3rd Symposium on Security and Privacy in Speech Communication
19 August 2023, Dublin, Irland

Privacy and Quality Improvements in Open Offices Using Multi-Device Speech Enhancement

Silas Rech[1], *Mohammad Hassan Vali*[1], *Tom Bäckström*[1]

[1]Department of Information and Communications Engineering, Aalto University, Espoo, Finland

silas.rech, mohammad.vali, tom.backstrom}@aalto.fi

Abstract

Teleconferencing has increased in popularity and often takes place around other people such as open offices. A particular problem of such environments is that multiple users can have independent conversations simultaneously, which leak into each others' devices. This poses problems of both privacy and quality. In this work, we introduce a multi-device, targeted speech separation network. We call this network IsoNet, as it isolates the dominant speech in a mixture of multiple speakers by generating a mask from interfering speakers. This mask is used to remove speech from other simultaneous conversations in the enhanced speech signal. The privacy improvement is measured by mutual information and the enhancement quality is evaluated with a MUSHRA test, PESQ, and SI-SNR. Our experiments show a statistically significant improvement with IsoNet from 27 to 75 in MUSHRA score and a decrease of mutual information of 60%. IsoNet improves privacy as sensitive speech content is effectively attenuated.

Index Terms: privacy-aware, multi-device, targeted speech separation, voice isolation

1. Introduction

The most natural telecommunication channel between humans is phone calls or teleconferences. However, often multiple people have conversations independently of each other at the same time in the same room. This presents a threat to privacy, as the sensitive information from one conversation can leak into another [1]. We view the privacy threat in this scenario for the leaked, also referred to as interfering speaker, not for the enhanced speaker. Next to the privacy threat, it is exhausting to focus on one speaker when there are other speakers or noises over longer periods of time [2]. Though state-of-the-art speech enhancement techniques are used in telecommunication services to e.g. attenuate background noises, they are unable to effectively remove speech content from a mixture of multiple overlapping speakers. In this work, we consider two speakers, Alice and Eve, in one room which have microphones closely positioned to their mouths, such as a headset, see Figure 1.

We propose to use the two available audio streams, one from each device, to improve the privacy of one of the speech signals with multi-channel speech enhancement. Specifically, we define the target, here Alice, of the separation algorithm to be the speaker closest to the microphone. This speaker will most likely dominate the speech mixture. The interfering audio, on the other hand, stems from a speaker who does not play a role in the ongoing conversation, in this case, Eve. By effectively attenuating Eve's signal in the transmission of Alice's speech, we are increasing privacy for Eve as the transmission of her speech is limited to the desired conversation. This work presents a solution that uses two microphone signals to isolate the dominant speaker in each microphone. We call this network IsoNet because of its main functionality, isolating the dominant speaker. IsoNet generates a mask from the interfering audio stream. The inverse of that mask is then used to remove undesired speech contents from the targeted input. After the separation, an additional speech enhancement model is used to compensate for any artefacts caused by the masking. This setup makes IsoNet independent of any prior information and it can be run in real-time.

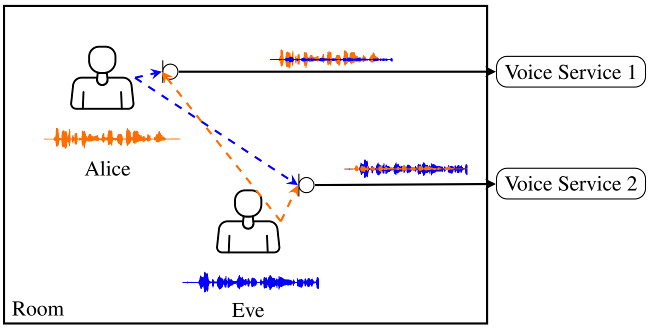

Figure 1: *Problem Scenario: Alice and Eve are in the same room but are connected to different voice services. The leakage of each speaker reduces privacy and degrades the speech quality.*

Convolutional time-domain audio separation networks such as Conv-TasNet [3] have shown great potential in speech separation tasks due to their quality and, more importantly, their real-time capability. Hence, IsoNet is also based on the basic architecture of the Conv-TasNet but improves two of its aspects further. Firstly, Conv-TasNet needs additional information about how many speakers need to be separated at all times, as it approximates an individual mask for each speaker. This requires additional methods to estimate the number of speakers in a room. Secondly, the reconstruction of the separated audio streams presents a challenge, as it is not clear how to connect separated speech channels to the right speakers, especially when the number of speakers changes over time. Although Conv-TasNet was not designed as a *targeted* speech separation network, it shows state-of-the-art separation performance in subjective measures such as the mean opinion score (MOS) compared to networks that target a speaker directly. Moreover, implementations of Conv-TasNet are readily available which makes it a good baseline for this work. Other recent works that are designed to target an individual speaker leverage additional information about a speaker such as visual cues [4] or a-priori speaker embeddings which define the targeted speaker beforehand like [5–9]. The methods proposed in WaveFilter, personal-

10.21437/SPSC.2023-9

ized PercepNet, and Voicefilter rely on prior recordings from the targeted speaker to isolate speech. Another approach is to use inter-channel phase differences in a multichannel setup [10]. As targeted speech separation is a subgroup of traditional speech separation, we are also including state-of-the-art models from blind speech separation to allow a better comparison to the overall research area such as SepFormer [11] and TDANet [12].

The novelty of our work is dominant speaker isolation, for example in an open office, where we can make the assumption that the most dominant speaker in a speech signal, so one microphone, is also the target of the isolation algorithm. The practical implementation of a two-stage convolutional time-domain audio separation network shows that we can isolate a speaker without any prior information using audio streams from two devices, which we introduce in Section 2.2. This network first masks the interfering speakers and then enhances the filtered signal. An implementation with listening examples and the source code is openly available.[1]

2. Method

2.1. Signal Model

A mixture of multiple speakers $x_{\text{mix}}(n)$ over time n can be defined as

$$x_{\text{mix}}(n) = \sum_{c=1}^{N_s} s_c(n) + s_n(n), \tag{1}$$

where $s_c(n)$ denotes the cth contributing speech source, $s_n(n)$ the potential background noise, N_s is the number of speakers, and $1 \leq c \leq N_s$. Each of the overlapping audio signals $s_c(n)$ can further be described as the convolution of a clean speech signal with a corresponding room impulse response such as

$$s_c(n) = x_c(n) * h_c(n), \tag{2}$$

where $x_c(n)$ denotes a clean audio signal, $h_c(n)$ the impulse response and the asterisk $*$ denotes the convolution operator.

Our scenario has two speakers in *one* room with *two* ongoing but independent conversations with other people using two separate devices, see Figure 1. In this scenario, speech from one speaker leaks into the opposite speaker's microphone. This presents a privacy risk for the leaked speaker and degrades the speech quality. We propose to extract the desired speaker from the sum of overlapping speech signals by adding a processing step, which has access to the audio streams of each of the devices in the room. As the two people in this scenario are in the same room, they are aware of the potential to be leaked into the other person's microphone. Thus, we do not see a shared processing step as a privacy risk. Access to another's speech signal has been given indirectly by being in a shared room.

2.2. IsoNet

We call the network, which is used to isolate the dominant speaker, IsoNet. It consists of an encoder, mask-generator, enhancer, and decoder which is modelled after the Conv-TasNet, see Figure 2. The two audio signals are first fed sequentially into the encoder to window the utterance into smaller time frames. The weights of the encoder for both inputs are shared. The encoding is achieved with one linear convolutional layer without any activation function using 50% stride. Then, the framed interfering speech signal is fed into a temporal block which consists of multiple convolutional blocks (1-D Conv Blocks), which are similar to the Conv-TasNet. This temporal block represents a mask for the interfering speaker. The output of the temporal block is activated with a sigmoid function, such that the values contained in the mask are between 0 and 1. To remove the undesired speech contents, the mask is first inverted by subtracting 1 from the calculated mask values and then multiplied with the targeted signal channel. Subsequently, the quality of the masked signal is enhanced with a second temporal block following the same architecture. The enhanced signal is finally reconstructed with a linear convolutional layer with 50 % stride into the original representation.

2.3. Loss Function

Following Conv-TasNet [3], we also use the scale-invariant signal-to-noise ratio (SI-SNR) as the loss function, calculated as

$$\text{SI-SNR} := 10 \cdot \log_{10} \frac{||s_{target}||^2}{||e_{noise}||^2}, \tag{3}$$

in which the target signal is calculated with $s_{target} := \frac{(\hat{s}^T s)s}{||s||^2}$. Here, s is the reference and $\hat{s}$ the estimated isolated signal. Furthermore, the noise is defined as $e_{noise} := \hat{s} - s_{target}$ and $||\cdot||^2$ refers to the power of the signal.

3. Evaluation

3.1. Datasets

We use the LibriSpeech dev-clean dataset for training and testing [13]. It consists of approximately 1000 hours of audiobook recordings in English, sampled at 16 kHz. In total, there are 40 different speakers in the dataset, which have between 20 and 60 speech samples. All samples that are less than six seconds long are discarded. All remaining samples serve as possible speech samples for the room simulation. We choose two random speakers from the Librispeech dataset and place them in a virtual room with a random distance between each other. Then, we place microphones close to the position of each speaker to simulate a phone call scenario with a distance of 0.15 m. For the room simulations, we use the PyRoomAcoustics package [14]. In total, we generate 20000 samples in this fashion with randomized room and speaker settings, which results in 33 hours of training data. The height of the audio source is fixed to 1.5 m. The size of the simulated room was randomly chosen to have a length and width in the range of 5 m to 10 m and a height of 2.5 m to 5 m. Figure 3 shows the distance between microphones in the simulated room in ms, thus the delay between one speech signal arriving at the targeted and interfering microphone respectively. This shows that the speech signal does not need to be time-aligned to each other for a successful speech enhancement. Other public datasets such as WSJ0-2mix, and Libri2Mix can not be used in this work, as they do not provide a correlated and accurate simulation of the room where the speakers are placed in. Additionally, these datasets provide a single-channel input. This complicates the comparison between IsoNet and blind single-channel speech separation models.

3.2. Metrics

We use mutual information (MI) as a measure of privacy. MI is a useful metric to measure how much of the speech of the interfering speaker is still contained in the isolated speech signal.

[1] https://github.com/Speech-Interaction-Technology-Aalto-U/IsoNet

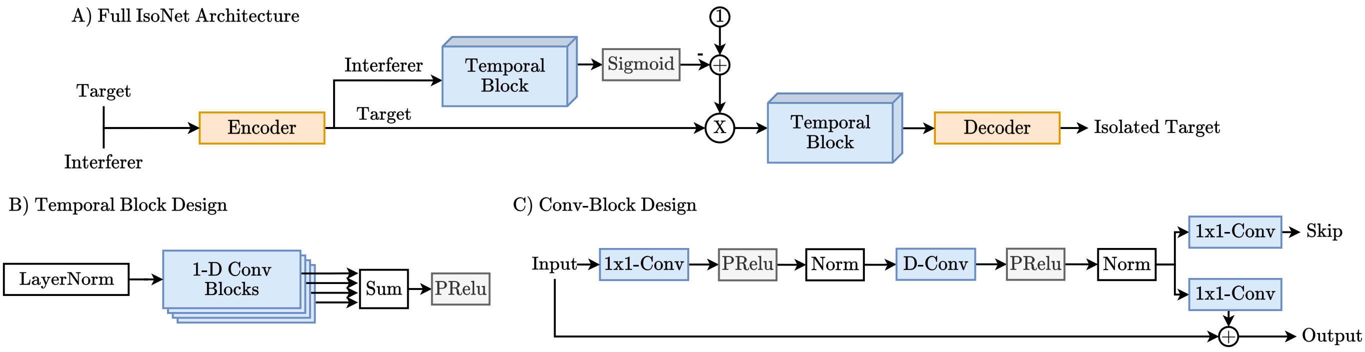

Figure 2: *Full IsoNet architecture, A) shows an abstract view of the whole architecture, while B) shows the structure of one temporal block. Lastly, C) shows the detailed architecture of one Conv-Block.*

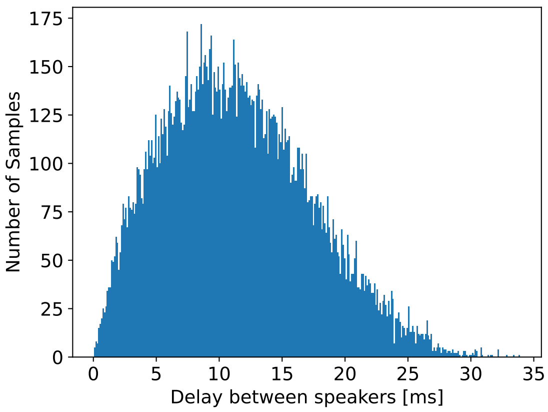

Figure 3: *Distribution of distances between microphones in ms.*

Method	PESQ	Δ SI-SNR	Δ SDR
WaveFilter	-	-	10.45
PercepNet	2.412	-	-
VoiceFilter	-	12.6	-
Conv-TasNet	3.22	12.2*	12.7*
SepFormer	-	16.5*	17*
TDANet	-	17.4*	17.9*
IsoNet	3.7	18.6	14.1

Table 1: *Comparison between state-of-the-art networks in speech separation tasks. IsoNet performs on par or better than the other baselines. The asterisk $*$ describes baseline models which are trained on the Libri2Mix Dataset. The reported values are the original ones from the paper, as there was no public model available for retraining.*

The lower the MI between the output of the network and the interfering speaker, the better the algorithm performed regarding privacy. MI is not often reported in other current research as the usual task is to separate a mixture of speech into its individual components, rather than optimizing for the smallest possible residue of one speaker in another speech signal [15].

We use three metrics to evaluate our network performance regarding its enhancements capability: SI-SNR, PESQ, and the MUSHRA listening tests. We also include metrics such as short-term objective intelligibility (STOI) [16], and signal-to-distortion ratio (SDR) to allow a better comparison with other baseline models. However, we identified perceptual evaluation of speech quality (PESQ) and the listening tests as methods that correlate better to the subjectively perceived speech quality. Prior work shows that the SDR tends to suffer from unaccounted channel errors and has a problem with scaling. [17]. PESQ is a widely used objective metric to evaluate speech quality [18] due to its simplicity to use and availability. This stands in contrast to its successor the Perceptual Objective Listening Quality Analysis (POLQA) which is not commonly available. The multiple stimuli with hidden reference and anchor (MUSHRA) listening test was designed to evaluate lossy audio coding algorithms [19]. In this test, see Table 4, listeners are presented simultaneously with different versions of the same audio signal in random order. Each of the degraded versions is evaluated against the reference so that even small differences between signals can be taken into account. Further, with the MUSHRA listening test, we can get statistically significant results even with a low number of listeners. Further, we calculate the mutual information between the clean leakage signal and the input and output of the Iso-Net respectively.

3.3. Experimental settings

The inputs to the network are four-second long utterances which are sampled at 16 kHz, high-pass filtered with a cutoff frequency of 70 Hz and mean normalized. There is no overlap between the frames of input samples. For all experiments, we used the Adam optimizer with a learning rate of 0.001 and a learning rate scheduler, as well as the SI-SNR loss function. We set the batch size to 64 and ran for a maximum of 200 epochs. For our activation of the masking functions we used a sigmoid activation and for all others a ReLU activation. The bias was deactivated in every layer. For a single temporal block, we used the found settings from Conv-TasNet [3].

4. Experimental results

Table 1 presents the result of the comparison between IsoNet and the baseline models. IsoNet outperforms all baseline models with the highest PESQ of 3.7 as well as the highest SI-SNR improvement of 18.6 dB. It has to be noted that a comparison between models can be difficult, as the initial signal-to-noise ratios in the training database differ for each model. Direct distribution of SNR as in other methods cannot be given in this

Method	No. Params	Fs
WaveFilter	5.9*	8 kHz
PercepNet	8.5/26.5	48 kHz
VoiceFilter	18.8*	8 kHz
Conv-TasNet	5.1	8 kHz
IsoNet	3.7	16 kHz
SepFormer	26	8 kHz
TDANet	2.3	8 kHz

Table 2: *Comparison between state-of-the-art networks in speech separation tasks. Just like the TDANet, IsoNet stands out with its small network size while showing the same speech quality. The asterisk * marks networks, which did not report network size and we estimated the complexity of the networks ourselves.*

Input Length	PESQ	std	STOI	std	SI-SNR	std
0.5 [s]	2.81	0.58	0.9	0.18	10.5	1.66
1 [s]	2.81	0.51	0.88	0.1	11.01	1.98
2 [s]	3.26	0.45	0.97	0.03	18.46	1.13
4 [s]	3.7	0.4	0.98	0.01	28.3	2.3

Table 3: *Evaluating the influence of the input length on the separation and speech quality performance of the network, std refers to the standard deviation.*

work, as the SNR is implicitly set by the room simulation.

Next to speech quality metrics, we evaluate resource consumption in terms of network size and hyperparameters. Table 2 shows that IsoNet is on par with the smallest network, the TDANet with 2.3 M parameters). This is a notable improvement since IsoNet has the added penalty of a higher sampling rate 16 kHz than TDANet, 8 kHz, which increases the complexity of IsoNet.

4.1. Varying Input Length

To find the optimal input length, we varied the input length by feeding speech utterances between 0.5 s and 4 s to the network. As the PESQ score does not drop in the same ratio as the input length, we conclude that despite the assumption that longer periods of audio files are necessary for effective speech separation, we can still achieve a sufficient separation ability even with sub-second input length (see Table 3).

4.2. Listening Tests

The enhancing performance is further evaluated with a subjective MUSHRA [19] listening test in which each subject listened to four versions of an audio signal; 1) the reference signal, 2) the output of IsoNet, 3) the original speech mixture that was recorded on the targeted microphone, and 4) an ideal ratio masked signal (IRM) [20]. Each audio signal was rated against the reference signal. The results in Figure 4 show that IsoNet improves the speech quality by 47.72 MUSHRA points compared to the original mixture. It also shows an improvement of 18.78 MUSHRA points compared to a speech mixture that was multiplied with an ideal ratio mask. To evaluate whether the improvements are statistically significant, we conducted a Kruskal-Wallis H-test between the Δ distributions. This test was chosen as the individual distributions were not normally distributed. The test shows p-values<0.05, which confirms that the differences in quality obtained in the listening test are statistically significant.

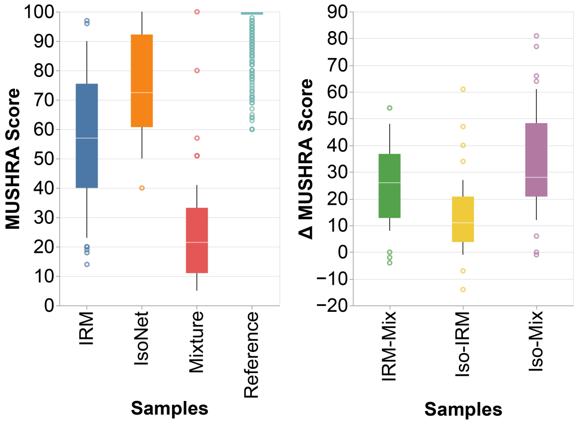

Figure 4: *MUSHRA listening test result with $N = 26$ listeners, where the bars represent the first and third quartiles. The mixture refers to the original mixed speech consisting of target and interferer. IRM refers to the ideal ratio masked signal. Δ refers to the difference between reference and IRM, mixture and IsoNet respectively.*

	Mean	Variance
Reference	96.38	7.85
IsoNet	75.08	18.01
IRM	56.3	24.57
Mixture	27.36	22.97

Table 4: *Perceived speech quality based on a MUSHRA test with $N = 26$ listeners. The test used four signals: 1) the reference, 2) the output of the IsoNet, 3) an IRM-masked signal and 4) the original speech mixture.*

No. Speakers	SI-SNR	std	PESQ	std
2	28.3	2.3	3.92	0.4
4	11	10.51	2.26	0.98
8	7.56	5.86	1.1	0.16

Table 5: *Influence of the number of overlapping speakers on the speech quality evaluated with SI-SNR and PESQ scores, std refers to the standard deviation. Adding more speakers clearly results in worse separation performance.*

4.3. Scaling with additional speakers

To further evaluate the separation performance in a scenario with more speakers than microphones, we evaluated the capability of IsoNet to remove multiple speakers from the speech mixture. The number of microphones was kept fixed at two. In Table 5 we can see that adding more speakers to the same mixture results in decreasing speech quality. Nonetheless, the degradation does not stand in a linear relation with the number of speakers in the room.

4.4. Privacy evaluation

As one targeted voice is isolated with the Iso-Net, the privacy for the removed speaker increases. We can measure how much of the interfering speech information is still contained in the target speech signal after enhancement by calculating the mutual information before and after the separation process, which can be seen in Table 6.

	MI [bits]	Variance
Before Iso-Net	3.52	0.177
After Iso-Net	1.39	0.151

Table 6: *Comparison of the mutual information before and after the targeted separation. 'Before', denotes the mutual information between the original speech mixture and the interfering speaker, whereas 'After Iso-Net' denotes the MI between the output of the network and the interfering speaker.*

The comparison between the mutual information in the original speech mixture and the isolated network output shows a reduction of 60 %. Here, the mutual information is calculated between the clean interfering speaker and the speech mixture and the network output, respectively. This leads to the conclusion that the removal of undesired speech information is successful as the number of bits which share similar information is reduced.

5. Conclusion

In this paper, we showed the ability to leverage multi-device audio for targeting the most dominant speaker in a speech signal. This is relevant in any scenario where multiple speakers have different conversations in the same room simultaneously. We proposed a novel convolutional neural network architecture based on the Conv-TasNet, called IsoNet. Listening tests confirmed the perceived speech quality is on par with state-of-the-art methods, but without requiring any additional or prior information such as target speaker embeddings or visual cues. Further, the IsoNet only uses 3.7 M model parameters which is only exceeded by the TDANet and increase the perceived speech quality by 48 MUSHRA points. The effective attenuation of a speaker who is not part of the ongoing interaction improves privacy, as only the relevant information for the interaction is transmitted which can be seen by the reduction of mutual information. Further improvement could potentially be gained by using room microphones instead of speaker-specific microphones to remove unwanted speech.

6. References

[1] P. Zarazaga, S. Das, T. Bäckström, V. V. R. Vegesna, and A. K. Vuppala, "Sound Privacy: A conversational speech corpus for quantifying the experience of privacy." in *Interspeech*, 2019, pp. 3720–3724.

[2] D. Wendt, R. K. Hietkamp, and T. Lunner, "Impact of noise and noise reduction on processing effort: A pupillometry study," *Ear and hearing*, vol. 38, no. 6, pp. 690–700, 2017.

[3] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[4] R. Gao and K. Grauman, "Visualvoice: Audio-visual speech separation with cross-modal consistency," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 15 490–15 500.

[5] R. Giri, S. Venkataramani, J. Valin, U. Isik, and A. Krishnaswamy, "Personalized PercepNet: Real-time, low-complexity target voice separation and enhancement," *arXiv preprint arXiv:2106.04129*, 2021.

[6] J. Zhao, S. Gao, and T. Shinozaki, "Time-domain target-speaker speech separation with waveform-based speaker embedding." in *Interspeech*, 2020, pp. 1436–1440.

[7] R. Rikhye, Q. Wang, Q. Liang, Y. He, and I. McGraw, "Multi-user VoiceFilter-Lite via attentive speaker embedding," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 275–282.

[8] T. Li, Q. Lin, Y. Bao, and M. Li, "Atss-Net: Target speaker separation via attention-based neural network," *arXiv preprint arXiv:2005.09200*, 2020.

[9] S. Mun, S. Choe, J. Huh, and J. Chung, "The sound of my voice: Speaker representation loss for target voice separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7289–7293.

[10] R. Gu, L. Chen, S. Zhang, J. Zheng, Y. Xu, M. Yu, D. Su, Y. Zou, and D. Yu, "Neural Spatial Filter: Target speaker speech separation assisted with directional information." in *Interspeech*, 2019, pp. 4290–4294.

[11] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," 2020. [Online]. Available: https://arxiv.org/abs/2010.13154

[12] K. Li, R. Yang, and X. Hu, "An efficient encoder-decoder architecture with top-down attention for speech separation," *arXiv preprint arXiv:2209.15200*, 2022.

[13] 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.

[14] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 351–355.

[15] T. M. Cover and J. A. Thomas, "Information theory and statistics," *Elements of information theory*, vol. 1, no. 1, pp. 279–335, 1991.

[16] 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.

[17] 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.

[18] 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.

[19] M. Schoeffler, S. Bartoschek, F. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, "webMUSHRA—A comprehensive framework for web-based listening tests," *Journal of Open Research Software*, vol. 6, no. 1, 2018.

[20] S. Xia, H. Li, and X. Zhang, "Using optimal ratio mask as training target for supervised speech separation," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 163–166.