# Adaptive time segmentation for improved signal model parameter estimation for a single-source scenario

Li, Changheng; C.Hendriks, Richard

**DOI**
10.1109/IEEECONF59524.2023.10476904

**Publication date**
2023

**Document Version**
Final published version

**Published in**
Proceedings of the 2023 57th Asilomar Conference on Signals, Systems, and Computers

# ADAPTIVE TIME SEGMENTATION FOR IMPROVED SIGNAL MODEL PARAMETER ESTIMATION FOR A SINGLE-SOURCE SCENARIO

*Changheng Li and Richard C. Hendriks*

Signal Processing Systems (SPS) Group, Delft University of Technology, Delft, The Netherlands

2023 57th Asilomar Conference on Signals, Systems, and Computers | 979-8-3503-2574-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/IEEECONF59524.2023.10476904

## ABSTRACT

Estimating the parameters that describe the acoustic scene is very important for many microphone array applications. For example, consider the power spectral densities (PSDs) or relative acoustic transfer functions (RTFs) that are required when estimating a particular sound source using multi-microphone noise reduction. State-of-the-art algorithms estimate the parameters per segment, where each segment consists of a fixed number of time frames. These algorithms exploit the assumption that PSDs are constant per time frame, and RTFs are constant per segment. However, in practice, sound sources will move relative to the microphone array. Improved performance is therefore expected when the actual time frames that are used to form the segments are adapted such that time frames all share the same (unknown) RTF. In this paper, we therefore present an algorithm to obtain an optimal adaptive time segmentation and combine this with our previously published joint maximum likelihood estimator (JMLE) for jointly estimating the RTF, source PSD and late reverberation PSD of a single source in a reverberant environment.

***Index Terms—*** Adaptive time segmentation, microphone array signal processing, RTF estimation, PSD estimation.

## 1. INTRODUCTION

In hand-free speech communication applications such as hearing aids and mobile phones, microphone arrays are commonly used to enhance the quality and intelligibility of the target signal as the microphone signals are typically corrupted by late reverberation and ambient noise. Typically, this is done using spatial filtering techniques. However, these techniques depend on acoustic scene-related parameters such as the relative transfer function (RTF) of the target signal and the power spectral densities (PSDs) of the target signal, the late reverberation and the ambient noise, which are typically unknown in practice. Therefore, it is essential to estimate these parameters.

Speech signals are non-stationary in nature, but can be assumed stationary for a very short duration of about 10$\sim$30 ms. This results in the fact that the PSD of each acoustic component is constant for only a short duration. However, the RTF can be assumed constant as long as the sound source does not move relative to the microphone array. Typically, the duration that the source is static (defined here as a time segment) is longer than the duration that the speech source is stationary (defined here as a time frame). Hence, each time segment might contain multiple time frames that share the same RTF.

Recently, several estimation methods have been proposed to estimate the RTF and the PSDs using multiple time frames [1, 2] instead of a single time frame [3–10]. The methods using multiple time frames always outperform the methods using a single time frame as long as the sound source is indeed static during the time segment. However, if the source or array changes position or the room acoustics change, the methods using time segments during which the RTF is time-varying have worse estimation performance than when the time segment would be selected such that the underlying RTF is time-invariant. Therefore, in this paper, we present an algorithm to obtain an adaptive time segmentation and combine this with our previously published joint maximum likelihood estimator (JMLE) [2] for jointly estimating the RTF, source PSD and late reverberation PSD of a single source in a reverberant environment. Notice that the use of an adaptive time segmentation in the speech enhancement context has been proposed before, e.g., [11], for improved estimation of the PSDs used in single-microphone noise reduction algorithm. In the current work, we present a different segmentation algorithm for the multi-microphone context based on the inner product of a sequence of initial RTF estimates. In combination with the recently proposed JMLE algorithm, this leads to improved estimates of the RTF and PSDs.

## 2. PRELIMINARIES

### 2.1. Signal model

We consider a single acoustic point source observed by a microphone array in a reverberant environment. The source changes to new positions at unknown moments, which means it is spatially fixed for unknown time durations. The time duration that the source does not move will be referred to as a time segment indexed by $\beta$. The $\beta$-th time segment consists of one or multiple time frames from $t_\beta$ to $(t_\beta + T_\beta - 1)$, where $T_\beta$ is the number of time frames for the $\beta$-th time seg-

Changheng Li is supported by the China Scholarship Council.

ment. Within a time frame, the speech source is assumed to be stationary. The time frame will be indexed by $t$. Each time frame $t$ contains multiple overlapping sub-time frames. See Fig. 1 for a visual interpretation. We use the short-time

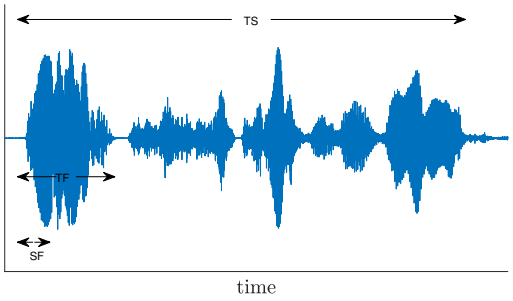



**Fig. 1**: Visualisation of the definition of time segment (TS), time frames (TF) and sub frames (SF).

Fourier transform (STFT) to transfer the signal received at the $m$-th microphone into the frequency domain, leading to

$$y_m(l,k) = x_m(l,k) + r_m(l,k), \tag{1}$$

with $l$ the sub-time frame index, $k$ the frequency bin index, and $m$ the microphone index. In Eq. (1), $x_m$ denotes the direct and early reflections of the source. Variable $r_m$ denotes the late reverberation, which is the sum of all the late reflections of the source. Using the relative transfer function (RTF) between the microphones, we can model $x_m$ as

$$x_m(l,k) = a_m(l,k)\, s(l,k), \tag{2}$$

with $s(l,k)$ the direct component and early reflections at the reference microphone (the $1^{st}$ microphone in this work) and $a_m(l,k)$ the RTF of the source from the reference microphone to the $m$-th microphone. In vector form, all $M$ microphone signals in the STFT domain can be expressed as

$$\mathbf{y}(l,k) = \mathbf{a}(l,k)\, s(l,k) + \mathbf{r}(l,k) \in \mathbb{C}^{M\times 1}. \tag{3}$$

Assuming that the early reflections and the late reverberation that fall in one sub-time frame are uncorrelated and zero-mean, we can write the covariance matrix of $\mathbf{y}(l,k)$ as

$$\mathbf{\Phi}_{\mathbf{y}}(l,k) \triangleq \mathbf{\Phi}_{\mathbf{x}}(l,k) + \mathbf{\Phi}_{\mathbf{r}}(l,k), \tag{4}$$

where $\mathbf{\Phi}_{\mathbf{y}} \triangleq E\{\mathbf{y}\mathbf{y}^H\}$ with $E\{\cdot\}$ the expectation operator. Matrices $\mathbf{\Phi}_{\mathbf{x}}$ and $\mathbf{\Phi}_{\mathbf{r}}$ are defined in the same way as $\mathbf{\Phi}_{\mathbf{y}}$. For $\mathbf{\Phi}_{\mathbf{x}}(l,k)$, we have

$$\mathbf{\Phi}_{\mathbf{x}}(l,k) = \phi_s(l,k)\, \mathbf{a}(l,k)\, \mathbf{a}^H(l,k), \tag{5}$$

where $\phi_s(l,k) \triangleq E\left\{|s(l,k)|^2\right\}$ is the PSD of the source at the reference microphone. For the late reverberation, we assume a spatially homogeneous sound field model

$$\mathbf{\Phi}_{\mathbf{r}}(l,k) = \phi_\gamma(l,k)\, \mathbf{\Gamma}(k), \tag{6}$$

where $\phi_\gamma(l,k)$ is the unknown time-varying PSD of the late reverberation and $\mathbf{\Gamma}(k)$ is the known time-invariant spatial coherence matrix, which can be calculated using the microphone array geometry [12].

### 2.2. Problem formulation

By using Eqs. (5) and (6), we formulate the noisy covariance matrix as

$$\mathbf{\Phi}_{\mathbf{y}}(t,k) = \phi_s(t,k)\, \mathbf{a}(\beta,k)\, \mathbf{a}^H(\beta,k) + \phi_\gamma(t,k)\mathbf{\Gamma}(k), \tag{7}$$

where we assumed the microphone signals are stationary over a time frame $t$ consisting of the $L_s$ sub-time frames indexed by $l = 1+(t-1)L_s$ till $l = tL_s$ and the RTF stays constant over a time segment $\beta$ consisting of the time frames indexed by $t = t_\beta$ till $t = t_\beta + T_\beta - 1$. Based on the stationarity assumption, we can estimate $\mathbf{\Phi}_{\mathbf{y}}(t,k)$ using the sample covariance matrix $\hat{\mathbf{\Phi}}_{\mathbf{y}}(t,k) = 1/L_s \sum_{l=1+(t-1)L_s}^{tL_s} \mathbf{y}(l,k)\, \mathbf{y}^H(l_s,k)$. Assuming that the RTF is constant for all time frames in a time segment (i.e., $t \in [t_\beta, t_\beta + T_\beta - 1]$), we can use the set $\{\mathbf{\Phi}_{\mathbf{y}}(t,k)\}_{t_\beta}^{t_\beta+T_\beta-1}$ jointly to estimate $\mathbf{a}(\beta,k)$.

The aim of this work is to estimate the time segment indices $\{t_\beta, T_\beta\}$ using the fact that the true but unknown RTFs are the same in the time frames from a single segment. Note that $t_1 = 1$ and the last time frame of the $(\beta-1)$-th time segment should be followed by the first time frame of the $\beta$-th time segment (i.e.,$t_\beta = t_{\beta-1} + T_{\beta-1}$). Since we determine the time segments sequentially, we know $\{t_{\beta-1}, T_{\beta-1}\}$ when determining the $\beta$-th time segment. Therefore, $t_\beta$ is known as well and we only need to estimate $T_\beta$.

## 3. JMLE

We first present the algorithm for joint MLE of the parameters $\mathbf{a}$ and $\{\phi_s(t), \phi_\gamma(t)\}_{t=t_\beta}^{t_\beta+T_\beta-1}$ for a given time segment $[t_\beta, t_\beta + T_\beta - 1]$. Note that this is based on our work recently published in [2]. In the next section, we will then propose the algorithm to determine $\{T_\beta\}$. Note that only in this section, we omit the frequency indexes for the simplicity of notation.

With the assumption that all the time frames are independent and the STFT coefficients are complex Gaussian distributed, we can write the negative log-likelihood function of the STFT coefficients (up to a constant and scale) as

$$L = -\sum_{t=t_\beta}^{t_\beta+T_\beta-1} \left[\log|\mathbf{\Phi}_{\mathbf{y}}(t)| + \mathrm{tr}\left(\hat{\mathbf{\Phi}}_{\mathbf{y}}(t)\, \mathbf{\Phi}_{\mathbf{y}}^{-1}(t)\right)\right]. \tag{8}$$

Then, using the reparameterization $\tilde{\mathbf{a}} = \frac{\mathbf{L}^{-1}\mathbf{a}}{\sqrt{\mathbf{a}^H\mathbf{\Gamma}^{-1}\mathbf{a}}}$ and $\tilde{\phi}_s(t) = \phi_s(t)\, \mathbf{a}^H\mathbf{\Gamma}^{-1}\mathbf{a}$, we reformulate the covariance

matrix in Eq. (7) as

$$\mathbf{\Phi_y}(t) = \mathbf{L}\left(\tilde{\phi}_s(t)\,\tilde{\mathbf{a}}\tilde{\mathbf{a}}^H + \phi_\gamma(t)\,\mathbf{I}\right)\mathbf{L}^H, \tag{9}$$

where $\mathbf{L}$ is the Cholesky factor of $\mathbf{\Gamma}$ (i.e. $\mathbf{\Gamma} = \mathbf{LL^H}$). The MLE cost function then becomes [2]

$$\begin{aligned}\underset{\tilde{\phi}_s(t),\tilde{\mathbf{a}},\phi_\gamma(t)}{\arg\min} \sum_{t=t_\beta}^{t_\beta+T_\beta-1} & \log\left[\left(\tilde{\phi}_s(t)+\phi_\gamma(t)\right)\left(\phi_\gamma(t)^{M-1}\right)\right] \\ & + \mathrm{tr}\left(\phi_\gamma(t)^{-1}\hat{\mathbf{P}}_{\mathbf{w}}(t)\right) \\ & - \frac{\phi_\gamma(t)^{-2}\tilde{\phi}_s(t)}{1+\phi_\gamma(t)^{-1}\tilde{\phi}_s(t)}\tilde{\mathbf{a}}^H\hat{\mathbf{P}}_{\mathbf{w}}(t)\,\tilde{\mathbf{a}},\end{aligned} \tag{10}$$

where $\hat{\mathbf{P}}_{\mathbf{w}} = \mathbf{L}^{-1}\hat{\mathbf{\Phi}}_{\mathbf{y}}\mathbf{L}^{-H}$.

To solve the above optimization, we first find initial estimates of the parameters by considering each time frame independently (as explained in Section 3.1). This initialisation step does thus not require a segmentation algorithm as it works on the individual time frames. After the initialisation, alternating estimation between $\mathbf{a}$ and $\{\phi_s(t),\phi_\gamma(t)\}_{t=t_\beta}^{t_\beta+T_\beta-1}$ is performed (see Section 3.2), which thus can benefit from a correct segmentation.

### 3.1. Initialisation

When considering a single time frame, the cost function reduces to

$$\begin{aligned}\underset{\tilde{\phi}_s(t),\tilde{\mathbf{a}},\phi_\gamma(t)}{\arg\min} & \log\left[\left(\tilde{\phi}_s(t)+\phi_\gamma(t)\right)\left(\phi_\gamma(t)^{M-1}\right)\right] \\ & + \mathrm{tr}\left(\phi_\gamma(t)^{-1}\hat{\mathbf{P}}_{\mathbf{w}}(t)\right) \\ & - \frac{\phi_\gamma(t)^{-2}\tilde{\phi}_s(t)}{1+\phi_\gamma(t)^{-1}\tilde{\phi}_s(t)}\tilde{\mathbf{a}}^H\hat{\mathbf{P}}_{\mathbf{w}}(t)\,\tilde{\mathbf{a}},\end{aligned} \tag{11}$$

where only the last term depends on $\tilde{\mathbf{a}}$ with a negative coefficient. Hence, the MLE-optimal $\tilde{\mathbf{a}}$ based on a single time frame is the solution to

$$\underset{\tilde{\mathbf{a}}}{\arg\max}\,\tilde{\mathbf{a}}^H\hat{\mathbf{P}}_{\mathbf{w}}(t)\,\tilde{\mathbf{a}}, \tag{12}$$

which is the principal eigenvector of $\hat{\mathbf{P}}_{\mathbf{w}}(t)$. Note that the initialization step does not use the prior information that all time frames in a time segment share the same RTF. Let $T$ denote the maximum size of a segment. For the $T$ time frames that could potentially form the $\beta$-th time segment, we will have $T$ different estimates of $\tilde{\mathbf{a}}$ at this step. These are denoted by $\left\{\hat{\tilde{\mathbf{a}}}(t)\right\}_{t=t_\beta}^{t_\beta+T-1}$. These estimates will be used for the time segmentation algorithm to find the actual length $T_\beta$ of the $\beta$-th time segment in Section 4.

With the estimated RTF $\hat{\tilde{\mathbf{a}}}(t)$, we can find the optimal estimates of $\tilde{\phi}_s(t)$ and $\tilde{\phi}_\gamma(t)$ by substituting $\hat{\tilde{\mathbf{a}}}(t)$ into Eq. (11) [2], that is,

$$\hat{\tilde{\phi}}_s(t) = \frac{M\lambda_{\max}(t) - \mathrm{tr}\left(\hat{\mathbf{P}}_{\mathbf{w}}(t)\right)}{M-1}, \tag{13}$$

$$\hat{\phi}_\gamma(t) = \frac{\mathrm{tr}\left(\hat{\mathbf{P}}_{\mathbf{w}}(t)\right) - \lambda_{\max}(t)}{M-1}, \tag{14}$$

where $\lambda_{\max}(t)$ is the principal eigenvalue of $\hat{\mathbf{P}}_{\mathbf{w}}(t)$.

### 3.2. Alternating estimation

The initial estimates of the PSDs $\left\{\hat{\tilde{\phi}}_s(t),\hat{\phi}_\gamma(t)\right\}_{t=t_\beta}^{t_\beta+T_\beta-1}$ can be substituted into the cost function in Eq. (10) to estimate the RTF using all time frames in the time segment jointly

$$\underset{\tilde{\mathbf{a}}}{\arg\max} \sum_{t=t_\beta}^{t_\beta+T_\beta-1}\left(\frac{\hat{\tilde{\phi}}_s(t)}{\hat{\phi}_\gamma(t)+\hat{\tilde{\phi}}_s(t)}\frac{1}{\hat{\phi}_\gamma(t)}\tilde{\mathbf{a}}^H\hat{\mathbf{P}}_{\mathbf{w}}(t)\,\tilde{\mathbf{a}}\right), \tag{15}$$

which is the principal eigenvector of

$$\sum_{t=t_\beta}^{t_\beta+T_\beta-1}\frac{\hat{\tilde{\phi}}_s(t)}{\hat{\phi}_\gamma(t)+\hat{\tilde{\phi}}_s(t)}\frac{1}{\hat{\phi}_\gamma(t)}\hat{\mathbf{P}}_{\mathbf{w}}(t). \tag{16}$$

Then, with the estimated RTF $\hat{\mathbf{a}}$, we can estimate the PSDs using [2]

$$\hat{\tilde{\phi}}_s(t) = \frac{M\hat{\tilde{\mathbf{a}}}^H\hat{\mathbf{P}}_{\mathbf{w}}(t)\,\hat{\tilde{\mathbf{a}}} - \mathrm{tr}\left(\hat{\mathbf{P}}_{\mathbf{w}}(t)\right)}{M-1} \tag{17}$$

and

$$\hat{\phi}_\gamma(t) = \frac{\mathrm{tr}\left(\hat{\mathbf{P}}_{\mathbf{w}}(t)\right) - \hat{\tilde{\mathbf{a}}}^H\hat{\mathbf{P}}_{\mathbf{w}}(t)\,\hat{\tilde{\mathbf{a}}}}{M-1}. \tag{18}$$

Note that $\hat{\phi}_\gamma(t)$ is positive but $\hat{\tilde{\phi}}_s(t)$ can be negative [2], while $\tilde{\phi}_s(t)$ is positive. We therefore replace the negative estimates $\hat{\tilde{\phi}}_s(t)$ with the initial estimates from Eq. (13).

We alternatingly estimate $\tilde{\mathbf{a}}$ and $\left\{\tilde{\phi}_s(t),\phi_\gamma(t)\right\}_{t=t_\beta}^{t_\beta+T_\beta-1}$ untile a certain number of iterations are executed. Finally, the RTF vector and the PSD of the target source have to be compensated for the reparameterization, and are given by $\hat{\mathbf{a}} = \frac{\mathbf{L}\hat{\tilde{\mathbf{a}}}}{\mathbf{L}\hat{\tilde{\mathbf{a}}}\mathbf{e}_1}$ and $\hat{\phi}_s = \frac{\hat{\tilde{\phi}}_s}{\hat{\mathbf{a}}^H\mathbf{\Gamma}^{-1}\hat{\mathbf{a}}}$, where $\mathbf{e}_1 = [1,0,\cdots,0]^T$.

## 4. TIME SEGMENTATION

In this section, we present the proposed algorithm for an adaptive time segmentation, where the number of time frames in a segment depends on how time-varying the RTF is. Due to the latency consideration, we consider that a maximum of

$T$ time frames can form a time segment. Hence, the maximum length of a segment is $T$. The minimum size is a single time frame. After the initialization step in the JMLE (Section 3.1), we have estimated reparameterized RTF vectors $\left\{\hat{\tilde{\mathbf{a}}}(t,k)\right\}_{t=t_\beta}^{t_\beta+T-1}$ for the $T$ time frames that could potentially be part of segment $\beta$, which will start at time frame $t_\beta$ (i.e., the first time frame after the previous time segment $\beta-1$). By analyzing the distance between these roughly estimated RTF vectors, we can find all time frames that should fall into the $\beta$-th time segment.

As the RTF estimation error is often expressed using the Hermitian angle, it would be natural to use this as a metric to determine the distance between two RTF vectors $\left\{\hat{\tilde{\mathbf{a}}}(i,k), \hat{\tilde{\mathbf{a}}}(j,k)\right\}$. We consider the $i$-th TF and $j$-th TF to belong to the same time segment if the Hermitian angle satisfies

$$\operatorname{acos}\left(\frac{\left|\hat{\tilde{\mathbf{a}}}(i,k)^H\hat{\tilde{\mathbf{a}}}(j,k)\right|}{\left\|\hat{\tilde{\mathbf{a}}}(i,k)\right\|_2\left\|\hat{\tilde{\mathbf{a}}}(j,k)\right\|_2}\right) < c_h, \tag{19}$$

where $c_h$ is a given constant threshold.

Alternatively, we could also construct a $M\times 2$ matrix $\mathbf{A}(i,j,k)=\left[\hat{\tilde{\mathbf{a}}}(i,k),\hat{\tilde{\mathbf{a}}}(j,k)\right]$ and analyze its second largest singular value $\sigma_2(i,j,k)$. In the ideal case that $\hat{\tilde{\mathbf{a}}}(i,k)$ and $\hat{\tilde{\mathbf{a}}}(j,k)$ are estimates of the same RTF vector without any errors, $\mathbf{A}(i,j,k)$ has rank 1 and $\sigma_2(i,j,k)=0$. Hence, we consider the $i$-th TF and the $j$-th TF to belong to the same time segment if

$$\sigma_2(i,j,k) < c_s, \tag{20}$$

where $c_s$ is 1:1 related to $c_h$ as we will show below.

We now show that these two methods are equivalent. Since $\operatorname{acos}(\cdot)$ is a monotonous decreasing function, Eq. (19) (i.e., the Hermitian angles) is equivalent to

$$\left|\hat{\tilde{\mathbf{a}}}(i,k)^H\hat{\tilde{\mathbf{a}}}(j,k)\right| > c, \tag{21}$$

where $\operatorname{acos}(c)=c_h$ and we used the fact that $\left\|\hat{\tilde{\mathbf{a}}}(t,k)\right\|_2=1$ for all $t$ and $k$. For the singular value method, the second largest singular value of $\mathbf{A}(i,j,k)$ is the square root of the second largest eigenvalue of

$$\mathbf{A}(i,j,k)^H\mathbf{A}(i,j,k) = \begin{bmatrix} 1 & \hat{\tilde{\mathbf{a}}}(i)^H\hat{\tilde{\mathbf{a}}}(j) \\ \hat{\tilde{\mathbf{a}}}(j)^H\hat{\tilde{\mathbf{a}}}(i) & 1 \end{bmatrix}, \tag{22}$$

which is given by $\sigma_2=\sqrt{1-\left|\hat{\tilde{\mathbf{a}}}(i,k)^H\hat{\tilde{\mathbf{a}}}(j,k)\right|}$. Hence, Eq. (20) is also equivalent to Eq. (21) with $\sqrt{1-c}=c_s$.

Note that if the $i$-th TF and the $j$-th TF belong to the same time segment, the inner products $\left|\hat{\tilde{\mathbf{a}}}(i,k)^H\hat{\tilde{\mathbf{a}}}(j,k)\right|$ are close to one for all frequencies. Therefore, we average the inner products for all frequency bins to express the similarities between the $i$-th and the $j$-th time frames with a single quality, that is,

$$B(i,j)=\frac{\sum_{k=1}^{K}\left|\hat{\tilde{\mathbf{a}}}(i,k)^H\hat{\tilde{\mathbf{a}}}(j,k)\right|}{K}. \tag{23}$$

Furthermore, by assuming the source does not change to other positions in between the $i$-th TF and the $j$-th TF when $B(i,j)$ is sufficiently large, the time frames in between them should also belong to the same time segment.

We assume the time segments before the $\beta$-the time segment have been determined (i.e., $\{t_i,T_i\}_{i=1}^{\beta-1}$ is known). Since $t_\beta=t_{\beta-1}+T_{\beta-1}$, $t_\beta$ is known. We only need to estimate the length $T_\beta$ of the $\beta$-th time segment. We first calculate $B(t_\beta,t_\beta+j-1)$ for $j=1,\cdots,T$. Then, we find $T_\beta$ by

$$\max\{j|B(t_\beta,t_\beta+j-1)>c, j=1,\cdots,T\}. \tag{24}$$

We then execute the JMLE algorithm using time frames from $t_\beta$ to $t_\beta+T_\beta-1$ jointly to estimate the RTF vector for the $\beta$-th time segment and the PSDs for time frames from $t_\beta$ to $t_\beta+T_\beta-1$.

The JMLE method combined with the adaptive time segmentation method is summarized in Algorithm 1.

**Algorithm 1:** TS-JMLE

**Input:** $\left\{\hat{\mathbf{\Phi}}_{\mathbf{y}}(t)\right\}_{t=t_\beta}^{t_\beta+T-1}$, $\hat{\mathbf{\Gamma}}$,$c$,$IterN$

**Output:** $T_\beta$,$\hat{\mathbf{a}}(\beta,k)$ and $\left\{\hat{\phi}_s(t,k),\hat{\phi}_\gamma(t,k)\right\}_{t=t_\beta}^{t_\beta+T_\beta-1}$

**for** *all* $k$, $t=t_\beta:t_\beta+T-1$ **do**
  Estimate $\hat{\tilde{\mathbf{a}}}(t,k)$, $\hat{\tilde{\phi}}_s(t)$ and $\hat{\phi}_\gamma(t)$ using Eqs. (12) to (14).
Calculate $B(t_\beta,j)$ for $j=t_\beta:t_\beta+T-1$ using Eq. (23);
Estimate $T_\beta$ by Eq. (24).
**for** *all* $k$ **do**
  **for** *iter=1:IterN* **do**
    Calculate $\mathbf{P}(\beta)=\sum_{t=t_\beta}^{t_\beta+T_\beta-1}\frac{\hat{\tilde{\phi}}_s(t)}{\hat{\phi}_\gamma(t)+\hat{\tilde{\phi}}_s(t)}\frac{1}{\hat{\phi}_\gamma(t)}\hat{\mathbf{P}}_{\mathbf{w}}(t)$
    Estimate $\hat{\tilde{\mathbf{a}}}(\beta)$ using the principal eigenvector of $\mathbf{P}(\beta)$.
    Estimate $\hat{\tilde{\phi}}_s(t)$ and $\hat{\phi}_\gamma(t)$ for $t=t_\beta,\cdots,t_\beta+T_\beta-1$ using Eqs. (17) and (18).
  Estimate $\hat{\mathbf{a}}(\beta)$ and $\hat{\phi}_s(t)$ by $\hat{\mathbf{a}}=\frac{\mathbf{L}\hat{\tilde{\mathbf{a}}}}{\mathbf{L}\hat{\tilde{\mathbf{a}}}\mathbf{e}_1}$ and $\hat{\phi}_s=\frac{\hat{\tilde{\phi}}_s}{\hat{\mathbf{a}}^H\mathbf{\Gamma}^{-1}\hat{\mathbf{a}}}$.

## 5. EXPERIMENTS

To evaluate the performance of the proposed method, we simulate the microphone signals by convolving the speech signal from the TIMIT data base [13] with the recorded room impulse responses (RIRs) from [14]. The setup for recording the RIRs is shown in Fig. 2, where 8 microphones are placed in a line with inter distance of 8 cm. The sound source is placed at a distance of 2 m from the center of the microphone array at different angles. We also add white Gaussian noise to the reverberant signals to simulate the microphone self noise even though the used JMLE method assumes the signals are noise free. The target signal-to-self noise ratio (SNR) is set to 50 dB, which is calculated over the whole time duration since the target signal is non-stationary. The noisy microphone signals are sampled at a rate of 16 kHz and processed by the STFT procedure. That is, we use the square-root Hann window with 50% overlap between adjacent sub-time frames and an FFT, both with a length of 512 samples (32 ms). Note that each time frame consists of $L_s = 40$ overlapping sub-time frames and thus has a duration of 0.64 s. The speed of sound is set to 344 m/s. The reverberation time is 0.61s. The threshold $c$ is set to 0.6 in the experiments based on some initial experiments.

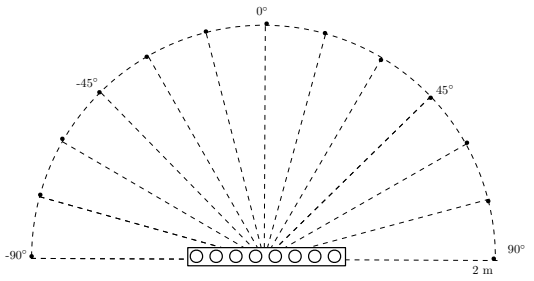



**Fig. 2**: Geometric setup for the real RIRs.

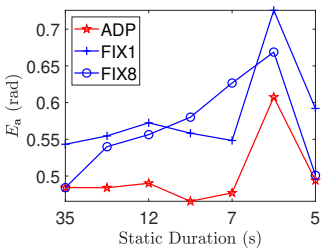
To compute the performance between the proposed and reference methods, we use the averaged Hermitian angle error (in rad) [15]

$$E_{\mathbf{a}} = \frac{\sum_{t=1}^{N}\sum_{k=1}^{K/2+1} \operatorname{acos}\left(\frac{\left|\mathbf{a}^H(t,k)\hat{\mathbf{a}}(t,k)\right|}{\left\|\mathbf{a}^H(t,k)\right\|_2\|\hat{\mathbf{a}}(t,k)\|_2}\right)}{N\,(K/2+1)}, \tag{25}$$

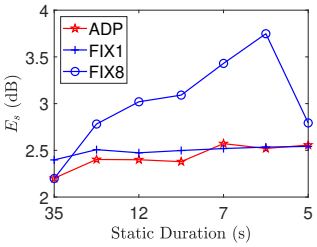
with $N$ the total number of time frames across all segments. Note that we average the errors over time frames instead of over time segments because the estimated time segments might have different duration. For the PSDs estimates, we use the symmetric log-error distortion measure [16]

$$E_i = \frac{10\sum_{t=1}^{N}\sum_{k=1}^{K/2+1}\left|\log\left(\frac{\phi_i(t,k)}{\hat{\phi}_i(t,k)}\right)\right|}{N\,(K/2+1)}, \tag{26}$$

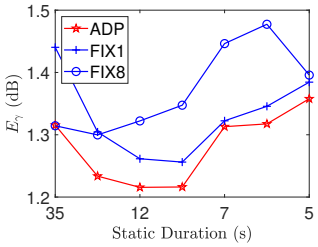
with $i \in \{s, \gamma\}$.




(a) RTF estimation error vs static duration at each position.




(b) Target source PSD estimation error vs static duration at each position.




(c) Late reverberation PSD estimation error vs static duration at each position.

**Fig. 3**: Performance comparison of the JMLE method combined with adaptive time segmentation and fix time segmentation.

In Fig. 3, we show the estimation performance comparison of the JMLE method combined with different time segmentation strategies. 'ADP' denotes our proposed adaptive time segmentation method. The maximum size $T$ of a time segment is set to 8. 'FIX1' denotes considering a single time frame (TF) as a time segment (TS), and 'FIX8' denotes considering every 8 TFs as a TS. The speech signal has a duration of 35 s. In the experiment, we simulate the time varying RTF by changing the source position from 0° to $(k-1)\times 15°$ by 15° every $\frac{35}{k}$ seconds. As $k$ increases from 1 to 7, the duration with which the source stays at the same position thus decreases from 35 s to $\frac{35}{7} = 5$ s along the x-axis in the graphs in Figure 3. For the RTF estimation error, the proposed ADP has the smallest error, which is about 0.1 rad smaller than FIX1 for different static time durations. The error for FIX8 fluctuates, but is always larger than ADP except for a static duration of 35 s and 5 s when ADP and FIX8 are approximately equal. For the source being static at 0° for 35s, all TFs share the same RTF. Therefore, ADP gives us the maximum size of the time segment, which is equal to FIX8 (considering 8 TFs as a TS) and much better than FIX1 (considering 1 TF as a TS). For the source staying at each position for 5 s, since the duration of a TF is 0.64 s, each TS contains about $\frac{5}{0.64} \approx 8$ TFs. therefore, the error of FIX8 is also close to ADP in this case. For the errors of the target PSD and the late reverberation PSD, ADP also has the best performance.

## 6. CONCLUSIONS

We presented an algorithm to obtain an optimal adaptive time segmentation and combined this with our previously published joint maximum likelihood estimator (JMLE) for jointly estimating the RTF, source PSD and late reverberation PSD of a single source in a reverberant environment. We proved that comparing the Hermitian angle of two RTF estimates to a threshold is equivalent to comparing the second largest singular value of the matrix combining these two RTF estimates or their inner product. We thus provided a thresholding method based on averaged inner products over all frequency bins. The JMLE combined with our adaptive time segmentation outperforms the JMLE combined with fixed time segmentation.

## 7. REFERENCES


[1] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, "Robust joint estimation of multimicrophone signal model parameters," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 7, pp. 1136–1150, 2019.

[2] C. Li, J. Martinez, and R. C. Hendriks, "Joint Maximum Likelihood Estimation of Microphone Array Parameters for a Reverberant Single Source Scenario," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 695–705, 2023.

[3] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum likelihood PSD estimation for speech enhancement in reverberation and noise," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 9, pp. 1599–1612, 2016.

[4] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 544–548.

[5] B. Schwartz, S. Gannot, and E. A. Habets, "Two model-based EM algorithms for blind source separation in noisy environments," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 11, pp. 2209–2222, 2017.

[6] I. Kodrasi and S. Doclo, "Analysis of eigenvalue decomposition-based late reverberation power spectral density estimation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 6, pp. 1106–1118, 2018.

[7] ——, "Joint Late Reverberation and Noise Power Spectral Density Estimation in a Spatially Homogeneous Noise Field," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 441–445.

[8] M. Tammen, S. Doclo, and I. Kodrasi, "Joint Estimation of RETF Vector and Power Spectral Densities for Speech Enhancement Based on Alternating Least Squares," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 795–799.

[9] Y. Laufer and S. Gannot, "Scoring-Based ML Estimation and CRBs for Reverberation, Speech, and Noise PSDs in a Spatially Homogeneous Noise Field," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 61–76, 2020.

[10] P. Hoang, Z.-H. Tan, J. M. de Haan, and J. Jensen, "Joint Maximum Likelihood Estimation of Power Spectral Densities and Relative Acoustic Transfer Functions for Acoustic Beamforming," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6119–6123.

[11] R. C. Hendriks, R. Heusdens, and J. Jensen, "Adaptive time segmentation for improved speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 14, no. 6, pp. 2064–2074, 2006.

[12] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 4, pp. 692–730, 2017.

[13] J. S. Garofolo, L. F. Lamel, W. M. Fisher, D. S. Pallett, N. L. Dahlgren, V. Zue, and J. G. Fiscus, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," 1993.

[14] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Proc. IEEE Int. Workshop Acoust. Signal Enhanc.*, Sept. 2014.

[15] R. Varzandeh, M. Taseska, and E. A. P. Habets, "An iterative multichannel subspace-based covariance subtraction method for relative transfer function estimation," in *Proc. IEEE Hands-free Speech Commun. Microphone Arrays*, 2017, pp. 11–15.

[16] R. C. Hendriks, J. Jensen, and R. Heusdens, "DFT domain subspace based noise tracking for speech enhancement," in *Proc. Interspeech*, 2007, pp. 830–833.