Varzandeh *et al.*
EURASIP Journal on Audio, Speech, and Music Processing (2025) 2025:5
https://doi.org/10.1186/s13636-025-00392-8

EURASIP Journal on Audio, Speech, and Music Processing

METHODOLOGY

Open Access

Improving multi-talker binaural DOA estimation by combining periodicity and spatial features in convolutional neural networks

Reza Varzandeh[1*], Simon Doclo[1] and Volker Hohmann[1]

Abstract

Deep neural network-based direction of arrival (DOA) estimation systems often rely on spatial features as input to learn a mapping for estimating the DOA of multiple talkers. Aiming to improve the accuracy of multi-talker DOA estimation for binaural hearing aids with a known number of active talkers, we investigate the usage of periodicity features as a footprint of speech signals in combination with spatial features as input to a convolutional neural network (CNN). In particular, we propose a multi-talker DOA estimation system employing a two-stage CNN architecture that utilizes cross-power spectrum (CPS) phase as spatial features and an auditory-inspired periodicity feature called periodicity degree (PD) as spectral features. The two-stage CNN incorporates a PD feature reduction stage prior to the joint processing of PD and CPS phase features. We investigate different design choices for the CNN architecture, including varying temporal reduction strategies and spectro-temporal filtering approaches. The performance of the proposed system is evaluated in static source scenarios with 2–3 talkers in two reverberant environments under varying signal-to-noise ratios using recorded background noises. To evaluate the benefit of combining PD features with CPS phase features, we consider baseline systems that utilize either only CPS phase features or combine CPS phase and magnitude spectrogram features. Results show that combining PD and CPS phase features in the proposed system consistently improves DOA estimation accuracy across all conditions, outperforming the two baseline systems. Additionally, the PD feature reduction stage in the proposed system improves DOA estimation accuracy while significantly reducing computational complexity compared to a baseline system without this stage, demonstrating its effectiveness for multi-talker DOA estimation.

Keywords Convolutional neural networks, Spatial feature, Periodicity feature, Binaural DOA estimation, Multiple talkers, Feature reduction, Hearing devices.

1 Introduction

Multi-talker direction of arrival (DOA) estimation is integral to acoustic signal processing and plays a pivotal role in many applications, from enhancing auditory experiences in assisted listening devices to improving voice command detection in smart devices [1, 2]. In hearing aids, accurate DOA information facilitates improved speech intelligibility through beamforming, enables the suppression of competing noise sources, and can increase environmental awareness. This ultimately helps users of hearing aids in navigating conversations in complex social environments. While the human auditory system is uniquely able to localize speech sources in noisy and reverberant environments, this remains a challenging task for machine listening systems such as hearing aids [3, 4]. This study addresses multi-talker DOA estimation in the context of binaural hearing aids.

*Correspondence:
Reza Varzandeh
reza.varzandeh@uol.de
[1] Department of Medical Physics and Acoustics and the Cluster of Excellence Hearing4all, University of Oldenburg, Oldenburg 26111, Germany

Binaural DOA estimation, the process of determining the direction of sound sources using signals received by two microphones (e.g., in a binaural hearing aid setup), primarily leverages binaural cues inspired by the human auditory system, namely interaural time difference (ITD), interaural phase difference (IPD), and interaural level difference (ILD) [3, 5]. While ITD and IPD pertain to the differences in time and phase of a sound arriving at two microphones, respectively, ILD concerns the difference in sound intensity levels captured by the microphone pair. ITD information can be defined either through an auditory-inspired approach in auditory gammatone filterbank channels [5, 6] or using the broadband generalized cross-correlation (GCC) function [5, 7]. Integrating ITD with ILD cues has been shown to enhance the accuracy of binaural DOA estimation compared to using ITD alone [6, 8]. Studies show that combining IPD and ILD information improves DOA estimation accuracy in multi-talker scenarios [9]. Most of these approaches usually match the estimated binaural features with pre-computed feature templates from head-related transfer function (HRTF) databases to obtain the DOA. Another class of conventional binaural DOA estimation approaches utilizes relative transfer function (RTF) vectors [10–12]. These methods employ a database of prototype RTF vectors, pre-computed for each direction using a measured HRTF database.

A major challenge for these approaches in real-world acoustic conditions is that background noise, interference, and reverberation introduce uncertainties into binaural cues. These uncertainties distort the extracted binaural features from microphone signals, leading to mismatches with the pre-computed templates and subsequently degrading DOA accuracy [6]. To address the limitations of these approaches, researchers have developed supervised learning techniques using deep neural networks (DNNs) [13–31]. When trained on diverse acoustic conditions, these techniques demonstrate more robust performance in adverse scenarios [14, 19]. In this paper, we address the multi-talker binaural DOA estimation using supervised DNN-based techniques.

Supervised DNN-based approaches commonly formulate the multi-talker DOA estimation task as a classification or regression problem [13, 30]. In classification-based approaches, the neural network predicts a spatial probability map for a discretized DOA range. Peaks within this map indicate the probable locations of active sound sources. A common assumption here is that the number of sources is known in advance, and their DOAs are found by peak detection [14, 15, 17, 28, 29]. In regression-based approaches, the neural network provides continuous estimates directly in the output [30, 31]. This offers potential for improved performance compared to the classification-based approaches, as the latter limits the DOA resolution depending on the discretized DOA range. However, in regression-based approaches, the number of simultaneously active sources must be known before training, which may not align with the number of active sources during testing. This is a significant limitation of regression-based approaches in multi-talker scenarios, as they require multiple regressors and distinct models for handling varying numbers of sources. Another drawback of regression-based approaches is the source permutation problem in multi-talker scenarios, where it becomes ambiguous which predicted output corresponds to which target speaker. While the choice between regression and classification depends on the specific application and requirements, several studies [30, 32, 33] suggest that classification may be a more suitable option for DOA estimation, particularly in multi-talker scenarios and challenging environments such as low SNR conditions or closely located sources. This paper focuses on the classification-based approach for binaural DOA estimation, which requires only a single network architecture, thereby simplifying system design and potentially enhancing performance in real-world hearing aid scenarios.

DNN-based methods for binaural DOA estimation typically utilize spatial features extracted from binaural signals [13]. Commonly used features include the ILD [14, 20, 21, 34], ITD (or IPD) [20, 23, 34, 35], RTF [26], cross-correlation function (CCF) [14, 21, 22], cross-power spectrum (CPS) [36–38], and generalized cross-correlation with phase transform (GCC-PHAT) [17, 32, 33, 38, 39]. Most of these methods directly map the input features for DOA estimation utilizing the network output [14, 20, 21, 38, 39], while some methods adopt a two-step approach, first refining the input features into enhanced representations using the DNN and then estimate the DOA from the enhanced features [18, 25, 35]. To address the challenge of sound target speech separation in scenarios with closely located sources, authors in [40] investigated 3D spatial features incorporating azimuth, elevation, and source distance, which could potentially benefit the binaural DOA estimation task. While most existing methods focus on DOA estimation in the azimuthal plane [14, 19, 21, 25], some employ multi-task learning to simultaneously estimate both azimuth and elevation [20, 22] or azimuth alongside distance [41]. Real-world hearing aid scenarios, e.g., multi-talker conversations, typically involve sources at the listener's elevation in distances much larger that the size of the hearing aid. Therefore, we focus specifically on DOA estimation within the azimuthal (horizontal) plane.

It is assumed that the human auditory system groups signal components according to information such as

periodicity of voiced speech and continuity of harmonics, and then ITD (or IPD) information is used to segregate the grouped components [3]. Motivated by that, a learning-based method for multi-talker DOA estimation [42] proposed to incorporate a monaural pathway including pitch-based analysis to group time-frequency units dominated by the same talker. The grouping provided constraints for the integration of binaural cues, improving azimuth estimation accuracy. It has also been shown in [43] that periodicity-based salient features yield a sparse auditory time-frequency representation capable of decoding complex auditory scenes.

While binaural features are widely used for DOA estimation, the benefit of their combination with monaural auditory-inspired spectral features, such as salient periodicity features as input features for DNN-based multi-talker DOA estimation, has not been investigated. In our previous studies on single-talker binaural DOA estimation [37–39], we investigated various representations of a periodicity feature called periodicity degree (PD), including narrowband, subband-averaged, and broadband versions, in combination with spatial features. These studies consistently demonstrated that including PD features allows for reliable speech detection and more accurate DOA estimation. In [37], we proposed a classification-based system based on convolutional neural networks (CNNs) for single-talker binaural DOA estimation using narrowband PD features in combination with spatial features as input to a two-stage CNN architecture. Due to its effectiveness in highlighting periodic components of speech, we selected the narrowband PD features for this study, anticipating that its rich frequency representation will aid in enhancing DOA estimation in environments with multiple talkers.

In this paper, we propose a DOA estimation system that builds upon our earlier work [37] by incorporating a unique feature combination within a computationally efficient two-stage CNN adapted for multi-talker DOA estimation. Our main objective is to explore the potential benefits of incorporating periodicity features, alongside spatial features for DNN-based multi-talker DOA estimation, as established for the single-talker scenarios in [37, 38]. As the spatial feature, we use the phase component of CPS, which is closely related to the IPD for a pair of microphones. We hypothesize that combining the CPS phase as the spatial feature with a compact representation of PD, obtained through a feature reduction stage inspired by [37], will outperform using the CPS phase alone in multi-talker scenarios.

To evaluate the benefit of combining PD features with CPS phase features, we consider baseline systems that utilize either only CPS features or a combination of CPS phase features and the magnitude spectrogram as spectral features. Experimental results clearly show the advantages of combining PD and CPS phase features within the proposed system, demonstrating consistent improvements in DOA estimation performance across various conditions and environments, compared to both baseline systems. Moreover, to investigate the benefit of the PD feature reduction stage in the proposed system, we compare it against a baseline system that processes the same features without a PD feature reduction stage. The results show that the proposed system not only improves DOA estimation accuracy but also significantly reduces computational complexity, underscoring the effectiveness of the PD feature reduction stage for multi-talker DOA estimation.

The remainder of this paper is structured as follows. First, in Section 2, the multi-talker DOA estimation is formulated and discussed as a classification problem. Section 3 introduces the input features employed in this study. In Section 4, comprehensive descriptions of the proposed and baseline systems are presented. The details of the experimental setup for training and evaluation of all systems including datasets, data generation, training and network hyperparameters, and evaluation metrics appear in Section 5. The proposed and baseline systems are evaluated, and the results are discussed in Section 6. Section 7 summarizes the results and presents the conclusion.

2 DOA estimation as a classification problem

In this work, we consider the problem of multi-talker DOA estimation in the azimuthal plane for a known number of talkers ($\mathcal{I}$) using a binaural hearing aid setup with M microphones, where the microphones are located close to the ears on both sides. The acoustic scenario consists of multiple speech sources and background noise, which are assumed to be mutually uncorrelated. The m-th microphone signal in the time domain at time t is given by

$$y_m(t) = \sum_{i=1}^{\mathcal{I}} x_m^i(t) + v_m(t), \tag{1}$$

where x_m^i and v_m denote the desired i-th speech source at DOA θ_i in the azimuthal plane, and noise signal components in the m-th microphone signal, respectively. In the short-time Fourier transform (STFT) domain, the m-th microphone signal at time frame n and frequency bin k (with K and D the STFT length and hop size, respectively) can be written as

$$Y_m(n,k) = \sum_{i=1}^{\mathcal{I}} X_m^i(n,k) + V_m(n,k). \tag{2}$$

Conventionally, by discretizing the azimuth range into C DOAs $\{\phi_1, \cdots, \phi_C\}$, multi-talker DOA estimation is formulated as a C-class classification task, where

output classes correspond to independent DOAs, i.e., sound source locations are mutually independent [15, 28, 29]. The goal is to assign the DOAs of multiple incoming sound sources to corresponding DOA classes. In this study, we use $C = 72$ classes spanning the full 360° azimuth range, yielding a DOA map with 5° resolution.

By taking a supervised approach, during training, each training example may belong to one or more output classes that are labeled using ground truth DOA information. In other words, each training example can represent situations where multiple speakers are active simultaneously. However, this approach can complicate the training data generation, as the differences in signal levels for these scenarios can significantly impact the performance of the DOA estimation system. In this work, we generate training examples involving only a single active speaker and evaluate the system's ability to generalize to multi-talker scenarios where each speaker contributes equally to the microphone signal.

During testing, the neural network predicts a posterior probability for each DOA class in the output. The generated posterior probability map $\boldsymbol{P} = [P_1, \cdots, P_C]$ represents the likelihood of the sound source being located at each of the C possible DOAs. As a common approach, with $\mathcal{I}$ active sources, the $\mathcal{I}$ DOA classes with the highest probability values in $\boldsymbol{P}$ are selected as the estimated DOAs. In this study, we will take a slightly different approach for DOA estimation in Section 4.

3 Input features

This section outlines the spatial and spectral features used as inputs for the classification-based DOA estimation methods in this study. Section 3.1 introduces narrowband CPS features as spatial features. Section 3.2 presents narrowband PD features (as introduced in [37]), alongside the magnitude spectrogram as an alternative spectral feature.

3.1 Spatial features

The CPS, which represents the frequency-domain counterpart of the CCF, encodes both the joint magnitude and the phase difference between signals from a pair of microphones. While the real/imaginary or magnitude/phase components of the CPS can be used as spatial features for DNN-based DOA estimation [37, 38], employing the CPS phase alone as the spatial cue reduces dimensionality and computational complexity, which is critical for resource-constrained hearing aids. Previous work on single-talker binaural DOA estimation [38] found no clear benefit from using both magnitude and phase components of the CPS compared to other spatial features as inputs to a CNN. Moreover, using only the CPS phase can mitigate the sensitivity to amplitude variations while offering a rich representation in the time-frequency domain. In this study, we aim to combine periodicity and spatial features for multi-talker DOA estimation. Because periodicity features already encode spectral amplitude information, adding CPS magnitude components may introduce undesired feature redundancy. Hence, since the CPS phase offers a good balance between spatial cue information, feature redundancy, and computational complexity, in this work we consider the CPS phase component as the spatial feature used for the baseline and proposed systems in Section 4. The instantaneous CPS between the r-th and q-th microphone is defined as

$$G_d(n,k) = Y_r(n,k)Y_q^*(n,k), \tag{3}$$

where $(\cdot)^*$ denotes complex conjugate and d denotes a microphone pair combination. As CPS input, we consider the phase components of $G_d(n,k)$ for all $M(M-1)/2$ unique microphone pairs for frequencies up to the Nyquist frequency, i.e., $k = 0, 1, \cdots, K/2$, for L consecutive time frames. This means that the shape of the CPS input is equal to $L \times (K/2+1) \times M(M-1)/2$. We note here that the first, second, and third dimensions represent the height, width, and depth of the input feature, respectively, with the depth corresponding to the number of input channels.

3.2 Spectral features

Periodicity is an important cue to segregate and localize different talkers [43, 44]. Periodicity features typically require an auditory pre-processing stage followed by feature extraction [43]. In [37–39], a periodicity feature called PD was used, which captures the salience of the periodic components in the input signal. In this work, we propose to use narrowband PD features [37] for multi-talker DOA estimation, computed for a set of N fundamental period candidates.

To compute PD features, we select one of the M microphones as the reference. Please note that this microphone is selected arbitrarily, and optimal microphone selection for PD estimation is beyond the scope of this study. In the pre-processing step, the reference microphone signal in the hearing aid setup is first decomposed into signals in different gammatone frequency bands using a complex-valued gammatone filterbank (GTFB) [45]. The real part of each signal then undergoes half-wave rectification, yielding signal $y(t, f)$ in the f-th gammatone frequency band. In each frequency band, $y(t, f)$ is processed with a fifth-order low-pass filter (770 Hz cutoff) and a second-order high-pass filter (40 Hz cutoff), resulting in band-pass-filtered signal envelopes $y_{env}(t,f)$. These envelopes serve as the basis for our PD feature extraction.

In the feature extraction step, we filter the signal envelopes using a set of N parallel infinite impulse response (IIR) comb filters. These filters are designed for a set of N fundamental period candidates $p_j, j = 1, \cdots, N$. The comb-filtered signals are computed by

$$s(j,t,f) = (1-\alpha) y_{env}(t,f) + \alpha s(j, t-p_j, f), \tag{4}$$

where α denotes the filter gain. The periodicity degree is defined as the mean amplitude of the comb-filtered signal, given by

$$PD(j,t,f) = (1-\beta_j)|s(j,t,f)| + \beta_j PD(j,t-1,f), \tag{5}$$

where the averaging parameter β_j for each fundamental period candidate is defined as $\beta_j = e^{-1/p_j}$.

To enable joint spectro-temporal processing of PD and CPS features, their time-frequency resolutions must be aligned. Since PD features in (5) initially have the temporal resolution of the time-domain signal, we achieve the necessary alignment with CPS features by temporally averaging them within each STFT frame as

$$\overline{PD}(j,n,f) = \frac{1}{K} \sum_{t=(n-1)D+1}^{(n-1)D+K} PD(j,t,f). \tag{6}$$

The non-uniform frequency resolution of gammatone bands (decreasing with frequency) contrasts with the linear spacing of STFT frequency bands. To align the frequency resolution of PD features with CPS features, we employ different strategies based on STFT frequency. For low frequencies, PD features from multiple gammatone bands corresponding to a single STFT band are averaged, while for high frequencies, PD features from each gammatone band are replicated and assigned to the associated STFT frequency bands.

As input PD feature used for the proposed system (cf. Fig. 3), we consider PD features in (6) for all N fundamental period candidates, for L consecutive time frames, and for all $K/2+1$ STFT frequency bands. This leads to an input PD feature of size $L \times (K/2+1) \times N$, which will be used as the input PD feature of the proposed DOA estimation system in Section 4.2.

For a 1s clean signal of a female talker, Fig. 1 depicts exemplary two-dimensional (2D) images of PD features, corresponding to a subset of fundamental frequency candidates. For a perfectly periodic signal characterized by a specific fundamental frequency, a high PD value will be captured for candidates associated with the harmonics and sub-harmonics of this fundamental frequency. While speech is not perfectly periodic, fundamental frequency variations and harmonics create a spectro-temporal structure visible in the PD features. The primary rationale for using PD features alongside spatial features is to leverage the periodicity features as a robust footprint of speech signals in a noisy mixture [46, 47]. This approach allows the neural network to pinpoint voiced speech segments while simultaneously mapping their associated CPS features to the talker's DOA.

Magnitude spectrograms provide rich spectro-temporal information about formant frequencies and harmonic content, making them common in DOA estimation systems [13]. They are also employed as monaural features

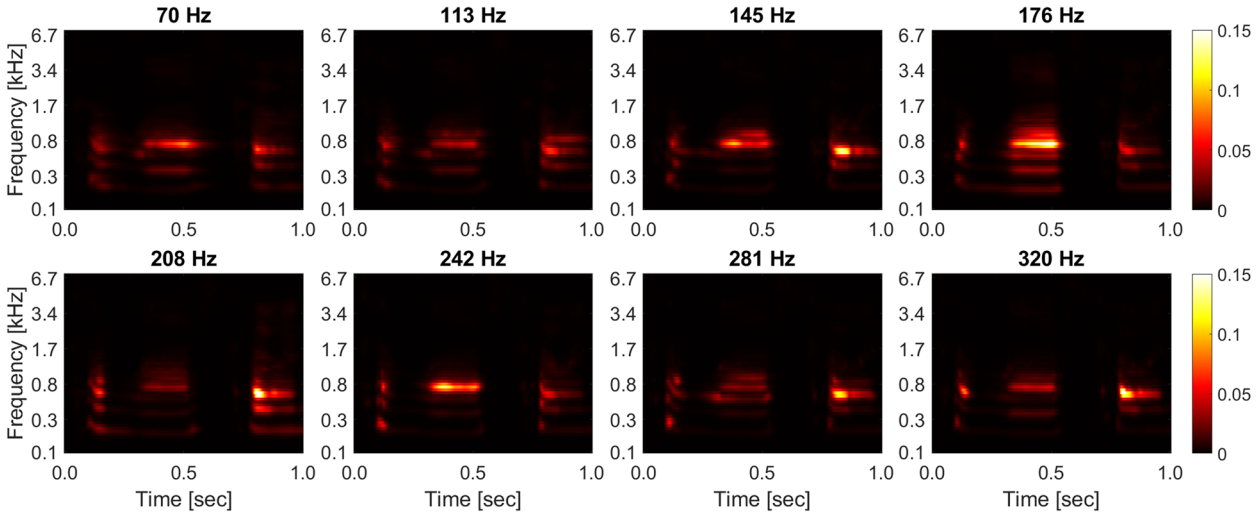


Fig. 1 PD features computed over $L = 199$ consecutive time frames and 61 gammatone bands for a clean female speech in an anechoic environment. A small set of fundamental frequency candidates (specified above each image) is shown for visualization. The sparse spectro-temporal structure of PD features contains sufficient information to decode complex auditory scenes and motivates using a feature reduction stage to learn the salient PD features prior to joint processing with the CPS phase

for target speech DOA estimation through the use of ratio masks [48]. To explore the benefit of spectral features in combination with spatial features for multi-talker DOA estimation, as an alternative to periodicity features, we use the magnitude spectrogram of the same microphone used for PD feature extraction. We take the magnitude of the microphone's STFT coefficients up to the Nyquist frequency, i.e., $k = 0, 1, \cdots, K/2$, for L consecutive time frames. This results in an input magnitude spectrogram of shape $L \times (K/2 + 1)$.

4 CNN-based DOA estimation systems

Neural networks based on CNNs are the most commonly used architectures and have proven highly effective for DOA estimation and sound source localization [13]. These architectures typically start by feeding input features into a cascade of convolutional blocks followed by fully connected blocks, ultimately leading to an output layer. The primary role of the convolutional blocks is to extract local patterns from the input data, generating intermediate features for different numbers of filters. These features are then flattened and serve as input to fully connected blocks, which process these intermediate features and learn global patterns essential for the DOA estimation task. Several spatial features have been used as inputs to CNNs for the DOA estimation task, including GCC-PHAT [17], multichannel phase spectrogram [15, 29], IPD and ILD [20], and CPS [36]. In our previous studies [37–39], we proposed to use the periodicity features in combination with different common spatial features as input features to different CNN architectures for single-talker binaural DOA estimation. In this study, we employ the combination of periodicity and spatial features to improve multi-talker DOA estimation within an efficient two-stage CNN architecture, compared to baseline systems using common input features and CNN architectures.

This section outlines the CNN-based DOA estimation systems. Section 4.1 describes baseline systems, while the proposed two-stage system is detailed in Section 4.2. The proposed system uses the CPS phase combined with PD features as input, incorporating a PD feature reduction stage. Three different baseline systems are considered in Section 4.1, including a single-stage baseline system that directly combines the PD and CPS features without PD feature reduction, in order to assess the benefit of the PD feature reduction stage in the proposed system. Furthermore, to evaluate the benefit of using PD features in combination with the CPS phase, we consider two other baseline systems that employ either only the CPS phase or a combination of the CPS phase and the magnitude spectrogram as input. Section 4.2 explores alternative design choices for the two-stage architecture compared to the proposed two-stage CNN. Finally, Section 4.3 analyzes the computational complexity of all considered systems.

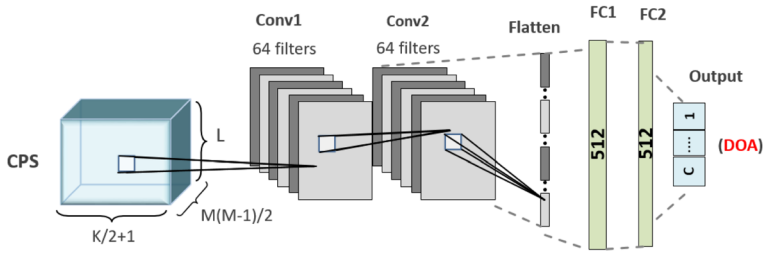
All systems share the same training and DOA estimation procedures. The key difference between our proposed system and the baselines lies in the two-stage CNN architecture and the combination of CPS phase and PD features. For training, each training example consists of a block of L consecutive time frames, i.e., we employ block-level labeling and each CNN generates its output for the whole block. A key assumption is that the DOA remains constant within this block of L frames when assigning a ground truth label.

For DOA estimation, with $\mathcal{I}$ active talkers, we first find the $\mathcal{I}$ DOA classes $\phi_j, j = 1, \cdots, \mathcal{I}$ with the highest probability values in the posterior probability map $\boldsymbol{P}$. To refine these discrete DOA classes into continuous estimates, for each ϕ_j, we estimate a talker's DOA by employing parabolic interpolation [49] on three DOA classes centered around ϕ_j, i.e., ϕ_{j-1}, ϕ_j and ϕ_{j+1}. As a result, this approach allows for a more accurate DOA estimation with a higher spatial resolution. Each system processes a block of L consecutive time frames, including the current frame and the past $L - 1$ frames, to produce DOA estimates. During testing, the systems continuously process the incoming data and generate a posterior probability map and corresponding DOA estimates for each frame, meaning that they can be used in an online fashion.

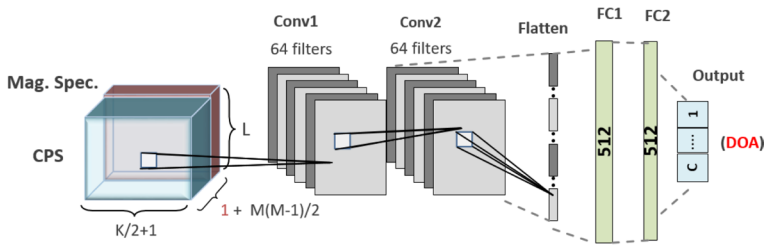
4.1 Baseline systems

Figure 2 depicts the baseline systems consisting of a single-stage CNN using only the CPS phase, a combination of the CPS phase and the magnitude spectrogram, or a combination of the CPS phase and PD features as input. The CNN architecture in all three baseline systems begins with a cascade of two convolutional blocks (*Conv1* and *Conv2*). Each block consists of a 2D convolutional layer, followed by batch normalization and a rectified linear unit (ReLU) activation layer. Only *Conv2* incorporates a max pooling layer after the ReLU. Next, the concatenated outputs of *Conv2* serve as an intermediate feature vector and are fed into two fully connected blocks (*FC1* and *FC2*). These blocks each comprise a fully connected dense layer with batch normalization, ReLU activation, and dropout layers. Finally, the output layer employs C sigmoid activation functions to generate the posterior probability map for the C independent DOA classes.

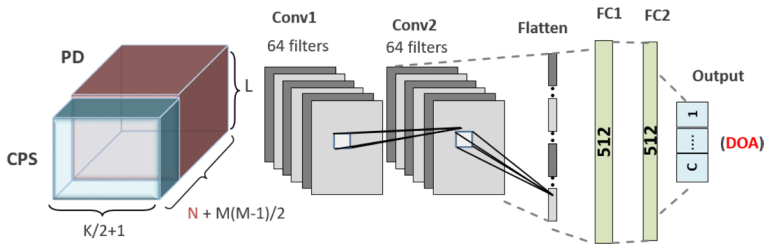
To improve CNN performance, we employ layer normalization [50] (without an affine transformation) directly on the input features before the first convolutional block. This normalization targets the CPS phase, magnitude spectrogram, and PD features separately. It's important to note that this has been implemented in



(a) With only CPS phase



(b) With CPS phase and magnitude spectrogram



(c) With CPS phase and PD

Fig. 2 Baseline DOA estimation systems using **a** only spatial feature (CPS phase), **b** spatial and spectral features (CPS phase and magnitude spectrogram), and **c** spatial and spectral features (CPS phase and PD)

addition to the batch normalization used within the convolutional and fully connected blocks of the CNNs.

4.2 Proposed system

Figure 3 illustrates our proposed multi-talker DOA estimation system, which combines PD features (see Section 3.2) with the CPS phase (see Section 3.1) as input to a two-stage CNN. Inspired by [37], our system features a PD feature reduction stage before joint processing with the CPS phase. Within this reduction stage, 1×1 convolutions [51] are used to transform the N-channel PD features into a single-channel PD saliency feature for each time-frequency bin. The 1×1 convolution applies a kernel with dimension 1×1 that extends across the entire depth of the input, i.e., N fundamental period candidates. Instead of using larger kernels that slide across spectro-temporal regions, the 1×1 kernel processes each time-frequency bin independently while aggregating information across the depth channels. Given the input of size $L \times (K/2+1) \times N$, the 1×1 kernel has dimensions $1 \times 1 \times N$ and performs a weighted sum over the N depth channels, producing a single value at each time-frequency bin. This mechanism makes this type of convolutions highly effective for the dimensionality reduction of PD features, as it compresses the depth while preserving the spectro-temporal resolution. In the subsequent stage, these PD saliency features are jointly processed with CPS features using convolutional filters.

The second stage of our proposed system shares the same architecture as our baseline systems (Section 4.1). To process the combined CPS phase and PD saliency features, we use convolutional blocks (*Conv1* and *Conv2*), each composed of a 2D convolutional layer, batch normalization, and a ReLU activation layer. *Conv2* also includes max pooling after the ReLU. The outputs of the *Conv2* block are then concatenated and fed as an intermediate feature vector into two fully connected blocks (*FC1* and *FC2*). Each block features a fully connected

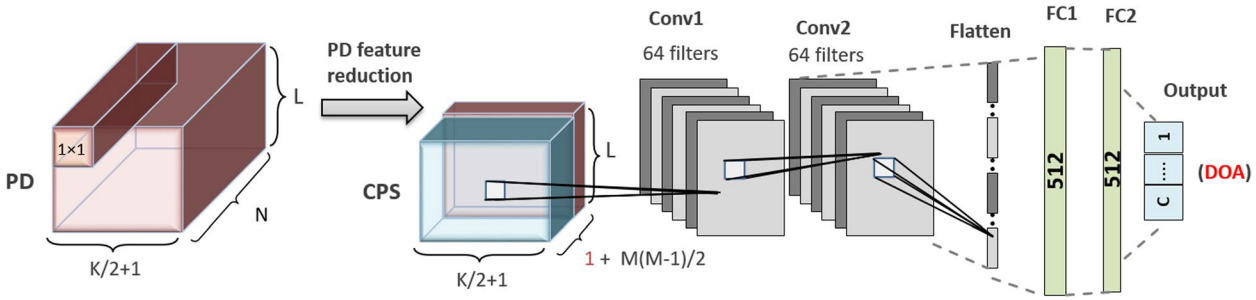


Fig. 3 Proposed system using the CPS phase and PD features as input. The PD features undergo dimensionality reduction via 1×1 convolutions to create compact PD saliency features. These are then combined with CPS phase features as input to the convolutional blocks for joint processing and extraction of local spectro-temporal information related to source DOA

dense layer, batch normalization, a ReLU activation layer, and a dropout layer. Finally, similar to the baseline systems, the output layer uses C sigmoid activation functions to generate the posterior probability map for the C independent DOA classes. We employ the same input layer normalization scheme as our baseline systems, normalizing the PD and CPS phase separately. It should be noted that both stages of the proposed system are trained jointly.

Employing convolutional kernels across L consecutive time frames and $K/2+1$ frequency bins allows for different design approaches to capturing spectro-temporal dependencies. In particular, we have made two main design choices for the two-stage CNN architecture employed for the proposed system.

As our first design choice, we consider a combination of kernel, dilation, and max pooling sizes that reduces the temporal dimension of the input features to a single value at the output of the last convolutional block. This essentially captures temporal correlation solely through the convolutional path. Consequently, the intermediate feature vector at the input of the fully connected path primarily contains elements representing different frequencies. Our CNN design is based on the assumption that convolutional blocks effectively capture temporal dependencies in the input features, while fully connected blocks best capture global patterns across frequencies.

In multi-talker scenarios, neighboring frequency bins may contain dominant activity from different speakers. Hence, previous works [15, 29] have used convolutional kernels that separately process each frequency bin to take benefit from the widely adopted assumption of W-disjoint orthogonality [52]. As the second design choice, we preserve frequency resolution at the output of the convolutional path (and hence, capturing global patterns across frequencies merely via the fully connected path). We expect that this approach may lead to a better DOA estimation performance compared to joint learning of time-frequency dependencies using 2D convolutional kernels.

With an interest in designing a computationally efficient system in this study, the proposed system utilizes a temporal kernel size of 7, a dilation rate of 2, and a temporal max pooling size of 2. This CNN architecture was chosen to help reduce computational costs while maintaining performance. The large kernel size of 7 captures long-range temporal dependencies in the input features. Using dilation in the convolutions expands the receptive field without increasing the number of parameters. The max pooling then further downsamples the temporal dimension to reduce computations in subsequent layers. This specific combination implements our first design choice while capturing large temporal contexts within just two convolutional blocks (*Conv1* and *Conv2* in Fig. 3). To implement our second design choice, we ensure convolutional kernels and max pooling operate exclusively across the time dimension. In the following, we investigate our main design choices by considering alternative approaches.

First, aiming at investigating different approaches to capture temporal dependencies, we employ convolutional blocks with different combinations of dilation and max pooling. This results in reducing the temporal dimension into different numbers of features for each filter at the output of the convolutional path (compared to the single feature in the proposed system). Consequently, both convolutional and fully connected paths contribute to capturing temporal dependencies. We compare our proposed system with two additional two-stage CNN configurations: one with a dilation size of 2 and no max pooling (temporal dimension of 2), and the other with neither dilation nor max pooling (temporal dimension of 8). Similar to the proposed system, these two two-stage configurations use convolutional kernels exclusively across time (see the second and third systems in Table 1).

Second, we explore the usage of kernels that span both time and frequency dimensions using different kernel sizes across frequencies. To do so, the proposed system with the kernel size of 7x1 (only temporal processing) is paired with two alternative two-stage systems that employ 2D kernels across both time and frequency with the sizes of 7x2 and 7x3 (see the fourth and fifth systems in Table 1). It should be noted that the temporal dimension in all three systems is reduced to a single value. To prevent information loss and focus on adjacent frequencies, we avoid dilated kernels across the frequency dimension.

To prevent temporal information loss, in all systems, the first convolutional block (*Conv1*) uses neither dilation nor max pooling. The subsequent convolutional block (*Conv2*) in the alternative systems may incorporate max pooling and/or a dilation rate of 2 across time as specified in Table 1. All systems include convolutional blocks with 64 filters and fullyconnected blocks with 512 neurons.

4.3 Computational complexity

Table 1 shows the number of trainable parameters and multiply-accumulate operations (MACs), both in millions for the proposed system, alternative two-stage systems using different temporal reduction and spectro-temporal strategies, as well as the two baseline systems. The number of parameters, i.e., the model size, influences the memory required to store the model, while MACs provide an estimate of the arithmetic computations, which inherently affects energy consumption.

To investigate the effect of different dilation and max pooling strategies (across time) on the complexity of the two-stage systems (cf. Fig. 3), we consider the first three systems in Table 1. Using different dilation rates and max pooling results in varying degrees of temporal reduction within the convolutional path (*Conv1* and *Conv2*), which consequently leads to intermediate feature vectors with different sizes. This manipulation yields configurations with differing computational complexities, where the complexity scales in proportion to the degree of reduction in the intermediate feature vector. For example, the proposed system, which utilizes the maximum temporal dimension reduction (temporal dimension size of 1) has the minimum computational complexity of 11.5 million MACs, compared to 14.2 and 44 million MACs for the two other two-stage CNN configurations. Similarly, it has the smallest memory footprint (3 million parameters) in contrast to 5.6 and 21.6 million parameters for the two other systems.

To compare spectro-temporal processing strategies, we pair the proposed system (using 1D temporal kernels of 7x1) with alternative two-stage CNNs employing 2D kernels spanning both time and frequency (sizes 7x2 and 7x3). As Table 1 reveals, while the number of parameters remains relatively comparable across different frequency kernel sizes, larger kernels lead to increased MACs. All three systems maintain the same temporal kernel size (7), resulting in a temporal dimension reduction to a single value. Their difference lies in the frequency kernel size (1, 2, and 3). Due to frequency dimension reduction within the convolutional path, the modified systems have fewer intermediate features. This translates to slightly fewer trainable parameters in the system using 7x3 kernels.

Apart from the PD feature reduction stage, the architecture of the proposed system (Fig. 3) closely mirrors the baseline systems (Fig. 2). Since the proposed and baseline systems use the same kernel, dilation and max pooling sizes, as well as the same number of filters (64), they have the same number of intermediate features. Additionally, all systems use 512 neurons in the fully connected blocks. This results in a comparable number of trainable parameters, predominantly determined by the fully connected path. Please note that the PD reduction stage has a negligible impact on the total number of trainable parameters (3 million). However, the inclusion of the PD feature reduction stage, along with the larger input dimensions to the *Conv1* block, only marginally increases the MACs of the proposed system (11.5 million) compared to the

Table 1 Number of trainable parameters and multiply-accumulate operations (MACs) of different systems

CNN architecture	Kernel/dilation	Max pool	Temp. dim.	MACs (M)	Param. (M)
Proposed two-stage	7x1/2x1	2x1	1	11.5	3
Two-stage w/ only Dil.	7x1/2x1	No	2	14.2	5.6
Two-stage w/o MaxP. & Dil.	7x1/No	No	8	44	21.6
Two-stage w/ Spectro-temp. proc. 1	7x2/2x1	2x1	1	19.3	3
Two-stage w/ Spectro-temp. proc. 2	7x3/2x1	2x1	1	26.8	2.9
Baseline w/ only CPS phase	7x1/2x1	2x1	1	10.7	3
Baseline w/ CPS phase & Mag. Spec.	7x1/2x1	2x1	1	11.2	3
Baseline w/ CPS phase & PD	7x1/2x1	2x1	1	102.2	3

baseline systems employing only the CPS phase (10.7 million) or a combination of the CPS phase and the magnitude spectrogram (11.2 million). When comparing the MACs of the proposed two-stage system using the PD and CPS phase as input (cf. Fig. 3) with the baseline system using the same features without the feature reduction stage (cf. Fig. 2c), the feature reduction stage results in a significant reduction in MACs by approximately a factor of 9. This reduction is primarily due to the lower dimensionality of the input to the *Conv1* block.

In summary, among the two-stage CNN configurations explored, the proposed system's focus on temporal feature reduction (to a single value) together with independent frequency processing yields the lowest computational cost and smallest model size. Furthermore, with a slight increase in computational need, the model size of the proposed system remains comparable to the baselines. While the model's low complexity suggests potential for real-time capability, especially given recent advancements in AI-based solutions for hearing aids, further model optimizations to reduce latency, model quantization, pruning, or efficient on-device inference engines are needed, which is beyond the scope of this study.

5 Experimental setup

This section presents experiments evaluating the performance of the multi-talker DOA estimation systems described in Sections 4.2 and 4.1. We detail the employed datasets in Section 5.1. Sections 5.2 and 5.3 describe procedures for generating training and evaluation data. In Section 5.4, implementation details of input features are provided. Section 5.5 presents CNN training procedures and hyperparameters. Finally, Section 5.6 outlines the evaluation metrics used to assess the performance.

5.1 Datasets

We utilized speech signals from 462 and 168 speakers in the TIMIT dataset [53] for training and validation, respectively, including both male and female speakers. For evaluation, speech signals from the testing TIMIT dataset were used as source signals. To generate data for training and evaluation, a database of multichannel binaural room impulse responses (BRIRs) [54] was used. We considered a binaural hearing aid setup consisting of $M = 4$ microphones, where the front and rear microphones (approximate microphone distance of 15 mm) in both left and right hearing aids were used. The database in [54] provides BRIRs for anechoic conditions for $C = 72$ directions in the azimuthal plane (5° resolution). It additionally includes BRIRs for different source-to-head distances in two reverberant environments: a cafeteria ($T_{60} \approx 1.3$ s) and a courtyard ($T_{60} \approx 0.9$ s). Noisy binaural microphone signals were generated by convolving source signals with BRIRs and mixing the resulting clean binaural microphone signals with background noise.

5.2 Training data

For training, clean binaural microphone signals were generated by convolving speech signals with anechoic BRIRs for each of the 72 directions at a fixed 3 m source-to-head distance. The noisy binaural microphone signals were generated by mixing the clean binaural microphone signals with simulated binaural diffuse noise at signal-to-noise ratios (SNRs) ranging from − 5 dB to + 20 dB in 5 dB steps. This noise was generated by convolving uncorrelated speech-shaped noise (ICRA database [55]) with anechoic BRIRs and summing all resulting binaural signals from 72 directions. Training examples included all 72 directions at six different SNRs. In a data pre-processing step, a simple oracle broadband energy-based voice activity detector (VAD) was used to select segments with sufficient speech content, ensuring meaningful data contributed to the loss function. Training examples consisted of blocks of $L = 20$ consecutive time frames (corresponding to 105 ms). We generated a *training set* of 1.9 million examples (approximately 55.4 hours) and a fixed *validation set* of 200000 examples (approximately 5.8 h). A summary of the training data is presented in Table 2.

5.3 Evaluation data

We evaluated the performance of all systems for static source scenarios in reverberant environments. Table 3 summarizes the evaluation setup and data generation. Source signals, background noise signals, acoustic conditions, and source positions used for evaluation were entirely distinct from those used during training. All systems were trained in noisy anechoic conditions with simulated binaural diffuse noise, whereas evaluation was performed in noisy reverberant environments with different recorded background noises [54]. For evaluation, we used entirely different environments from those used during training, namely a cafeteria and a courtyard with a reverberation time of approximately 1300 ms and 900 ms, respectively. Source positions were carefully selected to simulate realistic listening distances and scenarios. The positions were chosen within typical spatial distances from the listener to reflect practical scenarios.

Table 2 Summary of the training data

Source signals	Speech (TIMIT)
Environment	Anechoic [54]
Background noise	Simulated diffuse noise
SNR	− 5 dB to + 20 dB in 5 dB steps
Source-to-head distance	3 m
Source positions	72 positions in the horizontal plane

Table 3 Summary of the evaluation data

Source signals	Speech (TIMIT)
Environment	Cafeteria ($T_{60} \approx 1.3$s) and courtyard ($T_{60} \approx 0.9$s) [54]
Background noise	Recorded noise
SNR	−5 dB to +10 dB in 5 dB steps
Source-to-head distance	1 − 1.6 m
Source positions	4 positions with 2 head orientations in each environment

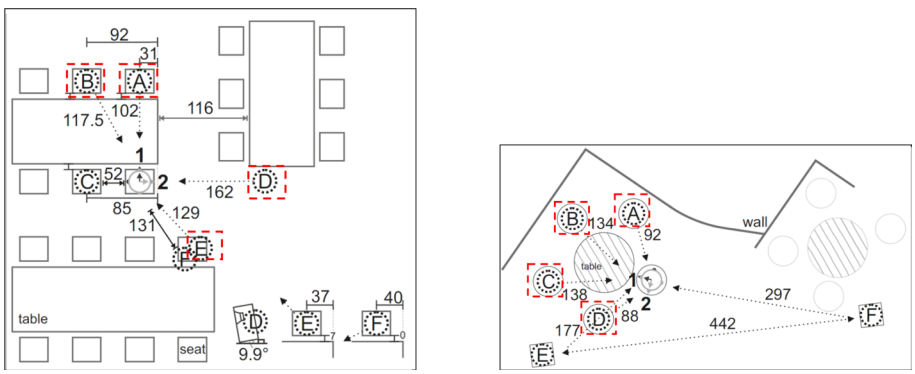
Figure 4a and b illustrate the room configurations for both environments. In each environment, we considered four source positions (depicted with dashed boxes) with two head orientations measured for each position. Two-source and three-source scenarios with constantly active talkers were created for each environment by combining all possible pairs and triplets of considered source positions across the two head orientations, resulting in 12 two-source and 8 three-source scenarios. This approach ensures a wide range of spatial configurations and allows testing the generalization ability of the model to different multi-talker situations in real-world environments. Clean binaural microphone signals were obtained by convolving the speech source signals with reverberant BRIRs [54]. Noisy binaural microphone signals were then generated by mixing these clean microphone signals with recorded binaural cafeteria babble noise or courtyard ambient noise [54] at SNRs ranging from − 5 dB to + 10 dB in 5 dB steps. These recordings were selected to represent diverse and complex real-world acoustic scenarios, allowing to investigate the generalization ability of the models under unmatched background noise conditions. A total number of 150 speech utterances (each with a length of 2 s) randomly chosen from 30 unique male and female speakers were selected from the testing TIMIT dataset.

5.4 Implementation details

All signals were sampled at 16 kHz. We used an STFT framework with a Hann window of length $K = 160$, equivalent to 10 ms, and a hop size of $D = 80$, equivalent to 5 ms. This setup resulted in 81 STFT frequency bins. For each of the 6 microphone pairs, CPS features were computed over a block of $L = 20$ consecutive time frames, resulting in a CPS input feature of size $20 \times 81 \times 6$.

In this paper, we consider the front microphone of the left hearing aid as the reference microphone for the PD and magnitude spectrogram feature extraction. The magnitude spectrogram was also calculated over 20 STFT time frames and 81 frequency bins, resulting in an input feature size of 20×81, aligning with the CPS input dimensions for joint spectro-temporal processing.

Feature extraction for PD employed a 4-th GTFB implementation [45], utilizing 61 bands. This setup included a group delay of 256 and center frequencies ranging from a minimum of 60 Hz to a maximum of 7200 Hz. The fundamental frequency settings were selected at a minimum of 70 Hz and a maximum of 320 Hz, establishing a fundamental period range for PD features from 14.3 ms to 3.1 ms across $N = 180$ candidate periods. The considered fundamental frequency candidates span the range typically observed in both male and female speech. The comb filter gain was chosen to be $\alpha = 0.7$. After adjusting the frequency resolution of the PD features to match the STFT (see Section 3.2), the input PD features had a size of $20 \times 81 \times 180$, aligning them with the spectro-temporal dimensions of the CPS input features.



(a) Cafeteria (static scenario) (b) Courtyard (static scenario)

Fig. 4 Evaluation setups for static scenarios, adapted from [54]. In the cafeteria, source positions *A*, *B*, *D*, and *E* were considered, while in the courtyard, source positions *A*, *B*, *C*, and *D* were considered. Dashed arrows extending from each source position towards the head indicate the head location. Head orientations are indicated by the numerals 1 and 2, which are placed close to the head icon

5.5 Training and network hyperparameters

All systems were implemented using PyTorch [56]. Convolutional blocks in all CNNs employed 64 filters with a stride size of 1×1. When used, max pooling had a size of 2×1 (i.e., across time) with strides of the same size. The training was conducted using the Adam optimizer [57], a binary cross-entropy loss function, an initial learning rate of 10^{-4}, a mini-batch size of 128, and a dropout rate of 0.5. We employed early stopping regularization that terminated training if the validation loss did not improve for 10 epochs. A variable learning rate scheduler was also used, halving the learning rate if the validation loss did not improve for 2 epochs. The maximum training epoch number was set to 100. Each epoch randomly selected 1.9 million non-repeating examples from the training set. Mini-batches were assembled randomly, drawing examples from various SNR conditions and DOA classes. The validation data were not seen by the network during the training.

5.6 Evaluation metrics

We evaluated the DOA estimation performance of all systems in terms of accuracy (Acc.) [14, 15]. For a signal block l containing a mixture of $\mathcal{I}$ simultaneously active sources, the DOA estimate for the i-th source ($\hat{\theta}_i^l$) is considered accurate if the absolute error between the DOA estimate and the oracle DOA (θ_i^l) is less than 5°, corresponding to the minimum angular resolution of the database in [54]. Since all sources are assumed to be constantly active, there is always a one-to-one matching between a DOA estimate and an oracle DOA. The assignment of DOA estimates to oracle DOAs is achieved using the Hungarian algorithm [58] by identifying the one-to-one matching combination that minimizes the total absolute error between the estimated and oracle DOAs [16, 29, 59]. The accuracy is defined as

$$\text{Acc.} = \frac{1}{\mathcal{L}\mathcal{I}} \sum_{l=1}^{\mathcal{L}} \sum_{i=1}^{\mathcal{I}} H\left(5 - \left|\hat{\theta}_i^l - \theta_i^l\right|\right) \times 100, \qquad (7)$$

where $\mathcal{L}$ denotes the total number of signal blocks, and H denotes the Heaviside step function. Please note that the Heaviside step function is defined here such that it returns 1 if the absolute error is less than 5° (an accurate estimate), and 0 otherwise (an inaccurate estimate).

6 Results and discussion

In this section, we will present and analyze the performance evaluation results of the proposed system and the alternative two-stage configurations employing PD and CPS phase features, along with baseline systems using either the CPS phase or the combination of the CPS phase and magnitude spectrogram. We assessed the performance of all systems in different reverberant environments with different background noises for both static two-source and three-source scenarios in terms of accuracy. Section 6.1 compares the proposed system to the two alternative two-stage systems using different temporal dilation and max pooling strategies. In Section 6.2, we compare the performance of the proposed system and the two alternative two-stage systems using different spectro-temporal processing strategies. Finally, Section 6.3 discusses the performance evaluation of the proposed and baseline systems.

6.1 Different temporal reduction strategies

For two-source and three-source scenarios in two reverberant environments (cafeteria and courtyard), Fig. 5 shows the accuracy at different SNRs for the proposed system, and the two alternative two-stage systems (Section 4.2) using different temporal feature reduction strategies. The blue bars indicate the evaluation results of the proposed system employing dilated kernels and max pooling across time that leads to a temporal dimension of size 1 for each filter output of the *Conv*2 block (cf. Fig. 3). The red and yellow bars correspond to alternative configurations, one with dilated kernels but no max pooling, leading to a temporal dimension of 2, and another with neither max pooling nor dilation across time, yielding a temporal dimension of 8 (as detailed in Table 1).

In particular, we intend to test the hypothesis that the temporal dependencies in the input features can be effectively captured merely by convolutional blocks, while the frequency resolution of features is preserved. The latter is ensured by using no max pooling and no convolutional kernels across frequencies. This would essentially mean that global patterns across frequencies are exclusively captured by the fully connected blocks.

In the cafeteria environment, the proposed system clearly outperforms the alternative two-stage configurations, particularly under non-negative SNR conditions. In the courtyard environment, all three systems perform comparably, with slightly higher accuracy observed in the two alternative two-stage configurations, albeit at the expense of significantly increased computational costs and model size, such as in the case of the system that does not use temporal max pooling or dilated kernels (cf. Table 1)

It is important to note that using a model with the highest computational complexity does not necessarily lead to a better performance in terms of accuracy. In fact, the proposed system, which captures temporal dependencies in the input features through convolutional blocks, proves to be a favorable approach overall due to its more

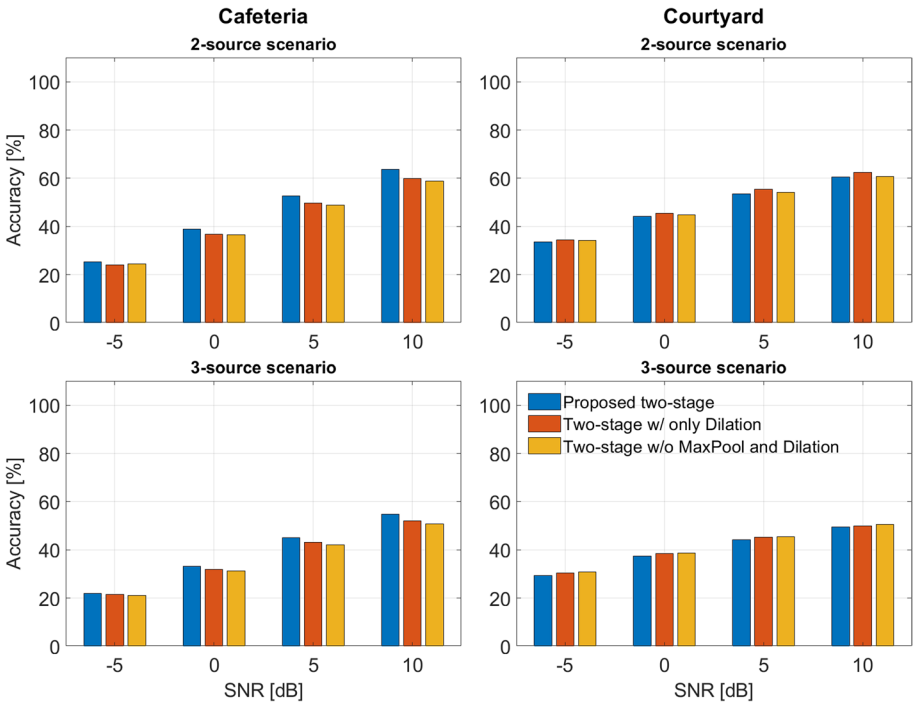


Fig. 5 Accuracy of the proposed system and the two two-stage CNN configurations using different temporal feature reductions for static source scenarios in two reverberant environments (cafeteria and courtyard) for different SNRs and two-source and three-source scenarios. The proposed system (indicated by the blue bar) employs a dilation and max pooling size combination that leads to a temporal dimension of size 1 for each filter output, while the red and yellow bars correspond to counterpart configurations, resulting in temporal dimensions of sizes 2 and 8, respectively. All systems use the combination of the PD and CPS phase as input, without any max pooling or convolutional kernels across frequencies

efficient configuration with less computational complexity and model size.

6.2 Different spectro-temporal filtering strategies

In this section, in comparison to the proposed system, which only employs dilated convolutional kernels of size 7 and dilation rate of 2 across time, we investigate the potential benefit of using kernels across both time and frequency in alternative two-stage systems using kernels of sizes 2 and 3 across frequencies, while utilizing the same temporal kernel sizes as the proposed system. In all systems, the temporal dimension across convolutional blocks is reduced to one.

For two-source and three-source scenarios in two reverberant environments (cafeteria and courtyard), Fig. 6 shows the accuracy at different SNRs for the three two-stage systems using the PD and CPS phase as inputs through different spectro-temporal feature processing. The blue bars indicate the evaluation results of the proposed system using the kernel size of 7x1, while the red and yellow bars represent the alternative two-stage system configurations using kernel sizes of 7x2 and 7x3, respectively.

It can be clearly observed that in all environments and SNR conditions for the two-source and three-source scenarios, the proposed system performs comparably to or better than the alternative two-stage systems. This demonstrates that the two-stage CNN using PD and CPS features does not benefit from the joint spectro-temporal filtering using 2D kernels. This also suggests that while capturing the temporal dependencies solely through the convolutional blocks (i.e., temporal dimension reduction to a single value), it is more effective to process each frequency independently through the convolutional path. In this way, the fully connected layers alone can effectively learn global patterns across frequencies, rather than having both the convolutional and fully connected blocks contribute to learning these patterns. It is particularly notable when considering the additional computational load from the joint spectro-temporal processing in the convolutional path (cf. Table 1), which further demonstrates the benefit of independently processing each frequency.

6.3 Comparison against baseline systems

This section evaluates the DOA estimation performance of the proposed system (cf. Fig. 3) against the three considered baseline systems (cf. Fig. 2). All baseline systems utilize the same convolutional kernel, dilation, and max pooling strategies as the proposed system (cf. Table 1).

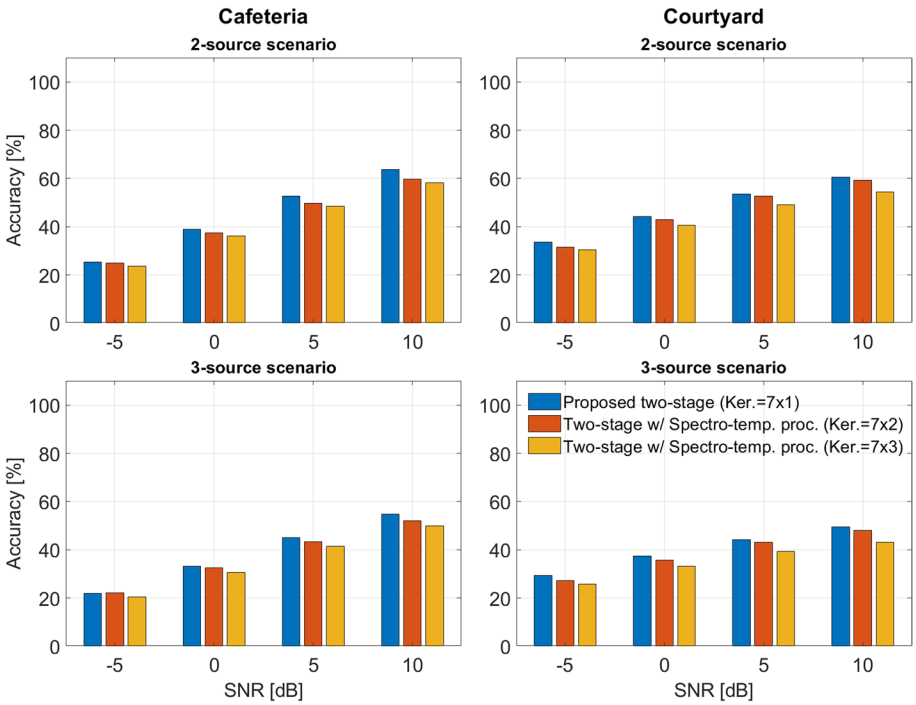


Fig. 6 Accuracy of the proposed system and the two two-stage CNN configurations using different spectro-temporal processing for static source scenarios in two reverberant environments (cafeteria and courtyard) for different SNRs and two-source and three-source scenarios. The proposed system (indicated by the blue bar) employs a kernel size of 1 across frequencies (no frequency correlation), while the red and yellow bars correspond to counterpart configurations, using kernel sizes of 2 and 3 across frequencies. All systems use the combination of PD and CPS phase as input, and employ kernels of size 7 with a dilation rate and max pooling size of 2 across time, and without any max pooling across frequencies

The key distinction is that the proposed system employs a two-stage CNN that includes a feature reduction stage before merging the PD saliency features and the CPS phase features. In contrast, the baseline systems directly employ the input features, either the CPS phase, the CPS phase and the magnitude spectrogram, or the CPS phase and the PD features.

6.3.1 Benefit of using PD features

In this section, we evaluate the advantage of incorporating PD features in combination with the CPS phase as a spatial feature in our proposed system (cf. Fig. 3), compared to baseline systems that use either the CPS phase or a combination of CPS phase and magnitude spectrogram as a spectral feature (cf. Fig. 2).

For two-source and three-source scenarios in two reverberant environments (cafeteria and courtyard), Fig. 7 depicts the accuracy at different SNRs for the proposed system and the two baseline systems. The blue bars indicate the evaluation results of the proposed system, while the red and yellow bars represent the baseline systems. For all conditions and environments, the proposed system clearly benefits from using the PD features in combination with CPS features, when compared to the two baseline systems. For example, for an SNR of 0 dB in the courtyard environment, for two-source scenarios, the benefit of using PD features is approximately 4% points compared to the baseline system using only the CPS phase, and 5% points compared to the baseline system using the magnitude spectrogram and CPS phase. For three-source scenarios, the benefit of using PD features is approximately 5% points compared to the baseline system using only the CPS phase and 3% points compared to the baseline system using the magnitude spectrogram and CPS phase. It is also evident from Fig. 7 that while the performance of all systems, regardless of input features, degrades with the number of overlapping talkers, the proposed system maintains its advantage in both environments. It is expected that further increasing the number of simultaneously active talkers will negatively impact the DOA estimation performance, especially in noisy and reverberant environments [14, 16, 29].

We can also observe from Fig. 7 that the benefit of using PD features increases with SNR. For example, comparing the performance of the proposed system and the baseline system using only the CPS phase in the cafeteria environment exhibits that, for the two-source scenarios, this benefit increases from about 1% points at -5 dB to about 5% points at 10 dB SNR condition. At higher SNRs, the impact of background noise is reduced, thus

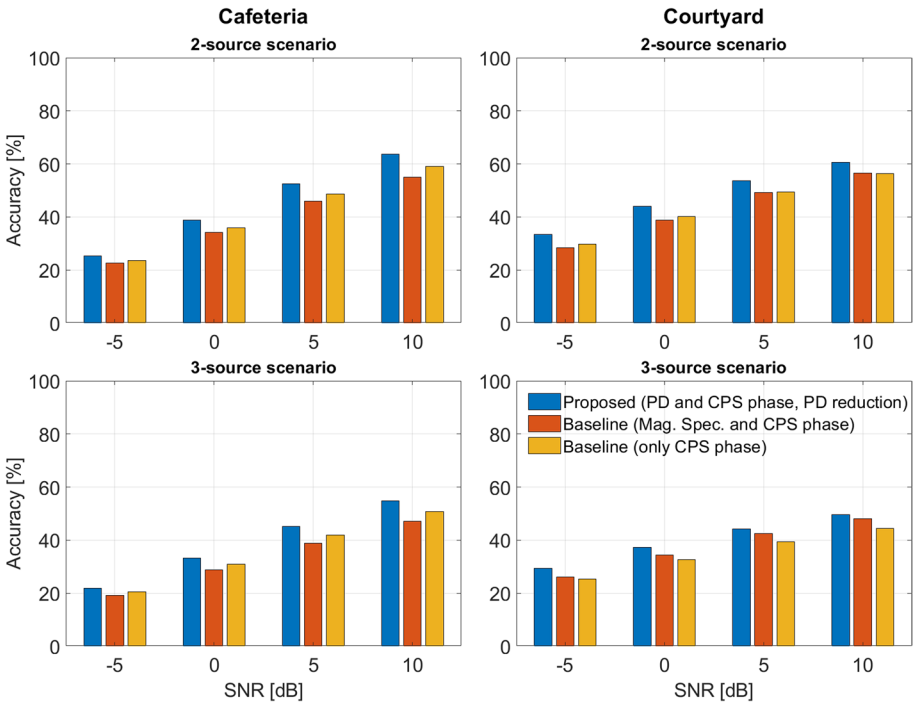


Fig. 7 Accuracy of the proposed system, and the two baseline systems for static source scenarios in two reverberant environments (cafeteria and courtyard) for different SNRs and two-source and three-source scenarios. The proposed system (indicated by the blue bar) employs the PD and CPS phase as input, while baseline systems specified by the red and yellow bars employ the combination of magnitude spectrogram and CPS phase, and the CPS phase, respectively. All systems employ a convolutional kernel size of 7 with dilation rate and max pooling size of 2 across time, without using any max pooling or convolutional kernels across frequencies

emphasizing the periodicity characteristics of speech and making the periodicity features more distinguishable. This improved discriminability enhances the ability of the proposed system to effectively use PD features in conjunction with spatial CPS features, allowing the PD features to contribute more meaningfully to the accuracy of the DOA estimation.

Including the magnitude spectrogram in combination with the CPS phase seems to be advantageous merely in the courtyard environment, in particular, for the theesource scenario. This observation suggests that, unlike PD features, the usage of the magnitude spectrogram as a spectral feature in combination with the CPS phase does not offer significant benefits for DOA estimation for the considered settings and environments when compared to using only CPS phase features. On the one hand, while PD features provide a clear indication of the source's harmonic structure, the magnitude spectrogram provides a broad spectral representation that may not be as effective in isolating the specific characteristics of speech needed for accurate DOA estimation or may require a much more sophisticated network architecture to capture these characteristics. On the other hand, PD features are less susceptible to noise that does not share the harmonic structure of the sound sources of interest, while magnitude spectrogram features are more general and can capture both the speech signal and noise without distinguishing between them, making it harder to identify speech sources in noisy environments.

6.3.2 Benefit of using the feature reduction stage

In this section, we compare the performance of the proposed two-stage system incorporating a PD feature reduction stage to the baseline system that directly combines the CPS phase and PD features without applying feature reduction.

Figure 8 illustrates the results for two-source and threesource scenarios in two reverberant environments across varying SNRs. The blue and red bars represent the performance of the proposed system and the baseline system, respectively. In the cafeteria environment with recorded cafeteria babble noise, a significant benefit of incorporating the feature reduction stage can be observed for both two-source and three-source scenarios. For instance, for an SNR of 0 dB, the benefit amounts to approximately 10% and 8% points for two-source and three-source scenarios, respectively. In the courtyard environment, a considerable benefit of the PD feature reduction is obtained for two-source scenarios (especially at low SNRs), while for three-source scenarios both systems demonstrate

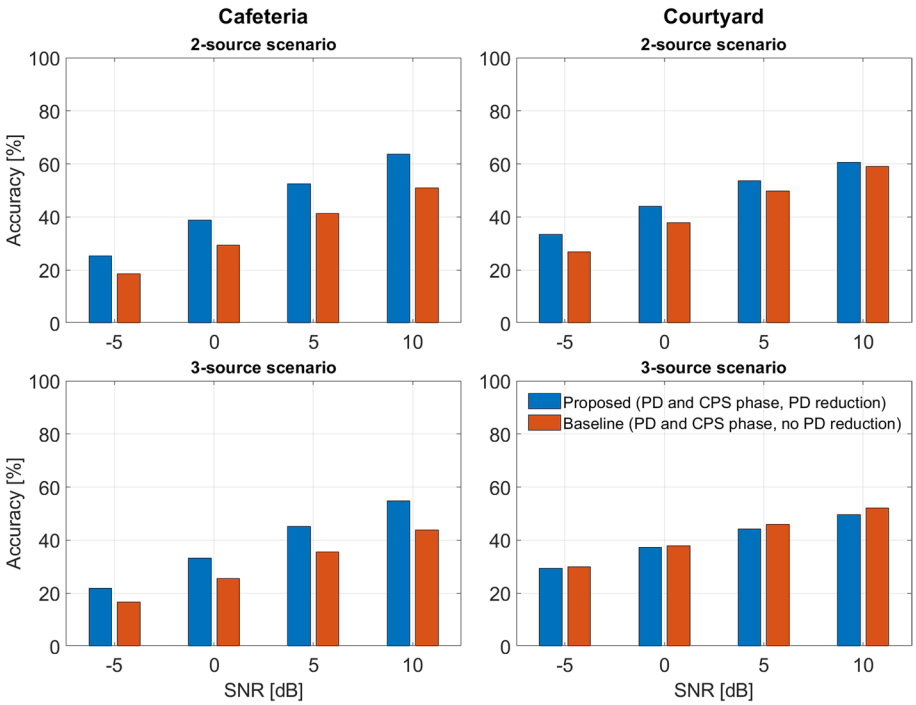


Fig. 8 Accuracy of the proposed two-stage system with PD feature reduction stage and the baseline system without PD feature reduction. Both systems are evaluated for static source scenarios in two reverberant environments (cafeteria and courtyard) under varying SNRs and two-source and three-source scenarios. Both systems employ a convolutional kernel size of 7 with dilation rate and max pooling size of 2 across time, without using any max pooling or convolutional kernels across frequencies

comparable performance. Hence, for all scenarios the proposed system yields a better or comparable performance than the baseline system at a significantly lower computational complexity (cf. Table 1).

These results hence show that the PD feature reduction stage is not only useful for single-talker scenarios, as previously shown in [37], but also for multi-talker scenarios. By reducing the sparse structure of narrowband PD features and focusing on the most salient glimpses of these features, the feature reduction stage can be viewed as an implicit attention mechanism, where certain spectro-temporal regions of the spatial features are weighted more heavily based on their alignment with the salient glimpses of the PD features. For the baseline system without PD reduction stage to reach the performance of the proposed two-stage system, a more sophisticated network structure may be needed, which would increase its computational complexity even more.

7 Conclusion

This paper investigated the effectiveness of combining periodicity and spatial features for multi-talker DOA estimation in binaural hearing aids using a two-stage convolutional neural network (CNN) architecture. The proposed system utilized periodicity degree (PD) features as spectral features in combination with cross-power spectrum (CPS) phase as spatial features.

Several design choices for the two-stage CNN architecture were explored, including different strategies for temporal feature reduction through dilation and max pooling, as well as spectro-temporal filtering approaches using convolutional kernels of varying sizes. Experimental results demonstrated that the proposed system, which effectively captures the temporal dependencies within the convolutional blocks alone while independently processing each frequency, leads to the best performance. Furthermore, the proposed system offers advantages in terms of computational complexity compared to alternative configurations.

The evaluation results in terms of DOA estimation accuracy for two-source and three-source scenarios across two reverberant environments and various SNRs demonstrated that the proposed system outperforms baseline systems that utilized either only CPS phase features or a combination of CPS phase and magnitude spectrogram features, without requiring significantly higher computational complexity. This underscores the advantage of incorporating PD features for multi-talker DOA estimation. The proposed system also outperformed a baseline system utilizing CPS phase and

PD features without a PD feature reduction stage while requiring significantly lower computational complexity, highlighting the benefit of the PD feature reduction stage.

This study paves the way for advancements in sound source localization and speech enhancement for binaural hearing aids. By combining periodicity and spatial features, the research demonstrates the potential for more accurate DOA estimation and broader improvements in various speech-related tasks. Moreover, this study underscores the importance of feature selection in designing systems for complex auditory scene analysis, particularly in noisy and reverberant environments and multi-talker scenarios.

Future work could explore the adaptation and integration of the proposed system for real-time processing pipelines. Additionally, further research could investigate using PD features to enhance the spatial features by taking alternative approaches other than the direct combination of features. For instance, by exploiting periodicity features for learning-based mask estimation techniques, which might potentially achieve even better DOA estimation performance.

Acknowledgements
Not applicable.

Authors' contributions
The contribution of the first author consists of developing the main algorithmic idea, performing simulations, analyzing the simulation results, and drafting the article. The contribution of the second and third authors consists of critically discussing the developed algorithms and the simulation results with the first author and proofreading and revising the article. All authors read and approved the final manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project ID 352015383 - SFB 1330 B2.

Data availability
All datasets used during this study are included in [53, 54]. Model code may be available from the corresponding author upon reasonable request.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 30 April 2024 Accepted: 6 January 2025
Published online: 01 February 2025

References

1. S. Gannot, E. Vincent, S. Markovich-Golan, A. Ozerov, A consolidated perspective on multimicrophone speech enhancement and source separation. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(4), 692–730 (2017)
2. S. Doclo, W. Kellermann, S. Makino, S.E. Nordholm, Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones. IEEE Signal Process. Mag. **32**(2), 18–30 (2015)
3. D. Wang, G.J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications* (Wiley-IEEE press, New Jersey, 2006)
4. J. Blauert, *The technology of binaural listening* (Springer-Verlag, Germany, 2013)
5. T. May, S. van de Par, A. Kohlrausch, *Binaural localization and detection of speakers in complex acoustic scenes* (Springer-Verlag, Germany, 2013), pp.397–425
6. T. May, S. van de Par, A. Kohlrausch, A probabilistic model for robust localization based on a binaural auditory front-end. IEEE Trans. Audio Speech Lang. Process. **19**(1), 1–13 (2011)
7. C. Knapp, G. Carter, The generalized correlation method for estimation of time delay. IEEE Trans. Acoust. Speech Signal Process. **24**(4), 320–327 (1976)
8. M. Raspaud, H. Viste, G. Evangelista, Binaural source localization by joint estimation of ILD and ITD. IEEE Trans. Audio Speech Lang. Process. **18**(1), 68–77 (2010)
9. M. Dietz, S.D. Ewert, V. Hohmann, Auditory model based direction estimation of concurrent speakers from binaural signals. Speech Commun. **53**(5), 592–605 (2011)
10. S. Braun, W. Zhou, E.A.P. Habets, Narrowband direction-of-arrival estimation for binaural hearing aids using relative transfer functions. in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (IEEE, New Paltz, 2015), pp. 1–5
11. M. Farmani, M.S. Pedersen, Z. Tan, J. Jensen, Bias-compensated informed sound source localization using relative transfer functions. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(7), 1275–1289 (2018)
12. D. Fejgin, S. Doclo, Assisted RTF-vector-based binaural direction of arrival estimation exploiting a calibrated external microphone array. in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, Rhodes Island, 2023), pp. 1–5
13. P.A. Grumiaux, S. Kitić, L. Girin, A. Guérin, A survey of sound source localization with deep learning methods. J. Acoust. Soc. Am. **152**(1), 107–151 (2022)
14. N. Ma, T. May, G.J. Brown, Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(12), 2444–2453 (2017)
15. S. Chakrabarty, E.A.P. Habets, Multi-speaker DOA estimation using deep convolutional networks trained with noise signals. IEEE J. Sel. Top. Signal Process. **13**(1), 8–21 (2019)
16. S. Adavanne, A. Politis, T. Virtanen, Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network, in *2018 26th European Signal Processing Conference (EUSIPCO)* (IEEE, Rome, 2018), pp. 1462–1466
17. W. He, P. Motlicek, J. Odobez, Deep neural networks for multiple speaker detection and localization. in *Proc. IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, Brisbane, 2018), pp. 74–79
18. Z.Q. Wang, X. Zhang, D. Wang, Robust speaker localization guided by deep learning-based time-frequency masking. IEEE/ACM Trans. Audio Speech Lang. Process. **27**(1), 178–188 (2019)
19. P. Vecchiotti, N. Ma, S. Squartini, G.J. Brown, End-to-end binaural sound localisation from the raw waveform. in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, Brighton, 2019), pp. 451–455
20. C. Pang, H. Liu, X. Li, Multitask learning of time-frequency CNN for sound source localization. IEEE Access **7**, 40725–40737 (2019)
21. J. Wang, J. Wang, K. Qian, X. Xie, J. Kuang, Binaural sound localization based on deep neural network and affinity propagation clustering in mismatched HRTF condition. EURASIP J. Audio Speech Music Process. **2020**(1), 1–16 (2020)
22. J. Ding, Y. Ke, L. Cheng, C. Zheng, X. Li, Joint estimation of binaural distance and azimuth by exploiting deep neural networks. J. Acoust. Soc. Am. **147**(4), 2625–2635 (2020)
23. L. Wang, Z. Jiao, Q. Zhao, J. Zhu, Y. Fu, Framewise multiple sound source localization and counting using binaural spatial audio signals. in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, Rhodes Island, 2023), pp. 1–5
24. Q. Yang, Y. Zheng, Deepear: Sound localization with binaural microphones. IEEE Trans. Mob. Comput. **23**(1), 359–375 (2024)

25. B. Yang, H. Liu, X. Li, Learning deep direct-path relative transfer function for binaural sound source localization. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 3491–3503 (2021)
26. H. Hammer, S.E. Chazan, J. Goldberger, S. Gannot, Dynamically localizing multiple speakers based on the time-frequency domain. EURASIP J. Audio Speech Music Process. **2021**(1), 1–10 (2021)
27. P. Goli, S. van de Par, Deep learning-based speech specific source localization by using binaural and monaural microphone arrays in hearing aids. IEEE/ACM Trans. Audio Speech Lang. Process. **31**, 1652–1666 (2023)
28. A.S. Subramanian, C. Weng, S. Watanabe, M. Yu, D. Yu, Deep learning based multi-source localization with source splitting and its effectiveness in multi-talker speech recognition. Comput. Speech Lang. **75**, 101360 (2022)
29. A. Bohlender, A. Spriet, W. Tirry, N. Madhu, Exploiting temporal context in CNN based multisource DOA estimation. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 1594–1608 (2021)
30. P. Cooreman, A. Bohlender, N. Madhu, CRNN-based multi-DOA estimator: Comparing classification and regression. in *Speech Communication; 15th ITG Conference* (VDE, Aachen, 2023), pp. 156–160
31. H. Sundar, W. Wang, M. Sun, C. Wang, Raw waveform based end-to-end deep convolutional network for spatial localization of multiple acoustic sources. in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, Barcelona, 2020), pp. 4642–4646
32. X. Xiao, S. Zhao, X. Zhong, D.L. Jones, E.S. Chng, H. Li, A learning-based approach to direction of arrival estimation in noisy and reverberant environments. in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, South Brisbane, 2015), pp. 2814–2818
33. F.B. Gelderblom, Y. Liu, J. Kvam, T.A. Myrvoll, Synthetic data for DNN-based DOA estimation of indoor speech. in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, Toronto, 2021), pp. 4390–4394
34. R. Roden, N. Moritz, S. Gerlach, S. Weinzierl, S. Goetze, On sound source localization of speech signals using deep neural networks. in *Deutsche Jahrestagung Akustik (DAGA)* (DEGA, Nuremberg, 2015), pp. 1510–1513
35. J. Pak, J.W. Shin, Sound localization based on phase difference enhancement using deep neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. **27**(8), 1335–1345 (2019)
36. W. Ma, X. Liu, Phased microphone array for sound source localization with deep learning. Aerosp. Syst. **2**(2), 71–81 (2019)
37. R. Varzandeh, S. Doclo, V. Hohmann, *A two-stage CNN with feature reduction for speech-aware binaural DOA estimation*. in *2023 31st European Signal Processing Conference (EUSIPCO)* (IEEE, Helsinki, 2023), pp. 241–245
38. R. Varzandeh, S. Doclo, V. Hohmann, Speech-aware binaural DOA estimation utilizing periodicity and spatial features in convolutional neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. **32**, 1198–1213 (2024)
39. R. Varzandeh, K. Adiloğlu, S. Doclo, V. Hohmann, *Exploiting periodicity features for joint detection and DOA estimation of speech sources using convolutional neural networks*. in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, Barcelona, 2020), pp. 566–570
40. R. Gu, S.X. Zhang, M. Yu, D. Yu, *3D spatial features for multi-channel target speech separation*. in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (IEEE, Cartagena, 2021), pp. 996–1002
41. D.A. Krause, G. García-Barrios, A. Politis, A. Mesaros, Binaural sound source distance estimation and localization for a moving listener. IEEE/ACM Trans. Audio Speech Lang. Process. **32**, 996–1011 (2024)
42. J. Woodruff, D. Wang, Binaural localization of multiple sources in reverberant and noisy environments. IEEE Trans. Audio Speech Lang. Process. **20**(5), 1503–1512 (2012)
43. A. Josupeit, V. Hohmann, Modeling speech localization, talker identification, and word recognition in a multi-talker setting. J. Acoust. Soc. Am. **142**(1), 35–54 (2017)
44. S. Popham, D. Boebinger, D.P. Ellis, H. Kawahara, J.H. McDermott, Inharmonic speech reveals the role of harmonicity in the cocktail party problem. Nat. Commun. **9**(1), 1–13 (2018)
45. Z. Chen, V. Hohmann, Online monaural speech enhancement based on periodicity analysis and a priori SNR estimation. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(11), 1904–1916 (2015)
46. A. Josupeit, N. Kopčo, V. Hohmann, Modeling of speech localization in a multi-talker mixture using periodicity and energy-based auditory features. J. Acoust. Soc. Am. **139**(5), 2911–2923 (2016)
47. J. Luberadzka, H. Kayser, V. Hohmann, Making sense of periodicity glimpses in a prediction-update-loop-A computational model of attentive voice tracking. J. Acoust. Soc. Am. **151**(2), 712–737 (2022)
48. W. Mack, J. Wechsler, E.A. Habets, Signal-aware direction-of-arrival estimation using attention mechanisms. Comput. Speech Lang. **75**, 101363 (2022)
49. J.O. Smith, X. Serra, PARSHL: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation. in *Proc. International Computer Music Conference (ICMC)* (Michigan Publishing, Champaign/Urbana, 1987), pp. 290–297
50. J.L. Ba, J.R. Kiros, G.E. Hinton, Layer normalization. 2016. arXiv preprint arXiv:1607.06450
51. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions. in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Boston, 2015), pp. 1–9
52. O. Yilmaz, S. Rickard, Blind separation of speech mixtures via time-frequency masking. IEEE Trans. Signal Process. **52**(7), 1830–1847 (2004)
53. J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, N.L. Dahlgren, V. Zue, TIMIT acoustic-phonetic continuous speech corpus. LDC93S1, Linguist. Data Consortium (1993)
54. H. Kayser, S.D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, B. Kollmeier, Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses. EURASIP J. Adv. Signal Process. **2009**(1), 298605 (2009)
55. W.A. Dreschler, H. Verschuure, C. Ludvigsen, S. Westermann, ICRA noises: Artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment. Audiology **40**(3), 148–157 (2001)
56. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: An imperative style, high-performance deep learning library. in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32 (Curran Associates, Vancouver, 2019), pp. 8026–8037
57. D.P. Kingma, J. Ba, Adam: A method for stochastic optimization. in *Proc. International Conference on Learning Representations (ICLR)* (San Diego, 2015)
58. H.W. Kuhn, The Hungarian method for the assignment problem. Nav. Res. Logist. Q. **2**(1–2), 83–97 (1955)
59. C. Evers, H.W. Löllmann, H. Mellmann, A. Schmidt, H. Barfuss, P.A. Naylor, W. Kellermann, The LOCATA challenge: Acoustic source localization and tracking. IEEE/ACM Trans. Audio Speech Lang. Process. **28**, 1620–1643 (2020)

Publisher's Note