# Real-time Speech Enhancement and Separation with a Unified Deep Neural Network for Single/Dual Talker Scenarios

Kashyap Patel\*, Anton Kovalyov, and Issa Panahi Electrical and Computer Engineering, University of Texas at Dallas Richardson, USA Email: \*patelkashyap@utdallas.edu

Abstract—This paper introduces a practical approach for leveraging a real-time deep learning model to alternate between speech enhancement and joint speech enhancement and separation depending on whether the input mixture contains one or two active speakers. Scale-invariant signal-to-distortion ratio (SI-SDR) has shown to be a highly effective training measure in time-domain speech separation. However, the SI-SDR metric is ill-defined for zero-energy target signals, which is a problem when training a speech separation model using utterances with varying numbers of talkers. Unlike existing solutions that focus on modifying the loss function to accommodate zero-energy target signals, the proposed approach circumvents this problem by training the model to extract speech on both its output channels regardless if the input is a single or dual-talker mixture. A lightweight speaker overlap detection (SOD) module is also introduced to differentiate between single and dual-talker segments in real time. The proposed module takes advantage of the new formulation by operating directly on the separated masks, given by the separation model, instead of the original mixture, thus effectively simplifying the detection task. Experimental results show that the proposed training approach outperforms existing solutions, and the SOD module exhibits high accuracy.

Index Terms—Speech separation, speech enhancement, real-time processing, multi-talker detection

## I. INTRODUCTION

In recent years, a lot of research has been done on deep learning (DL)-based speech enhancement (SE) and speech separation (SS) methods. Applications include automatic speech recognition (ASR) and hearing aids [1]–[3]. In the real world, speech signals are often a reverberant mixture of one or more speech sources and noise. When the number of speech sources is not known a priori, it can be hard for a real-time system to determine whether only SE, in the case of a single talker, or a combination of SE and SS, in the case of multiple overlapping talkers, should be applied. In practice, the number of overlapping talkers is rarely more than two [4]. Hence, for the purpose of this work, we consider only single and dual-talker scenarios.

Numerous techniques have been suggested to overcome the challenges of a practical system when dealing with different numbers of concurrently active talkers [5]–[9]. One approach is to train multiple networks for different speaker counts [5], [6]. The speaker count is then estimated by a separate module and the appropriate SE or SS model is selected to extract the speech signal/s. In a similar but more efficient approach, a noncausal neural network architecture was proposed that utilizes a shared encoder and separator but a different decoder for each speaker count [7], resulting in considerably less training parameters. Although the technique of speaker counting followed by selecting the appropriate SE or SS module was shown to work well in noncausal settings, it is unclear how it can be efficiently applied to latency-demanding applications.

In a more practical approach, a single-talker signal can be modeled as dual-talker by setting the second signal to zero energy, i.e., silent speech. With this formulation, an SS model can be trained to jointly perform SS and SE in either single or dual talker scenarios. However, time-domain SS models are typically trained using a signal-to-distortion ratio (SDR)-based loss function, which is ill-defined for zero-energy target signals. Although modifications to SDR-based loss functions have been proposed to handle zero-energy target signals [10], [11], they generally come at the cost of somewhat degraded performance in terms of the original SDR metric.

Motivated by the above observations, this study proposes a simple and efficient approach for training a DL-based SS model to handle both single and dual-talker scenarios. Given an input mixture, the proposed approach consists in training a model to output two channels. In a dual-talker scenario, these channels correspond to the two separated and enhanced speech signals, whereas in a single-talker case, both channels correspond to the same output, the enhanced speech. This approach allows leveraging the standard permutation invariant training (PIT) [12] with an SDR-based loss function without requiring any modifications. Additionally, taking advantage of the new formulation, a lightweight speaker overlap detection (SOD) post-processing module is introduced for detecting dual-talker instances in real time. This module simplifies the detection task by operating directly on the separated masks, estimated by the SS model, instead of the original mixture signal. The proposed methodology was tested on the recently introduced UX-Net model [13] for causal,

This work was supported by the National Institute on Deafness and Other Communication Disorders (NIDCD) of the National Institutes of Health (NIH) under Award 5R01DC015430-05. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

low-latency SS, revealing improved performance over methods that reformulate the SDR measure. The proposed SOD module is also shown to attain high accuracy.

# **II. PROBLEM FORMULATION**

Let us consider a scenario with one or two active speech sources in a noisy and reverberant environment. The time-domain signal captured by the microphone is modeled by

$$\mathbf{y} = \mathbf{s}_1 + \mathbf{s}_2 + \mathbf{v} \tag{1}$$

where  $s_1$  and  $s_2$  are the clean reverberant speech signals of the two sources, and v is background noise. In a dual-talker scenario, we wish to extract both  $s_1$  and  $s_2$  from y (SS task), whereas, in a single talker scenario,  $s_2$  is assumed to be a zero-energy signal, and we wish to extract only  $s_1$  from y (SE task).

Let  $S = {\mathbf{s}_1, \mathbf{s}_2}$  and  $\hat{S} = {\hat{\mathbf{s}}_1, \hat{\mathbf{s}}_2}$  denote sets grouping the speech sources of interest and their corresponding estimates, respectively. Let  $\mathcal{F}$  denote the SS model trained to extract  $\hat{S}$  given the input mixture  $\mathbf{y}$ .  $\mathcal{F}$  is trained applying PIT to minimize

$$\mathcal{L}(S, \hat{S}) = -\frac{1}{2} \max_{\pi} \sum_{n=1}^{2} \mathcal{D}(\mathbf{s}_{\pi(n)}, \hat{\mathbf{s}}_{n}) , \qquad (2)$$

where  $\pi$  represents the permutation set on S and  $\mathcal{D}(\mathbf{s}, \hat{\mathbf{s}})$  is a signal-level similarity measure between a target utterance s and its estimate  $\hat{\mathbf{s}}$ . The most commonly used similarity measures are SDR and scale-invariant SDR (SI-SDR) [14]. Both measures can be jointly expressed by letting

$$\mathcal{D}(\mathbf{s}, \hat{\mathbf{s}}) = 10 \log_{10} \left( \frac{\|\alpha \mathbf{s}\|^2}{\|\hat{\mathbf{s}} - \alpha \mathbf{s}\|^2 + \epsilon} + \epsilon \right) .$$
(3)

where  $\|\cdot\|$  denotes Euclidean norm,  $\epsilon$  is a constant for numerical stability and  $\alpha$  is a parameter selected to be either 1 for SDR or the scalar projection of  $\hat{s}$  onto s, i.e.,  $\frac{\hat{s}^T s}{\|s\|^2}$ , for SI-SDR. Among the two measures, SI-SDR is typically preferred due to its invariance to signal scaling. However, both measures are ill-defined when one of the target signals is zero energy, such as  $s_2$  in (1). Hence, the problem is to modify the training objective of  $\mathcal{F}$  to allow the same model to perform joint SE and SS in single and dual talker scenarios.

#### **III. EXISTING SOLUTIONS**

Let us first discuss previously proposed approaches in the literature for tackling the problem in this study.

# A. Softmax SDR

One approach is to add a small positive constant  $\epsilon$  to the numerator of the SDR measure. However, this introduces a bias in training since zero-energy target signals are easy to learn. However, this issue can be addressed by limiting the SDR value to a soft maximum [15], resulting in the following performance measure

$$\mathcal{D}_{\epsilon-\text{tSDR}}(\mathbf{s}, \hat{\mathbf{s}}) = 10 \log_{10} \frac{\|\mathbf{s}\|^2 + \epsilon}{\|\hat{\mathbf{s}} - \mathbf{s}\|^2 + \tau(\|\mathbf{s}\|^2 + \epsilon)} , \quad (4)$$

where  $\tau = 10^{-\text{SDR}_{\text{max}}/10}$  is a constant that restricts the maximum value of SDR to some threshold SDR<sub>max</sub>.

#### B. Source aggregated SDR

A recent study proposed a modified training objective for handling varying numbers of overlapping talkers, called source aggregated SDR (SA-SDR) [10]. Unlike traditional PIT in (2) paired with the SDR measure in (3), which, for a given permutation, computes the arithmetic mean of signal-level SDRs, SA-SDR aggregates the energies of the target signals and reconstruction errors to compute a global SDR measure. The modified PIT objective is given by

$$\mathcal{L}_{\text{SA-SDR}}(S, \hat{S}) = -\max_{\pi} 10 \log_{10} \frac{\sum_{n=1}^{N} \|\mathbf{s}_{\pi(n)}\|^2}{\sum_{n=1}^{N} \|\hat{\mathbf{s}}_n - \mathbf{s}_{\pi(n)}\|^2} , \quad (5)$$

where N is the maximum number of concurrent speakers, which in the context of this work is 2. This training objective is defined as long as the mixture contains at least one active speech source.

#### C. Multi-Objective Loss

An alternative approach is to introduce distinct loss functions during training for tackling different numbers of talkers [16]. In the context of this work, we can modify the training objective as follows

$$\mathcal{L}_{\text{MOL}}(S, \hat{S}) = \begin{cases} -\mathcal{D}(\mathbf{s}_1, \hat{\mathbf{s}}_1) - \lambda \mathcal{D}_{\text{log-MSE}}(\mathbf{s}_2, \hat{\mathbf{s}}_2), & \mathbf{s}_2 = \mathbf{0} \\ \mathcal{L}(S, \hat{S}), & \mathbf{s}_2 \neq \mathbf{0} \end{cases}$$
(6)

where  $\lambda$  is some positive constant, **0** is a vector of zeros, and

$$\mathcal{D}_{\text{log-MSE}}(\mathbf{s}, \hat{\mathbf{s}}) = -10 \log_{10}(\|\hat{\mathbf{s}} - \mathbf{s}\|^2 + \epsilon)$$
(7)

is the log mean squared error (log-MSE). In a single-talker scenario, this approach employs signal-level SDR/SI-SDR in (3) along with log-MSE for training the model to respectively output enhanced speech on the first channel and zeros on the second channel. The parameter  $\lambda$  is introduced to balance the gradients between the two objectives. In dual-talker scenarios, standard PIT combined with signal-level SDR/SI-SDR is employed.

## **IV. PROPOSED SOLUTION**

The methods discussed so far involve modifying (2) or (3) to handle zero-energy signals. However, modifying the desired performance measure may degrade the model's performance in terms of the original metric. Hence, in this work we prefer to avoid introducing any modifications to (2) and (3) and reformulate the target signals in S as follows

$$\bar{S} = \begin{cases} \{\mathbf{s}_1, \mathbf{s}_1\}, & \mathbf{s}_2 = \mathbf{0} \\ \{\mathbf{s}_1, \mathbf{s}_2\}, & \mathbf{s}_2 \neq \mathbf{0} \end{cases},$$
(8)

meaning that, in single-talker scenarios,  $\mathcal{F}$  is trained to extract  $s_1$  at both channels, whereas, in dual-talker scenarios,  $\mathcal{F}$  is trained to extract the distinct sources, just as in conventional SS. This approach is motivated by the observation that dual-talker SS models tend to output a similar version of the signal at both channels when the input mixture consists of just one speech source. Hence, the idea is to simplify training without the need of modifying the desired performance measure.

The proposed modification in (8) introduces the need for an additional post-processor to differentiate between single and dual talker scenarios. For this purpose, we propose a lightweight SOD module that would work in tandem with  $\mathcal{F}$  to detect speaker overlap in real time. This module can serve different purposes, e.g., preventing ASR systems from processing both output channels in single-talker scenarios; and, optionally, allowing a means to know when to replace the output at the second channel with zeros if consistency with the original target signals in *S* is desired.

In this work, we consider  $\mathcal{F}$ as a real-time time-domain SS neural network that follows the general encoder-separator-decoder design of the well-known time-domain audio speech separation network (TaSNet) [2]. As shown in Fig. 1, the SOD module operates in a frame-wise manner and takes as input the two-channel mask vectors from the separator module of  $\mathcal{F}$ . The masks are concatenated resulting in the vector  $\mathbf{m}^{(2K)}$ , where the superscript denotes the length of the vector and K is the dimension of individual-channel masks.  $\mathbf{m}^{(2K)}$  is then fed as input to the SOD module, which consists of the following three processing stages

$$\mathbf{m}_{1}^{(H)} = \operatorname{ReLU}(\operatorname{FF}(\mathbf{m}^{(2K)}))$$
  

$$\mathbf{m}_{2}^{(H)} = \operatorname{ReLU}(\operatorname{Stacked-GRU}(\mathbf{m}_{1}^{(H)})) \qquad (9)$$
  

$$\mathbf{m}_{3}^{(1)} = \sigma(\operatorname{FF}(\mathbf{m}_{2}^{(H)})) ,$$

where ReLU(·) and  $\sigma(\cdot)$  denote rectified linear unit and sigmoid activation functions, respectively, FF(·) denotes a feed-forward layer, and Stacked-GRU(·) denotes two stacked gated recurrent unit (GRU) layers. Recurrent processing is introduced to provide longer context awareness. It follows that the SOD module transforms an input vector of length 2K into a lower-dimensional hidden state vector of length H which is then further processed and mapped into a classification output denoted by a scalar value between 0 and 1. Frame-level outputs are averaged across an individual training utterance and the module is trained separately from  $\mathcal{F}$  to minimize the binary cross-entropy loss.

Despite its simplicity, the advantage of the proposed training approach in (8) compared to existing solutions is that it allows training the same model  $\mathcal{F}$  on single and dual-talker datasets without modifying the popular SI-SDR-based loss function. Furthermore, forcing both channels to output an audio signal, even in single-talker scenarios, enables the model to equally optimize the parameters associated with the output at both channels. Lastly, the new training approach paired with the proposed SOD post-processing module conveniently enable efficient real-time detection of speaker overlap by reusing the frame-level masks estimated by the separator in  $\mathcal{F}$  as input to the SOD block. This claim follows from the reasoning that SOD is simpler when the input consists of the already separated signals instead of the original mixture.

#### V. EXPERIMENTAL CONFIGURATIONS

We evaluate the performance of the proposed methodology on SE and SS tasks using an existing neural network architecture.

## A. Dataset

A dataset is generated to simulate single and dual-talker noisy mixtures in a reverberant room. This dataset consists of 36000, 10800, and 9000 4-second long utterances sampled at 16 kHz for training, testing, and validation, respectively. Clean speech and noise utterances are obtained from LibriSpeech [17] and WHAM! [18] datasets, respectively. For each utterance, the room dimensions are randomly sampled between 5 and 10 meters in length and width and 2 to 5 meters in height. The reverberation time is randomly sampled between 0.1 and 0.5 seconds. The number of talkers in the mixture is set to vary from 1 to 2. In dual-talker utterances, speech signals are mixed to have a randomly sampled signal-to-interference ratio between -5 and 5 dB. Speech and noise source positions are randomly sampled within the room with the constraint of being at least 50 cm away from the walls. The microphone is placed at the center of the room, and the image method [19] is used to generate the corresponding room impulse responses (RIRs). Reverberant speech and noise signals are added and mixed to have a signal-to-noise ratio (SNR) randomly sampled between 5 and 20 dB.

#### B. Network Architecture and Training

UG-Net [13] is adopted as the baseline model for  $\mathcal{F}$ . UG-Net is a casual TaSNet-like system designed for SS. The dimension of the encoder in UG-Net is set to 256 and the separator depth is set to 5. The frame size is set to 2 ms and 50% overlap is used, resulting in an algorithmic latency of only 3 ms. The network is trained using the Adam [20] optimizer for 70 epochs with a batch size of 8. The initial learning rate is set to  $10^{-3}$  and multiplied by 0.98 every epoch. Gradients are clipped to [-5, 5] during backpropagation to avoid the exploding gradient problem.

Once training of  $\mathcal{F}$  was completed, the SOD model was trained using as input the masks estimated by the separation module of UG-Net on the training dataset. The hidden state dimension H of the SOD module was set to 64.

# C. Evaluation

The SE and SS performance of the proposed method is evaluated using the following three performance measures: Perceptual Evaluation of Speech Quality (PESQ) [21], Short-Time Objective Intelligibility (STOI) [22], and



Fig. 1. Schematic of the proposed methodology consisting of a primary neural network  $\mathcal{F}$  for SE and SS tasks with an integrated SOD module. The model  $\mathcal{F}$  consists of encoder, separator, and a decoder modules. The arrows indicate the flow of gradients during training.

SI-SDR in dB. These metrics are reported separately for single and dual-talker scenarios using the formatting PESQ/STOI/SI-SDR. Additionally, we evaluate the SOD module's accuracy in detecting dual-talker segments using frame-level true negative (TNR) and true positive (TPR) rates.

#### VI. RESULTS

Three experiments were conducted. In the first experiment,  $\mathcal{F}$  is trained on the following three tasks: SE, SS, and SE-SS. For the SE task, the network is trained on the subset of the training set that includes only single-talker utterances using SDR and SI-SDR measures as training objectives, where only the first output channel is considered. For the SS task, the network is trained on the subset of the training set that includes only dual-talker utterances using the classic PIT with SDR ( $\mathcal{L}_{SDR}$ ) and SI-SDR ( $\mathcal{L}_{SI-SDR}$ ) training objectives. Finally, for the SE-SS task, the network is trained on the entire training set using the existing training objectives described in Section III and the proposed method described in Section IV. The parameters  $\epsilon$ , SDR<sub>max</sub> and  $\lambda$  are set to  $10^{-8}$ , 30 dB and 0.1, respectively. SI-SDR was used as the signal-level similarity measure in  $\mathcal{L}_{MOL}$  and the proposed method.

Table I reports the results of the first experiment. We note that training the model solely on the dual-talker set leads to a significant reduction in performance on the single-talker set when compared to all other methods, thus confirming the need for an improved training objective that can handle varying numbers of speakers. Among the training methods suitable for the SE-SS task, the proposed approach is shown to attain the best overall performance, especially in terms of SI-SDR. Fig. 2 further illustrates the SI-SDR performance gap in the learning curves of the different methods. This performance improvement is attributed to the use of an unaltered SI-SDR-based training objective, made possible by the proposed formulation in (8).

In the second experiment, we investigated the effect of various levels of SNR on the performance of the proposed training method on the SE-SS task and the TPR and TNR scores of the proposed SOD module. The results in Table

TABLE I COMPARISON WITH EXISTING SOLUTIONS.

Tasks	Training Objective	Single-talker	Dual-talker
Unprocessed	_	2.36/0.76/7.88	1.26/0.49/-0.13
SE	SDR	<b>2.93/0.91</b> /15.68	-
	SI-SDR	2.92/0.90/ <b>15.79</b>	_
SS	$\mathcal{L}_{ ext{SDR}}$	2.52/0.77/9.35	<b>1.85</b> /0.68/ <b>5.96</b>
	$\mathcal{L}_{ ext{SI-SDR}}$	2.58/0.78/9.10	1.84/ <b>0.69</b> /5.89
SE-SS	$\mathcal{L}_{\epsilon- ext{tSDR}}$	2.80/0.87/13.58	1.70/0.65/5.15
	$\mathcal{L}_{\text{SA-SDR}}$	<b>2.88/0.89</b> /14.30	1.74/0.67/5.49
	$\mathcal{L}_{ ext{MOL}}$	2.78/0.88/14.78	1.72/0.66/5.19
	Proposed	2.87/ <b>0.89/14.94</b>	1.80/0.68/5.68

 TABLE II

 Evaluation of the proposed methodology on the SE-SS task in varying SNR conditions.

SNR	TNR	TPR	Single-talker	Dual-talker
5-10 dB	94.6%	95.3%	2.77/0.86/14.52	1.68/0.63/5.13
10-15 dB	97.2%	97.6%	2.90/0.88/15.02	1.80/0.68/5.73
15-20 dB	98.7%	99.3%	2.94/0.90/15.28	1.92/0.70/6.10

VI show that TPR tends to be consistently higher than TNR, suggesting that the model finds detecting dual-talker segments easier than single-talker. As expected, we also note that the performance of the proposed methods tends to improve at increased SNR.

In the third experiment, we further quantify the detection performance of the SOD module in dual-talker scenarios. For this purpose, we replace the frame-wise output at the second channel of  $\mathcal{F}$  with a vector of zeros whenever the frame-level SOD value is below 0.5 and evaluate the resulting extracted



Fig. 2. Learning curves of the different training objectives targeting the SE-SS task. Results are reported in terms of SS performance on the dual-talker subset of the validation set.

 TABLE III

 EFFECT OF SOD MASKING ON THE QUALITY OF SEPARATED SIGNALS.

TNR	TPR	SOD masking	Dual-talker
97.9%	98.6%	× √	1.80/0.68/5.74 1.72/0.67/5.53

signals in terms of PESQ, STOI and SI-SDR. We refer to the procedure of replacing the second channel output frames with zeros when the SOD value is low as *SOD masking*. The initial 500 ms segment of the extracted signals is ignored to warm up the SOD module. Table III reports the results. We note that the use of SOD masking does not result in excessive degradation in signal quality, thus confirming the effectiveness of the proposed SOD module.

## VII. CONCLUSION

This paper proposed a simple and practical training approach for leveraging a real-time DL model to perform joint SE and SS in single and dual-talker scenarios. The proposed methodology circumvents the problem of zero-energy target signals by training the separation network to extract speech in both its output channels. The newly defined training targets facilitate the use of the SI-SDR measure during training as in conventional time-domain SS with fixed number of speakers. Additionally, taking advantage of the inherent speaker-counting property of the separation network, an efficient SOD module is introduced for differentiating between single and dual-talker scenarios in real time. Experimental results showed that the proposed training approach outperforms the existing solutions that consist in modifying the loss function to accommodate zero-energy target signals. Finally, the SOD module was shown to attain high performance.

#### REFERENCES

- A. Kovalyov, K. Patel, and I. Panahi, "Dfsnet: A steerable neural beamformer invariant to microphone array configuration for real-time, low-latency speech enhancement," *arXiv preprint arXiv:2302.13407*, 2023.
- [2] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 696–700.
- [3] G. S. Bhat, N. Shankar, C. K. Reddy, and I. M. Panahi, "A real-time convolutional neural network based speech enhancement for hearing impaired listeners using smartphone," *IEEE Access*, vol. 7, pp. 78 421–78 433, 2019.
- [4] Z. Chen, T. Yoshioka, L. Lu, T. Zhou, Z. Meng, Y. Luo, J. Wu, X. Xiao, and J. Li, "Continuous speech separation: Dataset and analysis," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics*, *Speech and Signal Processing (ICASSP)*, 2020, pp. 7284–7288.
- [5] Z.-Q. Wang and D. Wang, "Count and separate: Incorporating speaker counting for continuous speaker separation," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 11–15.
- [6] E. Nachmani, Y. Adi, and L. Wolf, "Voice separation with an unknown number of multiple speakers," in *International Conference on Machine Learning*. PMLR, 2020, pp. 7164–7175.
- [7] J. Zhu, R. A. Yeh, and M. Hasegawa-Johnson, "Multi-decoder dprnn: Source separation for variable number of speakers," in *ICASSP* 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 3420–3424.
- [8] N. Takahashi, S. Parthasaarathy, N. Goswami, and Y. Mitsufuji, "Recursive speech separation for unknown number of speakers," 2019.
- [9] F. Jiang and Z. Duan, "Speaker attractor network: Generalizing speech separation to unseen numbers of sources," *IEEE Signal Processing Letters*, vol. 27, pp. 1859–1863, 2020.
- [10] T. von Neumann, K. Kinoshita, C. Boeddeker, M. Delcroix, and R. Haeb-Umbach, "Sa-sdr: A novel loss function for separation of meeting style data," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6022–6026.
- [11] Y. Luo and N. Mesgarani, "Separating varying numbers of sources with auxiliary autoencoding loss," in *Proc. Interspeech 2020*, 2020, pp. 2622–2626.
- [12] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017, pp. 241–245.
- [13] K. Patel, A. Kovalyov, and I. Panahi, "Ux-net: Filter-and-process-based improved u-net for real-time time-domain audio separation," arXiv preprint arXiv:2210.15822, 2022.
- [14] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr-half-baked or well done?" in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.
- [15] T. von Neumann, K. Kinoshita, C. Boeddeker, M. Delcroix, and R. Haeb-Umbach, "Graph-PIT: Generalized Permutation Invariant Training for Continuous Separation of Arbitrary Numbers of Speakers," in *Proc. Interspeech 2021*, 2021, pp. 3490–3494.
- [16] S. Wisdom, H. Erdogan, D. P. Ellis, R. Serizel, N. Turpault, E. Fonseca, J. Salamon, P. Seetharaman, and J. R. Hershey, "What's all the fuss about free universal sound separation data?" in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). IEEE, 2021, pp. 186–190.
- [17] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2015, pp. 5206–5210.
- [18] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. Le Roux, "Wham!: Extending speech separation to noisy environments," *Proc. Interspeech 2019*, pp. 1368–1372, 2019.
- [19] E. Habets, "Room impulse response generator. technische universiteit eindhoven," Tech. Rep, 2 (2.4): 1, 2006. 5, Tech. Rep.
- [20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

- [21] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in 2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221), vol. 2. IEEE, 2001, pp. 749–752.
  [22] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time
- [22] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in 2010 IEEE international conference on acoustics, speech and signal processing. IEEE, 2010, pp. 4214–4217.