



Location-Aware Target Speaker Extraction for Hearing Aids

Daniel-José Alcalá Padilla¹, Nils L. Westhausen^{2,1}, Swati Vivekananthan¹, Bernd T. Meyer¹

¹Communication Acoustics and Cluster of Excellence Hearing4all, Carl von Ossietzky Universität Oldenburg, Germany

²Bose Corporation, USA

daniel-jose.alcala.padilla@uni-oldenburg.de,
nils.westhausen@bose.com, swati.vivekananthan@uni-oldenburg.de,
bernd.meyer@uni-oldenburg.de

Abstract

Target speaker extraction (TSE) using deep learning offers potential benefits for hearing-impaired listeners. However, their implementation in hearing aids requires low-latency, low-complexity algorithms capable of real-time operation. Existing models that comply with these requirements use multi-channel input from binaural hearing aids to improve speech intelligibility but are limited to extracting speech from a single fixed direction. In this work, we utilize the direction of arrival (DOA) of the target speaker to extract speech at arbitrary angles in complex acoustic environments based on a deep-learning model. We introduce a novel DOA encoding method based on complex exponentials which is compared to one-hot (*oh*) encoding. We explore three low-complexity methods for integrating DOA information into the model. The evaluation using objective measures demonstrates that our extended model outperforms the baseline system with our novel encoding method achieving superior performance in 16 out of 21 cases.

Index Terms: target speaker extraction, direction of arrival, steering, hearing aids

1. Introduction

Complex acoustic scenarios with multiple speakers, noise and reverberation are often very challenging for hearing impaired (HI) listeners. While hearing aids can increase speech intelligibility [1] and decrease listening effort [2] to some extent, they cannot fully eliminate the problems associated with speech-in-speech masking. In recent years, many neural network (NN)-based approaches for speaker separation have been proposed which are able to extract individual speech signals from superimposed speech signals [3–8]. While these models significantly extend the state-of-the-art for speaker separation, they are not compatible with the requirements of hearing aids, i.e., real-time capability and compatibility with small-footprint hardware. These requirements were considered with the Group Communication Binaural Filter and Sum network (GCBFSnet) [9], which performs real-time speaker separation with low latency and small model sizes. The latter is achieved by using data grouping and parameter sharing in combination with group communication (GC) [10]. The network extracts speech signals from two-speaker mixtures by applying filter-and-sum beamforming in the time-frequency domain. In listening experiments carried out in [9], GCBFSnet increased speech intelligibility in HI listeners and outperformed established processing methods found in hearing aids such as the minimum variance distortionless response (MVDR) beamformer [11]. GCBFSnet can be easily adapted to perform target speaker extraction (TSE). However, it can only extract speech from a single fixed direction, when the target speaker is in front of the listener (i.e. at

0° azimuth). In reality, conversation partners typically alternate between making eye contact and looking away, as part of our natural social behaviour [12]. Further, facing away from the speaker can maximize spatial release from masking and increase intelligibility [13]. Models for extracting target speech using directional information require the position of the target speaker, which is usually provided as the target’s direction of arrival (DOA). By additionally processing target DOA information, a model could be steered towards it in order to extract the target speech signal. However, the optimal method for integrating DOA information into a model and the interplay between model architecture and DOA input requires further investigation. In an approach to TSE using an autoencoder-type architecture, feature data was scaled by multiplication with the cosine of the azimuth DOA, which resulted in selective enhancement for sources restricted to the frontal hemisphere [14]. In [15], DOA was encoded in embeddings based on a one-hot vector. These embeddings then served as initial states for the first of two Long Short-Term Memory (LSTM) layers in order to achieve steerable speaker separation.

In this work, a critical extension for GCBFSnet is explored to provide spatial information based on the target speaker’s DOA. To the best of our knowledge, this work presents the first demonstration of DOA-based steering integrated into a binaural speech separation model that meets the real-time, low-latency, low complexity requirements for hearing aids. We explore several approaches of integrating DOA information to the model and benchmark it against a competitive baseline. Additionally, we propose a novel DOA encoding strategy that enables efficient representation of directional information while allowing speech extraction from all directions. All methods retain the properties of the original GCBFSnet, such as causal processing with low latency and a small model size.

2. Methods

2.1. Architecture

The network architecture, as illustrated in Fig. 1, consists of the original GCBFSnet architecture [9] combined with our proposed extensions for DOA processing. During training, oracle DOA information, represented as azimuth angle in degrees, is provided to the neural network. This information is encoded using one of two encoding strategies described in Section 2.2. The encoded DOA information is then processed in DOA modules and induced into the feature data of an intermediate stage within GCBFSnet. Four DOA processing strategies are explored in this work to integrate DOA information into the GCBFSnet architecture. The positions of the DOA module in the architecture are chosen as they proved to be most promising in preliminary tests.

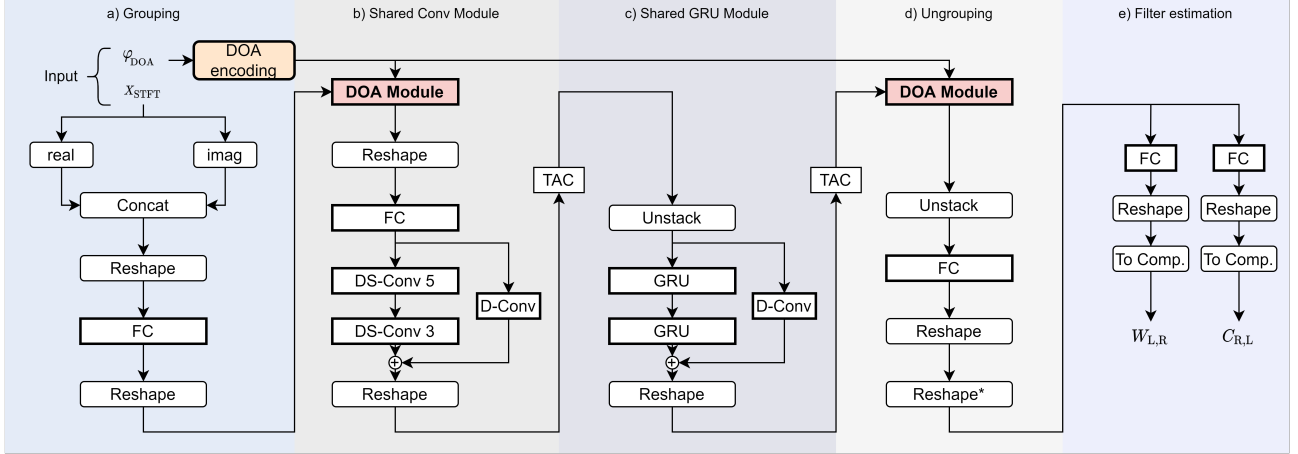


Figure 1: *GCBFSnet* architecture including DOA processing. The model contains five modules and uses group communication (GC) and weight sharing. The Grouping module (a), encodes DOA information and scales, projects and distributes the feature vector to different groups. The DOA module processes DOA information and is introduced in the Shared Conv Module (b) and in the Ungrouping module (c). The Shared Conv Module has two depth-wise separable convolution layers (DS-Conv) and a skip connection with a depthwise convolution (D-Conv) to capture short-term temporal changes. Group communication is implemented with transform average concatenate (TAC). The Shared conv module contains two Gated Recurrent Units (GRUs) to capture long-term temporal information. The ungrouping module combines the groups to form a latent representation. The filter estimation module estimates the real and imaginary parts of the filter weights W and C which are predicted by fully connected (FC) layers with tanh activations.

2.2. DOA encoding

Target DOA information is fed into the NN as the azimuth angle φ in degrees, with $\varphi \in \{0^\circ, 1^\circ, \dots, 359^\circ\}$. DOA is encoded using two techniques, namely one-hot (*oh*) encoding and exponential encoding. As proposed in [15], for *oh* encoding, a vector is created where each element represents one of the 360 discrete azimuth angles with 1° resolution. Elements are ordered from 0° to 359° and are all set to zero, except for the element corresponding to φ , which is set to 1. *Oh* encoding is common for handling multi-class data in machine learning applications and produced good results for DOA encoding [15]. However, to use a more efficient encoding method that makes use of a more compact representation, we propose a new encoding strategy based on the complex exponential $e^{i\varphi}$, referred to as *exp* in this work. *Exp* encoding consists of a vector

$$DOA_{\text{enc}} = [\Re\{e^{i\varphi}\}, \Im\{e^{i\varphi}\}], \quad (1)$$

containing the real and imaginary parts of the complex exponential. Since this representation is derived from the unit circle, the relation between angular directions is reflected realistically. Furthermore, it only uses two values to encode continuous-valued directions, whereas *oh* encoding requires as many values as the number of representable discrete directions.

2.3. DOA processing

This section introduces three methods for DOA processing as well as one baseline method. All methods receive the encoded DOA information DOA_{enc} and feature data X of an intermediate processing stage within *GCBFSnet* as input. The three proposed methods are applied at positions marked with “DOA Module” in Fig. 1, where the shape of X is (B, G, t, H) with batch size $B = 32$, number of groups $G = 8$, number of time frames $t = 2001$ and a hidden size $H = 32$. Unlike the baseline method, DOA modules retain the shape of X between their input and output, which allows them to be inserted without al-

tering existing parts of the NN architecture.

2.3.1. FILM

This method is based on feature-wise linear modulation (FiLM) [16]. First, embeddings γ and β are each separately generated by processing DOA_{enc} with a FC layer and parametric rectified linear unit (PReLU) for activation. The feature vector X is reshaped to (t, B, G, H) , allowing the feature-wise transformation to be applied uniformly across all time steps t . Both γ and β are reshaped to (B, G, H) and multiplied and added to X :

$$Y = \gamma \odot X + \beta, \quad (2)$$

where \odot and $+$ denote the elementwise multiplication and addition, respectively. Finally, the output feature data Y are reshaped to the original shape of X , (B, G, t, H) .

2.3.2. Scale

This method was derived from the idea to use a single value for scaling X , as described in [14], in order to minimize the required number of model parameters. In this study, we use an additional additive bias. This makes *Scale* a sparse version of FiLM with γ and β both being single scalar values, generated from a shared FC layer with PReLU activation. As with FiLM, equation 2 is applied where \odot and $+$ represent multiplication and addition of γ and β with all elements of X , respectively. This method requires the lowest number of model parameters, with only 7 for *exp* encoding.

2.3.3. Concat

As the name suggests, this method is based on the concept of concatenation. In this method, DOA_{enc} is first fed to an FC layer with PReLU activation to generate DOA embeddings (DOA_{emb}) of size $E = 10$ and shape (B, E) . The (DOA_{emb}) is inflated by copying values from the E dimension for every time frame (t times) to generate embeddings of size (B, t, E) .

X is reshaped to (B, t, GH) and the embedding is concatenated along the (GH) dimension. Finally, the result is given as input to a FC layer with PReLU activation, and X is reshaped back to its original shape (B, G, t, H) .

2.3.4. InitStates (baseline)

Originally applied in an LSTM-based architecture, the approach by [15] is used with GRU as a baseline method in this work. DOA_{enc} is processed with a FC layer and PReLU to yield an embedding which then serves as the initial hidden states for the first of two GRU layers.

3. Training

3.1. Training data

Training and validation data consist of 4-second-long, 4-channel mixture signals of three speakers and noise in a spatial setting. Speech and noise signals originate from the 2nd Deep Noise Suppression (DNS) Challenge [17]. 60k rooms were simulated with a room acoustics simulator- RAZR [18] of which 45k are used for training, 10k for validation and 5k for testing (evaluation). Room properties were drawn randomly from uniform distributions, i.e. with room sizes from 12 to 100 m². Three speakers, a noise source and a listener position were randomly placed inside each room. Reverberation times were taken from a lognormal distribution $T_R \sim \text{Lognormal}(\ln 0.45, (\ln 1.4)^2)$, which is loosely based on reverberation times found in common spaces of daily life, as reported by [19].

An in-house head-related transfer function (HRTF) database was used to simulate a listener wearing hearing aids. This HRTF was recorded through the four microphones of two hearing aids from the portable hearing laboratory [20] set on a head-and-torso simulator. After including this HRTF into the room simulation, 4-channel impulse responses (IR) were rendered for each sound source of each room. RAZR allows to separately access the full IR h as well as parts of the IR that correspond to the direct sound h_{direct} , the early h_{early} and late reflections h_{late} . Noise signals are convolved with h_{late} to generate diffuse-like noise. The three speech signals, including the target speech, are convolved with h and, together with the noise, added up at signal-to-noise ratios (SNR), drawn from a uniform distribution $U(-8, 8)$ dB. Overall, 40% of the mixtures contain both interferer speech and noise, whereas mixtures with only one make up 30% each. A clean and dereverberated target speech signal is created by first applying an exponential window to a concatenation of h_{direct} and h_{early} , and then convolving the result with the target speech signal. This way, only the earliest reflections, known to be beneficial for speech intelligibility [21], are included in an otherwise anechoic target speech signal.

Evaluation data is based on WHAM! [22], a popular dataset containing speech signals mixed with urban environmental noise. In this work, the signals were spatialized by using the 5k simulated rooms from the test split. A version with three-speaker mixtures was created from WSJ0-3mix [5] by adapting the python scripts for generating WHAM!, provided by [22]. Scaling factors for the speech files were also generated using a provided script and the WSJ0-mix3 corpus, which lead to normally distributed SNR with $\mathcal{N}(1, 4.5^2)$ dB.

3.2. Training procedure

Training is performed for 100 epochs on 80k training and 3k validation utterances. The compressed mean squared error (cMSE) is used as the loss function. It is defined as

$$L_{cMSE} = (1 - \alpha) \left| |\hat{X}|^c - |X|^c \right|^2 + \alpha |\hat{X}^c - X^c|^2 \quad (3)$$

where X and \hat{X} are the true and estimated short-time Fourier transform (STFT) of the target speech signal. The exponent $c = 0.3$ is used to apply power-law compression, which decreases the dominance of large values [9, 23]. A weight factor $\alpha = 0.3$ balances the influence between complex and magnitude STFT features. For the loss calculation, STFTs with a frame length of 20 ms and a frame shift of 10 ms were calculated. In contrast, the STFT features used for processing within the NN were extracted with a frame length of 4 ms and 2 ms frame shift. Model parameters are optimized using ADAM [24] with an initial learning rate of $1e^{-3}$. The learning rate is multiplied by 0.98 after every epoch and further multiplied by 0.8 if the validation loss did not decrease for five consecutive epochs. AutoClip [25] with a cutoff percentile $p = 10$ is used for gradient clipping.

4. Evaluation

4.1. Baselines

An adaption of the fully-convolutional time-domain audio separation network (Conv-TasNet) [3] is referred to as GCMiMoBiTasNet and is used as a baseline model, which was suggested by [9] to obtain a suitable baseline for comparison with GCBFSnet. Adaptations were made to ensure a comparable model size by using parameter sharing and GC as well as generating binaural estimates of the target speech. GCMiMoBiTasNet is used with $G = 8$ groups and eight convolutional blocks that are repeated four times. DOA processing methods are applied after the last block of every repeat. Since GCMiMoBiTasNet does not contain any recurrent NN layer, it cannot be used with the InitStates DOA processing method.

4.2. Evaluation metrics

Models are evaluated with three common objective metrics: Scale-invariant signal-to-distortion ratio (SI-SDR) [26] is a popular measure for source separation performance in dB. For estimates of speech intelligibility, the revised second version of the Hearing Aid Speech Perception Index (HASPI) [27] is used with hearing thresholds set to 0 dB HL to simulate normal hearing. HASPI is computed for both ears, after which the higher (better) value is reported as the final score to account for better-ear listening. The Modified Binaural Short-Time Objective Intelligibility (MBSTOI) [28] is a measure of speech intelligibility that takes binaural effects into account.

5. Results & Discussion

Mean improvement scores obtained for 1000 evaluation utterances are listed in Table 1. Overall, scores indicate that all DOA processing methods enable GCBFSnet to successfully perform TSE with clear improvements of SI-SDR scores and speech intelligibility estimates. For GCBFSnet, $FILM_{exp}$ achieves the highest scores with $\Delta SI-SDR = 9.93$ dB and absolute HASPI score improvements of 0.208. Even $Scale_{exp}$, which scores lowest for GCBFSnet, still achieves acceptable results, considering that it only adds 14 model parameters to GCBFSnet.

Table 1: Evaluation results with mean improvements of metric scores relative to those of the unprocessed input signals. Mean absolute scores for the unprocessed signals are -11.41 dB for SI-SDR, 0.068 for HASPI and 0.362 for MBSTOI. Arrows indicate whether the desired outcome is an increase or decrease of scores. Model size is stated as the number of parameters with k denoting units of one thousand. Best scores for each base model and performance metric are highlighted in bold type.

Model	DOA Embedding	Size ↓	ΔSI-SDR ↑	ΔHASPI ↑	ΔMBSTOI ↑	
GCBFSnet	none	276k	2.36	0.029	0.008	
	FILM	exp	279k	9.93	0.208	0.117
		oh	645k	9.06	0.160	0.090
	Scale	exp	276k	8.86	0.147	0.086
		oh	277k	8.92	0.150	0.087
	Concat	exp	412k	9.52	0.165	0.095
		oh	420k	9.20	0.179	0.098
	InitStates	exp	276k	9.46	0.168	0.098
oh		368k	9.24	0.155	0.088	
GCMiMoBiTasNet	none	247k	2.72	0.024	0.008	
	FILM	exp	248k	8.21	0.106	0.083
		oh	385k	8.10	0.085	0.070
	Scale	exp	247k	9.07	0.143	0.102
		oh	249k	7.27	0.062	-0.045
	Concat	exp	261k	9.18	0.204	0.115
		oh	272k	8.89	0.157	0.095

Therefore, the choice of a suitable DOA processing method is not critical for GCBFSnet, even if $FILM_{exp}$ is the preferred method among those evaluated in this work. This is, however, not the case for the model configurations based on GCMiMoBiTasNet, where differences in performance scores are much more pronounced. Here, $Concat_{exp}$ scores highest on all metrics with $\Delta SI-SDR = 9.18$ dB and HASPI improvements of 0.204. In comparison, $FILM_{exp}$, which scored highest with GCBFSnet, only achieves around half of this improvement for HASPI scores with $\Delta HASPI = 0.106$. $Scale_{oh}$ even leads to decreased MBSTOI scores, compared to the unprocessed input mixtures. From this, we conclude that the choice of a suitable DOA processing method heavily depends on the base model it is used with. We can also conclude from Fig. 2, that with larger azimuth distances between the target speaker and interfering speaker, the model performance improves.

All model configurations that use exp encoding yield smaller model sizes than their corresponding counter part using oh encoding, with size differences of up to a factor of 2.3. Despite this, in 16 out of 21 cases, exp encoding leads to better performance scores than oh encoding. These results indicate that exp is, indeed, a suitable encoding strategy for DOA. Even for the baseline method $InitStates$, which was originally used with LSTM and oh encoding [15], scores highest with exp encoding on all metrics.

Considering the strict requirements for a usage in hearing aids, the additional computational complexity and latency introduced to GCBFSnet by using $FILM_{exp}$ are expected to be minimal. Compared to original GCBFSnet, only 3,076 parameters are added, which equals a 1.12% increase in model size. Extended GCBFSnet uses STFT features at double the frame length and frame shift of original GCBFSnet, with 4 ms and 2 ms, respectively. The algorithmic latency of a hearing aid setup with original GCBFSnet was measured at 5.4 ms [11], meaning it would increase to at least 7.4 ms for the extended version. However, running extended GCBFSnet on shorter STFT frames should be tested in the future.

Another limitation of the current approach is the assumption of access to oracle DOA information, which is not feasible in real-world applications. The model may also be sensitive to

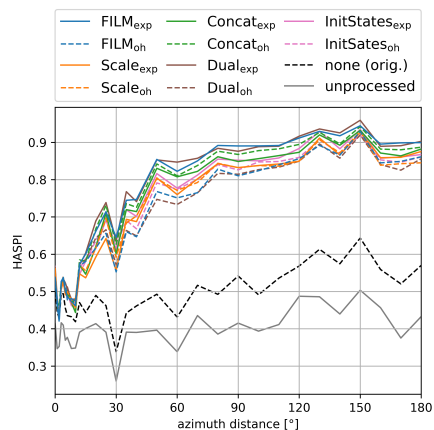


Figure 2: HASPI Scores for different azimuth distances in degrees between the target speaker and interfering speaker

DOA estimation inaccuracies, particularly in dynamic or multi-speaker environments. As the model has only been evaluated for three speaker conditions, the performance may vary when there are more speakers present. We will address these limitations in future work, including the integration of DOA estimation methods and evaluation of performance of the model in scenarios with more speakers present.

6. Conclusions

This paper introduced a crucial extension for TSE with deep learning in hearing aids: The original GCBFSnet, is a speaker separation model that does not utilize any directional information. As the GCBFSnet is compatible with hearing aid constraints, it has been adapted to perform TSE in this work. Results show a strong interaction between the NN architecture and how DOA information is fed to the model. A novel encoding strategy utilising exponentials was introduced in this work. The effectiveness of exp encoding could be confirmed as it lead to higher performance scores in comparison to oh encoding, despite smaller model sizes. We explored multiple methods of inducing DOA information into the model, all of which resulted in steerable TSE with noticeable improvements on estimates of speech intelligibility, despite some loss of speech quality. For the GCBFSnet, FILM with exp encoding was determined most suitable, outperforming the baseline method.

7. Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2177/1 - Project ID 390895286 and by - Project-ID 352015383 - SFB 1330 C6.

8. References

- [1] J. Löhler, B. Akcicek, B. Wollenberg, R. Schönweiler, L. Verges, C. Langer, U. Machate, R. Noppeney, K. Schultz, J. Kleeberg, B. Junge-Hülsing, L. E. Walther, P. Schlattmann, and A. Ernst, "Results in using the freiburger monosyllabic speech test in noise without and with hearing aids," *European archives of oto-rhino-laryngology : official journal of the European Federation of Oto-Rhino-Laryngological Societies (EUFOS) : affiliated with the German Society for Oto-Rhino-Laryngology - Head and Neck Surgery*, vol. 272, no. 9, pp. 2135–2142, 2015.
- [2] E. M. Picou, T. A. Ricketts, and B. W. Y. Hornsby, "How hearing aids, background noise, and visual cues influence objective listening effort," *Ear and hearing*, vol. 34, no. 5, pp. e52–64, 2013.
- [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] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: Efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 46–50.
- [5] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016, pp. 31–35. [Online]. Available: <https://www.merl.com/publications/TR2016-003>
- [6] Y. Luo, C. Han, N. Mesgarani, E. Ceolini, and S.-C. Liu, "Fasnet: Low-latency adaptive beamforming for multi-microphone audio processing," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 260–267.
- [7] Z.-Q. Wang, G. Wichern, S. Watanabe, and J. Le Roux, "Stft-domain neural speech enhancement with very low algorithmic latency," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 397–410, 2022.
- [8] R. Sinha, M. Tammen, C. Rollwage, and S. Doclo, "Speaker-conditioning single-channel target speaker extraction using conformer-based architectures," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2022, pp. 1–5.
- [9] N. L. Westhausen and B. T. Meyer, "Binaural multichannel blind speaker separation with a causal low-latency and low-complexity approach." [Online]. Available: <http://arxiv.org/pdf/2312.05173v1>
- [10] Y. Luo, C. Han, and N. Mesgarani, "Ultra-lightweight speech separation via group communication." IEEE, 2021.
- [11] N. L. Westhausen, H. Kayser, T. Jansen, and B. T. Meyer, "Real-time multichannel deep speech enhancement in hearing aids: Comparing monaural and binaural processing in complex acoustic scenarios," *arXiv preprint arXiv:2405.01967*, 2024.
- [12] A. Kendon, "Some functions of gaze-direction in social interaction," *Acta psychologica*, vol. 26, pp. 22–63, 1967.
- [13] J. A. Grange and J. F. Culling, "The benefit of head orientation to speech intelligibility in noise," *The Journal of the Acoustical Society of America*, vol. 139, no. 2, pp. 703–712, 2016.
- [14] A. Briegleb, M. M. Halimeh, and W. Kellermann, "Exploiting spatial information with the informed complex-valued spatial autoencoder for target speaker extraction," pp. 1–5, 2023. [Online]. Available: <http://arxiv.org/pdf/2210.15512v2>
- [15] K. Tesch and T. Gerkmann, "Multi-channel speech separation using spatially selective deep non-linear filters," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 542–553, 2024. [Online]. Available: <http://arxiv.org/pdf/2304.12023v2>
- [16] E. Perez, F. Strub, H. d. Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer." [Online]. Available: <http://arxiv.org/pdf/1709.07871v2>
- [17] C. K. A. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "Icassp 2021 deep noise suppression challenge," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6623–6627.
- [18] T. Wendt, S. van de Par, and S. Ewert, "A computationally-efficient and perceptually-plausible algorithm for binaural room impulse response simulation," *Journal of the Audio Engineering Society*, vol. 62, no. 11, pp. 748–766, 2014.
- [19] J. Traer and J. H. McDermott, "Statistics of natural reverberation enable perceptual separation of sound and space," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 113, no. 48, pp. E7856–E7865, 2016.
- [20] BatAndCat Sound Labs, "Portable hearing laboratory." [Online]. Available: <https://batandcat.com/portable-hearing-laboratory-phl.html>
- [21] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *The Journal of the Acoustical Society of America*, vol. 113, no. 6, pp. 3233–3244, 2003.
- [22] 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." [Online]. Available: <http://arxiv.org/pdf/1907.01160v1>
- [23] K. Wilson, M. Chinen, J. Thorpe, B. Patton, J. Hershey, R. A. Saurous, J. Skoglund, and R. F. Lyon, "Exploring tradeoffs in models for low-latency speech enhancement," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 366–370.
- [24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization." [Online]. Available: <http://arxiv.org/pdf/1412.6980v9>
- [25] P. Seetharaman, G. Wichern, B. Pardo, and J. Le Roux, "Auto-clip: Adaptive gradient clipping for source separation networks," in *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2020, pp. 1–6.
- [26] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr – half-baked or well done?" IEEE, 2019.
- [27] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (haspi) version 2," *Speech Communication*, vol. 131, pp. 35–46, 2021.
- [28] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions," *Speech Communication*, vol. 102, pp. 1–13, 2018. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0167639317302947>