DOI: 10.61782/fa.2025.0828

FORUM ACUSTICUM
EURONOISE 2025

STEERABLE NEURAL DIRECTIONAL FILTERING

Weilong Huang **Mhd Modar Halimeh** **Srikanth Raj Chetupalli**
Oliver Thiergart **Emanuël Habets**
International Audio Laboratories Erlangen†, Am Wolfsmantel 33, 91058 Erlangen, Germany
†A joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)
and Fraunhofer Institute for Integrated Circuits (IIS)

ABSTRACT

Sound capture with specific directivity patterns is essential in many far-field speech communication systems. Traditionally, fixed beamformers provide this capability through predefined directivity patterns. However, the fixed beamformers' characteristics, such as the white noise gain and directivity factor, highly depend on the number of microphones, and their directivity and robustness at low frequencies are often inadequate. Recent works have employed deep neural network-based approaches, such as neural directional filtering, to overcome the limitations of conventional fixed beamformers and demonstrate superior performance. This paper expands on the concept of neural directional filtering by incorporating steerable capabilities, termed steerable neural directional filtering. We propose a training strategy that uses the steering direction of the directivity pattern as a conditioning input for the neural network, allowing for the generation of directivity patterns aimed at any desired direction during inference. Additionally, we analyze the performance of the directivity patterns for various steering directions, revealing that the performance across different directions remains consistent.

Keywords: Steerable spatial filtering, directivity pattern, microphone array

*__Corresponding author__: weilong.huang@fau.de.

1. INTRODUCTION

Far-field sound capture is indispensable in many speech communication systems, such as televisions, video conferencing devices, and smart speakers. One prevalent technique is to capture sound from a specific speaker that forms the task of speaker extraction [1, 2]. This technique requires information describing the target speaker. Alternatively, leveraging spatial information rather than speaker information, deep neural network (DNN)-based approaches [3, 4] focus on extracting speeches from predefined spatial directions/regions. As a result, moving speakers close to a region's border often results in discontinuous directional filtering. Consequently, achieving sound capture with a controllable and smooth directivity pattern is critical.

Traditionally, fixed beamformers realize such patterns by filtering and then superimposing microphone signals. As one of the fixed beamformers, differential beamformers can achieve a frequency-invariant pattern as the desired pattern using the Jacobi-Anger expansion [5, 6] or define the null positions to control the shape of the directivity pattern through the null-constrained method [7, 8]. However, differential beamformers often suffer from low white noise gain (WNG) at low frequencies, leading to white noise amplification issues. Moreover, differential beamformers' directivity depends on the number of microphones. For example, the highest order of a directivity pattern for a uniform circular array (UCA) (which leads to the highest directivity) is upper-bounded by $\lfloor \frac{Q}{2} \rfloor$, where the Q is the number of microphones [9]. These limitations persist despite improvements made via exploiting directional microphones [10–12].

Recently, a neural directional filtering (NDF) method [13] is proposed, which overcomes the limitations of the fixed beamformer. As an example, a 3^{rd}-order differen-

tial microphone array (DMA) pattern was realized using a three-microphone UCA with an additional center microphone, which is unfeasible for a differential beamformer. Moreover, the NDF significantly outperforms conventional algorithms, such as [14, 15]. However, trained NDF models in [13] are limited to static pre-defined patterns. This paper extends the NDF to a steerable neural directional filtering (SNDF), which achieves flexible steering with a single trained model. We propose a training strategy to steer the directivity pattern using a conditioning input such that the trained model can steer the learned pattern to any desired direction during inference. Experimental results demonstrate the enhanced steerability of the proposed SNDF.

2. PROBLEM FORMULATION

We consider a compact microphone array of Q omnidirectional sensors that capture an anechoic acoustic scene with N sound sources located in the far field of the array. Let $X_{q,n}[f,t]$ represent the signal from the n-th source as captured by the q-th microphone in the short-time Fourier transform (STFT) domain, where f and t denote the frequency-bin and time-frame indices, respectively. The signal at the q-th microphone, denoted as $Y_q[f,t]$, is expressed as

$$Y_q[f,t] = \sum_{n=1}^{N} X_{q,n}[f,t] + V_q[f,t], \qquad (1)$$

where $V_q[f,t]$ represents the sensor noise, which is assumed to be spatially uncorrelated across the microphones, and $q \in 1, 2, \ldots, Q$. Furthermore, we have $X_{q,n}[f,t] = H_{\mathbf{p}_q,\mathbf{p}_n}[f]\, X_n[f,t]$, where $H_{\mathbf{p}_q,\mathbf{p}_n}[f]$ models the acoustic transfer function (ATF) between the n-th source $X_n[f,t]$ at position $\mathbf{p}_n$ and the q-th microphone located at position $\mathbf{p}_q$.

The objective of the *steerable neural directional filtering* task is to capture the acoustic scene with N sources at a position $\mathbf{p}_{\text{VDM}}$ with a steerable directivity pattern $\Psi_{\theta_s}[\theta]$, where θ denotes the direction-of-arrival (DOA) of a source in the far field and θ_s is the steering direction of this pattern. In this paper, we assume that all sound sources and microphones are positioned in the x-y plane, thereby simplifying our formulation to a two-dimensional pattern-learning scenario. For simplicity, we also assume that $\mathbf{p}_{\text{VDM}}$ is the origin of the coordinate system, the incident angle for the n-th source $\theta_n = \arctan2(y_{\mathbf{p}_n}, x_{\mathbf{p}_n})$, where $y_{\mathbf{p}_n}$ and $x_{\mathbf{p}_n}$ represent the coordinates of the position $\mathbf{p}_n$ on the y-axis and x-axis, respectively. One possible realization of this task is to mimic a virtual directional microphone (VDM) located at reference position $\mathbf{p}_{\text{VDM}}$ with the required directivity pattern steered towards the specified direction θ_s. Therefore, the output $Z_{\theta_s}[f,t]$ of the VDM represents the target signal for our task. In an anechoic room, there is only one direct-path impulse response (DPIR) $H_{\mathbf{p}_{\text{VDM}},\mathbf{p}_n}[f]$ between the n-th source and an omnidirectional microphone at the position of the VDM; then the VDM signal is expressed as

$$Z_{\theta_s}[f,t] = \sum_{n=1}^{N} \Psi_{\theta_s}[\theta_n]\, H_{\mathbf{p}_{\text{VDM}},\mathbf{p}_n}[f]\, X_n[f,t]. \qquad (2)$$

In this work, we propose a data-driven approach wherein a neural network utilizes microphone signals along with the desired steering direction to estimate the target signal.

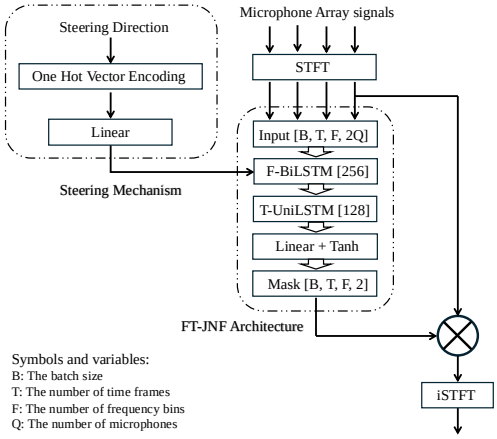


Figure 1. FT-JNF based neural directional filtering with the steering mechanism.

3. PROPOSED METHOD

3.1 Neural Network Architecture

In this work, we employ the spatially selective filter based on JNF neural network architecture (JNF-SSF) [4], shown in Fig. 1. In this architecture, the real and imaginary parts of the Q microphone signals in the STFT domain are stacked along the channel dimension, resulting in an input with dimensions of $[B, T, F, 2Q]$, where T represents the number of time frames, F denotes the number of frequency bins, and B indicates the batch size during training. Firstly, the input is reshaped into an input with dimensions of $[B \times T, F, 2Q]$ and then fed into a bidirectional long short-term memory (LSTM) layer, referred

to as F-BiLSTM, which models the spectro-spatial relationships in the data. In addition to the microphone array signals, we input an angle θ_s through the steering mechanism. This angle represents the steering direction of the directivity pattern. Within the steering mechanism, the input angle is encoded into a one-hot vector, the dimension of which is determined by the defined angular resolution. Subsequently, a linear layer processes the one-hot encoded vector, ensuring that its outputs are compatible with the input dimensions of the F-BiLSTM layer. We employ the steering direction-based information from the linear layer to initialize the forward and backward initial states of the F-BiLSTM layer for each time frame, as in [4]. The output from the F-BiLSTM layer is then reshaped into $[B \times F, T, 512]$ which is processed by a second unidirectional LSTM layer, denoted as T-UniLSTM, which can model the temporal relationships in the data. This study focuses on a frame-level causal scenario; therefore, the T-UniLSTM is configured to be unidirectional. The first F-BiLSTM layer contains 256 hidden units, while the second T-UniLSTM layer contains 128. Finally, the output of the second LSTM is reshaped to $[B, F, T, 256]$ and then passed through a linear layer with a hyperbolic tangent activation function, which computes a complex-valued single-channel mask, denoted $\mathcal{M}_{\theta_s}[f,t]$. The desired VDM signal is then estimated by applying this mask to the reference microphone signal (here chosen to be the first microphone positioned at the center of the array) as follows

$$\widehat{Z}_{\theta_s}[f,t] = \mathcal{M}_{\theta_s}[f,t]Y_1[f,t]. \tag{3}$$

3.2 Loss Function

In this work, we utilize a batch-aggregated normalized L_1 loss function to measure mean absolute error (MAE), formulated as:

$$\mathcal{L}_{\text{MAE}} = \frac{\sum_{b=1}^{B} \left\| \mathbf{z}_{\text{VDM}}^b - \widehat{\mathbf{z}}_{\text{VDM}}^b \right\|_1}{\sum_{b=1}^{B} \left\| \mathbf{z}_{\text{VDM}}^b \right\|_1 + \epsilon}, \tag{4}$$

where ϵ is a small constant value, and the signals $\mathbf{z}_{\text{VDM}}$ and $\widehat{\mathbf{z}}_{\text{VDM}}$ are the time-domain signals of the STFT representations $Z_{\theta_s}[f,t]$ and $\widehat{Z}_{\theta_s}[f,t]$, respectively.

3.3 Training Strategy

Our earlier research, as detailed in [13], has shown that the DNN models trained with two or more concurrently active speakers can effectively generalize to scenarios involving up to six speakers. Simultaneously, training with more than three speakers does not significantly enhance the model's performance. Therefore, in this study, we train our model using mixtures of up to three speakers.

We perform a discrete uniform sampling of the azimuth angle along a circle with a radius of d. This uniform sampling generates a number of P discrete admissible speaker positions, each position corresponding to a DOA having equal angular distances. Then, we locate our array in the circle's center and make the array coplanar and concentric with the circle, such that the radius d is equivalent to the source-array distance.

We define a specific speaker-array setup as one acoustic scene. In each acoustic scene, we randomly select N positions from the P discrete admissible speaker positions for N speech sources, where $N \in \{1,2,3\}$. Then we simulate $H_{\mathbf{p}_q,\mathbf{p}_n}[f]$ for all N sources and Q microphones using the room impulse response (RIR) generator [16] with a reflection order of zero. Following this, we compute Q microphone signals for this acoustic scene using (1). For each acoustic scene, we simulate M target VDM signals for steering directions uniformly spanning 0° to 360° degrees, where $M = \frac{360^\circ}{\vartheta}$ and ϑ denotes the angular resolution of the steerable network. The m-th VDM target signal $Z_{\theta_s^m}[f,t]$ corresponding to the steering direction θ_s^m is obtained using (2). During training, we consider microphone signals from each acoustic scene paired with one VDM target signal $Z_{\theta_s^m}[f,t]$ as a training sample. This process is then repeated for M target signals. Thus, the repeated utilization of the same microphone signals for training helps to emphasize the steerability function.

In addition, we introduce an enhanced mini-batch sampling approach. More specifically, for a mini-batch of B samples, at least one sample must contain a speaker from the target direction or its vicinity. This prevents the denominator calculation for L1-based loss functions from becoming excessively small. Therefore, the loss returned in each iteration is more stable, thereby enhancing the robustness of our model training process.

4. EXPERIMENTAL SETUP

4.1 Target directivity pattern

The frequency-independent directivity pattern of an Rth-order DMA for the target direction θ_s is defined as [17]

$$\Psi_{\theta_s}[\theta] = \sum_{r=0}^{R} a_r \cos^r(\theta - \theta_s), \tag{5}$$

where a_r, $r \in \{0,1,\cdots,R\}$ are real coefficients and determine the shape of DMA patterns. At the target direc-

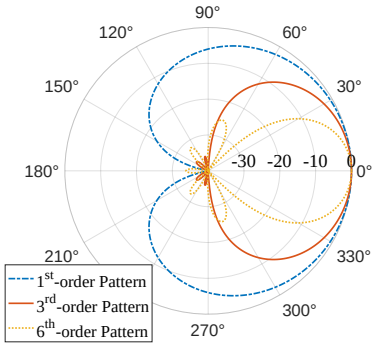


Figure 2. Three target DMA patterns for training our DNN models.

tion θ_s, the response must be equal to 1, i.e., $\Psi_{\theta_s}[\theta_s] = 1$. Therefore, we have $\sum_{r=0}^{R} a_r = 1$.

In this paper, we choose three DMA directivity patterns as target patterns for the DNN model to learn. The first pattern is the first-order cardioid pattern with coefficients $a_0 = \frac{1}{2}$ and $a_1 = \frac{1}{2}$; The second pattern is the third-order pattern with coefficients $a_0 = 0$, $a_1 = \frac{1}{6}$, $a_2 = \frac{1}{2}$, and $a_3 = \frac{1}{3}$; The third pattern is the sixth-order pattern with coefficients $a_0 = \frac{1}{49}$, $a_1 = \frac{8}{49}$, $a_2 = \frac{8}{49}$, $a_3 = -\frac{48}{49}$, $a_4 = -\frac{48}{49}$, $a_5 = \frac{64}{49}$, and $a_6 = \frac{64}{49}$. Polar plots of the respective patterns steered towards 0° are shown in Fig. 2. In the training and testing, we use these target DMA patterns to generate the target VDM signals via (2). Additionally, throughout this paper, we set the maximum suppression at -40 dB (linear scale: 0.01) for all target directivity patterns in training and testing when generating target VDM signals.

4.2 Performance Evaluation

We consider a test dataset with K test samples. For each test sample, we have N concurrent active sound sources from different directions. For the k-th sample, we use $\boldsymbol{\theta}^k$ to represent these directions

$$\boldsymbol{\theta}^k = \theta_1^k, \ldots, \theta_n^k, \ldots, \theta_N^k. \tag{6}$$

where θ_n^k represents the incident angle of the n-th speaker for the k-th test sample. The STFT representation $X_{1,n}^k[f,t]$ stands for n-th speaker of the k-th sample as received by the reference microphone $q = 1$.

4.2.1 *Estimated directivity patterns*

For the k-th test sample, we apply the estimated mask $\mathcal{M}_{\theta_s}[f,t]$ separately to the direct-path part of each individual source, such as the n-th speaker $X_{1,n}^k[f,t]$ at the reference microphone. The corresponding narrowband power ratio $\zeta_{\theta_s,n}^k[f]$ of the masked source signals to the unmasked source signals is then calculated as

$$\zeta_{\theta_s,n}^k[f] = \frac{\sum_{t=1}^{T} \left|\mathcal{M}_{\theta_s}[f,t]\; X_{1,n}^k[f,t]\right|^2}{\sum_{t=1}^{T} \left|X_{1,n}^k[f,t]\right|^2}, \tag{7}$$

and the wideband power ratio $\xi_{\theta_s,n}^k$ for the n-th source of the k-th sample is given as:

$$\xi_{\theta_s,n}^k = \frac{\sum_{f=1}^{F}\sum_{t=1}^{T} \left|\mathcal{M}_{\theta_s}[f,t]\; X_{1,n}^k[f,t]\right|^2}{\sum_{f=1}^{F}\sum_{t=1}^{T} \left|X_{1,n}^k[f,t]\right|^2}. \tag{8}$$

After we obtain the power ratios, we can obtain the learned patterns for the entire test dataset. The narrowband directivity pattern $\widehat{\mathcal{B}}_{\theta_s}[\theta,f]$ is given by:

$$\widehat{\mathcal{B}}_{\theta_s}[\theta,f] = \sqrt{\frac{1}{|\mathcal{H}_\theta|} \sum_{(k,n)\in\mathcal{H}_\theta} \zeta_{\theta_s,n}^k[f]}, \tag{9}$$

where $\mathcal{H}_\theta$ is a set of indices (k,n) that include all sources in the test dataset which are located in the direction θ and $|\mathcal{H}_\theta|$ represents the cardinality of the set $\mathcal{H}_\theta$. We define $\mathcal{H}_\theta$ as follows

$$\mathcal{H}_\theta = \left\{(k,n) \mid \theta_n^k = \theta\right\}. \tag{10}$$

The wideband directivity pattern $\widehat{\mathcal{P}}_{\theta_s}[\theta]$ is then given by:

$$\widehat{\mathcal{P}}_{\theta_s}[\theta] = \sqrt{\frac{1}{|\mathcal{H}_\theta|} \sum_{(k,n)\in\mathcal{H}_\theta} \xi_{\theta_s,n}^k}. \tag{11}$$

4.2.2 *Signal-to-distortion ratio*

We use the averaged signal-to-distortion ratio (SDR) [18, 19] to measure the distance between the estimated and target signal

$$\text{SDR} = \frac{10}{K}\sum_{k=1}^{K} \log_{10}\left(\frac{\left\|\mathbf{z}_{\text{VDM}}^k\right\|_2^2}{\left\|\mathbf{z}_{\text{VDM}}^k - \widehat{\mathbf{z}}_{\text{VDM}}^k\right\|_2^2 + \epsilon)}\right), \tag{12}$$

where $\mathbf{z}_{\text{VDM}}^k$ and $\widehat{\mathbf{z}}_{\text{VDM}}^k$ are the time-domain signals corresponding to the STFT representations $Z_{\theta_s}^k[f,t]$ and $\widehat{Z}_{\theta_s}^k[f,t]$ at the k-th sample in the test set, respectively.

4.3 Datasets

We followed the dataset preparation scheme in [13] for the microphone array signals. All speech sources were taken from the LibriSpeech database [20]. We used the subsets 'train-clean-360', 'dev-clean', and 'test-clean' for training, validation, and testing. We truncated each speech

source signal into a 4-second sample prior to convolution by the acoustic impulse response (AIR). If any sources from LibriSpeech were shorter than 4 seconds, we extended them by zero-padding.

The number of admissible speaker positions for the training set and validation set were restricted to $P_{\text{train}} = 72$ with $\theta \in \{0°, 5°, \cdots, 355°\}$ and $P_{\text{validate}} = 72$ with $\theta \in \{2.5°, 7.5°, \cdots, 357.5°\}$. The training and validation sets consisted of 11520 and 2880 acoustic scenes, respectively. Each acoustic scene corresponded to $M = 72$ VDM target signal with $\theta_s \in \{0°, 5°, \cdots, 355°\}$. Therefore, the number of samples for the training set was 11520×72, and the number for the validation set was 2880×72.

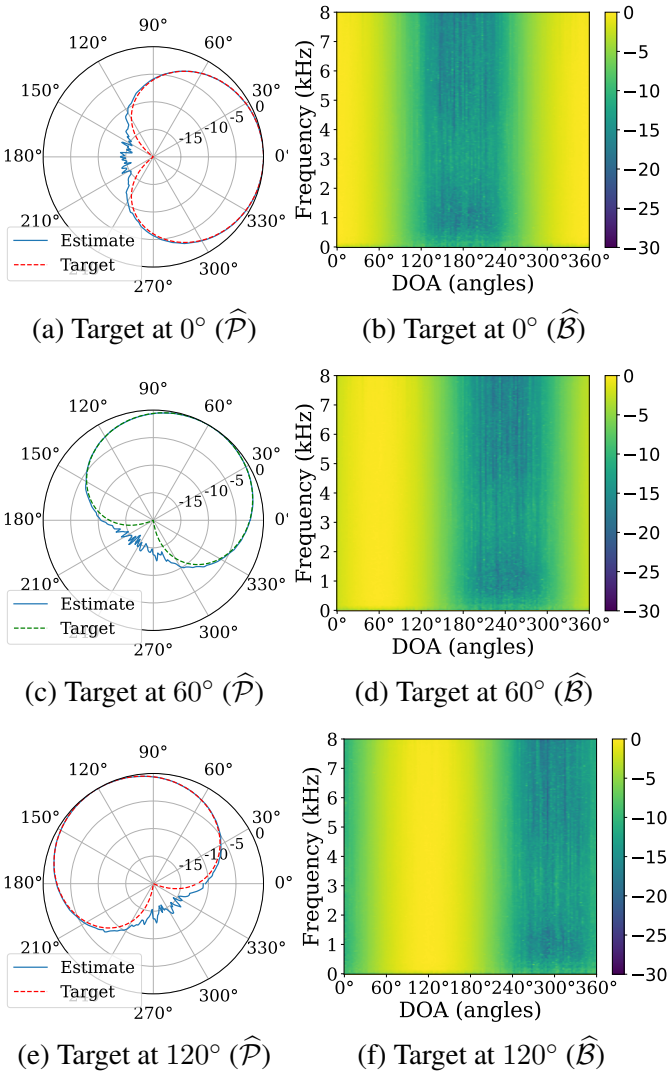
The number of admissible speaker positions for the test set was restricted to $P_{\text{test}} = 144$ with $\theta \in \{1.25°, 3.75°, \cdots, 358.75°\}$. During testing, each acoustic scene contained two concurrent speakers. The test set consisted of 3240 acoustic scenes. For each acoustic scene, we generated five target VDM signals with $\theta_s \in \{0°, 30°, 60°, 90°, 120°\}$. Therefore, the number of samples in the test set was 3240×5.

Similar to [13], we normalized all signals after convolution with the RIR to achieve a loudness within $[-33, -25]$ dBFS. Additionally, we added white Gaussian noise for the array's microphones as microphone self-noise at a signal-to-noise ratio of 30 dB with respect to the mixture of all speakers.

4.4 Configuration details

We employed a four-microphone configuration ($Q = 4$), consisting of a microphone at the center of the array ($q = 1$) and three microphones ($q = 2, 3, 4$) arranged in a UCA. The VDM was placed at the center microphone position, i.e., $\mathbf{p}_{\text{VDM}} = \mathbf{p}_1$. The diameter of the UCA was 3 cm.

For each directivity pattern, a DNN model was trained to a maximum of 100 epochs. We configure all models with a batch size of 10 and a learning rate of 0.001. The final model was selected based on the lowest validation loss observed throughout the training epochs. All trained DNNs have a total of 873K parameters. The STFT was computed on signal frames of 32 ms duration, using a square-root Hann window with a 50 % overlap at a sampling frequency of 16 kHz. We set $\epsilon = 1.2 \cdot 10^{-7}$.



(a) Target at 0° ($\widehat{\mathcal{P}}$) (b) Target at 0° ($\widehat{\mathcal{B}}$)

(c) Target at 60° ($\widehat{\mathcal{P}}$) (d) Target at 60° ($\widehat{\mathcal{B}}$)

(e) Target at 120° ($\widehat{\mathcal{P}}$) (f) Target at 120° ($\widehat{\mathcal{B}}$)

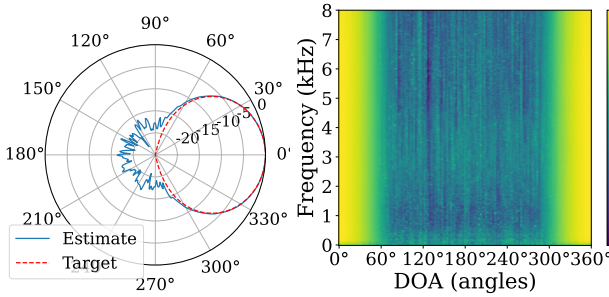
Figure 3. DNN estimated wideband pattern (marked as $\widehat{\mathcal{P}}$) and narrowband pattern (marked as $\widehat{\mathcal{B}}$) for the target cardioid pattern

5. EXPERIMENTAL RESULTS

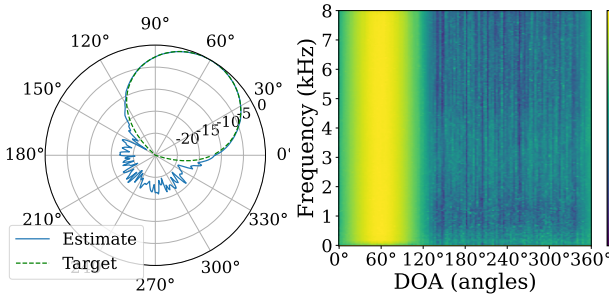
We study the performance of the proposed steerable neural directional filtering in terms of estimated directivity patterns, SDR, and audio spectrograms.

5.1 Steerable patterns

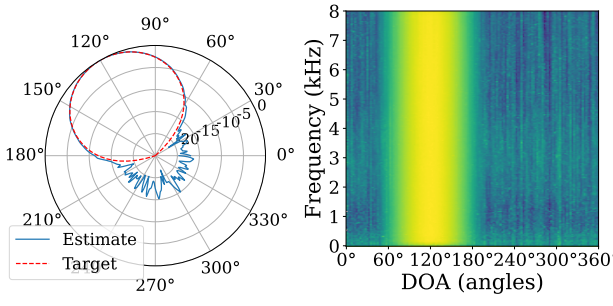
Figures 3, 4 and 5 show the estimated wideband directivity patterns $\widehat{\mathcal{P}}_{\theta_s}[\theta]$ and narrowband directivity patterns



(a) Target at 0° ($\widehat{\mathcal{P}}$) (b) Target at 0° ($\widehat{\mathcal{B}}$)

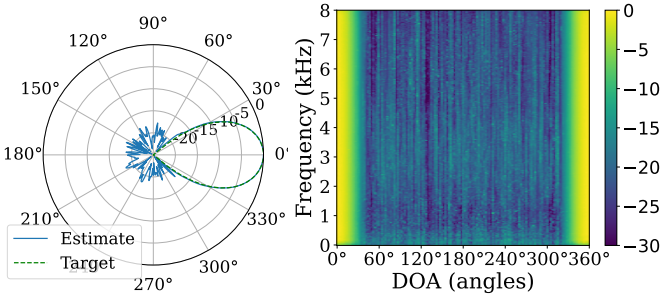


(c) Target at 60° ($\widehat{\mathcal{P}}$) (d) Target at 60° ($\widehat{\mathcal{B}}$)

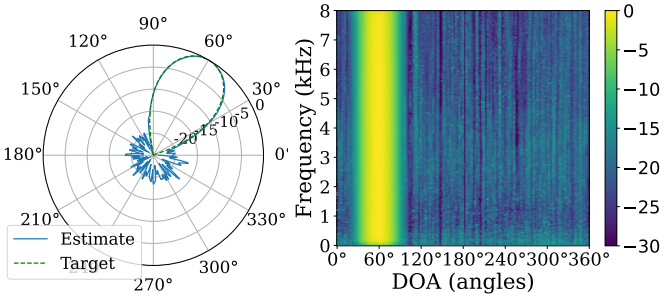


(e) Target at 120° ($\widehat{\mathcal{P}}$) (f) Target at 120° ($\widehat{\mathcal{B}}$)

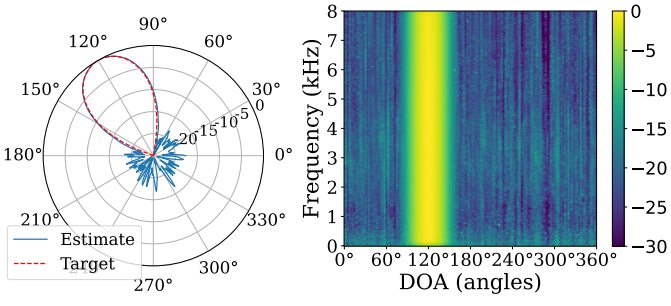
Figure 4. DNN estimated Wideband pattern (marked as $\widehat{\mathcal{P}}$) and Narrowband pattern (marked as $\widehat{\mathcal{B}}$) for the target 3$^{\text{rd}}$-order pattern



(a) Target at 0° ($\widehat{\mathcal{P}}$) (b) Target at 0° ($\widehat{\mathcal{B}}$)



(c) Target at 60° ($\widehat{\mathcal{P}}$) (d) Target at 60° ($\widehat{\mathcal{B}}$)



(e) Target at 120° ($\widehat{\mathcal{P}}$) (f) Target at 120° ($\widehat{\mathcal{B}}$)

Figure 5. DNN estimated Wideband pattern (marked as $\widehat{\mathcal{P}}$) and Narrowband pattern (marked as $\widehat{\mathcal{B}}$) for the target 6$^{\text{th}}$-order pattern

$\widehat{\mathcal{B}}_{\theta_s}[\theta, f]$ for the first-order cardioid, third-order, and sixth-order target patterns, respectively. The left-hand side (a, c, e) of each figure shows the estimated wideband directivity patterns and compares them with the corresponding target pattern in different steering directions $\theta_s = \{0°, 60°, 120°\}$. Meanwhile, each figure's right-hand side (b, e, f) shows the estimated narrowband directivity patterns in different steering directions $\theta_s = \{0°, 60°, 120°\}$.

Firstly, we can see that SNDF can learn similar patterns for different steering directions. The shape of the estimated patterns is steering invariant. Secondly, the main-lobe of the target pattern is well approximated, while the null direction has a limited attenuation. Lastly, we can also see that the estimated patterns are frequency invariant as desired.

Table 1. SDR [dB] over various steering directions.

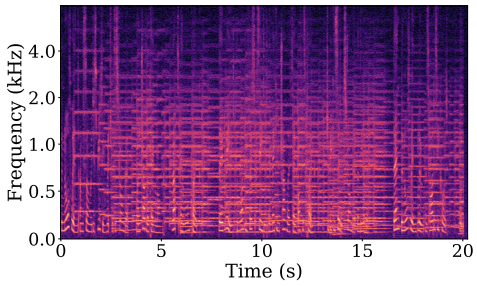
	0°	30°	60°	90°	120°
1^{st}-order pattern	25.81	25.90	25.89	25.96	25.95
3^{rd}-order pattern	20.16	20.22	20.21	20.20	20.29
6^{th}-order pattern	17.21	17.06	16.89	16.73	17.51

5.2 SDR results

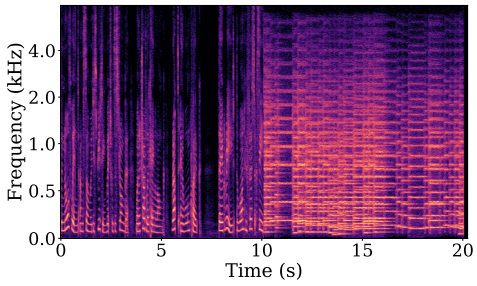
We show the averaged SDR for different steering directions $\theta_s \in \{0°, 30°, 60°, 90°, 120°\}$ in Table 1. It is clear that the SNDF achieves similar performance over different steering directions in terms of SDR. Since the input microphone array signals in the test set remain the same for different steering directions, this implicitly suggests that the target speakers differ for each direction, leading to minor differences in the SDRs over steering directions. Moreover, higher-order target patterns are increasingly difficult for DNN to learn. The target VDM signals for the sidelobes around the null positions are low-power signals, resulting in poor SDR, as observed also in [13].

5.3 Spectrogram Example

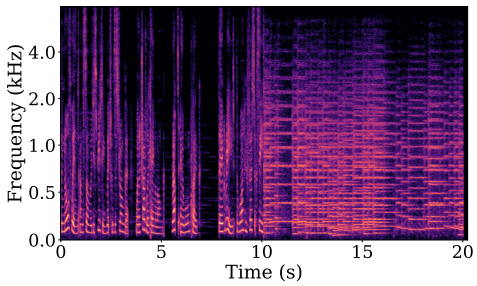
To illustrate the SNDF performance, we designed an acoustic scene of duration 20 seconds involving two sound sources, a speech source located at an angle of 60° and a music source located at an angle of 230° with respect to the center of the array. Both sources were active simultaneously, producing a fully overlapping mixture signal in the reference microphone, which is illustrated in Figure 6(a). During the first 10 seconds, we set the steering direction to 60° using the steering mechanism in our method. We adjust the steering direction for the subsequent 10 seconds to 230°. We processed this mixture using the SNDF model trained with a first-order pattern as the target pattern. The expected output (target VDM signal) and the output of our proposed method for this inference scenario are shown in Figure 6(b) and Figure 6(c), respectively. During the first 10 seconds, the music source is used as interference, which arrives from an angle close to the null direction and is consequently suppressed. A similar phenomenon is observed for the last 10 seconds where the speech source as interference is suppressed. There is no evident signal distortion comparing the spectrogram of the output signal by the proposed SNDF to the spectrogram of the VDM signal. It is worth noting that the SNDF works well for music even though it is trained using speech only.



(a) Reference microphone signal



(b) Target VDM signal



(c) Output signal by the proposed SNDF

Figure 6. Spectrograms comparison for a scenario with different steering direction of the SNDF.

6. CONCLUSIONS

In this paper, we have proposed a DNN-based steerable directional filtering method named SNDF. We propose a training strategy that considers the directivity pattern and steering direction. The steering direction is also used as an additional input to the model, which allows the model to form a directivity pattern steered in any desired direction during inference. The experimental results demonstrate the SNDF's steerability by comparing the learned and target patterns. Meanwhile, the SNDF achieves similar SDRs over different steering directions. Furthermore, SNDF can learn high-order patterns, even when the order exceeds the number of microphones.

7. REFERENCES

[1] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 5554–5558, IEEE, 2018.

[2] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. Černockỳ, "Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE Journal of Selected Topics in Signal Processing*, pp. 800–814, 2019.

[3] K. Tesch and T. Gerkmann, "Nonlinear spatial filtering in multichannel speech enhancement," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 29, pp. 1795–1805, 2021.

[4] K. Tesch and T. Gerkmann, "Multi-channel speech separation using spatially selective deep non-linear filters," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 32, pp. 542–553, 2023.

[5] G. Huang, J. Benesty, and J. Chen, "On the design of frequency-invariant beampatterns with uniform circular microphone arrays," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, pp. 1140–1153, 2017.

[6] G. Huang, J. Chen, and J. Benesty, "Insights into frequency-invariant beamforming with concentric circular microphone arrays," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, pp. 2305–2318, 2018.

[7] J. Benesty and C. Jingdong, *Study and design of differential microphone arrays*, vol. 6. Springer Science & Business Media, 2012.

[8] G. Huang, I. Cohen, J. Chen, and J. Benesty, "Continuously steerable differential beamformers with null constraints for circular microphone arrays," *The Journal of the Acoustical Society of America*, pp. 1248–1258, 2020.

[9] J. Benesty, J. Chen, and I. Cohen, *Design of circular differential microphone arrays*. Springer, 2015.

[10] W. Huang and J. Feng, "Differential beamforming for uniform circular array with directional microphones," in *Interspeech 2020*, pp. 71–75, 2020.

[11] W. Huang and J. Feng, "Minimum-norm differential beamforming for linear array with directional microphones," in *Interspeech 2021*, pp. 701–705, 2021.

[12] W. Huang and J. Feng, "Robust steerable differential beamformer for concentric circular array with directional microphones," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 319–323, 2022.

[13] J. Wechsler, S. R. Chetupalli, M. M. Halimeh, O. Thiergart, and E. A. Habets, "Neural directional filtering: Far-field directivity control with a small microphone array," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 459–463, IEEE, 2024.

[14] E. Rasumow, M. Hansen, S. van de Par, D. Püschel, V. Mellert, S. Doclo, and M. Blau, "Regularization approaches for synthesizing hrtf directivity patterns," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 24, no. 2, pp. 215–225, 2016.

[15] K. Kowalczyk, O. Thiergart, M. Taseska, G. Del Galdo, V. Pulkki, and E. A. P. Habets, "Parametric spatial sound processing: A flexible and efficient solution to sound scene acquisition, modification, and reproduction," *IEEE Sig. Proc. Magazine*, pp. 31–42, 2015.

[16] E. A. Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, p. 1, 2006.

[17] G. W. Elko, "Superdirectional microphone arrays," *Acoustic signal processing for telecommunication*, pp. 181–237, 2000.

[18] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 696–700, IEEE, 2018.

[19] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 46–50, IEEE, 2020.

[20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 5206–5210, 2015.