EXPLOITING PERIODICITY FEATURES FOR JOINT DETECTION AND DOA ESTIMATION OF SPEECH SOURCES USING CONVOLUTIONAL NEURAL NETWORKS

Reza Varzandeh[1,2,3] *Kamil Adiloğlu*[1,3] *Simon Doclo*[2,3] *Volker Hohmann*[2,3]

[1] HörTech gGmbH, Oldenburg, Germany
[2] University of Oldenburg, Department of Medical Physics and Acoustics
[3] Cluster of Excellence Hearing4all, Oldenburg, Germany
{r.varzandeh,k.adiloglu}@hoertech.de,{simon.doclo,volker.hohmann}@uni-oldenburg.de

ABSTRACT

While many algorithms deal with direction of arrival (DOA) estimation and voice activity detection (VAD) as two separate tasks, only a small number of data-driven methods have addressed these two tasks jointly. In this paper, a multi-input single-output convolutional neural network (CNN) is proposed which exploits a novel feature combination for joint DOA estimation and VAD in the context of binaural hearing aids. In addition to the well-known generalized cross correlation with phase transform (GCC-PHAT) feature, the network uses an auditory-inspired feature called periodicity degree (PD), which provides a broadband representation of the periodic structure of the signal. The proposed CNN has been trained in a multi-conditional training scheme across different signal-to-noise ratios. Experimental results for a single-talker scenario in reverberant environments show that by exploiting the PD feature, the proposed CNN is able to distinguish speech from non-speech signal blocks, thereby outperforming the baseline CNN in terms of DOA estimation accuracy. In addition, the results show that the proposed method is able to adapt to different unseen acoustic conditions and background noises.

Index Terms— convolutional neural networks, binaural DOA estimation, voice activity detection, periodicity

1. INTRODUCTION

In many applications, such as assistive listening devices, hands-free speech communication systems and robot audition, reliable DOA estimates of sound sources are required [1–4]. While the human auditory system has the remarkable ability to localize a speech source in noisy and reverberant environments, this remains a challenging task for machine listening systems such as hearing aids (HAs) [5].

A set of classical DOA estimation approaches is based on GCC-PHAT [6] or its generalization, called steered response power with phase transform (SRP-PHAT) [7]. Other state-of-the-art approaches apply model-based techniques such as maximum likelihood (ML) estimation [8], subspace-based techniques such as multiple signal classification (MUSIC) [9], or machine learning techniques such as deep neural networks (DNNs) [4, 10–13]. However, the performance of all methods degrades in noisy and reverberant environments.

Most of the aforementioned methods estimate the DOA of the sound sources without detecting of speech activity. Vecchiotti et al. [14, 15] proposed a CNN which was trained using logarithmic mel spectrogram (LogMel) and GCC-PHAT features to simultaneously estimate the activity and location of a single talker. This method estimates the coordinates of a single talker. To the best of our knowledge, there is no data-driven DNN-based approach which jointly addresses both problems of DOA estimation and VAD.

Many features have been used for VAD [16]. Among them, periodicity information is an important cue to discriminate between different talkers [17], and it has been shown to be useful for VAD at low signal-to-noise ratios (SNRs) [18]. In [18], an auditory-inspired feature called PD has been proposed for fundamental period detection and estimation.

In this paper, we propose to train a hybrid CNN in a single-label multi-class classification scheme using a combination of GCC-PHAT and PD features for joint binaural DOA estimation and VAD. In the proposed CNN, the DOA is estimated at block-level, i.e., feature vectors of multiple consecutive time frames are used as the input feature map of the CNN. This eliminates the necessity of an accurate VAD to remove silent frames of a speech signal in the training. By training with speech and non-speech signals, the network learns the harmonic structure of the signal from the PD features over consecutive frames, thereby detecting speech and non-speech signal blocks, and at the same time estimating the DOA of the detected speech signal blocks using GCC-PHAT features. It should be noted that the proposed method only needs feature maps of a block containing current and past frames, i.e., can be used in an online fashion. The trained CNN is evaluated for the task of joint DOA estimation and VAD for several single-talker scenarios in different reverberant environments with unseen background noises. Experimental results in various acoustic conditions show that the proposed CNN outperforms the baseline system which only uses GCC-PHAT features for DOA estimation.

2. INPUT FEATURES

We consider a binaural HA setup with $M = 4$ microphones (2 microphones on each HA), recording a single sound source at DOA θ in the azimuthal plane and background noise in a reverberant environment. The m-th microphone signal in the short-time Fourier transform (STFT) domain at time frame n and frequency bin k is given by

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

where X_m and V_m denote the mutually uncorrelated source signal and background noise signal, respectively. The sound source can be either a speech or a non-speech signal. In this paper, all signals were analyzed using a Hann window at a sampling frequency of 16 kHz with an STFT frame length of 10 ms and 50% overlap.

2.1. Generalized Cross Correlation With Phase Transform

GCC-PHAT has been sucessfully used as a robust feature for several data-driven DOA estimation methods in reverberant environments [4, 13, 19, 20]. The GCC-PHAT vector between the r-th and

This work was funded by the German Federal Ministry of Education and Research, FKZ 16SV7785 ”mEEGaHStim”. The authors would like to thank Dr. Bernd T. Meyer for discussions, and Dr. Hendrik Kayser for providing code.

978-1-5090-6631-5/20/$31.00 ©2020 IEEE

the q-th microphone at frame n is defined as the inverse Fourier transform (IFFT) of the cross-power spectrum phase [21], i.e.,

$$\boldsymbol{\rho}_{rq}^{(n)} = \mathcal{IFFT}\left(\frac{\boldsymbol{Y}_{\boldsymbol{r}}^{(\boldsymbol{n})} \odot \boldsymbol{Y}_{\boldsymbol{q}}^{(\boldsymbol{n})^*}}{|\boldsymbol{Y}_{\boldsymbol{r}}^{(\boldsymbol{n})} \odot \boldsymbol{Y}_{\boldsymbol{q}}^{(\boldsymbol{n})^*}|}\right), \tag{2}$$

where $(\cdot)^*$, $|\cdot|$ and $\odot$ denote complex conjugate, absolute value and element-wise multiplication, respectively. $\boldsymbol{Y}_{\boldsymbol{r}}^{(\boldsymbol{n})}$ and $\boldsymbol{Y}_{\boldsymbol{q}}^{(\boldsymbol{n})}$ represent STFT vectors of the r-th and q-th microphone signals for all frequency bins, i.e.,

$$\boldsymbol{Y}_{\boldsymbol{r}}^{(\boldsymbol{n})} = [Y_r(n,1),\cdots,Y_r(n,k_{\max})]^{\mathrm{T}}, \tag{3}$$

$$\boldsymbol{Y}_{\boldsymbol{q}}^{(\boldsymbol{n})} = [Y_q(n,1),\cdots,Y_q(n,k_{\max})]^{\mathrm{T}}, \tag{4}$$

where $k_{\max}$ and $(\cdot)^{\mathrm{T}}$ denote the maximum frequency bin and the vector transpose, respectively. From the vector $\boldsymbol{\rho}_{rq}^{(n)}$, we only consider $\boldsymbol{\phi}_{rq}^{(n)} = \left[\boldsymbol{\rho}_{rq}^{(n)}(-\tau_{rq}),\cdots,\boldsymbol{\rho}_{rq}^{(n)}(\tau_{rq})\right]^{\mathrm{T}}$, where τ_{rq} corresponds to the maximum possible delay between the r-th and the q-th microphone, depending on their distance. The GCC-PHAT feature vector at frame n is constructed by concatenating the vectors $\boldsymbol{\phi}_{rq}^{(n)}$ for all possible microphone pairs [20], i.e.,

$$\boldsymbol{\phi}^{(n)} = \left\{\boldsymbol{\phi}_{rq}^{(n)}\right\}_{r=1,\cdots,M;q=r+1,\cdots,M}. \tag{5}$$

For our considered binaural HA setup, we obtain a GCC-PHAT feature vector $\boldsymbol{\phi}$ of size 678 delay bins for each frame.

2.2. Periodicity Degree

The PD feature $PD(n,p)$ [18] at frame n is defined as the ratio of the harmonic signal power for a given period p and the total signal power. If there is no periodic content with period p in the signal, then $PD(n,p) = 0$. If a signal is fully harmonic with period p at frame n, then $PD(n,p) = 1$. The PD feature vector for N possible candidate periods can be written as $\boldsymbol{\psi}^{(n)} = [PD(n,1),\cdots,PD(n,N)]^{\mathrm{T}}$.

The PD feature vector can be computed for each of the M microphones. Without loss of generality, in this paper we will consider the front microphone of the left HA as the reference microphone. To compute the PD feature, the time-domain reference microphone signal is first decomposed into 60 frequency subbands by a gammatone filter bank (GTFB) with minimum and maximum center frequency (CF) of 70 Hz and 7200 Hz. Each subband signal is passed through a haircell processing stage encompassing half-wave rectification followed by a low-pass filter, which extracts the fine structure information (up to 1.5 kHz) and the envelope information of the signal. The subband-averaged PDs are eventually estimated with the same temporal resolution as the time-domain signal. The details of this method can be found in [18]. By choosing $N = 180$, the range of fundamental period candidates for PD feature extraction lies between 3.1 ms and 14.3 ms, corresponding to fundamental frequencies in the range from 320 Hz to 70 Hz.

Figure 1 depicts an exemplary representation of PD feature vectors for 1 s clean and noisy speech and non-speech (engine sound) signals. While for the clean and noisy speech signals, the fundamental period and its multiple harmonics are clearly identifiable as a two-dimensional (2D) structure over time, no such harmonic structure exists for the engine sound. This 2D structure for speech signals motivates the usage of 2D convolutional filters in the CNN (see Section 3).

2.3. Input Feature Maps

As input features for joint VAD and DOA estimation, we propose to use both GCC-PHAT and PD feature vectors ($\boldsymbol{\phi}^{(n)}$ and $\boldsymbol{\psi}^{(n)}$), after scaling to unit variance and zero mean. To capture relevant information over time, we propose to use feature vectors of a block of L consecutive frames (with length 10 ms and 50% overlap). The GCC-PHAT and PD feature maps at frame n can be defined as $\boldsymbol{\Phi}^{(n)} = [\boldsymbol{\phi}^{(n)},\cdots,\boldsymbol{\phi}^{(n-L+1)}]$ and $\boldsymbol{\Psi}^{(n)} = [\boldsymbol{\psi}^{(n)},\cdots,\boldsymbol{\psi}^{(n-L+1)}]$.

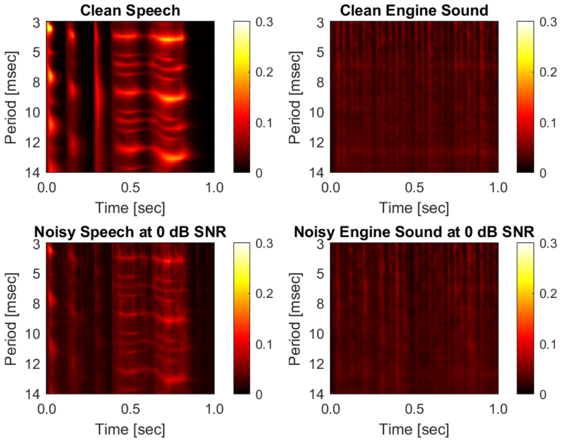

Fig. 1: An exemplary visualization of PD feature vectors for clean and noisy female speech and engine sound signals.

3. PROPOSED NETWORK ARCHITECTURE

The proposed hybrid network architecture is depicted in Figure 2. It consists of two parallel and independent branches of cascaded CNN layers which receive the PD and GCC-PHAT feature maps as *input1* and *input2*, respectively. Each CNN layer (*Conv1* to *Conv4*) embodies a cascade of 2D convolutional, activation and 2D max pooling layer. The outputs of both branches are concatenated after a flattening layer. The resulting output can be seen as a heterogeneous feature vector of PD and GCC-PHAT. The concatenated output is then used as an input for a cascade of fully connected layers (*FC1* to *FC3*), each representing a fully connected dense layer followed by batch normalization, activation and dropout layers. The classification in the output layer of the network is achieved by using a softmax layer as the activation function.

The main motivation to use the proposed combination of PD and GCC-PHAT features is to allow the network to capture the most relevant information required for joint VAD and DOA estimation. More in particular, the proposed CNN is expected to learn the harmonic structure and continuity of the sound source from the PD feature map, thereby discriminating between speech and non-speech blocks, and at the same time capture the spatial information of the sound source from the GCC-PHAT feature map, thereby finding the DOA. The advantage of using task-specific features instead of using the magnitude and phase spectrogram as generic input features [10, 12] is to realize the CNN layers with a small number of filters, and hence less trainable network parameters. As a baseline system we consider a CNN which only uses the GCC-PHAT feature map as input (see Section 6.2).

All implementations for training and evaluating the CNN networks were realized using Keras [22]. For all 2D convolutional layers of both branches 4-channel filters with filter size of 3×3 and with stride size of 1×1 were used. The max pooling size was 2×2 with strides of the same size. The rectified linear unit (ReLU) activation function was used for all fully connected and convolutional layers. All trainable network weights were initialized using the Glorot uniform initializer. Adam optimizer was used as the optimization algorithm for CNN training with categorical cross-entropy as loss function and a learning rate of 0.002. The dropout rate was set to 0.5. In addition, overfitting was prevented using early stopping, i.e., the training was stopped if no improvement in validation loss was observed for 10 epochs. A variable learning rate schedule was implemented to halve the learning rate if no validation loss improvement was seen for 3 epochs.

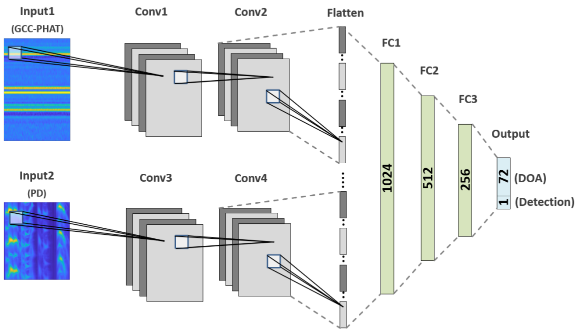

Fig. 2: Proposed hybrid CNN architecture. An exemplary visualization of the GCC-PHAT and PD feature maps of a speech signal is shown as input1 and input2, respectively.

4. JOINT DOA ESTIMATION AND SPEECH DETECTION

The task of joint VAD and DOA estimation is realized by training the CNN using oracle DOA and speech/non-speech detection labels. Each label is one-hot encoded, which means that each training example belongs to only one output class. The joint task can be formulated as a $C+1$-class classification task, where the first C classes (called DOA classes) refer to all possible discrete DOA values $\{\theta_1,\cdots,\theta_C\}$, and the last class refers to the speech/non-speech activity (called detection class). In this work, we consider $C=72$ for the full 360° azimuth range, which corresponds to a 5-degree resolution DOA map in the horizontal plane.

On the one hand, for a training feature map of a speech source coming from a certain 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, for a training feature map of a non-speech source the detection class is labeled by one, whereas all DOA classes are labeled by zero no matter which direction the non-speech source is coming from. In doing so, the network is expected to learn to map the PD and GCC-PHAT features to a posterior probability map $\boldsymbol{P}=[P_1,\cdots,P_C,P_{C+1}]$ for given number of directions C.

For the joint classification task, we formulate two hypotheses

$$\mathcal{H}_s: \quad \text{speech is detected,} \tag{6}$$

$$\mathcal{H}_{ns}: \quad \text{non-speech is detected,} \tag{7}$$

and define the decision rule as

$$\begin{aligned} &\text{decide} \quad \mathcal{H}_{ns} \quad \text{if} \operatorname*{argmax}_i P_i|\boldsymbol{P}=C+1 \\ &\text{decide} \quad \mathcal{H}_s \quad \text{otherwise.} \end{aligned} \tag{8}$$

The DOA class number corresponding to the estimated DOA is obtained as

$$I=\operatorname*{argmax}_i P_i|\mathcal{H}_s, \tag{9}$$

and the estimated DOA is given by $\hat{\theta}=\theta_I$. The joint classification task can be described as follows. At time frame n, given the posterior probability map $\boldsymbol{P}$ predicted from a block of L consecutive time frames, and under the single-source assumption, we take the largest value of the probability map. If the last class is detected, i.e., P_{C+1} is found as the maximum, our estimate at that frame is predicted as non-speech, and hence, we do not estimate the DOA. Otherwise, speech is detected, and the detected class determines the estimated DOA.

5. TRAINING AND VALIDATION DATA

For training and validation, corpora of 462 and 168 unique speakers from the TIMIT dataset [23] (including both male and female speakers) were used. The silences in the beginning and the end of the files were removed. For non-speech signals three categories (natural soundscapes and water sounds, interior and domestic sounds, and exterior and urban noises) of the ESC50 dataset [24] were used. A total number of 960 and 240 distinct sound files were used for training and validation, respectively. The clean binaural microphone signals used during training were generated for both speech and non-speech sources by convolving them with anechoic binaural room impulse responses (BRIRs) [25]. The front and rear microphones in both left and right HAs were used as the 4-channel microphone array. Multi-conditional training was performed in different SNRs ranging from -10 dB to $+20$ dB in 5 dB steps. The noisy binaural microphone signals were generated by mixing the clean binaural microphone signals with simulated binaural diffuse noise. This noise was generated by summing uncorrelated white Gaussian noise (WGN) sources from all 72 directions, using the same anechoic BRIRs as for the clean data generation.

Training feature maps were extracted for both speech and non-speech signals at different SNRs and for all 72 directions. By taking 300 and 100 unique speakers and sounds from the training and validation corpora, in total we have 86.4 million time frames as *training set* and 28.8 million time frames as *validation set*. The maximum epoch number was set to 100. In each epoch, 100 mini-batches of 720 blocks were randomly selected from the training set such that network did not see the same block twice. The block length L was set as an hyperparameter for training the network. Each mini-batch included all SNR conditions and DOA classes for both speech and non-speech signals in a uniform way. To calculate the validation loss at the end of each epoch, 21600 blocks were randomly selected from the validation set, and kept fixed throughout the training. The validation data were not seen by the network during the training.

6. EXPERIMENTAL EVALUATION

6.1. Experiment Design

To evaluate the generalizability of the trained network to unmatched acoustic conditions and unseen background noises, we evaluated the performance in two reverberant environments. The binaural microphone signals used for the evaluation were simulated by convolving the clean source signals from the validation TIMIT corpus with BRIRs [25] of two real environments (cafeteria and courtyard) with reverberation times of approximately 1300 ms and 900 ms, respectively. The room configurations are depicted in Figure 3, where in each room four source positions (specified with dashed boxes) and for each source position two head orientations were considered. Recorded cafeteria babble noise and courtyard ambient noise were used to generate noisy binaural microphone signals at SNRs ranging from -5 dB to $+10$ dB. A total number of 150 unique speakers (each with length 1 s) were selected from the validation TIMIT corpus. For each environment and room configuration, we extracted the feature maps of consecutive blocks. For each SNR condition and environment, we evaluated the CNNs trained for two different block lengths ($L=20$ and $L=50$, corresponding to 100 ms and 250 ms, respectively). A simple broadband energy-based VAD was used as oracle VAD to discard signal blocks with low speech energy from the evaluation data.

6.2. Baseline System

The performance of the proposed hybrid network using both GCC-PHAT as well as PD feature maps is compared with a CNN using only GCC-PHAT feature maps. The baseline CNN architecture is depicted in

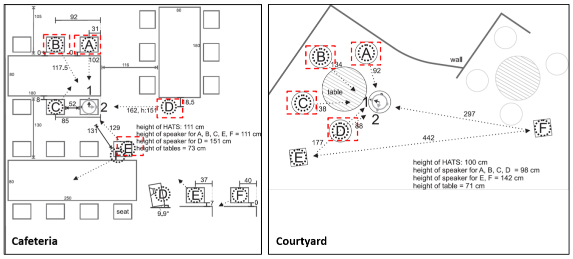

Fig. 3: Evaluation setup in two reverberant environments. In the cafeteria source positions *A*, *B*, *D*, *E* were considered, while in the courtyard source positions *A*, *B*, *C*, and *D* were considered.

Figure 4, which looks very similar to the upper branch of the proposed network in Figure 2. Since the baseline system only aims at DOA estimation and no speech detection, the baseline network was only trained with speech signals. In addition, instead of 73 output units in the hybrid network, the output layer of the baseline network consists of 72 units, which only predicts the DOA posterior probability map. Apart from these differences, the training parameters and network hyperparameters of the proposed and baseline networks are the same (see Section 3). During the evaluation, given the predicted posterior probability map $\boldsymbol{P} = [P_1, \cdots, P_C]$, the estimated DOA is determined by the DOA class given by $\mathrm{argmax}_i\, P_i | \boldsymbol{P}$. If a signal block is mostly dominated by silent frames, the resulting estimated DOA may lead to a large DOA estimation error (see Section 6.3).

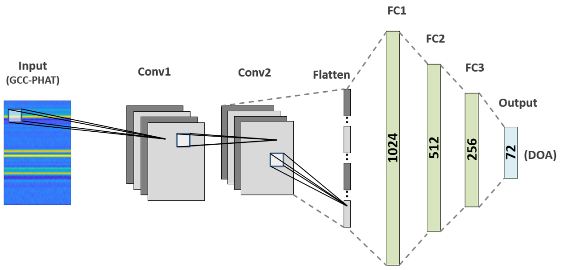

Fig. 4: Baseline CNN architecture using only the GCC-PHAT feature map as input.

6.3. Performance Measures

The DOA estimation performance of both CNNs is evaluated in terms of mean absolute error (MAE) and accuracy (Acc.) [10, 26]. An estimate in block d is considered accurate if the absolute error between the estimated DOA $\hat{\theta}_d$ and the oracle DOA θ_d is smaller than 5 degrees. The accuracy and the MAE (in degrees) are defined as

$$\mathrm{Acc.} = \frac{1}{\mathrm{D}} \sum_{\mathrm{d}=1}^{\mathrm{D}} \Theta\left(5 - \left|\hat{\theta}_{\mathrm{d}} - \theta_{\mathrm{d}}\right|\right) \times 100, \tag{10}$$

$$\mathrm{MAE} = \frac{1}{\mathrm{D}} \sum_{\mathrm{d}=1}^{\mathrm{D}} \left|\hat{\theta}_{\mathrm{d}} - \theta_{\mathrm{d}}\right|, \tag{11}$$

where D is the total number of blocks used for the evaluation in all room configurations, and Θ is the Heaviside step function. It should be realized that since the baseline system is merely designed for the DOA estimation task and there is no VAD integrated in the evaluation of this system, D includes all signal blocks of the evaluation data. On the other hand, the proposed system is only evaluated for blocks where it detects speech activity, i.e., the DOA estimation errors are only calculated when speech is detected to be present.

6.4. Results

Figure 5 depicts the accuracy and the mean absolute error of the proposed and the baseline system for different SNRs in two reverberant environments (cafeteria and courtyard). Both systems were trained and evaluated for two different block lengths.

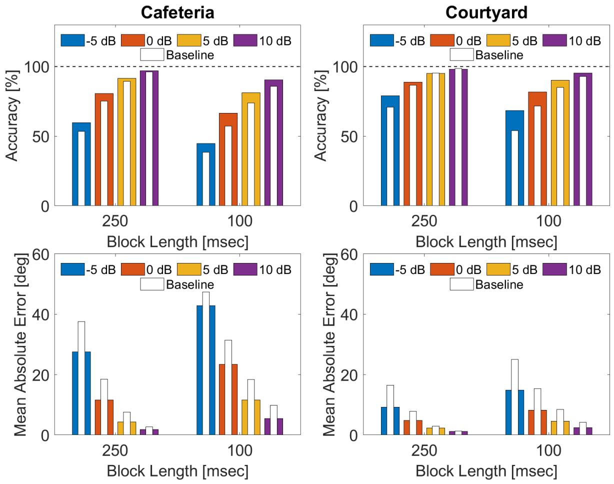

Fig. 5: Accuracy and MAE of the proposed and the baseline method for different SNRs in two reverberant environments. Coloured bars show the performance measures of the proposed method, whereas white bars show the performance of the baseline method.

For both environments, it can be clearly observed that the proposed method outperforms the baseline method in terms of accuracy and MAE. While both methods yield a better performance when using a larger block length, the benefit of joint VAD and DOA estimation appears to be more prominent for smaller block length. This can be explained by the fact that when using a smaller block length the proposed network correctly classifies more blocks with little or no speech activity and blocks that are dominated by noise as non-speech blocks. Since the DOA estimates of these blocks are often inaccurate, discarding them from the evaluation leads to a higher accuracy and a smaller MAE, compared to the baseline method which estimates the DOA in all blocks. This also explains why the benefit is larger at low SNRs, particularly at -5 dB.

It can also be observed that when the accuracy is close to 100%, the MAE is below 5°, which is to be expected. This mainly occurs for the larger block length at higher SNRs, where the speech energy is dominant in the most of the blocks and thus a smaller number of non-speech blocks are discarded by the proposed method. Although the cafeteria with 1300 ms reverberation time and babble noise is the most challenging acoustic condition, the proposed method is able to achieve a smaller MAE compared to the baseline method.

7. CONCLUSION

In this paper, we proposed a hybrid CNN architecture for joint DOA estimation and VAD in a single-talker scenario by exploiting both GCC-PHAT features as well as an auditory-inspired periodicity degree feature. The joint task was realized as a multi-class classification task, where each input feature map was assigned to only one output class. Experimental results in unseen reverberant environments with unseen background noises clearly show that the proposed hybrid CNN outperforms the baseline CNN which only uses GCC-PHAT features.

8. REFERENCES

[1] D. Marquardt and S. Doclo, "Noise power spectral density estimation for binaural noise reduction exploiting direction of arrival estimates," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz NY, USA, Oct 2017, pp. 234–238.

[2] M. Taseska and E. A. P. Habets, "DOA-informed source extraction in the presence of competing talkers and background noise," *EURASIP Journal on Advances in Signal Processing*, vol. 2017, pp. 60, 2017.

[3] K. Adiloğlu, H. Kayser, R. M. Baumgärtel, S. Rennebeck, M. Dietz, and V. Hohmann, "A binaural steering beamformer system for enhancing a moving speech source," *Trends in hearing*, vol. 19, pp. 2331216515618903, 2015.

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

[5] S. Braun, W. Zhou, and E. A. P. Habets, "Narrowband direction-of-arrival estimation for binaural hearing aids using relative transfer functions," in *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz NY, USA, Oct 2015, pp. 1–5.

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

[7] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*, Springer Science & Business Media, 2013.

[8] S. A. Vorobyov, A. B. Gershman, and K. M. Wong, "Maximum likelihood direction-of-arrival estimation in unknown noise fields using sparse sensor arrays," *IEEE Transactions on Signal Processing*, vol. 53, no. 1, pp. 34–43, Jan 2005.

[9] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, Mar 1986.

[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] N. Ma, J. A. Gonzalez, and G. J. Brown, "Robust binaural localization of a target sound source by combining spectral source models and deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2122–2131, Nov 2018.

[12] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, March 2019.

[13] E. L. Ferguson, S. B. Williams, and C. T. Jin, "Sound source localization in a multipath environment using convolutional neural networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, April 2018, pp. 2386–2390.

[14] P. Vecchiotti, E. Principi, S. Squartini, and F. Piazza, "Deep neural networks for joint voice activity detection and speaker localization," in *2018 26th European Signal Processing Conference (EUSIPCO)*, Rome, Italy, Sep. 2018, pp. 1567–1571.

[15] P. Vecchiotti, G. Pepe, E. Principi, and St. Squartini, "Detection of activity and position of speakers by using deep neural networks and acoustic data augmentation," *Expert Systems with Applications*, vol. 134, pp. 53–65, 2019.

[16] S. Graf, T. Herbig, M. Buck, and G. Schmidt, "Features for voice activity detection: a comparative analysis," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, pp. 91, 2015.

[17] A. Josupeit and V. Hohmann, "Modeling speech localization, talker identification, and word recognition in a multi-talker setting," *The Journal of the Acoustical Society of America*, vol. 142, no. 1, pp. 35–54, 2017.

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

[19] J. Anemüller and H. Schoof, "Deep network source localization and the influence of sensor geometry," in *2019 International Congress on Acoustics (ICA)*, Aachen, Germany, 2019.

[20] H. Kayser and J. Anemüller, "A discriminative learning approach to probabilistic acoustic source localization," in *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Juan-les-Pins, France, Sept 2014, pp. 99–103.

[21] M. Omologo and P. Svaizer, "Acoustic event localization using a crosspower-spectrum phase based technique," in *Proceedings of ICASSP '94. IEEE International Conference on Acoustics, Speech and Signal Processing*, Adelaide, Australia, April 1994, vol. ii, pp. II/273–II/276 vol.2.

[22] François Chollet et al., "Keras," `https://keras.io`, 2015.

[23] J. Garofolo, L.Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 11 1992.

[24] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd Annual ACM Conference on Multimedia*, Brisbane, Australia, pp. 1015–1018, ACM Press.

[25] 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.

[26] C. Evers, H. W. Löllmann, H. Mellmann, A. Schmidt, H. Barfuss, P. A. Naylor, and W. Kellermann, "Locata challenge-evaluation tasks and measures," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Tokyo, Japan, Sep. 2018, pp. 565–569.