POSITIONAL TRACKING OF A MOVING MICROPHONE IN REVERBERANT SCENES BY APPLYING PERFECT SEQUENCES TO DISTRIBUTED LOUDSPEAKERS

Fabrice Katzberg[1], *Marco Maass*[1], *René Pallenberg*[1], *and Alfred Mertins*[1,2]

[1]Institute for Signal Processing
University of Lübeck
Lübeck, Germany

[2]German Research Center for Artificial Intelligence (DFKI)
AI in Biomedical Signal Processing
Lübeck, Germany

ABSTRACT

Moving microphones allow for the fast acquisition of sound-field data. Such continuous sampling procedures require trajectory knowledge for relating the dynamic samples to a positional context and solving the involved spatio-temporal channel estimation problem. Having one non-uniformly moving microphone, the hardware effort is heavily dominated by additional tracking equipment. In this paper, we present a low-cost tracking method using the microphone signal itself, provided that loudspeakers at surrounding reference points are excited by perfect sequences. As this is often the actual measurement setup, the proposed technique can be seamlessly incorporated into such dynamic sampling procedures. The acoustic tracking system employs a polyphase decomposition and resampling scheme for obtaining low-frequency estimates of the varying impulse responses to the loudspeaker positions and applying multilateration.

Index Terms— Acoustic tracking, perfect sequences, interpolation of time-variant systems, bandwidth adaptation, time of arrival.

1. INTRODUCTION

In closed-room environments, localization and tracking systems are often based on acoustic wave propagation. These approaches usually extract spatial cues to the target object from audio data, e.g., time of arrival (TOA), time difference of arrival, angle of arrival, received signal strength, and feed them to adequate multilateration or multiangulation algorithms [1]. While pure localization algorithms rely on snapshots of static source-sensor setups, tracking algorithms deal with continuously moving targets that allow for smoothing estimates by using observation histories and motion models [2]. However, fast moving objects may lead to estimation biases due to limitations in following the time-varying system dynamics.

There are two general tasks: the localization of sound sources and the localization of microphones. The source localization involves either supervised sound sources where the emitted signals and their timings are controlled or roughly provided by close-distance microphones, or sound sources of widely unknown behavior, e.g., talkers and random acoustic events. The corresponding non-blind and blind source localization problems can be solved by using synchronous measurements from a self-contained microphone array [3], or using decentralized audio data from an ad-hoc wireless acoustic sensor network (WASN) [1]. At this, the microphone positions define the reference points for the multilateration and multiangulation algorithms. The extraction of position related parameters is basically an estimation problem regarding (relative) channel impulse responses that may consist of interfering multipath components due to reverberant surroundings. There are approaches relying on a simplified single-path propagation model, e.g., synchrony-based methods such as the generalized cross-correlation [4] and beamforming-like procedures [5], and the subspace methods MUSIC [6] and ESPRIT [7]. Other methods are based on the explicit formulation of a multichannel system identification problem, which may be solved, e.g., by applying adaptive-filtering techniques [8]. Range estimates for multilateration systems are then obtained from the dominant peaks in the particular impulse responses [9]. Data-driven strategies using supervised learning and neural networks have also been proposed, especially for tasks of audio event localization [10] and direction-of-arrival problems in multi-source scenarios [11]. For the source tracking issue, localization systems often use likelihood-based algorithms with an additional correction stage, e.g., by applying Kalman filters, particle filters, or probability hypothesis density filters [12].

Due to the emitter-receiver reciprocity, mathematical models for source localization can be equivalently used for the microphone localization task. However, for the latter, the spatial variety of accessible signals is often constrained: the number of microphones provided to estimate the position of the target object is strictly limited by the spatial extent of the target itself. Generic applications for microphone localization and tracking are the self-calibration and automatic ranging of nodes in WASNs [1], the spatial calibration of microphone arrays [13], and the (private) positional monitoring of sensor-equipped objects [14]. Usually, the spatial reference points are defined by controlled loudspeakers.

In recent years, increasing demand for high-precision microphone tracking has emerged due to new dynamic measurement techniques [15–24]. These techniques exploit spatio-temporal samples from moving microphones for the spatial reconstruction of acoustic impulse responses (AIRs). The key of such procedures is accurate knowledge of instantaneous sensor locations. For example, angular position data may be constituted by performing a circular trajectory at constant speed and measuring the round-trip time of the moving microphone array. This is often the practical case in dynamic setups acquiring head-related [15, 16], binaural [17], and room impulse responses [18, 19], where the uniformly rotating target array is driven by specific actuation hardware. Regarding non-uniform trajectories, microphone positions can be controlled by a robot or tracked by optical [17], acoustical [20], or inertial [21, 22] measurement systems. However, so far, there is no acoustical solution that provides accurate trajectory data by use of only one microphone channel. In particular, one arbitrarily moving microphone is the basis for the compressed-sensing framework proposed in [24], which allows for the spatial reconstruction of room impulse responses (RIRs) from sub-Nyquist sampling within three-dimensional volumes.

In this paper, we present a simple and effective tracking scheme

This work has been supported by the German Research Foundation under Grant No. ME 1170/10-2.

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

that is capable of providing accurate positional estimates based on the signal of a single moving microphone. The core of our contribution is a polyphase model that involves the bandwidth-adapted reconstruction of time-varying signal components for robust TOA estimations. To employ this model, shifted perfect sequences must be applied to multiple loudspeakers at distributed reference positions. Due to its perfectly flat spectrum, perfect-sequence excitation is often already a part of dynamic strategies for AIR-field acquisition. Thus, the proposed method can be easily incorporated into such continuous procedures, avoiding additional tracking hardware. Regarding this purpose, we consider reverberant environments and may assume clock synchronization, compensated internal delays, known speed of sound, non-real-time handling, existing direct paths (line of sight), and calibrated source positions suited to multilateration techniques.

2. EXCITATION AND SAMPLING MODEL

Assume a closed-room environment with constant atmospheric conditions. For one sound signal $s(t)$ transmitted from a fixed source position, the sound pressure dynamically observed along the trajectory $\tilde{\boldsymbol{x}}(t)$ inside the target volume $\Omega \in \mathbb{R}^3$, $\tilde{\boldsymbol{x}} : \mathbb{R} \to \Omega$, may be described by $\tilde{p}(t) = \int_{-\infty}^{\infty} s(t-\tau)h(t,\tau)\mathrm{d}\tau$, where $t \in \mathbb{R}$ defines the global time variable and $h(t,\tau)$ is the time-variant RIR subject to the relative delay time $\tau \in \mathbb{R}$ [25].

2.1. Perfect Sequences for Single-Channel Excitation

For the single-loudspeaker case, the repetitive use of deterministic L-shift cross-orthogonal sequences $\tilde{s}(n)$ leads to L-periodic excitation signals $s(n) = \tilde{s}(n \bmod L)$ having the perfect autocorrelation $r_{ss}(m) = \sigma_s^2\, \delta(m \bmod L)$, with σ_s^2 being the signal power and $\delta(m)$ denoting the unit impulse function. Since all L circularly shifted versions of $\tilde{s}(n)$ are orthogonal to each other, the system identification task, i.e. deconvolution, can be simplified to a pure correlation task exploiting the property

$$s(n) * \tilde{s}(-n) = L\sigma_s^2 \sum_{r=-\infty}^{\infty} \delta(n - rL). \tag{1}$$

The corresponding inverse convolution approach using perfect-sequence excitation (PSQE) is a common technique for stationary sound-field measurements [26, 27].

For the identification of rapidly time-varying systems by adaptive filtering, the normalized least-mean square algorithm (NLMS) achieves optimal tracking ability in PSQE situations [28, 29]. In fact, having a time-invariant system, the NLMS is equivalent to the inverse convolution technique based on (1) and reaches convergence after L iterations for noiseless steady-state conditions and unit step size [27]. Moreover, a system representation subject to the orthogonal basis formed by $\tilde{s}(n)$ may be used for an efficient NLMS variant, where each iteration reduces to a single assignment operation of the particular expansion coefficient, i.e., each coefficient is updated/replaced once per period [27, 30]. A similar technique has been proposed in [18] for the continuous measurement of (binaural) RIRs along a circle using PSQE. Here, the knowledge of positional data is exploited for embedding the time-varying expansion coefficients into the spatio-temporal context of the time-invariant environment. At this, the efficient NLMS algorithm is interpreted as nearest-neighbor-like interpolation along the trajectory, where its inherent online mode leads to a delayed rectangular synthesis function [19].

2.2. Considered Sampling Problem

Let us consider a fixed set of N spatially distributed loudspeakers. The i-th loudspeaker is fed by the source sequence $s_i(n)$ with $n \in \mathbb{N}_0$ being the discrete variable of the causal time signals. Supposing linearity, the measurement along the trajectory $\tilde{\boldsymbol{x}}(t)$ by use of one moving microphone provides the sampled signal

$$\tilde{p}(n) = \sum_{i=1}^{N} \sum_{m=0}^{L-1} s_i(n-m)h_i(n,m) + \eta(n), \tag{2}$$

with $\eta(n)$ denoting the measurement noise and $h_i(n,m)$ being the sampled time-variant RIR from the respective loudspeaker to the current sensor position. Here, the amplitudes of the RIRs disappear into the noise floor for sampled delays $m \geq L$. The microphone measures at uniform points $t_n = n/f_\text{s}$, where $f_\text{s} \geq 2(f_\text{cut} + \alpha f_\text{cut})$ is the sampling rate, f_cut is the cutoff frequency of the analog anti-aliasing prefilter, and α is the maximum shifting factor due to the Doppler effect [23, 25].

2.3. Sampling Model for Controlled Multichannel PSQE

As the microphone is non-stop moving, the estimation of $h_i(n,m)$ from (2) leads to an ill-posed problem. For solving this problem as best as possible, controlled PSQE can be used to apply signals $s_i(n)$ to the loudspeakers that achieve both the perfect autocorrelation property (1) and zero cross-correlation to each other [29]. Using the vector notations $\boldsymbol{s}_i(n) = [s_i(n), \dots, s_i(n-L+1)]^T$, $\boldsymbol{s}(n) = [\boldsymbol{s}_1^T(n), \dots, \boldsymbol{s}_N^T(n)]^T$, $\boldsymbol{h}_i(n) = [h_i(n,0), \dots, h_i(n, L-1)]^T$, and $\boldsymbol{h}(n) = [\boldsymbol{h}_1^T(n), \dots, \boldsymbol{h}_N^T(n)]^T$, the samples can be represented subject to the pooled time-variant impulse response $\boldsymbol{h}(n) \in \mathbb{R}^{LN}$,

$$\tilde{p}(n) = \sum_{i=1}^{N} \boldsymbol{s}_i^T(n)\, \boldsymbol{h}_i(n) + \eta(n) = \boldsymbol{s}^T(n)\, \boldsymbol{h}(n) + \eta(n). \tag{3}$$

By choosing a perfect sequence $\tilde{s}(n)$ of length $\mathcal{L} = LN$ for setting up the $\mathcal{L}$-periodic excitation sequence $s(n)$, phase-shifted loudspeaker excitations $s_i(n) = s(n + (i-1)L)$ allow for reformulating the steady-state sampling model for the noise-free case according to

$$\tilde{p}(n) = \boldsymbol{s}^T(n)\, \boldsymbol{S}^T \boldsymbol{S} \gamma^{-1} \boldsymbol{h}(n) = \boldsymbol{s}^T(n)\, \boldsymbol{S}^T \gamma^{-1} \boldsymbol{c}(n) = c_q(n), \tag{4}$$

where the columns of $\boldsymbol{S} \in \mathbb{R}^{\mathcal{L} \times \mathcal{L}}$ are built up by all $\mathcal{L}$ circularly delayed versions of $\tilde{s}(n)$ and satisfy the orthogonality $\boldsymbol{S}^T \boldsymbol{S} = \gamma \boldsymbol{I}$ with $\gamma = \mathcal{L}\sigma_s^2$ and $\boldsymbol{I}$ being the identity matrix. Correspondingly, $\boldsymbol{s}^T(n)\, \boldsymbol{S}^T$ yields a row vector containing only zeros except for the value γ at the q-th element with $q = n \bmod \mathcal{L} + 1$. Thus, the instantaneous sample $\tilde{p}(n)$ acquired by the moving microphone at the particular time point n can simply be assigned to the q-th time-varying expansion coefficient $c_q(n)$ in the vector $\boldsymbol{c}(n) \in \mathbb{R}^{\mathcal{L}}$.

3. BANDWIDTH-ADAPTED RECONSTRUCTION OF TIME-VARYING EXPANSION COEFFICIENTS

The full coefficient vector $\boldsymbol{c}(n) = [c_1(n), \dots, c_{\mathcal{L}}(n)]^T$, including its out-of-phase elements $c_k(n)$ (OPEs) at indices $k \neq q$, decodes the time-varying RIRs according to

$$\boldsymbol{h}(n) = \gamma^{-1} \boldsymbol{S}^T \boldsymbol{c}(n). \tag{5}$$

By following an NLMS-based online strategy, only outdated information would be available for determining the OPEs as the microphone is continuously moving and the current sample $\tilde{p}(n)$ covers $c_q(n)$ exclusively. Dropping real-time requirements, the method in [18] uses the already known trajectory to reconstruct these coefficients within a spatial context. In contrast to that, we propose a scheme that interpolates the OPEs explicitly along the global time index n, in order to specify the positional connections a posteriori. For reducing aliasing artifacts, this procedure requires a preprocessed bandwidth adaptation in practice.

3.1. Interpolation of Out-Of-Phase Coefficients

According to the sampling model (4), the ℓ-th expansion coefficient, $\ell \in \{1, \dots, \mathcal{L}\}$, is sampled along the time-varying dimension at equidistant time points $n_\ell \in \{n \,|\, n = (\ell-1) + r\mathcal{L}, r \in \mathbb{N}_0\}$. Thus, the underlying PSQE procedure can be interpreted as the decomposition of the microphone signal $\tilde{p}(n)$ sampled at f_s into $\mathcal{L}$ polyphase signals $c_\ell(n_\ell)$ actually sampled at $f_\text{s}/\mathcal{L}$. Our aim is to resample the polyphase signals to the original rate f_s. For this purpose, let us introduce the mapping $r_\ell(n) = (n - \ell + 1)/\mathcal{L}$, where integer values of the variable $r_\ell \in \mathbb{R}$ define the sampling positions of the ℓ-th expansion coefficient. Then, the reconstruction of the expansion coefficients in the time-varying dimension can be expressed as

$$c_\ell(n) = \sum_{r \in \mathbb{N}_0} \phi(r_\ell(n) - r)\, \tilde{p}((\ell-1) + r\mathcal{L}), \qquad (6)$$

where $\phi(r_\ell(n) - r)$ is a fractional delay filter approximating the sinc function. For $\ell = q$, the interpolation (6) reduces to $c_\ell(n) = \tilde{p}(n)$.

3.2. Aliasing in the Time-Varying Dimension

For rating the performance of the bandlimited OPE interpolation (6), an upper bound of the signal bandwidth in the time-varying dimension needs to be provided. Due to the constant environment, i.e., fixed temperature, static scatterers, etc., the target volume Ω contains a time-invariant configuration of scattering paths. Thus, the time-variance of the transmission system is actually a space-variance, fully determined by the spatial variations of the moving microphone inside Ω. Accordingly, sampling and reconstruction of the time-varying coefficients $c_\ell(n_\ell)$ is a problem that may be transferred to a spatial worst-case correspondence. First, let us bound the highest temporal frequency passing through the prefilter of the microphone according to $f_\text{max} \leq f_\text{cut}(1 + v_\text{max}/c)$, which considers the Doppler-shifted case where the microphone is moving at maximum speed v_max along the direction of the sound wave propagating at speed c. Since the sound signal is bandlimited in the temporal dimension, it is also bandlimited in the spatial dimension [31]. For the worst case of maximum spatial variations, which occurs when the microphone is constantly moving at v_max along a straight line, the received maximum wave number is $\kappa_\text{max} = f_\text{max}/c$ [31]. At this, the sampling rate $f_\text{s}/\mathcal{L}$ of the polyphase signals $c_\ell(n_\ell)$ corresponds to a maximum spatial sampling interval of $\Delta_\text{max} = \mathcal{L}v_\text{max}/f_\text{s}$. In order to allow for an aliasing-free OPE reconstruction according to the Nyquist-Shannon sampling theorem, i.e., $\Delta_\text{max} \leq (2\kappa_\text{max})^{-1}$, this worst-case scenario requires an audio bandwidth limited to

$$f_\text{c} \leq \frac{c f_\text{s}}{2\mathcal{L} v_\text{max}(1 + v_\text{max}/c)}. \qquad (7)$$

3.3. Bandwidth Adaptation for Robust TOA Extraction

Due to its flat magnitude spectrum, PSQE generally permits broadband measurements for the system identification task, e.g., for the estimation of RIRs having wide frequency ranges. However, within the tracking procedure, our only aim is to extract TOAs from the first dominant peaks in approximated time-varying RIRs. For that, the cutoff frequency f_c is considered as a tuning parameter that allows for reducing aliasing artifacts in the interpolation scheme (6) at the cost of a less sharp peak localization in the early time dimension of the resulting RIRs. For finding a suitable tradeoff between tolerable corruption by aliasing and sufficient sharpness of the direct peak, the worst-case aliasing bound (7) may be employed with reference to the maximum expected microphone speed v_max. Similar to the multiresolution strategy in [23], we can tune the bandwidth by applying a digital low-pass filter $g_{f_\text{c}}(n)$ with cutoff f_c to the dynamic microphone signal $\tilde{p}(n)$, leading to the bandwidth-adapted model

$$\mathcal{D}\{g_{f_\text{c}}(n) * \tilde{p}(n)\} = \tilde{p}'_{f_\text{c}}(n) = \boldsymbol{s}^T(n)\, \boldsymbol{h}'(n) = c'_q(n), \qquad (8)$$

where $\mathcal{D}$ compensates for the delay introduced by the filter and the superscript $(\cdot)'$ denotes the low-pass filtered signal equivalents. Since the microphone signal is dynamically shaped by the trajectory function, i.e., $\tilde{p}(\tilde{\boldsymbol{x}}(n))$, this actually corresponds to a spatio-temporal filtering. As the time-varying RIRs live on the same trajectory, i.e, $\boldsymbol{h}(\tilde{\boldsymbol{x}}(n))$, they are equally affected by this filtering operation.

Altogether, the proposed tracking scheme based on the microphone signal $\tilde{p}(n)$ can be condensed into the following simple steps:

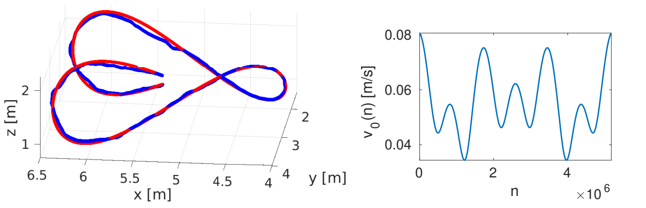
1) Calculation of $\tilde{p}'_{f_\text{c}}(n)$ for a chosen tradeoff f_c.
2) Interpolation of $c'_\ell(n)$ according to (6) for setting up $\boldsymbol{c}'(n)$.
3) Recovery of $\boldsymbol{h}'_i(n)$ using $\boldsymbol{c}'(n)$ and (5).
4) TOA estimations $\mathcal{T}_i(n)$ from the relevant peaks in $\boldsymbol{h}'_i(n)$.
5) Multilateration of $\tilde{\boldsymbol{x}}(n)$ using $\mathcal{T}_1(n), \dots, \mathcal{T}_N(n)$.

The key of this procedure is formed by the first three steps, yielding current RIR estimates (in low frequencies) that allow for TOA estimates being robust against spatial variations induced by the microphone velocity. The last steps can be achieved by standard peak-detection [9] and multilateration algorithms [1].

Finally, with reference to (7), it should be mentioned that the period length $\mathcal{L} = LN$ is another important parameter that could be adjusted to improve the OPE reconstruction along the time-varying dimension for a given bandwidth. For that, the considered RIR length L could be reduced. This increases undesired wrap-around artifacts (time aliasing) for $L < T_{60} f_\text{s}$, with T_{60} being the reverberation time. Also, the number N of involved loudspeakers could be reduced. However, for range-based methods localizing in D dimensions, at least $N = D + 1$ spatial reference points are required to solve the multilateration problem unambiguously.

4. EXPERIMENTS

For proving our concept, we first present data from numerical experiments based on the image source method [32]. We simulated dynamic measurements along the trajectory depicted in Fig. 1(a), considering PSQE by four randomly distributed loudspeakers in a box-shaped room of dimensions $6.94\,\text{m} \times 5.22\text{m} \times 2.61\text{m}$ with different reverberation times $T_{60} \in \{0.15, 0.3\}\,\text{s}$. Various microphone speeds were tested, using the slowest velocity profile shown in Fig. 1(b) and its accelerated versions $v_p(n) = 2^p v_0(2^p n)$ for $p = \{1, 2, 3\}$. The recordings were simulated at $f_\text{s} = 24\,\text{kHz}$ and corrupted by additive white Gaussian noise with a signal-to-noise ratio (SNR) of $40\,\text{dB}$. The delay dimension of the RIRs was limited to $L = \lceil T_{60} f_\text{s} \rceil$ taps. For the PSQE, we used maximum-length sequences with period length $\mathcal{L} = 4L$. The average positional errors (in centimeter) achieved by the proposed tracking method are summarized in Table 1 for the different simulated scenarios and different



(a) (b)

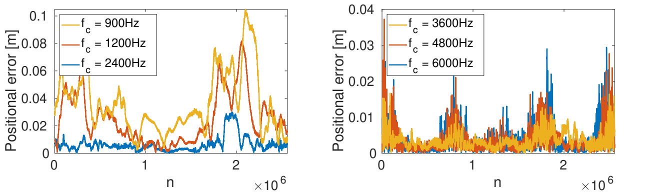
Fig. 1. Tracking scenario. (a) True trajectory (red) and tracked positions (blue) for the velocity profile $v_1(n) = 2v_0(2n)$ in the scenario $T_{60} = 0.3\,\mathrm{s}$ choosing $f_\mathrm{c} = 1200\,\mathrm{Hz}$. (b) Velocity profile $v_0(n)$.

Table 1. Mean (median) values $[\mathrm{m}]^{-2}$ of the tracking errors for different reverberation times T_{60}, velocities $v_p(n)$, and choices of f_c.

	Velocity profile			
Scenario	$v_0(n)$	$v_1(n)$	$v_2(n)$	$v_3(n)$
$\mathbf{T_{60} = 0.15\,s}$				
$f_\mathrm{c} = 900\,\mathrm{Hz}$	2.90 (3.55)	3.60 (2.94)	3.78 (3.16)	5.85 (4.22)
$f_\mathrm{c} = 1200\,\mathrm{Hz}$	2.44 (1.95)	2.49 (2.02)	2.72 (2.29)	6.35 (3.85)
$f_\mathrm{c} = 2400\,\mathrm{Hz}$	0.60 (0.44)	0.64 (0.51)	1.74 (0.88)	46.5 (5.37)
$f_\mathrm{c} = 3600\,\mathrm{Hz}$	0.25 (0.19)	0.35 (0.30)	3.82 (0.82)	94.5 (13.5)
$f_\mathrm{c} = 4800\,\mathrm{Hz}$	0.15 (0.12)	0.40 (0.27)	10.3 (1.15)	128 (81.1)
$f_\mathrm{c} = 6000\,\mathrm{Hz}$	0.11 (0.09)	0.41 (0.28)	24.1 (1.44)	171 (130)
$\mathbf{T_{60} = 0.30\,s}$				
$f_\mathrm{c} = 900\,\mathrm{Hz}$	5.68 (4.26)	6.41 (4.65)	17.2 (7.08)	128 (114)
$f_\mathrm{c} = 1200\,\mathrm{Hz}$	3.75 (3.07)	3.90 (3.16)	27.2 (6.78)	154 (129)
$f_\mathrm{c} = 2400\,\mathrm{Hz}$	0.99 (0.70)	7.84 (1.11)	89.4 (13.4)	291 (259)
$f_\mathrm{c} = 3600\,\mathrm{Hz}$	0.46 (0.37)	31.6 (1.16)	160 (107)	340 (328)
$f_\mathrm{c} = 4800\,\mathrm{Hz}$	3.09 (0.31)	45.7 (1.48)	241 (184)	361 (336)
$f_\mathrm{c} = 6000\,\mathrm{Hz}$	12.1 (0.31)	78.6 (2.32)	269 (216)	379 (345)

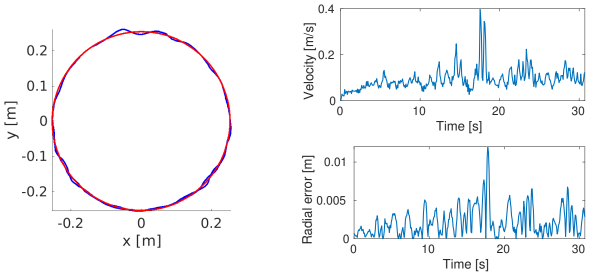
choices of the tuning parameter f_c. In each case, the signal filtering subject to f_c was accomplished by a Hamming windowed low-pass filter of order 1000 and the reconstruction according to (6) was performed by a simple linear interpolator. For the TOA estimations, we resampled the recovered RIRs by the factor 32 and determined each index of the earliest significant maximum. For solving the actual localization problem, we recast the non-linear multilateration equations to an unconstrained linear least-squares problem and used its closed-form solution [1].

Considering the room scenario $T_{60} = 0.15$ and the velocity profiles $v_0(n)$, $v_1(n)$, $v_2(n)$, $v_3(n)$, the (worst-case) aliasing bounds according to (7) are 3600 Hz, 1800 Hz, 900 Hz, and 450 Hz, respectively. In these scenarios, our tracking algorithm achieves robust positional accuracy in the range of a few centimeters for the lowest tested cutoffs $f_\mathrm{c} = 900\,\mathrm{Hz}$ and $f_\mathrm{c} = 1200\,\mathrm{Hz}$. Larger cutoffs lead to sharper peak resolutions in the estimated RIRs and high-precision tracking for $v_0(n)$ and $v_1(n)$ with small errors in the millimeter range. Generally, from a certain point where f_c exceeds the bound (7) by an excessive margin, errors due to aliasing artifacts become dominant. This can be observed, e.g., for $v_1(n)$, $f_\mathrm{c} > 3600\,\mathrm{Hz}$ and for $v_2(n)$, $f_\mathrm{c} > 2400\,\mathrm{Hz}$. For $v_1(n)$, the instantaneous tracking errors subject to f_c are presented in Fig. 2. For $v_2(n)$, $f_\mathrm{c} > 2400\,\mathrm{Hz}$, high-velocity segments yield high error outliers and, thus, increased mean tracking errors in Table 1, while the median error values remain reasonably low. Similar dependencies on f_c can be observed from the tracking results for $T_{60} = 0.3\,\mathrm{s}$. Here, the errors are generally higher. Due to the doubled reverberation time, the accordingly doubled RIR length halves the aliasing bound (2) for $v_0(n)$, $v_1(n)$, $v_2(n)$, $v_3(n)$ to 1800 Hz, 900 Hz, 450 Hz, and 225 Hz, respectively.



(a) (b)

Fig. 2. Instantaneous tracking errors at $T_{60} = 0.15\,\mathrm{s}$ and $v_1(n)$ for selections of (a) lower and (b) higher cutoff frequencies f_c.



(a) (b)

Fig. 3. Tracking results for the planar case. (a) Ideal (red) and tracked (blue) circular trajectory of the microphone. (b) Recovered velocity profile (top) and the corresponding radial mismatch (bottom).

Finally, we present the tracking results from a simple real-world experiment. For this, we used a MOTU 8M audio interface, three K+H M52 loudspeakers, and one microphone of the type beyerdynamic MM1. In an office room of size $4.02\,\mathrm{m} \times 4.64\,\mathrm{m} \times 3\,\mathrm{m}$ and reverberation time $T_{60} = 0.34\,\mathrm{s}$, we calibrated the three distributed loudspeaker positions by measuring their mutual distances and applying multidimensional scaling. Within the convex hull of the loudspeaker points, we mounted the microphone on a stand of the type K&M 210/9 and performed a circular trajectory of radius $0.253\,\mathrm{m}$ by freely rotating its boom arm. Simultaneously, the microphone signal was sampled at 48 kHz and PSQE was applied to the loudspeakers using a perfect sweep of length $\mathcal{L} = 3L$ with $L = 2^{14}$. In this setting, the SNR due to ambient noise is about 30 dB. For the tracking scheme, we used a Hamming windowed low-pass filter of order 4000 with cutoff $f_\mathrm{c} = 1500\,\mathrm{Hz}$, a linear interpolator, and a resampling factor of 16. The resulting positional estimates for 1.5 round trips of the microphone are depicted in Fig. 3(a). The recovered velocity profile is presented at the top of Fig. 3(b), leading to a mean velocity of about $0.08\,\mathrm{m/s}$. The radial error compared to the ideal circle is shown below. The mean radial mismatch is only $0.002\,\mathrm{m}$. The maximum radial error is $0.012\,\mathrm{m}$ and coincides with the maximum microphone speed.

5. CONCLUSIONS

In this paper, we presented a simple and effective tracking procedure for localizing the instantaneous positions of one continuously moving microphone. The method exploits the cross-orthogonality of perfect-sequence excitation for representing the involved RIRs by time-variant expansion coefficients. Using these coefficients, a bandwidth-adapted interpolation scheme allows for robust RIR estimates at low frequencies and reliable TOA extractions that can be used to solve the multilateration problem with very high accuracy.

6. REFERENCES

[1] M. Cobos, F. Antonacci, A. Alexandridis, A. Mouchtaris, and B. Lee, "A survey of sound source localization methods in wireless acoustic sensor networks," *Wirel. Commun. Mob. Comput.*, vol. 2017, Article ID 3956282, 2017.

[2] F. Gustafsson and F. Gunnarsson, "Mobile positioning using wireless networks: Possibilities and fundamental limitations based on available wireless network measurements," *IEEE Signal Process. Mag.*, vol. 22, no. 4, pp. 24–40, 2005.

[3] C. Evers, H. W. Löllmann, H. Mellmann, A. Schmidt, H. Barfuss, P. A. Naylor, and W. Kellermann, "The LOCATA challenge: Acoustic source localization and tracking," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 1620–1643, 2020.

[4] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 3, pp. 288–292, 1997.

[5] C. Rascon and I. Meza, "Localization of sound sources in robotics: A review," *Robot. Auton. Syst.*, vol. 96, pp. 184–210, 2017.

[6] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[7] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 984–995, 1989.

[8] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *J. Acoust. Soc. Am.*, vol. 107, no. 1, pp. 384–391, 2000.

[9] G. Defrance, L. Daudet, and J.-D. Polack, "Finding the onset of a room impulse response: Straightforward?," *J. Acoust. Soc. Am.*, vol. 124, no. 4, EL248-EL254, 2008.

[10] H. Phan, L. Pham, P. Koch, N. Duong, McLoughlin, and A. Mertins, "On multitask loss function for audio event detection and localization," in *Proc. Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, 2020, pp. 160–164.

[11] N. Ma, J. Gonzalez, and G. Brown, "Robust binaural localization of a target sound source by combining spectral source models and deep neural networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 11, pp. 2122–2131, 2018.

[12] S. Challa, M. Morelande, D. Mušicki, and R. Evans, *Fundamentals of Object Tracking*, Cambridge University Press, 2011.

[13] J. M. Sachar, H. F. Silverman, and W. R. Patterson, "Microphone position and gain calibration for a large-aperture microphone array," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 1, pp. 42–52, 2005.

[14] X. Cheng, H. Shu, Q. Liang, and D. H.-C. Du, "Silent positioning in underwater acoustic sensor networks," *IEEE Trans. Veh. Technol.*, vol. 57, no. 3, pp. 1756–1766, 2008.

[15] G. Enzner, "3D-continuous-azimuth acquisition of head-related impulse responses using multi-channel adaptive filtering," in *Proc. IEEE Workshop Applications of Signal Process. to Audio and Acoustics*, 2009, pp. 325–328.

[16] C. Urbanietz and G. Enzner, "Direct spatial-fourier regression of HRIRs from multi-elevation continuous-azimuth recordings," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, 2020.

[17] N. Hahn, W. Hahne, and S. Spors, "Dynamic measurement of binaural room impulse responses using an optical tracking system," in *Proc. Int. Conf. Spatial Audio*, 2017, pp. 16–21.

[18] N. Hahn and S. Spors, "Simultaneous measurement of spatial room impulse responses from multiple sound sources using a continuous moving microphone," in *Proc. Europ. Signal Process. Conf.*, 2018, pp. 2194–2198.

[19] N. Hahn and S. Spors, "Comparison of continuous measurement techniques for spatial room impulse responses," in *Proc. Europ. Signal Process. Conf.*, 2016, pp. 1638–1642.

[20] S. Nagel, T. Kabzinski, S. Kühl, C. Antweiler, and P. Jax, "Acoustic head-tracking for acquisition of head-related transfer functions with unconstrained subject movement," in *Proc. AES Int. Conf. Audio for Virtual and Augmented Reality*, 2018.

[21] S. Li and J. Peissig, "Fast estimation of 2D individual HRTFs with arbitrary head movements," in *Proc. Int. Conf. Digit. Signal Process.*, 2017.

[22] J. He, R. Ranjan, W.-S. Gan, N. K. Chaudhary, N. D. Hai, and R. Gupta, "Fast continuous measurement of HRTFs with unconstrained head movements for 3D audio," *J. Audio Eng. Soc.*, vol. 66, no. 1, pp. 884–900, 2018.

[23] F. Katzberg, R. Mazur, M. Maass, P. Koch, and A. Mertins, "Sound-field measurement with moving microphones," *J. Acoust. Soc. Am.*, vol. 141, no. 5, pp. 3220–3235, 2017.

[24] F. Katzberg, R. Mazur, M. Maass, P. Koch, and A. Mertins, "A compressed sensing framework for dynamic sound-field measurements," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 11, pp. 1962–1975, 2018.

[25] F. Hlawatsch and G. Matz, Eds., *Wireless Communications Over Rapidly Time-Varying Channels*, Academic Press, 2011.

[26] G.-B. Stan, J.-J. Embrechts, and D. Archambeau, "Comparison of different impulse response measurement techniques," *J. Audio Eng. Soc.*, vol. 50, no. 4, pp. 249–262, 2002.

[27] C. Antweiler, S. Kühl, B. Sauert, and P. Vary, "System identification with perfect sequence excitation – Efficient NLMS vs. inverse cyclic convolution," in *Proc. ITG Conf. Speech Communication*, 2014.

[28] C. Antweiler and G. Enzner, "Perfect sequence LMS for rapid acquisition of continuous-azimuth head related impulse responses," in *Proc. IEEE Workshop Applications of Signal Process. to Audio and Acoustics*, 2009, pp. 281–284.

[29] C. Antweiler, A. Telle, and P. Vary, "NLMS-type system identification of MISO systems with shifted perfect sequences," in *Proc. Int. Workshop Acoustic Echo and Noise Control*, 2008.

[30] A. Carini, "Efficient NLMS and RLS algorithms for perfect and imperfect periodic sequences," *IEEE Trans. Signal Process.*, vol. 85, no. 4, pp. 2048–2059, 2010.

[31] T. Ajdler, L. Sbaiz, and M. Vetterli, "The plenacoustic function and its sampling," *IEEE Trans. Signal Process.*, vol. 54, no. 10, pp. 3790–3804, 2006.

[32] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.