KU LEUVEN

STADIUS
Center for Dynamical Systems, Signal Processing and Data Analytics

| | |
|---|---|
| **Citation/Reference** | Authors: Kathleen MacWilliam, Thomas Dietzen, Toon van Waterschoot<br>Title: Tracking of Spatially Dynamic Room Impulse Responses Along Locally Linearized Trajectories<br>Conference: Forum Acusticum Euronoise<br>Date of conference: June 23-26, 2025 |
| **Archived version** | ☐ Submitted version<br>■ Accepted (Author manuscript: the content is identical to the content of the published paper, but without the final typesetting by the publisher)<br>☐ Published<br>☐ Published (Open Access) |
| **Acknowledgements** | In this manuscript, the following funding is acknowledged:<br>■ EU H22020 ☐ EU Horizon Europe ■ FWO ■ KULeuven Internal Funds ☐ VLAIO ☐ AI ☐ Other |
| **Published version** | `https://arxiv.org/pdf/2506.11703` |
| **Copyright statement** | |
| **Author contact** | kathleen.macwilliam@esat.kuleuven.be |



FORUM ACUSTICUM EURONOISE 2025

# TRACKING OF SPATIALLY DYNAMIC ROOM IMPULSE RESPONSES ALONG LOCALLY LINEARIZED TRAJECTORIES

**Kathleen MacWilliam**[1*] **Thomas Dietzen**[2] **Toon van Waterschoot**[1]

[1] Department of Electrical Engineering (ESAT-STADIUS), KU Leuven, Belgium
[2] Department of Electrical Engineering (ESAT-PSI), EAVISE, KU Leuven, Belgium

## ABSTRACT

Measuring room impulse responses (RIRs) at multiple spatial points is a time-consuming task, while simulations require detailed knowledge of the room's acoustic environment. In prior work, we proposed a method for estimating the early part of RIRs along a linear trajectory in a time-varying acoustic scenario involving a static sound source and a microphone moving at constant velocity. This approach relies on measured RIRs at the start and end points of the trajectory and assumes that the time intervals occupied by the direct sound and individual reflections along the trajectory are non-overlapping. The method's applicability is therefore restricted to relatively small areas within a room, and its performance has yet to be validated with real-world data. In this paper, we propose a practical extension of the method to more realistic scenarios by segmenting longer trajectories into smaller linear intervals where the assumptions approximately hold. Applying the method piecewise along these segments extends its applicability to more complex room environments. We demonstrate its effectiveness using the trajectoRIR database, which includes moving microphone recordings and RIR measurements at discrete points along a controlled L-shaped trajectory in a real room.

**Keywords:** *room impulse response estimation, time-varying systems, dynamic time warping*

## 1. INTRODUCTION

Knowledge of room impulse responses (RIRs) is an important aspect of room acoustics, with applications ranging from spatial audio rendering to speech enhancement and dereverberation [1, 2]. However, measuring RIRs at multiple spatial positions is time-consuming, and while computational simulations offer an alternative, they typically require detailed knowledge of the room's geometry and material properties, which is often unavailable or difficult to obtain. A practical alternative is to estimate RIRs along a trajectory using a moving microphone, allowing for efficient spatial sampling of the acoustic field.

In our previous work [3], we introduced a method for estimating the early part of RIRs along a linear trajectory in a time-varying acoustic scenario, involving a static source and a moving microphone. This can be framed as a time-varying system identification problem, where the system to be identified is referred to as a time-variant RIR. Here, we define a time-variant RIR as the linear time-invariant (LTI) system relating the source signal to the microphone signal if the microphone location is time-variant.

Time-variant RIRs have been estimated using fully data-driven methods, such as normalized least mean squares (NLMS) algorithms [4, 5], and Kalman filtering approaches [6]. In contrast, methods that operate with limited data often incorporate prior knowledge through an underlying room acoustic model. If the time-variant RIR is considered as a sequence of RIRs at discrete spatial positions, the estimation problem can be viewed as one of interpolation, a topic that has been explored in numerous studies [7–10]. Most relevant to our scenario is the task of interpolating an RIR at a location between two measured positions, given RIRs from a common static source. This has been addressed using techniques such as Dynamic Time Warping (DTW), which temporally aligns early reflections prior to linear interpolation [11, 12], and optimal transport-based methods, which allow non-bijective matching of reflection components [13].

In our prior work, we integrated RIR interpolation, derived from a room acoustic model, into a state-space framework. This approach combines physical modeling with data-driven estimation by using a first-order difference equation to model the evolution of early reflections over time. Unlike other approaches that use a state transition factor [6, 14], we incorporated interpolation into the state transition matrix between consecutive early RIR segments, derived from an analytical model for this transition based on the image source model (ISM) [15]. Along with a linear trajectory and constant velocity, our method assumes that the time intervals occupied by the direct sound and individual reflections over the entire trajectory do not overlap. These assumptions limit the method's applicability to small regions within a room, and its effectiveness in real-world conditions remains to be validated. Furthermore, the method relies on known RIRs at the start and end of the trajectory, from which the times of arrival (TOAs) of corresponding reflections must be accurately ex-

*__Corresponding author__: kathleen.macwilliam@esat.kuleuven.be.
__Copyright__: ©2025 First author et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

tracted and paired. To facilitate this, we proposed a DTW-based approach for estimating the parameters required to construct the derived transition matrix from the endpoint RIRs

In this paper, we extend our earlier work by introducing a segmentation-based approach to accommodate more complex microphone trajectories. Specifically, we divide longer trajectories into smaller linear segments where the original assumptions approximately hold, allowing for a piecewise application of our method. This segmentation strategy enables the handling of non-constant microphone velocities and curved paths while maintaining the model's validity. We validate both the original and extended methods using the trajectoRIR dataset [16], which includes moving microphone recordings along an L-shaped path in a real room, along with measured RIRs at discrete locations. We also emphasize the advantages of DTW-based transition matrix estimation in real-world scenarios, where inconsistencies in reflection amplitudes and the absence of certain reflections across RIRs are common.

## 2. SIGNAL MODEL

Consider a microphone moving along a trajectory and recording a signal from a static source. Let $k$ be the discrete-time index, and let $l$ be a one-dimensional index representing the microphone's location along the trajectory. The RIR is the LTI system relating the source signal $x(k)$ to the observed signal $y(l,k)$ at location $l$. In this paper, we focus only on the early part of the RIR, excluding the reverberant tail, though we will still refer to them as RIRs throughout for simplicity. Let $h(l,n)$ denote the early RIR at microphone location $l$, where $n$ indexes the time shift of the RIR samples. Defining the vectors:

$$\mathbf{x}(k) = (x(k) \quad x(k-1) \quad \dots \quad x(k-N+1))^T,$$

$$\mathbf{h}(l) = (h(l,0) \quad h(l,1) \quad \dots \quad h(l,N-1))^T.$$

The observed signal is then modeled as

$$y(l,k) = \mathbf{x}^T(k)\mathbf{h}(l) + v(k), \tag{1}$$

where $v(k)$ represents noise and may also account for the late reverberation excluded from the RIR model. If the microphone moves along a linear trajectory at a constant speed, the location index $l$ can be directly related to the time index $k$, as detailed in the next section. For a nonlinear trajectory, we can approximate it as a sequence of locally linear segments, where each segment adheres to the same model assumptions. In this paper, we aim to estimate $\mathbf{h}(l)$ using the known source signal $x(k)$ and the observed signal $y(l,k)$, while accounting for a time-variant microphone location.

## 3. PROPOSED STATE-SPACE MODEL

Now, consider a trajectory of length $d$, which may be nonlinear, indexed by $l \in \{0,\dots,L_{\text{total}}-1\}$, where $L_{\text{total}}$ is the total number of sampled locations. Let us partition the trajectory into $S$ approximately linear segments of varying lengths, indexed by $s \in \{1,\dots,S\}$. Within each segment, we assume the microphone moves with a constant velocity $\mu_s$. Let segment $s$ have length $d_s$ and contain $L_s$ uniformly spaced locations, where the first location of each segment coincides with the last location of the preceding segment. The spacing between consecutive locations within a segment is denoted by $\Delta d_s$. The set of location indices for segment $s$ is defined as

$$\mathcal{L}_s = \{l_{s|\text{st}}, \dots, l_{s|\text{en}}\}, \quad s \in \{1,\dots,S\}, \tag{2}$$

where $l_{s|\text{st}}$ and $l_{s|\text{en}}$ are, respectively, the start and end location indices of the segment. Figure 1 illustrates this setup. Given a temporal sampling period $T$, the smallest spatial increment within segment $s$ for a constant velocity $\mu_s$ is $\Delta d_s = \mu_s T$. In this case, the microphone crosses one location $l$ at each time instant $k$, so the location-variant RIR $\mathbf{h}(l)$ can equivalently be interpreted as a time-variant RIR $\mathbf{h}(k)$ with $k = l$. Accordingly, we define $y(l) = y(l,k)$, $\mathbf{x}(l) = \mathbf{x}(k)$ and $v(l) = v(k)$, allowing us to express the system in the following state-space form:

$$\mathbf{h}(l) = \mathbf{A}(l)\mathbf{h}(l-1) + \mathbf{w}(l), \tag{3}$$

$$y(l) = \mathbf{x}^T(l)\mathbf{h}(l) + v(l), \tag{4}$$

for $l \in \{1,\dots,L_{\text{total}}-1\}$. Here, $\mathbf{A}(l)$ is a location-variant transition matrix, and $\mathbf{w}(l)$ represents process noise modeling errors in $\mathbf{A}(l)$. Using this state-space model, an estimate $\hat{\mathbf{h}}(l)$ of $\mathbf{h}(l)$ can be obtained via a Kalman filter [17], with one recursion per location, as detailed in Section 3.1.

To define $\mathbf{A}(l)$, we assume that within each approximately linear segment, the RIR at location $l$ can be related to its predecessor through a segment-dependent transition matrix $\mathbf{A}_s$, which is location-*invariant* within the segment i.e.,

$$\mathbf{h}(l) \approx \mathbf{A}_s\mathbf{h}(l-1), \quad l \in \mathcal{L}_s \setminus \{l_{s|\text{st}}\}. \tag{5}$$

We thus propose that the location-variant transition matrix $\mathbf{A}(l)$ be defined as a piecewise constant function that follows the segment structure, such that

$$\mathbf{A}(l) = \begin{cases} \mathbf{A}_1, & l \in \mathcal{L}_1 \setminus \{l_{1|\text{st}}\} \\ \mathbf{A}_2, & l \in \mathcal{L}_2 \setminus \{l_{2|\text{st}}\} \\ \vdots & \\ \mathbf{A}_S, & l \in \mathcal{L}_S \setminus \{l_{S|\text{st}}\} \end{cases} \tag{6}$$

The segment-dependent matrix $\mathbf{A}_s$ can be derived using the ISM and estimated by interpolating the TOAs of the direct component and reflections between assumed known RIRs at the segment boundaries. This is detailed in Section 3.2.

### 3.1 Kalman filter update equations

For the state-space model in (3)–(4), let $\mathbf{Q} = \mathrm{E}[\mathbf{w}(l)\mathbf{w}^T(l)]$ and $R = \mathrm{E}[v^2(l)]$ denote the covariance of the process noise $\mathbf{w}(l)$ and the variance of the observation noise $v(l)$, respectively. While $R$ can potentially be estimated from background noise recordings, $\mathbf{Q}$ is typically a tuning parameter. The Kalman filter

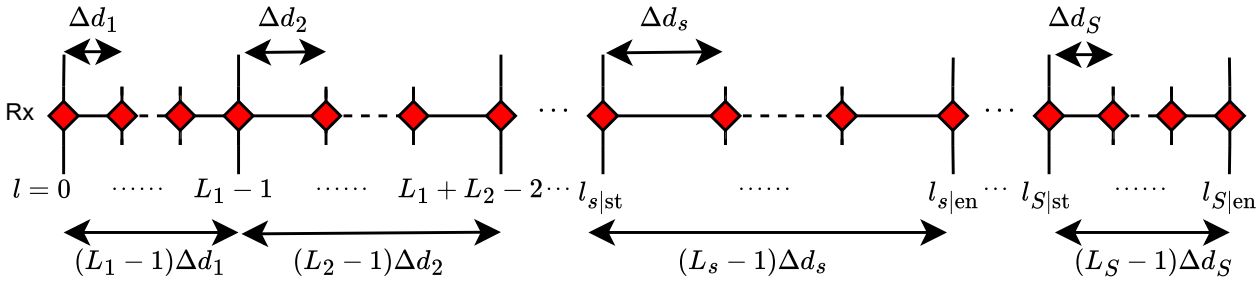



**Figure 1**. Illustration of trajectory partitioned into $S$ approximately linear segments of varying lengths.

update equations are given by:

$$\hat{\mathbf{h}}(l) = \mathbf{A}(l)\hat{\mathbf{h}}^{+}(l-1), \tag{7}$$

$$\mathbf{P}(l) = \mathbf{A}(l)\mathbf{P}^{+}(l-1)\mathbf{A}^{T} + \mathbf{Q}, \tag{8}$$

$$\mathbf{k}(l) = \frac{\mathbf{P}(l)\mathbf{x}(l)}{\mathbf{x}^{T}(l)\mathbf{P}(l)\mathbf{x}(l) + R}, \tag{9}$$

$$\hat{\mathbf{h}}^{+}(l) = \hat{\mathbf{h}}(l) + \mathbf{k}(l)\left(y(l) - \mathbf{x}^{T}(l)\hat{\mathbf{h}}(l)\right), \tag{10}$$

$$\mathbf{P}^{+}(l) = \left(\mathbf{I} - \mathbf{k}(l)\mathbf{x}^{T}(l)\right)\mathbf{P}(l), \tag{11}$$

Equations (7)–(8) compute the prior estimates (prediction step), while (9)–(11) refine them into posterior estimates, denoted by the superscript $^{+}$, using new observations (update step). Here $\hat{\mathbf{h}}(l)$ and $\hat{\mathbf{h}}^{+}(l)$ are the state estimates, $\mathbf{P}(l)$ and $\mathbf{P}^{+}(l)$ are the respective state-estimation error covariance matrix estimates, and $\mathbf{k}(l)$ is the Kalman gain. The filter is initialized with $\hat{\mathbf{h}}^{+}(0) = \mathbf{h}(0)$, assuming a known initial RIR, and $\mathbf{P}^{+}(0)$ is set to a small-norm matrix, reflecting negligible initial uncertainty. Unlike common state-space models, such as the one in [6], which use a scalar $\alpha$ instead of $\mathbf{A}(l)$, we estimate the segment-dependent transition matrix $\mathbf{A}_s$ needed to construct $\mathbf{A}(l)$ as in (6) using the ISM-based approach detailed in the next section. If no recorded signals are available, i.e., $y(l) = 0$ and $\mathbf{x}(l) = 0$, the recursion simplifies to

$$\hat{\mathbf{h}}(l) = \mathbf{A}(l)\hat{\mathbf{h}}(l-1) \approx \mathbf{A}_s^{(l-l_{s|\text{st}})}\mathbf{h}(l_{s|\text{st}}), \tag{12}$$

which corresponds to a linear interpolation of the RIRs, similar to the methods in [11, 13].

### 3.2 Segment-dependent transition matrix

Here we present a summary of the ISM-based model proposed in [3] for the transition matrix $\mathbf{A}_s \in \mathbb{R}^{N\times N}$ that approximates the mapping between two RIRs $\mathbf{h}(l_{s|\text{st}})$ and $\mathbf{h}(l_{s|\text{en}})$ as in equation (5). Using the ISM, the RIR can be expressed as the sum of contributions from the original sound source and additional image sources, which represent reflections within the room. Let the ISM representation of the discrete-time RIR at location $l$ with discrete time-shift index $n$ and sampling period $T$ be given by,

$$h(l,n) = \sum_{r\in\mathcal{R}} a_r(l)\,\text{sinc}\left(n - \frac{\tau_r(l)}{T}\right) \tag{13}$$

where $r \in \mathcal{R} = \{0, 1, ..R-1\}$ is the (image) source or reflection index, and $\tau_r(l)$ and $a_r(l)$ respectively denote the TOA and amplitude of the (image) source $r$ observed at location $l$. Let

$$\Delta_r(l) = \tau_r(l) - \tau_r(l-1) \tag{14}$$

be the time difference of arrival (TDOA) of reflection $r$ between locations $l$ and $l-1$. Recall that we are considering a linear segment of the trajectory with $l \in \mathcal{L}_s \setminus \{l_{s|\text{st}}\}$ where $\mathcal{L}_s$ is defined in (2). Under the assumption of far-field conditions, where the (image) sources are located much further away than the length of the segment, we get $a_r(l) = a_r(l-1)$ and $\Delta_r(l) = \Delta_{r,s}$. We can obtain an estimate of this constant TOA shift using

$$\Delta_{r,s} = \frac{\tau_r(l_{s|\text{en}}) - \tau_r(l_{s|\text{st}})}{L_s - 1}, \tag{15}$$

where $\tau_r(l_{s|\text{st}})$ and $\tau_r(l_{s|\text{en}})$ denote the (image) source TOAs at the initial and final locations on the segment respectively, and $L_s$ is the total number of locations on that segment. Let us now define the interval that the TOAs of non-negligible components of reflection $r$ occupy over the segment $l \in \mathcal{L}_s \setminus \{l_{s|\text{st}}\}$ as

$$\mathcal{T}_{r,s} = [\tau_{r,s|\,\text{min}},\ \tau_{r,s|\,\text{max}}] \tag{16}$$

$$\tau_{r,s|\,\text{min}} = \min\big(\tau_r(l_{s|\text{st}}+1),\ \tau_r(l_{s|\text{en}})\big) - \epsilon, \tag{17}$$

$$\tau_{r,s|\,\text{max}} = \max\big(\tau_r(l_{s|\text{st}}+1),\ \tau_r(l_{s|\text{en}})\big) + \epsilon. \tag{18}$$

where $2\epsilon$ is the effective temporal width of an individual reflection $r$. Based on this we define the set

$$\boxed{\tilde{\mathcal{R}}_s(n) = \left\{ r \in \mathcal{R} \;\middle|\; n \in \frac{\mathcal{T}_{r,s}}{T} \right\}}. \tag{19}$$

Under the further assumption that the intervals occupied by individual reflections along the segment do not overlap, we derived the relation between $h(l,n)$ and $h(l-1,n')$ in [3]. Here, $n$ and $n'$ denote the respective time shift indices, and the relation is given by

$$h(l,n) \approx \sum_{n'} \sum_{r\in\tilde{\mathcal{R}}_s(n)} \text{sinc}\left(n - \frac{\Delta_{r,s}}{T} - n'\right) h(l-1,n') \tag{20}$$

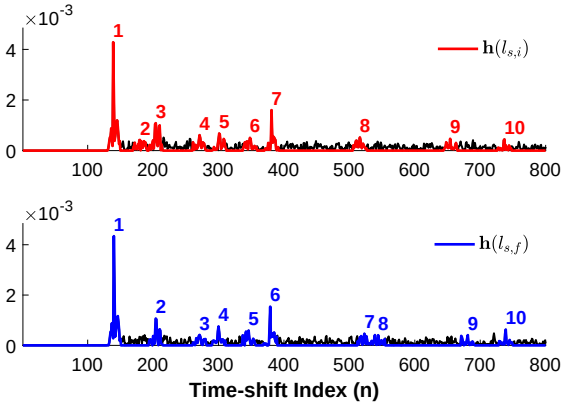



**Figure 2**. Example of mismatched TOA estimations using peak-picking and a matched filter approach as in [19, 20].

for $l \in \mathcal{L}_s \setminus \{l_{s|\text{st}}\}$. Consequently, the element of $\mathbf{A}_s$ at index $(n+1, n'+1)$ is:

$$A_{s(n+1,n'+1)} = \sum_{r \in \tilde{\mathcal{R}}_s(n)} \text{sinc}\left(n - \frac{\Delta_{r,s}}{T} - n'\right). \quad (21)$$

when $\tilde{\mathcal{R}}_s(n) \neq \varnothing$, and $A_{s(n+1,n'+1)} = 0$ otherwise. A detailed derivation is available in [3].

### 3.3 Practical estimation of transition matrix

The proposed transition matrix in (21) relies on accurately retrieving and pairing TOAs in $\mathbf{h}(l_{s|\text{st}})$ and $\mathbf{h}(l_{s|\text{en}})$. Assuming non-overlapping TOA intervals, reflections should arrive in the same order. A common method for identifying TOAs is to select the $R$ largest peaks using a peak-picking algorithm [18] after matched filtering [19], refined via DTW in [20] to improve peak detection of reflections that poorly correlate with the direct sound due to shape changes. However, owing to variations in reflection amplitudes or the absence of certain reflections at specific locations (particularly in real data), the selected peaks may differ between RIRs. Consequently, reflections may be incorrectly paired, as illustrated in Figure 2. Therefore, a method was proposed to directly estimate the parameters needed for the matrix $\mathbf{A}_s$ using DTW. This approach aligns elements in $\mathbf{h}(l_{s|\text{st}})$ and $\mathbf{h}(l_{s|\text{en}})$, leveraging their warp matrices to extract both $\tilde{\mathcal{R}}_s(n)$ and $\tau_r(l_{s|\text{en}}) - \tau_r(l_{s|\text{st}})$, the latter of which can be used to estimate $\Delta_{r,s}$ via (15).

The DTW algorithm [21] computes an accumulated cost matrix $\mathbf{D} \in \mathbb{R}^{(N+1)\times(N+1)}$, where each entry represents the cost of aligning $h(l_{s|\text{en}}, n)$ with $h(l_{s|\text{st}}, n')$. The cost function is defined as:

$$D_{n+2,n'+2} = \|h(l_{s|\text{en}}, n) - h(l_{s|\text{st}}, n')\|_2 + \min\left\{\begin{matrix} D_{n+1,n'+2} \\ D_{n+2,n'+1} \\ D_{n+1,n'+1} \end{matrix}\right\}$$

initialized [1] with $D_{1,1} = 0$ and boundary values set to $\infty$. The optimal alignment is represented by a warp path $w = \{(n+2, n'+2)\}_{i=1}^{I}$, a sequence of $I$ index pairs that defines correspondences between the samples in the RIR vectors. This path is found by backtracking from $D_{N+2,N+2}$ to $D_{2,2}$, maintaining a monotonically increasing order. Note that a single point in one RIR can correspond to multiple points in the other, thus $I \geq N$. Once the optimal alignment is determined, we can warp the RIRs to align them temporally. Specifically, $\mathbf{h}(l_{s|\text{st}})$ can be warped to align with $\mathbf{h}(l_{s|\text{en}})$, yielding $\tilde{\mathbf{h}}(l_{s|\text{st}})$, and vice versa, yielding $\tilde{\mathbf{h}}(l_{s|\text{en}})$. This warping process can be expressed as a matrix operation on each of the original vectors namely,

$$\tilde{\mathbf{h}}(l_{s|\text{st}}) = \mathbf{W}_{\text{st}}\mathbf{h}(l_{s|\text{st}}), \quad \tilde{\mathbf{h}}(l_{s|\text{en}}) = \mathbf{W}_{\text{en}}\mathbf{h}(l_{s|\text{en}}),$$

where $\mathbf{W}_{\text{st}}, \mathbf{W}_{\text{en}} \in \mathbb{R}^{I\times N}$ are sparse matrices constructed such that $[\mathbf{W}_{\text{st}}]_{i,n+2} = 1$ and $[\mathbf{W}_{\text{en}}]_{i,n'+2} = 1$ for each $i \in \{1, \ldots, I\}$, with $(n+2, n'+2)$ being the $i^{\text{th}}$ index pair in the warp path $w$; all other entries are zero. Now, in order to find a direct mapping from $\mathbf{h}(l_{s|\text{st}})$ to $\mathbf{h}(l_{s|\text{en}})$, we aim to satisfy the condition $\mathbf{W}_{\text{st}}\mathbf{h}(l_{s|\text{st}}) \approx \mathbf{W}_{\text{en}}\mathbf{h}(l_{s|\text{en}})$. This problem can be formulated as a least squares problem, leading to

$$\mathbf{h}(l_{s|\text{en}}) = \mathbf{W}_{\text{en}}^{+}\mathbf{W}_{\text{st}}\mathbf{h}(l_{s|\text{st}}) \quad (22)$$

where $\mathbf{W}_{\text{en}}^{+}$ is the generalized left inverse of $\mathbf{W}_{\text{en}}$. For a more detailed derivation of (22) we refer the interested reader to [20]. The warp matrix product $\mathbf{W}_{\text{en}}^{+}\mathbf{W}_{\text{st}}$, which we hereby denote as $\mathbf{W}$, serves as an estimate of $\mathbf{A}_s^{(l_{s|\text{en}}-l_{s|\text{st}})}$ for $l = l_{s|\text{en}}$ in Equation (12) i.e.,

$$\mathbf{W} = \mathbf{A}_s^{(l_{s|\text{en}}-l_{s|\text{st}})} = \mathbf{A}_s^{L_s-1}. \quad (23)$$

Finding an estimate of $\mathbf{A}_s$ by directly computing the $(L_s - 1)^{\text{th}}$ root of $\mathbf{W}$ is not straightforward due to potential non-diagonalizability and numerical instability, therefore we instead analyze its structural properties. When a single point in $\mathbf{h}(l_{s|\text{st}})$ maps to a single point in $\mathbf{h}(l_{s|\text{en}})$, the corresponding row or column in $\mathbf{W}$ contains only one nonzero entry. However, if multiple points align, multiple nonzero entries appear in that row or column. Therefore, diagonal segments in $\mathbf{W}$ typically indicate regions where similar patterns, such as reflections, align.

Using this observation, integer estimates of $(\tau_r(l_{s|\text{en}}) - \tau_r(l_{s|\text{st}}))/T$ can be obtained by measuring the 'distances' (number of elements) from sufficiently long diagonal segments in $\mathbf{W}$ to the main diagonal. These estimates are then used to find estimates of $\Delta_{r,s}$ using the relation in (15) and the known sampling period $T$ leading to:

$$\hat{\Delta}_{r,s} = \frac{\hat{\tau}_r(l_{s|\text{en}}) - \hat{\tau}_r(l_{s|\text{st}})}{L_s - 1}. \quad (24)$$

The start and end indices of the diagonal segments are also retrieved, denoted respectively as $(n_{r|\text{st}}, n'_{r|\text{st}})$ and $(n_{r|\text{en}}, n'_{r|\text{en}})$.

[1] An additional row and column are used exclusively for initialization. As matrices in this paper are indexed starting from (1,1) while RIR sequences are indexed from 0, the entries of the accumulated cost matrix are therefore indexed as $D_{n+2,n'+2}$.

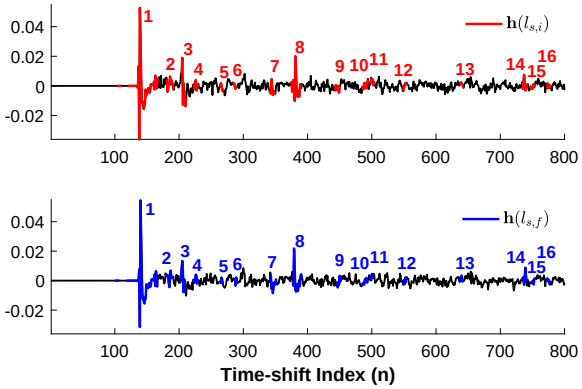



**Figure 3**. Illustration of correctly paired peaks in RIRs using the proposed DTW-based method.

These indices are then used to find a suitable estimate of the range $\mathcal{T}_{r,s}$ for the set $\tilde{\mathcal{R}}_s(n)$ in Equation (19) as,

$$\hat{\mathcal{T}}_{r,s} = [\hat{\tau}_{r,s|\,\min},\ \hat{\tau}_{r,s|\,\max}] \tag{25}$$

$$\hat{\tau}_{r,s|\,\min} = T \cdot \min\left(n'_{r|\mathrm{st}} + \frac{\hat{\Delta}_{r,s}}{T},\ n_{r|\mathrm{st}}\right), \tag{26}$$

$$\hat{\tau}_{r,s|\,\max} = T \cdot \max\left(n_{r|\mathrm{en}},\ n'_{r|\mathrm{en}} + \frac{\hat{\Delta}_{r,s}}{T}\right). \tag{27}$$

Figure 3 illustrates the correctly associated TOA ranges derived from the diagonal segments. Unlike the peak-picking method shown in Figure 2, our DTW-based approach does not require specifying the number of peaks, as it automatically aligns all potential reflection peaks across the RIRs. A minimum diagonal segment length is enforced to reduce incorrect reflection peak detections. Finally, we construct $\hat{\mathbf{A}}_s$ using Equations (21) and (19) with estimates $\hat{\Delta}_{r,s}$ and $\hat{\mathcal{T}}_{r,s}$ obtained from Equations (24) to (27). Additionally, as a robustness measure for sections where reflections are not identified, ones are added along the main diagonal.

## 4. EXPERIMENTAL EVALUATION AND RESULTS

In this experiment, we evaluate the proposed method using real data from the trajectoRIR database [16]. The setup we chose consists of a static sound source emitting a white noise signal while a microphone moves along an L-shaped trajectory at a constant velocity of 0.8 m/s, recording the signal. RIRs were also measured at fixed locations along the trajectory at 5 cm intervals, resulting in a total of 92 measurement points. We denote the set of all measurement indices as $\mathcal{P} = \{1, 2, \ldots, 92\}$. A schematic of the trajectory with corresponding RIR measurement points is shown in Figure 4. The initial recordings were made at a sampling rate of 48 kHz, but to align with typical speech processing applications, the signals were downsampled to 16 kHz. This downsampling also reduces the computational load as less data points need to be processed across the entire trajectory.

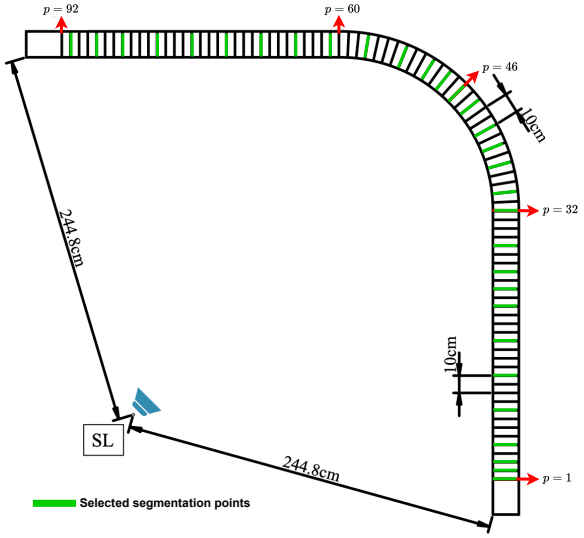



**Figure 4**. Schematic of trajectory with points used for the segmentation in green.

### 4.1 Trajectory segmentation

As the microphone moved through the fixed trajectory locations, the corresponding timestamps in the recorded signal, $y(l)$, were retrieved as detailed in [16]. We denote the corresponding location index of each measurement point $p \in \mathcal{P}$ as $l_p$. For segmentation purposes, we selected a subset of measurement points, $\tilde{\mathcal{P}} \subseteq \mathcal{P}$, with $\tilde{\mathcal{P}} = \{1, 2, 3, 5, 9, 13, 17, 21, 23, 25, 28, 32, 34, 37, 39, 41, 44, 46, 48, 50, 53, 57, 61, 65, 69, 73, 77, 81, 85, 88, 91\}$.

The selected measurement points (highlighted in green in Figure 4) are more densely spaced at the start and end of the trajectory, where velocity was not yet constant, and along the curved section. This ensures our modeling assumptions hold on those segments. Additionally, closer spacing was used in sections where the visibility of certain reflection peaks varied significantly if further-spaced RIRs were used (e.g., $p = [21, 25]$). Based on the measured RIRs at the selected points, the segment-dependent transition matrices $\mathbf{A}_s$ were then constructed as described in Section 3.3.

### 4.2 Kalman filter parameters

Given the assumption of knowledge of the RIR at the start of the trajectory $\mathbf{h}(0)$, the Kalman filter is initialized with $\hat{\mathbf{h}}^{+}(0) = \mathbf{h}(0)$. The measurement noise variance $R$ was not known in this case but after some initial testing of the Kalman filter was chosen as $R = 0.01$, while the process noise covariance matrix $\mathbf{Q}$ was a tuning parameter chosen as $\mathbf{Q} = \sigma_w^2\mathbf{I}$, where $\mathbf{I}$ is an identity matrix and $10\log(\sigma_w^2) = -50$ dB. The appropriate $\mathbf{A}(l)$ matrix is selected for each algorithm, as outlined in Table 1. It was found that for the reference Kalman filter, a transition factor of $\alpha = 1$ was suitable and varying this or applying a forgetting factor did not noticeably change the results.

**Table 1**. Description of the compared algorithms.

| Algorithm Name | Algorithm Description |
|---|---|
| **KF-$\alpha$** | Kalman Filter update equations (7)-(11) with $\mathbf{A}(l)$ replaced with $\alpha = 1$. |
| **KF-A**$(l)$ | Kalman Filter update equations (7)-(11) with $\mathbf{A}(l)$ estimated using the segment-dependent $\hat{\mathbf{A}}_s$ described in Section 3.3 |
| **LI-A**$(l)$ | Linear interpolation equivalent to Kalman Filter update equation (7) with $\mathbf{A}(l)$ estimated using the segment-dependent $\hat{\mathbf{A}}_s$ described in Section 3.3 |

### 4.3 Performance measures

Since the measured RIRs are only available at discrete points along the trajectory, we evaluate the accuracy of the estimated location- or time-variant RIR at these specific locations. Let $\hat{\mathbf{h}}^+(l_p)$ and $\mathbf{h}(l_p)$ denote the estimated and measured RIRs at location index $l_p$, respectively. To assess performance, we use a combination of two key metrics: normalized cross-correlation (NCC) for alignment and normalized misalignment (NM) for error quantification.

To address small uncertainties in positioning during RIR measurement and tracking errors during microphone movement, NCC is initially used to correct for any resulting inaccuracies in selecting the appropriate location-variant RIR index corresponding to the measured RIR at $l_p$. Specifically, we align $\hat{\mathbf{h}}^+(l_p)$ with $\mathbf{h}(l_p)$ by determining the delay $\lambda_{\max}(l_p)$ that maximizes the NCC function,

$$\lambda_{\max}(l_p) = \arg\max_{\lambda} \frac{\mathbf{h}(l_p) \cdot \hat{\mathbf{h}}^+(l_p - \lambda)}{\|\mathbf{h}(l_p)\|_2 \|\hat{\mathbf{h}}^+(l_p - \lambda)\|_2}.$$

We denote the resulting aligned estimated RIR as $\hat{\mathbf{h}}^+(l_p - \lambda_{\max}(l_p))$ and then compute the NM in decibels to quantify the error:

$$\mathcal{M}_{\mathrm{dB}}(l_p) = 20 \log_{10} \left( \frac{\|\hat{\mathbf{h}}^+(l_p - \lambda_{\max}(l_p)) - \mathbf{h}(l_p)\|_2}{\|\mathbf{h}(l_p)\|_2} \right). \tag{28}$$

with lower values indicating a better match between the estimated and measured RIRs.

Lastly, we evaluate how well the estimated location-variant RIR reproduces the recorded microphone signal. This is done by convolving the estimated location-variant RIR with the known input signal to obtain an estimated microphone signal, $\hat{y}(l) = \mathbf{x}^T(l)\hat{\mathbf{h}}^+(l)$, and computing its correlation with the recorded microphone signal $y(l)$.

### 4.4 Results and Discussion

The top panel of Figure 6 illustrates the normalized misalignment of the different algorithms as defined in (28), with vertical dashed lines indicating segmentation points. The bottom panel shows the estimated average velocity of the microphone between each known trajectory point, derived from timestamp data and interpoint distances. Figure 5 shows an example of the measured RIR and corresponding estimated RIRs at a selected trajectory point, allowing a qualitative comparison of the different algorithms. Lastly, Table 2 contains the correlation coefficient between the estimated and actual recorded microphone signal for each algorithm.

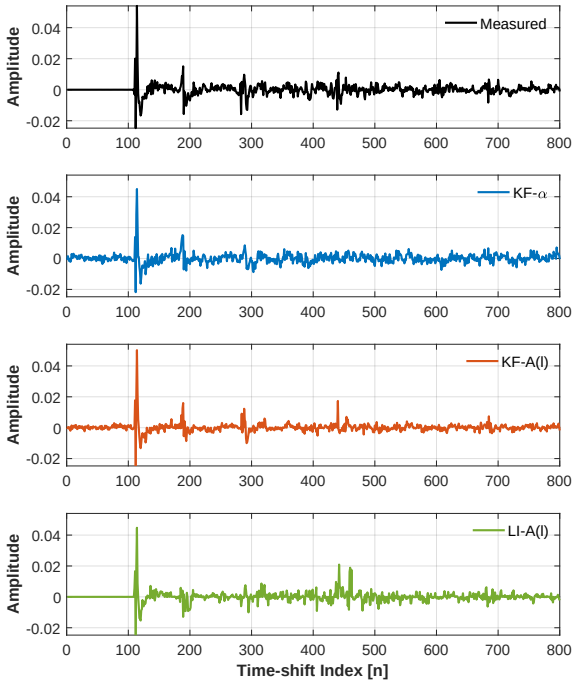



**Figure 5**. Measured RIR at a point along the trajectory compared to estimated RIRs from various algorithms.

**Table 2**. Correlation coefficients between estimated and recorded microphone signals for different algorithms.

| Algorithm | Correlation Coefficient |
|---|---|
| **KF-$\alpha$** | 0.9382 |
| **KF-A**$(l)$ | 0.9444 |
| **LI-A**$(l)$ | 0.3359 |

The results in Figure 6 indicate that the proposed algorithm, **KF-A**$(l)$, generally outperforms the reference methods **LI-A**$(l)$ and **KF-**$\alpha$. This performance advantage is particularly apparent in the linear sections of the trajectory, where velocity remains approximately constant, which is expected given these most strongly meet the original underlying model assumptions.

The **LI-A**$(l)$ algorithm consistently underperforms compared to both data-driven algorithms, underscoring the limitations of purely interpolative methods. However, between points $p = [60, 78]$, both **KF-**$\alpha$ and **KF-A**$(l)$ see a degrade in performance, resulting in **LI-A**$(l)$ outperforming **KF-**$\alpha$. This decline stems from a spike in noise in the recorded microphone signal along this section of the trajectory, though the cause re-

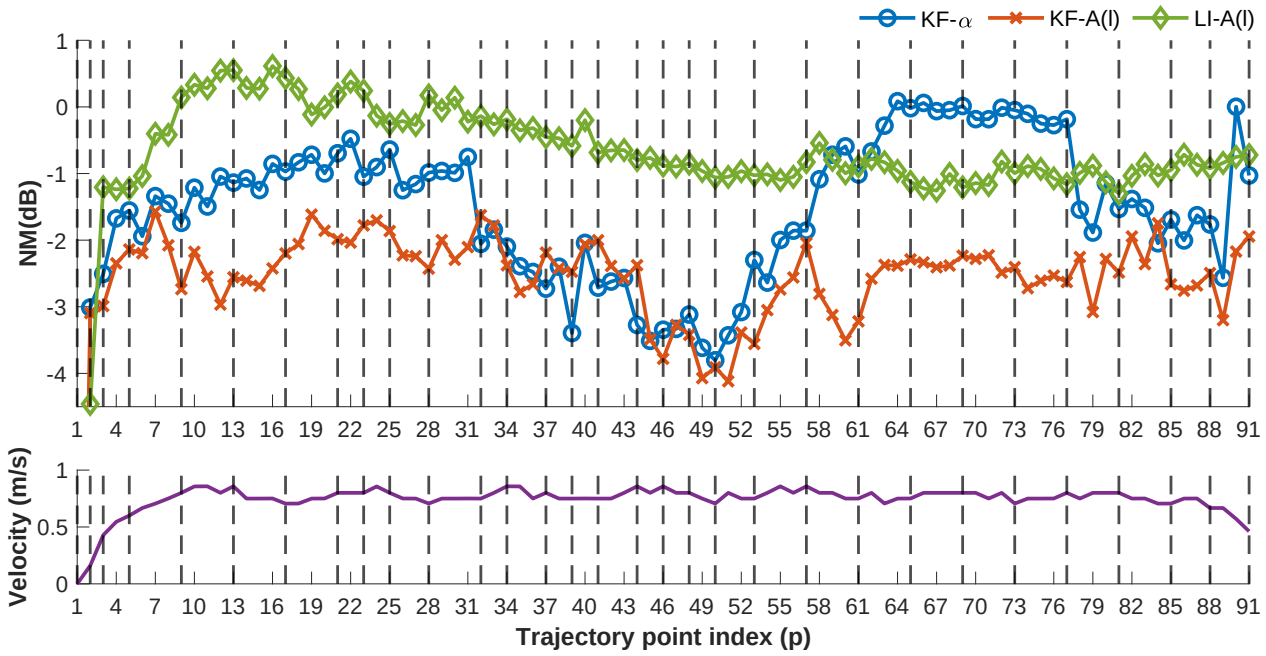



**Figure 6**. Top: Normalized misalignment between estimated and measured RIRs along the trajectory. Bottom: Estimated average velocity between trajectory points.

mains unknown. We retained the use of this particular signal to demonstrate our method's robustness to noise variation, as it still achieves the best performance in these circumstances.

Around the midpoint of the trajectory ($p = 46$), it can be observed that the performances of **KF-**$\alpha$ and **KF-A**$(l)$ somewhat converge. This can be attributed to the decrease in the rate of change in the TOA of the direct source as the microphone approaches the turning point $p = 46$ of its symmetrical curved path around the source position (see Figure 4). Under these conditions, the assumption of a transition factor of $\alpha = 1$ becomes a reasonable approximation, explaining the similar performance between the two Kalman-based approaches.

Performance differences are also evident in the curved sections of the trajectory. In the range $p = [46, 60]$, **KF-A**$(l)$ outperforms the other methods, however, in the range $p = [32, 46]$, this is not always the case. This discrepancy may stem from significant variations in the amplitudes of associated reflection peaks in this region. A closer inspection of the measured RIRs confirmed this, revealing that certain reflections were either absent or significantly dampened at specific points along this section of the trajectory. This is likely related to the surrounding acoustic environment, particularly the narrow planar underside of the staircase above the trajectory [16].

While all methods perform comparably in estimating the direct path component, **KF-A**$(l)$ seems to have better accuracy in capturing reflection peaks, as evident in Figure 5. This is particularly important as early reflections play a crucial role in intelligibility in speech applications, localization, and room geometry estimation [22]. Finally, the results in Table 2 confirm that **KF-A**$(l)$ achieves the highest correlation between estimated and recorded signals. Notably, the **LI-A**$(l)$ algorithm performs particularly poorly in this measure. These results reinforce the efficacy of the proposed approach in estimating RIRs in real-world applications.

## 5. CONCLUSIONS

In this paper, we extend our previous method [3] for estimating the early part of RIRs in a time-variant acoustic scenario using a state-space model with the ISM incorporated into the state transition matrix. By segmenting longer trajectories into smaller linear segments, we address more complex scenarios while preserving the original assumptions. We validated both the original and extended methods with real measurements, demonstrating that they outperform both RIR interpolation and a purely data-driven state-space model using a transition factor. Additionally, we showed that our DTW-based parameter estimation for the derived state transition matrix improves the pairing and detection of reflection TOAs across RIRs.

## 6. ACKNOWLEDGMENTS

This work was supported in part by 1) The European Research Council under the European Union's Horizon 2020 research and innovation program / ERC Consolidator Grant: SONORA (no. 773268). This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information, 2) KU Leuven Internal Funds C14/21/075 "A holistic approach to the design of integrated and distributed digi-

tal signal processing algorithms for audio and speech communication devices", 3) The FWO Research Project: "The Boundary Element Method as a State-Space Realization Problem" (G0A0424N)

## 7. REFERENCES

[1] A. Ratnarajah, I. Ananthabhotla, V. K. Ithapu, P. F. Hoffmann, D. Manocha, and P. T. Calamia, "Towards improved room impulse response estimation for speech recognition," *Proc. 2023 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '23)*, pp. 1–5, 2022.

[2] F. Borra, I. D. Gebru, and D. Markovic, "Soundfield reconstruction in reverberant environments using higher-order microphones and impulse response measurements," in *Proc. 2021 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '19)*, pp. 281–285, 2019.

[3] K. MacWilliam, T. Dietzen, R. Ali, and T. Van Waterschoot, "State-space estimation of spatially dynamic room impulse responses using a room acoustic model-based prior," *Frontiers in Signal Processing*, vol. 4, p. 1426082, 2024.

[4] G. Enzner, "Analysis and optimal control of lms-type adaptive filtering for continuous-azimuth acquisition of head related impulse responses," in *Proc. 2008 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '08)*, pp. 393–396, 3 2008.

[5] C. Antweiler, A. Telle, P. Vary, and G. Enzner, "Perfect-sweep NLMS for time-variant acoustic system identification," in *Proc. 2012 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '12)*, pp. 517–520, mar 2012.

[6] G. Enzner, "Bayesian inference model for applications of time-varying acoustic system identification," in *Proc. 18th European Signal Process. Conf. (EUSIPCO '10)*, pp. 2126–2130, 2010.

[7] M. Hahmann and E. Fernandez-Grande, "A convolutional plane wave model for sound field reconstruction," *J. Acoust. Soc. Amer.*, vol. 152 5, p. 3059, 2022.

[8] X. Karakonstantis, D. Caviedes-Nozal, A. Richard, and E. Fernandez-Grande, "Room impulse response reconstruction with physics-informed deep learning.," *J. Acoust. Soc. Amer.*, vol. 155 2, pp. 1048–1059, 2024.

[9] N. Antonello, E. De Sena, M. Moonen, P. A. Naylor, and T. Van Waterschoot, "Room impulse response interpolation using a sparse spatio-temporal representation of the sound field," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 10, pp. 1929–1941, 2017.

[10] D. Sundström, F. Elvander, and A. Jakobsson, "Estimation of impulse responses for a moving source using optimal transport regularization," in *Proc. 2024 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '24)*, pp. 921–925, 2024.

[11] G. Kearney, C. Masterson, S. Adams, and F. Boland, "Dynamic time warping for acoustic response interpolation: Possibilities and limitations," in *Proc. 17th European Signal Process. Conf. (EUSIPCO '09)*, pp. 705–709, 2009.

[12] V. Bruschi, S. Nobili, S. Cecchi, and F. Piazza, "An innovative method for binaural room impulse responses interpolation," in *Audio Engineering Society Convention 148*, 2020.

[13] A. Geldert, N. Meyer-Kahlen, and S. J. Schlecht, "Interpolation of spatial room impulse responses using partial optimal transport," in *Proc. 2023 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '23)*, pp. 1–5, 2023.

[14] M. Nophut, S. Preihs, and J. Peissig, "Velocity-controlled kalman filter for an improved echo cancellation with continuously moving microphones," *J. Audio Eng. Soc.*, vol. 72, pp. 33–43, 1 2024.

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

[16] S. Damiano, K. MacWilliam, V. Lorenzoni, T. Dietzen, and T. van Waterschoot, "The trajectorir database: Room acoustic recordings along a trajectory of moving microphones," *arXiv preprint,* arXiv:2503.23004, Mar 2025. [Online]. Available: `https://doi.org/10.48550/arXiv.2503.23004`.

[17] D. Simon, *Optimal state estimation: Kalman, H infinity, and nonlinear approaches*. John Wiley & Sons, 2006.

[18] M. Brookes *et al.*, "Voicebox: Speech processing toolbox for matlab," *Software, available [Mar. 2011] from www. ee. ic. ac. uk/hp/staff/dmb/voicebox/voicebox. html*, vol. 47, p. 45, 1997.

[19] F. Antonacci, J. Filos, M. R. P. Thomas, E. A. P. Habets, A. Sarti, P. A. Naylor, and S. Tubaro, "Inference of room geometry from acoustic impulse responses," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 10, pp. 2683–2695, 2012.

[20] I. J. Kelly and F. M. Boland, "Detecting arrivals in room impulse responses with dynamic time warping," *IEEE Trans. Audio Speech Lang. Process.*, vol. 22, no. 7, pp. 1139–1147, 2014.

[21] M. Müller, *Information retrieval for music and motion*, ch. Dynamic Time Warping, pp. 69–84. Springer, 2007.

[22] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *The Journal of the Acoustical Society of America*, vol. 113, no. 6, pp. 3233–3244, 2003.