

SIGNAL-INFORMED DNN-BASED DOA ESTIMATION COMBINING AN EXTERNAL MICROPHONE AND GCC-PHAT FEATURES

Ulrik Kowalk¹, Simon Doclo², and Joerg Bitzer¹

¹Jade University of Applied Sciences, Institute for Hearing Technology and Audiology, Oldenburg, Germany

²University of Oldenburg, Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4all, Oldenburg, Germany

ABSTRACT

Aiming at estimating the direction of arrival (DOA) of a desired speaker in a multi-talker environment using a microphone array, in this paper we propose a signal-informed method exploiting the availability of an external microphone attached to the desired speaker. The proposed method applies a binary mask to the GCC-PHAT input features of a convolutional neural network, where the binary mask is computed based on the power distribution of the external microphone signal. Experimental results for a reverberant scenario with up to four interfering speakers demonstrate that the signal-informed masking improves the localization accuracy, without requiring any knowledge about the interfering speakers.

Index Terms— signal-informed, source localization, GCC-PHAT, binary masking, external microphone

1. INTRODUCTION

In the last decades, a wide range of direction of arrival (DOA) estimation methods using microphone arrays have been proposed, ranging from correlation-based approaches, e.g., exploiting the generalized cross-correlation (GCC) [1, 2], beamforming-based approaches, e.g., steered response power with phase transform (SRP-PHAT) [3] or diagonal unloading [4], subspace-based approaches [5] to approaches based on deep neural networks (DNNs) [6–12]. A specific problem is the localization of a single desired speaker in a multi-talker scenario. Without any knowledge about the desired speaker, calculating a reliable DOA estimate is obviously a challenge [13]. The sparse nature of speech – temporally as well as spectrally – may provide a valuable leverage point. Several authors have proposed the use of masking to guide the DOA estimation towards a desired speaker [7, 8, 10, 11].

In this paper we assume the availability of an external microphone close to the desired speaker (see Fig. 1), which is exploited as an additional source of information (similarly as

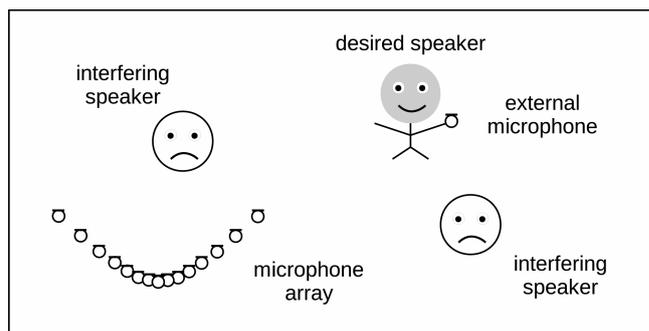


Fig. 1. Scenario with one desired speaker and two interfering speakers. The external microphone is attached to the desired speaker we would like to localize.

in [14, 15]). For example, one could consider a classroom scenario where the teacher (desired speaker) wears a microphone, e.g. to support a student with hearing disabilities, and the objective is to localize the teacher in the presence of multiple interfering speakers (students). We propose to utilize the power distribution from the external microphone signal to compute a binary mask that is applied to the input features of a DNN to calculate a signal-informed DOA estimate for the desired speaker. The algorithm is evaluated under reverberant conditions with a single desired speaker in a multi-talker environment. Results demonstrate that the signal-informed masking improves the localization accuracy, without requiring any knowledge about the acoustic scenario.

2. SIGNAL MODEL

We consider a reverberant scenario with one desired speaker, J interfering speakers, and a small amount of background noise. The speech signals are captured by a microphone array with M microphones and an external microphone close to the desired speaker (see Fig. 1). In the frequency domain, the desired speech signal is denoted by $D(\omega)$, while the m -th microphone signal and the external microphone signal are denoted by $Y_m(\omega)$ and $E(\omega)$, respectively. The $(M+1)$ -dimensional vector of all microphone signals $\mathbf{Y}(\omega) = [Y_0(\omega), Y_1(\omega), \dots, Y_{M-1}(\omega), E(\omega)]^T$, where $(\cdot)^T$

This work was funded by the German Federal Ministry of Education and Research under the funding program "Forschung an Fachhochschulen", Project ID: 13FH6661B6.

denotes transpose, can be written as

$$\mathbf{Y}(\omega) = D(\omega)\mathbf{H}_D(\omega) + \sum_{j=1}^J I_j(\omega)\mathbf{H}_j(\omega) + \mathbf{V}(\omega), \quad (1)$$

where $\mathbf{H}_D(\omega)$ denotes the vector of acoustic transfer functions between the desired speaker and the microphones, $I_j(\omega)$ denotes the j -th interfering speech signal, $\mathbf{H}_j(\omega)$ denotes the vector of acoustic transfer functions between the j -th interfering speaker and the microphones and $\mathbf{V}(\omega)$ denotes the background noise.

3. SIGNAL-INFORMED DOA ESTIMATION EXPLOITING AN EXTERNAL MICROPHONE

The proposed algorithm uses time-domain generalized cross-correlation with phase transform (GCC-PHAT) [1] features as input to a convolutional neural network. We exploit the external microphone signal to generate a binary mask that is applied to the input features. The DOA estimation is formulated as a multi-class classification task with $C=72$ classes representing a set of DOAs with 5° resolution in the horizontal plane. For each audio frame a set of GCC-PHAT features is calculated and a DOA estimate is produced. Figure 2 depicts an overview of the algorithm. The remainder of this section is divided into three parts. In Section 3.1 the input features are described. In Section 3.2 we explain how the external microphone signal can be used to compute a mask and guide the DOA estimation towards the desired speaker. Finally, in Section 3.3, the DNN architecture is presented.

3.1. Input Features

As input features to the DNN we use the well-known GCC-PHAT. The GCC-PHAT between microphones k and l at time lag τ is defined as

$$\gamma_{k,l}(\tau) = \mathcal{F}^{-1} \left\{ \frac{Y_k(\omega) \cdot Y_l^*(\omega)}{|Y_k(\omega) \cdot Y_l^*(\omega)|} \right\} \quad (2)$$

where $(\cdot)^*$ denotes complex conjugation and \mathcal{F}^{-1} denotes the inverse Fourier Transform. Through the PHAT weighting, the GCC-PHAT only depends on the phase difference between the microphone signals, i.e.

$$\gamma_{k,l}(\tau) = \mathcal{F}^{-1} \left\{ e^{i\phi_{k,l}(\omega)} \right\} \quad (3)$$

$$= \mathcal{F}^{-1} \left\{ e^{i \cdot (\arg \{Y_k(\omega)\} - \arg \{Y_l(\omega)\})} \right\} \quad (4)$$

and not on the magnitude spectrum of the sound source. In practice, we limit the time lag τ to the interval $[-\tau_{max}, \tau_{max}-1]$, where τ_{max} is an integer number of samples which depends on the maximum inter-microphone distance of the array. Considering all possible microphone combinations

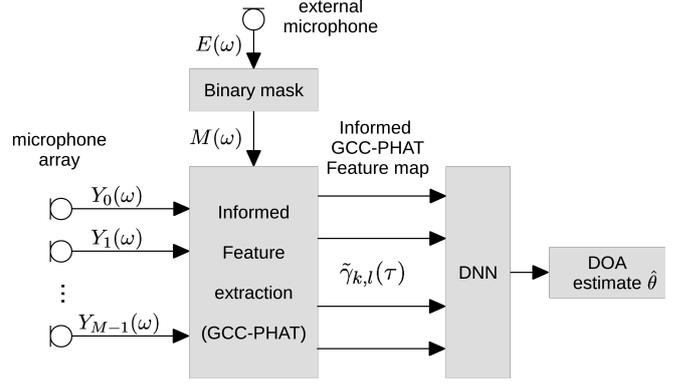


Fig. 2. Block diagram of the signal-informed DOA estimation algorithm

(including $k=l$), a complete GCC-PHAT feature map has dimensions $M \times M \times (2\tau_{max})$. The fact that every source direction results in a distinct GCC-PHAT pattern inspires the usage of a DNN for image classification (see Section 3.3).

3.2. Signal-informed Masking

To guide the algorithm towards localizing the desired speaker in a multi-talker scenario, we propose to utilize the information from the external microphone. Since the external microphone is assumed to be close to the desired speaker, its power relative to the other speakers can be assumed to be large in the external microphone signal. We propose to generate a binary mask by comparing the magnitude of the time-aligned external microphone signal $|E(\omega)|$ to a threshold, i.e.

$$M(\omega) = \begin{cases} 1 & |E(\omega)| \geq P_x(|E(\omega)|) \\ 0 & \text{else} \end{cases} \quad (5)$$

where $P_x(|E(\omega)|)$ denotes the x -th percentile of $|E(\omega)|$. It should be noted that to compute the mask $M(\omega)$, no knowledge about the interfering speakers nor the noise is required. Similar to [11], where a binary mask was applied to raw input phases, we propose to apply the binary mask in (5) to the phase differences and add random noise, i.e.

$$\tilde{\phi}_{k,l}(\omega) = M(\omega) \cdot (\arg \{Y_k(\omega)\} - \arg \{Y_l(\omega)\}) + (1 - M(\omega)) \cdot U(\omega), \quad (6)$$

which are then used to compute an informed GCC-PHAT

$$\tilde{\gamma}_{k,l}(\tau) = \mathcal{F}^{-1} \left\{ e^{i \cdot \tilde{\phi}_{k,l}(\omega)} \right\}. \quad (7)$$

Directly multiplying the mask with the phase differences, i.e. setting x percent of the phase bins to zero, would result in a biased source direction as a phase difference of 0 between two microphones implies equidistance to a source. Therefore, uniformly distributed random noise $U(\omega) \in [0, 2\pi]$ is added

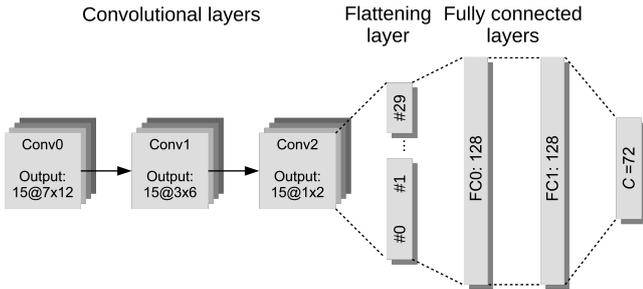


Fig. 3. DNN architecture: 3 convolution layers with dimensionality reduction, 2 fully connected (FC) layers followed by an output layer with $C=72$ classes.

in (6), which does not contribute to the DOA estimation but indirectly guides the estimation towards the spectral components dominated by $E(\omega)$. Because the mask is applied during feature generation, it may be changed without the need for retraining the DNN. Since the external microphone signal is transmitted directly, $E(\omega)$ is time-aligned with the microphone signals $Y_0(\omega)\dots Y_{M-1}(\omega)$ in order to account for the travel time of the sound wave from the source to the microphone array. We approximate this time difference as the time lag that maximizes the cross-correlation between $E(\omega)$ and the central microphone signal of the array.

3.3. DNN architecture

Figure 3 depicts the considered DNN architecture. Like in [6] we use a cascade of 3 convolution layers, but with kernel dimensions of 3×3 . Each stage includes batch normalization, max pooling (2×2 for the first two stages, 3×3 for the last stage), 50% dropout, and a leaky ReLU activation function. Through the continuous dimensionality reduction, the receptive field is gradually broadened, from small details in the feature (3×3) to the whole feature in the end. After a flattening layer, two fully connected layers with 128 neurons are followed by the output layer, resulting in $C=72$ neurons, one for each class. A DOA estimate is defined as the class with the largest value at the output of the DNN. Comprising only 36008 learnable parameters, this architecture is comparatively small.

4. EXPERIMENTAL VALIDATION

In this section we explain the acoustic setup and the training and evaluation of the proposed algorithm. It should be noted that training is only conducted using a single sound source and no masking is applied, whereas in the evaluation there is a desired speaker in a multi-talker environment and the binary mask is applied to guide the DOA estimation towards the desired speaker.

Room dimensions:	$[9.0, 5.0, 3.0] \text{ m} \pm [1.0, 1.0, 0.5] \text{ m}$
Array position:	$[4.5, 2.5, 1.5] \text{ m} \pm [0.5, 0.5, 0.5] \text{ m}$
Source distance:	1.0 - 3.0 m [within boundaries]
Source direction:	$0^\circ : 5^\circ : 355^\circ$
T_{60} :	0.13 s - 1.0 s
SNR:	0 - 30 dB

Table 1. Simulation parameters

4.1. Acoustic Setup

For the experimental validation, we consider a non-uniformly spaced two-dimensional array with $M=15$ microphones, where the microphones are log-log-spaced on an arc, with a width of approximately 0.4 m and a depth of approximately 0.13 m (see Fig. 1 for outline). According to the array geometry we assume $\tau_{max}=12$ smpls. To simulate sound sources with directional cues, we convolve clean monophonic signals with room impulse responses (RIRs), that we generate using *pyroomacoustics* [16]. All simulations are performed using Hann windowed non-overlapping frames with a length of 32 ms at a sampling rate of 8 kHz.

4.2. Training

Training is conducted on single frames, each containing a single sound source. As already mentioned, no masking is applied during training, therefore the external microphone signal is not used. Half of the source signals for training consist of white noise, whereas the other half consist of speech. The speech signals are taken from the “clean” section of the *LibriSpeech* corpus [17], comprising 2700 recordings of male and female speakers. Aiming at achieving a good level of generalization against unseen conditions, we generate training data with a high intrinsic variance for all parameters. Every training signal contains a new set of RIRs with randomized room dimensions, position of the microphone array, position of the sound source in the room, as well as reverberation time. Finally, we add either white noise or babble noise with diffuse-like [18] as well as spatially uncorrelated characteristics at different SNRs. All simulation parameters are given in Table 1. The network is trained using the cross-entropy loss function, together with the Adam optimizer operating at a learning rate of 10^{-4} with mini-batches of 32 training samples and 10^5 training samples per epoch.

4.3. Evaluation

During evaluation, we only consider speech signals. Every scenario consists of one desired speaker and a set of interfering speakers $J \in [0, 1, 2, 4]$, each interfering speaker having the same power as the desired speaker. For every J , the performance is evaluated for 5000 individual scenarios (trials), each 5 s long. The evaluation data are generated with the

same parameter variance as the training data (see Table 1), except for the reverberation time fixed to 0.5 s and the SNR fixed to 20 dB. To prevent sources from overlapping, we impose a minimum angular distance of 5 classes ($\hat{=}25^\circ$) between the desired speaker and the closest interfering speaker, and at least one class ($\hat{=}5^\circ$) between two interfering speakers. The external microphone shares the same coordinates as the desired speaker, but with an offset of 0.2 m on the vertical axis. It is important to note, that in some scenarios, an interfering speaker may be much closer to the microphone array than the desired speaker, resulting in a negative signal-to-interference-ratio at the microphone array.

4.4. Performance Measures

The DOA estimation performance is evaluated in terms of the absolute angular error (in degrees) as

$$\delta = \left| \arg \left\{ e^{i2\pi \cdot (\hat{\theta} - \theta_t) / 360^\circ} \right\} \right| \cdot \frac{360^\circ}{2\pi}, \quad (8)$$

where $\hat{\theta}$ denotes the estimated DOA and θ_t denotes the ground truth DOA (both in degrees). A single DOA estimate is obtained per trial by computing the median of the DOA estimates over all frames that are labeled as speech. Speech frames are defined as frames whose energy in the external microphone signal is larger than 4 dB below the global average.

5. RESULTS

In this section, we present the results for two sets of experiments. First, we investigate the influence of the proposed masking for a different number of interfering speakers J , using P_{50} as the masking threshold. Second, we investigate the influence of the masking threshold P_x for a different number of interfering speakers. The respective combination of J and P_x is referred to as ‘‘condition’’.

Figure 4 shows the impact of the proposed masking. For the condition with one interfering speaker, it can be observed that masking reduces the overall median error by about 14 %. For the conditions with 2 and 4 interfering speakers, masking reduces the overall median error by about 36 % and 28 %, respectively. In terms of the overall mean error, it can be observed that masking results in a reduction of about 45 % for the condition with one interfering speaker. For the conditions with 2 and 4 interfering speakers, masking reduces the overall mean error by about 30 % and 18 %, respectively. Comprehensively, it can be observed that for all considered conditions, using the proposed source-informed features as input to the DNN substantially improves the localization accuracy of the desired speaker.

Figure 5 shows the impact of the masking threshold P_x . For scenarios with a higher number of interfering speakers, a larger threshold appears to yield a benefit. For the condition with one interfering speaker, P_{33} presents itself as the

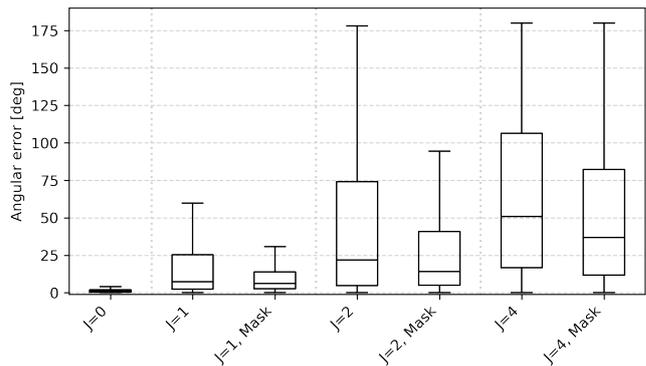


Fig. 4. Angular error for different number of interfering speakers J , with and without masking. Threshold: P_{50}

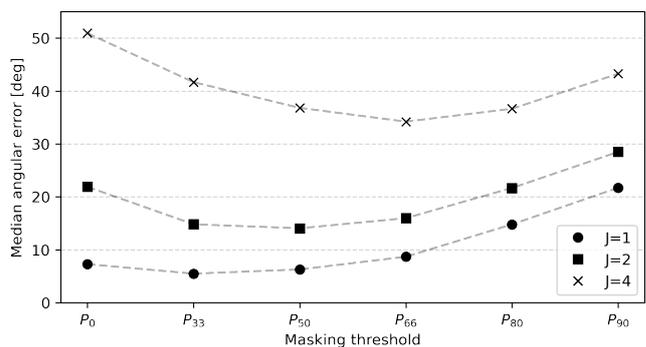


Fig. 5. Impact of the masking threshold on the overall median angular error for different number of interfering speakers J

optimal choice (overall median reduction of about 25 %). For the condition with 2 interfering speakers, where the desired source produces only about 33 % of the sound energy in the room, P_{50} delivers the best results (overall median reduction of about 36 %), while for the condition with 4 interfering speakers, P_{66} delivers the best results (overall median reduction of about 33 %). However, it can also be observed that a masking threshold between P_{50} and P_{66} yields good results for all considered acoustic scenarios.

6. CONCLUSION

This paper has demonstrated the benefit from integrating an external microphone signal in a DNN-based DOA estimation algorithm where a desired speaker is to be localized in the presence of interfering speakers. The proposed algorithm applies a binary mask to the GCC-PHAT input features, where this mask is computed based on the power distribution of the external microphone signal. Experimental results for a variety of acoustic scenarios show that the proposed algorithm significantly improves the DOA estimation accuracy without requiring any knowledge about the acoustic scenario. In future work we will investigate different ways of integrating the external microphone signal for informed DOA estimation.

7. REFERENCES

- [1] C.H. Knapp and G.C. 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, 1976.
- [2] G.C. Carter, "Coherence and time delay estimation," *Proc. of the IEEE*, vol. 75, no. 2, pp. 236–255, 1987.
- [3] J.H. DiBiase, *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*, PhD thesis, Brown University, 2000.
- [4] D. Salvati, C. Drioli, and G.L. Foresti, "Diagonal unloading beamforming in the spherical harmonic domain for acoustic source localization in reverberant environments," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 28, pp. 2001–2012, 2020.
- [5] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.
- [6] S. Chakrabarty and E.A.P. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 136–140.
- [7] Z.-Q. Wang, X. Zhang, and D. Wang, "Robust speaker localization guided by deep learning-based time-frequency masking," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 178–188, 2018.
- [8] Z. Wang, J. Li, and Y. Yan, "Target speaker localization based on the complex watson mixture model and time-frequency selection neural network," *Applied Sciences*, vol. 8, no. 11, 2018.
- [9] 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, 2019.
- [10] W. Zhang, Y. Zhou, and Y. Qian, "Robust DOA estimation based on convolutional neural network and time-frequency masking," in *Proc. Interspeech*, 2019, pp. 2703–2707.
- [11] W. Mack, U. Bharadwaj, S. Chakrabarty, and E.A.P. Habets, "Signal-aware broadband DOA estimation using attention mechanisms," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 4930–4934.
- [12] 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)*, 2020, pp. 566–570.
- [13] S. Sivasankaran, E. Vincent, and D. Fohr, "Keyword-based speaker localization: Localizing a target speaker in a multi-speaker environment," in *Proc. Interspeech*, 2018, pp. 2703–2707.
- [14] M. Farmani, M.S. Pedersen, Z.-H. Tan, and J. Jensen, "Informed sound source localization using relative transfer functions for hearing aid applications," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 611–623, 2017.
- [15] D. Fejgin and S. Doclo, "Comparison of binaural RTF-vector-based direction of arrival estimation methods exploiting an external microphone," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2021, pp. 241–245.
- [16] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 351–355.
- [17] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.
- [18] E.A.P. Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep.*, vol. 2, no. 2.4, pp. 1, 2006.