Audio Engineering Society

Convention Paper

Presented at the 148th Convention
2020 May 25 – 28, Vienna, Austria

*This paper was peer-reviewed as a complete manuscript for presentation at this convention. This paper is available in the AES E-Library (http://www.aes.org/e-lib) all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.*

# Acoustic Source Localization and High Quality Beamforming Using Coincident Microphone Arrays

Jonathan D. Ziegler[1,2], Hendrik Paukert[1], Andreas Koch[1], and Andreas Schilling[2]

[1]*Stuttgart Media University, Nobelstr. 10, 70569 Stuttgart*
[2]*Eberhard Karls Univerity, Sand 14, 72076 Tübingen*

Correspondence should be addressed to Jonathan D. Ziegler (zieglerj@hdm-stuttgart.de)

**ABSTRACT**

This paper presents an application-oriented approach to Acoustic Source Localization using a coincident microphone array. Multiple processing blocks are presented to generate a reactive, yet stable Direction of Arrival estimation tuned toward speaker tracking. Building on an energy based scanning method, individual characteristics, such as sound field directivity and static sound source positions are used for adaptive smoothing of the detected angle. The methods and resulting performance gain are discussed for the individual components of the algorithm. Objective performance is evaluated using simulated and recorded data. Audio quality is assessed using listening tests, which show a significant increase in subjective sound quality, noise suppression, and speech intelligibility when combining the tracker with a beamforming algorithm for coincident microphone arrays.

## 1 Introduction

With beamformers in mobile and smart-home devices gaining relevance, many applications focus on low-cost linear and circular arrays for Acoustic Source Localization (ASL) and tracking [1]. Advances in spherical array beamforming have enabled the creation of versatile, robust beamformers in three dimensions, often using spherical harmonics as an orthonormal base for beamforming [2–7]. Beamforming for professional high quality audio is still uncommon, as the large number of transducers needed for higher-order beams prevents the use of professional quality microphones [8]. Combinations of shotgun microphones and adaptive spectral beamformers have proven effective and can generate high quality audio [9]. The downside of such microphones is the need for manual, mechanical source tracking. Producing an audio signal of consistently high quality is difficult and requires skilled personnel. Recent research has shown that a first-order beamformer using a coincident microphone array can produce beam patterns similar to shotgun microphones - with a more linear frequency response for off angle sound incidence [10–12]. The combination of such beamformers with an effective algorithm for ASL can produce high quality audio signals of moving sources with Directivity Indices beyond the possibilities of classical first-order microphones [13]. This paper presents an application-oriented algorithm for real-time ASL using a coincident microphone array consisting of three high-end microphone capsules. The array configuration represents a simple setup that can be transformed onto a spherical harmonic base in two dimensions and create any beampattern expressible with spherical harmonics of $\mathcal{O}(1)$ [14]. The configuration of the capsules allows for first-order beamforming on the horizontal plane,

presenting an acceptable and well researched solution for many applications [15, 16]. A Steered Response Power (SRP) approach is chosen for initial Direction of Arrival (DOA) estimation, using virtual cardioid microphones as a scanning beam [17].

Details about the chosen microphone configuration, the resulting virtual microphone synthesis, and the ASL algorithm can be found in sections 2.1 and 2.2. A variable exponential smoothing algorithm increases the algorithm's angular stability, while maintaining high sensitivity for directional changes. The basic concept and the individual weighting factors are discussed in section 3. Performance is evaluated using objective error analysis and listening tests based on a set of subjective quality metrics. The experimental set up is discussed in section 4 and results are presented in section 5.

## 2 Acoustic Source Localization

### 2.1 Microphone Configuration

The described system uses a microphone configuration consisting of three high-end professional microphones. This guarantees a known and consistent frequency response of the individual capsules for on-axis as well as for off-axis pick up of audio events. Uniform frequency response for all angles is a critical requirement for high quality broadband beamforming [18]. To optimize coincidence relative to the horizontal plane, the capsules are stacked vertically, with a spacing of $\leq 30\,\text{mm}$. The capsules are mounted in a double-M/S configuration, with one cardioid capsule $F$ facing $0^\circ$, a second cardioid $R$ facing $180^\circ$, and a bidirectional figure-of-eight capsule $B$ facing $\pm 90^\circ$. Capsule correction filters $H_x$ are applied for further linearization of the signals. This step improves tracking and the beamformer's isotropic frequency response.

### 2.2 Steered Response Power ASL

Prior to ASL processing, a detection filter $H_d$ is applied to the linearized microphone signals. This filter is designed to reject non-speech signals. The corrected and filtered microphone signals can be transformed onto the base of horizontal b-format Ambisonics [14]:

$$W = F_{lin,d} + R_{lin,d} \tag{1}$$

$$X = F_{lin,d} - R_{lin,d} \tag{2}$$

$$Y = B_{lin,d} \tag{3}$$

As W, X and Y represent a two-dimensional, orthonormal basis, any arbitrary first-order microphone pattern $M(\theta, p)$ can be synthesized on the horizontal plane, using the WXY-decoded signals and

$$M(W,X,Y,\theta,p) = pW + (1-p)\left(X\cos\theta + Y\sin\theta\right), \tag{4}$$

with p representing the polar pattern shape and $\theta$ the orientation on the horizontal plane [4, 19, 20]. The factor p can be statically set or dynamically manipulated in a range between 0, which results in the polar pattern of a dipole, and 1, which results in an omnidirectional polar pattern. The most commonly used values for p in this paper are $p = 0.5$, resulting in the unidirectional polar pattern of a virtual cardioid and $p = \frac{1}{3}$, creating a virtual supercardioid.

Using (4) with $p = 0.5$, any number $n_M$ of virtual cardioid microphone signals can be synthesized. The virtual microphone with the highest relative Root Mean Square (RMS) value indicates the Direction of Arrival of the sound source:

$$\theta_{DOA} = \underset{\theta_i}{arg\,max}\left(\overline{M}(W,X,Y,\theta_i)\right), \tag{5}$$

with $\overline{M}$ representing Root Mean Square of M.

## 3 Tracker Stabilization

The described real-time setup uses audio buffers of 256 samples, sampled at 48 kHz. Figure 5a shows the raw directional information $\theta_{DOA}$. The large amount of noise in the angle detection requires additional filtering, since beamformers using $\theta_{DOA}$ as beam orientation perform poorly and produce strong audible artifacts. Filtering is performed using exponential smoothing [21]:

$$\theta_s^t = \alpha\,\theta_{DOA}^t + (1-\alpha)\,\theta_s^{t-1}, \tag{6}$$

with $\theta_{DOA}^t$ and $\theta_s^t$ representing the input and smoothed output angle for time frame t and $\alpha \in (0,1]$ as the reactivity factor. Circular continuity of the angle is ensured within a separate function. If $\alpha$ is set dynamically, a smoothing effect can be achieved that is directly connected to a set of signal characteristics. In the following paragraphs, these factors will be called Confidence Indices (CI) and the smoothing process will be defined as Confidence Weighting. Figure 1 shows three characteristics contributing to two stages of smoothing. The

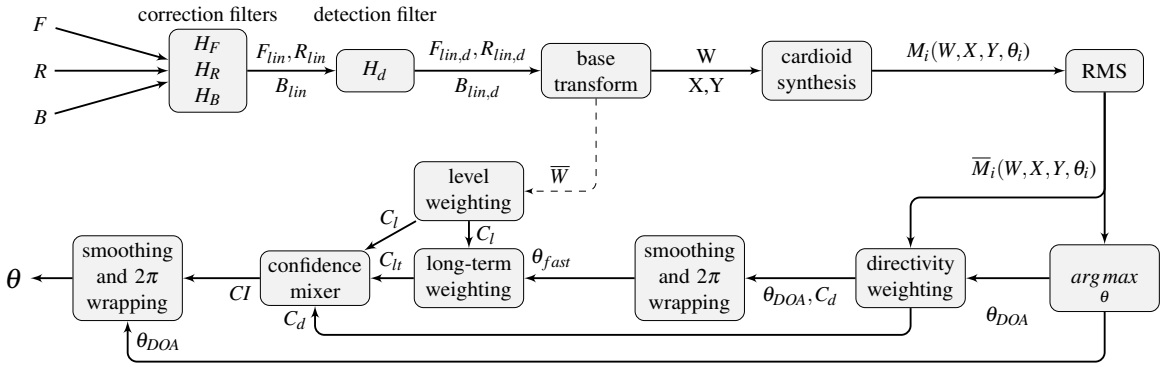


**Fig. 1:** Signal flow through the tracking algorithm. After compensating for nonlinear frequency responses of the microphones and filtering the incoming signals to a range of 200 Hz to 2000 Hz, the synthesis of virtual cardioids over $2\pi$ and RMS-maximization of the signals results in initial DOA estimations. Various weighting algorithms in combination with a variable exponential smoothing process create a more stable, yet reactive tracker.

initial filtering is performed using directivity weighting, a process that analyzes the level of directivity in the detected one-dimensional sound field. The output angle of this process $\theta_{fast}$ is passed on to the two following Confidence Weighting algorithms. A buffer is filled with multiple cycles of $\theta_{fast}$ to compare the detection angle with known source positions, which are dynamically learned and forgotten. Aditionally, a level weighting algorithm compares the omnidirectional level of the current buffer with the average level during speech. The Confidence Weighting processes create a combined CI, which is in turn used for a second filtering operation to compute the final tracker output $\theta$. The algorithms are described in detail in the following sections.

### 3.1 Directivity Weighting

Directivity weighting uses the level of directivity within the recorded sound field as an indicator as to whether a given buffer contains an actual audio event. Figure 2 shows examples of buffers with high directivity (*left*) and low directivity (*right*). The Confidence Index $C_d$ is obtained using the mean distance between the detected sound field, which is normalized, so that $\max(\overline{M}_i) = 1$, and the unidirectional level distribution $U$:

$$U_i = (0.5 + 0.5\cos(\theta_i - \theta_{DOA})), \tag{7}$$

$$C = \frac{1}{n_M}\sum_{i=1}^{n_M}\left(U_i - \overline{M}_i(W,X,Y,\theta_i)\right). \tag{8}$$

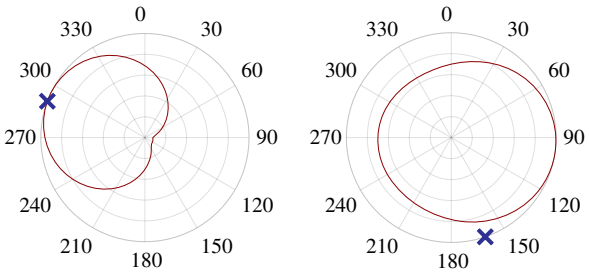


**Fig. 2:** Directivity of two audio buffers. The levels of 360 virtual cardioid microphones arranged with 1° spacing represent the detected, one-dimensional sound field. The closer the level distribution is to the optimal unidirectional distribution $U$, the higher the confidence index $C_d$. *Left*: Buffer with high level of directivity, *Right*: Buffer with low directivity. The marking indicates $\theta_{fast}$ for the displayed buffer. As $C_d$ is large for the buffer on the left, $\theta_{fast} \approx \theta_{DOA}$. For the buffer on the right, a large portion of $\theta_{fast}$ is contributed by $\theta_{fast}$ of the previous buffer, and not by $\theta_{DOA}$.

Scaling to the interval (0,1] is performed with

$$C_d = 10^{(\nu C)}, \tag{9}$$

with $\nu > 0$ representing a parameter controlling the reactivity of the tracker. Considering that for most cases

$$\overline{M}_i \geq U_i, \tag{10}$$

it is clear, that

$$C \leq 0 \text{ and } 0 < C_d \leq 1. \tag{11}$$

Using (6) and setting $\alpha = C_d$, an initial direction of arrival $\theta_{fast}$ can be computed. Figure 5b shows the effect of directivity weighting compared to the raw DOA data shown in Figure 5a.

### 3.2 Level Weighting

Level weighting analyzes the level of the current audio buffer and compares it to a threshold $L$. The signal used for level weighting is $\overline{W}$. The confidence index associated with level weighting $C_l$ interacts directly with long-term weighting, as shown in Figure 1. $C_l$ is computed as

$$C_l = \begin{cases} 1 & \text{for} \quad \overline{W} \geq L \\ 0 & \text{for} \quad \overline{W} < L \end{cases}. \tag{12}$$

### 3.3 Long-Term Weighting

In many acoustic scenarios the speaker positions remain quasi-static. Participants of a meeting mostly stay seated, a driver will remain in the driver's seat, etc. Long-term weighting makes use of this fact by assessing the sound field over a longer period of time. The initial DOA estimation $\theta_{fast}$ is stored in a buffer under the condition that the level confidence index $C_l$ is set to 1. If $C_l \neq 1$, $\theta_{DOA}$ of the previous buffer is used. An average over 50 buffers is passed to the long-term weighting algorithm. The directional information is then classified using a point system. Every incoming angle is quantized with a resolution of 5° and results in a point for the associated bin. The total number of points is limited to 72, resulting in one point per 5° bin in the initial state. For a point to be awarded to the most recent position, a point must be deducted from the least recent DOA. This procedure creates a type of long-term memory for the algorithm. With the parameters presented in this paper, the system "forgets" an audio event after 19.2 s and can adapt to a new static source in 1.5 s to 3 s. Figure 3 shows the point score after processing an excerpt of Scenario I, described in section 4. All five speaker positions listed in Table 1 are clearly discernible. For the determination of the associated confidence index $C_{lt}$, $\theta_{DOA}$ is quantized to 5° and the point distribution is normalized to 1. The relative point value at the quantized angle corresponds directly to $C_{lt}$. This comparison is performed every cycle of the algorithm. If $\theta_{DOA}$ is within $\pm 1°$ of a peak in the long-term angle distribution, an additional confidence bonus is awarded (*snap-to* process).

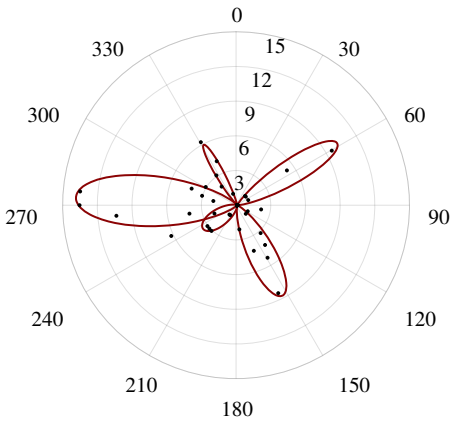


**Fig. 3:** Visualization of long-term confidence. The average detected angle $\theta_{fast}$ over the most current 267 ms, rounded to 5°, results in a point for the associated segment. As the total point count is limited, a point is deducted from the segment with the least recent position detection. With 72 available points, the algorithm learns a static position within 1.5 s to 3 s and forgets an audio event within 19 s. The point score is normalized to 1.

### 3.4 Confidence Mixing

Confidence mixing describes the process of combining all previously described Confidence Indices in the most effective way. Given (9), the final Confidence Index CI can be computed using $C_d, C_l, C_{lt}$ and the mixing parameter $\kappa$:

$$CI = (\kappa C_d + (1-\kappa) C_d C_{lt})\, C_l. \tag{13}$$

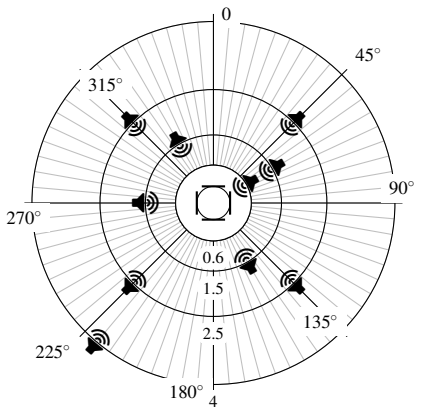


**Fig. 4:** Loudspeaker distribution for the test setup. Ten studio monitors are placed at four distances and varying angles around the microphone array. The configuration is used to record dialog, noise and ambiance for beamformer evaluation.

**Table 1:** Speaker Positions, Scenario I

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Angle | 60 | 150 | 220 | 270 | 330 |
| Distance | 1.5 m | 1.5 m | 4 m | 1.5 m | 1.5 m |
| Sum Duration | 7.1 s | 9.0 s | 4.0 s | 16.0 s | 8.9 s |

**Table 2:** Speaker Positions, Scenario II

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Angle | 60 | 150 | 270 | 330 |
| Distance | 1.5 m | 1.5 m | 1.5 m | 1.5 m |
| Sum Duration | 24.7 s | 1.0 s | 22.3 s | 9.1 s |

## 4 Experimental Setup

Both synthetic data and real recordings were used during development. The subjective results in section 5.2.1 are exclusively presented using real recordings of the setup described in subsection 4.2. Objective results are presented using synthetic and real data.

### 4.1 Synthetic Data

Convolving audio data with appropriate room impulse responses (RIR) can generate accurate simulations of auditory scenes [22]. The synthetic data used for the results in section 5 are generated using [23]. The speech data were randomly selected from the VCTK speech corpus, consisting of short passages read by 109 different speakers [24]. Optional noise interference was selected from the ESC50 corpus, consisting of 2000 recordings of environmental sounds [25]. The acoustic environment was randomly sampled with room geometries ranging from 3 m to 8 m and room heights of 2.5 m to 4 m. Absorption coefficients and RT60 reverberation times were uniformly sampled in different ranges, as described along with the results in table 3.

### 4.2 Virtual Conference

ASL and tracking were evaluated on real recordings using a reproducible multi-channel loudspeaker setup consisting of eight identical Genelec 1029A loudspeakers, arranged on two concentric rings around the microphone array, combined with a far-range and a close-range loudspeaker. The microphone array was constructed using two Schoeps CCM 4V and one Schoeps CCM 8, mounted within a dedicated double-M/S shock mount. The results described in the following sections were gathered using a RME Fireface UFX.

The angular positioning of the virtual sources is shown in Figure 4. The audio material played back was based on [26] and consisted of near-anechoic recordings of male and female speech in German and English, recordings of office and household noise sources such as cell phones, moving chairs, doors, etc. and multi-channel recordings of traffic and construction noise with open and closed windows. Three scenarios were recorded, each with and without interference of background and object noise. The room used for the results of this paper was 8.3 m by 8.2 m, with a total height of 3.8 m. No acoustic treatment or furniture was present, which resulted in an RT60 of 2.31 s, averaged over the 500 Hz and 1000 Hz frequency bands.

### 4.3 Listening Tests

Listening tests were performed to evaluate the impact of different types of signal degradation prior to the tracker design. The results are presented in [27] and were used to prioritize during the development process. In addition, a larger listening test was performed using the tracker output with various beamforming algorithms. A short summary of the listening test can be seen in Table 5, detailed methods and results can

be found in [28]. For the test, 59 test subjects were asked to grade various sound recordings which were recorded using the test setup described in section 4 and processed with the tracking algorithm and a selection of 3-, 2-, and 1-channel beamformers, both commercially available and currently under development.

## 5 Results

The following sections will present both subjective and objective evaluations of the system's performance. Objective results are presented using synthetic and real audio data. It is important to mention that the results are only partially comparable as the synthetic data contains no speech pauses, which prevents error accumulation due to $C_l$-driven static positions at unfavorable angles between speech sections. Additionally, the simulated data cannot make use of the long-term Confidence Weighting as every position is randomly sampled on a Cartesian grid.

Objective error analysis is performed using two connected error metrics, $\theta_{err}$ and $\Delta\theta_{err}$. The angular error $\theta_{err}$ is computed using the circular distance between the reference angles $\theta_r^t$ and the detected angles $\theta^t$, averaged over all time bins t:

$$\theta_{err}^t = \begin{cases} |\theta^t - \theta_r^t| & \text{for} \quad |\theta^t - \theta_r^t| \leq 180^\circ \\ 360 - |\theta^t - \theta_r^t| & \text{for} \quad |\theta^t - \theta_r^t| > 180^\circ \end{cases}. \tag{14}$$

The gradient is calculated using (14) and a two-point calculation:

$$d\theta^t = \frac{\theta^{t+1} - \theta^{t-1}}{2} \tag{15}$$

$$d\theta_r^t = \frac{\theta_r^{t+1} - \theta_r^{t-1}}{2} \tag{16}$$

$$\Delta\theta_{err}^t = \begin{cases} |d\theta^t - d\theta_r^t| & \text{for} \quad |\ \cdot\ | \leq 180 \\ 360 - |d\theta^t - d\theta_r^t| & \text{otherwise} \end{cases} \tag{17}$$

Table 4 shows the mean errors over all $n_T$ time bins:

$$\theta_{err} = \frac{1}{n_T} \sum_{t=1}^{n_T} \theta_{err}^t \tag{18}$$

$$\Delta\theta_{err} = \frac{1}{n_T} \sum_{t=1}^{n_T} \Delta\theta_{err}^t \,. \tag{19}$$

The error calculations shown in Table 4 are performed on reference information which was manually labeled using the session file of the digital audio workstation used for the playback and recording of the test scenarios. The results presented in Table 3 are calculated using the geometric parameters of every individual simulation.

Both $\theta_{err}$ and $\Delta\theta_{err}$ represent important quality metrics for the tracking algorithm. While accurate localization of an acoustic source is important, stable tracking of sources while maintaining high reactivity during change of speakers equally influences the system's real-world usefulness.

Subjective quality assessments are presented using listening tests, performed on recorded audio[1]. The test subjects were asked to grade the recordings with respect to speech intelligibility, noise suppression and subjective quality for German and English test scenarios using a MUSHRA test [29]. Speech intelligibility was additionally analyzed using the Short Time Objective Intelligibility Index proposed in [30]. STOI compares clean speech with processed versions of the same audio. In this case, the clean studio recordings of the speech used for the virtual scenarios were compared to the recorded multi-channel playback.

### 5.1 Simulated Data

| | $RT_{60}$ | SNR | DRR | $\theta_{err}$ | $\Delta\theta_{err}$ |
|---|---|---|---|---|---|
| C | <0.05 s | | 10.98 dB | 3.17° | 0.53° |
| N | <0.05 s | 6.06 dB | | 33.65° | 1.21° |
| C | 0.4 s to 0.6 s | | −7.20 dB | 21.15° | 0.68° |
| N | 0.4 s to 0.6 s | 5.96 dB | | 31.41° | 0.59° |
| C | 0.6 s to 1.5 s | | −9.37 dB | 37.75° | 0.63° |
| N | 0.6 s to 1.5 s | 6.05 dB | | 50.92° | 0.56° |

**Table 3:** ASL performance analysis on synthetic data. Reverberation and additional noise both have strong negative effects on ASL.

ASL performance is evaluated on six one-minute sets of synthetic data, each containing 15 scenes of 4 s. The six sets can be categorized into three subsets, each containing a clean (C) and a noisy (N) simulation of the same scenario. Within the clean sets, only speech and the corresponding reverberation are present, the noisy

[1] Audio examples can be found at zieglerj.home.hdm-stuttgart.de/aslt-companion.html.

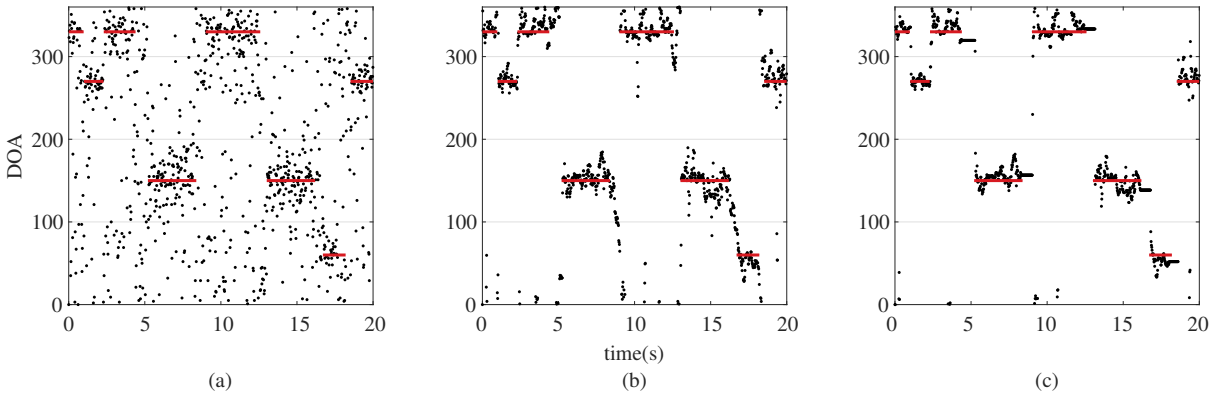


**Fig. 5:** Performance analysis of Confidence Weighting components. (a): Direct DOA estimate $\theta_{DOA}$. (b): Added directivity weighting results in considerably less noise in the output $\theta_{fast}$. (c): Additional level-dependent weighting reduces jumps during pauses. Long-term weighting further improves the accuracy and stability of the tracker output $\theta$. Solid lines represent labeled reference. The values on display were down-sampled by a factor of 4 for increased clarity.

sets contain three additional noise sources randomly placed within the same area as the speaker. The three sets differ in their level of reverberation, which is given as Direct to Reverberant Ratio (DRR) and span RT60 reverberation times from $\approx 0$ ms up to 1500 ms. Within the noisy sets, the signal to noise ratio is given for the virtual omnidirectional microphones.

Table 3 shows the results for the simulated audio data. ASL on clean speech in near-anechoic environments produces a mean error of 3.17° or approximately 1.8 %. An additional noise source within the simulated scenario increases the error by a factor of 10, mild reverberation causes a similar degradation of ASL performance. Furthermore, a clear correlation between the DRR and the localization error can be observed.

### 5.2 Recorded Data

The results shown in Table 4 were created using two different speech scenarios. Scenario I is a 45 s office scene in German, with one female and three different male speakers, located at five positions around the microphone. Scenario II is a 57 s dialog in English, between a female and a male speaker, with additional comments from two less prominently featured positions by the same speakers. Speaker positions and speech duration can be found in Tables 1 and 2[2]. Pauses and overlaps were intentionally added to simulate more realistic conversations. Two versions of each scene were recorded. The first version contains desired speech only. The second version contains interference consisting of office noises, such as cell phones, ripping paper, coffee cups and shifting chairs, being played back at 0.6 m to 1.5 m, whispered side-conversations being played back at 1.5 m and quadraphonic ambient recordings, such as traffic and construction noise, played back on a quadraphonic playback system, positioned at a radius of 2.5 m.

Figure 5 shows a 15 s extract from Scenario I at various steps of the signal processing chain, compared to the reference position. Figure 5a shows $\theta_{DOA}$, the raw output of the virtual cardioid maximization process. Figure 5b shows $\theta_{fast}$, the fast position estimation obtained using the variable exponential smoothing and only one Confidence Index, $C_d$, associated with directivity weighting. It can be seen that this step greatly reduces $\theta_{err}$ and $\Delta\theta_{err}$. For this reason, $\theta_{fast}$ is used throughout the processing chain as a good initial guess for $\theta$. Figure 5c shows the additional improvement realized through the use of the algorithms described in section 3. While the values for $\theta_{err}$ are large, it is worth noting that the calculation is performed over the entire recording. Pauses between words and phrases were not removed during evaluation. This operation would require a subjective threshold of pauses and seemed

[2]Some sources shown in figure 4 only contained interference and ambiance, hence they are not listed in tables 1 and 2.

| | German | | German (noisy) | | English | | English (noisy) | |
|---|---|---|---|---|---|---|---|---|
| | $\theta_{err}$ | $\Delta\theta_{err}$ | $\theta_{err}$ | $\Delta\theta_{err}$ | $\theta_{err}$ | $\Delta\theta_{err}$ | $\theta_{err}$ | $\Delta\theta_{err}$ |
| $\theta_{DOA}$ | 46.18 | 24.67 | 52.59 | 23.86 | 46.24 | 23.19 | 45.44 | 22.73 |
| $+C_d$ | 13.53 | 1.96 | 22.25 | 1.88 | 14.97 | 2.07 | 15.02 | 1.95 |
| $+C_{lt}$ | **11.20** | 0.97 | **17.84** | 0.96 | 11.55 | 1.24 | **12.27** | 1.20 |
| $+C_l$ | 12.08 | **0.84** | 17.90 | **0.93** | **11.54** | **1.23** | 12.30 | **1.19** |
| $+snap$ | 12.41 | 0.85 | 18.00 | **0.93** | 11.58 | **1.23** | 12.29 | **1.19** |
| $\Delta_{SC}$ | −0.08 dB | | −0.16 dB | | −0.07 dB | | −0.08 dB | |

**Table 4:** Tracker Performance Analysis on recorded audio. Adding Confidence Weighting components improves the performance. While $C_d$ globally improves stability and accuracy, other Confidence Indices show a more situation-dependent behavior.

| | omni | beamformers |
|---|---|---|
| Speech Intelligibility | $0.23 \pm 0.12$ | $0.61 \pm 0.16$ |
| Noise Suppression | $0.16 \pm 0.12$ | $0.52 \pm 0.19$ |
| Subjective Quality | $0.19 \pm 0.12$ | $0.65 \pm 0.22$ |

**Table 5:** Listening Test Results. In all categories the beamformed signal is preferred over the omnidirectional baseline. The mean result and pooled standard deviation over all tested beamformers are presented.

arbitrary and situation-dependent[3]. For the calculation of $\Delta\theta_{err}$, the pauses between labeled clips were additionally filled with the last available reference position of the preceding audio clip. This reflects the fact that a passive behavior of the tracker is desired during speech pauses.

#### 5.2.1 Listening Tests

Regardless which beamformer is used, the signal outperforms that of a virtual omnidirectional microphone $F_{lin} + R_{lin}$. Even a simple gradient synthesis beamformer creating a virtual supercardioid facing the tracked direction $\theta$ provides improved intelligibility, noise reduction and subjective quality, compared to the omnidirectional signal. Once confidence weighting is applied, the tracked supercardioid performs without any audible artifacts.

Table 5 shows the summarized results of a listening test performed with 59 test subjects. Possible scores in the categories *Speech Intelligibility*, *Noise Suppression*, and *Subjective Quality* range from zero to one. On average, the use of the tracking algorithm in combination with a beamformer improves *Speech Intelligibility* by 170 %, *Noise Suppression* by 225 % and *Subjective Quality* by 256 %, compared to the signal of a static omnidirectional microphone of equal quality. Detailed results can be found in [28].

#### 5.2.2 Speech Intelligibility

The Short Term Objective Intelligibility was calculated by comparing the dry voice recordings from the test scenarios with the signal of a virtual omnidirectional microphone and of a virtual supercardioid microphone synthesized towards the tracked angle $\theta$, combined with various beamforming algorithms. The use of a virtual omnidirectional microphone results in an average STOI of 0.593, while a virtual, tracked supercardioid produces a STOI of 0.744 and all beamformers used in the test produce a mean STOI of 0.745, a 26 % improvement to the virtual omnidirectional signal.

## 6 Discussion

The overall effect of the various Confidence Indices depends on the application scenario. While directivity weighting universally improves ASL performance, long term smoothing and position snapping improve performance in static environments such as meetings. The results in Table 4 reflect the mean errors in the four described scenarios. The last row of data provides insight into the performance of a first-order supercardioid

[3] Ex: calculating the error for $\theta_{DOA} + C_d$ in Scenario I (German) using only buffers with an rms larger than 20 % of the mean rms of the recording results in an error $\theta_{err}$ of 10.06°. This equals a performance increase of 24.8 % when only examining frames subjectively deemed relevant.

beamformer driven with the tracker output. On average, the level of the target signal deviates by 0.1 dB from the reference value. This is well below the threshold of just-noticeable amplitude difference measured by Zwicker and Fastl for common SPL [31]. The STOI measurements presented in section 5.2.2 show that the objective difference between a static omnidirectional signal and a simple first-order beamformer is significantly larger than the improvement gained by the introduction of more complex beamforming algorithms. This is, in part, due to the focus of STOI. For further investigations, a testing algorithm with stronger focus on high quality audio will be selected.

## 7 Conclusion

The described system for acoustic source localization and tracking provides real-time Direction of Arrival information for coincident beamforming. A set of processing blocks is introduced to provide application-specific improvement over the direct output of energy based scanning methods, resulting in a more accurate and stable DOA-detection. Listening tests show a strong increase in speech intelligibility, noise suppression and subjective quality, when comparing the combination of tracker and beamformer with static microphone signals. When using a simple synthesized supercardioid driven by the tracker, the resulting signal is not subjectively discernible from a signal based on the reference position as input. The algorithm generates artifact-free audio and makes the system suitable for professional audio production applications as well as high-end conferencing and on-set recording.

## Acknowledgments

This research was in part funded by the *Zentrales Innovationsprogramm Mittelstand*, a grant from the *Bundesministerium für Wirtschaft und Energie*.

The authors would like to thank Bernfried Runow for his contribution of beamformer test data.

## References

[1] Reinette, A., Cornejo, M., Rouchon, C., and Fester, M., "Benchmarking Microphone Arrays: ReSpeaker, Conexant, MicroSemi AcuEdge, Matrix Creator, MiniDSP, PlayStation Eye," *Snips Labs*, 2017.

[2] Abhayapala, T. D. and Ward, D. B., "Theory and design of high order sound field microphones using spherical microphone array," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pp. II–1949–II–1952, 2002, doi:10.1109/ICASSP.2002.5745011.

[3] Meyer, J. and Agnello, T., "Spherical microphone array for spatial sound recording," *NEW YORK*, p. 9, 2003.

[4] Meyer, J. and Elko, G. W., "Spherical Microphone Arrays for 3D Sound Recording," in Y. Huang and J. Benesty, editors, *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, pp. 67–89, Springer US, Boston, MA, 2004, ISBN 978-1-4020-7769-2, doi:10.1007/1-4020-7769-6_3.

[5] Li, Z. and Duraiswami, R., "Flexible and Optimal Design of Spherical Microphone Arrays for Beamforming," *IEEE Transactions on Audio, Speech, and Language Processing*, 15(2), pp. 702–714, 2007, ISSN 1558-7924, doi:10.1109/TASL.2006.876764.

[6] Rafaely, B., Koretz, A., Winik, R., and Agmon, M., "Spherical microphone array beampattern design for improved room acoustics analysis," 2008.

[7] Yan, S., Sun, H., Svensson, U., Ma, X., and Hovem, J., "Optimal Modal Beamforming for Spherical Microphone Arrays," *Audio, Speech, and Language Processing, IEEE Transactions on*, 19, pp. 361 – 371, 2011, doi:10.1109/TASL.2010.2047815.

[8] *mh acoustics - product catalog*, 2018.

[9] Wittek, H., Faller, C., Favrot, A., Langen, C., and Tournery, C., "Digitally Enhanced Shotgun Microphone with Increased Directivity," in *Audio Engineering Society Convention 129*, 2010.

[10] Benesty, J., Jingdong, C., and Huang, Y., *Microphone Array Signal Processing*, Springer Berlin Heidelberg, 2008, ISBN 978-3-540-78612-2.

[11] Runow, B. and Curdt, O., "Microphone Arrays for professional audio production," in *28th Tonmeistertagung - VDT International Convention*, p. 7, 2014.

[12] Runow, B., Curdt, O., and Schilling, A., "Shotgun Microphones versus Microphone Arrays," in *29th Tonmeistertagung - VDT International Convention*, 2016.

[13] Eargle, J., *From Mono to Stereo to Surround, A Guide to Microphone Design and Application*, Focal Press : [distributor] Elsevier Books Customer Services, Oxford, 2004, ISBN 978-0-240-51961-6, oCLC: 851974436.

[14] Wittek, H., Haut, C., and Keinath, D., "Double M/S – a Surround recording technique put to test," in *Tonmeistertagung*, Verband Deutscher Tonmeister eV, 2006.

[15] Benjamin, E. and Chen, T., "The Native B-Format Microphone," in *Audio Engineering Society Convention 119*, 2005.

[16] Benjamin, E. and Chen, T., "The Native B-Format Microphone: Part II," in *Audio Engineering Society Convention 120*, 2006.

[17] Jarrett, D. P., Habets, E. A. P., and Naylor, P. A., "3D source localization in the spherical harmonic domain using a pseudointensity vector," in *2010 18th European Signal Processing Conference*, pp. 442–446, 2010.

[18] Veen, B. D. V. and Buckley, K. M., "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine*, 5(2), pp. 4–24, 1988, ISSN 0740-7467, doi:10.1109/53.665.

[19] Gerzon, M. A., "The Design of Precisely Coincident Microphone Arrays for Stereo and Surround Sound," in *Audio Engineering Society Convention 50*, 1975.

[20] Freiberger, K., *Development and Evaluation of Source Localization Algorithms for Coincident Microphone Arrays*, Ph.D. thesis, Institute of Electronic Music and Acoustics (IEM), University of Music and Performing Arts, Graz, Austria, 2010.

[21] Brown, R. G., *Smoothing, forecasting and prediction of discrete time series*, Prentice-Hall Englewood Cliffs, N.J, 1963.

[22] Allen, J. B. and Berkley, D. A., "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, 65(4), pp. 943–950, 1979.

[23] Diaz-Guerra, D., Miguel, A., and Beltran, J. R., "gpuRIR: A Python Library for Room Impulse Response Simulation with GPU Acceleration," *arXiv e-prints*, p. arXiv:1810.11359, 2018.

[24] Veaux, C., Yamagishi, J., MacDonald, K., and others, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2017.

[25] Piczak, K. J., "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, pp. 1015–1018, ACM, New York, NY, USA, 2015, ISBN 978-1-4503-3459-4, doi:10.1145/2733373.2806390, event-place: Brisbane, Australia.

[26] Hirt, R., *Entwicklung einer virtuellen Konferenz unter besonderer Berücksichtigung der Reproduktion von zuvor aufgenommener Sprache - Development of a virtual conference with focus on optimal reproduction of pre recorded speech.*, Bachelor's Thesis, Stuttgart Media University, 2017.

[27] Paukert, H. and Ziegler, J., "Listening Tests in the Process of Microphone Development," in *29. Tonmeistertagung VdT International Convention*, p. 8, 2016.

[28] Paukert, H., Ziegler, J., and Koch, A., "Hörversuche zur Entwicklung eines neuartigen Mehrkapsel-Mikrofons," in *30th Tonmeistertagung VdT International Convention*, p. 8, 2018.

[29] "MUSHRA : Bs. 1534-1. method for the subjective assessment of intermediate sound quality," 2001.

[30] Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J., "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4214–4217, 2010, doi:10.1109/ICASSP.2010.5495701.

[31] Zwicker, E. and Fastl, H., *Psychoacoustics: Facts and Models*, Springer Series in Information Sciences, Springer Berlin Heidelberg, 2013, ISBN 978-3-662-09562-1.