Deep Learning-Based Speech Specific Source Localization by Using Binaural and Monaural Microphone Arrays in Hearing Aids

Peyman Goli and Steven van de Par

***Abstract*—A deep learning-based method is proposed for jointly detecting and localizing speech sources in a complex acoustic scene by using microphones of a hearing aid. Motivated by the human auditory system, peripheral preprocessing is applied on the microphone signals to obtain auditory subband signals that serve as input to the proposed deep neural network for detecting and localizing speech sources. In the proposed neural network, a combination of residual and dense aggregation learning is utilized rather than the conventional residual learning to preserve and reuse the spatial representations at the output layers. This process is performed to improve the gradient flow in deeper layers, in the training stage. The learning curves show that the proposed residual-dense aggregation mapping do improve the speed and accuracy of the convergence. The proposed model shows good performance in joint speech source detection and localization using a binaural microphone array (i.e., three channels at each side) but also using a monaural microphone array (i.e., four channels at the right side) despite of the short distances between the microphones. The proposed methods also outperform neural networks that are directly using STFT components of the binaural or monaural microphone arrays. In addition, the proposed models extended with learnable peripheral processing show slightly improved in detection and localization scores compared to the proposed models using the plain auditory subband signals, in both the binaural and monaural microphone arrays but only so, when the learnable peripheral processing is initialized with parameters stemming from human peripheral processing.**

***Index Terms*—Auditory system, direction of arrival estimation, speech source localization, deep learning, hearing aids.**

I. Introduction

CONTRARY to the remarkable capabilities of normal hearing listeners in understanding speech in a complex acoustic environment, hearing-impaired listeners can have substantial difficulties in speech source localization and in understanding speech [1]. To support hearing-impaired listeners, methods are developed that allow beamformers in hearing aids to selectively enhance speech sources [2], [3]. For this purpose, combined speech source detection (SSD) and localization (SSL) algorithms have been developed that can determine at what location a speech source is placed amid interfering noise sources placed within a reverberant environment.

Both for hearing aids (HA), as well as for human-robot interaction, numerous conventional methods for SSL (i.e., non-deep learning-based methods) have been proposed that use time-frequency representations of microphone signals as inputs. Some approaches in HAs use interaural features derived from pairs of microphone signals that resemble auditory localization cues such as interaural time and level differences. Many conventional SSL approaches in human–robot interaction rely on the sound having different propagation pathways to the multiple microphone channels located on the robots. These methods often extract the spatial information from the spectral representation of microphone signals to predict the direction of arrivals (DOA) of sound sources. For example, time-difference-based algorithms such as the generalized cross-correlation with phase transform (GCC-PHAT) [4] estimate the time differences between the Fourier transforms of pairs of microphone signals to localize sound sources. Beamforming-based approaches such as steered-response power phase transform (SRP-PHAT) [5] search for the candidate source location that maximizes the output of a steered delay-and-sum beamformer. GCC-PHAT is also a conventional feature for source localization in microphone configurations used in HAs [6]. Histogram analysis approaches estimate the angles of arrivals by scanning DOAs across all angles to determine that direction containing most energy [7]. Subspace-based methods such as multiple signal classification (MUSIC) algorithms rely on the maximum of the spatio-spectral covariance matrix of microphone channels to estimate DOA of speech sources [8].

Since the microphone configurations in HAs, contrary to robotics, are limited to the positions of the ears, the spatial distribution of the microphone positions in HAs may not provide sufficient time-of-arrival differences for the SSL methods to provide optimal performance. Therefore, some conventional methods, specifically in HAs, have used models of computational auditory scene analysis (CASA) to localize speech sources inspired by the human ability to detect and localize speech sources [9], [10], [11], [12], [13], [14], [15]. In the auditory system, starting from

Manuscript received 2 March 2022; revised 17 September 2022, 5 November 2022, 10 January 2023, and 3 March 2023; accepted 4 April 2023. Date of publication 24 April 2023; date of current version 1 May 2023. This work was supported by the Deutsche Forschungsgemeinschaft, German Research Foundation, through Germany's Excellence Strategy – EXC 2177/1 under Grant 390895286, titled Hearing4all. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. K. Kinoshita. *(Corresponding author:Peyman Goli.)*

The authors are with the Acoustics Group, Cluster of Excellence "Hearing4all", Department of Medical Physics and Acoustics, Oldenburg University, 26129 Oldenburg, Germany (e-mail: peyman.goli@uol.de; steven.van.de.par@uni-oldenburg.de).

Digital Object Identifier 10.1109/TASLP.2023.3268734

the outer ear and ending in the cortical areas of the brain, the spatial information of speech sources is processed by segmenting, grouping, and integrating the information extracted from the binaural signals [13]. Interaural time difference (ITD) and interaural level difference (ILD) between the ears are the two major cues that are exploited by the human auditory system to localize sound sources. ITDs result from the difference in arrival time of a sound between the ears, and ILDs result from the head shadow effect at high frequencies. It has been proposed that the interaural coherence is used as a selection criterion to select reliable interaural cues for speech source localization in mixtures of several speakers [13]. In [14], based on the different positions of sound sources, a decision rule system is used to separate a speech source from interfering sounds. In [15], a probabilistic localization model based on Gaussian mixture models (i.e., shallow learning models) was developed that estimated likelihoods of different azimuth angles for each time-frequency interval. These azimuth dependent likelihoods were then used to segregate sources in the time-frequency domain to support a missing data classifier to determine the presence of a speech source for various azimuth angles.

Using multiple microphone channels at each ear in HAs can further improve the spatial resolution because interchannel cues [i.e., interchannel time difference (ICTD) and interchannel level difference (ICLD)] can be determined across multiple pairs. Utilizing the contributions of the multiple interchannel cues in speech source localization and creating a mapping between the cues and corresponding azimuth angles at different acoustic conditions are challenging tasks that cannot be handled well by conventional methods and shallow learning models. Deep neural network (DNN)-based methods have been proposed for estimating and exploiting optimal representations of the interchannel cues for detecting and localizing speech sources [16], [17], [18], [19], [20], [21]. In addition, DNN-based approaches have been shown to be able to handle strong interfering sources and reverberation which are very challenging tasks in conventional methods.

Although deep learning-based methods that use high-level features in terms of pure interchannel cues may require neural networks with fewer layers and less complexity compared to neural networks using raw data, these interaural cues may not provide the most optimal representations of the raw microphone signals throughout the deep layers for localization. Accordingly, some recent DNN-based approaches in environment–robot interactions have used the raw time-frequency (TF) representations of the microphone signals as the inputs to preserve the optimal spatial representations throughout the deep layers [22], [23], [24]. In [22], the inputs of the deep neural network are the powers of the short-time Fourier transform (STFT) components of the signals received by 16 microphone channels placed around a robot body. In this method, a deep classification model for estimating the probability density across azimuth angles was trained using the gradient of a cross entropy loss function. In [23], the real and imaginary parts of STFT were used as input to a deep regression model in order to detect and localize speech sources. Here, the microphones were located on a robot device in a coplanar rectangular shape microphone array. In [24], the phase and magnitude of the STFT components of the signals received by circular and spherical microphone arrays were utilized as the input to a deep neural network to predict the azimuth angle of target speech sources. In all these methods, the microphone channels are positioned in a coplanar or noncoplanar microphone arrays around the robot devices. Compared to HAs, this microphone positioning can provide a better spatial resolution due to their more evenly distributed spatial configurations [25]. In a single HA, microphone channels are clustered more close to each other, and only small ICTDs and ICLDs are observed. Thus, although in robot devices, the STFT components may enable the extraction of sufficiently strong cues for localization, the closely spaced microphones in HAs usually allow only a poor representation of ICTDs and ICLDs.

In this study, we propose a deep learning-based method to jointly detect and localize speech sources in noisy and reverberant environments by using six behind-the-ear channels of a binaural HA microphone array. Rather than using STFT components extracted on a frame-by-frame basis, we use the time domain auditory subband signals obtained by an auditory peripheral preprocessing stage applied to each microphone signal. In the peripheral preprocessing, an auditory filter bank followed by half-wave rectification and square-root compression is applied on the microphone signals to prepare the inputs for the neural network. These preprocessing stages are inspired on the human auditory system. The interchannel cues are better represented in this way compared to using STFT components because the subband signals extracted by the auditory filter bank have better spectral resolution at lower frequencies. Note also that for a narrow-band signal, halfwave rectification will preserve the original (spectral) information that was present before rectification [26]. As a result, spatial information as well as periodicity information is preserved. Importantly, the onsets of individual periods may provide salient time-delay information between the auditory subband signals, which may be readily detected by convolutional layers. The square-root compression limits the total dynamic range of the subband signals while still preserving level information. Initial layers in the proposed deep neural network are designed to learn to extract the optimal interchannel cues from the raw auditory subband signals to be utilized in the rest of the network for detection and localization at various acoustic conditions.

Note that recently, some DNN-based speech recognition and sound classification approaches have suggested using learnable auditory filter banks to estimate the optimal TF representations of microphone signals [27], [28]. Estimating the optimal auditory filters (e.g., optimal center frequencies and bandwidths) through deep layers can potentially provide a better front end than the peripheral preprocessing we propose. Accordingly, we have also utilized a convolution-based learnable gammatone filter bank (LGFB) as a front-end SSD and SSL processing the raw microphone signals.

Utilizing the raw data (e.g., auditory subband signals or microphone signals) rather than the pure interchannel cues may be beneficial because more information of the input signals is preserved. It will, however, require a deeper neural network. Note that problems, such as gradient vanishing and overfitting, make the deeper neural networks more difficult to train [29]. Some approaches based on residual learning and dense aggregation

have been proposed to deal with this problem in DNN-based image recognition [30] and speech enhancement models [31]. Such approaches allow for the training of very deep neural networks by preventing the vanishing gradient problem and improving gradient flow during backpropagation [32], [33]. In addition, so called, dense aggregations use the identity mapping to concatenate the inputs of shallow layers with the outputs of deeper layers and allow for more layers to explore a larger set of features during training [34]. Based on these insights, in our study, we use a combination of identity mapping, consisting of the residual links and dense aggregation, to further improve gradient flow in deeper layers. This combination also enhances the performance of the neural network in terms of convergence speed and training loss during the training stage.

The contribution of this paper focuses on the following specific aspects of deep learning-based SSL and SSD in HAs: 1. The benefit of utilizing a peripheral preprocessing and a learnable peripheral processing to determine the auditory subband signals as the input of the neural network model and comparing the performance of the proposed model to the DNN-based models using a STFT-based preprocessing. 2. The use of only a monaural microphone array (i.e., four microphones of one hearing aid only) for joint speech source detection and localization. 3. Applying a combination of residual links and dense aggregation in deeper layers to further improve the gradient flow in deeper layers in comparison to the conventional residual blocks used in DNN-based methods for SSL and SSD.

The paper is organized as follows. Section II describes the CASA-based preprocessing stage for determining the auditory subband signals as the inputs of the neural network model. Section III explains the architecture and hyperparameters of the proposed neural network model and the proposed residual-dense learning. Section IV describes the experiments, the datasets prepared for training and testing the neural networks, and the performance of the trained models in SSD and SSL. Section V provides the conclusion and a summary of the main findings.

II. CASA-Based Preprocessing

In our proposed method, we apply a CASA-based preprocessing on the signals received by the microphones of the array to prepare the inputs for the neural network model. The preprocessing stage consists of an auditory peripheral model which results in a set of auditory subband signals, and a preparation stage which prepares the inputs of the neural network.

In the inner ear, the cochlea converts time dependent sound pressure patterns received from the ear drum into cochlear oscillations on the basilar membrane that separates the time signal in different bandpass filtered signals. The auditory nerves (driven by inner hair cells) then passes the interaural signals to the brain. Several approaches have been proposed to simulate the activities of the cochlea and the inner hair cells [35], [36], [37]. In the peripheral processing we used here, the activity of the human auditory system in the inner ear is simulated by applying an auditory filter bank followed by inner hair cell processing, which assumes half wave rectification and square root.

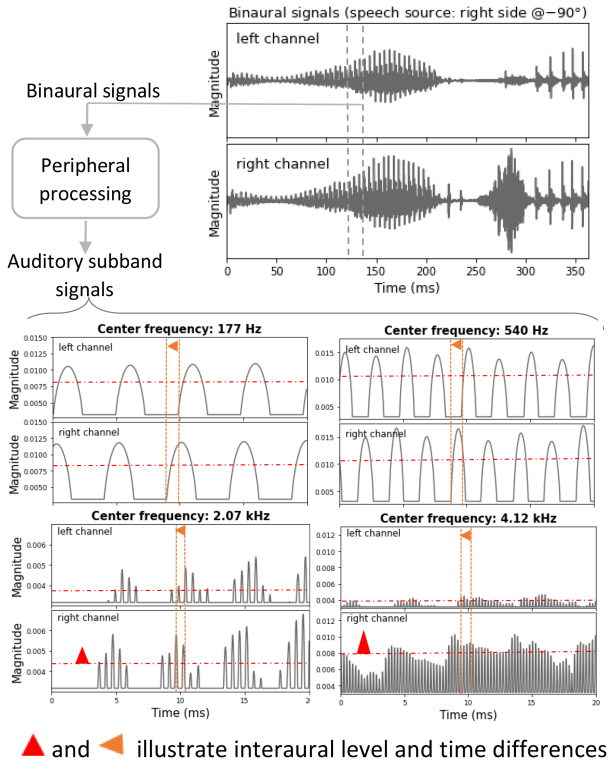

Fig. 1. Binaural signals of a speech source at −90° and the auditory subband signals produced by the peripheral processing in four frequency bands. ILDs increase at high frequencies because of the head shadow effect. ITDs are well detectable between the plain onsets at lower frequencies while conveyed in the modulated envelopes of the auditory subband signals at higher frequencies.

In more detail, the preprocessing stage first uses a rectangular window of 20 ms (882 samples at 44.1 kHz sampling rate) to segment the microphone signals in the time domain. Then a fourth-order gammatone filter bank is applied on the signals to simulate the activity of the human cochlea in the inner ear. The gammatone filter bank decomposes the signals into $N = 32$ subband signals. The band center frequencies are distributed uniformly on the equivalent rectangular bandwidth (ERB) scale between 80 Hz and 5 kHz and have bandwidth in accordance with the ERB estimates of [38]. The neural transduction process in the inner hair cells is simulated by applying halfwave rectification (i.e., simulating the firing rates of the auditory nerves) and square-root compression (i.e., simulating the compressive response of the early stages of auditory processing) on the subband signals

An exemplary binaural signal of a speech source located at −90° (i.e., a speech source at the right side) and the corresponding auditory subband signals in some frequency bands resulting after the peripheral processing are shown in Fig. 1. At lower frequencies, where the wavelength is larger than the diameter of the head, ILDs between the auditory subband signals are extremely small (i.e., near zero), and ITDs are well detectable between the onsets of the positive signal part.

At higher frequency bands, ILDs are well detectable due to the head shadow. Furthermore, in accordance with the processing of the human auditory system, the network can in principle

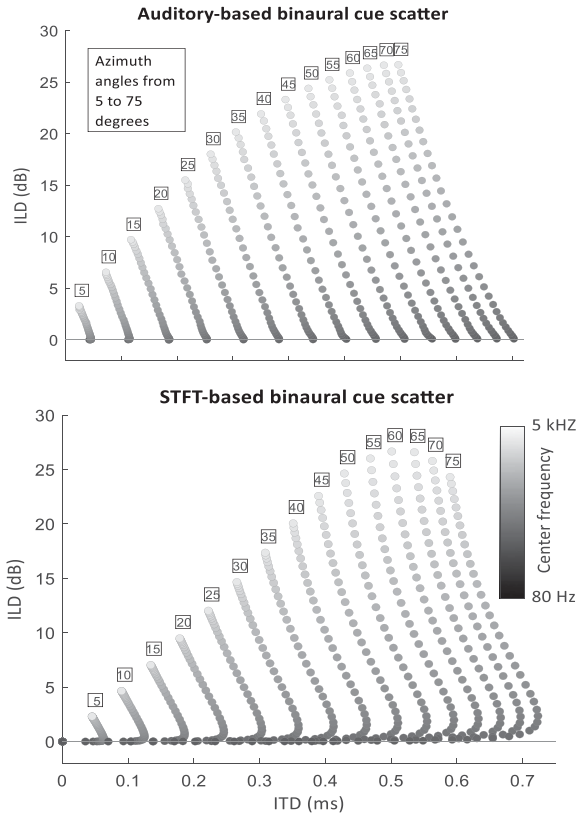

Fig. 2. Auditory-based and STFT-based binaural cue scatters extracted from the HRIRs of the DUDA head at different azimuth angles. Low bin resolution at lower frequencies in the STFT components leads to downscale ITDs at low frequency bands in comparison with the auditory-based ITDs.

exploit the modulated envelopes of the auditory subband signals to detect the time delays in the envelopes at higher frequencies.

Time delays play an important role in human localization, specifically, at lower frequencies. To get more insight in the representation of interaural time delays and level differences as represented in STFT-based approaches, in the lower panel of Fig. 2, scatter plots of STFT-based binaural cues are shown. In addition, in the upper panel 'auditory-based' ITDs and ILDs are shown that are calculated from the peripheral front-end that includes a gamma-tone filtering. For both cases, ITDs and ILDs are extracted using the head-related impulse responses (HRIRs) of the DUDA head model [39] at different azimuth angles. Auditory-based ILDs are calculated using the level differences between the auditory subband signals of HRIRs, and STFT-based ILDs are computed on the basis of the magnitude of Fourier transform of HRIRs at each frequency band. The auditory-based ITDs are determined by searching the maximum of normalized cross-correlation function between the auditory subband signals of HRIRs, and the STFT-based ITDs are calculated on the basis of the unwrapped IPDs of the Fourier transform of HRIRs at each frequency band.

The length of HRIRs is 1024 samples, which is the conventional size of segmented signals used in STFT-based methods for speech source localization. As can be seen in Fig. 2, STFT-based ITDs are downscale at lower frequency bands due to the low frequency-bin resolution of the STFT components. These downscaled time differences can lead to a reduced performance in SSL models at adverse acoustic conditions where the level differences cannot be detected by the model due to target speech signals being masked by interferers at somewhat higher frequencies. The auditory-based ITDs are detectable at all frequencies because the auditory subband signals no longer suffer the poor resolution at lower frequencies. The detectable time delays at lower frequency bands can improve the robustness of speech source localization when the target speech signals are masked by interferers at higher frequency bands.

In the model input preparation stage, a min-max scaling is applied on the auditory subband signals for each time frame separately to set the input value range between 0.1 and 0.9. This prevents that the large dynamic range of the input signal leads to the large gradient variance problem when updating the neural network parameters during the training stage. Four preceding context frames are added to the current input frame to extend the time interval for speech source localization (i.e., the input duration is 5×20 ms $= 100$ ms). The number of the time samples is reduced to $M = 2205$ by using a downsampling step with the rate of 2 to 1. The size of the 3D inputs is [frequency bands (N), time samples (M), microphone channels (P)].

III. Neural Network Model

We utilize a regression model based on a deep convolution neural network (CNN) to jointly detect and localize speech sources using auditory subband signals. Note that some SSL approaches have also proposed that use a temporal sequence of STFT representations combined with a recurrent structures to utilize the context information [40], [41]. However, since the context frames in our proposed method are concatenated in the subband signal samples anyway, using the convolutional learning, the temporal pattern of context information is available also for detection and localization.

In this section, the expected output used in the labeled data, the proposed residual-dense learning, and the architecture and hyper parameters of the neural network model proposed for joint speech source detection and localization are described.

A. Expected Output

In supervised learning-based methods, neural networks are trained by using well-labeled data based on the pairs of input feature and expected output. The specific training set we used will be detailed in Section IV where it will be explained how target speech sources and interfering noise sources are placed within a reverberant environment. In the proposed method, the expected output is an array with the size of 360 (i.e., the number of azimuth angles), and represents the likelihood that a speech source is present at a specific azimuth position. In the absence of speech sources (i.e., only nonspeech sources are found in the environment), the expected output is an all-zero array because no speech source should be detected. In the presence of speech sources (i.e., speech and nonspeech sources are found in the environment), the expected output is defined as a likelihood representation of speech source presence across different azimuth

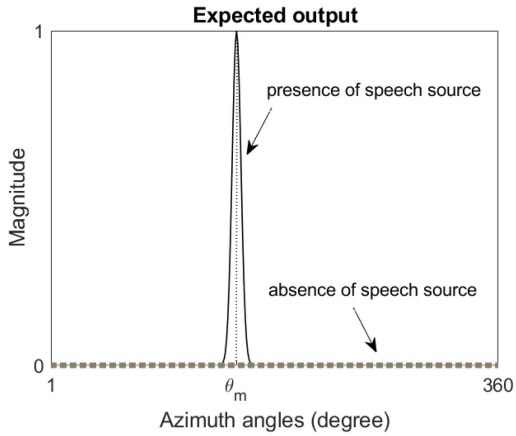

Fig. 3. Expected output for joint SSL and SSD. In the presence of speech source, the target output is the likelihood of azimuth angles around the target azimuth angle θ_m, and in the absence of speech source, the target output is an all-zero array.

angles by using a Gaussian function. The mean value of this Gaussian function is at the true azimuth angle of speech source:

$$\phi_m\ (\theta) = \begin{cases} 0 & ;\text{absence of speech source} \\ \exp\left[-(\theta-\theta_m)^2/\sigma^2\right] & ;\text{presence of speech source} \end{cases};$$
$$\theta\ =\ 1,\ \ldots,\ 360, \tag{1}$$

here $\phi_m(\theta)$ is the expected output at azimuth angle θ, for the mth time frame of a speech source positioned at an azimuth angle θ_m. Furthermore, σ^2 is the variance of the Gaussian function which determines the width of the curve around the target angle, as shown in Fig. 3. In our training, the parameter σ is set to 5° which corresponds approximately to human spatial resolution in moderately reverberant environments.

B. Residual-Dense Learning

Neural networks utilizing relatively raw data as inputs (e.g., TF components of binaural signals) may need more deep layers to extract optimal representations of the inputs. However, problems, such as gradient vanishing and overfitting, make deeper neural networks difficult to train. The residual connections have been proposed to deal with these problems also in the context of DNN-based SSL methods [22], [23]. One of the methods for implementing a residual block in CNNs is to use an identity mapping between the inputs and outputs of a stack of three convolutional layers (i.e., a convolutional layer with a kernel size of 3×3 is used between two convolutional layers with a kernel size of 1×1, maintaining the same number of channels). In neural networks, this compacts extracted representations to best fit in the available space. Residual links provide an identity mapping by adding the inputs of the stacked layers to the outputs.

Although utilizing the residual blocks in deeper layers can improve the convergence speed during the training stage, the spatial information still represented in earlier shallower layers may be lost after in case multiple summations with deeper layer outputs. Recently, dense aggregations have been proposed to deal with this problem in deep learning-based image processing methods [42], [43]. The dense aggregations utilize the identity mapping to concatenate the inputs of earlier shallow layers with the outputs after several deeper layers and allow layers to explore a larger set of features during training. In this study, we apply a combination of residual link and dense aggregation within three-layer stacks, as illustrated in Fig. 4. The dense aggregation can preserve the spatial information of the earlier shallow layer despite the multiple summations applied by the residual blocks. The kernel size of the middle layer is adopted to be 8×5 based on the dimension of the frequency bands and time samples at the input of the stacked layers. Five residual links are used in each residual-dense block. In this study, two neural network models are trained for detecting and localizing speech sources by using the residual and residual dense blocks in deep layers to compare the performance of the conventional residual block and the proposed residual dense block in the training phase.

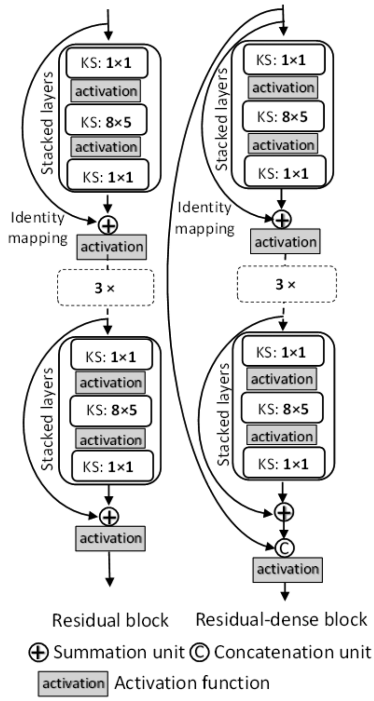

Fig. 4. Residual block and residual-dense block utilized by the proposed neural networks. The dense aggregation can preserve the spatial information of the input layer despite the multiple summations applied by the residual links.

C. Model Architecture

In this study, a deep regression model based on convolutional neural networks is used to create a complex mapping between the 3D inputs containing the auditory subband signals and corresponding expected outputs to simultaneously detect speech sources and predict the azimuth angle of available speech sources.

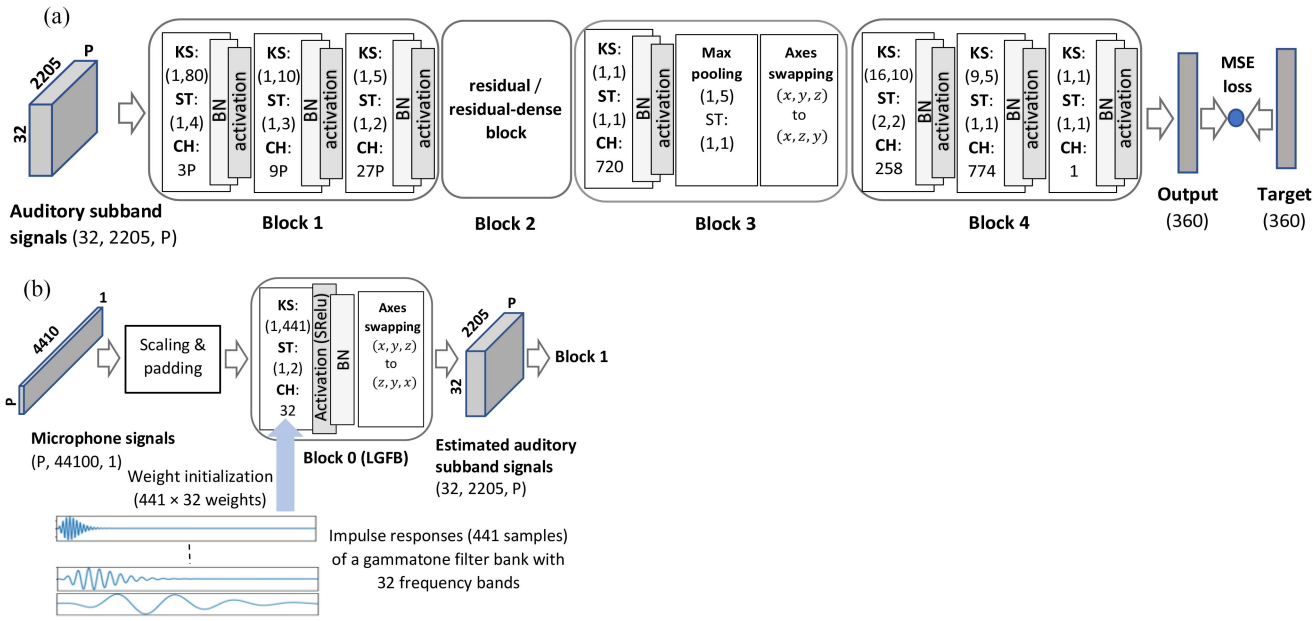

Fig. 5. (a) Architecture of the proposed neural network model for joint SSL and SSD using plain auditory subband signals. (b) Learnable gammatone filter bank (LGFB) for estimating auditory subband signals from microphone signals. The parameters of the convolutional layers are abbreviated as follows, **KS**: Kernel size, **ST**: Stride, and **CH**: Channels. **BN**, **activation**, and parameter **P** are, respectively, the batch normalization block, Swish activation function, and the number of microphone channels.

The mean square error (MSE) is used as the loss function between the expected and predicted outputs over batches:

$$MSE = \frac{1}{360B} \sum_{m=1}^{B} \sum_{\theta=1}^{360} \left[\phi_m(\theta) - \hat{\phi}_m^{\boldsymbol{w}}(\theta) \right]^2, \quad (2)$$

where $\boldsymbol{w}$ denotes the parameters of the neural network (i.e., weights and biases), which should be learned on the basis of the gradient of the loss function using backpropagation learning over batches during the training stage. $\hat{\phi}_m^{\boldsymbol{w}}(\theta)$ is the predicted output of the model at the mth time frame, with $\boldsymbol{w}$ representing the current model parameter set, and B is the batch size.

The architecture of the neural network model consists of four blocks, as shown in Fig. 5(a). Each block has a specific contribution in enabling the complex mapping of the auditory subband signals onto the azimuth dependent likelihood function defined in (1). Briefly, in Block 1, the optimal spatial representations of auditory subband signals are estimated at each frequency band. Block 2 improves the gradient flow in deeper layers (i.e., the parameters of the deeper layers in Block 1 can be updated based on the true gradient of the loss function.). Block 3 is used for removing the samples with less spatial information, and changing the convolution domain in Block 4. In the last block, the likelihood of azimuth angle is estimated by convolving the spatial representations across the frequency bands. The blocks are described in detail in the following.

In Block 1, three convolutional layers with the kernel size of 1×80, 1×10, and 1×5 are applied on time samples to estimate the optimal representations of the auditory subband signals throughout the expanded channels 3P, 9P, and 27P, where P is the number of channels in the microphone array. The dimension of time samples is reduced throughout the convolutional layers by using strided convolutions 1×4, 1×3, and 1×2.

In the next block, five three-layer stacks with residual/residual-dense mapping are used to improve gradient flow and prevent the vanishing gradient problem in deeper layers during backpropagation. The number of channels of the layers is 27P (i.e., the same as the number of channels of the last layer of Block 1). The output dimensions of the residual and residual-dense blocks are (32, 86, 27P) and (32, 86, 54P), respectively. Two models are designed and trained using the residual and residual-dense blocks described in details in Section *B*.

In Block 3, a 1×1 convolutional layer with 720 channels (i.e., twice the number of the azimuth angles) is applied to extract the optimal representations of previous layer throughout the azimuth angles, and then a max pooling layer with the size of 1×5 is applied on time samples to pool the most prominent samples with more energy. An axis swapping is used to change the convolution domain from frequency sample to frequency angle in the rest of the model [23].

In the last block, 2D convolutions are applied on the frequency and angle axes. A convolutional layer with the kernel size of 16×10 (i.e., 16 on frequency and 10 on angle domain), stride of 2×2, and the channel number of 258 is applied to create an output with the dimension of (9, 360, 258). A convolutional layer with 9×5 kernel size and 1×1 stride is applied to estimate the target azimuth angle based on the representations extracted across the frequency angle domain with filters having 774 expanded channels. The convolutional layers are applied by using circular padding along the azimuth angles (i.e., The azimuth angle 359° is connected with 0° by using circular padding). In

the last layer, a 1×1 convolutional layer with a single channel is applied to create the output array on the basis of the likelihood of the target azimuth angle. The batch normalization [44] and Swish activation function are applied after all convolutional layers. Note that the Swish activation function, $x/(1+e^{-x})$, recently proposed by the Google Brain Team [45] has shown better performance in deep learning-based image processing models [46], [47], [48], [49] compared to the activation functions conventionally used in DNN-based image and signal processing methods, such as the rectified linear unit (ReLU), $\max(0,x)$, and Sigmoid, $1/(1+e^{-x})$. The unboundedness in positive values and the nonmonotonicity of the Swish activation function lead to a better convergence speed and smoother gradients in deeper layers.

In addition to our proposed model for SSL and SSD which uses a fixed gammatone filter bank, another model was trained which uses a learnable gammatone filter bank (LGFB). As shown in Fig. 5(b), a convolutional layer with 32 channels (i.e., filters) is used in Block 0 to be able to model a gammatone filter bank with 32 filters that can be trained during the learning stage. It replaces the fixed peripheral preprocessing. The weights of the channels are initialized based on the impulse responses of the gammatone filters that were used in the fixed peripheral preprocessing. The kernel size (i.e., convolution size) applied on the microphone signals is 442 (i.e., the size of the impulse responses), and the stride is adapted to 2 to perform a downsampling by a factor of 2. The square root of ReLU (SReLU) is used as an activation function to implement the half-wave rectification (i.e., the zero function at negative values in SReLU) and square-root compression (i.e., the square root of the identity function at positive values in SReLU). An axes swapping unit is applied on the layer output to reshape the estimated auditory subband signal to $(32, 2205, 6)$. The rest of the model is the same as Blocks 1 to 4 shown in Fig. 5. The raw microphone signals (i.e., current frame and four preceding context frames) are scaled between -1 and 1, padded with zeros, and then used as input to the model in order to predict the azimuth angles of speech sources.

IV. Expriments

The purpose of the experiments was to evaluate performance of the proposed method in detecting and localizing speech sources in various acoustic environments. In order to perform the training, room simulations were made with the RAZR method [50] which allows to simulate the acoustics within a specified room for a specified sound source and listener. By providing the HRTFs of the hearing aid microphones, the actual signals received by the microphones can be simulated for arbitrary multi-source acoustic scenes. This allows for an extendable set of training data, and for creating test stimuli to evaluate performance of the proposed method.

A. Microphone Arrays and Dataset Preparation

The BKwHA HRTF dataset [51], recorded on Brüel & Kjær HATS, was used to simulate the head-related transfer functions of the microphone channels in an 8-channel behind-the-ear (BTE) HA [52]. Fig. 6 demonstrates the positions of the microphone channels at the right ear in the BTE HA. Six microphone channels behind the two ears are used for the binaural configuration, and four microphone channels (three behind, and one in the right ear) are used for the monaural configuration to capture the audio signals in a simulated noisy and reverberant environment.

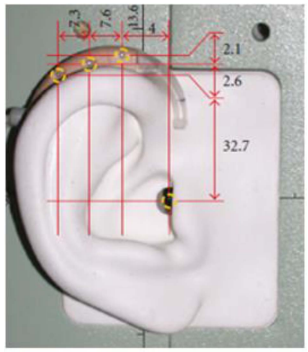

Fig. 6. The microphone positions of the right ear in the hearing aid. The distances between the microphones are given in mm [52].

The impulse responses of an acoustic room are simulated by using the perceptually plausible room acoustics simulator (RAZR [50]). In the version of RAZR used in this research, all sound sources are omnidirectional and increase the complexity of the reverberation conditions in the acoustic environment. The dimensions of the room are 5.4 m length, 7.1 m width, and 3 m height, and the head is positioned near the room center 2.7, 3.5, and 1.7 m. For a robust localization performance, sound sources are randomly located between 1 and 1.5 m from the head position at random azimuth angles between 1° and 360° with 5° resolution. A total of 2000 utterances from 200 different speakers (male and female) from the TIMIT database [53] and around 2.5 h nonspeech signals recorded in real rooms and public places (e.g., babble, restaurant, café, air conditioner, fan, kitchen, room ambience, office, etc.) [54] are used as speech sources and noise sources to prepare the training dataset at random signal-to-noise ratios (SNRs) from -6 dB to 30 dB with steps of 3 dB. Here, the spatial anechoic SNRs are used, determined at the center of the head. Because sound sources have different distances from the head position it is calculated according to:

$$SNR = 10\log_{10}\left\{\frac{[\text{rms}\,(speech)]^2}{[\text{rms}\,(noise)]^2}\frac{d_n^2}{d_s^2}\right\}, \tag{3}$$

where d_n and d_s are the distance of the noise source and speech source from the center of the head, respectively, and rms represents the root mean square. Three reverberation times, T60, are randomly adopted from the range of 0.01 s to 0.81 s with intervals of 0.2 s. T60 is varied by adjusting absorption coefficients for all surfaces at five octave band center frequencies 250, 500, 1000, 2000, 3000, and 4000 Hz. None or one speech source, and one noise source are simultaneously present in the acoustic room at random azimuth angles. The presence or absence of speech source is randomly selected during the preparation of the training dataset. The silent regions at the beginning and end of speech signals are removed. The training dataset is composed of 32400

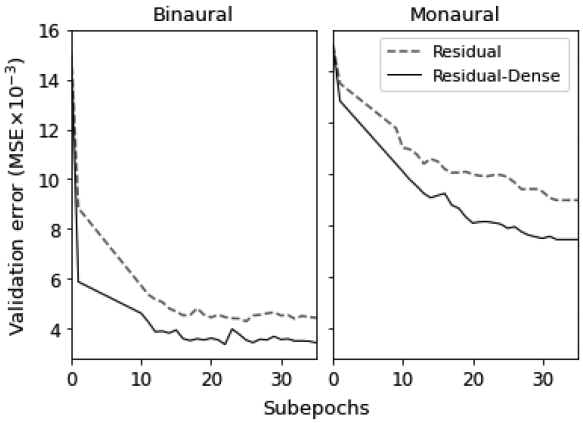

Fig. 7. Validation error (MSE) curves for the models employing residual and residual-dense blocks for binaural and monaural microphone array.

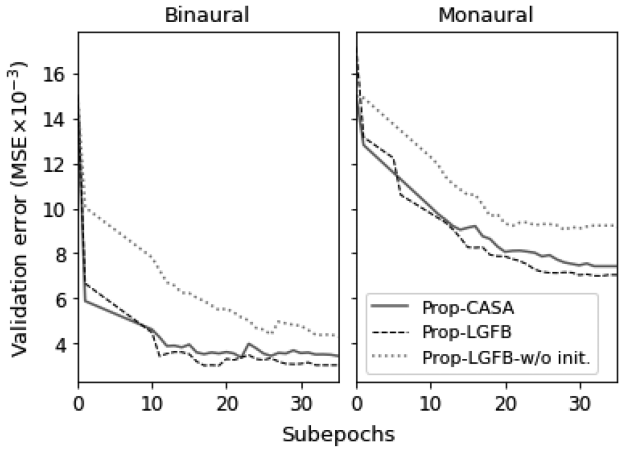

Fig. 8. Validation error curves for the model using the plain auditory subband signals (Prop-CASA, solid line), and the models with a learnable gammatone filter bank with initialization (Prop-LGFB, dashed line) and without initialization (Prop-LFGB-w/o init., dotted line).

different random combinations of positions, and each contains three utterances. The duration of the training dataset is around 59 h. In the test dataset, 200 utterances from 20 different speakers (male and female) and the nonspeech sources [54], which are all not used in the training dataset, are used at 2500 random positions. The number of the signals with speech sources and no speech source present are nearly balanced in the test dataset.

B. Training Stage and Learning Curve

The Adam optimizer [55] is used as an adaptive optimizer using the backpropagation algorithm to update the model parameters during the training stage. The minibatch size B is 32, and the learning rate is set to 10^{-4} for all epochs in all neural networks. Each epoch required approximately 4 days using one GPU, Tesla P100, and three CPUs, Intel Xeon E5-2650, available in the CARL cluster of Oldenburg University.

1) Residual/Residual-Dense Learning: The validation error (MSE) curves for the models using residual and residual-dense blocks are shown in Fig. 7. The models converged after around 3 epochs. For a better comparison, validation loss is computed every tenth of an epoch which we will refer to as subepoch, and a smoothing [56] is applied on the curves to smooth the rapid changes in the validation errors in subepochs. As illustrated, the residual-dense block in which the spatial representations of the input layer are reused at the output layer of the block converges to a lower validation error and leads to a better convergence speed in comparison to the conventional residual block in the training stage, both in binaural and monaural microphone arrays. In the rest of the paper, the proposed binaural and monaural models will use the proposed residual-dense structure.

2) Learnable Gammatone Filterbank: Two binaural models and two monaural models with LGFBs have been trained with weight initialization based on the peripheral preprocessing parameters (denoted as Prop-LGFB) and without such weight initialization (denoted as Prop-LGFB-w/o init.), as shown in Fig. 5(b). The validation error curves for the models with LGFB and the models using the plain auditory subband signals calculated by the peripheral processing (Prop-CASA) are shown in Fig. 8. The models initialized with the peripheral preprocessing parameters (Prop-LGFB) converge to a slightly lower MSE as compared to the CASA-based methods using a fixed peripheral processing stage, for both binaural and monaural microphone arrays. However, when not using the peripheral preprocessing-based initialization (Prop-LGFB-w/o init.), results show a larger MSE after convergence compared to the CASA-based method.

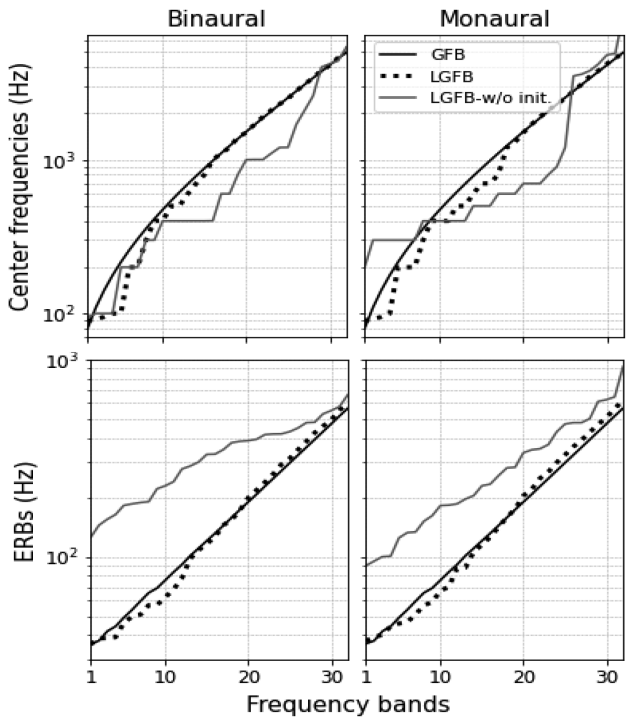

Fig. 9. Center frequencies and ERBs of the plain gammatone filter bank (GFB, solid black line), and a learnable gammatone filter bank trained with initialization (LGFB, dotted line) and without initialization (LGFB-w/o init. solid grey line).

Fig. 9 illustrates the center frequencies and the equivalent rectangular bandwidths (ERBs) of the plain gammatone filter

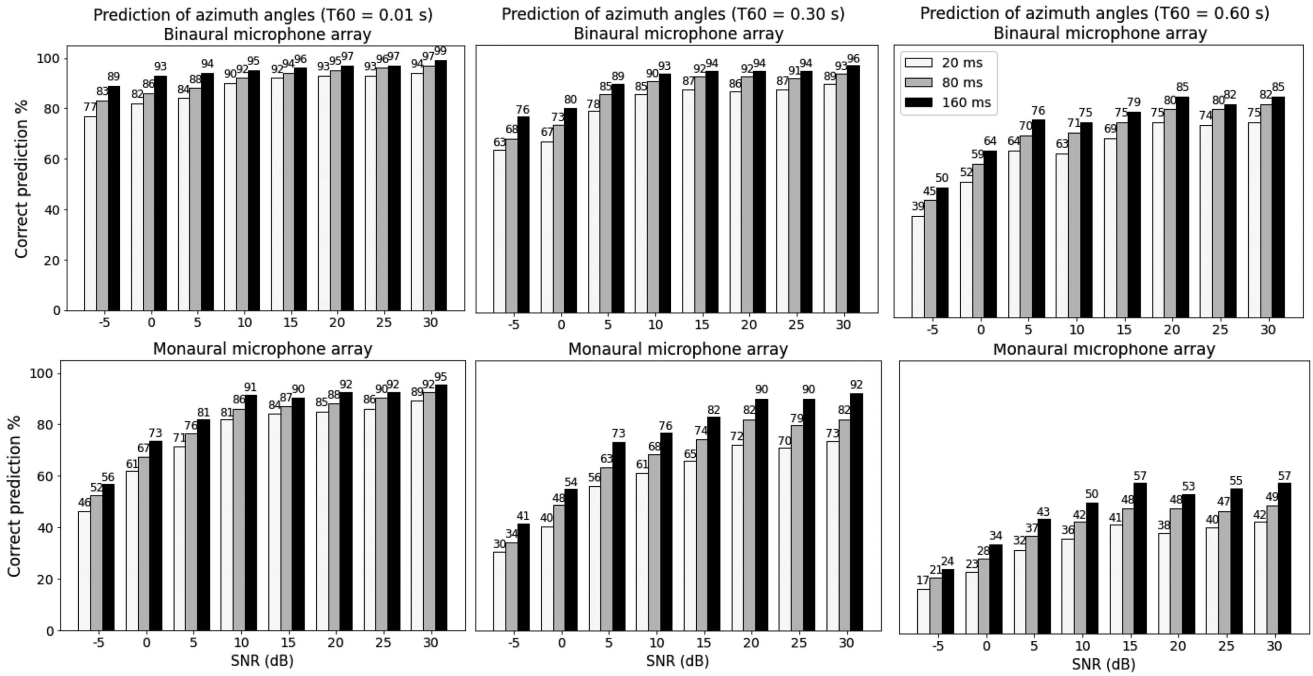

Fig. 10. Speech source localization scores for the proposed method using binaural and monaural microphone arrays in three time intervals, 20, 80, and 160 ms, at three reverberation times, T60 = 0.01, 0.30, and 0.60 s.

bank (GFB) and the learnable gammatone filter bank (LGFB) with and without weight initialization after convergence. In binaural and monaural models, the learnable gammatone filters, *with initialization*, converge to the center frequencies of the plain filters for center frequencies above 1000 Hz (i.e., frequency band 16), while the bandwidths increase somewhat. In lower frequency bands, the learnable filters converge to slightly lower center frequencies and narrower bandwidths in comparison to the plain gammatone filters. However, the filter learning *without initialization* diverges from the center frequencies of the plain gammatone filters specifically at frequencies between 400 and 3000 Hz, and attains very wider ERBs after convergence.

As shown in Fig. 8, the model with initialized LGFB slightly improves the validation MSE. It indicates that although the center frequencies of the plain gammatone filters at high frequency bands are efficient for speech source detection and localization, filters with somewhat wider bandwidths at high frequencies, and with slightly lower center frequencies and narrower bandwidths at low frequencies might be able to extract slightly more informative subband signals for detecting and localizing speech sources at adverse acoustic conditions. However, without weight initialization, the models using LGFB do not seem to converge to optimal center frequencies, specifically at higher frequencies.

C. Results

The performance of the trained neural networks in SSL and SSD is jointly evaluated based on the predicted outputs. A speech source is detected in mth frame if one of the values of the predicted output is above a threshold, δ. Otherwise, the frame is assumed to contain no speech source. Once a speech source is detected, its location is estimated by finding the argument of the maximum of the predicted output. The models is assumed to correctly estimate the azimuth angle of the detected speech source if the predicted azimuth angle does not differ more than 2° from the true azimuth angle. In addition, when the model detects a speech source while no speech source was present, or detects a nonspeech source while a speech source was present, this is counted as an incorrect response. The thresholding parameter, δ, is set to 0.2 for the two models using binaural and monaural microphone arrays. The trained models are chosen for evaluation based on the least validation error in the training stage.

Fig. 10 shows the speech source localization scores (i.e., percentage of correct localization) achieved by the proposed model using the plain auditory subband signals (i.e., calculated using the plain gammatone filters in the peripheral preprocessing stage) using the binaural (upper panels) and monaural (lower panels) microphone arrays for different SNRs and different reverberation times (panels from left to right). For each SNR, three localization scores are shown, calculated by averaging the estimated outputs across time intervals of 20 ms, 80 ms, and 160 ms, which corresponds to averaging across one, four, and eight successive time frames. As can be seen, the SSL performance of the models improves as the time interval duration increase. The model using the binaural microphone array outperforms the monaural microphone array in all noisy and reverberant conditions which is expected because the binaural positions of the microphones lead to larger time delays and level differences between the microphone signals compared to the monaural microphone positions.

As illustrated in Fig. 10, the model using the monaural microphone array still achieves a good localization performance at

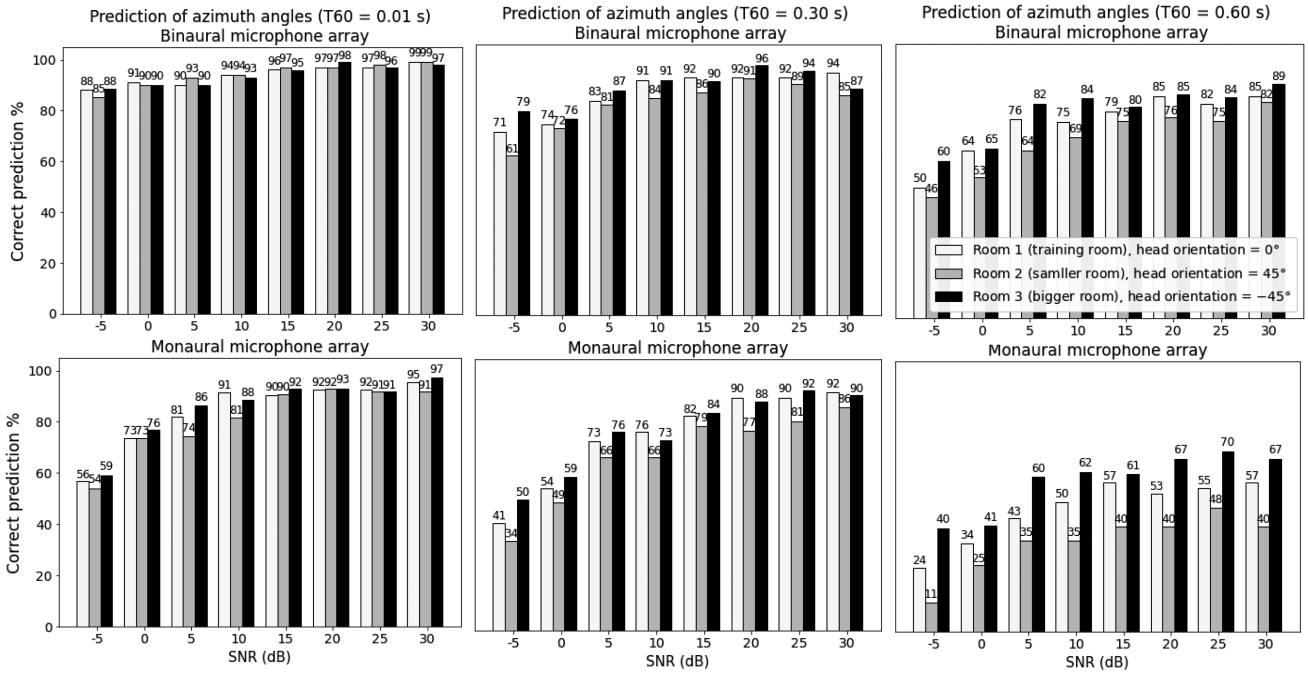

Fig. 11. Speech source localization scores for the proposed methods using binaural and monaural microphone arrays in three reverberant rooms and three head orientations. Room 1 where the head orientation is 0° has been used in the training stage, room 2 is a smaller room with a head orientation of 45°, and room 3 is a bigger room with a head orientation of -45°. Room 2 and 3 are not used in the training dataset.

high SNRs and short reverberation times despite the extremely small interchannel cues that occur in such a monaural microphone array due to the short distances between the microphones. It appears, the proposed neural network is still able to extract and exploit the extremely small interchannel cues from the monaural signals to a sufficient degree to be able to localizing speech sources in noisy and reverberant conditions. This achievement is relevant because it enables speech source localization for each single HAs without the need to send audio signals between the hearing aids.

The results show that the performance of the models is reduced by increasing the reverberation times and decreasing SNRs. The uncertainty of the interchannel cues caused by higher reverberation times and the masked cues at lower SNRs impairs the spatial information available for localizing speech sources, especially in monaural microphones.

In order to further assess the generalization of the proposed method in unseen acoustic conditions, we also evaluated the localization performance in two other acoustic rooms with different dimensions and different head orientations, which have not been seen during the training stage. The dimension of the acoustic room used in the training dataset (i.e., Room 1) was [5.4 m length, 7.1 m width, 3 m height] and the head orientation was 0°. Two test datasets were created using a smaller room (i.e., Room 2 with a dimension of [4.25 m length, 5.2 m width, 2.5 m height] and a head orientation of 45°) and a bigger room (i.e., Room 3 with a dimension of [12 m length, 10.3 m width, 3.2 m height] and head orientation of -45°). As illustrated in Fig. 11, the models show very similar localization performance in the three acoustic rooms at shorter reverberation times (left panels). However, at longer reverberation times, the models show lower and higher localization scores, in the smaller Room 2, and bigger Room 3, respectively, compared to Room 1. We assume, the differences between the localization scores between the different rooms are caused by the different critical distances of the rooms (i.e., reverberation radius). The critical distance in a reverberant rooms has a direct relationship with the reverberation time and an inverse relationship with the room dimensions [57]. When a listener is further away from a sound source than the critical distance, the level of the direct sound is lower than the reflected sound. In this case, the reflected sounds, having quasi random directions of arrival, will have more energy than the direct sound, which can be expected to reduce the localization performance. The shorter critical distance in the smaller Room 2 is therefore expected to more negatively affect the localization performance compared to the bigger Room 3 having a larger critical radius. Since our proposed method has shown good generalization in unseen acoustic conditions, in the remainder of this paper we will use Room 1 (i.e., with a medium dimension compared to Room 2 and Room 3) for comparing the performance of the proposed method against a number of baseline methods.

The performance of the proposed models, using different front ends, will now be evaluated in more detail; i.e., of using learnable versus plain auditory signals and in addition to compare to performance using STFT coefficients. Also, three methods from literature will be evaluated on the same data set. The evaluated methods are listed below.

Proposed methods (Prop-CASA and Prop-LGFB): we evaluated the performance of the primary methods proposed in this study in SSL and SSD. The first method detects and localizes speech sources based on the plain auditory signals calculated in the peripheral preprocessing stage (i.e., the model using

the CASA-based representations of microphone signals) (Prop-CASA). The second method utilizes the learnable gammatone filter bank with weight initialization as shown in Fig. 5 using the unprocessed microphone signals as input (Prop-LGFB).

Proposed model using STFT coefficients (Prop-STFT): to facilitate a proper comparison between the performance of methods using CASA-based and STFT-based TF representations, regardless the influence of neural network models, we also trained the model, proposed for localization based on the plain auditory subband signals, using STFT coefficients (Prop-STFT). We used the real and imaginary parts of the STFT as the model input for both the binaural and monaural microphone arrays in a similar way as in [23]. We edited some hyperparameters of the convolutional layers (e.g., kernel sizes and strides) to match the dimensions of the STFT-based inputs.

DNN-based methods using STFT coefficients (Spec-CL and STFT-RG): two deep learning-based models, which also use STFT-based TF representations as model inputs, are used to serve as baseline methods available from literature. The methods have reported good performance in joint SSL and SSD by using coplanar/noncoplanar microphone arrays located around the head of robot devices. Firstly, in [22], a classification model based on the spectrum power (Spec-CL) was proposed for joint SSL and SSD in an environment–robot interaction. In that method, a STFT-based preprocessing stage was applied on the signals received from a 16-channel microphone array located around a robot device to estimate the power of the TF components. Then a CNN-based classification model with cross entropy loss was used to estimate the probability density of the azimuth angles and the binary detection of speech sources in the output based on the power spectrum of the input signals. Secondly, a STFT-based regression model (STFT-RG) was proposed in [23] to detect and localize speech sources using a rectangular four-channel microphone array placed on a robot head. The model inputs contain the real and imaginary parts of the STFT components of each microphone signal. A CNN-based regression model with MSE loss was trained to estimate the labeled outputs. In the training we used, the labeled outputs corresponded to 360 speech source azimuth angles that could occur.

Conventional feature (GCC-PHAT)-MLP: a localization approach was also evaluated which is based on GCC-PHAT coefficients. This is a conventional feature for SSL in both HAs and robot devices. GCC-PHAT is a temporal feature used for estimating the overall time of arrival delay between microphone channels. GCC-PHAT coefficients are calculated using the spectrally-weighted wideband cross correlation derived from Fourier transforms of microphone signals for different time delays. The time delay of the speech source is estimated by finding the peak position in the GCC-PHAT coefficients. In this evaluation, for the binaural array, GCC-PHAT coefficients were calculated between the three microphone pairs on the left and on the right HAs, and between two corresponding microphone pairs on the front left/right and on the back right/left HAs (i.e., totally five pairs). In the monaural microphone array, the GCC-PHAT coefficients between the three behind-the-ear microphones and the in-the-ear microphone (i.e., totally three pairs) are calculated as the localization features. A regression model based on a shallow neural network [i.e., a multi-layer perceptron (MLP) with 3 hidden layers] was utilized to create a mapping between the GCC-PHAT coefficients and the corresponding azimuth angles, similar to the method used in [16]. This localization method has shown good performance in localizing speech sources compared to other conventional methods [16].

For all models that are evaluated, the hyperparameters (e.g., the number of the channels, and kernel and stride sizes of the convolutional layers) have been adapted to achieve the best performance on the available datasets.

1) Speech Source Localization Performance: In this evaluation, localization scores are determined as a function of SNR for three different reverberation times. The localization scores of all models are estimated over 180 ms time intervals. As shown in Fig. 12, Prop-CASA and Prop-LGFB show significantly better localization performance for both the binaural and monaural microphone arrays compared to the neural network models using the STFT components (Prop-STFT and STRF-RG), especially so at lower SNRs and longer reverberation times. In addition, the models using LGFB seem to show slightly improved localization scores at lower SNRs in comparison to the model using the plain the plain auditory subband signals (Prop-CASA).

The performance gap between the models using the CASA-based and STFT-based representations increases in the monaural microphone array, where the interchannel cues are extremely small. This finding indicates that the proposed model, using the CASA-based representations can much better estimate and exploit the much smaller interchannel cues that are present in the auditory subband signals for SSL than the models that use the STFT components. The results also show that the proposed model based on the STFT components (Prop-STFT) outperforms the regression-based baseline model using the same inputs (STFT-RG), especially for the binaural microphone array. Results indicate that the proposed neural network model (i.e., the model architecture, residual-dense learning, Swish activation, etc.) can better estimate the localization representations from the STFT components and exploit them for localization, in comparison with the deep regression model proposed in STFT-RG.

The results in Fig. 12 also show that the method using the conventional GCC-PHAT feature can well localize speech sources by using the binaural microphone array at higher SNRs and shorter reverberation times. However, the performance of the (GCC-PHAT)-MLP method dramatically decreases at lower SNRs and longer reverberation times, especially when using the monaural microphone array. GCC-PHAT is a temporal feature that only contains the information of the overall time delays between the microphone channels (i.e., it equally sums over all frequencies). Therefore, when the peak regions of the GCC-PHAT coefficients are corrupted by noise and reverberation, the true time delays cannot be estimated by the model. When, however, spatial features occurring within auditory bands are maintained, such as is the case for the CASA-based features, and for the STFT features, the neural network still has information available to estimate the speech source location. Also, the low localization scores obtained by the (GCC-PHAT)-MLP method in monaural microphone array show that the very small time

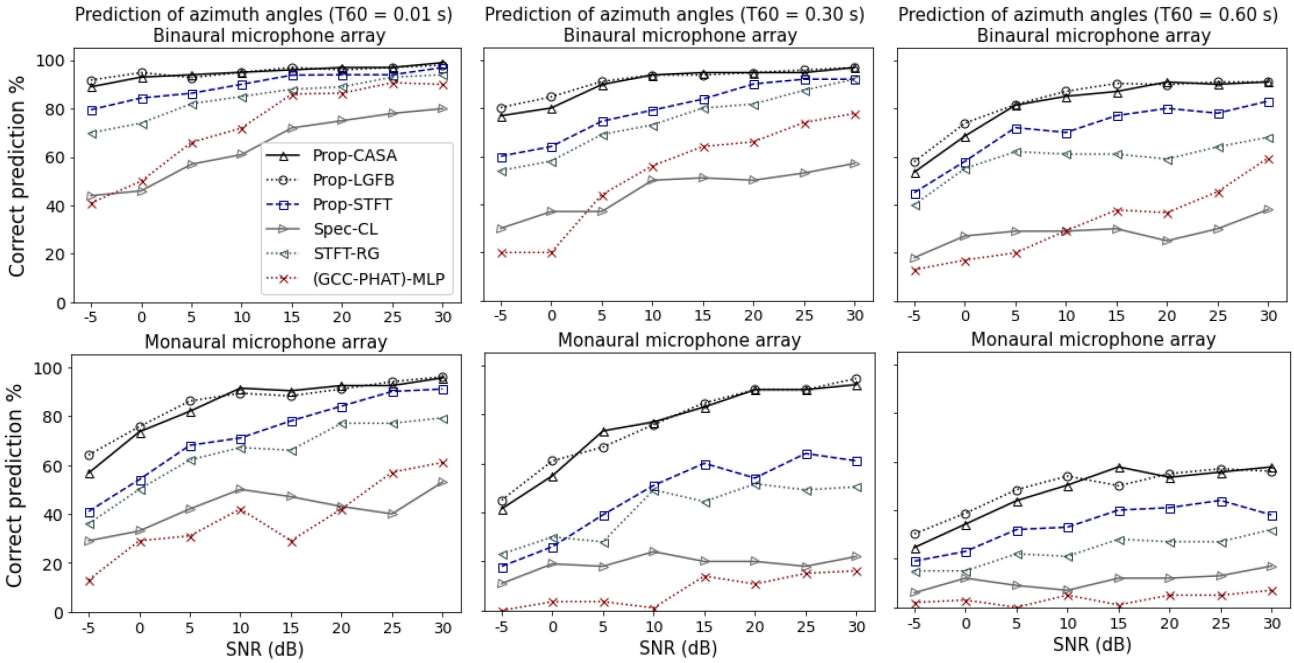

Fig. 12. Comparing the performance of the proposed models using CASA-based inputs (Prop) and STFT-based inputs (Prop-STFT), and the baseline methods [Spec-CL, STFT-RG, and (GCC-PHAT)-MLP] in speech source localization by using the binaural and monaural microphone arrays at three reverberation times, T60 = 0.01, 0.30, and 0.60 s.

delays detected by the GCC-PHAT coefficients between the monaural microphones may not be sufficient for localization.

The results in Fig. 12 also show that the exclusive use of the power spectrum of the microphone signals (Spec-CL) is insufficient to localize speech sources when using the spatial microphone configuration of a HA. In Spec-CL, the model cannot extract the ICTD representations from the input TF representations throughout the deep layers because the power spectrum does not have the information of the time delays between the microphone signals. Utilizing the power spectrum, which contains only the information of the level differences, is insufficient for SSL based on the microphone arrays of HAs, especially at longer reverberation time and lower SNRs. Accordingly, Spec-CL has shown the worst performance in SSL in comparison to the other DNN-based methods.

Contrary to the Spec-CL method, the Prop-STFT and STFT-RG methods use the real and imaginary parts of the Fourier transform of the microphone signals. The ICTD and ICLD representations can in principle be extracted and exploited by the neural networks because the phase and magnitude of the complex values of Fourier transform contains the information of time and level differences between the microphone signals. When using Spec-CL in adverse acoustic conditions, interchannel level differences at high frequencies may not be reliable due to the masked speech sources, and in addition, the ICLDs are expected to be near to zero at lower frequencies. In that case, exploiting the ICTD representations would be essential for being able to localizing speech sources. Accordingly, Prop-STFT and STFT-RG can achieve better performance than Spec-CL, on the one hand, which uses only the power spectrum of the microphone signals and also compared to (GCC-PHAT)-MLP, on the other hand, which only uses the time delay patterns between the microphone channels for localization.

The results show that although the baseline methods using STFT-based components can adequately estimate the interchannel representations of the microphone channels placed around the robot body for localizing speech sources, these methods perform relatively poorly in microphone arrays in HAs because microphones are positioned in two groups of very closely spaced microphones.

Note that the sizes of the models using the raw microphone signals and auditory subband signals as inputs are bigger than the model using the STFT components (i.e., almost twice) because extracting localization information from raw data requires a deeper neural network to learn relevant signal features.

2) Speech Source Detection Performance: In Table I, the performance in terms of speech source detection is shown for the proposed methods (i.e., Prob-CASA, Prop-LGFB, and Prob-STFT) and the three baseline methods both for binaural and monaural microphone arrays. The average of the recall, precision, and accuracy scores among the SNR range are shown for three different reverberation times. As can be seen in the table, the proposed models using the CASA-based TF representations can well detect speech sources in case a speech sources is present (i.e., the recall scores) and in case speech and nonspeech sources are present (i.e., the precision and accuracy scores) when using the binaural microphone array at different reverberation times. The scores are reduced for the proposed methods with increasing reverberation times. Although the proposed model using STFT components (Prop-STFT) and the two baseline methods (STFT-RG and Spec-CL) show similar recall scores compared to the models using the CASA-based representations

TABLE I
THE AVERAGE SCORES OF SPEECH SOURCE DETECTION OBTAINED BY THE PROPOSED METHOD AND BASELINE APPROACHES USING BINAURAL AND MONAURAL MICROPHONE ARRAY AT THREE REVERBERATION TIMES, T60 = 0.01, 0.30, AND 0.60 S

Method	Relevancy	Reverberation time (T60)					
		Binaural			Monaural		
		0.01 s	0.30 s	0.60 s	0.01 s	0.30 s	0.60 s
Prop-CASA	Recall	0.99	0.93	0.87	0.95	0.83	0.77
	Precision	0.98	0.96	0.90	0.92	0.86	0.80
	Accuracy	0.97	0.92	0.88	0.91	0.87	0.79
Prop-LGFB	Recall	0.97	0.93	0.85	0.96	0.89	0.80
	Precision	0.98	0.98	0.97	0.98	0.97	0.96
	Accuracy	0.95	0.92	0.87	0.93	0.88	0.81
Prop-STFT	Recall	0.97	0.91	0.82	0.91	0.79	0.66
	Precision	0.95	0.87	0.79	0.65	0.79	0.77
	Accuracy	0.96	0.89	0.80	0.70	0.79	0.73
Spec-CL	Recall	0.99	0.98	0.96	0.71	0.50	0.44
	Precision	0.76	0.65	0.58	0.88	0.93	0.86
	Accuracy	0.80	0.70	0.64	0.80	0.73	0.69
STFT-RG	Recall	0.98	0.95	0.88	0.91	0.72	0.58
	Precision	0.93	0.89	0.78	0.88	0.87	0.68
	Accuracy	0.95	0.91	0.82	0.85	0.80	0.62
(GCC-PHAT)-MLP	Recall	0.84	0.61	0.39	0.76	0.26	0.13
	Precision	0.78	0.76	0.74	0.54	0.52	0.49
	Accuracy	0.77	0.71	0.64	0.55	0.51	0.30

for the binaural microphone array, they cannot detect speech sources in the presence of speech and nonspeech sources as well as Prop-CASA and Prop-LGFB across the different acoustic conditions (i.e., they achieve lower precision scores).

In addition, when using the monaural microphone array, the proposed methods using the CASA-based inputs outperforms the models using the STFT components for all reverberation times. (GCC-PHAT)-MLP shows the worst detection scores in comparison to the other methods especially in monaural microphone arrays. It indicates that the model using GCC-PHAT coefficients cannot properly differentiate between the wideband cross correlations of speech and nonspeech sources in order to perform speech source detection.

V. CONCLUSION

A deep learning-based method was proposed for jointly detecting and localizing speech sources in hearing aids using auditory subband signals of microphone channels. Motivated by the human auditory system, a peripheral preprocessing was applied on the signals received from a binaural set of hearing aid microphones to determine auditory subband signals. A deep regression model, based on the convolution neural network was trained to create a complex mapping between the auditory subband signals and a 360° range of azimuth angles at which speech sources could be positioned. In addition, another neural network was trained using only a monaural microphone array (i.e., four channels of a single hearing aid). The models were based on a combination of residual and dense aggregation mapping to preserve and reuse the spatial representations of the earlier shallow layers at the deep layers. This helps to improve the gradient flow, in the training stage. Indeed, the learning curves showed that the proposed residual-dense block does improve the speed and accuracy of the model convergence during the training stage compared to using only the residual blocks conventionally used.

Surprisingly, the proposed neural network model, using the monaural microphone array achieves prediction scores over 90% in joint speech source detection and localization at high SNRs and low reverberation times despite the short distances between the microphone channels for which only extremely low time delays and level differences can occur.

Replacing the *plane* gammatone filters in the peripheral front end with learnable filters that were initialized with the impulse responses of the plain gammatone filters showed slightly improved localization scores at lower SNRs compared to using the CASA-based auditory preprocessing. The learnable filters converged to filters with slightly wider bandwidths at frequencies higher than 1000 Hz, while the center frequencies for low-frequency bands were lower than the plain filters.

The performance of the proposed method was also compared to two baseline DNN-based methods using the STFT components of all microphone signals, and was also compared to a shallow model utilizing the conventional GCC-PHAT feature. Results demonstrated that the proposed method, trained based on the auditory subband signals significantly outperforms the model utilizing the STFT components, and also outperformed the three baseline methods in terms of joint detection and localization for both the binaural and monaural microphone arrays. Note that the sizes of neural networks using raw microphone signals and/or processed subband signals as inputs are larger than the models utilizing STFT components. These large networks are necessary to be able to extract all information present in the raw input signals needed to detect and localize speech sources.

ACKNOWLEDGMENT

The authors would like to thank Dr. Stephan Ewert for providing the HRIRs of the DUDA head used in Fig. 2 in this paper. In addition, we would like to thank the reviewers and editor for their very helpful comments that helped improve the paper.

REFERENCES

[1] W. Noble and G. Stuart, "Effects of bilateral versus unilateral hearing aid fitting on abilities measured by the speech, spatial, and qualities of hearing scale (SSQ)," *Int. J. Audiol.*, vol. 45, no. 3, pp. 172–181, 2006.
[2] A. Xenaki, J. Bünsow Boldt, and M. Græsbøll Christensen, "Sound source localization and speech enhancement with sparse Bayesian learning beamforming," *J. Acoust. Soc. Amer.*, vol. 143, no. 6, pp. 3912–3921, 2018.
[3] Z. Q. Wang and D. Wang, "All-neural multi-channel speech enhancement," in *Proc. Interspeech*, 2018, pp. 3234–3238.
[4] 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.
[5] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin, Germany: Springer, 2013.

[6] A. W. Archer-Boyd, W. M. Whitmer, W. O. Brimijoin, and J. J. Soraghan, "Biomimetic direction of arrival estimation for resolving front-back confusions in hearing aids," *J. Acoust. Soc. Amer.*, vol. 137, no. 5, pp. EL360–EL366, 2015.

[7] T. N. T. Nguyen, S. K. Zhao, and D. L. Jones, "Robust DOA estimation of multiple speech sources," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 2287–2291.

[8] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[9] R. Venkatesan and A. B. Ganesh, "Full sound source localization of binaural signals," in *Proc. IEEE Int. Conf. Wireless Commun., Signal Process. Netw.*, 2017, pp. 213–217.

[10] M. Zohourian, G. Enzner, and R. Martin, "Binaural speaker localization integrated into an adaptive beamformer for hearing aids," *IEEE/Assoc. Comput. Machinery Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 3, pp. 515–528, Mar. 2018.

[11] G. A. Courtois, P. Marmaroli, H. Lissek, Y. Oesch, and W. Balande, "Development and assessment of a localization algorithm implemented in binaural hearing aids," in *Proc. IEEE 23rd Eur. Signal Process. Conf.*, 2105, pp. 2321–2325.

[12] T. May, S. van de Par, and A. Kohlrausch, "A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 7, pp. 2016–2030, Sep. 2012.

[13] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *J. Acoust. Soc. Amer.*, vol. 116, no. 5, pp. 3075–3089, 2004.

[14] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Amer.*, vol. 114, no. 4, pp. 2236–2252, 2003.

[15] T. May, S. van de Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 1–13, Jan. 2011.

[16] W. He, P. Motlicek, and J. M. Odobez, "Deep neural networks for multiple speaker detection and localization," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2018, pp. 74–79.

[17] R. Takeda and K. Komatani, "Discriminative multiple sound source localization based on deep neural networks using independent location model," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2016, pp. 603–609.

[18] C. Pang, L. Hong, and L. Xiaofei, "Multitask learning of time-frequency CNN for sound source localization," *IEEE Access*, vol. 7, pp. 40725–40737, 2019.

[19] Y. Yang, J. Xi, W. Zhang, and L. Zhang, "Full-sphere binaural sound source localization using multi-task neural network," in *Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2020, pp. 432–436.

[20] 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.

[21] F. Zhao, L. Ruwei, and P. Dongmei, "Deep learning for binaural sound source localization with low signal-to-noise ratio," *J. Phys.: Conf. Ser.*, vol. 1828, no. 1, 2021, Art. no. 12017.

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

[23] W. He, P. Motlicek, and J. M. Odobez, "Neural network adaptation and data augmentation for multi-speaker direction-of-arrival estimation," *IEEE/Assoc. Comput. Machinery Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1303–1317, 2021.

[24] T. N. T. Nguyen, W. S. Gan, R. Ranjan, and D. L. Jones, "Robust source counting and DOA estimation using spatial pseudo-spectrum and convolutional neural network," *IEEE/Assoc. Comput. Machinery Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2626–2637, 2020.

[25] V. Tourbabin and B. Rafaely, "Theoretical framework for the optimization of microphone array configuration for humanoid robot audition," *IEEE/Assoc. Comput. Machinery Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1803–1814, Dec. 2014.

[26] J. L. Lawson and G. E. Uhlenbeck, *Threshold Signals*. New York, NY, USA: McGraw-Hill, 1950.

[27] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform CLDNNs," in *Proc. Interspeech*, 2015, pp. 1–5.

[28] H. Park and C. D. Yoo, "CNN-based learnable gammatone filterbank and equal-loudness normalization for environmental sound classification," *IEEE Signal Process. Lett.*, vol. 27, pp. 411–415, 2020.

[29] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.

[30] Z. Zhu, Z. P. Bian, J. Hou, Y. Wang, and L. P. Chau, "When residual learning meets dense aggregation: Rethinking the aggregation of deep neural networks," 2020, *arXiv:2004.08796*.

[31] M. Nikzad, A. Nicolson, Y. Gao, J. Zhou, and K. K. Paliwal, "Deep residual-dense lattice network for speech enhancement," in *Proc. Assoc. Adv. Artif. Intell. Conf. Artif. Intell.*, 2020, pp. 8552–8559.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 630–645.

[33] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.

[34] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

[35] B. C. J. Moore, B. R. Glasberg, and T. Baer, "A model for the prediction of thresholds, loudness, and partial loudness," *J. Audio Eng. Soc.*, vol. 45, no. 4, pp. 224–240, 1997.

[36] A. G. Katsiamis, E. M. Drakakis, and R. F. Lyon, "Practical gammatone-like filters for auditory processing," *Eur. Assoc. Signal Process. J. Audio, Speech, Music Process.*, vol. 2007, pp. 1–15, 2007.

[37] A. Venkitaraman, A. Adiga, and C. S. Seelamantula, "Auditory-motivated Gammatone wavelet transform," *Signal Process.*, vol. 94, pp. 608–619, 2014.

[38] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, vol. 47, no. 1/2, pp. 103–138, 1990.

[39] C. P. Brown and R. O. Duda, "A structural model for binaural sound synthesis," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 5, pp. 476–488, Sep. 1998.

[40] L. Perotin, A. Défossez, E. Vincent, R. Serizel, and A. Guérin, "Regression versus classification for neural network based audio source localization," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2019, pp. 343–347.

[41] D. Krause, A. Politis, and K. Kowalczyk, "Comparison of convolution types in CNN-based feature extraction for sound source localization," in *Proc. IEEE 28th Eur. Signal Process. Conf.*, 2021, pp. 820–824.

[42] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2472–2481.

[43] X. Xu, M. Li, W. Sun, and M. H. Yang, "Learning spatial and spatio-temporal pixel aggregations for image and video denoising," *IEEE Trans. Image Process.*, vol. 29, pp. 7153–7165, 2020.

[44] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[45] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," in *Proc. Int. Conf. Learn. Representations (Workshop Track)*, 2018.

[46] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[47] L. Wang, Q. Li, and H. Guo, "A research on deep learning model for face emotion recognition based on Swish activation function," *J. Image Signal Process.*, vol. 8, no. 3, pp. 110–120, 2019.

[48] J. Zhu and Z. Chen, "Comparative analysis of various new activation functions based on convolutional neural network," *J. Phys.: Conf. Ser.*, vol. 1676, no. 1, 2020, Art. no. 12228.

[49] A. M. Alhassan and W. M. N. W. Zainon, "Brain tumor classification in magnetic resonance image using hard swish-based RELU activation function-convolutional neural network," *Neural Comput. Appl.*, vol. 33, pp. 9075–9087, Aug. 2021.

[50] T. Wendt, S. van de Par, and S. D. Ewert, "A computationally-efficient and perceptually-plausible algorithm for binaural room impulse response simulation," *J. Audio Eng. Soc.*, vol. 62, no. 11, pp. 748–766, 2014.

[51] J. Thiemann and S. van de Par, "A multiple model high-resolution head-related impulse response database for aided and unaided ears," *Eur. Assoc. Signal Process. J. Adv. Signal Process.*, vol. 2019, no. 1, pp. 1–9, 2019.

[52] 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," *Eur. Assoc. Signal Process. J. Adv. Signal Process.*, vol. 2009, pp. 1–10, 2009.

[53] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continuous speech corpus cd-rom. nist speech disc 1–1.1," *NASA STI/Recon technical report n*, 1993, Art. no. 27403.
[54] E. Fonseca et al., "Freesound datasets: A platform for the creation of open audio datasets," in *Proc. 18th Int. Symp. Music Inf. Retrieval Conf.*, 2017. pp. 486–493.
[55] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2014.
[56] W. H. Press and S. A. Teukolsky, "Savitzky-Golay smoothing filters," *Comput. Phys.*, vol. 4, no. 6, pp. 669–672, 1990.
[57] A. Nowoświat and M. Olechowska, "Investigation studies on the application of reverberation time," *Arch. Acoust.*, vol. 41, no. 1, pp. 15–26, 2015.

Peyman Goli received the Ph.D. degree in electrical engineering from the Babol Noshirvani University of Technology, Babol, Iran, in 2017. Since 2020, he has been a Postdoctoral Researcher with the Carl von Ossietzky University of Oldenburg, Oldenburg, Germany. He is currently working on deep learning-based sound source detection and localization, and speech intelligibility improvement in hearing aids. His research interests include audio signal processing, speech intelligibility and quality improvement, machine learning, and sound source localization.

Steven van de Par studied Physics from the Eindhoven University of Technology, Eindhoven, The Netherlands, and did a Ph.D. project on binaural detection followed by a postdoc on auditory-visual synchrony perception. In 2000, he joined Philips Research to work on low-bitrate audio coding, music information retrieval, and computational auditory scene analysis. Since 2010, he has been a Professor of acoustics with the Carl-von-Ossietzky University, Oldenburg, Germany. His research interests include spatial perception, applied psycho-acoustics, computational auditory scene analysis, and virtual acoustics