INFORMED VS. BLIND BEAMFORMING IN AD-HOC ACOUSTIC SENSOR NETWORKS FOR MEETING TRANSCRIPTION

Tobias Gburrek[*], *Joerg Schmalenstroeer*[*], *Jens Heitkaemper, Reinhold Haeb-Umbach*

Paderborn University, Department of Communications Engineering, Paderborn, Germany
{*gburrek, schmalen, heitkaemper, haeb*}*@nt.uni-paderborn.de*

ABSTRACT

We consider a network of initially unsynchronized microphone arrays to be used to capture a meeting which is afterwards transcribed. Beamforming is applied to exploit the spatial diversity of the setup for signal enhancement. We propose and compare two approaches to compute the beamformer coefficients. The first, informed beamforming, localizes speakers and sensors and computes speaker activity information, from which beamformer filter coefficients are derived, while the second, blind beamforming, estimates the beamforming coefficients in an unsupervised manner employing a spatial mixture model. We discuss the pros and cons of the two approaches and experimentally assess their sensitivity to synchronization errors, localization errors, erroneous activity information, etc. Simulations show that the informed beamforming achieves a promising performance as measured by the word error rate of a downstream speech recognizer.

***Index Terms*—** beamforming, meeting transcription, ad-hoc acoustic sensor network

1. INTRODUCTION

Transcribing the conversation of a meeting is a challenging task for at least two reasons. There is first the interaction dynamics among the speakers, which articulate themselves in an intermittent manner with alternating segments of speech inactivity, single-, and multi-talker speech. Second, the speech signal is usually captured by microphones from a distance resulting in noisy and reverberated recordings.

What comes to the rescue is the use of multiple microphones that can be combined to form spatial filters for the extraction of the signals of the individual speakers. In the considered scenario, an ad-hoc acoustic sensor network consisting of two or more microphone arrays is used for signal capture. Neither the positions of the speakers nor the positions and orientations of the microphone arrays are, however, known in advance. Furthermore, the microphone arrays are initially asynchronous having an unknown sampling rate offset (SRO) and sampling time offset (STO) w.r.t. each other. While the first is time-varying [1] and originates from the oscillators driving the sampling processes [2], the latter is due to the fact that each device starts recording at a different point in time [3].

To be able to combine the multi-channel recordings to spatial filters for signal extraction, the data streams have to be first synchronized and, second, beamformer filter coefficients have to be estimated, which account for the intermittent nature of speech activity. The main purpose of this contribution is to compare two approaches to speech activity estimation for beamformer coefficient computation in a meeting setup with initially unsychronized microphone arrays. The first is an informed approach to beamforming, where estimated source position and activity information controls (informs) the estimation of the spatial covariance matrices (SCMs) of the desired and interfering signals. It is based on [4], but unlike there, where the activity information is used to initialize a spatial mixture model, we here employ it to directly compute SCMs for each of the speakers. The second is a blind approach, where a spatial mixture model is learned in an unsupervised manner to obtain activity information, from which the beamformer coefficients are derived.

Note, that there exists a third approach, neural-network-based speech activity estimation [5, 6]. It relies on supervised training with the need for "parallel" data, where an utterance to be enhanced is presented at the input of the network, while the training target is derived from the clean, undistorted version of the very same utterance. As parallel data may not always be available, we exclude this approach from our study here.

Once activity masks have been obtained, actual spatial filtering can be done with different beamformer designs. Here, we employ the minimum variance distortionless response (MVDR) beamformer, which is a particularly popular front-end to automatic speech recognition (ASR) [7]. The above two approaches are compared w.r.t. the word error rate performance of a downstream ASR engine. We are also interested in assessing their sensitivity to remaining clock synchronization errors.

Several approaches to sampling rate offset estimation have been proposed in the literature, e.g., [8, 9, 10]. Here, we employ the method of [3], because it can track a time-varying SRO. It further allows to discern the STO and the time of flight (ToF) contributions to the time-difference of arrival (TDoA), which comes in handy when computing a beamformer steering vector based on estimates of the speakers' and sensors' positions.

Informed beamforming using narrow-band DoA or position estimates to control the computation of the desired and undesired signal SCMs has been developed in [11, 12], and studied in the context of meetings in [13]. In this contribution we assume that more than one microphone array is available and we employ steered-response power phase transform (SRP-PhaT) to infer speech activity at estimated speaker positions. Spatial mixture models for computing time-frequency activity masks have been extensively investigated both for beamforming [5] and source separation [14]. In the context of meeting recognition it has been investigated in [15, 16]. While studied individually, those works do not allow for a fair comparison of informed vs. blind activity estimation in the meeting setup. This contribution is meant to fill this gap.

The paper is organized as follows: After giving a coarse overview of the overall system and the synchronisation component in Sec. 2 and Sec. 3, we describe informed and blind source extraction in Sec. 4. Section 5 compares them experimentally, and we draw some conclusions in Sec. 6.

[*] These authors contributed equally.

978-1-6654-6867-1/22/$31.00 ©2022 IEEE

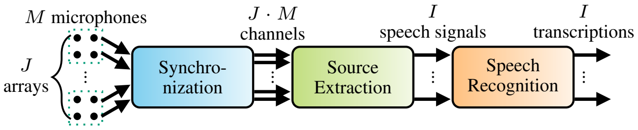

Fig. 1: Meeting transcription for ad-hoc acoustic sensor networks

2. MEETING TRANSCRIPTION SYSTEM

Figure 1 gives an overview of the considered meeting transcription system. The signals of I speakers at positions $\boldsymbol{s}_i$, $i\in\{1,\dots,I\}$, are recorded by an ad-hoc acoustic sensor network (ASN), which consists of $J \geq 2$ compact microphone arrays with M microphones each. While the geometric arrangement within an array is assumed to be known, the location and orientation of the arrays relative to the speakers is not known in advance.

The signals stemming from different microphone arrays are first synchronized. Next, the signals of the individual speakers are extracted by spatial filtering and transcribed by the speech recognition engine.

3. SYNCHRONIZATION

Synchronization is concerned with estimating the SROs and STOs of the microphone signals relative to a reference channel, e.g., the first microphone of the first array. Then, the signals are resampled to remove the offsets. To estimate the SROs we employ the dynamic weighted average coherence drift (DWACD) method [3] which is able to track a time-varying SRO. Furthermore, the approach to STO estimation from [3] is utilized which allows to discern between the STO, stemming from the different recording start times of the arrays, and the delays caused by the propagation times from the speakers to the microphones. This distinction is important for the subsequent position-based diarization used in the informed approach to beamforming, which relies on TDoA information that reflects the true physical positions of the speakers and the microphones. SRO and STO compensation is achieved with the short-time Fourier transform (STFT) based resampling method from [17].

4. SOURCE EXTRACTION

We employ an MVDR beamformer to extract the single speakers' signals from the meeting recordings. In the following the concept of the MVDR beamformer is shortly recapitulated. Afterwards an informed and a blind manner to estimate the SCMs needed to calculate the beamformer coefficients are introduced.

4.1. MVDR beamforming

The individual speakers' signals are extracted by employing a beamformer for each speaker. Let $\boldsymbol{Y}(\ell,k)=[Y_{1,1}(\ell,k),\dots,Y_{J,M}(\ell,k)]^T$ denote the stacked STFTs of the synchronized multi-channel input data of the J arrays with M microphones each, at time frame ℓ and frequency bin k. The STFT of the extracted signal of the i-th speaker is computed as

$$\widehat{X}_i(\ell,k) = \boldsymbol{W}_i^H(\ell,k)\cdot\boldsymbol{Y}(\ell,k), \tag{1}$$

with $\boldsymbol{W}_i(\ell,k)$ denoting the beamformer coefficients to extract the signal of the i-th speaker.

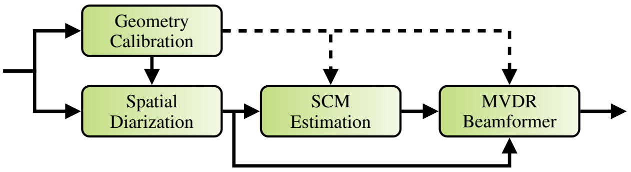

Fig. 2: Informed source extraction via MVDR beamforming. Dashed lines present the flow of position information used to estimate the target speaker's SCMs based on the steering vector.

In this work the MVDR beamformer in the formulation of [18] is employed:

$$\boldsymbol{W}_i(\ell,k) = \frac{\left(\overline{\boldsymbol{\Phi}}_i(\ell,k)\right)^{-1}\cdot\boldsymbol{\Phi}_i(\ell,k)}{\operatorname{tr}\left\{\left(\overline{\boldsymbol{\Phi}}_i(\ell,k)\right)^{-1}\cdot\boldsymbol{\Phi}_i(\ell,k)\right\}}\cdot\boldsymbol{u}. \tag{2}$$

Here, $\boldsymbol{\Phi}_i(\ell,k)$ and $\overline{\boldsymbol{\Phi}}_i(\ell,k)$ are the SCMs of the desired signal to be extracted and the SCMs of the interference, respectively. Further, $\operatorname{tr}\{\cdot\}$ is the trace operator, and $\boldsymbol{u}$ is a unit vector pointing to a reference microphone.

To compute the beamformer coefficients, we describe two alternative approaches in the following. The first, called informed source extraction, derives information about the spatial arrangement of the sources and sensors and the temporal activity of each speaker, and computes the beamformer coefficients from it. The second is a blind approach, that is agnostic to the spatial arrangement and is based on fitting a spatial mixture model to the observations. It computes the beamformer coefficients from the estimated posterior probabilities of source activity.

4.2. Informed SCM estimation

Figure 2 gives an overview of the proposed informed source extraction method. It is based on the spatial diarization component presented in [4]. However, the usage of the diarization information to estimate the SCMs differs from [4].

First, geometry calibration is conducted. It determines both the positions $\widehat{\boldsymbol{s}}_i$, $i\in\{1,\dots,I\}$, of the speakers and the positions and orientations of the microphone arrays, via the iterative data set matching method from [19]. Based on the estimated microphone and speaker positions a spatial diarization, i.e., an estimation of the information when and at which position a speaker is active, is performed via SRP-PhaT based multi-speaker tracking. For more details on the spatial diarization procedure we refer to [4].

Given the information about "who speaks when and where", SCMs of each speaker can be computed. Let $\mathcal{I}_i$ be the set of time frame indices, where the speaker at position $\boldsymbol{s}_i$ is active. Then

$$\tilde{\mathcal{I}}_i = \mathcal{I}_i \backslash \left\{\bigcup_{i'\neq i}(\mathcal{I}_i\cap\mathcal{I}_{i'})\right\} \tag{3}$$

is the set of indices, where the speaker is solely active. We compute the speaker's SCM $\boldsymbol{R}_i(k)$ for $i\in\{1,\dots,I\}$ as

$$\boldsymbol{R}_i(k) = \frac{1}{|\tilde{\mathcal{I}}_i|}\sum_{\ell\in\tilde{\mathcal{I}}_i}\boldsymbol{Y}(\ell,k)\cdot\boldsymbol{Y}^H(\ell,k), \tag{4}$$

with $|\tilde{\mathcal{I}}_i|$ denoting the cardinality of the set $\tilde{\mathcal{I}}_i$. Further, the time-varying SCM of the interference $\overline{\boldsymbol{\Phi}}_i(\ell,k)$ in (2) are calculated as sum over the SCMs of the interfering speakers in frame ℓ:

$$\overline{\boldsymbol{\Phi}}_i(\ell,k) = \sum_{\nu\in A(\ell)\backslash\{i\}}\mathbf{R}_\nu(k), \tag{5}$$

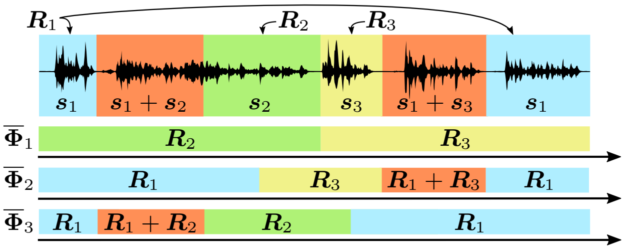

Fig. 3: Estimation of the speakers' SCMs from single speaker segments and their temporal application as interference SCMs per target speaker.

with $A(\ell)$ being the set of speaker indices for which an activity is detected in frame ℓ.

Fig. 3 visualizes the SCM estimation by means of an example with three target speaker positions ($\boldsymbol{s}_1$ blue, $\boldsymbol{s}_2$ green, $\boldsymbol{s}_3$ yellow), including the selected SCMs of the interference per frame and speaker position. Audio segments with more than one active speaker are marked in orange. For example, the beamformer used to extract the signal of the speaker at position $\boldsymbol{s}_1$, starts with the interference SCM $\overline{\boldsymbol{\Phi}}_1(\ell,k)$=$\boldsymbol{R}_2(k)$ and then continues with $\overline{\boldsymbol{\Phi}}_1(\ell,k)$=$\boldsymbol{R}_3(k)$. Note that this computation assumes that at least one interfering speaker is always active. Therefore, the previously and subsequently used interference SCMs are re-utilized in periods in time, in which no interfering speaker is active (see $\overline{\boldsymbol{\Phi}}_2$ and $\overline{\boldsymbol{\Phi}}_3$ in Fig. 3).

Concerning the computation of the SCM of the target speaker $\boldsymbol{\Phi}_i(\ell,k)$ to be used in (2) there are two options. While the obvious choice is $\boldsymbol{\Phi}_i(\ell,k)$=$\boldsymbol{R}_i(k)$, an alternative is to compute it via the outer product of the steering vector $\boldsymbol{d}(\widehat{\boldsymbol{s}}_i,k)$ pointing to the target speaker position $\boldsymbol{s}_i$, i.e., $\boldsymbol{\Phi}_i(\ell,k)$=$\boldsymbol{d}(\widehat{\boldsymbol{s}}_i,k)\boldsymbol{d}^H(\widehat{\boldsymbol{s}}_i,k)$. Assuming anechoic signal propagation, the steering vector can be computed from the time differences of flight (TDoFs) calculated from the estimated positions of the speakers and microphones.

In order to mitigate the effect of wrong decisions of the spatial diarization an energy-based voice activity detection (VAD) is used to decide when the target speaker is active. The energy threshold is calculated based on the target speaker's energy and the energy of the interfering speaker that is suppressed worst and thus shows the highest energy. The energies of the single speakers are estimated based on the periods in time for which the speaker is solely active according to the spatial diarization. The resulting activity information is used in the ASR system to discard periods in time without activity of the target speaker.

4.3. Blind SCM estimation

In the blind approach the speakers' SCMs are estimated using a spatial mixture model. To be specific, the complex Angular Central Gaussian Mixture Model (cACGMM) [20] is employed with time-varying instead of frequency-dependent mixture weights [21]:

$$p(\boldsymbol{Z}(\ell,k)) = \sum_{i=1}^{I} \pi_i(\ell)\cdot\mathcal{A}\left(\boldsymbol{Z}(\ell,k);\boldsymbol{B}_i(k)\right), \qquad (6)$$

with $\boldsymbol{Z}(\ell,k) = \boldsymbol{Y}(\ell,k)/\|\boldsymbol{Y}(\ell,k)\|$. Here, $\mathcal{A}(\cdot)$ denotes the complex angular central Gaussian distribution [20]. The mixture model's parameters are estimated with the Expectation Maximization (EM) algorithm. It is initialized by drawing the posterior probabilities of speakers, who are active at time frame ℓ and frequency bin k, $\gamma_i(\ell,k)$, from a Dirichlet distribution. First informal experiments have shown that this initialization tends to be more robust against remaining synchronization errors than the clustering-based initialization from [22], which leads to better results for perfectly synchronous signals. The parameter matrix $\boldsymbol{B}_i(k)$ is initialized with the identity matrix.

Once training is completed, the estimated time-varying mixture weights $\pi_i(\ell)$ are smoothed over time and thresholded to zero or one (with a threshold of 0.2), indicating an inactive or active speaker, respectively.

The time frame index set $\mathcal{I}_i$, where the i-th speaker is active is now divided into continuous time intervals $\mathcal{I}_{i,\zeta}$, where the i-th speaker is permanently active, with $\mathcal{I}_i = \bigcup_{\zeta} \mathcal{I}_{i,\zeta}$. For each interval $\mathcal{I}_{i,\zeta}$, the target speaker SCM is computed as

$$\boldsymbol{R}^{\mathrm{B}}_{i,\zeta}(k) = \frac{1}{|\mathcal{I}_{i,\zeta}|}\sum_{\ell\in\mathcal{I}_{i,\zeta}} \gamma_i(\ell,k)\cdot\boldsymbol{Y}(\ell,k)\cdot\boldsymbol{Y}^H(\ell,k), \qquad (7)$$

while the interferer's SCM is computed over the same interval, however summing over all other active speakers, weighted by their posterior probabilities $\gamma_\nu(\ell,k), \nu\neq i$.

This way of SCM computation, that is adopted from [22], is further modified to align it with the way it is done in the informed approach of Sec. 4.2. Therefore, the target SCM is computed on all frames where the target speaker is active:

$$\boldsymbol{R}^{\mathrm{B}}_{i}(k) = \frac{1}{|\mathcal{I}_{i}|}\sum_{\ell\in\mathcal{I}_{i}} \gamma_i(\ell,k)\cdot\boldsymbol{Y}(\ell,k)\cdot\boldsymbol{Y}^H(\ell,k), \qquad (8)$$

which is similar to (4), except that the set $\mathcal{I}_i$ instead of $\tilde{\mathcal{I}}_i$ is used and that the outer products are weighted by the posterior probability of speaker activity. The interferer SCMs are computed as in the informed approach (see (5)) on a per-frame basis, however again weighted by the posterior probabilities of speaker activity.

The EM algorithm assumes knowledge of the number of speakers. However, since this information is in general not available, we proceed as follows: we choose a value $I^{\max}$ in a way that it can be safely assumed that the true number of speakers is smaller, and start the EM iterations with the assumption of $I^{\max}$ speakers. During training the number of mixture components is reduced by merging classes with similar estimated mixture weights, as is measured by an Intersection-over-Union ratio above 0.8, following the class fusion suggested in [22].

5. EXPERIMENTS

For the experiments we use our database from [4] that consists of 100 simulated meetings. It simulates 5 min long meetings in a randomly generated room with speakers sitting around a table. Hereby, a single speaker is active in 66% of the total meeting duration, while two speakers are concurrently active in 21% of the total meeting duration, and in the remaining time no speaker is active. Audio signals are captured by J=3 independent microphone arrays whose M=4 microphones are arranged in a quadratic layout with edge length of 5 cm. For more details on the database we refer to [4].

We use the concatenated minimum-permutation word error rate (cpWER) [23] as the performance measure to evaluate the systems, whereby the ASR results are obtained using the acoustic model configuration from [24]. The model is trained on 16 kHz SMS-WSJ data [24] to match the sampling frequency used in the systems for synchronization, geometry calibration and beamforming.

5.1. Comparison of blind and informed beamforming

In Table 1 a comparison of the transcription performances is shown, which are achieved with the informed and the blind source extraction

Table 1: Comparison of the transcription performance of the informed and blind source extraction system

System	WER / %
Clean audio	6.42
W/o enhancement	31.51
Blind beamfoming	9.41
Informed beamforming	7.08

systems, respectively. Both approaches significantly improve the WER performance compared to the ASR result obtained on a single microphone input without enhancement. It can also be seen that the informed system outperforms the blind system. In the following the informed and blind source extraction systems are separately investigated in order to gain deeper insights into their individual advantages and disadvantages.

5.2. Ablation study for informed beamforming

Table 2 shows the influence of errors, which are made by the different subsystems of the informed source extraction system, on the transcription performance. It becomes obvious that a beamformer using the target speaker's SCMs ($\boldsymbol{\Phi}_i(\ell, k){=}\boldsymbol{R}_i(k)$), which is directly estimated from the microphone signals, is able to outperform a beamformer, which estimates the target speaker's SCMs based on a steering vector calculated from information about the microphones' and speakers' positions ($\boldsymbol{\Phi}_i(\ell, k){=}\boldsymbol{d}(\widehat{\boldsymbol{s}}_i, k)\boldsymbol{d}^H(\widehat{\boldsymbol{s}}_i, k)$). We hypothesize that this results from the fact that the system, which directly estimates the target speaker's SCMs from the microphone signals, is less sensitive to the quality of the synchronization and imperfections of the diarization information than the system based on the steering vector.

A major advantage of estimating the target speaker's SCMs directly from the microphone signals compared to the version based on the steering vector is that the beamformer does not explicitly require knowledge of the microphone and speaker positions. It uses, though, the activity information given by the spatial diarization component. In contrast to that the steering vector based SCM estimation relies on estimates of the signals' TDoFs and thus is heavily influenced by errors of the position estimates. The largest disadvantage of the steering vector based estimation of the SCMs of the target speaker is the accumulation of errors of all subsystems. For example, the accumulated localization and synchronization errors might significantly deteriorate the ability of the beamformer to focus on the correct position. However, the localization-based estimation of the target speaker's SCMs via the steering vector might be advantageous if a spatial broad-band noise source is continuously active. In this case there are no periods in time in which the target speaker is solely active so that (4) cannot be used anymore to estimate the SCMs of the target speaker.

Table 2: Ablation study for the informed source extraction system

Target speaker's	Oracle values for		WER / %	
SCM	Activity	TDoFs	Perf. sync.	Est. sync.
$\boldsymbol{d}(\widehat{\boldsymbol{s}}_i, k)\boldsymbol{d}^H(\widehat{\boldsymbol{s}}_i, k)$	✓	✓	6.66	7.14
$\boldsymbol{d}(\widehat{\boldsymbol{s}}_i, k)\boldsymbol{d}^H(\widehat{\boldsymbol{s}}_i, k)$	✓		7.09	8.36
$\boldsymbol{d}(\widehat{\boldsymbol{s}}_i, k)\boldsymbol{d}^H(\widehat{\boldsymbol{s}}_i, k)$		✓	7.00	7.36
$\boldsymbol{d}(\widehat{\boldsymbol{s}}_i, k)\boldsymbol{d}^H(\widehat{\boldsymbol{s}}_i, k)$			7.63	8.97
$\boldsymbol{R}_i(k)$	✓	-	6.82	7.00
$\boldsymbol{R}_i(k)$		-	6.96	7.08

Table 3: Ablation study for the blind source extraction system

Known number	SCM	WER / %	
of speakers	estimation	Perf. sync.	Est. sync.
✓	(7)	8.63	8.49
✓	(8)	7.52	7.51
	(7)	10.02	10.32
	(8)	8.95	9.41

5.3. Ablation study for blind beamforming

An ablation study for the blind source extraction system is presented in Table 3. It can be seen that the proposed estimation of the interference SCMs based on the complete meeting via (8) leads to a better transcription compared to the segment-wise SCM estimation via (7). A significant performance degradation is observed if the number of speakers is not known in advance but has to be estimated from the data. There is clearly room for improvement here by using better estimators, e.g., the infinite mixture model of [25]. Overall, the influence of the synchronization quality on the blind source extraction system is rather small.

5.4. Discussion

It is to be mentioned that all presented results are given for a batch offline processing of the meeting. However, it is straightforward to adapt the building blocks of the informed source extraction system to block online processing. Also the blind approach can be made (block) online by employing a recursive EM algorithm. Moreover, the informed source extraction system is computationally less demanding than the spatial mixture model although it consists of more small building blocks.

The advantage of the spatial mixture model that it provides "narrowband" activity information, i.e., at a time-frequency bin resolution, appears to play a minor role in the investigated scenario. However, in settings with permanently active spatial noise sources it could be beneficial. Given such a noise source the informed system would integrate it into the interference SCMs and the target SCMs would be based on the steering vector.

6. CONCLUSIONS

We compared an informed and a blind source extraction system via beamforming for meeting transcription using an ad-hoc ASN. The informed source extraction system relies on estimates of the microphones' and speakers' positions and information of the speakers' activities, which is estimated based on the position knowledge. In contrast to that, the blind source extraction system calculates the beamformer coefficients based on a spatial mixture model. On simulated meetings it was shown that both approaches achieve a promising transcription performance. However, the informed approach to source extraction is able to outperform the blind approach slightly.

Both approaches rely on the assumption of fixed speaker positions. Thus, future work has to adapt the source extraction systems to handle temporally moving speakers.

7. ACKNOWLEDGEMENTS

Computational resources were provided by the Paderborn Center for Parallel Computing. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Projects 282835863.

8. REFERENCES

[1] "High-precision audio drift measurements with GPS," https://protyposis.net/clockdrift/high-precision-audio-drift-measurements-with-gps/, Aug. 2021.

[2] Fred L. Walls and Jean-Jacques Gagnepain, "Environmental sensitivities of quartz oscillators," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 39, pp. 241–9, 02 1992.

[3] Tobias Gburrek, Joerg Schmalenstroeer, and Reinhold Haeb-Umbach, "On synchronization of wireless acoustic sensor networks in the presence of time-varying sampling rate offsets and speaker changes," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[4] Tobias Gburrek, Christoph Boeddeker, Thilo von Neumann, Tobias Cord-Landwehr, Joerg Schmalenstroeer, and Reinhold Haeb-Umbach, "A meeting transcription system for an ad-hoc acoustic sensor network," *arXiv preprint arxiv.2205.00944*, 2022.

[5] Takuya Higuchi, Nobutaka Ito, Takuya Yoshioka, and Tomohiro Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5210–5214.

[6] Jahn Heymann, Lukas Drude, and Reinhold Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[7] Reinhold Haeb-Umbach, Jahn Heymann, Lukas Drude, Shinji Watanabe, Marc Delcroix, and Tomohiro Nakatani, "Far-field automatic speech recognition," *Proceedings of the IEEE*, vol. 109, no. 2, pp. 124148, 2021.

[8] Shmulik Markovich-Golan, Sharon Gannot, and Israel Cohen, "Blind sampling rate offset estimation and compensation in wireless acoustic sensor networks with application to beamforming," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2012.

[9] Lin Wang and Simon 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.

[10] Aleksej Chinaev, Philipp Thüne, and Gerald 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.

[11] Maja Taseska and Emanul A. P. Habets, "Informed spatial filtering for sound extraction using distributed microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 7, pp. 1195–1207, 2014.

[12] M. Taseska and E.A. Habets, "DoA-informed source extraction in the presence of competing talkers and background noise," *EURASIP Journal on Advanced Signal Processing*, 2017.

[13] Aswin Shanmugam Subramanian, Chao Weng, Shinji Watanabe, Meng Yu, and Dong Yu, "Deep learning based multi-source localization with source splitting and its effectiveness in multi-talker speech recognition," *Computer Speech & Language*, vol. 75, pp. 101360, 2022.

[14] Lukas Drude and Reinhold Haeb-Umbach, "Integration of neural networks and probabilistic spatial models for acoustic blind source separation," *IEEE Journal of Selected Topics in Signal Processing*, 2019.

[15] Shoko Araki, Nobutaka Ono, Keisuke Kinoshita, and Marc Delcroix, "Meeting recognition with asynchronous distributed microphone array using block-wise refinement of mask-based MVDR beamformer," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[16] Takuya Yoshioka, Dimitrios Dimitriadis, Andreas Stolcke, William Hinthorn, Zhuo Chen, Michael Zeng, and Xuedong Huang, "Meeting transcription using asynchronous distant microphones," in *Proc. Interspeech*, September 2019.

[17] Joerg Schmalenstroeer and Reinhold Haeb-Umbach, "Efficient sampling rate offset compensation - an overlap-save based approach," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2018.

[18] Mehrez Souden, Jacob Benesty, and Sofiène Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Transactions on audio, speech, and language processing*, vol. 18, no. 2, pp. 260–276, 2010.

[19] Tobias Gburrek, Joerg Schmalenstroeer, and Reinhold Haeb-Umbach, "Geometry calibration in wireless acoustic sensor networks utilizing DoA and distance information," *EURASIP Journal on Audio, Speech, and Music Processing*, 2021.

[20] Nobutaka Ito, Shoko Araki, and Tomohiro Nakatani, "Complex angular central gaussian mixture model for directional statistics in mask-based microphone array signal processing," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2016.

[21] Nobutaka Ito, Shoko Araki, and Tomohiro Nakatani, "Permutation-free convolutive blind source separation via full-band clustering based on frequency-independent source presence priors," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[22] Christoph Boeddeker, Tobias Cord-Landwehr, Thilo von Neumann, and Reinhold Haeb-Umbach, "An initialization scheme for meeting separation with spatial mixture models," in *Accepted for Annual Conference of the International Speech Comunication Association (INTERSPEECH) (arXiv preprint arxiv.2204.01338)*, 2022.

[23] Shinji Watanabe, Michael Mandel, Jon Barker, Emmanuel Vincent, Ashish Arora, Xuankai Chang, Sanjeev Khudanpur, Vimal Manohar, Daniel Povey, Desh Raj, David Snyder, Aswin Shanmugam Subramanian, Jan Trmal, Bar Ben Yair, Christoph Boeddeker, Zhaoheng Ni, Yusuke Fujita, Shota Horiguchi, Naoyuki Kanda, Takuya Yoshioka, and Neville Ryant, "Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," in *Proc. International Workshop on Speech Processing in Everyday Environments (CHiME)*, 2020.

[24] Lukas Drude, Jens Heitkaemper, Christoph Boeddeker, and Reinhold Haeb-Umbach, "SMS-WSJ: Database, performance measures, and baseline recipe for multi-channel source separation and recognition," *arXiv preprint arXiv:1910.13934*, 2019.

[25] Oliver Walter, Lukas Drude, and Reinhold Haeb-Umbach, "Source counting in speech mixtures by nonparametric bayesian estimation of an infinite gaussian mixture model," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.