frontiers | Frontiers in Signal Processing

TYPE Original Research
PUBLISHED 10 September 2024
DOI 10.3389/frsip.2024.1413983

OPEN ACCESS

EDITED BY
Tsubasa Ochiai,
Nippon Telegraph and Telephone, Japan

REVIEWED BY
Chengshi Zheng,
Chinese Academy of Sciences (CAS), China
Li Li,
CyberAgent, Japan

*CORRESPONDENCE
Mingsian R. Bai,
msbai@pme.nthu.edu.tw

RECEIVED 08 April 2024
ACCEPTED 21 August 2024
PUBLISHED 10 September 2024

CITATION
Chang H, Hsu Y and Bai MR (2024) Deep beamforming for speech enhancement and speaker localization with an array response-aware loss function.
Front. Sig. Proc. 4:1413983.
doi: 10.3389/frsip.2024.1413983

Deep beamforming for speech enhancement and speaker localization with an array response-aware loss function

Hsinyu Chang[1], Yicheng Hsu[2] and Mingsian R. Bai[1,2]*

[1]Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan, [2]Department of Power Mechanical Engineering, National Tsing Hua University, Hsinchu, Taiwan

Recent research advances in deep neural network (DNN)-based beamformers have shown great promise for speech enhancement under adverse acoustic conditions. Different network architectures and input features have been explored in estimating beamforming weights. In this paper, we propose a deep beamformer based on an efficient convolutional recurrent network (CRN) trained with a novel ARray RespOnse-aWare (ARROW) loss function. The ARROW loss exploits the array responses of the target and interferer by using the ground truth relative transfer functions (RTFs). The DNN-based beamforming system, trained with ARROW loss through supervised learning, is able to perform speech enhancement and speaker localization jointly. Experimental results have shown that the proposed deep beamformer, trained with the linearly weighted scale-invariant source-to-noise ratio (SI-SNR) and ARROW loss functions, achieves superior performance in speech enhancement and speaker localization compared to two baselines.

KEYWORDS

multichannel speech enhancement, speaker localization, loss function, deep learning, beamforming

1 Introduction

Speech enhancement (SE) (Zheng et al., 2023) aims at extracting the clean speech signals from the noisy mixture, which is essential for various applications such as hands-free communication, hearing aids, teleconferencing, etc. However, under adverse acoustic conditions such as reverberation and interference, the enhancement performance can be significantly degraded. Thanks to the advent of deep neural network (DNN) technology, learning-based monaural SE algorithms (Valin, 2018)– (Schröter et al., 2022) have emerged with great promise in noise reduction.

DNN-based beamformers can be divided into two categories. One category is to integrate the DNN with a beamformer, referred to in this study as the two-stage weight estimation approach (Heymann et al., 2015; Nakatani et al., 2017)– (Boeddeker et al., 2017; Boeddeker et al., 2018). In the first stage, the spatial covariance matrices (SCM) of speech and noise signals are computed through time-frequency (T-F) masking estimated by a DNN. The computed SCMs are then used in the second stage to compute array weights according to various optimal beamforming design criteria (Capon, 1969)– (Warsitz and Haeb-Umbach, 2007; Souden et al., 2010). However, numerical instability may arise if matrix inversion is required. To mitigate this problem, an All Deep Learning MVDR (ADL-MVDR) network is proposed in (Zhang et al., 2021), where the matrix operations are replaced by two recurrent neural networks (RNNs). Another category (Xu et al., 2021)

attempts to estimate array weights directly through the DNN. Many DNN architectures have been suggested for estimating optimal filter weights, e.g., the multiple-in-multiple-out (MIMO) U-net structure (Ren et al., 2021) and the complex-valued spatial autoencoder (COSPA) structure (Halimeh and Kellermann, 2022). Several input features that carry spatio-spectral information for weight estimation have also been investigated (Xiao et al., 2016a; Xiao et al., 2016b)– (Li et al., 2022; Liu et al., 2022). However, these learning-based methods focus only on speech enhancement and do not consider localization issues.

Chen et al. (Chen et al., 2022) integrate an auxiliary localization module into MIMO-Deep Complex Convolution Recurrent network (MIMO-DCCRN) to perform speech enhancement and localization jointly. The signal processing-based localization module (SPLM) and the neural localization module (NLM) are compared under different conditions. However, both localization modules require grid search. A localization error may occur if the speaker is not located at one of the preselected grid points.

In this study, we propose a deep beamformer capable of jointly performing speech enhancement and speaker localization. We believe that addressing speech enhancement together with speaker localization provides mutually beneficial information from an array signal processing perspective. The system is based on a convolutional recurrent network (CRN) (Braun et al., 2021). Instead of using an auxiliary module NLM as in (Chen et al., 2022), we train the DNN with a loss function of weighted objectives including a scale-invariant source-to-noise ratio (SI-SNR) and an array response-aware (ARROW) loss. From the point of view of array signal processing (Stoica and Moses, 2005), the ARROW loss adopts the ground truth relative transfer functions (RTFs) of the target speaker and interferer for better enhancement and localization performance. In particular, the weighting parameters used in the ARROW loss function are thoroughly examined from the perspectives of enhancement and localization. The main contributions of this paper can be summarized as follows.

1) We present a combination of SI-SNR and ARROW loss functions designed for multichannel speech enhancement and speaker localization.
2) We investigate the impact of that the weighting parameters in the proposed loss function on speech enhancement and speaker localization.
3) We show that the introduction of the ground truth RTFs improves the performance and the robustness of localization in the presence of unseen room impulse responses (RIRs).

The remainder of this paper is organized as follows. In Section 2, the problem formulation and the signal model are introduced. In Section 3, the proposed method is presented in detail. The experimental setup and results are described in Section 4. The paper is concluded in Section 5.

2 Problem formulation and signal model

Consider an array of M microphones receiving speech signal and noise signal from a farfield speaker and an interferer. The noisy signal $\mathbf{Y} \in \mathbb{C}^{M\times 1}$ captured by the microphone array can be written in the short-time Fourier transform (STFT) domain as

$$\mathbf{Y}(l,f) = \mathbf{R}_s(f)S(l,f) + \mathbf{R}_n(f)N(l,f) + \mathbf{v}(l,f), \tag{1}$$

where $S(l,f)$ and $N(l,f)$ denote the target speech signal and the interferer corresponding to the frequency bin index f and the time frame index l, $\mathbf{R}_s \in \mathbb{C}^{M\times 1}$ and $\mathbf{R}_n \in \mathbb{C}^{M\times 1}$ denote the relative transfer functions (RTFs) associated with the target speaker and the interferer, respectively. $\mathbf{v} \in \mathbb{C}^{M\times 1}$ denotes the noise term comprising diffuse noise such as late reverberation.

We seek to enhance the fgsignal $\hat{S}$ by using a filter-and-sum beamformer with array weights, $\mathbf{W} \in \mathbb{C}^{M\times 1}$:

$$\hat{S}(l,f) = \mathbf{W}^H(l,f)\mathbf{Y}(l,f), \tag{2}$$

where superscript "^{H}" denotes the conjugate-transpose operator.

3 Proposed system

In this section, we describe a deep beamformer (DB) that is capable of performing jointly the enhancement and localization tasks. Figure 1 shows the DB system diagram, where a DNN is used to directly estimate the beamforming weights for subsequent enhancement and localization. In the training phase (indicated by the dashed blue box), the ground truth RTFs, the time-frequency domain target speaker signal, and the time-domain target speech received by the reference microphone are used to compute the weighted loss, as detailed next.

3.1 Loss function

To perform jointly learning-based enhancement and localization, we propose an ARray RespOnse-aWare (ARROW) loss function for training the DNN unit in Figure 1. We motivate the development of the ARROW by starting with the scale-invariant source-to-noise ratio (SI-SNR) (Luo and Mesgarani, 2019) loss function for the multichannel speech enhancement:

$$\mathcal{L}_{\text{SI-SNR}} = -10\log_{10}\frac{\|\eta\mathbf{s}\|_2^2}{\|\hat{\mathbf{s}}-\eta\mathbf{s}\|_2^2}, \eta = \frac{\langle\hat{\mathbf{s}},\mathbf{s}\rangle}{\|\mathbf{s}\|_2^2}, \tag{3}$$

where $\{\hat{\mathbf{s}},\mathbf{s}\} \in \mathbb{R}^{1\times T}$ are the vectors of the inverse STFTs of $\{\hat{S}(l,f), S(l,f)\}$, respectively, $\langle\cdot\rangle$ denotes the inner product between two vectors, and $\|\cdot\|_2$ is the Euclidean norm. Using the array signal model in Equation 1, the equivalent optimal solution of the SI-SNR in the frequency domain can be written as

$$\begin{aligned}\hat{S}(l,f) &= \mathbf{W}^H(l,f)\mathbf{Y}(l,f)\\ &= \mathbf{W}^H(l,f)[\mathbf{R}_s(f)S(l,f) + \mathbf{R}_n(f)N(l,f) + \mathbf{v}(l,f)]\\ &= \eta S(l,f) \Rightarrow [\mathbf{W}^H(l,f)\mathbf{R}_s(f) - \eta]S(l,f)\\ &+ \mathbf{W}^H(l,f)[\mathbf{R}_n(f)N(l,f) + \mathbf{v}(l,f)] = 0\end{aligned}, \tag{4}$$

Thus, minimizing the SI-SNR loss would only partially fulfill the distortionless constraint

$$\mathbf{W}^H(l,f)\mathbf{R}_s(f) \approx \eta, \tag{5}$$

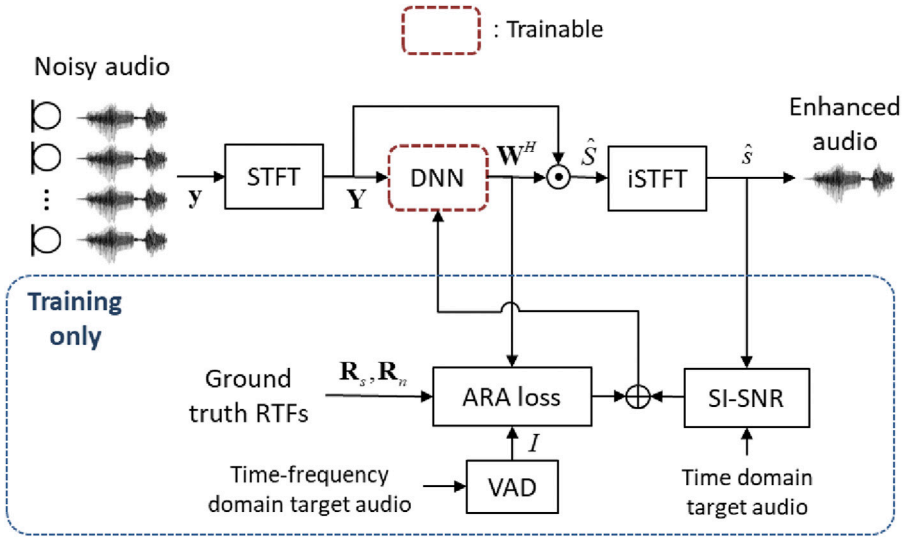

FIGURE 1
The proposed deep beamformer. O and indicate the operations of Equation 2, 5. The processing units in the dashed blue box are used only in the training phase.

with some of the effort going into reducing the interference and noise.

To further improve the enhancement performance and to provide localization information, an ARray RespOnse-aWare (ARROW) loss function is introduced as follows:

$$\mathcal{L}_{\text{ARROW-}\alpha} := \alpha \frac{1}{L_{tp}F} \sum_{if} \left[I(l) \left| \text{Im}\{\mathbf{W}^H(l,f)\mathbf{R}_s(f)\} \right| \right] + (1-\alpha) \frac{1}{L_{ta}F} \sum_{lf} \Big[(1-I(l)) \left(\left(\text{Re}\{\mathbf{W}^H(l,f)\mathbf{R}_n(f)\} \right) + \left(\text{Im}\{\mathbf{W}^H(l,f)\mathbf{R}_n(f)\} \right| \right) \Big] \quad (6)$$

where Re{·} and Im{·} denote the real and imaginary part operators, $I(l) \in \{0,1\}$ is the indicator of a voice activity detector (VAD), $\alpha \in [0,1]$ is a weighting factor that weights the target and interference terms, L_{tp} and L_{ta} are the number of frames corresponding to the target speech present and absent periods, and F is the number of frequency bins. In the VAD module, time-frequency bins with a signal-to-interference ratio (SIR) greater than 0 dB are considered speaker active bins ($\alpha = 1$), while those with an SIR below this threshold are considered speaker silent bins ($\alpha = 0$). Note that the first term of the loss function in Equation 5 is intended to "clean up" the imaginary part of the distortionless constraint in Equation 6, while the second term is intended to further reduce the array response associated with the unwanted directional interference. A natural question is why the distortionless constraint is not directly incorporated into the loss function in Equation 7. We found it difficult to train our DNN model with this setting due to the scaling problem and some potential conflicts with the SI-SNR loss.

To formulate the complete loss function, we combine the SI-SNR and ARROW loss functions with linear weighting

$$\mathcal{L} = \beta \mathcal{L}_{\text{SI-SNR}} + (1-\beta)\mathcal{L}_{\text{ARROW-}\alpha}, \quad (7)$$

where the weighting factor $\beta \in [0,1]$.

3.2 Localization

For localization of the target speaker, the following beampattern function is defined:

$$P(\theta) = \frac{1}{L_{tp}F} \sum_{f} \sum_{l} I(l) \left| \mathbf{W}^H(l,f)\mathbf{a}_\theta(f) \right|, \quad (8)$$

where $\mathbf{W}(l,f)$ is the array weights obtained from DNN, $\mathbf{a}_\theta$ denotes the free-field plane-wave steering vector at the angle θ which ranges from 30° to 150° in 15° increments and L_{tp} are the number of frames corresponding to the speech present periods. Note that we only consider the time when the target speaker is active ($I(l) = 1$).

It follows that the direction of arrival (DOA) of the speaker can be obtained by finding the peak of the beampattern function:

$$\hat{\theta}_s = \underset{\theta}{\text{argmax}}\, P(\theta), \quad (9)$$

3.3 Deep beamforming network (DBnet)

The DNN unit in Figure 1 is implemented in a convolutional recurrent neural network (CRNN) architecture illustrated in Figure 2, hereafter referred to as the deep beamforming network (DBnet). The beamformer weights can be estimated directly from the microphone signals using DBnet. The stacked real and imaginary parts of the microphone signals are the input data to the encoder. The decoder layer produces the array weights as output. In Figure 2, the DBnet structure consists of four symmetric convolutional and deconvolutional encoder and decoder layers with a 16-32-64-64 filter. To reduce computational complexity, the separable convolution (Howard et al., 2017) is chosen for each convolutional block. Each convolutional block is followed by a batch normalization and ReLU activation. Tanh activation is used at the last layer. The 1 × 1 pathway convolutions are used with

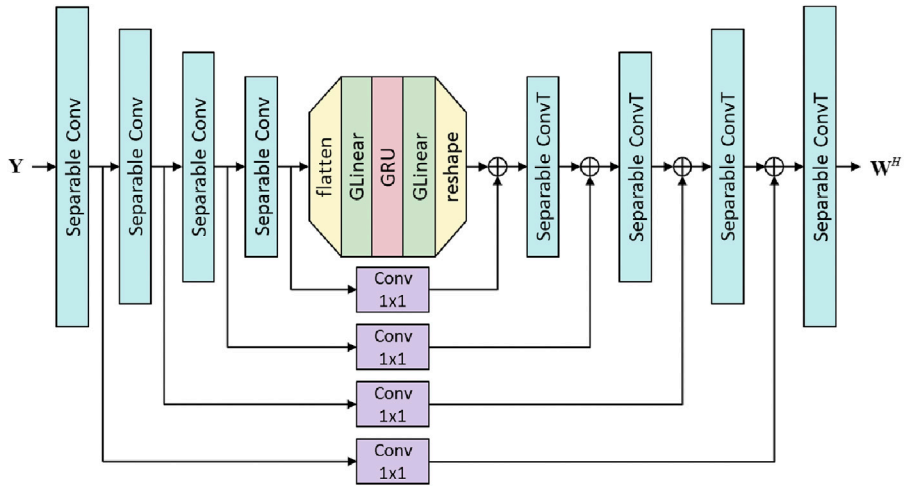

FIGURE 2
The architecture of DBnet.

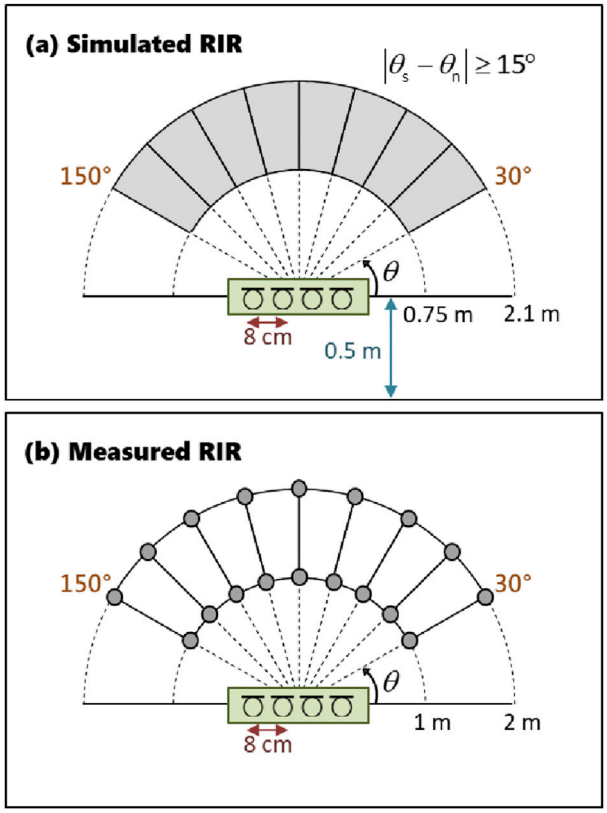

FIGURE 3
Experimental setup for **(A)** training and **(B)** testing of the proposed deep beamformer.

add-skip connections (Braun et al., 2021; Schröter et al., 2022), which allows for considerable parameter reduction with little performance degradation. The bottleneck consists of a grouped linear (GLinear) layer (Schröter et al., 2022). A single 256-unit GRU layer is used to capture the temporal information. The proposed DBnet has only 688.32 K parameters and 177.08 MMACs per second. It is worth noting that the proposed method focuses only on the loss function to improve the enhancement and localization performance. Consequently, the model size and computational complexity remain unchanged during the test phase.

4 Experimental study

The proposed DB system is evaluated through the tasks of speech enhancement and speaker localization. To see the robustness of the proposed system to unseen acoustic conditions, we train our neural network using the simulated RIRs, but test it using the measured RIRs.

4.1 Datasets

Clean speech utterances are selected from the LibriSpeech corpus (Panayotov et al., 2015), where the subsets *train-other-500*, *dev-clean*, and *test-clean* are adopted for training, validation, and testing. The noise clip used as the directional interferer is selected from the Microsoft Scalable Noisy Speech Dataset (MS-SNSD) (Reddy et al., 2019) and the Free Music Archive (FMA) (Defferrard et al., 2017). In the MS-SNSD dataset, only non-speech and directional interferences are considered in the data preparation. These include Air Conditioner, Copy Machine, Munching, Shutting Door, Squeaky Chair, Typing, Vacuum Cleaner, and Washer Dryer. Each training and testing signal mixture is prepared in the form of a 6-s clip randomly inserted with a 4-s clean speech clip. The training and validation sets comprise the signals with signal-to-interference ratio (SIR) randomly selected between −10 and 15 dB. The testing set

consists of noisy signals with SIR = −5, 0, 5, and 10 dB. In addition, sensor noise is added with signal-to-noise ratio (SNR) = 20, 25, and 30 dB. A four-element uniform linear array (ULA) with an inter-element spacing of 8 cm is used in the experiment. Reverberant speech signals are simulated by convolving the clean signals with RIRs generated by the image source method (Habets, 2010). Various reverberation times, (T60) = 0.2, 0.3, 0.4, 0.5, 0.6, and 0.7 s, are used. As illustrated in Figure 3A, the distance between the target speaker and interferer is randomly selected in the frontal plane at the ring sector bounded by radius = 0.75 and 2.1 m. In addition, any two sources are separated at least 15° apart from each other. The Multichannel Impulse Response Database (Hadad et al., 2014), recorded at Bar-Ilan University using an eight-element ULA with an inter-element spacing of 8 cm for T60 = 0.16 s, 0.36 s, and 0.61 s, is adopted as the test set. In this study, we use only the RIRs of the four center microphones to generate the reverberant signals for testing. As shown in Figure 3B, the target speaker and the interferer appear randomly in any two of nine angular directions equally spaced between 30° and 150° in 15° increments. A total of 30,000, 3,000 and 7,200 samples are used for training, validation and testing.

4.2 Baseline methods

The results presented in Ref. (Chen et al., 2022). have shown that the MIMO DCCRN method can significantly outperform the single-channel DCCRN (Hu et al., 2020) and MIMO U-net (Ren et al., 2021) when implemented with SPLM. Therefore, this study will focus on comparing the MIMO DBnet systems with different loss functions. Two baselines are used for comparison with the proposed system. All models are implemented in the DBnet architecture. The first baseline is the DBnet trained with the SI-SNR loss. For a fair comparison, a DBnet cascaded with SPLM (Chen et al., 2022) is used as the second baseline, because SPLM does not require additional parameters for training. Here, SPLM-9 refers to the SPLM with nine predefined zones. All datasets are generated at a sampling rate of 16 kHz. The signals are transformed to the STFT domain using a 25-m Hamming window with a 10-m hop size, and 512-point fast Fourier transform. The Adam optimizer is utilized in the training phase, with a learning rate of 0.001.

4.3 Enhancement performance evaluation

We use DNSMOS P.835 (Reddy et al., 2022) to evaluate the speech enhancement performance. Three mean opinion scores based on P.835 human ratings are used to assess the speech quality (SIG), background noise quality (BAK), and overall quality (OVRL). These metrics can be used to comprehensively investigate the trade-off between noise reduction and distortion caused by the tunable factors of the proposed ARROR loss. First, we examine the effects of weighting β between the SI-SNR loss and the ARROW loss on enhancement performance. As can be seen in Figure 4, a large β leads to an

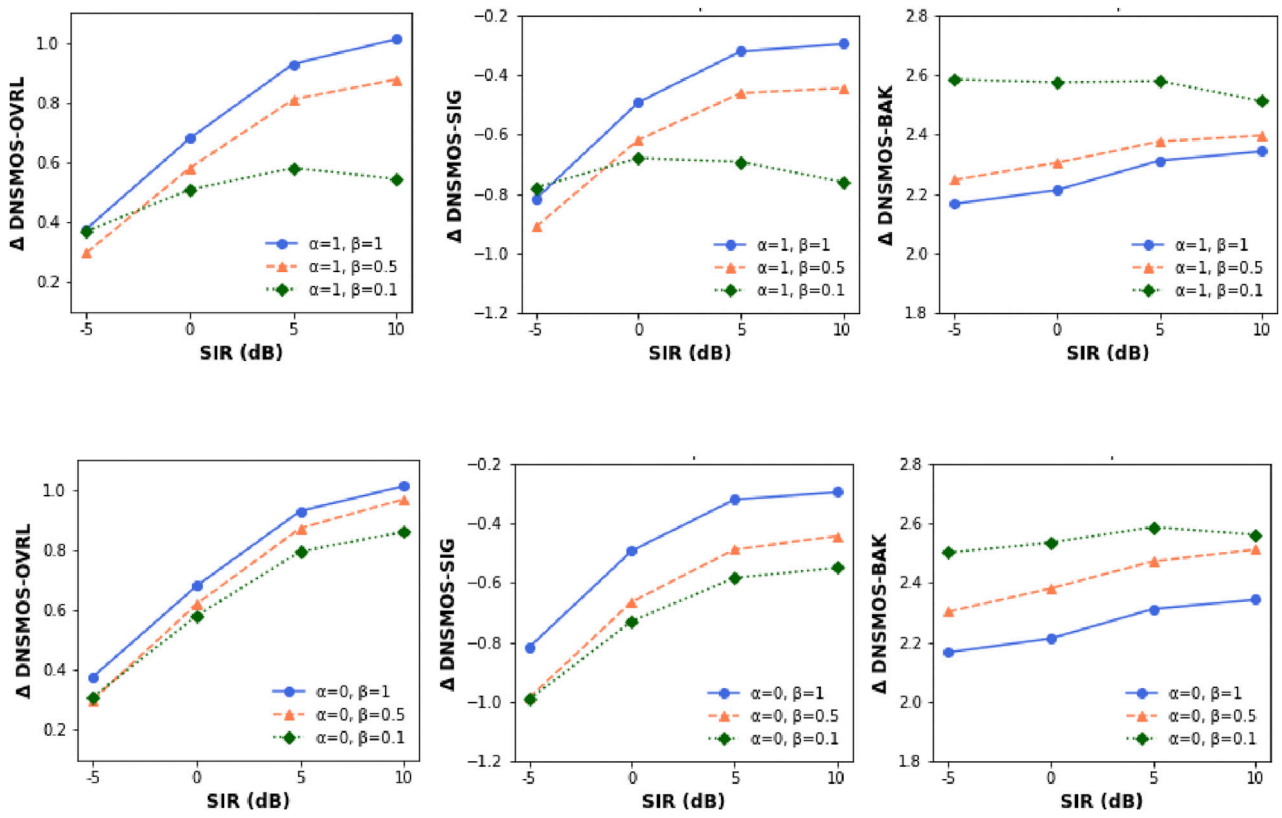

FIGURE 4
Enhancement performance for different ß factors to weight SI-SNR and ARROW loss.

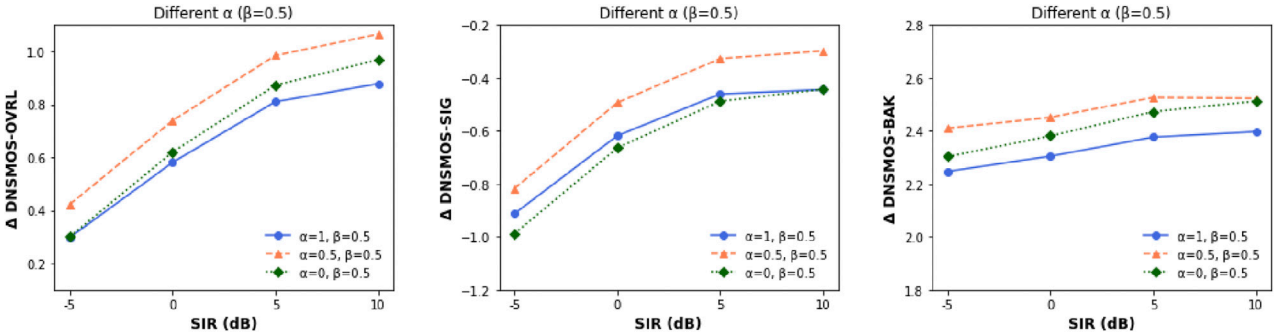

FIGURE 5
Enhancement performance for different a factors to weight the ARROW loss.

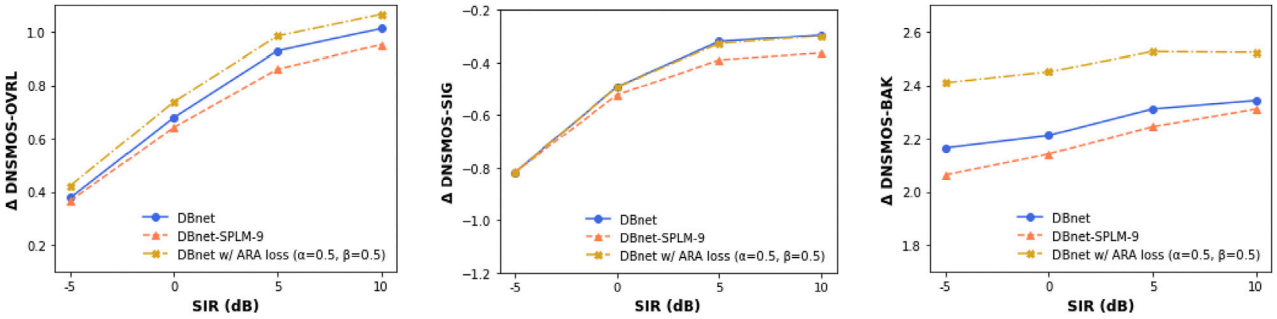

FIGURE 6
Enhancement performance of the proposed method and the baselines.

increased overall quality (OVL) and a signal quality (SIG) at the expense of increased background noise (BAK). Next, we examine the ARROW loss with different α factors, with a fixed weighting factor $\beta = 0.5$. The results in Figure 5 show that the optimal enhancement performance is achieved when both weighting factors are set to 0.5. These results suggest that the target speech and the interference terms in the loss function are equally important for speech enhancement.

Next, we compare the enhancement performance of the proposed system when $\alpha = 0.5, \beta = 0.5$ with baselines. The results in Figure 6 show that the proposed DB system performs the best in terms of all evaluation indices. Note that DBnet with SPLM performs worse than the original DBnet. This is due to the fact that the steering vector used in SPLM is based on the freefield plane wave model, which can lead to mismatch when applied to a reverberant environment. In summary, the method trained with the proposed ARROW loss can lead to much improved enhancement performance compared to the original DBnet method, by choosing appropriate weighting factors. In addition, to further suppress interference, time-frequency masking can indeed be applied after deep filtering. However, the trade-off between distortion and noise reduction needs to be carefully considered based on the specific requirements of subsequent applications.

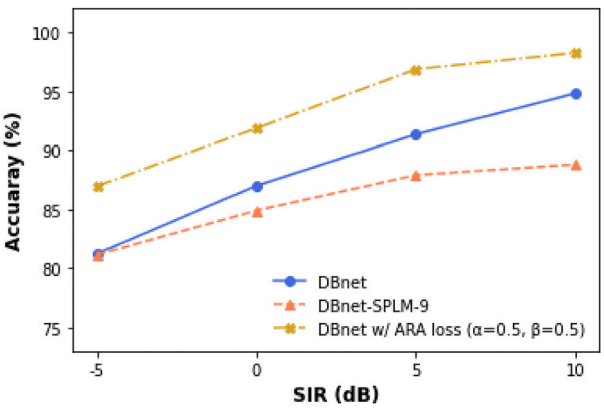

FIGURE 7
Localization performance of the proposed method and the baselines.

4.4 Localization performance evaluation

In this section, we evaluate the localization performance of the proposed DBnet with ARROW loss in comparison with two baselines (DBnet with SI-SNR loss and DBnet with SPLM).

To quantify the localization performance, we use the accuracy metric defined as

$$\text{Accuracy} = \frac{L_{true}}{L_{tp}} \times 100\%, \tag{10}$$

where L_{true} is the number of frames for which the angle estimation error is less than 15°, and L_{tp} is the total number of frames with speaker active. As shown in Figure 7, incorporating the ARROW loss results in superior speaker localization, with an average improvement of 5%. In addition, the DBnet with the SPLM is outperformed by the DBnet trained with only SI-SNR loss due to the free-field steering vector used in training. Therefore, training the DBnet with the proposed ARROW loss allows for more robust localization than cascading with an SPLM. Furthermore, the results presented in Figure 7 show that the DBnet trained with the proposed ARROW loss maintains approximately 98% localization accuracy at an SIR of 10 dB over various RT60s. This indicates that the localization performance of the model trained with the ARROW loss is mainly affected by the interferer, but less affected by the RT60.

5 Conclusion

In this study, we have proposed a deep beamforming system capable of speech enhancement and localization. A novel ARROW loss inspired by the distortionless constraint is proposed to effectively address these two tasks. The results have shown that the model trained with SI-SNR and ARROW loss provides superior enhancement and localization even when RIRs are not included in the training set. The future research agenda includes challenging scenarios with moving and multiple speakers. In future work, we will extend the proposed ARROR loss to address multiple sources. This extension will allow the model trained with this loss to handle scenarios with increased interference and multiple speakers.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

HC: Conceptualization, Formal Analysis, Investigation, Methodology, Writing–original draft, Writing–review and editing. YH: Conceptualization, Methodology, Validation, Writing–original draft, Writing–review and editing. MB: Supervision, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by the National Science and Technology Council (NSTC), Taiwan, under the project number 113-2221-E-007-057-MY3.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Boeddeker, C., Erdogan, H., Yoshioka, T., and Haeb-Umbach, R. (2018). *Exploring practical aspects of neural mask-based beamforming for far field speech recognition*. Proc. IEEE ICASSP, 6697–6701.

Boeddeker, C., Hanebrink, P., Drude, L., Heymann, J., and Haeb Umbach, R. (2017). Optimizing neural-network supported acoustic beamforming by algorithmic differentiation. *Proc. IEEE ICASSP.*, 171–175.

Braun, S., Gamper, H., Reddy, C. K. A., and Tashev, I. (2021). Towards efficient models for real-time deep noise suppression. *Proc. IEEE ICASSP.*, 656–660.

Capon, J. (1969). High-resolution frequency-wavenumber spectrum analysis. *Proc. IEEE* 57 (8), 1408–1418. doi:10.1109/proc.1969.7278

Chen, Y., Hsu, Y., and Bai, M. R. (2022). *Multi-channel end-to-end neural network for speech enhancement, source localization, and voice activity detection*. arXiv preprint arXiv:2206.09728.

Defferrard, M., Benzi, K., Vandergheynst, P., and Bresson, X. (2017). FMA: a dataset for music analysis. *Proc. Int. Soc. Music Inf. Retr. Conf.*, 316–323.

Erdogan, H., Hershey, J. R., Watanabe, S., Mandel, M., and Le Roux, J. (2016). Improved MVDR beamforming using single channel mask prediction networks. *Proc. Interspeech*, 1981–1985.

Habets, E. A. (2010). Room impulse response generator. *Tech. Univ. Eindh. Tech. Rep.*

Hadad, E., Heese, F., Vary, P., and Gannot, S. (2014) "Multichannel audio database in various acoustic environments," in *2014 14th international workshop on acoustic signal enhancement (IWAENC)*, 313–317.

Halimeh, M. M., and Kellermann, W. (2022). Complex-valued spatial autoencoders for multichannel speech enhancement. *Proc. IEEE ICASSP*, 261–265.

Heymann, J., Drude, L., Chinaev, A., and Haeb-Umbach, R. (2015). "BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge," in IEEE Workshop on Automatic Speech Recognition and Understanding, 444–451.

Heymann, J., Drude, L., and Haeb-Umbach, R. Neural network based spectral mask estimation for acoustic beamforming. *Proc. IEEE ICASSP*, 196–200.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., et al. (2017). *Mobilenets: efficient convolutional neural networks for mobile vision applications*. arXiv preprint arXiv:1704.04861.

Hu, Y., Liu, Y., Lv, S., Xing, M., Zhang, S., Fu, Y., et al. (2020). DCCRN: deep complex convolution recurrent network for phase-aware speech enhancement. *Proc. Interspeech*, 2472–2476.

Li, A., Liu, W., Zheng, C., and Li, X. (2022). Embedding and beamforming: all-neural causal beamformer for multichannel speech enhancement. *Proc. IEEE ICASSP.*, 6487–6491.

Liu, W., Li, A., Wang, X., Yuan, M., Chen, Y., Zheng, C., et al. (2022). A neural beamspace-domain filter for real-time multi-channel speech enhancement. *Symmetry* 14 (6), 1081. doi:10.3390/sym14061081

Luo, Y., and Mesgarani, N. (2019). Conv-TasNet: surpassing ideal time-frequency magnitude masking for speech separation. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 27, 1256–1266. doi:10.1109/taslp.2019.2915167

Nakatani, T., Ito, N., Higuchi, T., Araki, S., and Kinoshita, K. (2017). Integrating DNN-based and spatial clustering-based mask estimation for robust MVDR beamforming. *Proc. IEEE ICASSP*, 286–290.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: an ASR corpus based on public domain audio books. *Proc. IEEE ICASSP*, 5206–5210.

Reddy, C. K., Beyrami, E., Pool, J., Cutler, R., Srinivasan, S., and Gehrke, J. (2019). A scalable noisy speech dataset and online subjective test framework. *Proc. Interspeech*, 1816–1820.

Reddy, C. K. A., Gopal, V., and Cutler, R. (2022). DNSMOS P.835: a non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. *Proc. IEEE ICASSP*, 886–890.

Ren, X., Zhang, X., Chen, L., Zheng, X., Zhang, X., Guo, L., et al. (2021). A causal U-net based neural beamforming network for real-time multi-channel speech enhancement. *Proc. Interspeech*, 1832–1836.

Schröter, H., Escalante-B, A. N., Rosenkranz, T., and Maier, A. (2022). Deepfilternet: a low complexity speech enhancement framework for full-band audio based on deep filtering. *Proc. IEEE ICASSP.*, 7407–7411.

Souden, M., Benesty, J., and Affes, S. (2010). On optimal frequency-domain multichannel linear filtering for noise reduction. *IEEE Trans. Audio, Speech, Lang. Process.* 18, 260–276. doi:10.1109/tasl.2009.2025790

Stoica, P., and Moses, R. L. (2005). *Spectral Analysis of signals*. Upper Saddle River, NJ, USA: Prentice-Hall.

Valin, J.-M. (2018) "A hybrid DSP/deep learning approach to real-time full band speech enhancement," in *2018 IEEE 20th international workshop on multimedia signal processing (MMSP)*, 1–5.

Valin, J.-M., Isik, U., Phansalkar, N., Giri, R., Helwani, K., and Krishnaswamy, A. (2020). A perceptually-motivated approach for low-complexity, real-time enhancement of fullband speech. *Proc. Interspeech*, 2482–2486.

Warsitz, E., and Haeb-Umbach, R. (2007). Blind acoustic beamforming based on generalized eigenvalue decomposition. *IEEE Trans. Audio, Speech, Lang. Process.* 15, 1529–1539. doi:10.1109/tasl.2007.898454

Xiao, X., Watanabe, S., Chng, E. S., and Li, H. (2016b) "Beamforming networks using spatial covariance features for far-field speech recognition," in *Proc. 2016 asia-pacific signal and information processing association annual s ummit and conference (APSIPA)*, 1–6.

Xiao, X., Watanabe, S., Erdogan, H., Lu, L., Hershey, J., Seltzer, M. L., et al. (2016a). Deep beamforming networks for multi-channel speech recognition. *Proc. IEEE ICASSP*, 5745–5749.

Xu, Y., Zhang, Z., Yu, M., Zhang, S., and Yu, D. (2021). Generalized spatial-temporal RNN beamformer for target speech separation. *Proc. Interspeech*, 3076–3080.

Zhang, Z., Xu, Y., Yu, M., Zhang, S.-X., Chen, L., and Yu, D. (2021). ADL-MVDR: all deep learning MVDR beamformer for target speech separation. *Proc. IEEE ICASSP*, 6089–6093.

Zheng, C., Zhang, H., Liu, W., Luo, X., Li, X. D., Moore, B. C. J., et al. (2023). Sixty years of frequency-domain monaural speech enhancement: from traditional to deep learning methods. *Trends Hear.* 27, 23312165231209913. Trends in Hearing. doi:10.1177/23312165231209913