (Aalborg University) and Olga Slizovskaia (Universitat Pompeu Fabra) for the insightful discussions related to source separation methods and practices.

REFERENCES

[1] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel music separation with deep neural networks," in *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, 2016, pp. 1748–1752.

[2] E. M. Grais, G. Roma, A. J. Simpson, and M. D. Plumbley, "Combining mask estimates for single channel audio source separation using deep neural networks," in *17th Annual Conference of the International Speech Communication Association*, 2016, pp. 3339–3343.

[3] G. Roma, O. Green, and P. A. Tremblay, "Improving single-network single-channel separation of musical audio with convolutional layers," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2018, pp. 306–315.

[4] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," in *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, September 23-27, 2018*, 2018, pp. 334–340.

[5] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Demucs: Deep extractor for music sources with extra unlabeled data remixed," *arXiv preprint arXiv:1909.01174*, 2019.

[6] S. Park, T. Kim, K. Lee, and N. Kwak, "Music source separation using stacked hourglass networks," in *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, September 23-27, 2018*, 2018, pp. 289–296.

[7] P. Seetharaman, G. Wichern, S. Venkataramani, and J. Le Roux, "Class-conditional embeddings for music source separation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 301–305.

[8] F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, "Open-unmix-a reference implementation for music source separation," *Journal of Open Source Software*, vol. 4, no. 41, p. 1667, 2019.

[9] N. Takahashi, N. Goswami, and Y. Mitsufuji, "Mmdenselstm: An efficient combination of convolutional and recurrent neural networks for audio source separation," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2018, pp. 106–110.

[10] D. Samuel, A. Ganeshan, and J. Naradowsky, "Meta-learning extractors for music source separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 816–820.

[11] T. Li, J.-W. Chen, H. Hou, and M. Li, "Sams-net: A sliced attention-based neural network for music source separation," *arXiv: Audio and Speech Processing*, 2019.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[13] A. Jansson, E. J. Humphrey, N. Montecchio, R. M. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep u-net convolutional networks," in *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017, Suzhou, China, October 23-27, 2017*, 2017, pp. 745–751.

[14] D. Stoller, S. Ewert, and S. Dixon, "Adversarial semi-supervised audio source separation applied to singing voice extraction," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2391–2395.

[15] L. Prétet, R. Hennequin, J. Royo-Letelier, and A. Vaglio, "Singing voice separation: A study on training data," in *ICASSP 2019-2019 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 506–510.

[16] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1871–1880.

[17] G. Meseguer-Brocal and G. Peeters, "Conditioned-u-net: Introducing a control mechanism in the u-net for multiple source separations," in *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019, Delft, The Netherlands, November 4-8, 2019*, 2019, pp. 159–165.

[18] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[19] J. Oh, D. Kim, and S.-Y. Yun, "Spectrogram-channels u-net: a source separation model viewing each channel as the spectrogram of each source," *arXiv preprint arXiv:1810.11520*, 2018.

[20] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[21] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 570–586.

[22] D. Michelsanti, Z.-H. Tan, S. Sigurdsson, and J. Jensen, "On training targets and objective functions for deep-learning-based audio-visual speech enhancement," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 8077–8081.

[23] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "The MUSDB18 corpus for music separation," Dec. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.1117372

[24] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[25] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis, "Mir_eval: A transparent implementation of common MIR metrics," in *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR 2014, Taipei, Taiwan, October 27-31, 2014*, 2014, pp. 367–372.