ACOUSTIC EVENT SOURCE LOCALIZATION FOR SURVEILLANCE IN REVERBERANT ENVIRONMENTS SUPPORTED BY AN EVENT ONSET DETECTION

Peter Transfeld[1]*, Uwe Martens*[2]*, Harald Binder*[2]*, Thomas Schypior*[2]*, and Tim Fingscheidt*[1]

[1]Institute for Communications Technology, Technische Universität Braunschweig, Germany
[2]artec technologies AG, Diepholz, Germany
{transfeld, fingscheidt}@ifn.ing.tu-bs.de, {martens, binder, schypior}@artec.de

ABSTRACT

This contribution presents a robust approach to acoustic event source localization for surveillance under reverberant environmental conditions. In particular, we support the classical generalized cross-correlation algorithm with phase transform weighting (GCC-PHAT) and the steered response power (SRP) algorithm by a sound activity detection and an event *onset* detector. The proposed algorithmic framework including spatial minimum tracking and smoothing for the suppression of artifacts in the spatial likelihood function significantly outperforms a respective reference approach, decreasing *both* the miss ratio by up to 9% absolute, and the average angular estimation error by up to 4°.

Index Terms— acoustic event source localization, surveillance, reverberant environment, onset detection

1. INTRODUCTION

Acoustic source localization and acoustic speaker localization have been intensively investigated in the last years. In several aspects, the knowledge of the position of speakers or sound sources can be useful and is consequently employed within a wide range of applications. In teleconferencing and videoconferencing the speaker's position can be exploited by steering a microphone beamformer or automatically pointing a camera at him [1]. These two applications can be as well found in the field of smart rooms, where the room itself is aware of the people inside it [2]. Another promising field is ambient assisted living, where distributed microphones are employed to help elderly people, e. g., by fall detection [3]. Furthermore, these techniques can also be used for security aspects [4] or to automatically direct a robot [5] into the direction of a (moving) sound source.

A considerable number of acoustic sound source localization methods are available. Most common is the estimation of the time difference of arrival (TDOA) between two microphone signals. A state-of-the-art TDOA estimation approach is the generalized cross-correlation (GCC) method, which is based on the cross-correlation between two microphone signals. Within this method, several weighting factors can be employed, of which the phase transform (PHAT) [6] gives good results under reverberant conditions [7]. A similar approach is the crosspower-spectrum phase (CSP) method [8, 9]. To improve the resulting position estimate, optionally, one could make use of the steered response power algorithm (SRP) [10]. Another option is to employ acoustic beamforming and exploit the directional pattern of a microphone array [11].

In reverberant environments, the precision of the position estimate may severely be affected by sound reflections. Highest precision can be achieved in the moment that the direct sound arrives at the microphone, which requires then to detect the onset of an acoustic event. A typical application is musical analysis using phase- and energy-based onset detection [12, 13]. It is also employed for the analysis of sounds by taking psychoacoustic knowledge into account [14]. A further application is the field of auditory scene analysis, using acoustic event onsets for audio segmentation [15]. Furthermore, localization tasks make use of signal amplitude-based onset detection to, e. g., steer a robot [16].

In [17] a GCC-PHAT-SRP-based framework for acoustic speaker localization with distributed microphones is presented. The obtained spatial likelihood function (SLF) is spatially filtered and smoothed. Within these two steps a very simple (not further described) voice activity detection in the time domain is used to trigger the noise floor estimation for a Wiener-type filtering. Potential speaker positions are deleted from the noise floor by a complex threshold operation.

In our present work we rearrange and simplify the approach presented in [17] to obtain a framework suitable for acoustic *event* localization with a *microphone array* in a far-field context. The technique is used for surveillance purposes, where the task is to estimate an event sound source location and then steer a camera to it. The acoustic event source localization presented in this paper should augment respective image-based processing for higher robustness in bad visual conditions. As in this work the sound source has to be located in the far field, a further step is to change the geometry behind the computational framework to a *spherical coordinate* search space. In addition, we simplify the noise floor estimation in the computation of the total spatial likelihood function towards a *spatial minimum statistics* approach. Furthermore the simple time-domain voice activity detection is replaced by a frequency-domain *sound activity detection* (SAD) and an *event onset detection* (EOD).

The organization of the paper is as follows: In Section 2 we present the employed new search space, and briefly revisit the GCC-PHAT and SRP methods. Section 3 details our new event onset detection, the spatial minimum tracking and smoothing process. In Section 4 the evaluation methodology is presented and the results are analyzed. Final conclusions are made in Section 5.

2. BASELINE ALGORITHMIC APPROACHES

In this section we introduce the employed signal model and far-field assumption, the new search space resulting from this, and briefly revisit the generalized cross-correlation (GCC) approach, phase transform weighting (PHAT), and steered response power (SRP).

2.1. Signal Model

Given a room, and a rectangular microphone array with a camera placed in the array center. The position of the acoustic event sound

source to be localized is assumed in the camera's field of view. The array consists of M microphones $\mu \in \mathcal{M} = \{1, 2, \ldots, M\}$, providing output signals $y_\mu(t)$. The microphones are equidistant and located at positions $\mathbf{r}_\mu = (r_{x\mu}, r_{y\mu}, r_{z\mu})^\mathsf{T}$, respectively, with $(\cdot)^\mathsf{T}$ being the (vector) transpose. The array center is in the origin $\mathbf{r} = \mathbf{0}$, while the array itself is in the x-y-plane. From a sound source position $\mathbf{r}_s$ a signal $s(t)$ is emitted and then convolved with the impulse response $h_\mu(t)$ of the room towards microphone μ. Environmental noise $n_\mu(t)$ is superimposed leading to the microphone signal

$$y_\mu(t) = h_\mu(t) * s(t) + n_\mu(t). \tag{1}$$

The time needed for an arbitrary sound wave to travel from a position $\mathbf{r}$ to the microphone $\mathbf{r}_\mu$ at a velocity of $c = 343$ m/s is

$$\tau_\mu = \tau_\mu(\mathbf{r}) = \frac{\|\mathbf{r}_\mu - \mathbf{r}\|}{c}, \tag{2}$$

with $\|\cdot\|$ being the Euclidean norm. Neglecting reverberation for the moment and setting $\mathbf{r} = \mathbf{r}_s$ in (2), the microphone signal can be written as

$$y_\mu(t) = a_\mu \cdot s(t - \tau_\mu) + n_\mu(t), \tag{3}$$

whereby a_μ denotes an attenuation factor which is related to air absorption. The time difference of arrival (TDOA) between two microphones $\mu, \nu \in \mathcal{M}$, and an arbitrary position $\mathbf{r}$ can be written as

$$\tau_{\mu\nu}(\mathbf{r}) = \tau_\mu(\mathbf{r}) - \tau_\nu(\mathbf{r}) = \frac{1}{c}(\|\mathbf{r}_\mu - \mathbf{r}\| - \|\mathbf{r}_\nu - \mathbf{r}\|). \tag{4}$$

2.2. Far Field Assumption and Direction of Arrival

Under the far field assumption made in this work, the array is not able to resolve the distance $\|\mathbf{r}_s\|$ to the sound source [10], which generally leads to inaccurate localization results. The solution to this problem is to transfer the sound source position from a rectangular coordinate representation (depending on the distance $\|\mathbf{r}_s\|$), to a representation depending on the direction of arrival (DOA) of a sound wave to the origin. For a position $\mathbf{r} \in \mathbb{R}^3$ this is accomplished by a spherical coordinate transformation. Neglecting the length $\|\mathbf{r}\|$, the so-called propagation vector is defined as

$$\boldsymbol{\zeta}(\theta, \phi) = \frac{\mathbf{r}}{\|\mathbf{r}\|} \begin{pmatrix} \cos\phi\cos\theta \\ \cos\phi\sin\theta \\ \sin\phi \end{pmatrix} \in \mathcal{C} \text{ with } \|\boldsymbol{\zeta}(\theta, \phi)\| = 1, \tag{5}$$

with $\mathcal{C}$ being the set of all possible vectors $\boldsymbol{\zeta}$, the azimuth angle $-\pi \leq \theta \leq \pi$, and the elevation angle $-\pi/2 \leq \phi \leq \pi/2$. The definition of the angles follows the standard geographic convention. The angle-dependent TDOA (far field) can now be written as

$$\tau_{\mu\nu}(\boldsymbol{\zeta}(\theta, \phi)) = \frac{1}{c}\left[(\mathbf{r}_\mu - \mathbf{r}_\nu)^\mathsf{T} \cdot \boldsymbol{\zeta}(\theta, \phi)\right]. \tag{6}$$

2.3. GCC-PHAT

In this paper the generalized cross-correlation (GCC) method in combination with a phase transform (PHAT) weighting [6] is used for TDOA estimation. For a pair of microphones (μ, ν), a Hann window of length K is applied to the sampled signals $y_\mu(n)$ and $y_\nu(n)$ with the discrete time index n, and the discrete Fourier transforms (DFTs) $Y_\mu(\ell, k)$ and $Y_\nu(\ell, k)$ with frequency bin k and frame index ℓ are computed (the frame index ℓ will be omitted in the following). The GCC-PHAT function is then computed by [6]

$$\varphi_{\mu\nu}^{\mathrm{PHAT}}(\tau) = \frac{1}{K}\sum_{k=0}^{K-1} \frac{Y_\mu(k)Y_\nu^*(k)e^{j2\pi\frac{k\tau}{K}}}{|Y_\mu(k)Y_\nu^*(k)|} \tag{7}$$

with $(\cdot)^*$ denoting the complex conjugate.

2.4. Steered Response Power (SRP)

Due to sound reflections within the room, the TDOA estimate computed by maximizing (7) w. r. t. τ can be inaccurate or even wrong, as more than one local maximum can exist. To overcome this problem the steered response power (SRP) [10] method is used here, which is based on the variation of τ in (7). Under the far-field assumption the search space $\mathcal{C} \subset \mathbb{R}^3$ is expressed in discretized angles to represent different directions of arrival (DOAs): The resulting search space is then given by $\mathcal{Q} = \mathcal{A} \times \mathcal{E} = \{(\theta, \phi) | \theta \in \mathcal{A}, \phi \in \mathcal{E}\} \subset \mathbb{R}^2$ with $\mathcal{A} = \{\theta_{\min}, \ldots, \theta_{\max}\}$, $\mathcal{E} = \{\phi_{\min}, \ldots, \phi_{\max}\}$, and $\times$ denoting the Cartesian product. Each pair of angles $(\theta, \phi) \in \mathcal{Q}$, corresponds to a specific $\tau_{\mu,\nu}(\boldsymbol{\zeta}(\theta, \phi))$ for each microphone pair $(\mu, \nu) \in \mathcal{P} \subset \mathcal{M}^2$. The ranges of $\mathcal{A}$ and $\mathcal{E}$ need to be chosen task-dependent. Employing $\tau = \tau_{\mu,\nu}(\boldsymbol{\zeta}(\theta, \phi))$ in (7), leads to $\varphi_{\mu\nu}^{\mathrm{PHAT}}(\boldsymbol{\zeta}(\theta, \phi))$, which, expressed as a function of $\boldsymbol{\zeta}$ (and in this way of θ and ϕ), can be interpreted as a spatial likelihood function (SLF) for each pair of microphones (μ, ν). The SLF should show a maximum value belonging to an explicit DOA $\boldsymbol{\zeta}$. The sum over all microphone pairs (or at least more than one) gives a more precise DOA estimate, but on the other hand increases the computational complexity. We will call the sum of the SLFs over all microphone pairs the total spatial likelihood function (TSLF) which is expressed as

$$S_\mathcal{P}(\boldsymbol{\zeta}) = \frac{1}{|\mathcal{P}|}\sum_{(\mu,\nu)\in\mathcal{P}} \varphi_{\mu\nu}^{\mathrm{PHAT}}(\tau_{\mu\nu}(\boldsymbol{\zeta})). \tag{8}$$

3. PROPOSED ACOUSTIC EVENT LOCALIZATION

After having revisited some basics to DOA estimation in the previous section, this section introduces additional new components of our algorithmic framework.

3.1. Sound Activity and Event Onset Detection (SAD, EOD)

In the following the sound activity detection (SAD) and the event onset detection (EOD) are described, as they will be used later in the localization process (Section 3.2). Based on noise variance tracking, they are derived on the basis of a voice activity detection [18]. For each microphone μ (for better readability the microphone index will be omitted in wide parts of this subsection), a SAD and an EOD decision have to be made. Each frame ℓ of any microphone signal $y(n)$ is divided into subframes $\ell' \in \mathcal{L}' = \{1, \ldots, L'\}$. The noise variance $\sigma_N^2(\ell', k')$ is then estimated by a 3–state sound activity detector (SAD) in the DFT domain, with frequency bin $k' \in \{0, \ldots, K'/2\}$. Let the smoothed periodogram of the microphone signal be

$$\overline{|Y(\ell', k')|^2} = \beta_Y \cdot \overline{|Y(\ell'-1, k')|^2} + (1-\beta_Y) \cdot |Y(\ell', k')|^2$$

with $\beta_Y \in [0, 1]$ and $\Theta(\ell', k')$ a dynamic threshold. In each subframe one sound activity hypothesis $H(\ell', k')$ is determined out of three options $\mathcal{H} = \{H_{\mathrm{SP}}, H_{\mathrm{SA}}, H_{\mathrm{ST}}\}$:
Sound *presence* H_{SP} is assumed if

$$\overline{|Y(\ell', k')|^2} > 2 \cdot \Theta(\ell'-1, k').$$

Sound *absence* H_{SA} is assumed if

$$\overline{|Y(\ell', k')|^2} \leq 2 \cdot \Theta(\ell'-1, k') \ \wedge \ \overline{|Y(\ell', k')|^2} < \widehat{\sigma_N^2}(\ell'-1, k').$$

Sound *transition* H_{ST} is assumed if

$$\overline{|Y(\ell', k')|^2} \leq 2 \cdot \Theta(\ell'-1, k') \ \wedge \ \overline{|Y(\ell', k')|^2} \geq \widehat{\sigma_N^2}(\ell'-1, k').$$

The noise variance estimate $\widehat{\sigma_N^2}(\ell', k')$ is updated as

$$\widehat{\sigma_N^2}(\ell', k') = \epsilon(\ell', k') \cdot \widehat{\sigma_N^2}(\ell'-1, k') + \left[1 - \epsilon(\ell', k')\right] \cdot \overline{|Y(\ell', k')|^2}$$

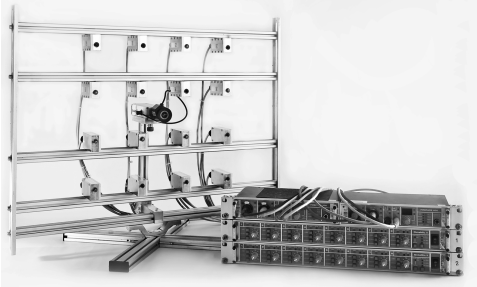

Fig. 1. Microphone array with camera and recording equipment.

with the initial value $\widehat{\sigma_N^2}(\ell'=0,k')=0$ and $\epsilon(\ell',k')$ denoting a time-varying smoothing factor, depending on the sound activity hypothesis of the previous frame $H(\ell'-1,k')$. The time-varying smoothing factor $\epsilon(\ell',k')$ and thy dynamic threshold $\Theta(\ell',k')$ are chosen according to [18].

Now for each subframe ℓ' (and channel μ) the decision $\mathrm{SAD}_\mu(\ell')$ is made upon the sound activity hypotheses $H(\ell',k')$ in that channel. We decide for $\mathrm{SAD}_\mu(\ell')=1$ if at least 60% of its frequency bins in the range [500 Hz, 5000 Hz] are classified by the 3–state SAD as H_{SP}, otherwise $\mathrm{SAD}_\mu(\ell')=0$.

The single decisions for each subframe ℓ' need to be joined to a decision for each frame ℓ. This is done by

$$\mathrm{SAD}_\mu(\ell)=\begin{cases}1, & \text{if } \sum_{\ell'\in\mathcal{L}'}\mathrm{SAD}_\mu(\ell')\geq\delta_{\mathcal{L}'}\\ 0, & \text{else.}\end{cases} \tag{9}$$

Finally, the sound activity decisions $\mathrm{SAD}_\mu(\ell)$ are joined to an overall sound activity decision

$$\mathrm{SAD}(\ell)=\begin{cases}1, & \text{if } \sum_{\mu\in\mathcal{M}}\mathrm{SAD}_\mu(\ell)\geq\delta_{\mathcal{M}}\\ 0, & \text{else.}\end{cases} \tag{10}$$

Based upon the *subframe* sound activity decision $\mathrm{SAD}_\mu(\ell')$, we propose an event onset detector (EOD). In a first step the subframe event onset decision

$$\mathrm{EOD}(\ell')=\begin{cases}1, & \text{if } \prod_{\lambda'\in\{\ell',\ldots,\ell'+L_{\min}-1\}}\mathrm{SAD}_\mu(\lambda')=1\\ 0, & \text{else,}\end{cases}$$

is made. This operation requires a lookahead of $L_{\min}-1$ subframes and ensures that there are at least $L_{\min}$ consecutive future subframes marked as active sound. By this means the performance of the whole framework can be optimized, as single subframes marked as active sound are ignored and a minimum event length is ensured. Following the hierarchy of the SAD, the frame- and channel-wise onset decision $\mathrm{EOD}_\mu(\ell)$ (c. f. (9)) and the overall event onset decision $\mathrm{EOD}(\ell)$ (c. f. (10)) are computed. The parameters $\delta_{\mathcal{L}'}$, $\delta_{\mathcal{M}}$, and $L_{\min}$ have to be chosen dependent on the task.

3.2. Spatial Minimum Tracking, Smoothing, and Localization

Estimating the DOA by maximizing (8) w. r. t. $\boldsymbol{\zeta}$ may still be affected by acoustic disturbances. Within this work a Wiener-type filter is used to suppress spatial noise and reverberation.

In case of sound *absence* ($\mathrm{SAD}(\ell)=0$) and using frame index ℓ, the noise floor (NF) of the SLF is simply estimated by (c. f. (8))

$$S_{\mathrm{NF},\ell}(\boldsymbol{\zeta})=S_{\mathcal{P},\ell}(\boldsymbol{\zeta}). \tag{11}$$

In case of sound *presence* ($\mathrm{SAD}(\ell)=1$), potential sound source positions are deleted from the desired noise floor by applying the following spatial minimum tracking

$$S_{\mathrm{NF},\ell}(\boldsymbol{\zeta})=\min\left(S_{\mathcal{P},\ell}(\boldsymbol{\zeta}),\ \frac{1}{|\mathcal{C}_{\boldsymbol{\zeta}}|}\sum_{\boldsymbol{\zeta}'\in\mathcal{C}_{\boldsymbol{\zeta}}}S_{\mathcal{P},\ell}(\boldsymbol{\zeta}')\right), \tag{12}$$

with $\mathcal{C}_{\boldsymbol{\zeta}}=\{\boldsymbol{\zeta}(\theta',\phi')|\theta-\delta\leq\theta'\leq\theta+\delta,\phi-\delta\leq\phi'\leq\phi-\delta\}$ being the space of vectors $\boldsymbol{\zeta}(\theta',\phi')$ belonging to a squared vicinity of $\boldsymbol{\zeta}(\theta,\phi)$ in the 2-dimensional spherical coordinate space. Independent of the SAD decision, in both cases a (temporal) first-order IIR filter is employed

$$\widetilde{S}_{\mathrm{NF},\ell}(\boldsymbol{\zeta})=\beta_{\mathrm{NF}}\cdot\widetilde{S}_{\mathrm{NF},\ell-1}(\boldsymbol{\zeta})+(1-\beta_{\mathrm{NF}})\cdot S_{\mathrm{NF},\ell}(\boldsymbol{\zeta}), \tag{13}$$

with the initial value $\widetilde{S}_{\mathrm{NF},0}(\boldsymbol{\zeta})=0$, and forgetting factor $\beta_{\mathrm{NF}}\in[0,1]$. In case of no detected event onset ($\mathrm{EOD}(\ell)=0$), processing for the current frame ℓ stops here.

In case of a detected event onset ($\mathrm{EOD}(\ell)=1$), a spatial *a priori* SNR ($\xi_\ell(\boldsymbol{\zeta})\geq 0$ c. f. (12))

$$\xi_\ell(\boldsymbol{\zeta})=\frac{S^2_{\mathcal{P},\ell}(\boldsymbol{\zeta})-\widetilde{S}^2_{\mathrm{NF},\ell}(\boldsymbol{\zeta})}{\widetilde{S}^2_{\mathrm{NF},\ell}(\boldsymbol{\zeta})} \tag{14}$$

is calculated. This *a priori* SNR can now be used to compute a Wiener-type spatial weight and to obtain an enhanced SLF

$$S^{\mathrm{opt}}_{\mathcal{P},\ell}(\boldsymbol{\zeta})=S_{\mathcal{P},\ell}(\boldsymbol{\zeta})\cdot\frac{\xi_\ell(\boldsymbol{\zeta})}{1+\xi_\ell(\boldsymbol{\zeta})}. \tag{15}$$

Even the enhanced spatial likelihood function $S^{\mathrm{opt}}_{\mathcal{P},\ell}(\boldsymbol{\zeta})$ may still show several local maxima, therefore, we propose to smooth it with a 2-dimensional Gaussian lowpass filter in the spherical coordinate system leading to the smoothed SLF

$$\overline{S}^{\mathrm{opt}}_{\mathcal{P},\ell}(\theta,\phi)=\left(\frac{1}{2\pi\sigma^2}e^{-\frac{\theta^2+\phi^2}{2\sigma^2}}\right)*S^{\mathrm{opt}}_{\mathcal{P},\ell}(\boldsymbol{\zeta}(\theta,\phi)) \tag{16}$$

with $*$ denoting the convolution operation. Finally, the estimated DOA in terms of θ and ϕ is given by

$$\hat{\boldsymbol{\zeta}}=\boldsymbol{\zeta}(\hat{\theta},\hat{\phi}),\ \text{ with } (\hat{\theta},\hat{\phi})=\arg\max_{(\theta,\phi)\in\mathcal{Q}}\overline{S}^{\mathrm{opt}}_{\mathcal{P},\ell}(\theta,\phi) \tag{17}$$

and is estimated only in frames with $\mathrm{EOD}(\ell)=1$. As the search space $\mathcal{Q}$ is spanned by only two angles (θ,ϕ) we in fact end up with a 2-dimensional optimization.

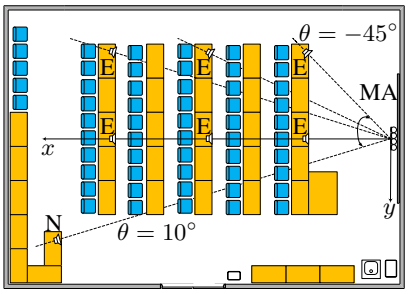

Fig. 2. Lecture hall with positions of the microphone array (MA), event sources (E), and the noise source (N).

4. EVALUATION SETUP AND RESULTS

4.1. Array Setup and Data Acquisition

For our experiments, a microphone array (see Fig. 2, MA) was placed in a medium size lecture hall. The array consists of $4\times 4=16$

	SNR					
	0 dB	5 dB	10 dB	15 dB	20 dB	∞
REF-str	28.04	24.37	22.84	20.37	19.72	18.59
REF-bab	63.59	57.92	53.31	49.15	45.63	
SAD-str	24.26	22.21	21.63	19.37	19.01	17.51
SAD-bab	61.70	56.34	52.11	48.57	45.06	
EOD-str	**18.76**	**17.33**	**17.62**	**15.67**	**16.08**	**17.10**
EOD-bab	**59.17**	**53.55**	**49.04**	**45.26**	**41.21**	

Table 1. Miss ratio (MR) in percent, for the reference framework REF and our two new approaches.

	SNR					
	0 dB	5 dB	10 dB	15 dB	20 dB	∞
REF-str	15.07	14.63	14.31	13.20	12.80	10.16
REF-bab	30.76	30.69	30.06	29.65	28.30	
SAD-str	14.20	13.92	13.86	12.83	12.42	10.16
SAD-bab	31.83	31.63	30.92	30.34	28.52	
EOD-str	**11.06**	**10.85**	**11.27**	**10.26**	**10.30**	**9.90**
EOD-bab	**30.75**	**30.51**	**29.50**	**28.88**	**26.58**	

Table 2. Average estimation error (AEE) *of missed frames* in degrees, for the reference framework REF and our new approaches.

microphones, equidistantly arranged with 10 cm spacing, with a camera being placed in the center (see Fig. 1). The recordings were made at a sampling frequency of 48 kHz and later downsampled to 16 kHz for simulations. At 6 room positions (E) 10 files of 50 classes from the RWCPSSDRAE database [19] were played back by a broadband loudspeaker positioned with the membrane facing the array. In addition two types of noise (babble noise (denoted as 'bab') from the NOISEX-92 database [20] and street noise (denoted as 'str') from the NTT Ambient Noise database [21]) were played back at a seventh position in the back of the room (N).

4.2. Algorithm Setup and Evaluation Methodology

For our evaluation, the recorded acoustic events were split into three sets: development (20%), development-test (20%), and test (60%). In consequence, all results given in this paper are averaged over 6 positions $\times$ 6 files $\times$ 50 classes $=$ 1800 different single acoustic events. To explore the influence of noise to the proposed algorithm, different signal-to-noise ratios (SNRs) where chosen and processed.

For evaluation we use two common metrics, based on the Euclidean distance of the estimated $(\hat{\theta}, \hat{\phi})$ to the original DOA (θ, ϕ) in degrees

$$\Delta_{\mathrm{DOA}} = \sqrt{(\hat{\theta} - \theta)^2 + (\hat{\phi} - \phi)^2}.$$

At first the miss ratio (MR) is calculated for each test case, giving the percentage of position estimates where $\Delta_{\mathrm{DOA}} > 3^\circ$. Second, the average estimation error (AEE) is calculated as the average over all Δ_{DOA}. All parameters were optimized on clean and 10dB SNR data, minimizing the miss ratio on the development-test dataset in each case, as in our use case it is most important to recognize an event within $\Delta_{\mathrm{DOA}} \leq 3^\circ$.

The reference (baseline) results are calculated by the framework from [17], modified for DOA estimation using the DOA search space and our frequency-domain sound activity detection, henceforth listed 'REF'. Our new framework without using the event onset detection is listed as 'SAD': Localization is then performed in any frame with $\mathrm{SAD}(\ell) = 1$. The third framework dubbed by 'EOD' is the whole framework presented in this work, including all functions as presented in Section 3: SAD, EOD, spatial minimum tracking, smoothing, and then localization. All three frameworks use the following parameters: GCC-PHAT-SRP framelength $K = 4096$ (no overlap), SAD/EOD framelength $K' = 512$ (overlap 256 samples) resulting in $L' = 16$, $\beta_y = 0.1$, $\delta_{\mathcal{L}'} = 9$, $\delta_{\mathcal{M}} = 2$, $L_{\min} = 3$, the vicinity of $\boldsymbol{\zeta}$ is given by $\delta = 3$, spatial noise floor smoothing constant $\beta_{\mathrm{NF}} = 0.9$, smoothing filter standard deviation $\sigma = 3^\circ$. Through the optical specifications of the camera in the array's origin, and as the source position should be used to steer a second camera the searchspace $\mathcal{Q}$ is spanned by $\mathcal{A} = \{-55^\circ, -54^\circ, \ldots, 55^\circ\}$ and $\mathcal{E} = \{-47^\circ, -46^\circ, \ldots, 47^\circ\}$. To reduce computational costs, only $M = 4$ microphones at the four corners of the array were used for the experiments.

4.3. Evaluation Results and Discussion

At first, have a look at the miss ratio (MR) in Table 1, a measure for the robustness of the localization algorithm. It is clearly visible that our new SAD framework decreases the miss ratio by about 1% in clean condition, compared to the REF framework. Also in noisy conditions the miss ratio is decreased. For street noise the improvement amends to $0.7\% \ldots 3.8\%$, and for babble noise to $0.6\% \ldots 1.9\%$.

Now compare the SAD to the EOD framework. Through the introduction of the event onset detection the miss ratio can be further decreased. Under clean condition the gain amounts to 0.4%. For street noise the miss ratio is further decreased by about $3\% \ldots 5.5\%$. Under the influence of babble noise it amounts to $2.5\% \ldots 3.8\%$.

Directly compared to the REF framework, the EOD framework significantly decreases the miss ratio under noisy conditions, gaining an absolute improvement of about $3.6\% \ldots 9.3\%$ under street noise and $3.9\% \ldots 4.5\%$ under babble noise. Obviously the EOD framework clearly outperforms the REF framework in all SNR and in all noise conditions.

As the main goal of the proposed algorithm is to gain a high recognition rate of events and to localize them close to the original location, the miss ratio is the most important performance measure. Nevertheless, now have a look at the average estimation error (AEE) *for missed frames* in Table 2 for the evaluation of the precision of the proposed algorithm. Here the goal should be to lower this error as in this way *all* position estimates become more precise. We observe that the REF framework and the SAD framework perform in the same range. Under street noise the SAD framework is slightly more precise gaining about $0.4^\circ \ldots 0.8^\circ$, whereas for babble noise the REF framework is slightly better by about $0.4^\circ \ldots 1^\circ$.

For the EOD framework compared to the REF framework, a significant improvement in street noise (up to 4°), and still a slight improvement in babble noise are observed.

Altogether we can summarize, that the proposed new EOD framework clearly outperforms the REF framework both in miss ratio and average estimation error. For street noise the new EOD approach performs approximately equally in the whole investigated SNR range.

5. CONCLUSION

In this paper we derived a framework for acoustic event source localization with a microphone array. A GCC-PHAT-SRP framework is supported by a frequency-domain sound activity detection and event onset detector. For the suppression of artifacts in the spatial likelihood function, spatial minimum tracking and smoothing are employed. The new framework clearly outperforms the reference approach, by reducing *both* the miss ratio up to 9% absolute, and increasing the overall precision in non-stationary noise and several signal-to-noise ratios by up to 4°.

6. REFERENCES

[1] H. Wang and P. Chu, "Voice Source Localization for Automatic Camera Pointing System in Videoconferencing," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Munich, Germany, Apr. 1997, vol. 1, pp. 187–190.

[2] C. Busso, S. Hernanz, C.W. Chu, S. Kwon, S. Lee, P. G. Georgiou, I. Cohen, and S. Narayanan, "Smart Room: Participant and Speaker Localization and Identification," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, PA, USA, Mar. 2005, vol. 2, pp. 1117–1120.

[3] C. Doukas and I. Maglogiannis, "Advanced Patient or Elder Fall Detection Based on Movement and Sound Data," in *Proc. of Pervasive Health*, Tampere, Finland, Jan. 2008, pp. 103–107.

[4] A.R. Abu-El-Quran and R.A. Goubran, "Security-Monitoring using Microphone Arrays and Audio Classification," in *Proc. of IEEE Instrumentation and Measurement Technology Conference (IMTC)*, Ottawa, Canada, May 2005, vol. 2, pp. 1144–1148.

[5] J.-M. Valin, F. Michaud, J. Rouat, and D. Letourneau, "Robust Sound Source Localization using a Microphone Array on a Mobile Robot," in *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Las Vegas, NV, USA, Oct. 2003, vol. 2, pp. 1228–1233.

[6] C. Knapp and G. C. Carter, "The Generalized Correlation Method for Estimation of Time Delay," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[7] C. Zhang, D. Florencio, and Z. Zhang, "Why Does PHAT Work Well in Low Noise, Reverberative Environments?," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, USA, Mar. 2008, pp. 2565–2568.

[8] M. Omologo and P. Svaizer, "Acoustic Event Localization Using a Crosspower-Spectrum Phase Based Technique," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Adelaide, Australia, Apr. 1994, vol. 2, pp. 273–276.

[9] P. Svaizer, M. Matassoni, and M. Omologo, "Acoustic Source Location in a Three-Dimensional Space Using Crosspower Spectrum Phase," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Munich, Germany, Apr. 1997, vol. 1, pp. 231–234.

[10] J. H. DiBiase, *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays*, Ph.D. thesis, Brown University, Providence, RI, USA, May 2000.

[11] A. Ikeda, H. Mizoguchi, Y. Sasaki, T. Enomoto, and S. Kagami, "2D Sound Source Localization in Azimuth and Elevation from Microphone Array by Using a Directional Pattern of Element," in *Proc. of. IEEE Sensors Conference*, Atlanta, GA, USA, Oct. 2007, pp. 1213–1216.

[12] J.P. Bello, C. Duxbury, M. Davies, and M. Sandler, "On the Use of Phase and Energy for Musical Onset Detection in the Complex Domain," *IEEE Signal Processing Letters*, vol. 11, no. 6, pp. 553–556, June 2004.

[13] J.P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and Mark B. Sandler, "A Tutorial on Onset Detection in Music Signals," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 1035–1047, Sept. 2005.

[14] A. Klapuri, "Sound Onset Detection by Applying Psychoacoustic Knowledge," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Phoenix, AZ, USA, Mar. 1999, vol. 6, pp. 3089–3092.

[15] G. Hu and D.L. Wang, "Auditory Segmentation Based on Onset and Offset Analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 396–405, Feb. 2007.

[16] J. Huang, N. Ohnishi, and N. Sugie, "A Biomimetic System for Localization and Separation of Multiple Sound Sources," *IEEE Transactions on Instrumentation and Measurement*, vol. 44, no. 3, pp. 733–738, June 1995.

[17] F. Hummes, J. Qi, and T. Fingscheidt, "Robust Acoustic Speaker Localization with Distributed Microphones," in *Proc. of European Signal Processing Conference (EUSIPCO)*, Barcelona, Spain, Aug. 2011, pp. 240 – 244.

[18] B. Fodor and T. Fingscheidt, "Reference-Free SNR Measurement for Narrowband and Wideband Speech Signals in Car Noise," in *Proc. of 10th ITG Conference on Speech Communication*, Braunschweig, Germany, Sept. 2012, pp. 199–202.

[19] S. Nakamura, K. Hiyane, F. Asano, and T. Endo, "Sound Scene Data Collection in Real Acoustical Environments," *The Journal of the Acoustic Society of Japan*, vol. 20, no. 3, pp. 225–231, May 1999.

[20] A. Varga and H. J. Steeneken, "Assessment for Automatic Speech Recognition: II. NOISEX-92: A Database and an Experiment to Study the Effect of Additive Noise on Speech Recognition Systems ," *Speech Communication*, vol. 12, no. 3, pp. 247 – 251, 1993.

[21] "Ambient Noise Database for Telephonometry," NTT Advanced Technology Corporation (NTT-AT), 1996.