2D Acoustic Source Localisation Using Decentralised Deep Neural Networks on Distributed Microphone Arrays

Stijn Kindt, Alexander Bohlender, Nilesh Madhu [1]

IDLab, Department of Electronics and Information Systems, Ghent University - imec, Ghent, Belgium
Email: {stijn.kindt,alexander.bohlender,nilesh.madhu}@ugent.be

Abstract

This paper takes a previously proposed convolutional recurrent deep neural network (DNN) approach to direction of arrival (DoA) estimation and extends this to perform 2D localisation using distributed microphone arrays. Triangulation on the individual DoAs from each array is the most straightforward extension of the original DNN. This paper proposes to allow more cooperation between the individual microphone arrays by sharing part of their neural network, in order to achieve a higher localisation accuracy. Two strategies will be discussed: one where the shared network has narrowband information, and one where only broadband information is shared. Robustness against slight clock offsets between different arrays is ensured by only sharing information at deeper layers in the DNN. The position and configuration of the microphone arrays are assumed known, in order to train the network. Simulations will show that combining information between neural network layers has a significant improvement over the triangulation approach.

1 Introduction

Wireless acoustic sensor networks (WASNs) is a field that is gaining a lot of attraction recently. Multiple microphones or microphone arrays are distributed around a room, in order to have a bigger coverage. This can aid in a variety of applications [1], like speech enhancement [2], beamforming [3], environment monitoring [4], hands-free communication and speaker localisation [5].

When only a single microphone array is present, it is very possible to estimate the direction of arrival (DoA) of the signal coming from a speaker as long as the direct path from the source to the array is dominant. In fact, a lot of research has already been done on this topic and an overview of the classical methods for DoA estimation may be found in [6, Ch. 6, P. 135-170]. Neural networks led to recent advances in the filed [7–10]. However, DoA estimation does not provide the distance of the source to the microphone array, and thus does not localise the source in 2D space. Techniques based on the use of data-based approaches such as deep neural networks (DNN), have tried to address this problem [11, 12]. However, distance information can not easily be inferred from the phase or amplitude. Other measures like the coherent-to-diffuse power ratio need to be used [13].

This is where a WASN, composed of distributed microphone arrays can help. Such configurations capture the source signal from widely different positions in the room, thereby permitting a better localisation estimate in the 2D space when the data from these arrays is combined.

One should, however, also be aware of the extra challenges that come with localisation using the distributed nodes in a WASN. Typically the different nodes (microphones or microphone arrays) are only weakly synchronised, meaning that the clock signals can differ fractions of a sample, up to a few samples. Comparing signals with unsynchronised clocks will lead to incorrect localisation [14–16]. WASNs generally also have limitations on the bandwidth of each node.

In ad-hoc WASNs, an additional challenge occurs: the positions of the sensor nodes are also unknown. In that case, both the source and the array positions need to be estimated, mostly leading to an iterative approach [17–19], making the system a lot more complex. This will not be further discussed here. The focus of this paper is on the use of WASNs with a known configuration (i.e., the location and configuration of each node is known and remains static).

[1] This work is supported by the Research Foundation - Flanders (FWO) under grant numbers G081420N and 11G0721N

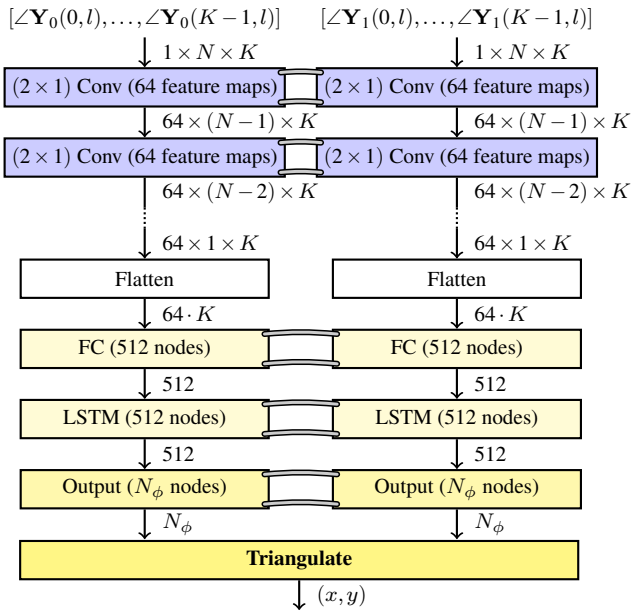


Figure 1: This figure shows the triangulation approach, consisting of two networks that performs DoA estimation. The grey bands represent that the weights of these layers are identical. After each convolutional layer batch normalisation is done, dropout with rate 0.5 is used before each FC layer and ReLU activation is used after each hidden layer.

Triangulation is an example of source localisation in 2D space using a WASN with multiple microphone arrays: each array estimates a DoA, and these estimates can be aggregated under the appropriate geometric constraints to provide an intersection point where the source is located [20–22].

Triangulation approaches easily fit the synchronisation and bandwidth limitations of WASNs: Triangulation is inherently robust against asynchronous clocks on different nodes, since the DoA computations can be done independently on each node. furthermore, by only sending the DoA to the central unit, the throughput is very low. However, not sharing more information between nodes limits the potential for higher localisation accuracy. Also, triangulation fails when the DoAs do not intersect, which can occur e.g., when the cumulative errors in the DoA estimates are large and in opposing directions.

This paper proposes three architectures that extend a deep convolutional recurrent neural network DoA approach [10]. The model will be expanded from using one single microphone array to being a WASN with two microphone arrays. A first, rather straightforward, approach triangulates two DoAs from the two different arrays. This will serve as the reference method. In order to improve the localisation accuracy, we also propose two architectures where the DNN structure is suitably modified in order to mix information between the different nodes at different depths in the DNN. These architectures will be referred to as co-operative

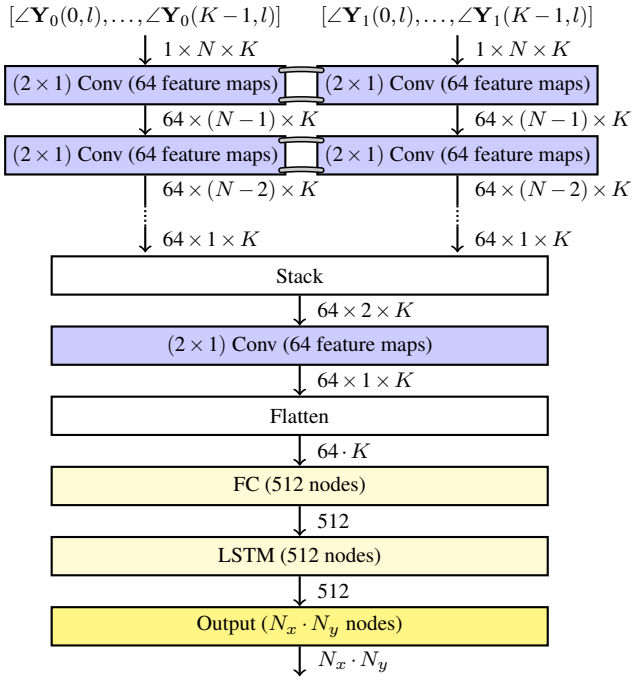


Figure 2: The narrowband mixing co-operative localisation architecture (NM-CLA). The architecture from Figure 1 is altered to mix information of both arrays right after the intra-array convolutional layers, by performing an inter-array convolution. This is narrowband in the sense that the frequency dimension is still present at that time.

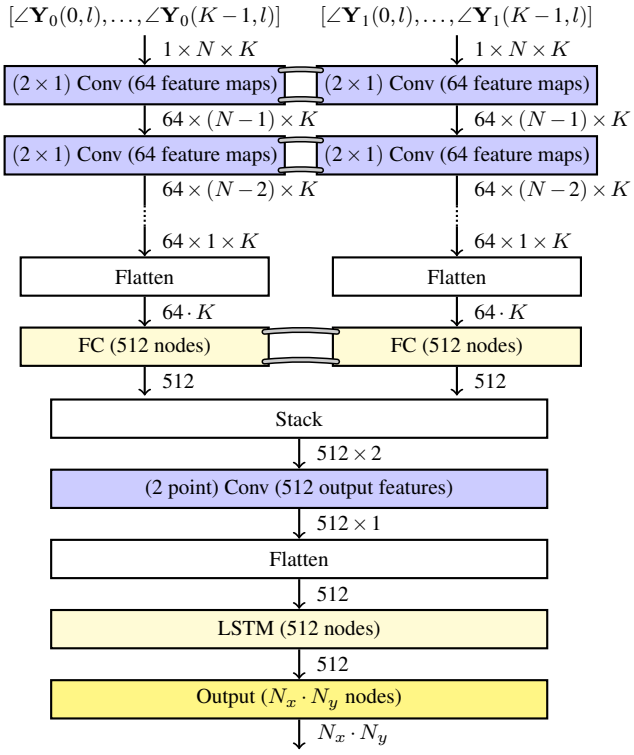


Figure 3: the broadband mixing co-operative localisation architecture (BM-CLA). Here the mixing of information happens after the FC layer, and is also performed by an inter-array convolution.

localisation architectures (CLAs).

The architectures are designed such that they do not directly combine the information at the microphone level between the two arrays. Instead, the information exchange occurs at a deeper layer. Thereby they are robust to clock offsets, even without explicitly training for these offsets. In order to share the information, the different nodes send it to a central processing unit. While such information sharing requires a larger bandwidth, it may be an acceptable tradeoff against the higher localisation accuracy obtained.

Further, the output of the two proposed architectures will be changed. Instead of yielding two DoAs for subsequent triangulation, the network will directly output a 2D estimate of the source position. This solves the problem where triangulation does not come up with a intersection point.

The rest of the paper will be structured as follows. In Section 2, the conventions for the signals at different microphones will be laid out, as well as the way in which the architectures predict the positions. The reference method will also be explained in more detail and then the two co-operative architectures will be shown. In Section 3, an evaluation of all three methods will be given. It will also be shown that the proposed methods work equally well under the condition of weakly synchronised clocks. Section 4 concludes the paper.

2 Models

2.1 Signal Model

First the conventions of this paper will be described. M microphone arrays will be used, which all consist of N microphones. The signals at microphone n of microphone array m at time sample i is the combination of J target speakers $x_{m,n,j}(i)$ and noise $v_{m,n}(i)$. $x_{m,n,j}(i)$ consists of the direct path of the speech signals as well as the reverberation of the room. In the short-time Fourier transform (STFT) domain, the microphone signals are then written as:

$$Y_{m,n}(k,l) = \sum_{j=0}^{J-1} X_{m,n,j}(k,l) \quad + V_{m,n}(k,l) \tag{1}$$

where k is the frequency index and l the time index of the STFT. This representation is often chosen for speaker localisation and separation due to the well-known properties of sparsity and disjointness of speech signals in the STFT representation [23].

In vector notation, the signals at microphone array m are represented by $\mathbf{Y}_m(k,l) = [Y_{m,0}(k,l),\ldots,Y_{m,N-1}(k,l)]^T$ and the signals from all the microphone arrays are represented by $\mathbf{Y} = [\mathbf{Y}_0(k,l),\ldots,\mathbf{Y}_{M-1}(k,l)]^T$.

2.2 Prediction models

The starting point of all proposed methods is the convolutional recurrent DNN proposed by Bohlender et al. [10], which extends the convolutional DNN (CNN) based approach of [9] with temporal context. Their network uses a single microphone array to estimate the direction(s) of arrival (DoA) of target speaker(s). It is assumed, here, that the target speaker is in the far field of the microphone array, meaning that the amplitude carries less information regarding the speaker location(s). That is why the input features of the neural network are the phases for all N microphones and K frequencies: $[\angle\mathbf{Y}_m(0,l),\ldots,\angle\mathbf{Y}_m(K-1,l)]$.

To execute the 2D localisation, the newly proposed systems make use of two microphone arrays ($M=2$), whose relative positions are known. This means that the distance between the arrays and their relative orientation are fixed. However, the position in the room and the orientation with respect to the room can be arbitrary but static. This knowledge is needed because the relation between the input features and the positions of the arrays is not linear. The localisation will be with respect to the centre of the chosen array configurations. For absolute localisation, the array positions should therefor also be known.

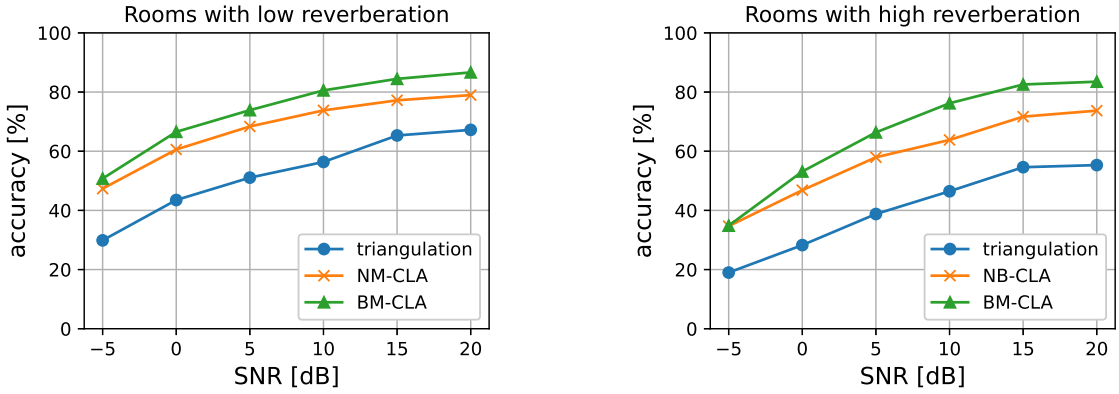


Figure 4: Accuracy of the localisation approaches for different SNRs. A position is considered to be correctly estimated if it is within one meter of the actual source position. The left figure shows the accuracy for low reverbrant rooms (RT60 $< 0.5\,\mathrm{s}$). On the right, the figure shows the accuracy of the highly reverberant rooms.

The phases $\angle\mathbf{Y}(k,l)$ are still chosen to be the input features, despite the fact that the signal amplitudes at different arrays now carry useful information. This is done in order to have a fair comparison between the triangulation and the CLAs. However, in order to still satisfy the far field assumption, combined with spacial aliasing concerns, the input features from different microphone arrays are kept separate in the first few network layers.

2.3 Reference method

The baseline of this paper is a triangulation approach using the DoA architecture of [10]. Two arrays are used individually to yield two independent DoA estimates. The estimated 2D source position is the intersection point of the rays extended in the direction of the estimated DoAs from each array. The triangulation architecture is shown in Figure 1. This can be implemented as an almost completely distributed approach: all processing can be done on the individual microphone arrays separately, and only the two DoAs should be sent to the central processing hub which does the triangulation. This implementation has a low demand on the bandwidth to the central processor.

We present below a brief overview of the DNN used for estimating the DoA by each array. Further details may be found in the original base paper [9], and [10] for the temporal context. First, convolutional (conv) layers are used to combine information from different microphones. This is done in multiple layers, where each layer only combines information form two neighbouring microphones at a time. At this point, each frequency is processed separately. The output of the convolutional layers is then put through a fully connected (FC) layer followed by a recurrent layer: a long short-term memory (LSTM) layer. The FC layer is the first layer which has information from all frequencies. After the LSTM, the output layer classifies in which of the N_ϕ angular sectors the source lies. More output classes can allow for a finer DoA estimate. This is done for the two arrays separately. Triangulation then gives the intersection point at which the reference method estimates the location of the target speaker.

2.4 Proposed methods

This paper proposes to increase the localisation accuracy by mixing information from both arrays within the DNN. However, for practical implementations, we need to ensure that there are no strict synchronisation requirements between the two arrays. This is done by combining the information at a deeper layer of the DNN instead of at the input feature level. This omits the need for asynchronous training data. Two co-operative localisation architectures are discussed: narrowband mixing CLA (NM-CLA) and broadband mixing CLA (BM-CLA). The narrowband variant combines the information of the different microphone arrays right after the last convolutional layer of both arrays, doing an inter-array convolution instead of an intra-array convolution. Each frequency bin up until this point is still considered separate, which explains the naming choice.

The output is also defined as a classification problem, where each class represents a rectangular region in 2D space. The output of the DNN may be interpreted as the the probability of the speaker being present within that region. The number of classes depends on how precisely we want to localise the speaker. the total number of classes is given by $N_x \cdot N_y$ where N_x represents the number of regions in the x direction and N_y that of the y direction. In Figure 2, the described architecture is shown.

This NM-CLA however does increase the minimum bandwidth requirements substantially: each node has to send $64 \cdot K$ features to the central node. In this work, the STFT has $K = 257$ frequency bins.

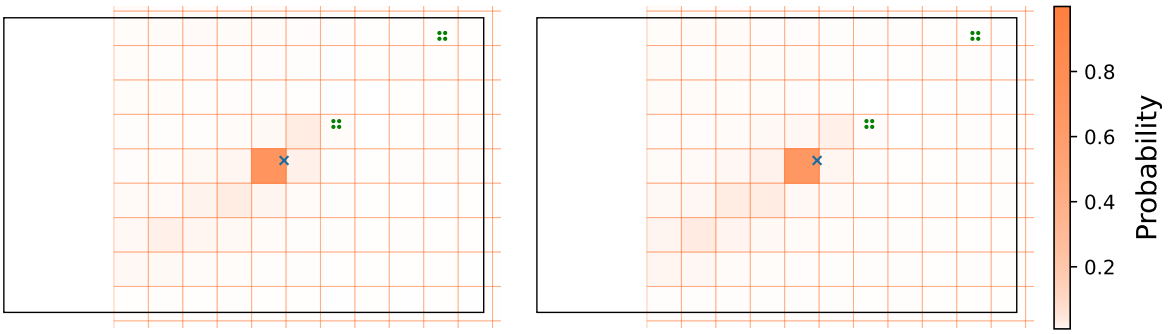
The second proposed variant, the BM-CLA, lowers the bandwidth requirements compared to NM-CLA. In this case, there is an intra-array convolutional layer after the FC layer. This FC layer already combines the different frequency bins, which makes the newly added convolutional layer work on broadband information. As an alternative, it was also tested to use a FC layer instead. The stack operation should then also be changed to a flatten operations, where the output dimension is then 1024. The convolutional layer was empirically found to perform slightly better. The output of the broadband architecture is again a classifier with $N_x \cdot N_y$ classes. This architecture only needs to send 512 features per microphone array to the central unit. The BM-CLA is depicted in Figure 3. Both CLAs do not increase the amount of trainable parameters with that much, since both of them only add one convolutional layer, which is small compared to the FC layer already present in the triangulation networks. The amount of parameters are 10.6×10^6 for triangulation, 10.7×10^6 for NM-CLA and $11,2 \times 10^6$ for BM-CLA.

In a WASN, the clocks of different nodes (here microphone arrays) cannot be assumed to be perfectly synchronised. Triangulation approaches are inherently robust against asynchronous nodes. For the proposed architectures, this needs to be verified. Section 3 will also present some results where asynchronicity is simulated.

3 Evaluation

3.1 Training

The training data set is generated in a similar manner as described in [10]. We want to account for time-variant source activity. This includes moving sources, or speakers becoming inactive and new speakers becoming active. A Markov model, $A_j(t) \in \{0,1\}$, is used to generate this dynamic setting. $A_j(t)$ indicates if source number j is active ($A_j(t) = 1$) or not ($A_j(t) = 0$). The training



(a) Synchronised clocks (b) Asynchronous clocks

Figure 5: A depiction of the broadband mixing co-operative source estimation. The left figure is the case where the network is perfectly synchronised. On the right, the clock signals of both arrays differ by 10 samples. The big black box represents the room. The green dots are single microphones: two microphone arrays with 4 microphones each, are present. The blue cross is the true source position. The orange boxes are the possible areas where the co-operative DNNs can localise the source (only locations within the room are depicted). the opaqueness indicates how high the probability is that the network has given to the corresponding location.

data set contains both instances where $J \in \{0, 1, 2\}$. The average Markov transition probability between the two states is set so that there is on average one transition per 1.5s. Each time a source becomes active, it is assigned a random 2D position within the room to simulate source movement. The source signals are randomly sampled speech signals coming from the TIMIT [24] or PTDB-TUG [25] databases.

The training set consists of 10 different room dimensions paired with thier own reverberation times (RT60). All the room impulse responses are generated using pyroomacoustics [26]. Spatially diffuse and temporally uncorrelated noise is added with SNRs ranging from 0 dB till 30 dB. During training, the Adam optimiser is used to minimise the binary cross entropy cost function. The batch size is chosen to be 20, where each sample consists of a sequence generated by the Markov model of length 2 s.

3.2 Evaluation

For the evaluation, two microphone arrays are used that are spaced 2 meters apart. Both arrays have four microphones, placed at each corner of a square of side 21mm. The output of the model can classify $N_x \cdot N_y$ different locations, where for this evaluation $N_x = N_y = 16$ is chosen. Each class represents a square of $0.5\,\text{m} \times 0.5\,\text{m}$, where the network output should ideally be 1 for the class where the true source position lies. For our configuration this restricts the true source position to a maximum of 4 m from the centre of the two arrays in the x and y directions.

The evaluation is carried out in unseen room dimensions and random reverberation times between 0.2 s and 0.8 s. 6 SNR levels are simulated between -5 dB till 20 dB, in steps of 5 dB. For each SNR and architecture, 1000 simulations are carried out. The results are divided in two subsets: one with lower reverberation times (RT60 $<$ 0.5 s), and one with higher reverberation times (RT60$\geq$ 0.5 s). For the evaluation, we focus on the case where only one source is active.

The results can be seen in Figure 4. A localisation is deemed successful if the estimate is within one meter of the real speaker location. The figure clearly shows that both proposed CLAs have significantly higher localisation accuracies at all the SNR levels compared to the reference triangulation approach. This trend is even greater in the highly reverberant and high SNR case. One specific case where the CLAs outperform triangulation substantially, is where the source and the microphone arrays lie on the same line. The intersection point can then be non existing, even without error in the DoA estimates, due to their discrete nature.

Another interesting result is that BM-CLA outperforms NM-CLA in accuracy at every SNR. This in combination with the lower bandwidth requirements makes BM-CLA superior to NM-CLA. This is somewhat surprising, since we hypothesised that microphone arrays sharing more information would increase the accuracy. One possible explanation could be that combining the narrowband information between microphone arrays, can increase the risk of spatial aliasing since the arrays are far apart. In the broadband variant, the DNN is already forced to mix all the frequencies before the information from different arrays are combined, which reduces the spatial aliasing problem.

3.3 Robustness against clock asynchronicity

In order to be able to deploy the proposed systems, they should also be robust against a slight misalignment of the clock signals as it is hard to perfectly synchronise two different nodes in WASNs. The evaluation process is done by repeating the experiment from Section 3.2, where subsample delays are added to the RIRs of the second microphone array. The delays are uniformly sampled from 0 to 2 samples. The accuracy plots with this added asynchronicity are almost identical to those of Figure 4, indicating that the proposed methods are inherently robust to sampling inaccuracies between the nodes. Instead of showing this result, therefore, we consider a specific case from BM-CLA with and without perfect synchronicity which are shown in Figure 5. The room is 7 by 4.3 m, and 2.6 m high. The RT60 is 0.72 s and diffuse noise is added 15 dB SNR. Here the clock of the second array is actually 10 samples apart from the first array. It is clear that in both cases, the network gives the highest probability to the same (correct) position. Similar results are present for different scenarios.

4 Conclusion

This paper showed that the localisation accuracy in a WASN can be significantly increased by sharing information early between microphone arrays, compared to triangulation. This is done by expanding upon a convolutional recurrent DNN based approach for DoA estimation. Two different co-operative multi-array localisation methods were discussed and compared: NM-CLA and BM-CLA. NM-CLA mixes information between microphones where narrowband information is still present, while BM-CLA only does the inter-array mixing after the broadband information has already been mixed by the individual microphone arrays. BM-CLA has the best accuracy and also has a lower bandwidth requirement on the WASN. Both CLAs are robust against small deviations in clock synchronicity between different nodes. This comes inherently since the features between nodes are only mixed at deeper stages of the DNN, where they are more abstract and no longer dependent on the exact clock samples. Future work includes extending the proposed methods for use in ad-hoc WASN applications, including amplitude information for an even greater localisation accuracy and doing mask based source separation, similar to, e.g. [27].

References

[1] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *2011 18th IEEE symposium on communications and vehicular technology in the Benelux (SCVT)*, pp. 1–6, IEEE, 2011.

[2] A. Bertrand, J. Callebaut, and M. Moonen, "Adaptive distributed noise reduction for speech enhancement in wireless acoustic sensor networks," in *Proc. of the International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2010.

[3] S. Markovich-Golan, A. Bertrand, M. Moonen, and S. Gannot, "Optimal distributed minimum-variance beamforming approaches for speech enhancement in wireless acoustic sensor networks," *Signal Processing*, vol. 107, pp. 4–20, 2015.

[4] F. Alías and R. M. Alsina-Pagès, "Review of wireless acoustic sensor networks for environmental noise monitoring in smart cities," *Journal of sensors*, vol. 2019, 2019.

[5] M. Cobos, F. Antonacci, A. Alexandridis, A. Mouchtaris, and B. Lee, "A survey of sound source localization methods in wireless acoustic sensor networks," *Wireless Communications and Mobile Computing*, vol. 2017, 2017.

[6] U. and Heute and C. Antweiler, *Advances in digital speech transmission*. John Wiley & Sons, 2008.

[7] P. Pertilä and E. Cakir, "Robust direction estimation with convolutional neural networks based steered response power," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6125–6129, IEEE, 2017.

[8] Z.-Q. Wang, X. Zhang, and D. Wang, "Robust speaker localization guided by deep learning-based time-frequency masking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 178–188, 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] A. Bohlender, A. Spriet, W. Tirry, and N. Madhu, "Exploiting temporal context in cnn based multisource doa estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[11] H. Sundar, W. Wang, M. Sun, and C. Wang, "Raw waveform based end-to-end deep convolutional network for spatial localization of multiple acoustic sources," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4642–4646, IEEE, 2020.

[12] T. Gburrek, J. Schmalenstroeer, A. Brendel, W. Kellermann, and R. Haeb-Umbach, "Deep neural network based distance estimation for geometry calibration in acoustic sensor networks," in *2020 28th European Signal Processing Conference (EUSIPCO)*, pp. 196–200, IEEE, 2021.

[13] A. Brendel and W. Kellermann, "Distributed source localization in acoustic sensor networks using the coherent-to-diffuse power ratio," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 61–75, 2019.

[14] D. Cherkassky, S. Markovich-Golan, and S. Gannot, "Performance analysis of mvdr beamformer in wasn with sampling rate offsets and blind synchronization," in *2015 23rd European Signal Processing Conference (EUSIPCO)*, pp. 245–249, IEEE, 2015.

[15] W. Su and I. F. Akyildiz, "Time-diffusion synchronization protocol for wireless sensor networks," *IEEE/ACM transactions on networking*, vol. 13, no. 2, pp. 384–397, 2005.

[16] Y.-C. Wu, Q. Chaudhari, and E. Serpedin, "Clock synchronization of wireless sensor networks," *IEEE Signal Processing Magazine*, vol. 28, no. 1, pp. 124–138, 2010.

[17] T. Gburrek, J. Schmalenstroeer, and R. Haeb-Umbach, "Iterative geometry calibration from distance estimates for wireless acoustic sensor networks," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 741–745, IEEE, 2021.

[18] A. Plinge, F. Jacob, R. Haeb-Umbach, and G. A. Fink, "Acoustic microphone geometry calibration: An overview and experimental evaluation of state-of-the-art algorithms," *IEEE Signal Processing Magazine*, vol. 33, no. 4, pp. 14–29, 2016.

[19] T. Gburrek, J. Schmalenstroeer, and R. Haeb-Umbach, "Geometry calibration in wireless acoustic sensor networks utilizing doa and distance information," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, pp. 1–17, 2021.

[20] Á. Lédeczi, G. Kiss, B. Feher, P. Volgyesi, and G. Balogh, "Acoustic source localization fusing sparse direction of arrival estimates," in *2006 International Workshop on Intelligent Solutions in Embedded Systems*, pp. 1–13, IEEE, 2006.

[21] A. Griffin, A. Alexandridis, D. Pavlidi, Y. Mastorakis, and A. Mouchtaris, "Localizing multiple audio sources in a wireless acoustic sensor network," *Signal Processing*, vol. 107, pp. 54–67, 2015.

[22] S. V. Sibanyoni, D. T. Ramotsoela, B. J. Silva, and G. P. Hancke, "A 2-d acoustic source localization system for drones in search and rescue missions," *IEEE Sensors Journal*, vol. 19, no. 1, pp. 332–341, 2019.

[23] S. Rickard and O. Yilmaz, "On the approximate w-disjoint orthogonality of speech," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. I–529, IEEE, 2002.

[24] J. S. Garofolo, "Timit acoustic phonetic continuous speech corpus," *Linguistic Data Consortium, 1993*, 1993.

[25] G. Pirker, M. Wohlmayr, S. Petrik, and F. Pernkopf, "A pitch tracking corpus with evaluation on multipitch tracking scenario," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[26] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 351–355, IEEE, 2018.

[27] A. Bohlender, A. Spriet, W. Tirry, and N. Madhu, "Neural networks using full-band and subband spatial features for mask based source separation," in *2021 29th European Signal Processing Conference (EUSIPCO)*, IEEE, accepted for publication in 2021.