ROBUST ACOUSTIC SPEAKER LOCALIZATION WITH DISTRIBUTED MICROPHONES

Frithjof Hummes, Junge Qi, Tim Fingscheidt

TU Braunschweig, Institute for Communications Technology,
Schleinitzstr. 22, 38106 Braunschweig, Germany
{hummes, qi, fingscheidt}@ifn.ing.tu-bs.de

ABSTRACT

This contribution to acoustic source localization presents a robust approach verified with ten distributed microphones in a laboratory apartment under reverberant acoustic conditions. Based on the classical steered response power phase transform (SRP-PHAT) algorithm, three optional extensions are presented: A method for selecting suitable microphone pairs, a spatial Wiener-type filtering for the suppression of artifacts *in the spatial likelihood function* (stemming from background noise), and finally smoothing of the spatial likelihood function. Simulation results show a significant improvement compared to SRP-PHAT in all noise conditions.

1. INTRODUCTION

The knowledge of a speaker's position can be useful in many respects and has therefore been investigated intensively during the last two decades. Applications include teleconferencing, smart rooms [1], and acoustic surveillance for safety and security purposes [2]. A recent topic finding increasing attention is ambient assisted living, i. e., technologies supporting the elderly in their home environment. Here, position tracks of the inhabitant can directly deliver information about the health status or untypical behavior, e. g., if the person has tumbled [3]. Additionally, the speaker's position offers the opportunity for further processing, like beamforming and distant speech recognition. A person's position inside a room can be obtained in many different ways, all of which suffer from specific problems. Camera solutions are often rejected due to privacy reasons, pressure sensitive carpets are relatively expensive, other systems do even need active elements worn at the person's body or clothes. Given the inhabitant is speaking or even crying for help, microphones, however, are a good compromise since they are cheap, small, and already ubiquitous (think of fixed and mobile telephones). They can also be used for further applications in parallel, e.g., voice control or hands-free telephony (emergency call). But they suffer from adverse acoustical conditions like reflections, reverberation and noise.

The minimum number of microphones for obtaining the position of an acoustic source is three. However, often times significantly more microphones are used [4]. They can be clustered yielding microphone arrays as proposed in [1, 5] or distributed loosely on walls or on the ceiling [6]. Other approaches use distributed microphone arrays [7]. In order to reduce complexity and to increase robustness, a spatial observability function has been introduced in [7] as a confidence measure. Another recent approach is to draw even advantage of the first reflections [8]. In some contributions, e. g., [9], an energy detector is proposed for a pre-classification of suitable speech frames. Estimation results can be further enhanced and misleading maxima be reduced by smoothing the spatial likelihood function in space and time. In recent years, also particle filters have been proposed for localization purposes [10].

Our approach for speaker localization has been developed in the context of an installation of ten single microphones distributed approximately equidistantly on the walls surrounding a room of a laboratory apartment. We chose the steered response power phase transform (SRP-PHAT) algorithm [11] as baseline which in principle has been shown to be quite robust against reverberation [12].

Based on that we propose three extensions, each of which can be applied separately or in combination. First, we present an efficient spatial observability function which is derived from the cross-correlation function. We use this function for selecting good microphone pairs before fusing their contribution prior to localization. Second, we present a spatial Wiener-type filtering approach which suppresses the influence of noise sources on the speaker position estimation. Third, smoothing of the spatial likelihood function is applied. It is worth mentioning that the proposed approaches are suitable for real-time speaker tracking.

This paper proceeds with introducing the applied signal model and revisiting the fundamental localization algorithm SRP-PHAT in Section 2. Based on this, we propose our new three approaches in Section 3, comprising a microphone pair evaluation and selection, the Wiener-type filtering, and spatial smoothing. Simulation results on recordings in a laboratory apartment are presented in Section 4 before drawing conclusions in Section 5.

2. BASELINE ALGORITHMIC APPROACHES

2.1 Signal Model

Given a room with distributed microphones and a talking person who shall be localized. A certain number M of microphones μ, $1 \leq \mu \leq M$, with outputs $y_\mu(t)$ are located at position (vector) $\mathbf{r}_\mu$, respectively. A speech signal $s(t)$ is emitted from the sound source position $\mathbf{r}_s$, which for the moment is assumed to be time-invariant. It is then convolved with the impulse response $h_\mu(t)$ and subject to additive environmental noise $n(t)$ at microphone μ, which leads to

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

Neglecting the reverberation we obtain

$$y_\mu(t) = a_\mu \cdot x(t - \tau_\mu) + n_\mu(t), \tag{2}$$

where a_μ is an attenuation factor and

$$\tau_\mu = \tau_\mu(\mathbf{r} = \mathbf{r}_s) = \frac{|\mathbf{r}_\mu - \mathbf{r}_s|}{c} \tag{3}$$

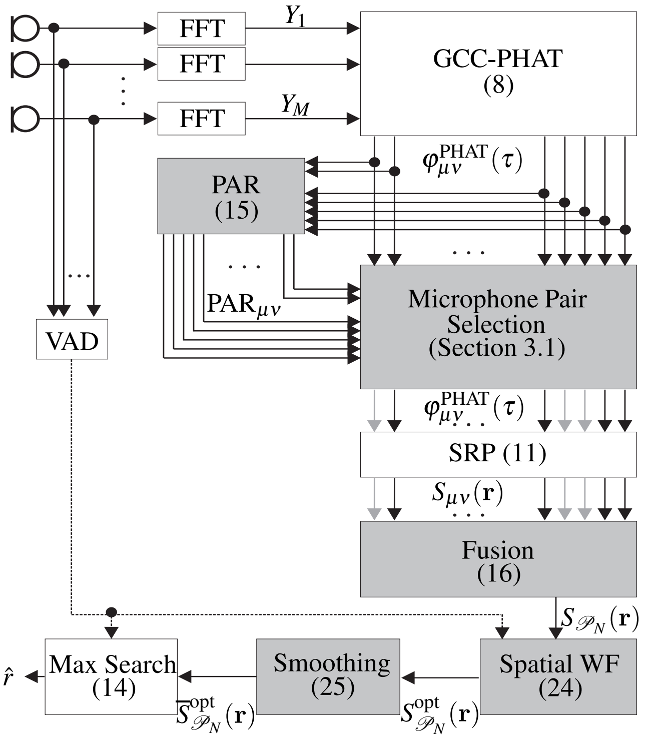
FFT
Y_1
FFT
FFT
Y_M
GCC-PHAT
(8)
$\varphi_{\mu\nu}^{\mathrm{PHAT}}(\tau)$
PAR
(15)
...
...
$\mathrm{PAR}_{\mu\nu}$
...
Microphone Pair
Selection
(Section 3.1)
VAD
$\varphi_{\mu\nu}^{\mathrm{PHAT}}(\tau)$
SRP (11)
$S_{\mu\nu}(\mathbf{r})$
Fusion
(16)
$S_{\mathscr{P}_N}(\mathbf{r})$
$\hat{r}$
Max Search
(14)
Smoothing
(25)
Spatial WF
(24)
$\bar{S}_{\mathscr{P}_N}^{\mathrm{opt}}(\mathbf{r})$
$S_{\mathscr{P}_N}^{\mathrm{opt}}(\mathbf{r})$

Figure 1: Block diagram of the entire localization system

being the time needed by the sound waves to travel from $\mathbf{r}_s$ to $\mathbf{r}_\mu$ at a velocity of $c = 343\,\mathrm{m/s}$. Regarding two microphones μ and ν,

$$\tau_{\mu\nu} \overset{\mathrm{def}}{=} \tau_\nu - \tau_\mu \tag{4}$$

is the time difference of arrival (TDOA).

2.2 GCC-PHAT

In order to estimate the TDOA for a given microphone pair the cross-correlation function of the sampled microphone signals $y_\mu(n)$ and $y_\nu(n)$ with discrete time index n has to be computed. This can be achieved by applying a rectangular window of length K, computing the discrete Fourier transforms $Y_\mu(k)$ and $Y_\nu(k)$ (a frame index ℓ is mostly omitted in the rest of the paper) with frequency bin k, and computing the following inverse DFT:

$$\varphi_{\mu\nu}(\tau) = \frac{1}{K}\sum_{k=0}^{K-1} Y_\mu(k)Y_\nu^*(k)e^{j2\pi\frac{k\tau}{K}}, \tag{5}$$

with $()^*$ denoting the complex conjugate. To achieve a generalized cross-correlation (GCC) function a complex-valued factor $G_{\mu\nu}$ is included into (5) leading to [13]:

$$\varphi_{\mu\nu}^{\mathrm{GCC}}(\tau) = \frac{1}{K}\sum_{k=0}^{K-1} G_{\mu\nu}(k)Y_\mu(k)Y_\nu^*(k)e^{j2\pi\frac{k\tau}{K}} \tag{6}$$

In a reverberant near-field speaker-microphone scenario, it is not reasonable to take signal attenuation into account. Hence, a phase transform (PHAT) is applied by choosing [13]

$$G_{\mu\nu}(k) = G_{\mu\nu}^{\mathrm{PHAT}}(k) = \frac{1}{|Y_\mu(k)Y_\nu^*(k)|}, \tag{7}$$

which finally leads to

$$\varphi_{\mu\nu}^{\mathrm{PHAT}}(\tau) = \sum_{k=0}^{K-1} \frac{Y_\mu(k)Y_\nu^*(k)e^{j2\pi\frac{k\tau}{K}}}{|Y_\mu(k)Y_\nu^*(k)|}. \tag{8}$$

The estimated TDOA between signals μ and ν is then given by

$$\hat{\tau}_{\mu\nu} = \arg\max_{\tau} \varphi_{\mu\nu}^{\mathrm{PHAT}}(\tau). \tag{9}$$

2.3 Steered Response Power (SRP)

Due to reflections the GCC function (8) usually shows several local maxima which can lead to a wrong estimation of $\tau_{\mu\nu}$ in (9). Varying τ in (8) corresponds to steering a beamformer over the search space and measuring the output power (steered response power (SRP)). As each point $\mathbf{r}$ in the spatially discretized search space $\mathscr{R}$, with $\mathscr{R} \subset \mathbb{R}^2$ or $\mathscr{R} \subset \mathbb{R}^3$, for instance, corresponds to a certain TDOA $\tau_{\mu\nu}$ for each microphone pair $(\mu,\nu) \in \mathscr{P}$, a specific generalized cross-correlation function $\varphi_{\mu\nu}^{\mathrm{PHAT}}(\tau = \tau_{\mu\nu}(\mathbf{r}))$ with

$$\tau_{\mu\nu}(\mathbf{r}) = \tau_\mu(\mathbf{r}) - \tau_\nu(\mathbf{r}) = \frac{1}{c}(|\mathbf{r}_\mu - \mathbf{r}| - |\mathbf{r}_\nu - \mathbf{r}|) \tag{10}$$

can be obtained. Expressed as a function of $\mathbf{r}$, the GCC function can be interpreted as a spatial likelihood function (SLF)

$$S_{\mu\nu}(\mathbf{r}) = \varphi_{\mu\nu}^{\mathrm{PHAT}}(\tau_{\mu\nu}(\mathbf{r})) \tag{11}$$

$$= \frac{1}{K}\sum_{k=0}^{K-1} G_{\mu\nu}^{\mathrm{PHAT}}(k)Y_\mu(k)Y_\nu^*(k)e^{-j2\pi\frac{k\tau_{\mu\nu}(\mathbf{r})}{K}} \tag{12}$$

for microphone pair (μ,ν). Please note, that especially (11) contributes significantly to computational complexity. Having more than one microphone pair enables us therefore to postpone the actual localization decision as the total spatial likelihood function at point $\mathbf{r}$ regarding all microphone pairs can first be expressed as

$$S_{\mathscr{P}}(\mathbf{r}) = \frac{1}{|\mathscr{P}|}\sum_{(\mu,\nu)\in\mathscr{P}} S_{\mu\nu}(\mathbf{r}). \tag{13}$$

The estimated sound source position

$$\hat{\mathbf{r}}_s = \arg\max_{\mathbf{r}\in\mathscr{R}} S_{\mathscr{P}}(\mathbf{r}) \tag{14}$$

is then typically chosen as maximum of the total spatial likelihood function.

3. NEW APPROACH

Fig. 1 shows the block diagram of the entire localization system. Prior to all further processing, a voice activity detection (VAD) is applied on the microphone signals. The signals are then transformed frame-by-frame into the frequency domain by a fast Fourier transform (FFT) and fed into the GCC-PHAT, where the cross-correlation between each microphone pair (μ,ν) is computed. Based on this, a peak-to-average ratio (PAR) can be calculated, which can be used for selecting suitable microphone pairs in the next block (Section 3.1). Their cross-correlation functions $\varphi_{\mu\nu}^{\mathrm{PHAT}}(\tau)$ are the inputs for achieving the steered response powers (SRPs) in the form of spatial likelihood functions (SLFs), which are then joined to a global SLF $S_{\mathscr{P}_N}(\mathbf{r})$. Now, an optional spatial Wiener-like filtering can be applied to reduce artifacts stemming from disturbing noise sources (Section 3.2), followed by an optional spatial smoothing function (Section 3.3). The speaker position is finally estimated by choosing the maximum value of the resulting spatial likelihood function.

3.1 Microphone Pair Selection and Fusion

Regarding changing acoustical properties like room impulse response or speaker directivity, it is recommended to continuously compute a confidence measure for each microphone pair $(\mu,\nu)\in\mathscr{P}$ in order to identify the most promising microphone pairs. This allows to focus the costly computation of (11) only on these microphone pairs. We propose the squared ratio of the (strongest) maximum to the average of the cross-correlation function (8) and call this kind of spatial observability function a peak-to-average ratio (PAR):

$$\mathrm{PAR}_{\mu\nu} = \left(\frac{\max\limits_{\tau\in\mathscr{D}} \varphi^{\mathrm{PHAT}}_{\mu\nu}(\tau)}{\frac{1}{|\mathscr{D}|}\sum_{\tau\in\mathscr{D}} \varphi^{\mathrm{PHAT}}_{\mu\nu}(\tau)} \right)^2, \tag{15}$$

with $\mathscr{D} = [\tau_{\min}\ \tau_{\max}]$ being the set of all possible values of τ for the given microphone positions (μ,ν) in search space $\mathscr{R}$. A single maximum leads to a high (good) PAR, whereas reflections decrease the confidence in the TDOA estimate for the respective microphone pair. This approach can also deal well with defective microphones as the PAR for affected pairs decreases significantly. As there exist $|\mathscr{P}| = 0.5(M^2-M)$ possible combinations for M applied microphones[1], it might also be useful to exclude obviously unsuitable pairs for efficiency purposes. We suggest to take only the N microphone pairs having the N highest PAR values, with $2\leq N\leq|\mathscr{P}|$ and define the resulting subset as $\mathscr{P}_N$.

Using only the selected microphone pairs in (13), we achieve

$$S_{\mathscr{P}_N}(\mathbf{r}) = \frac{1}{|\mathscr{P}_N|}\sum_{(\mu,\nu)\in\mathscr{P}_N} S_{\mu\nu}(\mathbf{r}). \tag{16}$$

At this point, one could already apply

$$\hat{\mathbf{r}}_s = \arg\max_{\mathbf{r}\in\mathscr{R}} S_{\mathscr{P}_N}(\mathbf{r}) \tag{17}$$

to achieve a potentially better localization performance compared to using (13) and (14).

3.2 Spatial Wiener-type Filtering

In order to reduce acoustic disturbances, we propose a spatial noise and reverberation suppression by using a Wiener-type filter. At first, a voice activity detection $\mathrm{VAD}_\mu\in\{0,1\}$ for each channel μ has to distinguish between frames where speech is present or absent. We use a very simple energy-based algorithm in the time domain. Its decisions are then joined for a global voice activity decision

$$\mathrm{VAD} = \begin{cases} 1, & \text{if } \sum_{\mu=1}^{M}\mathrm{VAD}_\mu \geq \theta^{VAD} \\ 0, & \text{else.} \end{cases} \tag{18}$$

In case of global speech *absence* (VAD=0), we estimate a noise floor (NF) of the SLF simply by

$$S_{\mathrm{NF}}(\mathbf{r}) = S_{\mathscr{P}_N}(\mathbf{r}). \tag{19}$$

In case of global speech *presence* (VAD=1), an upper threshold

$$\check{S}_{\mathscr{P}_N} = \min\left(\frac{1}{|\mathscr{R}|}\sum_{\mathbf{r}\in\mathscr{R}} S_{\mathscr{P}_N}(\mathbf{r}),\ \alpha_1\cdot\max_{\mathbf{r}\in\mathscr{R}}(S_{\mathscr{P}_N}(\mathbf{r}))\right) \tag{20}$$

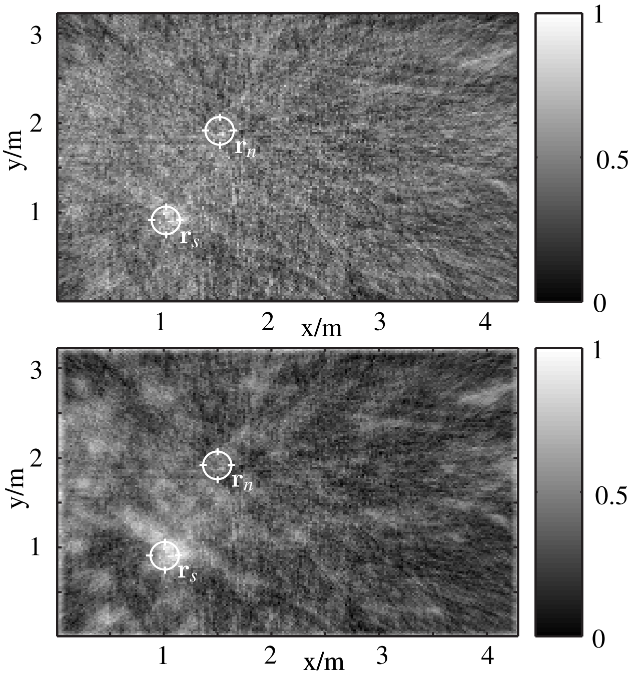

Figure 2: Example of a spatial likelihood function for speech presence before (upper) and after (lower) Wiener-type filtering at an SNR of 0 dB. The noise source's position is $\mathbf{r}_n = (1.50\,\mathrm{m}, 1.90\,\mathrm{m})$, the speaker is at $\mathbf{r}_s = (1.01\,\mathrm{m}, 0.89\,\mathrm{m})$.

with $0<\alpha_1<1$ is defined first. It is then further used to delete potential speaker positions from the desired noise floor

$$S_{\mathrm{NF}}(\mathbf{r}) = \min(S_{\mathscr{P}_N}(\mathbf{r}),\ \alpha_2\cdot\check{S}_{\mathscr{P}_N}) \tag{21}$$

with $\alpha_2>1$.

In both cases, speech absence and speech presence, a (spatial) lowpass filter is applied on $S_{\mathrm{NF}}(\mathbf{r})$ leading to $\overline{S}_{\mathrm{NF}}(\mathbf{r})$, followed by a (temporal) first-order IIR filter with forgetting factor β and frame index ℓ, according to

$$\widetilde{S}_{\mathrm{NF},\ell}(\mathbf{r}) = \beta\cdot\widetilde{S}_{\mathrm{NF},\ell-1}(\mathbf{r}) + (1-\beta)\cdot\overline{S}_{\mathrm{NF},\ell}(\mathbf{r}). \tag{22}$$

We initialize with $\widetilde{S}_{\mathrm{NF},0}(\mathbf{r})=0$. In case of speech absence, processing for the current frame ℓ stops and subsequent functions in Fig. 1 are not executed.

In case of global speech presence, a spatial *a posteriori* signal-to-noise ratio (SNR)

$$\mathrm{SNR}(\mathbf{r}) = \frac{S^2_{\mathscr{P}_N}(\mathbf{r})}{\widetilde{S}^2_{\mathrm{NF}}(\mathbf{r})} \tag{23}$$

can then be defined with the denominator being the squared noise floor of the SLF and the numerator being the squared *noisy* SLF from (16). This *a posteriori* SNR is then used in a Wiener-type filter to obtain the enhanced spatial likelihood function[2]

$$S^{\mathrm{opt}}_{\mathscr{P}_N}(\mathbf{r}) = S_{\mathscr{P}_N}(\mathbf{r})\cdot\frac{\max(\mathrm{SNR}(\mathbf{r})-1,\ 0)}{\mathrm{SNR}(\mathbf{r})}. \tag{24}$$

An example of the SLF in search space $\mathscr{R}$ is shown in Fig. 2, $S_{\mathscr{P}_N}(\mathbf{r})$ being drawn in the upper part, $S^{\mathrm{opt}}_{\mathscr{P}_N}(\mathbf{r})$ being shown in the lower subfigure.

[1] For $M=3$ microphones, we would get $\mathscr{P}=\{(1,2),(1,3),(2,3)\}$.

[2] Note that the respective *a priori* SNR would be $\mathrm{SNR}^a = \mathrm{SNR}-1$ leading to the typical Wiener filter formulation $\mathrm{SNR}^a/(1+\mathrm{SNR}^a)$.

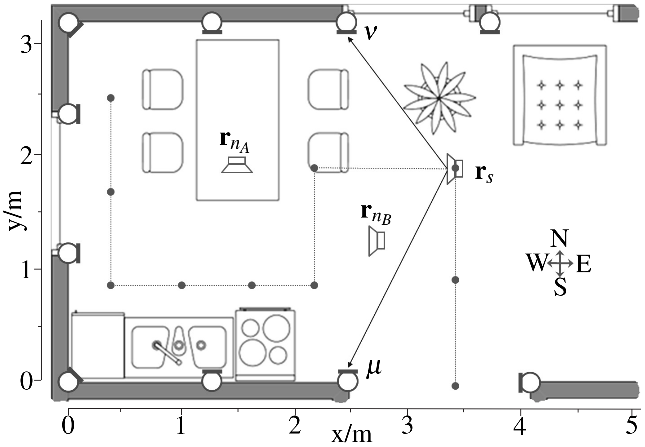

Figure 3: Room in the laboratory apartment with $M = 10$ microphones and 10 speaker position markers $\mathbf{r}_s$ (solid circles). The signal model for one position and one microphone pair (μ, ν) is depicted exemplarily. $\mathbf{r}_{n_A}$ and $\mathbf{r}_{n_B}$ represent the noise source positions.

3.3 Spatial Smoothing

Without loss of generality we assume a 2-dimensional localization task. As the spatial likelihood function (SLF) may show many local maxima, we propose to smooth the SLF $S^{\text{opt}}_{\mathscr{P}_N}(\mathbf{r})$ resulting in

$$\overline{S}^{\text{opt}}_{\mathscr{P}_N}(r_x, r_y) = H(r_x, r_y) * S^{\text{opt}}_{\mathscr{P}_N}(r_x, r_y), \qquad (25)$$

using the 2-dimensional Gaussian lowpass filter transfer function

$$H(r_x, r_y) = \frac{1}{2\pi\sigma^2} e^{-\frac{r_x^2 + r_y^2}{2\sigma^2}}. \qquad (26)$$

Note that if spatial Wiener-type filtering has been omitted, $S^{\text{opt}}_{\mathscr{P}_N}(r_x, r_y)$ in (25) just has to be replaced by $S_{\mathscr{P}_N}(r_x, r_y)$.

4. SIMULATION RESULTS

4.1 Data Acquisition

The data was recorded in a furnished laboratory apartment using $M = 10$ small omnidirectional electrostatic microphone capsules that are stuck directly to the wall at a height of 1.5 m. Ten representative speaker positions within an area of about $4 \times 3 = 12\,\text{m}^2$ were chosen as shown as solid circles in Fig. 3. The overall area of the combined kitchen and living room amounts to $29\,\text{m}^2$ and has a measured reverberation time of $\text{RT}_{60} \approx 0.6\,\text{s}$. Four utterances (2 male / 2 female) from the 16 kHz NTT speech database [14] were played back by a broadband loudspeaker located at the given speaker positions $\mathbf{r}_s$ and a height of 1.5 m. To simulate different head orientations of a human speaker, the recordings were repeated with the membrane directed towards the four cardinal directions, respectively. The microphone signals were sampled at a sampling frequency of $f_s = 48\,\text{kHz}$ for optimal hardware performance and were later downsampled to $f_s = 16\,\text{kHz}$ for the simulations. To investigate the influence of noise, white noise was played back from two different positions $\mathbf{r}_{n_A}$ and $\mathbf{r}_{n_B}$ separately. The desired signal-to-noise ratios (SNRs) from $-20\,\text{dB}$ to $+5\,\text{dB}$ were chosen in steps of 5 dB and are related to the outputs of the speech and noise loudspeakers, respectively. The SNR can be calculated as $\text{SNR} = \text{ASL}_s - \text{RMS}_n$ with ASL_s being the active speech level of the clean speech signal and RMS_n being the root mean square of the noise signal. Note that both are computed using ITU-T Recommendation P.56 [15].

4.2 Performance Metrics

As proposed in [5] we chose two different metrics for evaluating the performance of the localization algorithm. At first, the miss ratio (MR) is defined as the percentage of processed frames, where the estimated position has a Euclidean distance of more than 0.25 m from the correct position. Secondly, the average estimate error (AEE) is the average distance to the correct position*for all non-missed* frames.

4.3 Experimental Results

The common settings for the performed experiments were as follows: The discrete search stepsize was 2 cm and the length of the SRP-PHAT (or DFT) frames was $K = 4096$ samples with no overlap and all $0.5(M^2 - M) = 45$ possible microphone pairs were taken in (16). The parameters for estimating the noise floor in (20) and (21) are set to $\alpha_1 = 0.9$ and $\alpha_2 = 1.05$, and $\beta = 0.8$ in (22). For spatial smoothing based on the lowpass filter in (26), we chose $\sigma = 6\,\text{cm}$.

For each of the ten speaker positions, marked with solid circles in Fig. 3, the localization was performed on 16 clean speech samples. Additionally, each of the samples is subject to additive noise in 6 different SNR conditions for the two different noise source positions $\mathbf{r}_{n_A}$ and $\mathbf{r}_{n_B}$, separately, resulting in $16 \times 6 \times 2 = 192$ noisy samples per speaker position. This leads to $10 \times (16 + 192) = 2080$ processed samples, each of which results in approximately 20 position estimates. We did not differentiate between gender, cardinal directions, and noise source positions, which led to averaged results for each speaker position depending only on the SNR as shown in Fig. 4.

The dotted line represents the standard SRP-PHAT algorithm (13), (14), which performs acceptably under clean speech conditions, with 7% of the frames being missed. The miss ratio (MR) increases continuously with decreasing SNR. At an SNR of 0 dB, more than 51% of the frames are missed.

The Wiener-type filtering (WF, (24)) reduces the MR to 2%–40%, while spatially smoothing the SLF (Smooth, (25)) yields results between 5%–23%, both with significantly decreasing miss ratio towards higher SNR values. An almost SNR-independent localization method is achieved by combining both approaches (solid line) showing an additive gain. A MR between only 1%–8% is achieved.

The MR is a measure for the robustness of the localization algorithm. To evaluate the precision of the proposed approach, Table 1 shows the average estimate error (AEE) for the respective cases. There is only a slight decrease of precision using the proposed approaches. When both Wiener-type filtering and smoothing are applied, the AEE remains almost constant at 5.4 cm for all SNR conditions. The reason for higher precision at low SNR values is, that only non-missed frames are evaluated for obtaining the AEE.

In a second simulation we evaluated the microphone pair selection as proposed in Section 3.1. We reduced the maximal amount of $0.5(M^2 - M) = 45$ pairs to those $N = 30$ and $N = 15$ pairs revealing the highest $\text{PAR}_{\mu\nu}$ values, respec-

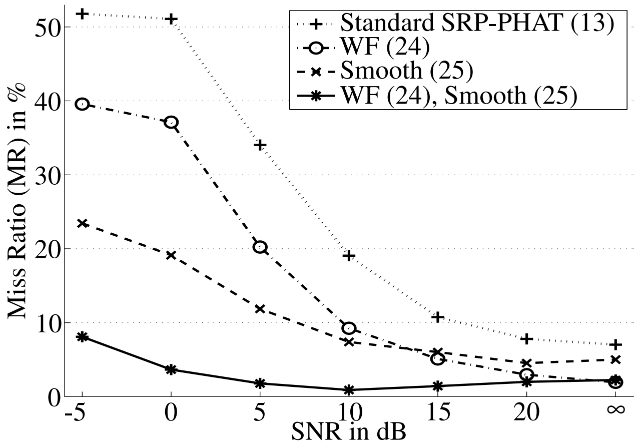

Figure 4: Average miss ratios (MR) for different SNR conditions.

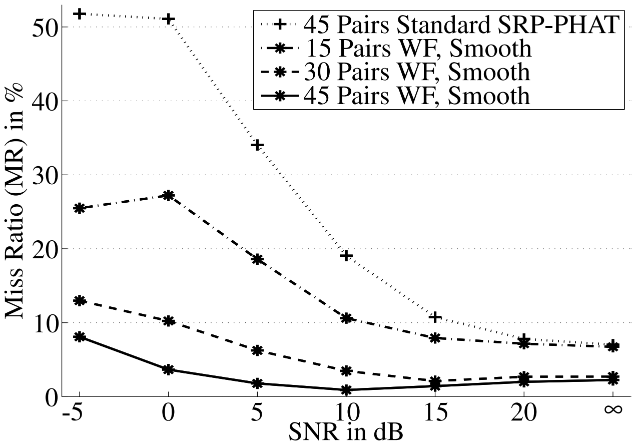

Figure 5: Influence of the microphone pair selection on the miss ratio (MR) when using Wiener-type filtering and smoothing

SNR/dB	-5	0	5	10	15	20	∞
Standard SRP-PHAT	2.0	1.9	2.5	3.1	3.4	3.6	3.6
WF	2.5	2.5	3.2	3.5	3.7	3.8	3.7
Smooth	4.0	4.4	4.7	4.9	5.0	5.0	4.9
WF, Smooth	5.1	5.5	5.5	5.5	5.5	5.4	5.4

Table 1: Average estimate error (AEE) in cm for different SNR conditions, only for non-missed frames.

tively. The results in Fig. 5 show that for the WF, Smooth approach a selection of only $N = 15$ pairs is still better than the standard SRP-PHAT with $N = 45$ pairs. For moderate noise (SNR ≥ 15 dB), one third of the pairs can be left out with only marginally increasing the miss ratio. The additional computational complexity of all investigated algorithmic options does not exceed 5% of standard SRP-PHAT.

5. CONCLUSIONS

We proposed a speaker localization system using distributed microphones based on the well-known SRP-PHAT. Besides an efficient procedure to select relevant pairs of microphones, our approach comprises a Wiener-like filtering of the spatial likelihood function, as well as spatial smoothing. The latter two proposed algorithms allowed for a significantly lower miss ratio (1%–8%) than SRP-PHAT (7%–51%), while the estimation error of non-missed frames is only marginally raised to approximately 5 cm.

The whole system is easily scalable by using the microphone pair selection option and suitable for real-time speaker localization.

REFERENCES

[1] C. Busso, S. Hernanz, Chi-Wei Chu, Soon il Kwon, Sung Lee, P. G. Georgiou, I. Cohen, and S. Narayanan, “Smart Room: Participant and Speaker Localization and Identification,” in *Proc. IEEE ICASSP*, Philadelphia, PA, USA, 2005.

[2] A. R. Abu-El-Quran, R. A. Goubran, and A. D. C. Chan, “Security Monitoring using Microphone Arrays and Audio Classification,” *IEEE Transactions on Instrumentation and Measurement*, vol. 55, no. 4, pp. 1025–1032, 2006.

[3] C. Doukas and I. Maglogiannis, “Advanced Patient or Elder Fall Detection based on Movement and Sound Data,” in *Proc. PervasiveHealth*, Tampere, Finland, 2008, pp. 103–107.

[4] H. F. Silverman, W. R. Patterson, J. L. Flanagan, and D. Rabinkin, “A Digital Processing System for Source Location and Sound Capture by Large Microphone Arrays,” in *Proc. IEEE ICASSP*, Munich, Germany, 1997, pp. 251–254.

[5] Pasi Pertilä, *Acoustic Source Localization in a Room Environment and at Moderate Distances*, Ph.D. thesis, Tampere University of Technology, Finland, 2009.

[6] C. Bartsch, A. Volgenandt, T. Rohdenburg, and J. Bitzer, “Evaluation of Different Microphone Arrays and Localization Algorithms in the Context of Ambient Assisted Living,” in *IWAENC*, Tel Aviv, Isreal, August 2010.

[7] P. Aarabi, “The Fusion of Distributed Microphone Arrays for Sound Localization,” *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 4, pp. 338 – 347, January 2003.

[8] F. Ribeiro, D. Ba, C. Zhang, and D. Florêncio, “Turning Enemies into Friends: Using Reflections to Improve Sound Source Localization,” in *Proc. IEEE ICME*, Singapore, July 2010, pp. 731–736.

[9] T. Machmer, A. Swerdlow, and K Kroschel, “Robust Impulsive Sound Source Localization by Means of an Energy Detector for Temporal Alignment and Pre-Classification,” in *Proc. EUSIPCO*, Glasgow, Scotland, August 2009, pp. 1409–1412.

[10] A. Löytynoja and P. Pertilä, “A Real-Time Talker Localization Implementation Using Multi-PHAT and Particle Filter,” in *Proc. EUSIPCO*, Glasgow, Scotland, August 2009, pp. 1418–1422.

[11] 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.

[12] C. Zhang, D. Florencio, and Z. Zhang, “Why Does PHAT Work Well in Low Noise, Reverberative Environments?,” in *Proc. IEEE ICASSP*, Las Vegas, NV, USA, 2008, pp. 2565–2568.

[13] C. H. 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, August 1976.

[14] NTT Advanced Technology Corporation, “Multi-Lingual Speech Database for Telephonometry 1994,” .

[15] ITU-T Recommendation P.56, “Objective Measurement of Active Speech Level,” 1993.