



Spatially Weighted Contrastive Learning for Robust Sound Source Localization

Hyun-Soo Kim, Da-Hee Yang, Joon-Hyuk Chang[†]

Department of Electronic Engineering, Hanyang University, Republic of Korea

{hyuns00803, douxi15, jchang}@hanyang.ac.kr

Abstract

We propose a spatially weighted contrastive loss (SWeC loss) for sound source localization in real-world scenarios using multi-channel speech data. In multi-channel localization, phase differences between microphone channels provide critical cues for estimating the azimuth angle of incoming speech. To effectively extract azimuth information, we leverage contrastive learning and introduce a novel loss function that incorporates spatial relationships between azimuth classes. Specifically, our loss assigns weights to negative pairs based on their angular distance, penalizing high similarity between embeddings corresponding to distant angles. Furthermore, we propose a contrastive data generation method tailored to multi-channel localization, enhancing the effectiveness of contrastive learning. Experimental results demonstrate that the proposed loss function and data generation strategy significantly improve localization performance.

Index Terms: sound source localization, contrastive learning

1. Introduction

Sound source localization (SSL) is the task of estimating the direction of an incoming sound using multi-channel audio signals. Traditional SSL techniques leverage spatial cues, such as phase differences across microphone channels, to estimate the azimuth angle of the sound source. For instance, generalized cross-correlation with phase transform [1] and steered response power [2, 3], have been widely used for multi-channel sound localization. Additionally, subspace-based methods like multiple signal classification [4, 5] improve localization accuracy by estimating signal subspaces. However, in real-world environments, factors like noise and reverberation significantly degrade localization performance, necessitating the development of more robust SSL methods. Therefore various SSL approaches have been explored to address these challenges. Deep learning-based approaches [6, 7, 8, 9, 10, 11] have been introduced to enhance SSL robustness by learning spatial representations from multi-channel inputs. While these models demonstrate improved generalization in complex acoustic environments, SSL under noisy and reverberant conditions remains an open challenge.

One promising direction for improving SSL robustness is contrastive learning (CL), which has been successfully applied to audio-visual localization. In video-based sound source localization, contrastive learning methods [12, 13, 14] align audio and visual embeddings to identify sounding objects in a scene. By pushing apart irrelevant audio-visual pairs while bringing related pairs closer in the embedding space, these methods effectively capture the correlation between sound and visual cues.

Inspired by this, we hypothesize that a contrastive approach can also be beneficial in multi-channel SSL. Specifically, it can help distinguish speech-related signals from interfering sound sources, such as reflections caused by reverberation or coherent noise, by clearly separating them in the embedding space. At the same time, contrastive learning should bring together embeddings associated with the same azimuth direction, reinforcing the correlation between directional cues.

Although contrastive learning has not been extensively explored for SSL, Li *et al.* [15] introduced a contrastive learning framework to extract azimuth embeddings from multi-channel signals. However, their approach was limited to narrow-band synthetic signals in controlled environments without reverberation or coherent noise. As a result, its applicability to real-world multi-channel speech localization remains uncertain. This raises the key question: *How can contrastive learning be effectively adapted to SSL in noisy and reverberant environments?*

To address this, we propose a spatially weighted contrastive loss (SWeC loss) tailored for incorporating spatial characteristics across azimuth classes. Our approach introduces three key advancements:

- **Spatially weighted negative sampling:** We incorporate azimuth-aware weighting in the contrastive loss to consider the similarity between embeddings of adjacent negative angles, preserving local azimuth relationships.
- **Multi-level sharpness control:** To balance class separation and inter-class alignment, we introduce adaptive sharpness weights that regulate embedding distribution.
- **Contrastive data augmentation:** We generate hard positive and negative samples to enhance the robustness of contrastive learning in noisy and reverberant environments.

To the best of our knowledge, this is the first study to apply contrastive learning to multi-channel speech-based SSL under realistic conditions. Our proposed method effectively learns azimuth embeddings that are robust to challenging acoustic conditions, paving the way for more accurate and generalizable SSL models.

2. Preliminaries

2.1. Contrastive learning

Contrastive learning aims to generate effective data representations by encouraging *positive* pairs to be closer and *negative* pairs to be farther apart in the embedding space. Chen *et al.* [16] introduced a widely adopted self-supervised contrastive learning framework, later extended by Khosla *et al.* [17] to incorporate supervised contrastive learning. The general framework consists of the following steps: *data augmentation module* - *feature encoder (FE)* - *projection head* - *contrastive loss function*.

[†] Corresponding author.

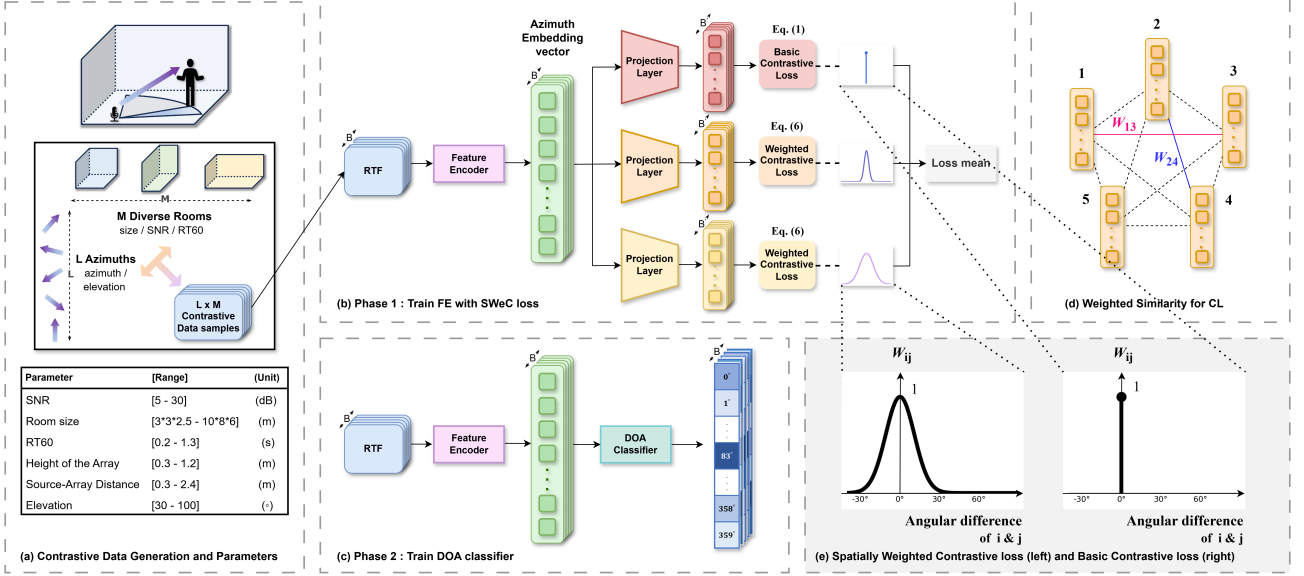


Figure 1: Illustration of the proposed approach pipeline. The dataset is generated in a contrastive manner by combining L azimuths and M room configurations, resulting in a total of B samples. Contrastive loss is computed by evaluating the similarity between all projection output pairs. Spatial weighting is applied for the SWeC loss.

Supervised contrastive learning leverages label information to determine positive and negative pairs. Specifically, all samples sharing the same class label are considered *positives*, while samples from different classes are treated as *negatives*. The contrastive loss function maximizes the similarity between positive pairs by placing their combined similarity in the numerator and normalizing it with the sum of all similarities (including positives and negatives) in the denominator:

$$\sum_{i \in I} \mathcal{L}_i = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_p / \tau)}{\sum_{a \in A(i)} \exp(\mathbf{z}_i \cdot \mathbf{z}_a / \tau)}. \quad (1)$$

Here, $I = \{1, \dots, 2N\}$ is the set of all sample indices in the mini-batch, $2N$ represents the batch size, and $P(i)$ denotes the set of positive sample indices (denoted as p) for anchor i . The cardinality of this set is denoted as $|P(i)|$. The set $A(i)$ includes all indices except i , and \mathbf{z} represents the output of the projection head; \mathbf{z}_i for anchor, \mathbf{z}_p for positive, \mathbf{z}_a for all samples except anchor i . The inner product operation (\cdot) computes the similarity between embedding, and τ is a temperature parameter that scales the contrastive loss.

2.2. Spatial feature extraction

We approximate the relative transfer functions (RTFs) [18] using a time-compressed instantaneous RTF [19, 20] under the assumption of a static speaker.

Let $Y_c(t, f)$ denote the time-frequency representation of the signal captured by microphone c at the time frame t and frequency bin f . Assuming a dominant single-source in each frame, the captured signal can be approximated as:

$$Y_c(t, f) \approx H_c(f) S(t, f) + N_c(t, f), \quad (2)$$

where $H_c(f)$ represents the acoustic transfer function (ATF) from the source to microphone c , $S(t, f)$ is the source signal in the short-time Fourier transform (STFT) domain, and $N_c(t, f)$ accounts for noise or reverberant residue, respectively. To extract spatial information, inspired by [21], we estimate the spatial covariance matrix $R(f) \in \mathbb{C}^{C \times C}$, where C is the number

of microphone channels. The (i, j) -th entry of this matrix is computed as:

$$R_{i,j}(f) = E \left[Y_i(f) Y_j(f)^H \right] = \frac{1}{T} \sum_t Y_i(t, f) Y_j(t, f)^* \quad (3)$$

where $(\cdot)^H$ denotes the Hermitian transpose, and $(\cdot)^*$ represents the complex conjugate. T is the number of STFT time frame.

Under the single-source assumption, the cross-term $R_{c,0}(f)$ mainly captures $H_c(f) H_0(f)^*$ scaled by the source power spectrum $P_s(f) = E[|S(t, f)|^2]$, while $R_{0,0}(f)$ corresponds to $|H_0(f)|^2 P_s(f)$. Thus, we define the RTF for microphone c relative to the reference channel 0 as:

$$Z_c(f) = \frac{R_{c,0}(f)}{R_{0,0}(f)} \approx \frac{H_c(f) H_0(f)^* P_s(f)}{|H_0(f)|^2 P_s(f)} = \frac{H_c(f)}{H_0(f)}. \quad (4)$$

This serves as an approximation of the true RTF:

$$\text{RTF}_c(f) = \frac{H_c(f)}{H_0(f)}. \quad (5)$$

As a result, $Z_c(f)$ approximates $\text{RTF}_c(f)$.

Since $Z_c(f)$ provides a robust estimation of the relative transfer function, it is well-suited for learning spatial cues. To form the final feature representation, we extract the real and imaginary components of $Z_c(f)$ across frequency bins and stack them to construct a matrix of size $2(C-1) \times F$, where F is the number of frequency bins. Normalizing with respect to the reference channel's covariance term ensures that the RTF captures essential relative spatial cues while mitigating variations due to source power and mild noise conditions.

3. Proposed Method

3.1. Spatially Weighted Contrastive Loss

Traditional supervised contrastive learning establishes a clear boundary between positive and negative pairs. However, when azimuth is treated as a class label, the angular distance between

neighboring classes should be taken into account. Specifically, the difference between 0° and 1° is significantly smaller than that between 0° and 180° . Standard contrastive learning does not consider such relationships, potentially leading to suboptimal feature learning. To address this limitation, we propose a spatially weighted contrastive loss that incorporates a weighting function based on the angular distances between class pairs. The azimuth angles are treated as classes, thereby having 360 classes. By applying greater weights to class pairs with smaller angular separations, the contrastive FE learns not only distinct azimuth directions but also their spatial relationships. The SWeC loss is defined as:

$$\sum_{i \in I} \mathcal{L}_i = \sum_{i \in I} - \frac{1}{\sum_a W_{ia}} \sum_{a \in A(i)} \log \frac{\exp(W_{ia} z_i \cdot z_a / \tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a / \tau)}. \quad (6)$$

The weighting approach W_{ij} between data samples i and j is derived from the spatial gain function proposed in [22, 23], which is defined as follows:

$$g(\xi_c, \sigma_c) = \exp(\kappa_c (\cos(|\phi_i - \phi_j|) - 1)), \quad (7)$$

$$\kappa_c = \frac{\ln(\xi_c)}{\cos(\sigma_c) - 1}. \quad (8)$$

Here, $g(\xi_c, \sigma_c)$ represents the spatial weighting function, which equals 1 when the azimuth difference ($\phi_i - \phi_j$) is 0. The parameter ξ_c and σ_c denote the attenuation factor and cutoff boundary, respectively. The attenuation factor ξ_c is set to 0.7071, consistent with the value used in [22].

Since the function is Gaussian-like, its smoothness can be controlled by adjusting σ_c . To further improve feature learning, we introduce multiple projection layers with varying smoothness levels. This enables the FE to differentiate clear azimuth directions while also capturing spatial relationships among them, as illustrated in Figure 1. One projection layer is trained with the standard supervised contrastive loss (Eq. (1)), while additional layers incorporate weighted contrastive losses (Eq. (6)) with different smoothness levels. This design is inspired by the multi-target learning strategy in [23].

3.2. System architecture

As shown in Figure 1, our proposed contrastive learning-based speech localization system consists of two phases. In Phase 1, the FE is trained using the proposed spatially weighted contrastive loss. In Phase 2, the learned azimuth embedding from the FE is normalized and then used as input to a direction-of-arrival (DOA) classifier, which is then further trained. During this phase, the classifier and FE are jointly optimized, allowing gradients to update the FE. This fine-tuning step enhances the FE's ability to accurately localize speech even in challenging environments with coherent noise. The architecture of both the FE and classifier consists of four 1D point-wise convolutional layers. Each layer is followed by batch normalization (BN) and an exponential linear unit (ELU) activation function. The projection head network, responsible for generating the outputs used in contrastive learning, consists of a two-layer 1D point-wise convolutional network, also with BN and ELU activations. The outputs from the projection head are also normalized and then utilized to compute the contrastive loss. By leveraging spatially weighted contrastive learning and a structured two-phase training strategy, our proposed method effectively enhances speech localization performance in complex acoustic environments.

3.3. Contrastive Data Generation

To ensure robust contrastive learning in reverberant and noisy environments, we take advantage of a data generation strategy that introduces diverse and challenging training samples. Figure 1(a) illustrates the data generation process, where each mini-batch consists of B samples with room impulse responses (RIRs) drawn from multiple environments. This setup ensures that azimuth representations are learned in a manner that generalizes across different acoustic conditions. In real-world scenarios, the same speech source may be captured in multiple reverberant conditions, and non-speech noises may introduce directional interference. To model these conditions, we employ a two-stage data generation process: Step 1: constructing azimuth-variant contrastive samples with diverse RIRs and Step 2: incorporating coherent noise to enhance robustness.

Step 1: Diverse RIR Generation. The contrastive data generation process ensures diversity in room environments and reverberation conditions. We employ two strategies for RIR sampling: the first involves selecting random room environments for each sample, while the second constrains selection to a predefined set of M room environments. Each predefined room is associated with L azimuth angles, generating $L \times M$ RIRs per mini-batch. Importantly, samples with the same azimuth always originate from different rooms, while some pairs with different azimuths share the same room, naturally introducing hard positive and hard negative pairs. The predefined room configurations are characterized by three primary factors: room size, reverberation time (RT60), and speech sample. Other parameters such as source elevation, source-array distance, microphone array position, and signal-to-noise ratio (SNR) are randomly assigned within predefined ranges to ensure sufficient variation in acoustic conditions.

Step 2: Coherent Noise Augmentation. To simulate realistic directional interference, we introduce coherent non-speech noise sources during fine-tuning. These noise sources exhibit spatial characteristics, which affect the model's ability to extract azimuth representations. The pretrained FE is further trained in the presence of coherent noise to enhance robustness. Aside from adding noise, the dataset construction process remains consistent with the initial FE training phase.

4. Experimental setup

4.1. Datasets

To train and evaluate our proposed method, we constructed a multi-channel dataset using simulated RIRs with `gpuRIR` toolkit [24] based on the tetrahedral microphone geometry [25] and 16 kHz speech recordings from the LibriSpeech corpus [26]. The training set was derived from the *train-clean-100* subset, the validation set from *dev-clean*, and the test set from *test-clean*. Figure 1(a) lists the parameter ranges used in the simulation, which were adopted from [7]. Each training sample was generated by selecting a random speech utterance and convolving it with an RIR, which was synthesized based on randomly chosen parameters described in Figure 1(a). To ensure realistic conditions, both coherent and white noise were added to the generated samples. For coherent noise, Microsoft Scalable Noisy Speech Dataset (MS-SNSD) [27] was adopted. Training data were generated on-the-fly to introduce variations across training iterations, while validation and test datasets were pre-generated to ensure consistency in evaluation.

Table 1: Performance comparison of DOA classifiers trained with and without contrastive learning. System 1 uses only the classifier, while system 2 includes a non-trained feature encoder. Systems 3–6 incorporate contrastive learning, with system 6 combining multiple spatial weightings.

No.	CL	Method	Acc. (10°)(%)	Acc. (5°)(%)	Acc. (1°)(%)	MAE (°)	Prec. (%)	Recall (%)
1.	×	classifier only	92.56	85.42	50.38	4.80	93.03	92.68
2.	×	non-trained FE	92.40	82.98	46.68	5.20	92.80	92.27
3.	○	basic CL	93.05	84.69	50.92	4.79	93.67	93.10
4.	○	SWeC ($\sigma_c = 3$)	92.14	83.51	46.30	5.34	92.61	92.17
5.	○	SWeC ($\sigma_c = 9$)	93.28	85.15	51.03	4.95	94.16	93.21
6.	○	3-SWeC (basic, $\sigma_c = 3, 9$)	93.55	86.26	50.65	5.10	93.55	93.21

Table 2: Performance comparison between contrastive and random data generation method. Experiments are conducted under the system of No. 6. in Table 1.

Method	Acc.(%)	MAE(°)	Prec.(%)	Recall(%)
Contrastive	93.55	5.10	93.55	93.21
Random	92.44	5.19	93.09	92.41

4.2. Training Specifications

Each training sample had a duration of 4 seconds and was sampled at 16 kHz. To extract speech-active frames, we applied the WebRTC VAD [28] as a voice activity detector. These active time frames were then used to compute a RTF. The STFT was performed with a window length of 16 ms and a hop length of 8 ms. The classifier network was trained using cross-entropy (CE) loss as the objective function. We used the Adam optimizer [29] with gradient clipping in the range of ± 5 . The initial learning rate was set to 1×10^{-4} for the FE network and 1×10^{-3} for the classifier.

4.3. Evaluation Metrics

The SSL performance was evaluated using four metrics: mean absolute error (MAE), accuracy (Acc.), precision (Prec.), and recall. MAE was computed as the angular distance between the ground truth and the estimated azimuth angles, measured in degrees (°). For Acc., a sample was considered correctly classified if the angular distance was below a predefined threshold. Precision and recall were computed based on Acc. with the 10° threshold. The overall performance was evaluated by averaging the results across all test samples.

5. Results and Analysis

Table 1 presents the performance comparison of various training configurations for the FE and classifier in the speech localization task. Overall, incorporating contrastive learning led to notable improvements over models trained without it. System 1 serves as a baseline, where Phase 2 of the pipeline is used without a FE. In this setup, the RTF features are directly input into the DOA classifier without additional transformations. In contrast, system 2 introduced a FE that processes the RTF features into an embedding vector before classification. However, since the FE in system 2 is not pre-trained with contrastive learning, its ability to extract meaningful spatial representations remains limited. Comparing these two systems, we observed that simply introducing a FE without pretraining does not lead to performance gains, as both systems showed similar results.

Next, because [15] focuses on narrow-band signals, we did not treat it as a direct baseline. Instead, we replaced it with our

basic contrastive learning approach (system 3). This improved accuracy across all azimuth thresholds and achieved the lowest MAE, suggesting that pretraining the FE with contrastive learning refines spatial representations for more precise azimuth estimation. When spatial weighting was incorporated into contrastive learning (systems 4 and 5), we observed mixed results. Specifically, system 4 ($\sigma_c = 3$) showed a slight decrease in accuracy compared to basic contrastive learning, likely due to excessive smoothing. However, system 5 ($\sigma_c = 9$) improved accuracy across all thresholds, especially at the strictest 1° criterion. This suggests that allowing greater azimuth ambiguity during training helps the model capture fine-grained spatial relationships more effectively. Finally, system 6, which combines basic contrastive learning with multiple spatial weightings ($\sigma_c = 3, 9$), achieved the highest accuracy at 10° and 5° thresholds while maintaining competitive performance at 1°. This indicates that leveraging different levels of spatial sharpness enhances the model’s ability to generalize across varying azimuth resolutions.

Table 2 further highlights the impact of contrastive data augmentation. The model trained with contrastively generated samples consistently outperformed the one trained with randomly generated samples across all metrics. In particular, the accuracy and recall improvements suggest that introducing hard positive and hard negative samples strengthens the model’s ability to distinguish between similar azimuth classes, thereby improving both robustness and localization precision.

6. Conclusion

In this study, we proposed a spatially weighted contrastive learning framework for sound source localization, designed to enhance localization accuracy in noisy and reverberant environments. The core of our approach was the SWeC loss, which incorporates spatial continuity between neighboring azimuth angles. By dynamically adjusting the weight of negative pairs based on the angular distance, the model effectively penalizes false positives and improves localization robustness. Additionally, we introduced a contrastive data generation strategy that ensures diverse and challenging training samples, further enhancing the model’s ability to generalize.

7. Acknowledgment

This work was supported by the Technology Innovation Program (1415178807, Development of Industrial Intelligent Technology for Manufacturing, Process, and Logistics) funded by the Ministry of Trade, Industry & Energy(MOTIE, Korea) and by Artificial Intelligence Graduate School Program(Hanyang University))

8. References

- [1] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, 1976.
- [2] J. H. DiBiase, *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*. Ph.D. thesis, Brown University, 2000.
- [3] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone arrays: signal processing techniques and applications*. Springer, 2001, pp. 157–180.
- [4] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas, Prop.*, vol. 34, no. 3, pp. 276–280, 1986.
- [5] J. P. Dmochowski, J. Benesty, and S. Affes, "Broadband MUSIC: Opportunities and challenges for multiple source localization," in *Proc. IEEE Workshop Appl., Signal Process. Audio Acoust.*, 2007, pp. 18–21.
- [6] S. Chakrabarty and E. A. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 8–21, 2019.
- [7] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "Robust sound source tracking using SRP-PHAT and 3D convolutional neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 300–311, 2021.
- [8] D. A. Krause, A. Politis, and A. Mesaros, "Sound event detection and localization with distance estimation," *arXiv preprint arXiv:2403.11827*, 2024.
- [9] A. S. Roman, I. R. Roman, and J. P. Bello, "Robust DoA estimation from deep acoustic imaging," in *Proc. IEEE Int Conf. Acoust., Speech Signal Process.*, 2024, pp. 1321–1325.
- [10] S. Advanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *Proc. 26th Eur. Signal Processing Conf.*, 2018, pp. 1462–1466.
- [11] J.-H. Cho and J.-H. Chang, "SR-SRP: Super-resolution based SRP-PHAT for sound source localization and tracking," in *Proc. Interspeech*, 2023, pp. 3769–3773.
- [12] W. Sun *et al.*, "Learning audio-visual source localization via false negative aware contrastive learning," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2023, pp. 6420–6429.
- [13] Y.-B. Lin, H.-Y. Tseng, H.-Y. Lee, Y.-Y. Lin, and M.-H. Yang, "Unsupervised sound localization via iterative contrastive learning," *Computer Vision and Image Understanding*, vol. 227, p. 103602, 2023.
- [14] Z. Song, J. Zhang, Y. Wang, J. Fan, and Z. Zhang, "Enhancing sound source localization via false negative elimination," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10 499–10 514, 2024.
- [15] Y. Li, Z. Zhou, C. Chen, P. Wu, and Z. Zhou, "An efficient convolutional neural network with supervised contrastive learning for multi-target DOA estimation in low SNR," *Axioms*, vol. 12, no. 9, 2023.
- [16] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.
- [17] P. Khosla *et al.*, "Supervised contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 18 661–18 673.
- [18] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Trans. Speech, Audio Process.*, vol. 12, no. 5, pp. 451–459, 2004.
- [19] H. Hammer, S. E. Chazan, J. Goldberger, and S. Gannot, "Dynamically localizing multiple speakers based on the time-frequency domain," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, p. 16, 2021.
- [20] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.
- [21] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *Proc. IEEE Int Conf. Acoust., Speech Signal Process.*, 2015, pp. 544–548.
- [22] J. Choi and J.-H. Chang, "Supervised learning approach for explicit spatial filtering of speech," *IEEE Signal Proc. Lett.*, vol. 29, pp. 1412–1416, 2022.
- [23] M.-S. Baek, J.-Y. Yang, and J.-H. Chang, "Deeply supervised curriculum learning for deep neural network-based sound source localization," in *Proc. INTERSPEECH*, 2023, pp. 3744–3748.
- [24] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "gpuRIR: A python library for room impulse response simulation with GPU acceleration," *Multimedia Tools Appl.*, vol. 80, no. 4, pp. 5653–5671, 2021.
- [25] K. Shimada *et al.*, "STARSS23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 72 931–72 957.
- [26] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int Conf. Acoust., Speech Signal Process.*, 2015, pp. 5206–5210.
- [27] C. K. Reddy *et al.*, "A scalable noisy speech dataset and online subjective test framework," *Proc. Interspeech 2019*, pp. 1816–1820, 2019.
- [28] J. Wiseman, "Wiseman/py-webrtcvad," *GitHub Repository*, 2019. [Online]. Available: <https://github.com/wiseman/py-webrtcvad>
- [29] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.