Speech-Aware Binaural DOA Estimation Utilizing Periodicity and Spatial Features in Convolutional Neural Networks

Reza Varzandeh, Simon Doclo, *Senior Member, IEEE*, and Volker Hohmann

***Abstract*—In recent years, several supervised learning-based approaches have been proposed for estimating the direction of arrival (DOA) of a single talker in noisy and reverberant environments. In the absence of auxiliary information, such as a voice activity detector (VAD), the estimated DOA may be erroneous due to speech pauses or noise dominance. In this paper, we consider a speech-aware DOA estimation system for binaural hearing aids, which does not require a separate VAD. This system utilizes a combination of spatial features with an auditory-inspired periodicity feature called periodicity degree (PD) as input features of a convolutional neural network (CNN). Using speech and non-speech signals during the training, the CNN can capture the harmonic structure encoded in the PD features, thereby distinguishing speech from non-speech portions and simultaneously mapping spatial features to sound source DOA upon speech detection. To investigate the benefit of using PD features for speech-aware DOA estimation, we evaluated the performance of speech-aware systems that utilized either broadband or narrowband feature combinations compared to baseline systems. We propose to use a novel narrowband feature combination consisting of the narrowband cross-power spectrum (CPS) as the spatial feature and a new subband-averaged representation of PD features. The broadband feature combination consisted of the generalized cross-correlation with phase transform (GCC-PHAT) and the broadband PD features. The baseline systems considered in this work consisted of a CNN that exploits only a spatial feature, cascaded with a VAD. Evaluations in reverberant environments with different background noises for both static and dynamic single-talker scenarios demonstrate that incorporating the PD feature in conjunction with any type of spatial feature provides an advantage for binaural DOA estimation in terms of accuracy and angular error.**

***Index Terms*—Convolutional neural networks, spatial feature, periodicity feature, binaural DOA estimation, hearing devices.**

I. Introduction

RELIABLY estimating the direction of arrival (DOA) of a target speech source is a crucial task in applications such as binaural hearing aids. Several DOA estimation approaches have addressed this task. The model-based approaches [1], [2], [3], [4], [5], [6] typically rely on specific assumptions about the signal, noise, or reverberation model, which can be violated in adverse noisy and reverberant conditions, leading to a degraded DOA estimation performance. In addition to model-based DOA estimation approaches, in recent years several supervised learning-based DOA estimation approaches based on deep neural networks (DNNs) have been proposed [7], [8], [9], [10], [11], [12], [13], [14], [15], which can provide more robust performance in adverse scenarios when trained in different acoustic conditions [8], [10].

Most DNN-based binaural DOA estimation methods directly map features extracted from the signal to the sound source DOA [8], [12], [13], while some methods follow a two-step approach by first transforming signal features into enhanced features [11], [15]. The most frequently-used (spatial) features for binaural DOA estimation are the interaural level difference (ILD), the interaural time difference (ITD), the cross-correlation function (CCF), and the generalized cross-correlation with phase transform (GCC-PHAT) [16]. The complete CCF or the GCC-PHAT are typically used as the input feature for the DNN [8], [13], as this was shown to outperform using the ITD as the input feature [8]. Whereas most methods estimate the DOA in the azimuthal plane [8], [10], [13], [15], a few methods use multi-task learning approaches to jointly estimate the sound source azimuth together with elevation [12], [14]. In this work, we only consider binaural DOA estimation in the azimuthal plane.

As a common DNN-based approach, the binaural DOA estimation task is often formulated as a classification problem, aiming at determining a mapping from input to a spatial probability map for a discretized azimuth range [8], [10], [13]. For instance, a binaural sound localization system was proposed in [10], which employs a convolutional neural network (CNN) to find a mapping from the raw binaural signal waveforms to a posterior probability map. Although this system has been successfully able to outperform the baseline system with the GCC-PHAT input feature, it has only been trained and evaluated for noiseless scenarios, which is unrealistic in practical situations. Another category of DNN-based approaches involves the task of sound event localization and detection, which aims to identify and localize specific sound events in audio recordings, including both speech and non-speech events [17], [18]. In this paper, we focus on classification-based binaural DOA estimation, specifically aiming at DOA estimation of a single speech source.

Manuscript received 17 May 2023; revised 31 October 2023 and 27 December 2023; accepted 8 January 2024. Date of publication 23 January 2024; date of current version 3 February 2024. This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Project 352015383—SFB 1330 B2. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Daniele Salvati. *(Corresponding author: Reza Varzandeh.)*

The authors are with the Department of Medical Physics and Acoustics and the Cluster of Excellence Hearing4all, University of Oldenburg, 26111 Oldenburg, Germany (e-mail: reza.varzandeh@uol.de; simon.doclo@uol.de; volker.hohmann@uol.de).

Digital Object Identifier 10.1109/TASLP.2024.3356987

A challenge when applying DOA estimation systems in real-life scenarios arises from speech inactivity, which can result in unreliable DOA estimates [5]. A general approach to deal with estimation errors due to speech inactivity in both model-based and DNN-based systems is to utilize a voice activity detector (VAD) [19] in parallel or cascaded with a DOA estimation system [5], [20], [21], [22]. It should be realized that a separate VAD nonetheless usually requires manual and time-consuming parameter tuning which may entail readjustments when the system is used in different acoustic conditions. Moreover, a separate VAD itself can introduce errors that can restrict the overall performance of the system. In [20], a VAD was integrated into a CNN-based DOA estimation system for hearing aid applications, ensuring that the system avoids DOA estimation during noise-only frames. We will adopt a similar VAD-informed approach in the baseline systems considered in this paper. To address the speech inactivity problem in single-talker binaural DOA estimation, we will consider an alternative approach in this study. We treat it as a DOA estimation task without the need for a separate VAD, which we refer to as speech-aware DOA estimation.

To mitigate estimation errors caused by speech inactivity, classification-based systems commonly employ temporal averaging of the posterior probability map in the output over a relatively long duration [8], [9], [10], [13], [23]. Although this approach helps to smooth out unreliable estimates and improve the overall accuracy, it can compromise the reliability of the DOA estimation system when a new speech source emerges or becomes inactive. It also prevents the system from quickly detecting a change in the trajectory of a moving sound source.

A limited number of systems detect periods of silence within the output of a neural network [24], [25]. However, these approaches, primarily utilized in robot audition scenarios, have not been evaluated against the conventional classification-based approach, leaving their benefits unclear. Furthermore, some of these approaches have only been evaluated under unrealistic background noise conditions [24], while others are tailored for specific source distance and heights and have shown limited performance when tested in conditions that were not included in the training data [25].

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 information is used to segregate the grouped components [1]. It is also known that about 75% of speech in spoken English is voiced and periodic [26]. This motivates the usage of an auditory-inspired periodicity feature in combination with spatial features as input features of a neural network for DOA estimation of a single speech source.

In [27], an auditory-inspired feature called periodicity degree (PD) was proposed for fundamental period detection and estimation and was shown to be useful for VAD in low-SNR conditions. In [28], we proposed a classification-based speech-aware binaural DOA estimation system based on CNNs, which does not require a separate VAD. The proposed speech-aware system was compared to a baseline system that used a conventional classification-based approach. This study showed the benefit of using broadband PD features in combination with GCC-PHAT features as input features of the CNN for speech-aware binaural DOA estimation in static source scenarios.

In this paper, we extend our earlier study [28] by incorporating novel narrowband feature combinations. Our objective is to investigate the advantages of employing PD features in both narrowband and broadband feature combinations for speech-aware binaural DOA estimation across different static and dynamic source scenarios. We propose the novel narrowband feature combinations as follows: First, we introduce a formulation of the PD that incorporates an auditory pre-processing with an adjustable frequency resolution. This formulation generates a subband-averaged representation of the PD, allowing us to take advantage of the frequency selectivity of the human auditory system. Second, we propose to use narrowband cross-power spectrum (CPS) features (as spatial features) in combination with the subband-averaged PD feature as input features for the CNN. For the CPS feature, we consider either using the real and imaginary or the magnitude and phase components of the CPS. In summary, this study aims to investigate the benefits of PD features in the context of novel narrowband feature combinations, as established for the broadband feature combination in [28].

We conduct evaluations to compare the performance of the proposed narrowband systems with narrowband baseline systems consisting of a CNN utilizing only the CPS feature, cascaded with a state-of-the-art pitch-based VAD [19]. Additionally, We evaluate the performance of speech-aware and baseline systems that use broadband features as input features. All systems have been evaluated for static-source scenarios in reverberant environments with matched and unmatched background noise conditions. Furthermore, experiments were conducted for dynamic scenarios with a single moving speech source at different velocities for different signal-to-noise ratio (SNR) conditions. Our experimental results demonstrate the advantage of using the auditory-inspired PD feature in combination with any type of spatial feature (including the GCC-PHAT, real and imaginary parts, or magnitude and phase components of the CPS) for binaural DOA estimation.

The remainder of this paper is organized as follows. In Section II, the single-talker DOA estimation problem is formulated as a classification problem and different approaches are discussed. In Section III, we introduce the input features employed in this study. Section IV provides a comprehensive description of the proposed and baseline systems. 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 V. The proposed and baseline systems are evaluated, and the results are discussed in Section VI. Section VII summarizes the results and presents the conclusion.

II. DOA Estimation as a Classification Problem

In this work, we consider the problem of single-talker DOA estimation in the azimuthal plane 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

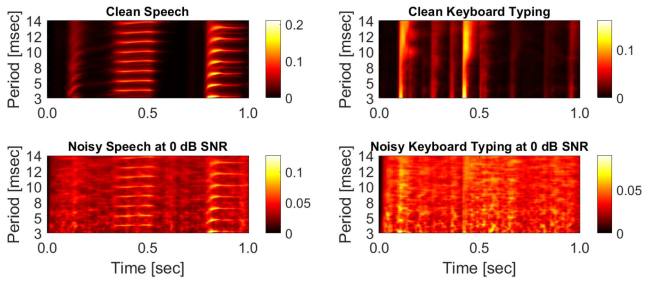


Fig. 1. Exemplary visualization of broadband PD computed for $N = 180$ fundamental period candidates and $L = 199$ consecutive time frames for clean and noisy (a) female speech and (b) keyboard typing signals in an anechoic environment with simulated diffuse noise at 0 dB SNR condition.

of a (possibly moving) sound source at DOA θ in the azimuthal plane and background noise. The m-th microphone signal in the time domain at time t is given by

$$y_m(t) = x_m(t) + v_m(t), \tag{1}$$

where x_m and v_m denote the desired speech and noise signal components in the m-th microphone signal, respectively, which are assumed to be uncorrelated. 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

$$Y_m(n,k) = X_m(n,k) + V_m(n,k). \tag{2}$$

By dividing the azimuth range into a set of C discrete DOAs $\{\theta_1, \ldots, \theta_C\}$, DOA estimation can be considered as a classification problem, where the DOA of a sound source should be assigned to one of the DOA classes. In this work, we consider $C = 72$ classes for the full 360° azimuth range, corresponding to a DOA map with 5° resolution. In the next subsections, two different classification-based approaches for binaural DOA estimation will be discussed.

A. Conventional DOA Estimation

Conventionally, DOA estimation is formulated as a C-class classification task, where each output class corresponds to a DOA [10], [13]. During training, each training example belongs to only one output class that has been labeled using oracle DOA information. During testing, the neural network predicts a posterior probability map in the output. Under the single-source assumption, the DOA is then estimated by finding the DOA class with the highest posterior probability. To deal with erroneous DOA estimates (e.g., during speech pauses), a VAD can be cascaded to this system [20], [21], where a DOA is only estimated from the probability map if the VAD detects the signal as speech. In this work, we adopt the VAD-informed DOA estimation approach to design the baseline systems depicted in Fig. 3.

B. Speech-Aware DOA Estimation

In contrast to the VAD-informed classification-based approach, in [28] we proposed a classification-based approach referred to as speech-aware DOA estimation, which can estimate the DOA of a single talker, without needing a separate VAD. This problem is formulated as a $C + 1$-class classification task, where the first C classes represent the DOA classes and the last class represents the non-speech activity, regarded as the detection class. During training, via a one-hot encoding scheme, if a training example belongs to a speech source from a given direction, the DOA class corresponding to that direction is labeled by one, whereas all other classes (including the detection class) are labeled by zero. On the other hand, if a training example belongs to a non-speech source, regardless of its direction, all DOA classes are labeled by zero, whereas the detection class is labeled by one. During testing, if the class with the highest posterior probability is a DOA class, the direction corresponding to that class indicates the sound source DOA. Otherwise, no reliable DOA could be estimated. In this work, we adopt the speech-aware DOA estimation approach in our proposed systems depicted in Fig. 4.

III. Input Features

This section provides an overview of the spatial and periodicity features utilized as input features for various classification-based DOA estimation methods in this study. In Section III-A, we present the broadband GCC-PHAT feature, which was also employed in [28], in addition to the newly introduced narrowband CPS features, as spatial features. In Section III-B, we introduce the novel subband-averaged representation of the PD, along with the broadband PD used in [28]. Furthermore, we present the rationale for the incorporation of PD through exemplary visualizations that demonstrate different PD representations.

A. Spatial Features

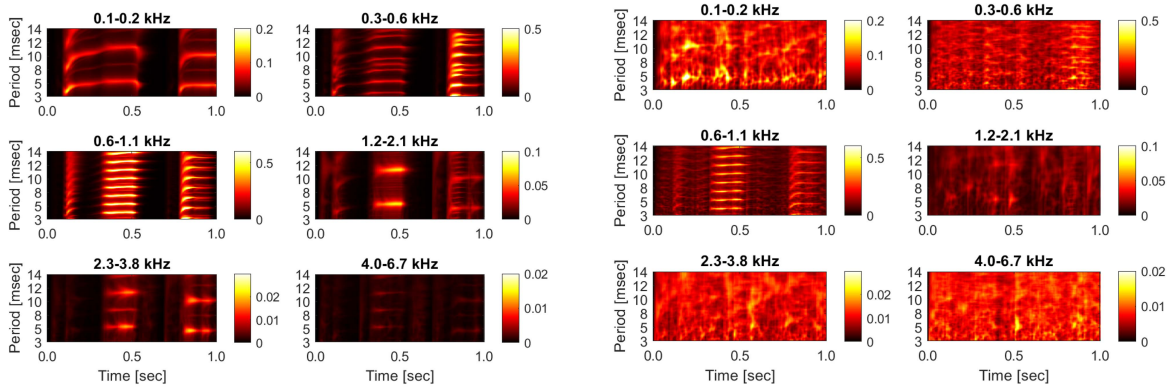
The GCC-PHAT has been successfully used as a feature for several data-driven DOA estimation methods [29], [30], [31], [32]. In this work, the broadband GCC-PHAT between the i-th pair of microphones is defined as the inverse Fourier transform of the phase of the instantaneous narrowband CPS which is given by

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

where microphones r and q constitute the i-th microphone pair and $(\cdot)^*$ denotes complex conjugate. We note that there are $M(M-1)/2$ microphone pairs, i.e., $i \in [1, M(M-1)/2]$. The GCC-PHAT for the i-th microphone pair at time frame n is computed as

$$\tau_i(n,d) = \mathcal{IFFT}\left(\frac{G_i(n,k)}{|G_i(n,k)|}\right), \tag{4}$$

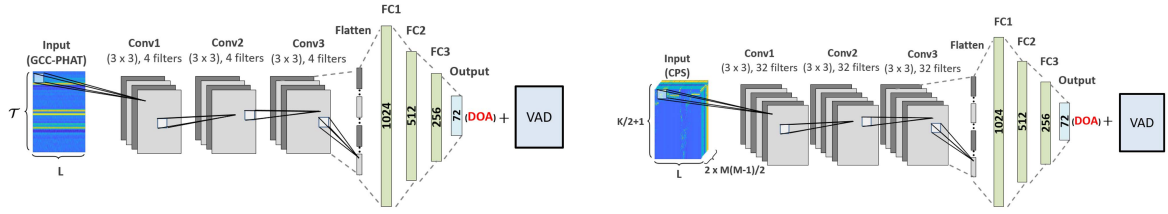
where $|\cdot|$ denotes absolute value, and d represents the index of the time delay. In order to resolve fractional signal delays occurring for microphone pairs with a small distance (e.g., microphones on a hearing aid), it is useful to interpolate the GCC-PHAT function by using an oversampled inverse Fourier transform [2]. With an upsampling factor of κ, the relevant discrete time delays lie in the range $[-\kappa\tau_i^{\max}, \kappa\tau_i^{\max}]$, where $\tau_i^{\max}$ denotes the maximum delay in samples, considered for



(a) Clean subband-averaged PD features

(b) Noisy subband-averaged PD features at 0 dB SNR

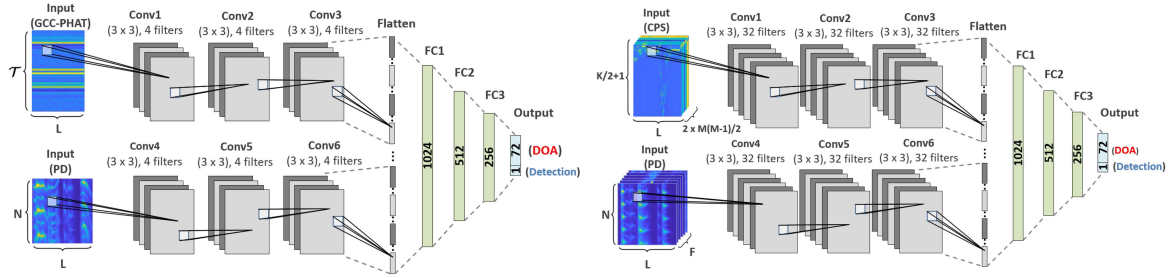
Fig. 2. Exemplary visualization of the subband-averaged PD feature shown for (a) clean female speech, and (b) noisy female speech at 0 dB SNR, computed for $N = 180$ fundamental period candidates, $L = 199$ consecutive time frames, and $F = 6$ frequency bands in an anechoic environment with simulated diffuse noise as background noise. The frequency range corresponding to each frequency band is specified at the top of the images.



(a) With only broadband spatial feature (GCC-PHAT)

(b) With only narrowband spatial feature (CPS)

Fig. 3. Baseline VAD-informed DOA estimation systems using only spatial features: (a) Broadband spatial feature (GCC-PHAT), and (b) narrowband spatial feature (CPS).



(a) With broadband feature combination (GCC-PHAT and PD)

(b) With narrowband feature combination (CPS and subband-averaged PD)

Fig. 4. Proposed systems with (a) broadband feature combination (GCC-PHAT and PD), and (b) narrowband feature combination (CPS and subband-averaged PD). The architecture of the convolutional branch with a spatial input feature (top branch) in each proposed system is identical to the architecture of the convolutional branch in a baseline system using the same spatial feature depicted in Fig. 3.

the i-th microphone pair. The GCC-PHAT vector of the i-th microphone pair is defined as

$$\boldsymbol{\tau}_i(n) = [\tau_i(n,1), \ldots, \tau_i(n,\mathcal{T}_i)]^{\mathrm{T}}, \tag{5}$$

where $(\cdot)^{\mathrm{T}}$ denotes the vector transpose. The first and last elements in (5) correspond to $-\kappa\tau_i^{\max}$ and $+\kappa\tau_i^{\max}$, respectively. Therefore, the length of the GCC-PHAT vector is obtained by $\mathcal{T}_i = 2\kappa\tau_i^{\max} + 1$. By concatenating the GCC-PHAT vectors $\boldsymbol{\tau}_i(n)$ for all possible microphone pairs, and considering L consecutive time frames (including the current frame n and the previous $L-1$ frames), we obtain the two-dimensional (2D) GCC-PHAT input feature with dimensions $\mathcal{T} \times L$, where $\mathcal{T} = \sum_{i=1}^{M(M-1)/2} \mathcal{T}_i$. This 2D feature will be used as a spatial input feature for broadband systems in Section IV.

As can be seen in (4), the PHAT weighting eliminates the effect of spectral magnitude, such that phases contribute equally for all frequencies. Hence, as an alternative to the broadband GCC-PHAT, in this work, we will also consider the narrowband CPS [7], encoding both spectral magnitudes and phase differences, as an input feature.

As the CPS input feature, we consider either the magnitude and phase (denoted as MagPhase) or the real and imaginary parts (denoted as ReIm) of the complex-valued CPS $G_i(n,k)$ for all $M(M-1)/2$ unique microphone pairs, for $K/2+1$

frequencies (up to the Nyquist frequency, i.e., $k = 0, 1, \ldots, K/2$), and for L consecutive time frames. This means that the shape of the CPS input feature is equal to $(K/2+1) \times L \times 2M(M-1)/2$. We note here that the first, second, and third dimensions represent the height, width, and depth of the input feature, respectively, where the depth corresponds to the number of input channels. For the CPS input feature, $2M(M-1)/2$ input channels are constructed by stacking either the MagPhase or the ReIm for all microphone pairs. The CPS features will be used as spatial input features for narrowband systems in Section IV.

B. Periodicity Features

Periodicity is an important cue to segregate and localize different talkers [33], [34]. Periodicity features are often estimated through an auditory pre-processing step followed by a feature extraction step [34] where they are estimated independently for each pre-processed subband signal. In [27] a periodicity feature called PD was introduced, which captures the salience of the periodic components in the input signal. In this work, we propose to use a subband-averaged representation of PD features, estimated for a set of N fundamental period candidates. Similar to spatial features, we will consider PD features from L consecutive time frames as input PD features. This section focuses on PD computation for time samples (t) spanning a block of L consecutive time frames.

To compute PD features, we use one of the M microphones, referred to as the reference microphone in this paper. It is important to note that the choice of the reference microphone is arbitrary, and determining the optimal microphone for PD estimation is not within the scope of this study. In the following, we present signal processing steps to compute the subband-averaged PD. In the pre-processing step, the reference microphone signal in the hearing aid setup is first decomposed into a set of subband signals using a complex-valued gammatone filter bank GTFB [27]. The real part of each subband signal is then passed through half-wave rectification, yielding the half-wave rectified signal $y(t, f)$ in the f-th gammatone subband. Although the PD is usually computed for each subband [34], in this paper we introduce a subband averaging step, enabling us to estimate the PD for frequency regions with adjustable bandwidths. The subband-averaged signal is computed as

$$y_{avg}(t, \bar{f}) = \frac{1}{\sigma S} \sum_{f=(\bar{f}-1)S+1}^{\bar{f}S} y(t, f), \tag{6}$$

where $\bar{f}$ denotes the (averaged) frequency band index and S denotes the number of averaged subbands. The normalization parameter σ represents the standard deviation computed over all subbands and times (time samples of L consecutive STFT frames) for the signal $y(t, f)$. This subband averaging step results in F frequency bands. Subsequently, a fifth-order low-pass filter with 770 Hz cutoff frequency and a second-order high-pass filter with 40 Hz cutoff frequency are applied to $y_{avg}(t, \bar{f})$, resulting in bandpass-filtered signal envelopes $y_{env}(t, \bar{f})$.

In the feature extraction step, a set of N parallel infinite impulse response (IIR) comb filters designed for a given set of N fundamental period candidates $p_j, j = 1, \ldots, N$, filter the signal envelopes as

$$s(j, t, \bar{f}) = (1-\alpha) y_{env}(t, \bar{f}) + \alpha s(j, t-p_j, \bar{f}), \tag{7}$$

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

$$PD(j, t, \bar{f}) = (1-\beta_j)|s(j, t, \bar{f})| + \beta_j PD(j, t-1, \bar{f}), \tag{8}$$

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

The PD features in (8) have the same temporal resolution as the time-domain signal. Since we aim at processing of the PD and CPS features by the neural network, it is desirable to represent both features at the same time resolution, which is the frame resolution of the STFT framework. Hence, the high-resolution PD features are temporally averaged as

$$\overline{PD}(j, n, \bar{f}) = \frac{1}{K} \sum_{t=(n-1)D+1}^{(n-1)D+K} PD(j, t, \bar{f}). \tag{9}$$

As the subband-averaged input PD feature, we consider PD features in (9) for all N fundamental period candidates, for L consecutive time frames, and for all F frequency bands. This means that the shape of the subband-averaged input PD is equal to $N \times L \times F$, where the first, second, and third dimensions represent the height, width, and depth of the input feature, respectively. This three-dimensional feature will be used as the periodicity input feature of the proposed narrowband speech-aware DOA estimation systems in Section IV-B.

As mentioned earlier, in this work, we also use the broadband PD feature from [28] for the broadband speech-aware DOA estimation system in Section IV-B. To obtain the broadband PD feature, signals of all gammatone subbands are averaged in (6), i.e., $F = 1$. The resulting broadband signal in (6) is utilized for PD feature extraction. Consequently, the broadband input PD has a shape of $N \times L$.

Fig. 1 depicts an exemplary representation of the broadband PD feature computed for $N = 180$ fundamental period candidates over a 1-second duration ($L = 199$) of both clean and noisy speech, as well as non-speech (keyboard typing) signals. While for the clean and noisy speech signals, the fundamental period variation, its multiple harmonics and their temporal continuity are identifiable as a 2D structure over time, no such structure exists for the keyboard signal, and in general for non-speech signals. Although speech signals are not perfectly harmonic, we hypothesize that utilizing the fundamental period information encoded in the harmonic structure of the PD feature could facilitate a neural network's ability to differentiate between signal portions that are predominantly speech (and periodic) versus non-speech, particularly when trained with a combination of speech and non-speech signals.

The 2D structure of the pitch modulations and harmonics can also be identified in the subband-averaged PD features. Fig. 2 illustrates the subband-averaged PD features computed

for $N = 180$ fundamental period candidates across $F = 6$ frequency bands ($S = 10$), using a 1-second duration ($L = 199$) of clean and noisy speech signals. As can be seen in Fig. 2(a), the harmonic structure of the pitch information is captured in most frequency bands for the clean signal, particularly in frequency bands with a high degree of periodicity, e.g., in $0.3 - 0.6$ kHz and $0.6 - 1.1$ kHz frequency bands (with a maximum PD value of 0.5). However, as can be seen in Fig. 2(b), for the noisy signal this information is substantially masked by the noise except for the frequency regions mainly in the $0.6 - 1.1$ kHz frequency band. By using subband-averaged PD features as input features, the neural network is expected to be able to select the most robust and salient periodicity information, particularly for those frequency bands in which speech signals have more energy, and hence are less susceptible to noise.

The primary rationale for employing PD features in conjunction with spatial features is to leverage the salient periodicity features as a footprint of speech signals in a noisy mixture [35], [36]. This approach enables the neural network to detect voiced speech portions of a signal while simultaneously mapping the CPS features of these portions to the talker's DOA.

IV. CNN-Based DOA Estimation Systems

This section outlines the CNN-based DOA estimation systems. The baseline systems are discussed in Section IV-A, which adopt a VAD-informed DOA estimation approach, utilizing only spatial features. The proposed systems are presented in Section IV-B, which adopt a speech-aware DOA estimation approach, utilizing a combination of spatial and PD features as input features. Finally, we discuss the computational complexity of the proposed and baseline systems in Section IV-C.

A. Baseline VAD-Informed Systems

Neural network architectures based on CNNs have been widely and successfully used for DOA estimation and sound source localization [7]. Fig. 3 depicts the baseline systems consisting of a CNN using only spatial features (cf. Section III-A) as input, cascaded with a pitch-based binary VAD [19]. We consider three baseline systems:

- Broadband (Fig. 3(a)) using GCC-PHAT features,
- Narrowband-ReIm (Fig. 3(b)) using the real and imaginary parts of the CPS,
- Narrowband-MagPhase (Fig. 3(b)) using the magnitude and phase of the CPS.

The CNN architecture in all considered baseline systems starts with a cascade of three convolutional blocks, with each block (*Conv1* to *Conv3*) comprising a sequence of 2D convolutional, batch normalization, rectified linear unit (ReLU) activation, and 2D max-pooling layer. The outputs of the last pooling layer in *Conv3* are concatenated and then used as an input for a cascade of three fully-connected blocks (*FC1* to *FC3*), each representing a fully-connected dense layer followed by batch normalization, ReLU activation, and dropout layers. In the output layer, a softmax activation function predicts the posterior probability map for the C DOA classes.

In addition to the batch normalization layer implemented in the convolutional and fully-connected blocks of the CNNs, we applied a normalization scheme only in the input layer and directly on the input features before the first convolutional block to improve the performance of the CNNs. We applied layer normalization [37] on the GCC-PHAT features. Concerning the CPS input features, layer normalization was applied on all $2M(M-1)/2$ channels of the real and imaginary parts of the CPS to preserve phase information encoded by these features. As for the magnitude and phase parts of the CPS, group normalization [38] was applied to the two groups of magnitude and phase features each including $M(M-1)/2$ channels. This means that within each group, features are normalized separately. The reason for this is that the magnitude and phase have different statistical properties, and hence, joint normalization of the magnitude and phase may not be optimal. We note that all layer normalizations and group normalizations have been implemented without an affine transformation.

Each training example consists of a block of L consecutive time frames, i.e., we employ block-level labeling and the CNN generates its output for each block. We adopt a one-hot encoding scheme during the training, i.e., each training example belongs to only one output class that has been labeled using oracle DOA information. It is important to note that we assume a constant DOA when assigning a ground truth DOA label to a training example of a speech signal, which implies that the DOA remains consistent throughout the block of L consecutive time frames. During the testing phase, the CNN in the baseline system generates a posterior probability map $\boldsymbol{P} = [P_1, \ldots, P_C]$, which represents the likelihood of the sound source being located at each of the C possible DOA classes. It should be noted that we obtain consecutive input features with an overlap of $L - 1$ frames for all systems. As input features consist of L consecutive time frames, this approach results in the generation of a new posterior probability map for each new frame.

To mitigate the effects of erroneous DOA estimates that can arise during periods of speech pauses, the system is augmented with a cascaded VAD. This configuration enables the DOA estimation process to be conditioned on the presence of speech, as determined by the VAD. Specifically, the DOA is estimated solely from the probability map when the VAD indicates the presence of speech, which is expected to lead to more robust and accurate DOA estimates. We note that the VAD decision is made using the same reference microphone signal that is used to compute PD features. As a common approach, the sound source DOA can be estimated as θ_I for the DOA class I with the highest posterior probability, i.e.

$$I = \arg\max_i P_i. \tag{10}$$

In this work, to obtain continuous DOA estimates from discrete DOA classes, we estimate the sound source DOA by employing parabolic interpolation [39] on three DOA classes centered around θ_I, i.e., θ_{I-1}, θ_I and θ_{I+1}. As a result, this approach allows for a more precise estimation of the DOA with a higher spatial resolution.

B. Proposed Speech-Aware Systems

Instead of using a separate pitch-based VAD in combination with a CNN, we adopt the speech-aware approach described in Section II-B. Fig. 4 depicts the proposed speech-aware DOA estimation systems, which use PD features (cf. Section III-B) in combination with spatial features (cf. Section III-A) as input features of the CNN. In this work, we consider three speech-aware systems:

- Broadband (Fig. 4(a)) using GCC-PHAT features and the broadband PD features as input features,
- Narrowband-ReIm (Fig. 4(b)) using the real and imaginary parts of the CPS and the subband-averaged PD features as input features,
- Narrowband-MagPhase (Fig. 4(b)) using the magnitude and phase of the CPS and the subband-averaged PD features as input features.

Each proposed system in Fig. 4 consists of two parallel independent branches of three cascaded convolutional blocks. The top branch receives the spatial features as input features, whereas the bottom branch receives the PD features as input features. The outputs of both branches are then concatenated, which serves as a hybrid intermediate feature vector used by a cascade of three fully-connected blocks. Similar to the baseline systems described in Section IV-A, for the proposed systems each convolutional block consists of a 2D convolutional, batch normalization, ReLU activation, and 2D max-pooling layer. Each fully-connected block is comprised of a fully-connected dense layer followed by batch normalization, ReLU activation, and dropout layers. In the output layer, a softmax activation function predicts the posterior probability map for the $C+1$ classes. We applied layer normalization without an affine transformation on the PD input features of each proposed system. As for the spatial input features, we used the same normalization scheme that was applied to spatial features of the baseline systems (cf. Section IV-A). Please note that for each proposed system, the PD and spatial features have been normalized separately.

We note here that for each spatial feature, GCC-PHAT or CPS, the architecture of the convolutional branch with spatial input features in the proposed system and the architecture of the convolutional path in the baseline system are the same. This can be seen, for instance, by comparing Fig. 3(a) with the top branch in Fig. 4(a), and also by comparing Fig. 3(b) with the top branch in Fig. 4(b). As a result, the convolutional path in the baseline system and the top branch of a proposed system that use the same spatial features will learn the same number of parameters and filters, and ultimately contribute to the input of the fully-connected path by the same amount of (intermediate) features. Consequently, we can consider the contribution of the spatial features in the fully connected path of the proposed system to be equivalent to that of the baseline system. This allows us to analyze the benefits of using PD features and compare the two systems using the same spatial features.

We expect that by training the proposed systems with speech and non-speech signals, the network is able to capture the harmonic structure of the signal encoded in the PD features over consecutive frames. This allows the proposed system to discern between speech and non-speech portions, while simultaneously mapping the spatial features to a sound source DOA when speech portions in the signal are detected.

The proposed systems were trained using oracle DOA and detection labels for speech and non-speech signals. All $C+1$ output classes were labeled as a single label, meaning each training example belonged to only one output class. This was achieved as follows: During the training phase, training examples of input features are provided for both speech and non-speech sources. For a given training example of a speech source, the direction of the speech source is associated with a particular DOA class, which is labeled by one. The remaining DOA classes, along with the detection class, are labeled by zero. In contrast, for a training example of a non-speech sound source, regardless of its direction, all DOA classes are labeled by zero, except for the detection class which is labeled by one.

During the testing phase, the proposed system generates a posterior probability map given by $\boldsymbol{P} = [P_1, \ldots, P_C, P_{C+1}]$ for a given number of directions C. We note here again that input features (each including L time frames) are consecutively obtained with an overlap of $L-1$ frames, i.e., a new posterior probability map is generated for every new frame. The process of speech-aware DOA estimation can be formulated by first introducing two hypotheses

$$\mathcal{H}_s: \quad \text{speech DOA detected}, \tag{11}$$

$$\mathcal{H}_{ns}: \quad \text{no speech DOA detected}, \tag{12}$$

and then defining the decision rule as

$$\begin{aligned} &\text{decide} \quad \mathcal{H}_{ns} \quad \text{if} \quad \arg\max_i P_i = C+1 \\ &\text{decide} \quad \mathcal{H}_s \quad \text{otherwise}. \end{aligned} \tag{13}$$

For the DOA estimation, we first consider the direction θ_I corresponding to the DOA class I with the highest posterior probability when speech DOA is detected, i.e.

$$I = \arg\max_i P_i | \mathcal{H}_s. \tag{14}$$

Then, we estimate the sound source DOA by employing parabolic interpolation [39] on three DOA classes centered around θ_I, i.e., θ_{I-1}, θ_I and θ_{I+1}. The process of speech-aware DOA estimation can be described as follows: the output class with the highest probability in the predicted probability map is selected. If the highest probability corresponds to the last class, which represents the detection class, it indicates that no reliable DOA estimation is possible. On the other hand, if the highest probability corresponds to a DOA class, the sound source DOA is estimated as the parabolic approximation of the direction associated with that DOA class.

It should be noted that the proposed speech-aware DOA estimation systems integrate both DOA estimation and VAD into a unified framework, with speech detection regarded as an implicit result of the proposed systems. Whereas for the VAD-informed systems, the DOA estimation is conditioned on the VAD decision (i.e., an *explicit* speech detection), for the speech-aware systems, the DOA estimation is merely conditioned on the joint probability distribution in the CNN output,

TABLE I
NUMBER OF TRAINABLE PARAMETERS AND MULTIPLY-ACCUMULATE OPERATIONS (MACS) OF THE BASELINE AND PROPOSED SYSTEMS

CNN	Baseline broadband	Proposed broadband	Baseline narrowband	Proposed narrowband
Parameters/M	1.37	1.55	1.36	2.82
MACs/M	2.54	3.03	11.61	29.79

including the detection class (i.e., an *implicit* speech detection which we refer to as speech DOA detection in (11)). Therefore, in addition to the performance evaluation of all systems for DOA estimation, our study will assess their speech detection capabilities. This could offer a more comprehensive insight into these systems.

C. Computational Complexity

Table I shows the number of trainable parameters and multiply-accumulate operations (MACs), both in millions for the baseline and proposed DOA estimation 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. By analyzing Table I, we can observe that the size of the two CNNs employed for the baseline systems using the broadband (1.37 M) and narrowband (1.36 M) spatial features are comparable. It is important to note that the number of convolutional filters in each baseline system (cf. Fig. 3) was chosen to ensure that both systems have a comparable number of parameters. However, the proposed narrowband system exhibits a higher number of trainable parameters (2.82 M) in comparison to the broadband counterpart (1.55 M). Table I also shows that while the proposed broadband and narrowband systems exhibit a larger number of trainable parameters compared to their respective baseline counterparts, the difference in the number of trainable parameters is especially noticeable for the narrowband systems. To the best of our knowledge, it is not possible to directly implement the considered systems in current hearing devices. This may be possible after model size optimization, model quantization and pruning, which is however not the main topic of this study.

V. EXPERIMENTAL SETUP

In this section, we conduct experiments to assess the performance of the speech-aware systems proposed in Section IV-B in comparison to the baseline systems described in Section IV-A. Furthermore, we provide details of the datasets utilized in this study in Section V-A, and describe the procedures for generating training and evaluation data in Sections V-B and V-C, respectively. Additionally, we present implementation details of the input features and the VAD in Section V-D, and describe the training procedure and hyperparameters of the CNNs used in this study in Section V-E. Evaluation metrics employed to assess the performance of all systems are described in Section V-F.

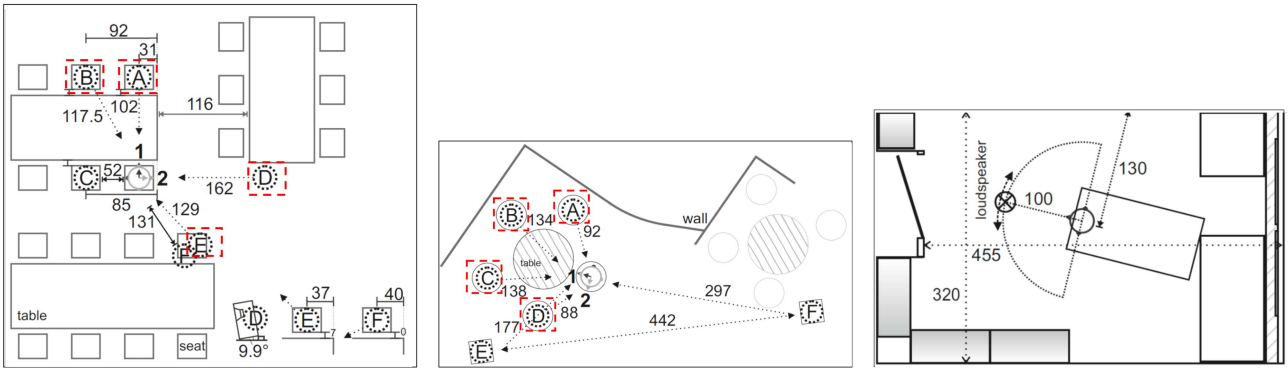
A. Datasets

Signals from speech and non-speech datasets were used as sound source signals to generate the training and validation data required during the training of all systems. In particular, speech signals of 462 and 168 speakers from the TIMIT dataset [40] (including both male and female speakers) were used for training and validation, respectively. In addition, three categories (natural soundscapes and water sounds, interior and domestic sounds, exterior and urban noises) of the ESC50 dataset [41] were used as non-speech signals, where we used 960 and 240 distinct sound files for training and validation, respectively. For evaluation, only speech signals from the validation TIMIT dataset were used as source signals.

We used a database of multichannel binaural room impulse responses (BRIRs) [42] to generate data for training and evaluation. The considered binaural hearing aid setup consists 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 [42] contains BRIRs measured in anechoic conditions for different source-to-head distances, and for $C = 72$ directions in the azimuthal plane, i.e., with a resolution of 5°. This dataset also contains BRIRs in three reverberant environments (cafeteria with $T_{60} \approx 1.3$ s, courtyard with $T_{60} \approx 0.9$ s, office with $T_{60} \approx 0.3$ s). We generated the noisy binaural microphone signals by convolving the source signals with BRIRs and mixing the resulting clean binaural microphone signals with background noise. All systems were trained in noisy anechoic conditions and evaluated in noisy reverberant environments.

B. Training Data

For training, the clean binaural microphone signals were generated by convolving both speech and non-speech source signals with anechoic BRIRs for each of the 72 directions with a source-to-head distance of 3 m. The noisy binaural microphone signals were generated by mixing the clean binaural microphone signals with simulated binaural diffuse noise at SNRs ranging from -5 dB to $+20$ dB in 5 dB steps. The noise at the microphones was generated by convolving uncorrelated speech-shaped noise from the ICRA noise database [43] with anechoic BRIRs, and summing all resulting binaural signals from 72 directions. Training examples were constructed for both speech and non-speech signals for all 72 directions at six different SNRs. It is important to note that in a data pre-processing step, a simple oracle broadband energy-based VAD was employed to identify segments containing enough speech content. This step ensures that for training examples associated with a speech source, only those containing meaningful speech content contribute to the loss function. Each training example consists of a block of $L = 20$ consecutive time frames (corresponding to 105 ms). In total, we obtained 5.9 million examples (about 172 hours) as *training set* and 2.4 million examples (about 70 hours) as *validation set*. A summary of the training data is presented in Table II.



(a) Cafeteria (static scenario) (b) Courtyard (static scenario) (c) Office (dynamic scenario)

Fig. 5. Evaluation setups for static scenarios (a and b), and dynamic scenarios (c), adapted from [42]. In the cafeteria, static source positions *A*, *B*, *D*, *E* were considered, while in the courtyard, static source positions *A*, *B*, *C*, and *D* were considered. In both environments, the cafeteria and courtyard, the head position is indicated by dashed arrows extending from the source positions to the head. The two head orientations are denoted by the numbers 1 and 2 positioned in proximity to the head. For the dynamic scenarios in the office, the source position traveled from left to right with respect to the look direction of the head, either between $-45°$ to $+45°$ ($6\,°/s$ angular velocity) or between $-60°$ to $+60°$ ($8\,°/s$ angular velocity). The head position in the office environment is depicted in the middle of the room. All distances are specified in centimeters.

TABLE II
SUMMARY OF THE TRAINING DATA

Source signals	Speech (TIMIT) and non-speech (ESC50)
Environment	Anechoic [42]
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

C. Evaluation Data

The performance of the baseline and proposed systems was evaluated for static and dynamic source scenarios in reverberant environments. As already mentioned, only speech signals from the validation TIMIT dataset were used as source signals. It should be noted that the source and background noise signals, acoustic conditions and source positions used during evaluations were different from those used during training and validation. A summary of the evaluation setup and data generation is presented in Table III.

1) Static Source Scenario: For the static source scenario, we considered two real environments (cafeteria and courtyard) with a reverberation time of approximately 1300 ms and 900 ms, respectively. The clean binaural microphone signals were generated by convolving the speech source signals with reverberant BRIRs [42]. The room configurations of both environments are depicted in Fig. 5(a) and (b). In each environment, we considered four source positions (specified with dashed boxes), with two head orientations for each source position. All systems were evaluated at SNRs ranging from -5 dB to $+10$ dB either with matched or unmatched background noise. The same binaural diffuse noise as that used during training was utilized for the matched noise condition, whereas recorded cafeteria babble noise and courtyard ambient noise [42] were used for the unmatched noise condition. A total number of 150 speech segments randomly chosen from 30 unique male and female speakers (each with a length of 1 s) were selected from the validation TIMIT dataset.

2) Dynamic Source Scenario: In [42], BRIRs of a reverberant office environment with a reverberation time of approximately 300 ms (specified in [42] as *Office I*) are provided, which cover the frontal azimuth range from $-90°$ to $+90°$ with a $5°$ resolution. Since only the BRIRs measured in the office environment allow to simulate moving sources, simulations for the dynamic source scenario were only performed for the office environment. To simulate a moving source, a time-aligned interpolation method [44] with shape-preserving piecewise cubic interpolation was used to interpolate the original BRIRs from a $5°$ resolution to a $0.5°$ resolution. A total number of 10 speech segments were randomly chosen from 30 unique male and female speakers (each with a length of 15 s) from the validation TIMIT corpus. The clean binaural microphone signals were simulated for two source velocities ($6\,°/$s and $8\,°/$s angular velocity) by partial convolution of the interpolated BRIRs with the clean speech signal using a frame length of 10 ms and 50% overlap. The office room configuration and the source movement trajectory are depicted in Fig. 5(c). Simulated binaural diffuse noise was used to generate noisy binaural microphone signals at SNRs ranging from -5 dB to $+10$ dB.

D. Implementation Details

All signals were sampled at 16 kHz. To compute the GCC-PHAT and CPS features, the microphone signals were transformed to the STFT domain using a Hann window of length $K = 160$ (corresponding to 10 ms), and a hop size of length $D = 80$ (corresponding to 5 ms), resulting in 81 STFT frequency bins. To compute GCC-PHAT features, we used an upsampling factor of $\kappa = 4$. In the case of a pair of microphones located on the same side of the head (left or right), the corresponding maximum delay $\tau_i^{\max}$ is considered as 2, which translates to a time delay of 125 μs and $\mathcal{T}_i = 17$. Conversely, for a pair of microphones located on opposite sides, the maximum delay $\tau_i^{\max}$ is considered as 20, corresponding to a time delay of 1.25 ms and $\mathcal{T}_i = 161$. We note that the chosen maximum delays

TABLE III
SUMMARY OF THE EVALUATION DATA

	Static scenario	dynamic scenario
Source signals	Speech (TIMIT)	Speech (TIMIT)
Environment	Cafeteria ($T_{60} \approx 1.3$ s) and courtyard ($T_{60} \approx 0.9$ s) [42]	Office ($T_{60} \approx 0.3$ s) [42]
Background noise	Simulated diffuse noise and recorded noise	Simulated diffuse noise
SNR	-5 dB to $+10$ dB in 5 dB steps	-5 dB to $+10$ dB in 5 dB steps
Source-to-head distance	$1-1.6$ m	1 m
Source positions	4 source positions with 2 head orientations in each environment	Trajectories from $-90°$ to $+90°$

are deliberately set to be greater than the maximum possible delay that can occur for each microphone pair by approximately a factor of two. In the considered binaural hearing aid setup, there are a total of four microphone pairs on opposite sides and two microphone pairs on the same side. As a result, GCC-PHAT feature vectors of size $\mathcal{T} = 678$ are obtained. For feature extraction, a block of $L = 20$ consecutive time frames is employed, leading to a GCC-PHAT input feature of size 678×20. For each pair of CPS features (real and imaginary parts, magnitude and phase components), the size of the input features is equal to $81 \times 20 \times 12$.

In this paper, we consider the front microphone of the left hearing aid as the reference microphone for the PD feature extraction, and also for the binary VAD decision employed in the baseline systems. To obtain a binary VAD decision on a block of L consecutive time frames, a majority vote rule is applied, where the block is classified as speech if at least 50% of the time frames are detected as such. In the baseline systems, we used the pitch-based binary VAD [19] (rVAD) with its original frame length but adjusted the frame hop size to 5 ms, aligning it with the proposed systems while keeping its spectral resolution unchanged.

PD features were computed using a 4-th order gammatone filter bank (GTFB) implementation [27] with 61 subbands, a group delay of 256, and minimum and maximum center frequencies of 60 Hz and 7200 Hz, respectively. By choosing the maximum and minimum fundamental frequencies as 320 Hz and 70 Hz, respectively, the range of fundamental period candidates for PD feature extraction lies between 3.1 ms and 14.3 ms for $N = 180$ period candidates. To compute the subband-averaged PD features, $F = 6$ frequency bands are obtained by averaging every $S = 10$ subband signals. The comb filter gain was chosen to be $\alpha = 0.7$. The size of the broadband and subband-averaged input PD features is equal to 180×20 and $180 \times 20 \times 6$, respectively.

E. Training and Network Hyperparameters

All systems were implemented using PyTorch [45]. For all CNNs, we used a 2D convolutional filter size of 3×3 with a stride size of 1×1. In each convolutional layer of the CNNs with broadband (GCC-PHAT) and narrowband (CPS) input, 4 and 32 filters were used, respectively. The max-pooling size was 2×2 with strides of the same size. The CNNs were trained using the Adam optimizer [46], a cross-entropy loss function, an initial learning rate of 10^{-5}, a mini-batch size of 128 and a dropout rate of 0.5. We used an early stopping regularization method which stopped the training if no improvement in validation loss was observed for 4 epochs, and a variable learning rate scheduler to halve the learning rate if the validation loss did not improve for 2 epochs.

The maximum epoch number for training all CNNs was set to 100. In each epoch, 1.63 million examples were randomly selected from the training set such that the network did not see the same example twice. Each mini-batch included 128 examples that were randomly chosen from different SNR conditions, DOA classes, and speech and non-speech signals. To calculate the validation loss at the end of each epoch, 200000 examples were randomly selected from the validation set and kept fixed throughout the training. The validation data were not seen by the network during the training.

F. Evaluation Metrics

We evaluated the DOA estimation performance of the proposed and baseline systems in terms of mean absolute error (MAE) and accuracy (Acc.) [8], [9]. A DOA estimate in block l is considered accurate if the absolute error between the estimated DOA $\hat{\theta}_l$ and the oracle DOA θ_l is smaller than $5°$, i.e., the minimum angular resolution of the database in [42]. The MAE (in degrees) and accuracy are defined as

$$\text{MAE} = \frac{1}{\mathcal{L}} \sum_{l=1}^{\mathcal{L}} \left| \hat{\theta}_l - \theta_l \right|, \tag{15}$$

$$\text{Acc} = \frac{\mathcal{L}_{acc}}{\mathcal{L}} \times 100, \tag{16}$$

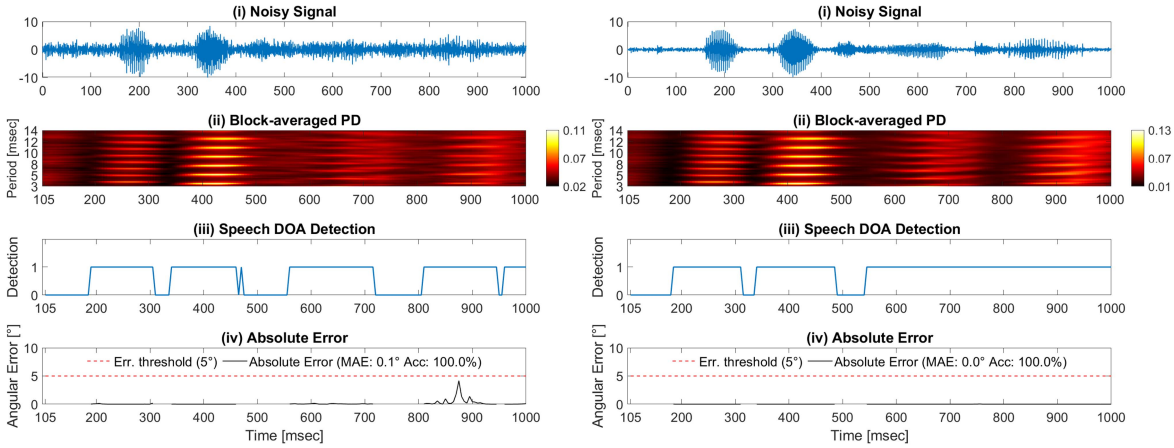
where $\mathcal{L}$ denotes the total number of estimates, i.e., the number of signal blocks with positive speech detections, and $\mathcal{L}_{\text{acc}}$ denotes the total number of accurate estimates.

We evaluated the speech detection performance of the VAD used in the baseline systems and the performance of the speech DOA detection in the proposed system using the precision (P) and recall (R) metrics defined as

$$\text{P} = \frac{TP}{(TP + FP)}, \tag{17}$$

$$\text{R} = \frac{TP}{(TP + FN)}, \tag{18}$$

where for each evaluated system, the number of true positives (TP) represents the total number of signal blocks detected as speech by both the system and the oracle VAD, while the number of false positives (FP) represents the total number of signal blocks detected as speech by the system but detected as non-speech by the oracle VAD. Conversely, the number of



(a) 0 dB SNR condition (b) 10 dB SNR condition

Fig. 6. Exemplary illustration of DOA estimation of the proposed system using the broadband feature combination in (a) SNR = 0 dB and (b) SNR = 10 dB: (i) Noisy reference microphone signal of a source at $-20°$ with simulated diffuse noise, (ii) Estimated block-averaged broadband PD, (iii) Speech DOA detection, (iv) Absolute angular error of the estimated DOAs over time specified by black lines, and the $5°$ error threshold specified by a red dashed line.

false negatives (FN) denotes the total number of signal blocks detected as non-speech by the system but detected as speech by the oracle VAD. While precision indicates the proportion of detected speech blocks that are actually correct, recall represents the proportion of actual speech blocks that are detected by the system. Both metrics range from 0 to 1.

VI. Results and Discussion

In this section, we will present and analyze the performance evaluation results of speech-aware systems employing either broadband or narrowband feature combinations, in comparison to baseline systems. The baseline systems consist of a CNN that uses only spatial features, combined with a pitch-based VAD. We assessed the performance of all systems in various reverberant environments with different background noises for both static and dynamic single-talker scenarios in terms of accuracy and mean absolute error for DOA estimation, as well as precision and recall for speech detection. Section VI-A serves as an exemplary demonstration of the proposed speech-aware DOA estimation using broadband input features. The evaluation results for static source scenarios in both matched and unmatched noise conditions are discussed in Section VI-B. The evaluation results for dynamic source scenarios are presented in Section VI-C. Finally, we discuss the limitations of this study and suggest potential future works in Section VI-D.

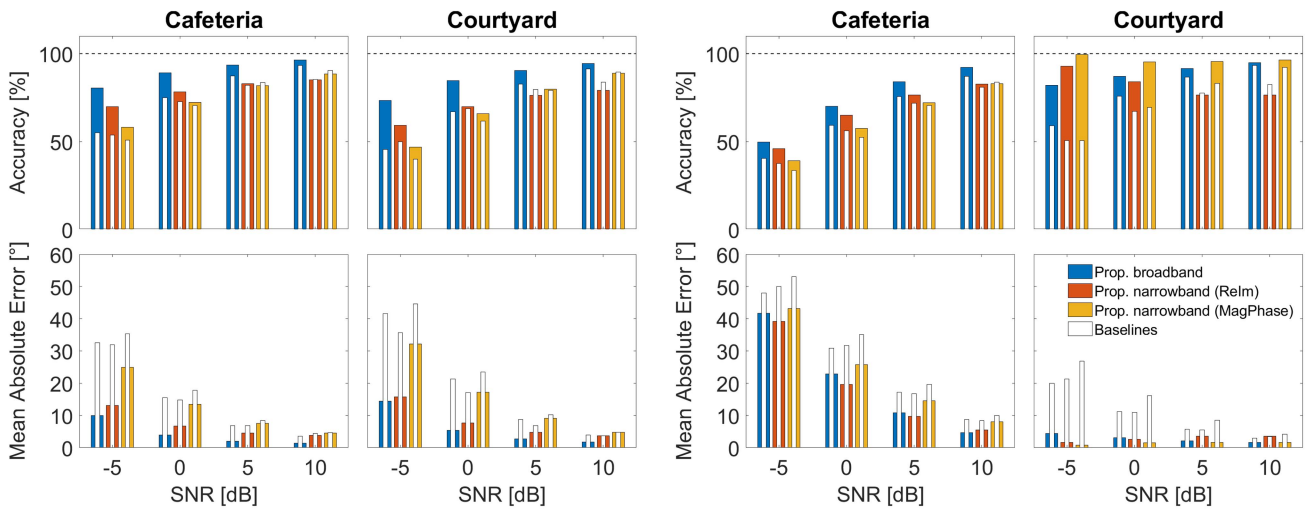
A. Speech-Aware DOA Estimation

To illustrate speech-aware DOA estimation, we consider an exemplary static source scenario in the courtyard (cf. Fig. 5(b)) for a female speech source at position C and head orientation 1 (corresponding to a DOA of $-20°$) with simulated diffuse noise at 0 dB and 10 dB SNR conditions. The proposed system with broadband input features (Fig. 4(a)) is chosen for DOA estimation in this scenario. Fig. 6 depicts the noisy reference microphone signal with a duration of 1 s, the corresponding block-averaged representation of the PD feature, the speech DOA detection (cf. (11) and (13)), and the DOA estimation error. Please note that the difference in the starting times between the reference microphone signal in subfigure (i) and the subsequent subfigures (ii-iv) is due to the design of the proposed system, which requires input features from consecutive time frames over a period of 105 ms before generating the first prediction. To aid in visualization, we obtained the block-averaged PD by averaging the PD values over consecutive time frames used by the CNN for each prediction.

When analyzing Fig. 6, several key observations emerge. First, comparing speech DOA detection results in the two SNR conditions (Fig. 6(a).(iii) and (b).(iii) shows that the speech-aware DOA estimation results in fewer signal blocks with DOA detections in the low SNR condition compared to the high SNR condition. Second, comparing the DOA detection results with the absolute error in either of the SNR conditions, e.g., in the low SNR condition (Fig. 6(a).(iii) and 6(a).(iv), demonstrates that for this example, all estimated DOAs result in absolute errors below $5°$, i.e., 100% accuracy. These findings illustrate the primary objective in designing the speech-aware DOA estimation systems, which is to reliably detect speech DOAs while excluding signal blocks prone to poor DOA estimation performance, without needing a separate VAD. As expected, such a system detects fewer signal blocks with reliable speech DOA in the low SNR condition. Moreover, when comparing block-averaged PD with DOA detection results, especially in the low SNR condition (Fig. 6(a).(ii) and (a).(iii), it becomes evident that the proposed system predominantly estimates the DOA for blocks with a high degree of periodicity. These observations are noteworthy because they demonstrate that the proposed system automatically selects the most reliable signal blocks for DOA estimation, primarily those with a high degree of periodicity, which are less susceptible to noise.

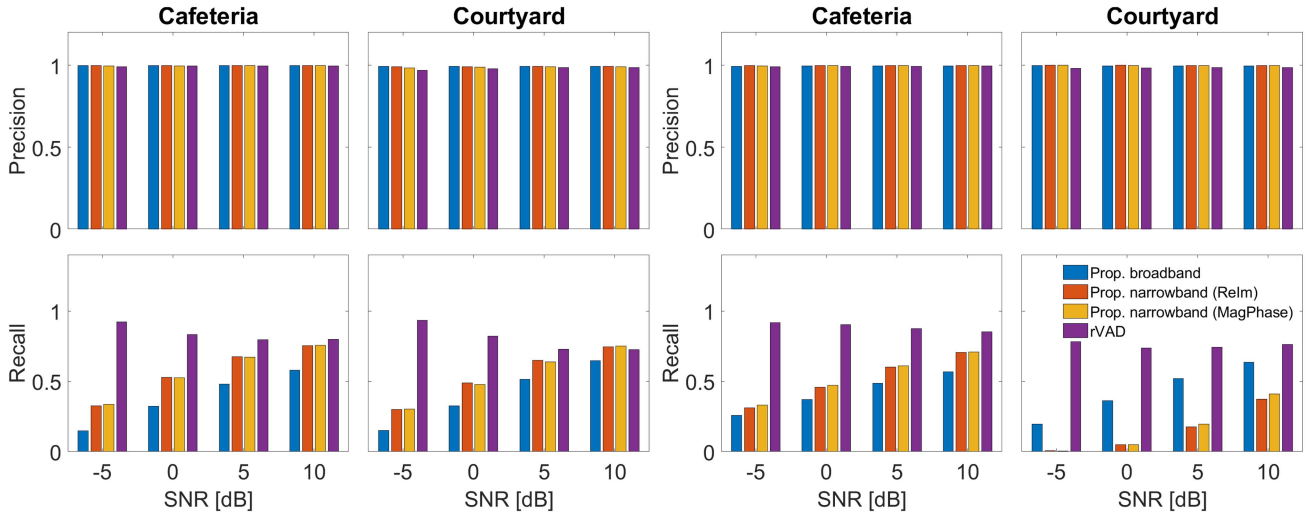
B. Evaluation Results for Static Source Scenarios

For the static source scenarios in two reverberant environments (cafeteria and courtyard), Fig. 7 shows the accuracy and mean absolute error at different SNRs for three proposed systems



(a) simulated diffuse background noise (matched condition) (b) recorded background noise (unmatched condition)

Fig. 7. Accuracy and mean absolute error of the proposed and baseline systems for the static source scenarios in two reverberant environments (cafeteria and courtyard) for different SNRs: (a) Matched noise condition with simulated diffuse background noise and (b) unmatched noise condition with recorded background noise. Colored bars show the performance of the proposed systems using broadband feature combination (GCC-PHAT and broadband PD) or narrowband feature combination (either CPS ReIm or MagPhase and subband-averaged PD), whereas white bars show the performance of the baseline systems using only broadband (GCC-PHAT) or narrowband (either CPS ReIm or MagPhase) spatial features.



(a) simulated diffuse background noise (matched condition) (b) recorded background noise (unmatched condition)

Fig. 8. Speech detection performance of the proposed systems and rVAD in terms of the precision and recall for the static source scenarios in two reverberant environments (cafeteria and courtyard) for different SNRs: (a) Matched noise condition with simulated diffuse background noise and (b) unmatched noise condition with recorded background noise.

(Section IV-B) and three baseline systems (Section IV-A) using either broadband or narrowband features. Performance measures of three proposed systems, i.e., the proposed system with broadband PD and GCC-PHAT (Prop. broadband), the proposed system with subband-averaged PD and real and imaginary parts of CPS (Prop. narrowband ReIm), and the proposed system with subband-averaged PD and magnitude and phase parts of CPS (Prop. narrowband MagPhase) are depicted by colored bars. To facilitate the direct comparison between each proposed system and the corresponding baseline system using the same spatial feature, white narrow bars in front of the colored bars show the performance measures of the corresponding baseline system. A dashed line in the top plots of each figure shows the maximum accuracy of 100% that each system can achieve. In addition to the DOA estimation metrics depicted in Fig. 7, Fig. 8 shows the speech detection evaluation results in terms of precision and recall at different SNR conditions for three proposed systems (Section IV-B) and the rVAD [19] used in the baseline systems (Section IV-A).

1) Matched Noise Condition: Fig. 7(a) depicts the performance measures for the matched noise condition with simulated diffuse background noise (also used during training). Comparing the performance of the proposed systems (colored bars) with the corresponding baseline systems (white bars), we can clearly

observe the benefit of using PD in combination with a spatial feature in both environments for all systems at low SNRs (-5 dB and 0 dB), whereas the benefit also persists for the broadband system at high SNRs (5 dB and 10 dB). For example, for an SNR of 0 dB in the cafeteria environment, the benefit of using PD features in terms of accuracy is approximately 14% points for the broadband system, 6% points for the narrowband system (ReIm), and 1% points for the narrowband system (MagPhase), whereas the benefit in terms of MAE is 11.7° for the broadband system, 8° for the narrowband system (ReIm), and 3.8° for the narrowband system (MagPhase).

Fig. 8(a) depicts the speech detection performance measures for the matched noise condition. It can be observed that all proposed systems exhibit nearly perfect precision, approaching 1. This suggests a low likelihood of falsely detecting a signal portion for DOA estimation (i.e. low false positive). It can be clearly seen for all conditions that the proposed systems demonstrate either better or comparable precision compared to the rVAD, but a lower recall. It is important to emphasize once more that the rVAD is specifically designed for speech detection, whereas the proposed systems are designed for speech DOA detection. This distinction is crucial, as the proposed systems leverage an output class which indeed serves as an uncertainty measure for DOA estimation. Although this class is regarded as a detection class, it has not been merely trained for the speech detection task.

2) Unmatched Noise Condition: Fig. 7(b) depicts the performance measures for the unmatched noise condition with recorded background noise (not seen during training). Except for the narrowband system (MagPhase) in the cafeteria environment at 10 dB SNR and the narrowband system (ReIm) in the courtyard environment at 10 dB SNR, the proposed systems using PD in combination with a spatial feature outperform the corresponding baseline systems for all SNRs in both environments. For example, for an SNR of 0 dB in the cafeteria environment, the benefit of using PD features in terms of accuracy is approximately 10% points for the broadband system, 9% points for the narrowband system (ReIm), and 5% points for the narrowband system (MagPhase), whereas the benefit in terms of MAE is 8° for the broadband system, 12.1° for the narrowband system (ReIm), and 9.2° for the narrowband system (MagPhase).

Fig. 8(b) depicts the speech detection performance measures for the unmatched noise condition. It can be observed that in the courtyard environment, the proposed narrowband systems result in notably low recall, particularly in low SNR conditions. The very low recall in this condition corresponds to a high number of missed detections (i.e. high false negative). However, as observed in Section VI-A, it's essential to emphasize that the primary objective of speech-aware systems is to detect the speech DOA for reliable localization, rather than solely focusing on speech activity detection. The good results for the proposed narrowband systems at low SNRs in the courtyard in the unmatched condition (Fig. 7(b)) can be attributed to the fact that these systems use only a small fraction of speech signal blocks for DOA estimation.

When comparing the performance measures between the matched and unmatched noise conditions (Fig. 7(a) and (b)),

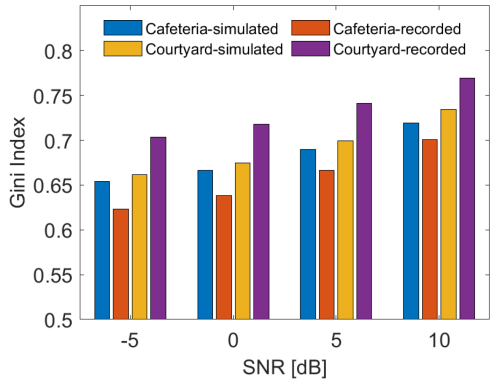


Fig. 9. Average spectro-temporal Gini index for two environments (cafeteria, courtyard), two background noise types (simulated, recorded) and different SNRs.

it can be clearly observed that in the cafeteria environment the performance for the recorded babble noise is worse than that for the simulated diffuse noise, whereas (somewhat surprisingly) in the courtyard environment the performance for the recorded ambient noise is better than that for the simulated diffuse noise. This can be explained by investigating the spectro-temporal sparsity of the signals for the different conditions. For the sparsity analysis, we use the Gini index [47], where a large Gini index (close to 1) corresponds to high sparsity, and a small Gini index (close to 0) corresponds to low sparsity. More in particular, we consider the joint spectro-temporal Gini index according to [48], computed on the STFT spectrogram of the noisy reference microphone signal. For each environment and background noise type and for different SNRs, Fig. 9 depicts the spectro-temporal Gini index averaged over all 150 speech segments. On the one hand, in the cafeteria environment, it can be observed for all SNRs that the spectro-temporal sparsity of the microphone signals with recorded babble noise is less than the spectro-temporal sparsity with simulated diffuse noise. On the other hand, in the courtyard environment, it can be observed for all SNRs that the microphone signals with recorded ambient noise exhibit a sparser spectro-temporal structure than with the simulated diffuse noise. Hence, in conjunction with the DOA estimation performance in Fig. 7, we can deduce that signals with sparser spectro-temporal structure appear to lead to better speech-aware DOA estimation.

Taking a closer look at Fig. 7(b), it becomes evident that, in the courtyard environment with recorded background noise, the two proposed narrowband systems perform best under the lowest SNR condition (0 dB SNR). The Gini index, however, does not provide a comprehensive explanation for this particular case. Unlike the simulated diffuse noise and cafeteria babble noise, the courtyard ambient noise energy predominantly falls within the first frequency band of PD features. This means that at low SNRs, especially at -5 dB, the noise can mask the harmonic structure of speech signals in this frequency band. This masking potentially aids the CNN in almost perfectly identifying segments with prevalent noise, enhancing DOA estimation accuracy. As SNR increases, enhanced harmonics in

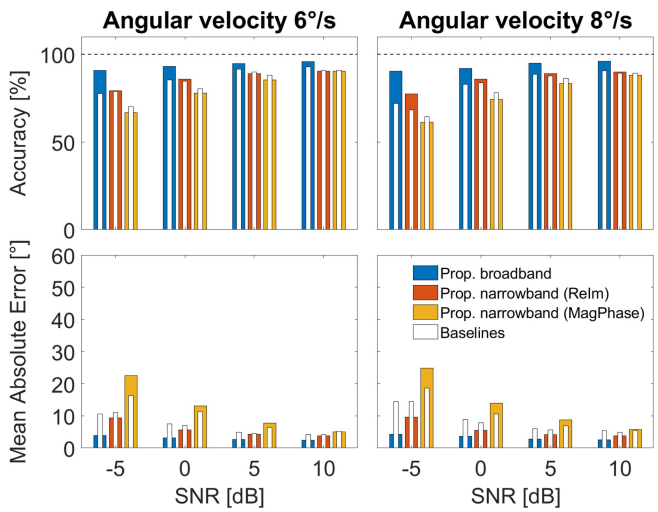


Fig. 10. Accuracy and mean absolute error of the proposed and baseline systems using broadband and narrowband features for the dynamic source scenarios in the office environment for different SNRs and angular velocities.

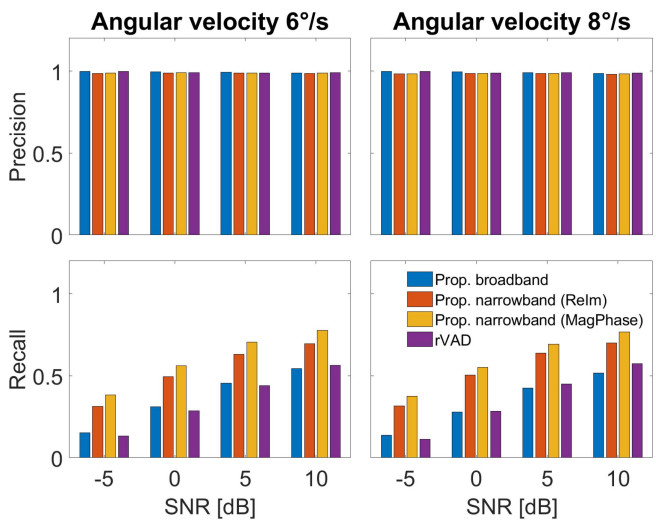


Fig. 11. Speech detection performance of the proposed systems and rVAD in terms of the precision and recall for the dynamic source scenarios in the office environment for different SNRs and angular velocities.

low frequencies may introduce uncertainties, potentially compromising DOA estimation accuracy. However, this does not affect our main findings and conclusions for speech-aware DOA estimation.

C. Evaluation Results for Dynamic Source Scenarios

For the moving source scenario in the office environment, Fig. 10 depicts the DOA estimation performance measures of the proposed systems (colored bars) and the corresponding baseline systems using the same spatial feature (white bars) for different SNRs and two angular velocities. Similarly as for the static source scenario (Fig. 7), a clear benefit of using PD features can be observed, especially for the broadband system at all SNRs and for the narrowband system (ReIm) at low SNRs. For the narrowband system (MagPhase), whose performance is anyway lower than the narrowband system (ReIm), the baseline system (using MagPhase) exhibits comparable or better performance. These results further reveal the benefit of using PD features in the proposed speech-aware DOA estimation systems compared to the baseline systems using merely spatial features. This benefit even increases with angular velocity, particularly at low SNRs.

For the moving source scenario, the evaluation results of speech detection performance for all considered systems are illustrated in Fig. 11. It becomes evident that in dynamic scenarios across all conditions, the proposed narrowband systems yield a higher recall when compared to all other systems (including rVAD), while maintaining a high level of precision. This is particularly noteworthy, as the higher recall facilitates speech source tracking by generating more observations of the dynamic scene.

Evaluation results in Figs. 7 and 10 show that, except for the matched condition in the static source scenario, the proposed broadband system outperforms the proposed narrowband systems, while indicating a larger benefit from the inclusion of PD features. The results also demonstrate that the broadband baseline system using GCC-PHAT features typically outperforms narrowband baseline systems using CPS features. Despite a similar number of trainable parameters (cf. Table I), the narrowband baseline systems must learn more intricate patterns from CPS features, whereas GCC-PHAT directly provides time delay information. This suggests that the narrowband baseline systems may need more capacity (trainable parameters) to match the performance of the broadband one. As our main goal was to study the benefit of using PD features in the proposed narrowband and broadband systems, we didn't optimize the narrowband systems for performance parity with the broadband system, potentially causing performance limitations when combining CPS and PD features.

D. Limitations and Future Works

This study only considered binaural DOA estimation of a far-field speech source. For a speech source in the near field of a microphone array, accurate estimation of the time delay (and phase) involves considering both the range and the DOA of the sound source. The normalization inherent in the PHAT weighting (see (4)) eliminates the effect of the signal level (and hence range information) due to the source-microphone distance. Consequently, a model trained solely on the GCC-PHAT may have limited capability to leverage range-dependent information in the near-field scenarios. Although this study only considered binaural DOA estimation in the azimuthal plane, the proposed systems, in principle, can be extended for DOA estimation in terms of both azimuth and elevation as azimuth and elevation information are encoded by the spatial input features [7], [12].

In this study, we examined single-talker speech-aware DOA estimation in the presence of background noise. Future research may explore the potential benefits of using PD features for speech-aware DOA estimation in the presence of non-speech interference and binaural DOA estimation in multi-talker scenarios.

VII. Conclusion

In this study, we proposed novel feature combinations for speech-aware DOA estimation in the context of binaural hearing aids. The proposed systems utilize CNNs and receive a spatial feature and an auditory-inspired periodicity feature as inputs to two parallel branches of convolutional layers. In particular, we introduced a subband-averaged PD feature as the periodicity feature, and combined it with either the real and imaginary or the magnitude and phase components of the narrowband CPS as the spatial feature. The performance of speech-aware systems was evaluated against CNN-based baseline systems which only use spatial features and a pitch-based VAD.

Comprehensive evaluations in static single-talker scenarios with different background noise types and SNRs demonstrate that for any type of spatial feature, the proposed method outperforms baseline systems in terms of DOA estimation accuracy and mean absolute error, particularly in adverse SNR conditions and in conditions with higher degrees of spectro-temporal sparseness. This study also shows that the proposed method using PD features is effective for speech-aware DOA estimation of a moving talker, and is robust to changes in talker velocity. Our proposed speech-aware system is able to estimate the sound source DOA when a high degree of periodicity is captured by the CNN, without any need for a separate VAD or pitch period estimation.

The primary finding of this study was that the usage of PD features in both narrowband and broadband feature combinations benefits the speech-aware binaural DOA estimation in different static and dynamic scenarios. It was also found that the proposed system employing the broadband feature combination typically demonstrated better performance than the proposed systems using the narrowband feature combinations in the specific system configuration employed in this study.

Overall, this study demonstrates the potential benefits of utilizing periodicity-based features in conjunction with spatial features for speech-related applications such as DOA estimation. The results also suggest that these features may have wider applications in other speech-related tasks. The findings of this study can contribute to the development of improved methods for sound source localization and speech enhancement in binaural hearing aids.

References

[1] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ, USA: Wiley, 2006.

[2] T. May, S. v. d. Par, and A. Kohlrausch, *Binaural Localization and Detection of Speakers in Complex Acoustic Scenes*. Berlin, Germany: Springer, 2013, pp. 397–425.

[3] S. Braun, W. Zhou, and E. A. P. Habets, "Narrowband direction-of-arrival estimation for binaural hearing aids using relative transfer functions," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2015, pp. 1–5.

[4] M. Farmani, M. S. Pedersen, Z. Tan, and J. Jensen, "Bias-compensated informed sound source localization using relative transfer functions," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 7, pp. 1275–1289, Jul. 2018.

[5] C. Evers et al., "The LOCATA challenge: Acoustic source localization and tracking," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1620–1643, 2020.

[6] D. Fejgin and S. Doclo, "Comparison of binaural RTF-vector-based direction of arrival estimation methods exploiting an external microphone," in *Proc. IEEE 29th Eur. Signal Process. Conf.*, 2021, pp. 241–245.

[7] P. A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *J. Acoust. Soc. Amer.*, vol. 152, no. 1, pp. 107–151, 2022.

[8] N. Ma, T. May, and 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.*, vol. 25, no. 12, pp. 2444–2453, Dec. 2017.

[9] S. Chakrabarty and E. A. P. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 8–21, Mar. 2019.

[10] P. Vecchiotti, N. Ma, S. Squartini, and G. J. Brown, "End-to-end binaural sound localisation from the raw waveform," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 451–455.

[11] Z. Q. Wang, X. Zhang, and D. Wang, "Robust speaker localization guided by deep learning-based time-frequency masking," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 178–188, Jan. 2019.

[12] C. Pang, H. Liu, and X. Li, "Multitask learning of time-frequency CNN for sound source localization," *IEEE Access*, vol. 7, pp. 40725–40737, 2019.

[13] J. Wang, J. Wang, K. Qian, X. Xie, and J. Kuang, "Binaural sound localization based on deep neural network and affinity propagation clustering in mismatched HRTF condition," *EURASIP J. Audio, Speech, Music Process.*, vol. 2020, no. 1, pp. 1–16, 2020.

[14] J. Ding, Y. Ke, L. Cheng, C. Zheng, and X. Li, "Joint estimation of binaural distance and azimuth by exploiting deep neural networks," *J. Acoust. Soc. Amer.*, vol. 147, no. 4, pp. 2625–2635, 2020.

[15] B. Yang, H. Liu, and X. Li, "Learning deep direct-path relative transfer function for binaural sound source localization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3491–3503, 2021.

[16] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[17] W. He, P. Motlicek, and J.-M. Odobez, "Joint localization and classification of multiple sound sources using a multi-task neural network," in *Proc. Interspeech*, 2018, pp. 312–316.

[18] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen, "Overview and evaluation of sound event localization and detection in DCASE 2019," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 684–698, 2021.

[19] Z. H. Tan, A. K. Sarkar, and N. Dehak, "rVAD: An unsupervised segment-based robust voice activity detection method," *Comput. Speech Lang.*, vol. 59, pp. 1–21, 2020.

[20] A. Küçük, A. Ganguly, Y. Hao, and I. M. S. Panahi, "Real-time convolutional neural network-based speech source localization on smartphone," *IEEE Access*, vol. 7, pp. 169969–169978, 2019.

[21] H. Hammer, S. E. Chazan, J. Goldberger, and S. Gannot, "Dynamically localizing multiple speakers based on the time-frequency domain," *EURASIP J. Audio, Speech, Music Process.*, vol. 2021, no. 1, pp. 1–10, 2021.

[22] D. Salvati, C. Drioli, and G. L. Foresti, "Localization and tracking of an acoustic source using a diagonal unloading beamforming and a Kalman filter," in *Proc. LOCATA Challenge Workshop - Satell. Event IWAENC*, 2018.

[23] A. Bohlender, A. Spriet, W. Tirry, and N. Madhu, "Exploiting temporal context in CNN based multisource DOA estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1594–1608, 2021.

[24] N. Yalta, K. Nakadai, and T. Ogata, "Sound source localization using deep learning models," *J. Robot. Mechatronics*, vol. 29, no. 1, pp. 37–48, 2017.

[25] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 405–409.

[26] G. Hu and D. Wang, "Segregation of unvoiced speech from nonspeech interference," *J. Acoust. Soc. Amer.*, vol. 124, no. 2, pp. 1306–1319, 2008.

[27] Z. Chen and V. Hohmann, "Online monaural speech enhancement based on periodicity analysis and a priori SNR estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 11, pp. 1904–1916, Nov. 2015.

[28] R. Varzandeh, K. Adiloğlu, S. Doclo, and V. Hohmann, "Exploiting periodicity features for joint detection and DOA estimation of speech sources using convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 566–570.

[29] H. Kayser and J. Anemüller, "A discriminative learning approach to probabilistic acoustic source localization," in *Proc. IEEE 14th Int. Workshop Acoustic Signal Enhancement*, 2014, pp. 99–103.
[30] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 2814–2818.
[31] W. He, P. Motlicek, and J. Odobez, "Deep neural networks for multiple speaker detection and localization," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2018, pp. 74–79.
[32] E. L. Ferguson, S. B. Williams, and C. T. Jin, "Sound source localization in a multipath environment using convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 2386–2390.
[33] S. Popham, D. Boebinger, D. P. Ellis, H. Kawahara, and J. H. McDermott, "Inharmonic speech reveals the role of harmonicity in the cocktail party problem," *Nature Commun.*, vol. 9, no. 1, pp. 1–13, 2018.
[34] A. Josupeit and V. Hohmann, "Modeling speech localization, talker identification, and word recognition in a multi-talker setting," *J. Acoust. Soc. Amer.*, vol. 142, no. 1, pp. 35–54, 2017.
[35] A. Josupeit, N. Kopčo, and V. Hohmann, "Modeling of speech localization in a multi-talker mixture using periodicity and energy-based auditory features," *J. Acoust. Soc. Amer.*, vol. 139, no. 5, pp. 2911–2923, 2016.
[36] J. Luberadzka, H. Kayser, and V. Hohmann, "Making sense of periodicity glimpses in a prediction-update-loop—a computational model of attentive voice tracking," *J. Acoust. Soc. Amer.*, vol. 151, no. 2, pp. 712–737, 2022.
[37] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.
[38] Y. Wu and K. He, "Group normalization," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
[39] J. O. Smith and X. Serra, "PARSHL: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation," in *Proc. Int. Comput. Music Conf.*, 1987, pp. 290–297.
[40] J. Garofolo et al., "Timit acoustic-phonetic continuous speech corpus," in *Linguistic Data Consortium*, 1993.
[41] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. ACM Conf. Multimedia*, 2015, pp. 1015–1018.
[42] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," *EURASIP J. Adv. Signal Process.*, vol. 2009, no. 1, Jul. 2009, Art. no. 298605.
[43] W. A. Dreschler, H. Verschuure, C. Ludvigsen, and S. Westermann, "ICRA noises: Artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment," *Audiology*, vol. 40, no. 3, pp. 148–157, 2001.
[44] M. Park, "Models of binaural hearing for sound lateralisation and localisation," Ph.D. dissertation, Univ. Southampton, Southampton, U.K., Oct. 2007.
[45] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2019, pp. 8026–8037.
[46] D. P. Kingma and L. J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.
[47] N. Hurley and S. Rickard, "Comparing measures of sparsity," *IEEE Trans. Inf. Theory*, vol. 55, no. 10, pp. 4723–4741, Oct. 2009.
[48] N. K. Desiraju, S. Doclo, and T. Wolff, "Efficient multichannel acoustic echo cancellation using constrained tap selection schemes in the subband domain," *EURASIP J. Adv. Signal Process.*, vol. 2017, no. 1, pp. 1–16, 2017.

Reza Varzandeh received the B.Sc. degree in electrical engineering from the Shahid Chamran University of Ahvaz, Ahvaz, Iran, in 2008, and the M.Sc. degree in electrical engineering from the Amirkabir University of Technology, Tehran, Iran, in 2012. From 2017 to 2020, he was a research and development Engineer with HörTech Oldenburg (Hörzentrum Oldenburg). Since 2020, he has been with the Auditory Signal Processing Group, University of Oldenburg, Oldenburg, Germany. His research interests include computational auditory scene analysis using machine learning and deep learning, in particular speech detection and localization.

Simon Doclo (Senior Member, IEEE) received the M.Sc. degree in electrical engineering and the Ph.D. degree in applied sciences from KU Leuven, Leuven, Belgium, in 1997 and 2003, respectively. From 2003 to 2007, he was a Postdoctoral Fellow with KU Leuven and McMaster University, Hamilton, ON, Canada. From 2007 to 2009, he was a Principal Scientist with NXP Semiconductors in Leuven, Belgium. Since 2009, he has been a Full Professor with the University of Oldenburg, Oldenburg, Germany, and scientific Advisor of the Branch Hearing, Speech and Audio Technology HSA of the Fraunhofer Institute for Digital Media Technology IDMT. His research interests include signal processing for acoustical and biomedical applications, more specifically microphone array processing, speech enhancement, active noise control, acoustic sensor networks and hearing aid processing. Prof. Doclo was the recipient of the several best paper awards (International Workshop on Acoustic Echo and Noise Control 2001, EURASIP Signal Processing 2003, IEEE Signal Processing Society 2008, VDE Information Technology Society 2019). He was a Member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing and the EURASIP Technical Area Committee on Acoustic, Speech and Music Signal Processing and also a Member of the EAA Technical Committee on Audio Signal Processing. Since 2021, he has been the Senior Area Editor of IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING.

Volker Hohmann received the Diploma-Phys. in physics and the doctorate degree in physics (Dr. rer. nat.) from the University of Göttingen, Göttingen, Germany, in 1989 and 1993, respectively. From 1993 to 2012, he was a faculty Member of the Physics Institute, Oldenburg University, Oldenburg, Germany, and was appointed a Full Professor with the Faculty for Medicine and Health Sciences, Oldenburg University in 2012. His research interests include acoustics and digital signal processing with applications to signal processing in speech processing devices, such as, hearing aids. He is a Consultant with the Hörzentrum Oldenburg GmbH and with HörTech gGmbH Oldenburg. He was the Guest Researcher with Boston University, Boston, MA, USA, (Prof. Dr. Colburn) in 2000 and with the Technical University of Catalonia, Barcelona, Spain, in 2008. Prof. Hohmann was the recipient of the Lothar-Cremer price of the German acoustical society (DEGA) in 2008 and the German President's Award for Technology and Innovation in 2012.