A Two-stage CNN with Feature Reduction for Speech-aware Binaural DOA Estimation

Reza Varzandeh, Simon Doclo, Volker Hohmann
Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4all, University of Oldenburg, Germany
{reza.varzandeh,simon.doclo,volker.hohmann}@uol.de

***Abstract*—In recent years, several supervised learning-based approaches have been proposed to estimate the direction of arrival (DOA) of a single talker in noisy and reverberant environments. In this paper, we consider a speech-aware DOA estimation system for binaural hearing aids, which does not require a separate voice activity detector (VAD). We propose the combination of two narrowband features as the input features of a convolutional neural network (CNN), namely the cross-power spectrum as spatial features and narrowband auditory-inspired periodicity features. Prior to the joint processing of both features, we propose to reduce the dimensionality of the narrowband periodicity features using a feature reduction stage based on 1×1 convolutions. Simulation results for two reverberant environments with different background noises demonstrate the benefit of the feature reduction stage in terms of DOA estimation accuracy while significantly reducing the number of trainable parameters. In addition, simulation results show that the proposed system outperforms a baseline system consisting of a CNN using only spatial features and a pitch-based VAD.**

***Index Terms*—convolutional neural network, binaural DOA estimation, feature reduction, periodicity degree.**

I. Introduction

Reliably estimating the direction of arrival (DOA) of a talker is a crucial task in applications such as binaural hearing aids [1], [2]. In addition to model-based DOA estimation approaches [3]–[6], in recent years several supervised learning-based DOA estimation approaches based on deep neural networks (DNNs) have been proposed [7]–[13]. In these approaches, the DOA estimation task is often formulated as a classification problem, aiming at determining a mapping from input features to a spatial probability map for a discretized DOA range. Without auxiliary information, e.g., a voice activity detector (VAD), such approaches also provide a DOA estimate during speech pauses or when the signal is dominated by noise, which typically results in erroneous DOA estimates. Hence, a VAD is often cascaded with a DOA estimation system [13], [14]. However, a separate VAD usually requires manual and time-consuming parameter tuning, and may introduce errors that propagate through the DOA estimation system.

In [15], we proposed a speech-aware binaural DOA estimation system based on convolutional neural networks (CNNs), which does not require a separate VAD. Simulation results showed the benefit of using broadband periodicity degree (PD) features in combination with generalized cross-correlation with phase transform (GCC-PHAT) features as input features for the CNN. However, the frequency integration of the cross-power spectrum (CPS) phase employed in the calculation of the GCC-PHAT feature [3], [15] does not allow the CNN to effectively exploit the sparsity property of speech signals in the time-frequency domain [16]. In addition, broadband PD only offers a coarse representation of the harmonic structure of a signal.

In this paper, we extend the speech-aware binaural DOA estimation system of [15] in two ways. First, aiming at exploiting the sparsity property of speech signals, we propose to use a narrowband representation of PD features in combination with narrowband CPS features (as spatial features) as input features for the CNN. Second, the key contribution of this paper is introducing a PD feature reduction stage before the joint processing of both narrowband features, resulting in a two-stage CNN architecture. We postulate here that the feature reduction stage better guides the DOA estimation by reducing the sparse structure of narrowband PD features to a set of more compact spectro-temporal features, referred to as PD saliency features. Evaluation results for a single talker in two reverberant environments for different signal-to-noise ratios (SNRs) show the benefit of using the proposed PD feature reduction stage compared to a system without feature reduction. Evaluation results also show that the proposed systems combining narrowband CPS and PD features outperform a baseline system, consisting of a cascade of a CNN using only narrowband CPS features and a pitch-based VAD.

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. In the short-time Fourier transform (STFT) domain, the m-th microphone signal at time frame n and frequency bin k (with K the STFT length) can be written as

$$Y_m(n,k) = X_m(n,k) + V_m(n,k), \tag{1}$$

where X and V denote the sound source (at direction θ) and the uncorrelated background noise, respectively. By dividing the azimuth range into a set of C discrete DOAs $\{\theta_1,\cdots,\theta_C\}$, 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 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 [8], [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

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project ID 352015383 – SFB 1330 B2.

probability map in the output. The DOA is usually estimated as the DOA class with the highest posterior probability. In this work, to obtain continuous DOA estimates from discrete DOA classes, we estimate the sound source DOA by employing parabolic interpolation [17] on three DOA classes centered around 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 [13], [14], 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 system in Section IV-A.

B. Speech-aware DOA estimation

In contrast to the VAD-informed classification-based approach, in [15] we proposed a classification-based approach referred to as speech-aware DOA estimation. The purpose of speech-aware DOA estimation is to estimate the DOA of a sound source only for speech sources, 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, we consider the class with the highest posterior probability. If this class is a DOA class, we estimate the sound source DOA by employing parabolic interpolation [17] on three DOA classes centered around this class. Otherwise, no reliable DOA could be estimated. In this work, we adopt the speech-aware DOA estimation approach in our proposed systems in Section IV-B.

III. Narrowband input features

Aiming at exploiting speech sparsity in the STFT domain, in this section we describe the narrowband features that are used as input features for the DOA estimation, namely the cross-power spectrum (Section III-A) and the periodicity degree (Section III-B).

A. Cross-power spectrum (CPS)

In [15] the broadband GCC-PHAT, defined as the inverse Fourier transform of the CPS phase, was used as the spatial input feature. In this work, we propose to directly use the narrowband CPS. The instantaneous CPS between the r-th and q-th microphone is defined as

$$G_i(n,k) = Y_r(n,k)Y_q^*(n,k), \tag{2}$$

where $(\cdot)^*$ denotes complex conjugate and i denotes a microphone pair combination. From (2) it can be seen that the CPS encodes both the phase difference and the levels of a microphone pair. As CPS input feature, we consider the real and imaginary parts of $G_i(n,k)$ for all $M(M-1)/2$ unique microphone pairs for frequencies up to the Nyquist frequency, i.e., $k=0,1,\cdots,K/2$, for L consecutive time frames. This means that the shape of the CPS input feature is equal to $L\times(K/2+1)\times 2M(M-1)/2$. We note here that the first, second, and third dimension 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 the real and imaginary parts for all microphone pairs.

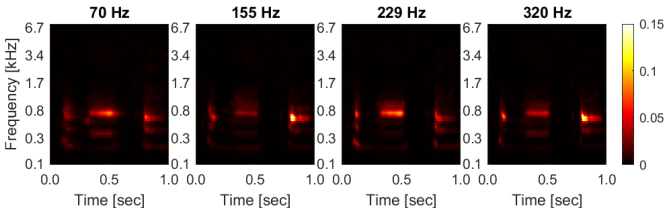


Fig. 1: Illustrative visualization of narrowband PD features for a set of fundamental frequency candidates. The sparse spectro-temporal structure of these features motivates using a feature reduction stage prior to the joint processing of the CPS and PD features by the CNN.

B. Periodicity degree (PD)

In [15] broadband PD features, which only offer a coarse representation of the harmonic structure of a signal, were used as input features. In this work, we propose to use a narrowband formulation of the PD features, estimated for a set of N fundamental period candidates. The PD features are computed by first decomposing a reference microphone signal into a set of bandpass-filtered time signals using a gammatone filter bank (GTFB) [18]. The real part of each bandpass-filtered signal is then passed through a half-wave rectification, followed by a fifth-order low-pass filter with 770 Hz cutoff frequency and a second-order high-pass filter with 40 Hz cutoff frequency, resulting in bandpass-filtered signal envelopes $y_{env}(t,f)$ in time t and subband f. In the next step, a set of N parallel infinite impulse response (IIR) comb filters designed for a set of fundamental period candidates $p_j, j=1,\cdots,N$, filter the signal envelopes as

$$s(t,f,j) = (1-\alpha)y_{env}(t,f) + \alpha s(t-p_j,f,j), \tag{3}$$

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

$$PD(t,f,j) = (1-\beta_j)|s(t,f,j)| + \beta_j PD(t-1,f,j), \tag{4}$$

where $|\cdot|$ denotes the absolute value and the parameter β_j for each fundamental period candidate is defined as $\beta_j = e^{-1/p_j}$.

Since we aim at joint spectro-temporal processing of the PD and CPS features, it is required to represent both features at the same time-frequency resolution. To obtain the same time resolution as the CPS features, the PD features are averaged in each STFT frame. Unlike the linearly-spaced STFT frequency bands, the gammatone bands have a non-uniform frequency resolution that decreases with frequency. To obtain the same frequency resolution for the PD features as for the CPS features, for low STFT frequencies we average the PD features in gammatone bands associated with one STFT frequency band. In contrast, for high STFT frequencies we replicate the PD features of each gammatone band and assign them to those STFT frequency bands associated with one gammatone band. Similarly as for the

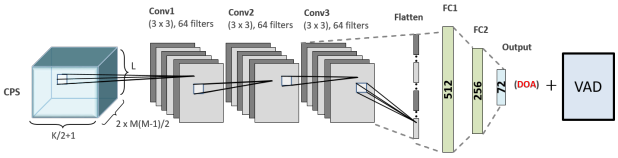


Fig. 2: Baseline VAD-informed DOA estimation system using only CPS features.

CPS features, we consider L consecutive frames, such that the shape of the PD input feature is equal to $L \times (K/2+1) \times N$.

For a 1s clean signal of a female talker, Fig. 1 depicts exemplary two-dimensional (2D) narrowband PD features, corresponding to a subset of fundamental frequency candidates (each representing an input channel). For a perfectly periodic signal with a certain fundamental frequency, a high PD value will be captured in each time-frequency bin across the N input channels associated with the harmonics and sub-harmonics of this fundamental frequency. Even though speech signals are not perfectly harmonic, their fundamental frequency variations and multiple harmonics exhibit a spectro-temporal structure that can be identified in the input channels of the PD features. The main idea of using PD features in combination with CPS features is to use the salient periodicity features as a footprint of speech signals in a noisy mixture [19], [20]. This enables the CNN to detect voiced speech portions of a signal, at the same time mapping the CPS features of these portions to the DOA of the talker.

IV. CNN-based DOA estimation systems

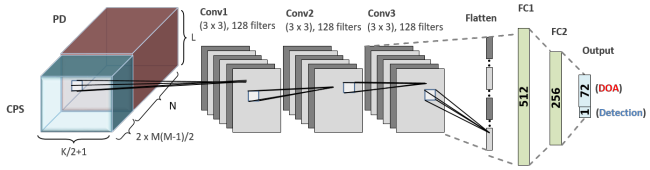
In this section, we describe the CNN-based DOA estimation systems. Section IV-A discusses the baseline system, which adopts a VAD-informed DOA estimation approach and uses only the CPS features. Section IV-B presents the proposed systems, which adopt a speech-aware DOA estimation approach and use a combination of the CPS features and the narrowband PD features as input features.

A. Baseline VAD-informed system

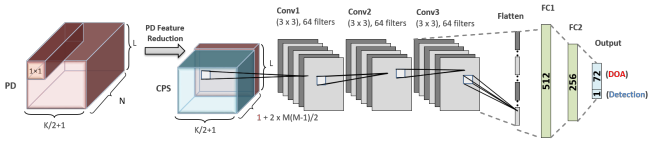
Fig. 2 depicts the baseline system consisting of a CNN using only spatial CPS features as input cascaded with a pitch-based binary VAD [21]. In the baseline CNN architecture, each convolutional block (*Conv1* to *Conv3*) consists of a cascade of 2D convolutional, batch normalization, rectified linear unit (ReLU) activation, and 2D max-pooling layers. The outputs of the last pooling layers in *Conv3* are concatenated and then used as an input for a cascade of two fully-connected blocks (*FC1* to *FC2*), each representing a fully-connected dense layer followed by batch normalization, ReLU activation, and dropout layers. A softmax activation function predicts the posterior probability map for the C DOA classes.

B. Proposed speech-aware systems

Fig. 3 depicts the proposed speech-aware DOA estimation systems, which use narrowband PD features in combination with spatial CPS features as input features of the CNN. We expect that by training these systems with speech and non-speech signals, the CNN can capture the spectro-temporal structure of the signal encoded in the PD features, thereby distinguishing between speech and non-speech portions while simultaneously



(a) without PD feature reduction



(b) with PD feature reduction stage (two-stage CNN)

Fig. 3: Proposed speech-aware DOA estimation systems: (a) CPS and PD features are jointly processed by the CNN, (b) PD features are reduced to PD saliency features using 1×1 convolutions before being jointly processed with CPS features by the CNN.

mapping the CPS features to a sound source DOA when speech portions in the signal are detected.

The system in Fig. 3a directly employs 2D convolutional filters to the time-frequency regions of each input channel, i.e., PD and CPS features belonging to the same time-frequency bins are jointly processed, ensuring a proper association of both features. However, the spectro-temporal sparsity of the PD features (as visualized in Fig. 1) may complicate this task when a relatively large number of PD channels are correlated to the CPS features by the CNN. This motivates the usage of a PD feature reduction stage prior to the joint feature processing by the CNN.

Fig. 3b depicts the proposed two-stage CNN architecture including a PD feature reduction stage. The PD feature reduction stage aims at reducing the PD input depth, i.e., the number of channels, while keeping its width and height, i.e., the time-frequency resolution fixed. We propose to use 1×1 convolutions [22] to reduce the N-channel PD features to a 1-channel PD feature, which can be interpreted as a PD saliency feature for each time-frequency bin. In the next stage, the PD saliency features are jointly processed with the CPS features using 2D convolutional filters. It should be noted that both stages are jointly trained.

The CNN architecture of the proposed systems in Fig. 3 is very similar to the CNN architecture of the baseline system in Fig. 2. However, since the input features of the first convolutional block (*Conv1*) in the considered systems are different (CPS only, CPS and PD, CPS and PD saliency), the number of input channels is obviously different. In addition, the VAD-informed baseline system has C nodes in the output layer, whereas the speech-aware systems have $C+1$ nodes in the output layer. Finally, after hyperparameter optimization the best performance was obtained when using 64 convolutional filters in the baseline system and the two-stage CNN (both corresponding to about 5.5 million trainable parameters), and using 128 convolutional filters in the proposed system without feature reduction (corresponding to about 11.2 million trainable parameters).

V. Experimental Evaluation

In this section, we conduct experiments to evaluate the performance of the baseline system and the proposed speech-aware sys-

tems described in Section IV-A and Section IV-B, respectively.

A. Datasets and data generation for training and evaluation

We used a database of multichannel binaural room impulse responses (BRIRs) [23] to generate data for training and evaluation. The considered binaural hearing aid setup consists of $M=4$ microphones, where the front and rear microphones in both left and right hearing aids were used. We used sound source signals from speech [24] and non-speech [25] datasets to generate the training and validation data required during the training of all CNNs. For evaluation, only speech signals from the validation TIMIT [24] data were used as source signals. Source signals were randomly chosen from unique speakers (both male and female) and from three categories [15] of non-speech signals. We generated the noisy binaural microphone signals by convolving the source signals with BRIRs and mixing the resulting clean binaural signals with a background noise at different SNRs. All systems were trained in noisy anechoic conditions and evaluated in noisy reverberant environments.

During training, we used a simulated binaural diffuse noise to generate noisy binaural microphone signals at SNRs ranging from -5 dB to $+20$ dB in 5 dB steps. This diffuse noise was generated by convolving uncorrelated speech-shaped noise taken from the ICRA noise database [26] with anechoic BRIRs and summing all resulting binaural signals from 72 directions. In total, we obtain 3.85 million training examples. To calculate the validation loss at the end of each epoch, 200000 examples were randomly selected from the validation data and kept fixed throughout training.

We generated the evaluation data for static-source scenarios in two real environments [23] (cafeteria and courtyard) with a reverberation time of approximately 1300 ms and 900 ms, respectively. The recorded cafeteria babble noise and courtyard ambient noise [23] were used to generate noisy binaural microphone signals. All systems were evaluated at SNRs ranging from -5 dB to $+10$ dB in 5 dB steps. A total of 150 speech segments randomly chosen from 30 unique male and female speakers (each with a length of 1 s) were used as source signals. In each environment, we considered BRIRs of two head orientations for four source positions [23]. It should be noted that the source and background noise signals, acoustic conditions, and source positions used during evaluation were different from those used during training and validation.

B. Implementation details

In our simulations, we used a sampling frequency $f_s=16$ kHz and an STFT framework with a Hann window of length $K=160$ (corresponding to 10 ms) and 50% overlap, resulting in 81 STFT frequency bins. Each training example includes a block of $L=20$ consecutive time frames. For the PD feature computation, we used a 4-th order GTFB implementation [18] with 61 frequency subbands, a group delay of 256, and minimum and maximum center frequency of 60 Hz and 7200 Hz. For PD features, we chose $N=180$ fundamental period candidates corresponding to minimum and maximum fundamental frequencies of 70 Hz and 320 Hz, respectively. The comb filter gain in (3) was chosen to be $\alpha=0.7$. We considered the front microphone of the left hearing aid as the reference microphone for the PD feature extraction.

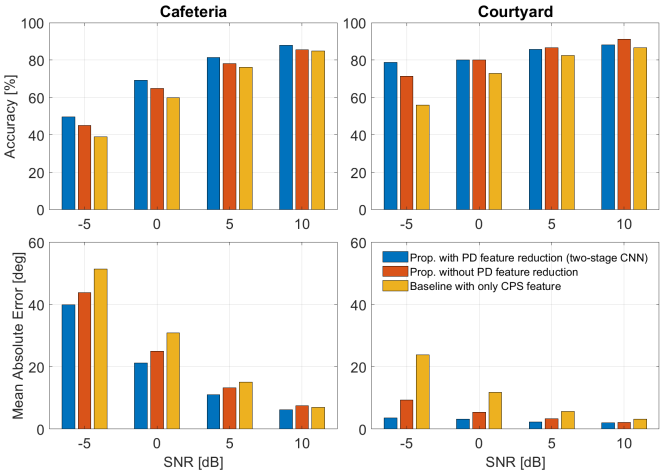


Fig. 4: Accuracy and MAE of the proposed systems with narrowband feature combination evaluated against the baseline system using only CPS features in static-source scenarios for different SNR conditions in the cafeteria and courtyard environments.

All systems were implemented using PyTorch [27]. For all CNNs, we used a 2D convolutional filter size of 3×3 with a stride size of 1×1. The max-pooling size was 2×1, i.e., no pooling is applied across frequencies. In addition to the batch normalization used in the convolutional and fully-connected blocks of the CNNs, the layer normalization [28] was applied on the CPS and PD features separately at the input. The CNNs were trained using the Adam optimizer [29], 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. An early stopping regularization method on the validation loss and a variable learning rate scheduler with a factor of 0.5 were also employed. A softmax activation function is used at the output layer of all systems.

C. Performance measures

To evaluate the DOA estimation performance, we used mean absolute error (MAE) and accuracy (Acc). A DOA estimate in frame 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°. The MAE (in degrees) and accuracy are defined as

$$\text{MAE}=\frac{1}{\mathcal{L}}\sum_{l=1}^{\mathcal{L}}\left|\hat{\theta}_l-\theta_l\right|, \tag{5}$$

$$\text{Acc}=\frac{\mathcal{L}_{\text{acc}}}{\mathcal{L}}\times100, \tag{6}$$

where $\mathcal{L}$ and $\mathcal{L}_{\text{acc}}$ denote the total number of estimates and the total number of accurate estimates, respectively.

D. Results and discussion

Fig. 4 shows the performance of all considered systems in terms of accuracy and MAE. By comparing the proposed two-stage CNN with PD feature reduction to the proposed system without PD feature reduction, it can be observed that the two-stage CNN generally results in a better or similar performance. The benefit of using PD feature reduction is especially clear in challenging acoustic conditions, i.e., in the highly-reverberant

cafeteria environment and in adverse SNR conditions in both environments. Although in terms of accuracy this benefit decreases with increasing SNR in the courtyard environment in favor of the proposed system without feature reduction, the proposed two-stage CNN maintains a lower MAE in all conditions.

The results in Fig. 4 clearly show that both proposed systems consistently outperform the baseline system in both environments and for all SNR conditions. This benefit decreases towards high SNR conditions, which is expected as there are fewer signal portions dominated by noise, which PD features can detect.

Considering the number of trainable parameters (cf. Section IV-B), compared to the baseline system the proposed two-stage CNN requires a comparable number of parameters while achieving a better performance. Moreover, the proposed two-stage CNN outperforms the proposed system without feature reduction while requiring significantly fewer parameters. This further highlights the benefit of employing the proposed feature reduction stage before the joint processing of the proposed narrowband feature combination.

VI. Conclusion

In this paper, we proposed two speech-aware DOA estimation systems that use a combination of narrowband periodicity features and spatial CPS features as inputs of a CNN. In particular, we introduced a two-stage CNN with a periodicity feature reduction stage employing 1×1 convolutions. Evaluation results showed that the proposed systems yield a better DOA estimation performance than a baseline system using CPS features and a pitch-based VAD. While offering a lower computational complexity, the proposed two-stage CNN with feature reduction outperforms a system that jointly processes the feature combination without feature reduction. This study suggests that a feature reduction stage can effectively map the sparse periodicity features into more compact salient periodicity features, which combined with spatial features, provide robust features to guide speech-aware DOA estimation.

References

[1] G. Grimm, H. Kayser, M. Hendrikse, and V. Hohmann, "A gaze-based attention model for spatially-aware hearing aids," in *Proc. ITG Symposium on Speech Communication*, Oldenburg, Germany, 2018, pp. 1–5.

[2] A. Aroudi and S. Doclo, "Cognitive-driven binaural beamforming using EEG-based auditory attention decoding," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 28, pp. 862–875, 2020.

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

[4] J. H. Dibiase, *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*, Ph.D. thesis, Brown University, Providence RI, USA, Aug. 2000.

[5] 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 on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz NY, USA, Oct. 2015, pp. 1–5.

[6] D. Fejgin and S. Doclo, "Coherence-based frequency subset selection for binaural RTF-vector-based direction of arrival estimation for multiple speakers," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Bamberg, Germany, Sep. 2022.

[7] P. A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *The Journal of the Acoustical Society of America*, 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. on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2444–2453, Dec. 2017.

[9] W. He, P. Motlicek, and J. Odobez, "Deep neural networks for multiple speaker detection and localization," in *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, Brisbane, Australia, May 2018, pp. 74–79.

[10] S. Chakrabarty and E. A. P. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 8–21, March 2019.

[11] P. Vecchiotti, N. Ma, S. Squartini, and G. J. Brown, "End-to-end binaural sound localisation from the raw waveform," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 2019, pp. 451–455.

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

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

[14] C. Evers, H. W. Löllmann, H. Mellmann, A. Schmidt, H. Barfuss, P. A. Naylor, and W. Kellermann, "The LOCATA challenge: Acoustic source localization and tracking," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 28, pp. 1620–1643, 2020.

[15] 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 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, Apr. 2020, pp. 566–570.

[16] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[17] J. O. Smith and X. Serra, "PARSHL: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation," in *Proc. International Computer Music Conference (ICMC)*, Champaign/Urbana, IL, USA, Aug. 1987, pp. 290–297.

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

[19] A. Josupeit, N. Kopčo, and V. Hohmann, "Modeling of speech localization in a multi-talker mixture using periodicity and energy-based auditory features," *The Journal of the Acoustical Society of America*, vol. 139, no. 5, pp. 2911–2923, 2016.

[20] J. Luberadzka, H. Kayser, and V. Hohmann, "Making sense of periodicity glimpses in a prediction-update-loop—a computational model of attentive voice tracking," *The Journal of the Acoustical Society of America*, vol. 151, no. 2, pp. 712–737, 2022.

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

[22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1–9.

[23] 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 Journal on Advances in Signal Processing*, vol. 2009, no. 1, pp. 298605, Jul. 2009.

[24] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[25] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proc. ACM Conference on Multimedia*, Brisbane, Australia, Oct. 2015, pp. 1015–1018.

[26] 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.

[27] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*. 2019, vol. 32, Curran Associates, Inc.

[28] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[29] D. Kingma P and J. Ba, "Adam: A method for stochastic optimization," in *Proc. International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, May 2015.