SAMbA: Speech enhancement with Asynchronous ad-hoc Microphone Arrays

Nicolas Furnon, Romain Serizel, Slim Essid, Irina Illina

SAMbA: Speech enhancement with Asynchronous ad-hoc Microphone Arrays

Nicolas Furnon[1]*, Romain Serizel*[1]*, Slim Essid*[2]*, Irina Illina*[1]

[1]Université de Lorraine, CNRS, Inria, Loria, F-54000 Nancy, France
[2]LTCI, Télécom Paris, Institut Polytechnique de Paris, Palaiseau, France
[1]{firstname.lastname}@loria.fr, [2]slim.essid@telecom-paristech.fr

Abstract

Speech enhancement in ad-hoc microphone arrays is often hindered by the asynchronization of the devices composing the microphone array. Asynchronization comes from sampling time offset and sampling rate offset which inevitably occur when the microphones are embedded in different hardware components. In this paper, we propose a deep neural network (DNN)-based speech enhancement solution that is suited for applications in ad-hoc microphone arrays because it is distributed and copes with asynchronization. We show that asynchronization has a limited impact on the spatial filtering and mostly affects the performance of the DNNs. Instead of resynchronising the signals, which requires costly processing steps, we use an attention mechanism which makes the DNNs, thus our whole pipeline, robust to asynchronization. We also show that the attention mechanism leads to the asynchronization parameters in an unsupervised manner.

Index Terms: speech enhancement, ad-hoc microphone arrays, asynchronization

1. Introduction

Due to their increased number of microphones, their spatial coverage and their flexibility of use, ad-hoc microphone arrays offer great potential to speech enhancement. This potential however may be limited by a series of challenges. One of the main challenges is the need for a distributed strategy which does not rely on a fusion center as most of classic beamformers do. Distributed algorithms have been proposed for speech enhancement in ad-hoc microphone arrays [1, 2, 3, 4] and recently, a deep neural network (DNN)-based distributed solution has also been introduced to combine the increased modelling capacity of DNNs with the flexibility of use of ad-hoc microphone arrays [5]. Besides, because the microphones embedded in different devices do not share the same hardware and software implementations, they are acquired at different sampling rates (SRs), causing a sampling rate offset (SRO), and triggered at different starting times, causing a sampling time offset (STO). These phenomena cause asynchronization [6, 7]. Asynchronization can have a negative impact on speech enhancement [8], especially on solutions relying on an accurate estimation of the direction-of-arrival, like the minimum variance distortionless response beamformer [9, 10, 11].

Solutions to asynchronization can be broadly classified into two categories. In the first category, specific signals are sent among the nodes. These signals can be either calibration signals [8, 12] or time stamps [13, 6, 7]. The second category gathers so-called *blind* approaches, because no signals are exchanged but the ones captured by the microphones of the nodes. Out of these observations, the SRO and STO can be estimated and compensated for based on the coherence [14, 15], the correlation [16] or the cross-correlation [17, 18, 15, 19] between signals of different nodes. These solutions proved to be efficient, but they suffer from two limitations. The first one is that they require extra computing steps, which might overload the small devices of ad-hoc microphone arrays and add some latency to the processing. It is here interesting to note that some works showed that asynchronization could be tolerated in speech enhancement tasks without any attempt to resample the signals [20, 21]. The second limitation is that none of these works studies the impact of asynchronization on DNNs although DNNs take a more and more important place in speech enhancement.

In this paper, we propose to study the impact of asynchronization on a speech enhancement solution for ad-hoc microphone arrays based on DNNs. We show that the impact of SR and sampling time offsets on spatial filtering is limited, but that their impact on the DNN performance is not negligible. To cope with this, instead of resampling the signals, we decide to use an attention mechanism which implicitly realigns the input signals of the DNN. This avoids an explicit search of the asynchronization parameters.

This paper is organized as follows. In Section 2 we describe the problem, the notations used throughout the paper and we introduce our speech enhancement system. The experimental setup is described in Section 3. In Section 4 we analyse the impact of STO and SRO on our speech enhancement system. In Section 5, we introduce a solution to cope with the asynchronization effects on the DNN performance in our system. Lastly, Section 6 concludes this paper.

2. Problem formulation

2.1. Notations

In the following, signals are considered in the short-time Fourier transform (STFT) domain, where time and frequency indices are dropped for the sake of conciseness. Bold lowercase letters represent vectors. Bold uppercase letters represent matrices. Regular lowercase represent scalars. We consider an ad-hoc microphone array of K nodes of M_k microphones each. The m-th microphone of the k-th node records a noisy mixture $y_{k,m} = s_{k,m} + n_{k,m}$ according to an additive noise model, where $s_{k,m}$ and $n_{k,m}$ are respectively the target speech and the noise components recorded by the microphone. The signals recorded by node k are stacked in a vector $\mathbf{y}_k = [y_{k,1}, \cdots, y_{k,M_k}]^T$.

The SRO of a node k relatively to a reference node will be denoted by ϵ_k. The STO of a node k relatively to a reference node will be denoted by τ_k.

2.2. Distributed speech enhancement in ad-hoc microphone arrays

In a previous work, we introduced a distributed speech enhancement system for ad-hoc microphone arrays, called Tango [5].

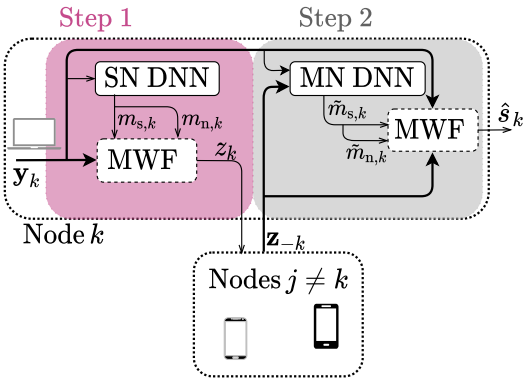


Figure 1: *Graphical representation of our distributed speech enhancement solution. Bold arrows represent multichannel signals, simple arrows represent single-channel signals.*

It processes in two steps, highlighted in Figure 1. In the first step, at each node k, a multichannel Wiener filter (MWF) $\mathbf{w}_{kk}$ is applied on the local signals $\mathbf{y}_k$. To do this, a single-node DNN (SNDNN) is used to predict a time-frequency (TF) mask m_k out of the reference signal $y_{k,1}$. The TF mask is used to compute the spatial covariance matrices of the speech and noise required by the spatial filter.

Filtering the mixture with this beamformer yields a so-called compressed signal $z_k = \mathbf{w}_{kk}^H \mathbf{y}_k$. The compressed signals are exchanged among nodes, so node k receives $K-1$ compressed signals $\mathbf{z}_{-k}$: $\mathbf{z}_{-k} = [z_1, ..., z_{k-1}, z_{k+1}, ..., z_K]^T$. In the second step, a global MWF $\mathbf{w}_k$ is applied on $\tilde{\mathbf{y}}_k = \left[\mathbf{y}_k^T, \mathbf{z}_{-k}^T\right]^T$. The compressed signals $\mathbf{z}_{-k}$ are used for the spatial filtering operation, but they are also fed to a multi-node DNN (MNDNN) to predict the TF mask $\tilde{m}_k$ required by the spatial filter. The spatial filters at both filtering steps are computed following the rank-1 generalized eigenvalue decomposition (GEVD) of the covariance matrices of the mixture and of the noise proposed by Serizel et al. [22].

We showed that this algorithm could efficiently process the spatial information conveyed by the compressed signals and outperforms an oracle voice activity detector (VAD)-based MWF [23]. We also showed that it performs comparatively well to FaSNet [24], while allowing for a trade-off between noise reduction and speech distortion, and relying on a much simpler DNN architecture [5]. For these reasons, we continue using this system for the current work.

3. Experimental setup

3.1. Signal setup and DNN settings

All the signals are sampled at 16 kHz and last between 5 s and 10 s. The STFT is computed with a Hann window of 32 ms with an overlap of 16 ms. The convolutional recurrent neural network (CRNN) architecture is composed of three convolutional layers followed by a recurrent layer and a fully-connected layer. The convolutional layers have 32, 64 and 64 filters, with kernel size 3×3 and stride 1×1. Each convolutional layer is followed by a batch normalisation and a maximum-pooling layer of kernel size 4×1 so that no pooling is applied over the time axis. The recurrent layer is a 256-unit GRU. The fully-connected layer has 257 units with a sigmoid activation function. The input of the model are the magnitudes of the STFT windows of 21 consecutive frames and the ground truth labels are the corresponding frames of the ideal ratio mask. At test time, only the middle frame of the predicted window is considered to estimate the mask, so sliding windows of the input are fed to the DNN. The mask of the whole signal is predicted before being used to enhance the speech in a batch mode.

3.2. Training and evaluation data

The data used to train and evaluate our systems is extracted from the DISCO dataset.[1] Room impulse responses in shoebox-shaped rooms are simulated. The rooms have a length, width and height randomly picked in the ranges $[\![3; 8]\!]$ m, $[\![3; 5]\!]$ m and $[\![2; 3]\!]$ m respectively. 2 sources, one target source and one noise source, are randomly laid in the room. 4 nodes of 4 microphones each are randomly laid in the room and record the scene. The only constraint is that the sources and the microphones should not be closer than 50 cm from each other and from the walls.

In our experiments, the effects of SRO and STO are considered separately. Their joint impact is left for future work. To simulate asynchronization, in each simulated configuration, one node k among the four is chosen as the reference node. Its SR (resp. sampling start) is left unchanged, this is why we have $\epsilon_k = 0$ (resp. $\tau_k = 0$) for this node. The SRO is simulated by resampling the signals of nodes $j \neq k$ at various sampling frequencies. The STO is simulated by padding zeros at the beginning of the signals of nodes $j \neq k$. Because of the symmetry of the SRO and STO effects, only positive SROs and STOs will be considered. For each node $j \neq k$, the SRO ϵ_j is randomly taken between 0 parts per million (ppm) and a maximum value $\mathrm{SRO}_{\max}$. 6 different values of $\mathrm{SRO}_{\max}$ are considered, leading to 6 evaluation conditions. The STO is randomly taken between 0 ms and a maximum value $\mathrm{STO}_{\max}$. 5 different values of $\mathrm{SRO}_{\max}$ are considered, leading to 5 evaluation conditions. Because the signals of one node share the same hardware and software implementation, we assume that they are synchronized. As a consequence, asynchronization can only affect the second filtering step of Tango.

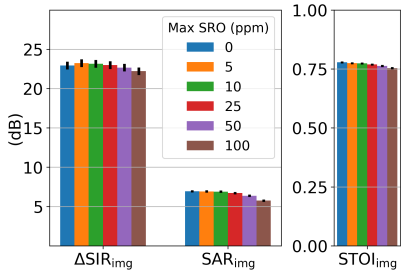
4. Impact of asynchronization on DNN-based speech enhancement

In this section, the system described in section 2.2 is evaluated on the data described in section 3.2. However, the MNDNNs of the second filtering step are trained with synchronized data: at train time, the compressed signals received by a given node are perfectly synchronized with the mixtures recorded by the receiving node.

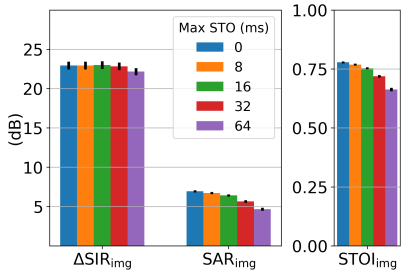
The speech enhancement performance of our system under such conditions is reported in Figure 2 in terms of SIR, SAR [25] and STOI [26], where the bars represent the 95 % confidence interval. It seems from Figure 2a that the SRO has a limited impact on the performance of our system, even for high values of SROs. This is probably explained by the fact that the signals are rather short, so that the effect of the SRO on the signals alignment is limited. Thus, given a rough estimation of the SRO, resampling the signals with this estimation at large intervals is enough to cope with SROs.

The impact of STO on our system is stronger. The SIR seems robust to STO, probably because of the rank-1 decompo-

[1] https://github.com/nfurnon/disco/tree/master/dataset_generation



(a) *SRO impact*



(b) *STO impact*

Figure 2: *Impact of SRO and STO on the speech enhancement performance of our system.*

sition of the MWF. However, the other metrics, especially the STOI, are sensitive to this kind of asynchronization, in particular when the STO exceeds 16 ms, corresponding to the duration of one frame.

As a conclusion of this section, asynchronization does have a negative impact on our system, especially because of STO. In the sequel, we will therefore focus on the impact of STO on our distributed speech enhancement system and consider that no SRO affects the recordings. A solution to compensate for STO without resampling the signals is proposed in the next section.

5. Solution to asynchronization of the input signals of DNNs

We propose to use an attention mechanism to compensate for the negative impact of STO on our speech enhancement system. Since the consequence of asynchronization is that the signals recorded on different devices are not aligned in time, we propose to use an alignment mechanism to implicitly shift the asynchronized signals [27, 28]. To this effect, a temporal alignment attention mechanism is used, which is inspired by the one introduced by Schulze-Forster et al. in a different application field [29]. It is described in the next section.

5.1. Temporal alignment attention mechanism

Let $\mathbf{C}_k$ be a reference channel and $\mathbf{C}_j$ an input channel, both of size $T \times F$; let $\mathbf{c}_k(m)$ and $\mathbf{c}_j(n)$ be their m-th and n-th column respectively. A score between these two columns is computed as:

$$\tilde{s}_{k,j}(m,n) = \mathbf{c}_k(m)\mathbf{W}\mathbf{c}_j(n)^T \,,$$

where $\cdot^T$ denotes the transpose operator and $\mathbf{W}$ is a learnable matrix. In the sequel, we will always consider the first channel of the MNDNNs as the reference channel, so we will drop the index k. We have: $\tilde{s}_j(m,n) = \tilde{s}_{k,j}(m,n)$. All the elements $\tilde{s}_j(m,n)$ are gathered in the matrix $\tilde{\mathbf{S}}_j$ of size $T \times T$. A softmax operation is applied on the rows of $\tilde{\mathbf{S}}_j$ to obtain the so-called

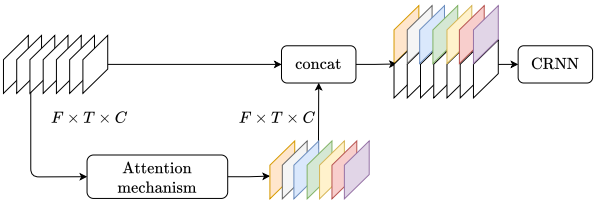


Figure 3: *Illustration of the attention-based CRNN.*

similarity matrix $\mathbf{S}_j$:

$$\mathbf{S}_j = \text{softmax}\big(\tilde{\mathbf{S}}_j\big) \,. \tag{1}$$

The idea of this mechanism is that $\mathbf{S}_j$ should contain the probability of the frames of $\mathbf{C}_j$ being aligned with the frames of the reference channel $\mathbf{C}_1$. These probabilities are multiplied with $\{\mathbf{c}_1(i)\}_{i=1..T}$, the columns of $\mathbf{C}_1$ following:

$$\mathbf{p}_j(m) = \sum_{i=1}^{T} s_j(m,i)\mathbf{c}_1(i) \,.$$

The output matrix $\mathbf{P}_j$ of columns $\{\mathbf{p}_j(m)\}_{m=1..T}$ is concatenated with the input matrix $\mathbf{C}_j$ over the frequency axis.

5.2. Integration of the temporal alignment mechanism in Tango

The previously described attention mechanism is used at the input of the CRNN. Since only the second filtering step is affected by asynchronization, they are integrated into the MNDNN only. The new architecture is represented in Figure 3. Since the input data has twice more features on the frequency axis compared to the initial MNDNN, the last maximum-pooling layer has a kernel size 8×1 to keep the size of the GRU layer the same. On each node, the first channel, corresponding to the local mixture, is taken as the reference channel.

5.3. Quantitative evaluation and analysis

Three systems are compared to evaluate our solution. The first system is the same as in section 4, where the MNDNNs are trained on synchronized data only. The second system has MNDNNs trained on asynchronized data. The third system has MNDNNs with the attention mechanism, trained on asynchronized data. In the asynchronous training set, the SRO is set to 0 ppm and the STO is randomly taken between 0 ms and 32 ms. During evaluation, the STO is randomly taken between 0 ms and a maximum value STO$_{\text{max}}$. The same values of SRO$_{\text{max}}$ as in Section 4 are considered. The results obtained with these three systems are represented in Figure 4. The first conclusion from this experiment is that training the MNDNN in matching conditions brings robustness to the system in terms of SAR and STOI. However, it does not have any significant impact on the SIR. With the attention mechanism, even if the differences are not significant, there is a noticeable improvement in terms of SIR over the system where the MNDNNs are trained in matched conditions but without attention mechanism. This experiment confirms that this attention mechanism is adapted to the misalignment problem, and that it makes our speech enhancement system robust to asynchronization. Another advantage of using such a mechanism is introduced in the next section.

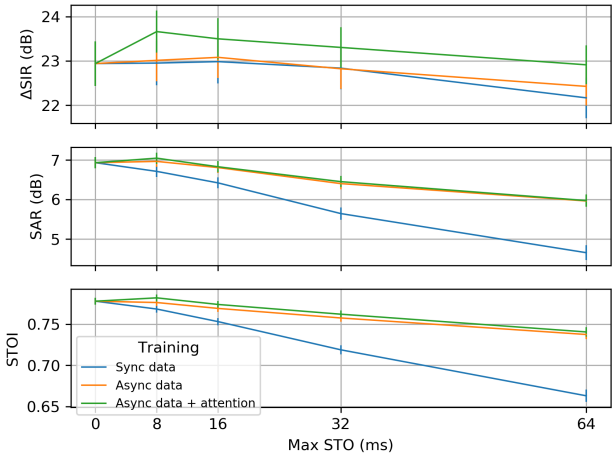


Figure 4: *Speech enhancement performance of three different systems.*

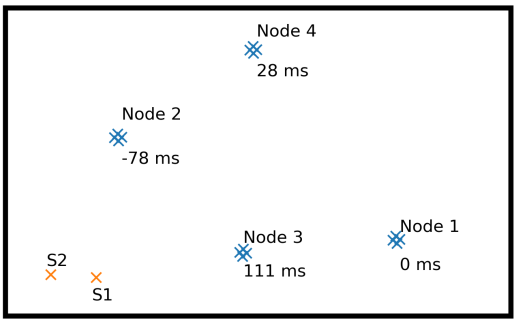


Figure 5: *2D view of the evaluation configuration. The STO of all nodes, relatively to the first node, it also mentioned.*

5.4. Qualitative evaluation and analysis

In this section, in order to highlight another advantage of using the introduced temporal alignment mechanism, we simulate a specific evaluation room to enhance the behaviour of the attention mechanism. The room configuration is represented in Figure 5 where the STOs of all nodes, relatively to the first node, are also mentioned. We represent in Figure 6 the values of the similarity matrices $\{\mathbf{S}_j\}_{j=1..4}$ in Equation 1 computed on the first node of this room configuration. These matrices are the weights applied by the first node on the channels at the input of the MNDNN. It can be seen that these weights seem correlated with the value of the STO. For the weights applied on the second channel for example, an upper diagonal can be clearly seen, linking the i-th output frames with the $(i+5)$-th input frames. Interestingly enough, the time duration of 5 frames, equal to 80 ms,[2] corresponds approximatively to the STO value of the second channel. Similarly for the weights applied on the third node, the clear diagonal dynamic indicates a correlation between the i-th output frames with the $(i-7)$-th input frames, corresponding to a negative delay of approximatively 112 ms,

[2]To recall, one frame lasts 16 ms.

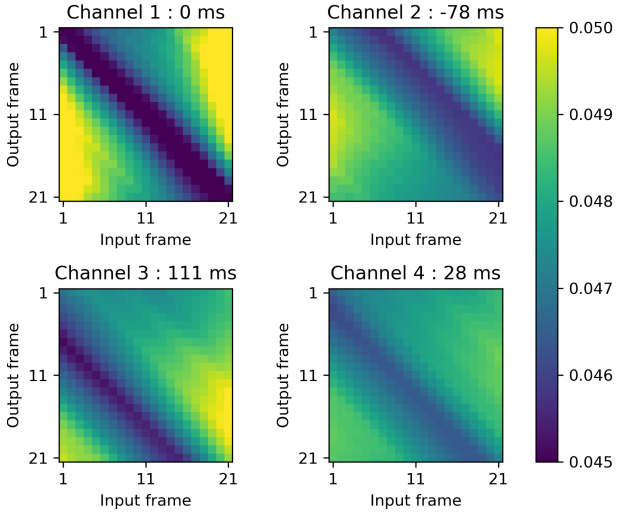


Figure 6: *Weights of the similarity matrices $\{\mathbf{S}_j\}_{j=1..4}$ (see Eq. 1) applied on the four input channels the MNDNN of the first node of the configuration represented in Figure 5.*

which is almost the STO value of the third node relatively to the first node. The same qualitative analysis can be conducted on the similarity matrix applied on the last channel.

As a conclusion of this analysis, the attention mechanism leads to a coarse estimation of the STO between asynchronized nodes in an unsupervised manner. This information could be useful to some applications which rely on a rough alignment of the signals [15].

6. Conclusions

We addressed the issue of asynchronization in a distributed speech enhancement system based on DNNs. We showed that SRO had a limited impact on our experiments, but that the influence of STO was detrimental to the speech enhancement performance of our system. To cope with it, we introduced a temporal alignment attention mechanism that makes the DNNs of our system robust to STO. In addition, we show that the hidden values of the attention mechanism can be interpreted and that they lead to a coarse estimation of the STO at all nodes. We believe that our work introduces a novel and interesting use of attention mechanisms for speech enhancement in ad-hoc microphone arrays. It would be interesting to apply this kind of attention mechanisms on signals in the time domain rather than in the time-frequency domain, where they would probably lead to more precise results and higher performance.

7. Acknowledgements

This work was made with the support of the French National Research Agency, in the framework of the project DiSCogs "Distant speech communication with heterogeneous unconstrained microphone arrays" (ANR-17-CE23-0026-01). Experiments presented in this paper were partially carried out using the Grid5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000).

8. References

[1] A. Bertrand and M. Moonen, "Distributed adaptive node-specific signal estimation in fully connected sensor networks — Part I: Sequential node updating," pp. 5277–5291, Oct 2010.

[2] M. O'Connor and W. B. Kleijn, "Diffusion-based distributed MVDR beamformer," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 810–814.

[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] T. Sherson, W. B. Kleijn, and R. Heusdens, "A distributed algorithm for robust LCMV beamforming," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 101–105.

[5] N. Furnon, R. Serizel, S. Essid, and I. Illina, "Dnn-based mask estimation for distributed speech enhancement in spatially unconstrained microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2310–2323, 2021.

[6] J. Schmalenstroeer, P. Jebramcik, and R. Haeb-Umbach, "A combined hardware–software approach for acoustic sensor network synchronization," *Signal Processing*, vol. 107, pp. 171–184, 2015.

[7] E. Ceolini, I. Kiselev, and S.-C. Liu, "Evaluating multichannel multi-device speech separation algorithms in the wild: a hardware-software solution," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1428–1439, 2020.

[8] R. Lienhart, I. Kozintsev, S. Wehr, and M. Yeung, "On the importance of exact synchronization for distributed audio signal processing," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 4. IEEE, 2003, pp. IV–840.

[9] 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)*. IEEE, 2015, pp. 245–249.

[10] Y. Zeng, R. C. Hendriks, and N. D. Gaubitch, "On clock synchronization for multi-microphone speech processing in wireless acoustic sensor networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 231–235.

[11] J. Schmalenstroeer and R. Haeb-Umbach, "Insights into the interplay of sampling rate offsets and mvdr beamforming," in *Speech Communication; 13th ITG-Symposium*. VDE, 2018, pp. 1–5.

[12] S. Wehr, I. Kozintsev, R. Lienhart, and W. Kellermann, "Synchronization of acoustic sensors for distributed ad-hoc audio networks and its use for blind source separation," in *IEEE Sixth International Symposium on Multimedia Software Engineering*. IEEE, 2004, pp. 18–25.

[13] L. Schenato and F. Fiorentin, "Average TimeSynch: A consensus-based protocol for clock synchronization in wireless sensor networks," *Automatica*, vol. 47, no. 9, pp. 1878–1886, 2011.

[14] S. Markovich-Golan, S. Gannot, and I. Cohen, "Blind sampling rate offset estimation and compensation in wireless acoustic sensor networks with application to beamforming," in *IWAENC 2012; International Workshop on Acoustic Signal Enhancement*. VDE, 2012, pp. 1–4.

[15] J. Schmalenstroeer, J. Heymann, L. Drude, C. Boeddecker, and R. Haeb-Umbach, "Multi-stage coherence drift based sampling rate synchronization for acoustic beamforming," in *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2017, pp. 1–6.

[16] L. Wang and S. Doclo, "Correlation maximization-based sampling rate offset estimation for distributed microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 571–582, 2016.

[17] S. Miyabe, N. Ono, and S. Makino, "Blind compensation of interchannel sampling frequency mismatch with maximum likelihood estimation in stft domain," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 674–678.

[18] D. Cherkassky and S. Gannot, "Blind synchronization in wireless sensor networks with application to speech enhancement," in *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2014, pp. 183–187.

[19] A. Chinaev, P. Thüne, and G. Enzner, "Double-cross-correlation processing for blind sampling-rate and time-offset estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1881–1896, 2021.

[20] H. Chiba, N. Ono, S. Miyabe, Y. Takahashi, T. Yamada, and S. Makino, "Amplitude-based speech enhancement with nonnegative matrix factorization for asynchronous distributed recording," in *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2014, pp. 203–207.

[21] R. M. Corey and A. C. Singer, "Speech separation using partially asynchronous microphone arrays without resampling," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2018, pp. 1–9.

[22] R. Serizel, M. Moonen, B. Van Dijk, and J. Wouters, "Low-rank Approximation Based Multichannel Wiener Filter Algorithms for Noise Reduction with Application in Cochlear Implants," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 785–799, 2014.

[23] N. Furnon, R. Serizel, I. Illina, and S. Essid, "DNN-based distributed multichannel mask estimation for speech enhancement in microphone arrays," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 4672–4676.

[24] Y. Luo, C. Han, N. Mesgarani, E. Ceolini, and S.-C. Liu, "FaSNet: Low-latency adaptive beamforming for multi-microphone audio processing," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 260–267.

[25] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[26] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2010, pp. 4214–4217.

[27] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *International Conference on Learning Representations (ICLR)*, 2014.

[28] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *Conference on Empirical Methods in Natural Language Processing*, 2015.

[29] K. Schulze-Forster, C. S. Doire, G. Richard, and R. Badeau, "Joint phoneme alignment and text-informed speech separation on highly corrupted speech," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7274–7278.