Stochastic Analysis of Time-Difference and Doppler Estimates for Audio Signals

Gabrielle Flood(✉), Anders Heyden, and Kalle Åström

Centre for Mathematical Sciences, Lund University, Lund, Sweden
{gabrielle.flood,anders.heyden,kalle.astrom}@math.lth.se

Abstract. Pairwise comparison of sound and radio signals can be used to estimate the distance between two units that send and receive signals. In a similar way it is possible to estimate differences of distances by correlating two received signals. There are essentially two groups of such methods, namely methods that are robust to noise and reverberation, but give limited precision and sub-sample refinements that are more sensitive to noise, but also give higher precision when they are initialized close to the real translation. In this paper, we present stochastic models that can explain the precision limits of such sub-sample time-difference estimates. Using these models new methods are provided for precise estimates of time-differences as well as Doppler effects. The developed methods are evaluated and verified on both synthetic and real data.

Keywords: Time-difference of arrival · Sub-sample methods
Doppler effect · Uncertainty measure

1 Introduction

Audio and radio sensors are increasingly used in smartphones, tablet PC's, laptops and other everyday tools. They also form the core of internet-of-things, e.g. small low-power units that can run for years on batteries or use energy harvesting to run for extended periods of time. If the locations of the sensing units are known, they can be used as an ad-hoc acoustic or radio sensor network. There are several interesting cases where such sensor networks can come into use. One such application is localization, cf. [5–7,9]. Another possible usage is beam-forming, i.e. to improve sound quality, [2]. Using a sensor network one can also determine who spoke when through speaker diarisation, [1]. If the sensor positions are unknown or if they are only known to a certain accuracy, the performance of such use-cases are inferior as is shown in [18]. It is, however, possible to perform automatic calibration, i.e. to estimate both sender and receiver positions, even without any prior information, as illustrated in Figs. 1 and 2. This can be done up to a choice of coordinate system, [8,12,13,19,22], thus providing accurate sensor positions for improved use. A key component for all of these methods is the process of obtaining and assessing estimates of e.g. time-difference of arrival

M. De Marsico et al. (Eds.): ICPRAM 2018, LNCS 11351, pp. 116–138, 2019.
https://doi.org/10.1007/978-3-030-05499-1_7

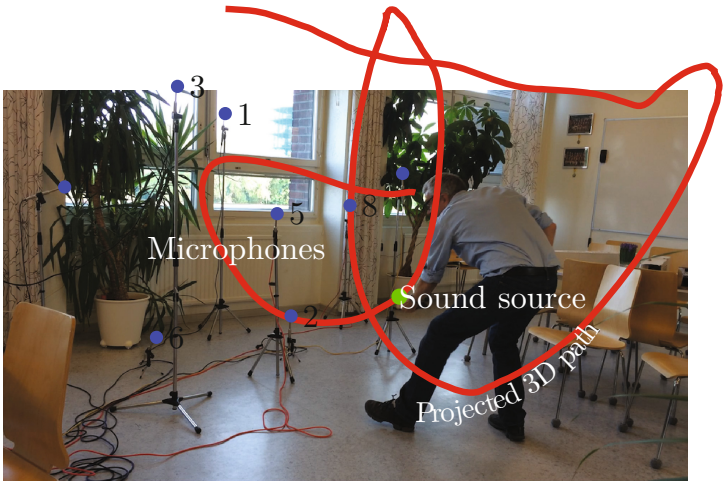

Fig. 1. Precise time-difference of arrival estimation can be used for many purposes, e.g. diarization, beam-forming, positioning and anchor free node calibration. The figure illustrates its use for anchor free node calibration, sound source movement and room reconstruction. The image is taken from [10].

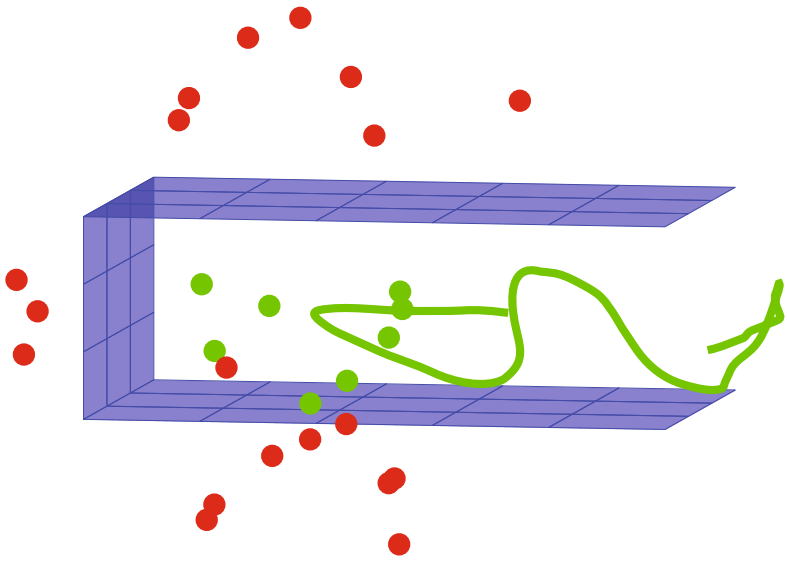

Fig. 2. The figure examplifies one usage of precise time-difference of arrival estimation. The image illustrates the estimated microphone positions (dots), estimated mirrored microphone positions (dots) and sound source motion (solid curve) from Fig. 1. The estimated reflective planes are also shown in the figure. These three planes correspond to the floor, the ceiling and the wall. The image is taken from [10].

of transmitted signals as they arrive in pairs of sensors. In this paper the focus is primarily on acoustic signals, but the same principles are useful for the analysis of radio signals [4].

All of these applications depend on accurate methods to extract features from the sound (or radio) signals. The most common feature is the time-difference-of-arrival, which is then used for subsequent processing. For applications, it is important to find as precise estimates as possible. In [23] time-difference estimates were improved using sub-sample methods. It was also shown empirically that the estimates of the receiver-sender configurations were improved by this. However, no analysis of the uncertainties of the sub-sample time-differences was provided.

This paper is an extended version of [10]. The main content is thus similar. However this version has been developed and is more thorough. E.g. the derivations in Sect. 3.1 have been extended, a comparison between different models has been added, see Sects. 3 and 4.1, and the experiments on real data in Sect. 4.2 have been changed and improved. In addition we have also performed stochastic analysis for the real data experiments. This is presented in Sect. 4.2. Then follows Sect. 4.3 which is partly new. Furthermore, most of the figures have been updated, even if a few remain from the original paper.

The main contributions of [10] and this paper are:

- A scheme for computing time-difference estimates and for estimating the precision of these estimates.
- A method to estimate minute Doppler effects, which is motivated by an experimental comparison between the models.
- An extension of the framework to capture and estimate amplitude differences in the signals.
- An evaluation on synthetic data to evince the validity of the models and provide knowledge of when the method fails.
- An evaluation on real data which demonstrates that the estimates for time-difference, minute Doppler effects and the amplitude changes contain relevant information. This is shown for speeds as small as 0.1 m/s.

2 Modeling Paradigm

2.1 Measurement and Error Model

In this paper, discretely sampled signals are studied. These could e.g. be audio or radio signals. Here, the sampling rate is assumed to be known and constant. Furthermore, we assume that the measured signal y has been ideally sampled after which noise – e.g. from the receivers – has been added, s.t.

$$y(k) = Y(k) + e(k)\,. \tag{1}$$

The original, continuous signal is denoted $Y : \mathbb{R} \mapsto \mathbb{R}$ and the noise, which is a discrete stationary stochastic process, is denoted e.

Let the set of functions $Y : \mathbb{R} \to \mathbb{R}$ that are (i) continuous (ii) square integrable and (iii) with a Fourier transform equal to zero outside $[-\pi, \pi]$ be denoted $\mathbb{B}$. Furthermore, denote the set of discrete, square integrable functions $y : \mathbb{Z} \to \mathbb{R}$ by ℓ. Now, define the discretization operator $D : \mathbb{B} \to \ell$ by

$$y(i) = D(Y)(i) = Y(i). \tag{2}$$

Moreover, we introduce the interpolation operator $I_g : \ell \to \mathbb{B}$, as

$$Y(x) = I_g(y)(x) = \sum_{i=-\infty}^{\infty} g(x-i)y(i)\,. \tag{3}$$

It has been shown that interpolation using the normalized sinc function, i.e. with $g(x) = \mathrm{sinc}(x)$, restores a sampled function for functions in $\mathbb{B}$, see [20] Thus, we call $I_{\mathrm{sinc}} : \ell \to \mathbb{B}$ the ideal interpolation operator and we have that

$$I_{\mathrm{sinc}}(D(Y)) = Y\,. \tag{4}$$

In the same way other interpolation methods can be expressed similarly. E.g. we obtain Gaussian interpolation by changing sinc in the expression above to

$$G_a(x) = \frac{1}{\sqrt{2\pi a^2}} \mathrm{e}^{x^2/(2a^2)}. \tag{5}$$

2.2 Scale-Space Smoothing and Ideal Interpolation

A measured and interpolated signal is often smoothed for two reasons. Firstly, there is often more signal as compared to noise for lower frequencies, whereas for higher frequencies there is usually less signal in relation to noise. Therefore smoothing can be used in order to remove some of the noise, while keeping most of the signal.

Secondly, patterns in a more coarse scale are easier captured after smoothing has been applied, [15]. A Gaussian kernel G_{a_2}, with standard deviation a_2, has been used for the smoothing. We will also refer to a_2 as the *smoothing parameter*.

Given a sampled signal y, the ideally interpolated and smoothed signal can be written as

$$Y(x) = (G_{a_2} * I_{\mathrm{sinc}}(y))(x) = I_{G_{a_2} * \mathrm{sinc}}(y)(x). \tag{6}$$

If a_2 is large enough the approximation $G_{a_2} * \mathrm{sinc} \approx G_{a_2}$ holds. Thus, one can use interpolation with the Gaussian kernel as an approximation for ideal interpolation followed by Gaussian smoothing, [3], s.t.

$$Y(x) = I_{G_{a_2} * \mathrm{sinc}}(y)(x) \approx I_{G_{a_2}}(y)(x). \tag{7}$$

What *large enough* means will be studied in Sect. 4.1.

Moreover, we will later use the fact that discrete w.s.s. Gaussian noise interpolates to continuous w.s.s. Gaussian noise, as is shown in [3].

3 Time-Difference and Doppler Estimation

Assume that we have two signals, $W(t)$ and $\bar{W}(t)$. The signals are measured and interpolated as described above. Also assume that the two signals are similar, but with one e.g. translated and compressed in the time domain. This could occur when two different receivers pick up an audio signal from a single sender. Then the second signal can be obtained from the other and a few parameters. We describe the relation as follows

$$W(t) = \bar{W}(\alpha t + h), \tag{8}$$

where h describes the time-difference of arrival, or translation in the signals. In a setup where the sound source has equal distance to both microphones $h = 0$. The second parameter, α, is a Doppler factor. This parameter is needed for example if the sound source or the microphones are moving. For a stationary setup $\alpha = 1$.

When the two microphones pick up the signals these are disturbed by Gaussian w.s.s. noise. Thus, the received signals can be written

$$V(t) = W(t) + E(t) \quad \text{and} \quad \bar{V}(t) = \bar{W}(t) + \bar{E}(t)\,. \tag{9}$$

Here, $E(t)$ and $\bar{E}(t)$ denotes the two independent noise signals after interpolation.

Assume that the signals V and $\bar{V}$ are given. Also, denote by $\boldsymbol{z} = \begin{bmatrix} z_1 & z_2 \end{bmatrix}^T = \begin{bmatrix} h & \alpha \end{bmatrix}^T$, the vector of unknown parameters. Then, the parameters for which (8) is true can be estimated by the $\boldsymbol{z}$ that minimizes the integral

$$G(\boldsymbol{z}) = \int_t (V(t) - \bar{V}(z_2 t + z_1))^2 \,\mathrm{d}t\,. \tag{10}$$

Comparing with Cross Correlation. If we only estimate a time delay h, the minimization of the error function (10) would in practice be the same as maximizing the cross correlation of V and $\bar{V}$. The cross-correlation for real signals is defined as

$$(V \star \bar{V})(h) = \int_t V(t)\bar{V}(t + h)\,\mathrm{d}t\,. \tag{11}$$

Thus, the h that maximize this cross-correlation is given by

$$\mathrm{argmax}_h (V \star \bar{V})(h) = \mathrm{argmax}_h \int_t V(t)\bar{V}(t + h)\,\mathrm{d}t\,. \tag{12}$$

If we expand the error function (10), while neglecting the Doppler factor we obtain the minimizer

$$\begin{aligned} \mathrm{argmin}_h \int_t (V(t) - \bar{V}(t + h))^2 \,\mathrm{d}t &= \mathrm{argmin}_h \int_t (V(t))^2 + (\bar{V}(t + h))^2 - 2V(t)\bar{V}(t + h)\,\mathrm{d}t \\ &= \mathrm{argmin}_h \int_t -2V(t)\bar{V}(t + h)\,\mathrm{d}t = \mathrm{argmax}_h \int_t V(t)\bar{V}(t + h)\,\mathrm{d}t. \end{aligned} \tag{13}$$

Note that since we integrate over t, the integral $\int_t (\bar{V}(t+h))^2 \, \mathrm{d}t$ is almost constant, ignoring edge effects.

We choose to use (10) for estimation of the parameters since it is simple to expand and is valid even if we add more parameters.

3.1 Estimating the Standard Deviation of the Parameters

If $\boldsymbol{z}_T = \begin{bmatrix} h_T & \alpha_T \end{bmatrix}^T$ is the "true" parameter for the data and $\hat{\boldsymbol{z}}$ is the parameter that has been estimated by minimizing (10), the estimation error can be expressed as

$$X = \hat{\boldsymbol{z}} - \boldsymbol{z}_T. \tag{14}$$

Assume, without loss of generality, that $\boldsymbol{z_T} = \begin{bmatrix} 0 & 1 \end{bmatrix}^T$. The standard deviation of $\hat{\boldsymbol{z}}$ will be the same as the standard deviation of X and the mean of those two will only differ by $\boldsymbol{z}_T$. Thus, it is sufficient to study X to get statistical information about the estimate $\hat{\boldsymbol{z}}$.

Linearizing $G(\boldsymbol{z})$ around the true displacement $\boldsymbol{z}_T = \begin{bmatrix} 0 & 1 \end{bmatrix}^T$ gives

$$G(\boldsymbol{z}) \approx F(X) = \frac{1}{2} X^T a X + bX + f, \tag{15}$$

with

$$a = \nabla^2 G(\boldsymbol{z}_T)\,, \qquad b = \nabla G(\boldsymbol{z}_T)\,, \qquad f = G(\boldsymbol{z}_T)\,. \tag{16}$$

Using (9) and (8), we get

$$\begin{aligned} f &= G(\begin{bmatrix} 0 & 1 \end{bmatrix}^T) = \int_t (V(t) - \bar{V}(1 \cdot t - 0))^2 \, \mathrm{d}t = \int_t (W(t) + E(t) - (\bar{W}(t) + \bar{E}(t)))^2 \, \mathrm{d}t \\ &= \int_t (E - \bar{E})^2 \, \mathrm{d}t = \int_t E^2 + 2E\bar{E} + \bar{E}^2 \, \mathrm{d}t. \end{aligned} \tag{17}$$

To find the coefficients a and b we first calculate the derivatives $\nabla G(\boldsymbol{z})$ and $\nabla^2 G(\boldsymbol{z})$.

$$\begin{aligned} \nabla G &= \begin{bmatrix} \int_t 2(V(t) - \bar{V}(\alpha t + h)) \cdot (-\bar{V}'(\alpha t + h)) \, \mathrm{d}t \\ \int_t 2(V(t) - \bar{V}(\alpha t + h)) \cdot (-\bar{V}'(\alpha t + h) \cdot t) \, \mathrm{d}t \end{bmatrix} \\ &= -2 \begin{bmatrix} \int_t (W(t) + E(t) - \bar{W}(\alpha t + h) - \bar{E}(\alpha t + h)) \cdot (\bar{W}'(\alpha t + h) + \bar{E}'(\alpha t + h)) \, \mathrm{d}t \\ \int_t (W(t) + E(t) - \bar{W}(\alpha t + h) - \bar{E}(\alpha t + h)) \cdot (\bar{W}'(\alpha t + h) + \bar{E}'(\alpha t + h)) \cdot t \, \mathrm{d}t \end{bmatrix}. \end{aligned} \tag{18}$$

Inserting the true displacement $\boldsymbol{z}_T$, at the point of linearization, gives

$$\begin{aligned} b =& \nabla G(\boldsymbol{z}_T) \\ =& -2 \begin{bmatrix} \int_t (W(t) + E(t) - \bar{W}(t) - \bar{E}(t)) \cdot (\bar{W}'(t) + \bar{E}'(t)) \, \mathrm{d}t \\ \int_t (W(t) + E(t) - \bar{W}(t) - \bar{E}(t)) \cdot (\bar{W}'(t) + \bar{E}'(t)) \cdot t \, \mathrm{d}t \end{bmatrix} \\ =& -2 \begin{bmatrix} \int_t (E - \bar{E})(\bar{W}' + \bar{E}') \, \mathrm{d}t \\ \int_t (E - \bar{E})(\bar{W}' + \bar{E}') t \, \mathrm{d}t \end{bmatrix} = -2 \begin{bmatrix} \int_t E\bar{W}' + E\bar{E}' - \bar{E}\bar{W}' - \bar{E}\bar{E}' \, \mathrm{d}t \\ \int_t (E\bar{W}' + E\bar{E}' - \bar{E}\bar{W}' - \bar{E}\bar{E}') t \, \mathrm{d}t \end{bmatrix}. \end{aligned} \tag{19}$$

To simplify further computations, we introduce

$$\hat{\varphi} = E\bar{W}' + E\bar{E}' - \bar{E}\bar{W}' - \bar{E}\bar{E}', \tag{20}$$

such that

$$b = -2\begin{bmatrix} \int_t \hat{\varphi}\,\mathrm{d}t \\ \int_t t\hat{\varphi}\,\mathrm{d}t \end{bmatrix}. \tag{21}$$

Furthermore,

$$\begin{aligned}
&\nabla^2 G = \\
&\begin{bmatrix} \int_t -2\bar{V}'(\alpha t+h)\cdot(-\bar{V}'(\alpha t+h)) + 2(V(t)-\bar{V}(\alpha t+h))(-\bar{V}''(\alpha t+h))\,\mathrm{d}t \\ \int_t -2\bar{V}'(\alpha t+h)\cdot t\cdot(-\bar{V}'(\alpha t+h)) + 2(V(t)-\bar{V}(\alpha t+h))(-\bar{V}''(\alpha t+h)\cdot t)\,\mathrm{d}t \end{bmatrix} \cdots \\
&\begin{bmatrix} \int_t -2\bar{V}'(\alpha t+h)\cdot(-\bar{V}'(\alpha t+h))\cdot t + 2(V(t)-\bar{V}(\alpha t+h))\cdot(-\bar{V}''(\alpha t+h)\cdot t)\,\mathrm{d}t \\ \int_t -2\bar{V}'(\alpha t+h)\cdot t(-\bar{V}'(\alpha t+h))\cdot t + 2(V(t)-\bar{V}(\alpha t+h))\cdot(-\bar{V}''(\alpha t+h)\cdot t^2)\,\mathrm{d}t \end{bmatrix} \\
&= 2\begin{bmatrix} \int_t (\bar{V}'(\alpha t+h))^2 - V(t)\bar{V}''(\alpha t+h) + \bar{V}(\alpha t+h)\bar{V}''(\alpha t+h)\,\mathrm{d}t \\ \int_t t\cdot(\bar{V}'(\alpha t+h))^2 - t\cdot V(t)\bar{V}''(\alpha t+h) + t\cdot\bar{V}(\alpha t+h)\bar{V}''(\alpha t+h)\,\mathrm{d}t \end{bmatrix} \cdots \\
&\begin{bmatrix} \int_t \cdot(\bar{V}'(\alpha t+h))^2 - t\cdot V(t)\bar{V}''(\alpha t+h) + t\cdot\bar{V}(\alpha t+h)\bar{V}''(\alpha t+h)\,\mathrm{d}t \\ \int_t t^2\cdot(\bar{V}'(\alpha t+h))^2 - t^2\cdot V(t)\bar{V}''(\alpha t+h) + t^2\cdot\bar{V}(\alpha t+h)\bar{V}''(\alpha t+h) \end{bmatrix}.
\end{aligned} \tag{22}$$

Now, introducing the notation

$$\phi(\boldsymbol{z}) = (\bar{V}'(\alpha t+h))^2 - V(t)\bar{V}''(\alpha t+h) + \bar{V}(\alpha t+h)\bar{V}''(\alpha t+h), \tag{23}$$

we can write $\nabla^2 G$ shorter as

$$\nabla^2 G = \begin{bmatrix} \int_t \phi\,\mathrm{d}t & \int_t t\phi\,\mathrm{d}t \\ \int_t t\phi\,\mathrm{d}t & \int_t t^2\phi\,\mathrm{d}t \end{bmatrix}. \tag{24}$$

If we let $\hat{\phi}$ be the value of ϕ for $\boldsymbol{z}_T$

$$\begin{aligned}
\hat{\phi} =& \phi(\boldsymbol{z}_T) = (\bar{W}'(t) + \bar{E}'(t))^2 - (W(t) + E(t))(\bar{W}''(t) + \bar{E}''(t)) \\
&+ (\bar{W}(t) + \bar{E}(t))(\bar{W}''(t) + \bar{E}''(t)) \\
=& (\bar{W}')^2 + 2\bar{W}'\bar{E}' + (\bar{E}')^2 - E\bar{W}'' - E\bar{E}'' + \bar{E}\bar{W}'' + \bar{E}\bar{E}''
\end{aligned} \tag{25}$$

we get

$$a = \nabla^2 G(\boldsymbol{z}_T) = \begin{bmatrix} \int_t \hat{\phi}\,\mathrm{d}t & \int_t t\hat{\phi}\,\mathrm{d}t \\ \int_t t\hat{\phi}\,\mathrm{d}t & \int_t t^2\hat{\phi}\,\mathrm{d}t \end{bmatrix}. \tag{26}$$

We also have that $F(X) = 1/2 \cdot X^T aX + bX + f$. To minimize this error function, we find the X for which the derivative of $F(X)$ is zero. Since a is symmetric we get

$$\nabla F(X) = aX + b = 0 \quad \Leftrightarrow \quad X = g(a,b) = -a^{-1}b. \tag{27}$$

In the calculations below, we assume that a is invertible.

Now we would like to find the mean and covariance of X. For this, Gauss' approximation formulas are used. If we denote the expected value of a and b with $\mu_a = \mathbf{E}[A]$ and $\mu_b = \mathbf{E}[b]$ respectively the expected value of X can be approximated to

$$\begin{aligned}\mathbf{E}[X] =&\mathbf{E}[g(a,b)] \approx \mathbf{E}[g(\mu_a,\mu_b) + (a-\mu_a)g'_a(\mu_a,\mu_b) + (b-\mu_b)g'_b(\mu_a,\mu_b)]\\ =&g(\mu_a,\mu_b) + (\mathbf{E}[a]-\mu_a)g'_a(\mu_a,\mu_b) + (\mathbf{E}[b]-\mu_b)g'_b(\mu_a,\mu_b)\\ =&g(\mu_a,\mu_b) = -\mu_a^{-1}\mu_b = -\mathbf{E}[a]^{-1}\mathbf{E}[b].\end{aligned} \tag{28}$$

In a similar manner the covariance of X is

$$\begin{aligned}\mathbf{C}[X] =&\mathbf{C}[g(a,b)] \approx g'_a(\mu_a,\mu_b)\mathbf{C}[a]g'_a(\mu_a,\mu_b)^T + g'_b(\mu_a,\mu_b)\mathbf{C}[b]g'_b(\mu_a,\mu_b)^T\\ &+ 2g'_a(\mu_a,\mu_b)\mathbf{C}[a,b]g'_b(\mu_a,\mu_b)^T,\end{aligned} \tag{29}$$

where $\mathbf{C}[a,b]$ denotes the cross-covariance between a and b. For further computations $g'_a(a,b)$, $g'_b(a,b)$, $\mathbf{E}[a]$, $\mathbf{E}[b]$, $\mathbf{C}[b]$ and $\mathbf{C}[a,b]$ are needed.

By computing the expected value of $\hat{\varphi}$

$$\begin{aligned}\mathbf{E}[\hat{\varphi}] =&\mathbf{E}[E\bar{W}' + E\bar{E}' - \bar{E}\bar{W}' - \bar{E}\bar{E}']\\ =&\mathbf{E}[E]\bar{W}' + \mathbf{E}[E]\mathbf{E}[\bar{E}'] - \mathbf{E}[\bar{E}]\bar{W}' - \mathbf{E}[\bar{E}]\mathbf{E}[\bar{E}'] = 0\end{aligned} \tag{30}$$

we get

$$\mathbf{E}[b] = \mathbf{E}\left[-2\begin{bmatrix}\int_t \hat{\varphi}\ dt\\ \int_t t\hat{\varphi}\,\mathrm{d}t\end{bmatrix}\right] = -2\begin{bmatrix}\int_t \mathbf{E}[\hat{\varphi}]\,\mathrm{d}t\\ \int_t t\mathbf{E}[\hat{\varphi}]\,\mathrm{d}t\end{bmatrix} = -2\begin{bmatrix}\int_t 0\,\mathrm{d}t\\ \int_t t\cdot 0\,\mathrm{d}t\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix}. \tag{31}$$

In the second step of the computation of $\mathbf{E}[\hat{\varphi}]$ we have used the fact that for a weakly stationary process the process and its derivative at a certain time are uncorrelated, and thus $\mathbf{E}[\bar{E}\bar{E}'] = \mathbf{E}[\bar{E}]\mathbf{E}[\bar{E}']$, [16]. Hence,

$$\mathbf{E}[X] = -\mathbf{E}[a]^{-1}\mathbf{E}[b] = -\mathbf{E}[a]^{-1}\begin{bmatrix}0\\0\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix}. \tag{32}$$

For the partial derivative of $g(a,b)$ w.r.t. b we get [17]

$$g'_b(a,b) = \frac{\partial}{\partial b}\left(-a^{-1}b\right) = -(a^{-1})^T = -(a^T)^{-1} = -a^{-1} \tag{33}$$

and thus $g'_b(\mu_a,\mu_b) = -(\mathbf{E}[a])^{-1}$. Since $\mathbf{E}[b] = \mathbf{0}$, we get that $g'_a(\mu_a,\mu_b) = \mathbf{0}$, [17]. Hence the first and the last term in (29) cancel, leaving

$$\begin{aligned}\mathbf{C}[X] =&g'_b(\mu_a,\mu_b)\mathbf{C}[b]g'_b(\mu_a,\mu_b)^T = (-\mathbf{E}[a]^{-1})\mathbf{C}[b](-\mathbf{E}[a]^{-1})^T\\ =&\mathbf{E}[a]^{-1}\mathbf{C}[b](\mathbf{E}[a]^{-1})^T.\end{aligned} \tag{34}$$

To find the expected value of a the expected value of $\hat{\phi}$ is needed. This is obtained from

$$\begin{aligned}\mathbf{E}[\hat{\phi}] &= (\bar{W}')^2 + 2\bar{W}'\mathbf{E}[\bar{E}'] + \mathbf{E}[(\bar{E}')^2] - \bar{W}''\mathbf{E}[E] - \mathbf{E}[E]\mathbf{E}[\bar{E}''] + \bar{W}''\mathbf{E}[\bar{E}] + \mathbf{E}[\bar{E}\bar{E}''] \\ &= (\bar{W}')^2 + \mathbf{E}[(\bar{E}')^2] + \mathbf{E}[\bar{E}\bar{E}''] = (\bar{W}')^2. \end{aligned} \tag{35}$$

In the last equality we have used that $\mathbf{E}[\bar{E}\bar{E}''] = -\mathbf{E}[(\bar{E}')^2]$, [16]. Thus, the two last terms cancel out. The expected value of a is therefore

$$\mathbf{E}[a] = 2\begin{bmatrix} \int_t \mathbf{E}[\hat{\phi}]\,\mathrm{d}t & \int_t \mathbf{E}[t\hat{\phi}]\,\mathrm{d}t \\ \int_t \mathbf{E}[t\hat{\phi}]\,\mathrm{d}t & \int_t \mathbf{E}[t^2\hat{\phi}]\,\mathrm{d}t \end{bmatrix} = 2\begin{bmatrix} \int_t (\bar{W}')^2\,\mathrm{d}t & \int_t t(\bar{W}')^2\,\mathrm{d}t \\ \int_t t(\bar{W}')^2\,\mathrm{d}t & \int_t t^2(\bar{W}')^2\,\mathrm{d}t \end{bmatrix}. \tag{36}$$

Now, since the expected value of b is zero, the covariance of b is

$$\mathbf{C}[b] = (-2)^2 \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}, \tag{37}$$

with

$$\begin{aligned} C_{11} &= \mathbf{E}\left[\int_{t_1} \hat{\varphi}(t_1)\,\mathrm{d}t_1 \cdot \int_{t_2} \hat{\varphi}(t_2)\,\mathrm{d}t_2\right] \\ C_{12} &= \mathbf{E}\left[\int_{t_1} t_1\hat{\varphi}(t_1)\,\mathrm{d}t_1 \cdot \int_{t_2} \hat{\varphi}(t_2)\,\mathrm{d}t_2\right] \\ C_{21} &= \mathbf{E}\left[\int_{t_1} \hat{\varphi}(t_1)\,\mathrm{d}t_1 \cdot \int_{t_2} t_2\hat{\varphi}(t_2)\,\mathrm{d}t_2\right] \\ C_{22} &= \mathbf{E}\left[\int_{t_1} t_1\hat{\varphi}(t_1)\,\mathrm{d}t_1 \cdot \int_{t_2} t_2\hat{\varphi}(t_2)\,\mathrm{d}t_2\right]. \end{aligned} \tag{38}$$

Note that by changing the order of the terms in C_{12} it is clear that $C_{21} = C_{12}$. Furthermore, we obtain

$$\begin{aligned} C_{11} =& \mathbf{E}\left[\int_{t_1} \hat{\varphi}(t_1)\,\mathrm{d}t_1 \cdot \int_{t_2} \hat{\varphi}(t_2)\,\mathrm{d}t_2\right] \\ =& \mathbf{E}\Big[\Big(\int_{t_1} (E-\bar{E})(\bar{W}'+\bar{E}')\,\mathrm{d}t_1\Big) \cdot \Big(\int_{t_2} (E-\bar{E})(\bar{W}'+\bar{E}')\,\mathrm{d}t_2\Big)\Big] \\ =& \mathbf{E}\Big[\int_{t_1}\int_{t_2} (E(t_1)-\bar{E}(t_1))(\bar{W}'(t_1)+\bar{E}'(t_1))\cdot \\ & (E(t_2)-\bar{E}(t_2))(\bar{W}'(t_2)+\bar{E}'(t_2))\,\mathrm{d}t_1\mathrm{d}t_2\Big]. \end{aligned} \tag{39}$$

Denoting $\mathbf{E}[(E(t_1)-\bar{E}(t_1))(E(t_2)-\bar{E}(t_2))] = r_{E-\bar{E}}(t_1-t_2)$ and assuming that $\mathbf{E}[\bar{E}'(t_1)\bar{E}'(t_2)]$ is small gives

$$
\begin{aligned}
C_{11} =& \mathbf{E}\left[\int_{t_1} \hat{\varphi}(t_1)\,\mathrm{d}t_1 \cdot \int_{t_2} \hat{\varphi}(t_2)\,\mathrm{d}t_2\right] \\
=& \int_{t_1}\int_{t_2} \mathbf{E}[(E(t_1)-\bar{E}(t_1))(E(t_2)-\bar{E}(t_2))]\cdot(\bar{W}'(t_1)\bar{W}'(t_2) \\
&+ \bar{W}'(t_1)\mathbf{E}[\bar{E}'(t_2)] + \mathbf{E}[\bar{E}'(t_1)]\bar{W}'(t_2) + \mathbf{E}[\bar{E}'(t_1)\bar{E}'(t_2)])\,\mathrm{d}t_2\mathrm{d}t_1 \qquad (40)\\
\approx& \int_{t_1}\int_{t_2} r_{E-\bar{E}}(t_1-t_2)\bar{W}'(t_1)\bar{W}'(t_2)\,\mathrm{d}t_2\mathrm{d}t_1 \\
=& \int_{t_1} \bar{W}'(t_1)(\bar{W}' * r_{E-\bar{E}})(t_1)\,\mathrm{d}t_1.
\end{aligned}
$$

The time t is a deterministic quantity and the other elements in $\mathbf{C}[b]$ can be computed similarly. Finally we have

$$
\begin{aligned}
C_{11} &= \int_t \bar{W}'(t)(\bar{W}' * r_{E-\bar{E}})(t)\,\mathrm{d}t \\
C_{12} = C_{21} &= \int_t t\bar{W}'(t)(\bar{W}' * r_{E-\bar{E}})(t)\,\mathrm{d}t \qquad (41)\\
C_{22} &= \int_t t\bar{W}'(t)((t\bar{W}') * r_{E-\bar{E}})(t)\,\mathrm{d}t
\end{aligned}
$$

and through (34) we get an expression for the variance and thus also the standard deviation of X.

3.2 Expanding the Model

It is easy to change or expand the model (8) to contain more (or fewer) parameters. If we keep h and α and add an extra amplitude parameter γ, we get the model

$$W(t) = \gamma\bar{W}(\alpha t + h). \qquad (42)$$

The error integral (10) would then be changed accordingly and the optimization would instead be over over $\boldsymbol{z} = \begin{bmatrix} z_1 & z_2 & z_3 \end{bmatrix} = \begin{bmatrix} h & \alpha & \gamma \end{bmatrix}$.

The computations for achieving the estimations does in practice not get harder when we add more parameters. However, the analysis from the previous section gets more complex.

4 Experimental Validation

For validation we perform experiments on both real data and synthetic data. The purpose of using synthetic data is to demonstrate the validity of the model, but also to verify the approximations used. In the latter case we have studied at what signal-to-noise ratio the approximations are valid. Furthermore, to show that the parameter estimations contain useful information, we have done experiments on real data. This is well-known for time-difference, but less explored for the Doppler effects and amplitude changes.

4.1 Synthetic Data - Validation of Method

The model was first tested on simulated data in order to study when the approximations in the model derivation hold. The linearization using Gauss' approximation formula, e.g. (28) and (29), is one example of such approximations. Another is the usage of Gaussian interpolation as an approximation of ideal interpolation followed by convolution with a Gaussian, (7).

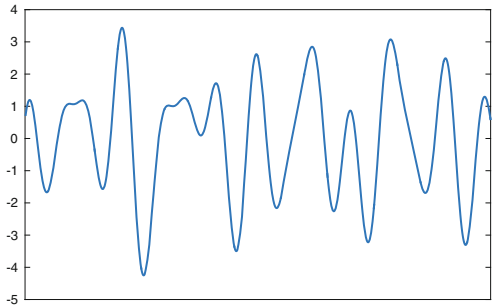

Fig. 3. The simulated signal that was used for the experimental validation. To achieve a more realistic signal noise of different levels was added later on. The plot is taken from [10].

To do these studies we compared the theoretical standard deviations of the parameters calculated according to Sect. 3.1 with empirically computed standard deviations. The agreement of these standard deviations makes us conclude that our approximations are valid.

First we simulated an original continuous signal $W(x)$, see Fig. 3. The second signal was then created according to (8) s.t. $\bar{W} = W(1/\alpha \cdot (x - h))$. The signals were ideally sampled after which Gaussian white discrete noise with standard deviation σ_n was added. After smoothing with a Gaussian kernel with standard deviation a_2 (see Sect. 2.2) the signals can be described by $V(t)$ and $\bar{V}(t)$ as before.

The two signals V and $\bar{V}$ were simulated anew 1000 times to investigate the effect of a_2 and σ_n. Each time the same original signals W and $\bar{W}$ were used, but with different noise realizations. Then, we computed the theoretical standard deviation of the parameter vector $\boldsymbol{z}$, $\sigma_{\boldsymbol{z}} = \begin{bmatrix}\sigma_h & \sigma_\alpha\end{bmatrix}$. This was done in accordance with the presented theory. We also computed an empirical standard deviation $\hat{\sigma}_{\boldsymbol{z}} = \begin{bmatrix}\hat{\sigma}_h & \hat{\sigma}_\alpha\end{bmatrix}$ from the 1000 different parameter estimations.

When studying the effect of a_2 the noise level was kept constant, with $\sigma_n = 0.03$. The translation was set to $h = 3.63$ and the Doppler factor was $\alpha = 1.02$. However, the exact numbers are unessential. While varying the smoothing parameter $a_2 \in [0.3, 0.8]$ the standard deviation was then computed according to the procedure above.

The results from these simulations can be seen in Fig. 4. When a_2 is below $a_2 \approx 0.55$ the theoretical values $\sigma_{\boldsymbol{z}}$ and the empirical values $\hat{\sigma}_{\boldsymbol{z}}$ do not agree,

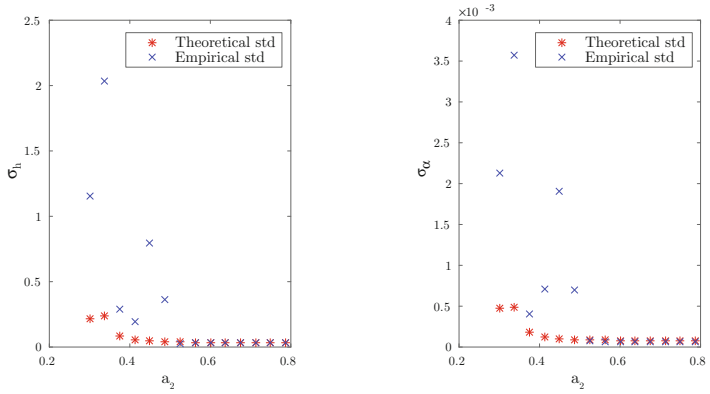

Fig. 4. The plots show the standard deviation of the parameters in $\boldsymbol{z}$ for different values of the smoothing parameter a_2. The stars ($*$) represent the theoretical values $\sigma_{\boldsymbol{z}}$ and the crosses (x) the empirical values $\hat{\sigma}_{\boldsymbol{z}}$. The left plot shows the results for the translation $z_1 = h$ and the right plot for the Doppler factor $z_2 = \alpha$. It is clear that the approximation is valid approximately when $a_2 > 0.55$. The plots are taken from [10].

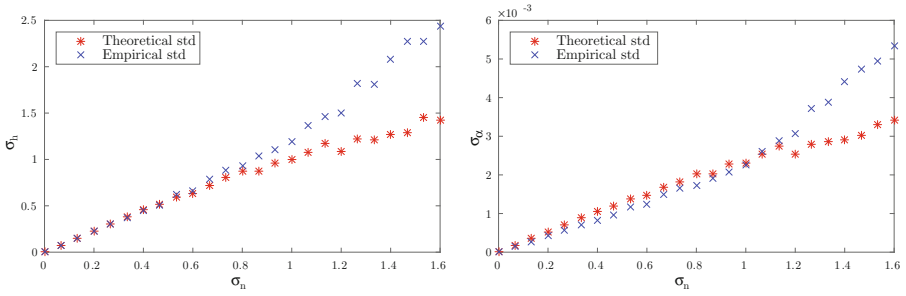

Fig. 5. The standard deviation of the translation (to the left) and Doppler factor (to the right) for different levels of noise in the signal. The stars ($*$) mark the theoretical values $\sigma_{\boldsymbol{z}}$ and the crosses (x) the empirical $\hat{\sigma}_{\boldsymbol{z}}$. For the translation the values agree for signals with a noise level up to $\sigma_n \approx 0.8$. For the Doppler factor the theoretical values follow the empirical values when $\sigma_n < 1.1$. The plots are taken from [10].

while they do for $a_2 > 0.55$. Therefore we draw the conclusion that the approximation (7) of ideal interpolation should only be used when $a_2 > 0.55$.

Secondly, the effect of changing the noise level was investigated. The smoothing parameters was fixed to $a_2 = 2$ and the translation and the Doppler factor were kept on the same level as before. Instead we varied the noise level s.t. $\sigma_n \in [0, 1.6]$. Then the standard deviations of the parameters $\sigma_{\boldsymbol{z}}$ and $\hat{\sigma}_{\boldsymbol{z}}$ were computed in the same way as in the previous section.

The results from this run can be seen in Fig. 5, with the results for the translation parameter h to the left and for the Doppler parameter α to the right. When σ_n is lower than $\sigma_n \approx 0.8$ the theoretical and empirical values for

the translation parameter are similar. For higher values of σ_n they do not agree. The same goes for the Doppler factor when the noise level is below $\sigma_n \approx 1.1$.

By this, we reason that noise with a standard deviation up to $\sigma_n \approx 0.8$ can be handled. The original signal W have an amplitude that varies between 1 and 3.5 and using the standard deviation of that signal, σ_W, we can compute the signal-to-noise ratio that the system can manage. We get the result

$$\text{SNR} = \frac{\sigma_W^2}{\sigma_n^2} \approx 4.7. \tag{43}$$

Comparing Different Models. In this paper we have chosen to work with the models (8) and (42). However, we have so far not presented any comparison between different models. To investigate this, we studied two models, namely (8), which we call model B and a slightly simpler model which we call model A,

$$W(x) = \bar{W}(x + h). \tag{44}$$

To begin with, we simulated data according to model A. We call this data A. During the simulation the standard deviation of the noise in the signals was set to $\sigma_n = 0.02$ and the smoothing parameter was $a_2 = 2.0$. Furthermore, we studied this data both using model A, i.e. by minimizing $\int_t (V(t) - \bar{V}(t+h))^2 \, \mathrm{d}t$ and using model B, see (10). The results can be seen in the first column (Data A) of Table 1.

Secondly, a similar test was made but this time we simulated data according to model B. We call this data B. We then studied this data using both model A and B. The results are shown in the second column (data B) of Table 1.

Table 1. Comparison between model A from (44) and model B from (8). Data A consists of signals with only translational differences while the second signal in data B is affected by both translation and a Doppler effect. The standard deviations for model B in the table regards the theoretical values that were derived in Section 3.1, and a similar analysis has been performed for model A.

		Data A	Data B
True values	Translation, h_T	3.63	3.63
	Doppler factor, α_T	1.00	1.02
Model A	Est. h, $\hat{h}^{(A)}$	3.63	13.4
	Std. of h, $\sigma_h^{(A)}$	$1.01 \cdot 10^{-2}$	$1.02 \cdot 10^{-2}$
Model B	Est. h, $\hat{h}^{(B)}$	3.63	3.66
	Std. of h, $\sigma_h^{(B)}$	$2.30 \cdot 10^{-2}$	$2.30 \cdot 10^{-2}$
	Est. α, $\hat{\alpha}^{(B)}$	1.00	1.02
	Std. of α, $\sigma_\alpha^{(B)}$	$5.32 \cdot 10^{-5}$	$5.33 \cdot 10^{-5}$

Studying the first column of Table 1 we see that model B estimates the parameters as good as model A – which in this case is the most correct model –

does. Though, for model B the standard deviation σ_h is more than twice as big as for model A.

In the second column of the table we see that since model A cannot estimate the Doppler effect, the translation parameter is erroneously estimated. The standard deviation σ_h is however still lower for model A. To minimize the error function model A estimates the translation such that the signal is fitted in the middle, see Fig. 6. This means that even though the standard deviation is low, the bias is high.

If we know that our collected data has only been affected by a translation it is clearly better to use model A. However, the loss for using a more simple model is larger on complex data than the loss for using a larger model for simple data. Thus, based on the results from Table 1 we conclude that it is better to use a larger model for the real data in the following section.

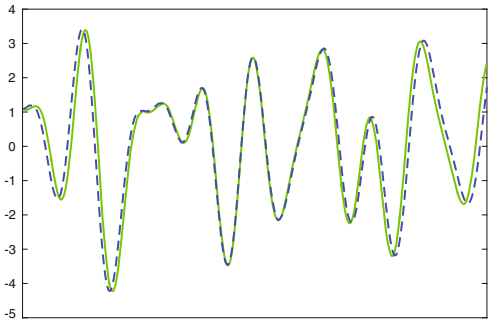

Fig. 6. The results after using model A on data B, where the second signal is affected both by a translation and Doppler effect. Since the model does not estimate any Doppler factor, the estimated translation will be biased. The two signals agree well in the middle, while there is a gap between them at the beginning and the end. This gap cannot be captured by a translation.

4.2 Real Data - Validation of Method

The experiments on real data were performed in an anaechoic chamber and the recording frequency was $f = 96\,\text{kHz}$. We used 8 T-Bone MM-1 microphones and these were connected to an audio interface (M-Audio Fast Track Ultra 8R) and a computer. Furthermore, the microphones were placed so that they spanned 3D, approximately 0.3–1.5 m away from each other. As a sound source we used a mobile phone which was connected to a small loudspeaker. The mobile phone was moved around in the room while playing a song.

We used the technique described in [22] and refined in [21] to achieve ground truth consisting of a 3D trajectory for the sound source path $s(t)$ and the 3D positions of the microphones $r_1, \ldots, r_8$. The method uses RANSAC algorithms which are based on minimal solvers [14] to find initial estimates of the sound trajectory and microphone positions. Then, these are refined using non-linear

optimization of a robust error norm, including a smooth motion prior, to reach the final estimates.

However, to make ground truth independent from the data that we used for testing we chose to only take data from microphone 3–8 into account during the first two thirds of the sound signal. Thus, by that we estimated $s(t)$ for certain t and $r_3, \ldots, r_8$. For the final third of the signal we added the information from microphone 1 and 2 as well, such that our solution would not drift compared to ground truth. By that we estimated the rest of $s(t)$, r_1 and r_2.

We only used data from microphone 1 and 2 for the validation of the method presented in this paper. The sound was played for around 29 s and the loudspeaker was constantly moving during this time. Furthermore, both the direction and the speed of the sound source changed.

Since our method assume a constant parameter $\boldsymbol{z}$ in a window we divided the recording into 2834 patches of 1000 samples each (i.e. about 0.01 s). Within these patches the parameters were approximately constant. Each of the patches could then be investigated and compared to ground truth separately. From ground truth we had a constant loudspeaker position $s^{(i)}$, its derivative $\frac{\partial s^{(i)}}{\partial t}(i)$ and the receiver positions r_1 and r_2 for each signal patch i.

Estimating the Parameters. If we call signal patch i from the first microphone $V^{(i)}(t)$ and let $\bar{V}^{(i)}(t)$ be the patch from the second microphone we can estimate the parameters using (8) to model the received signals.

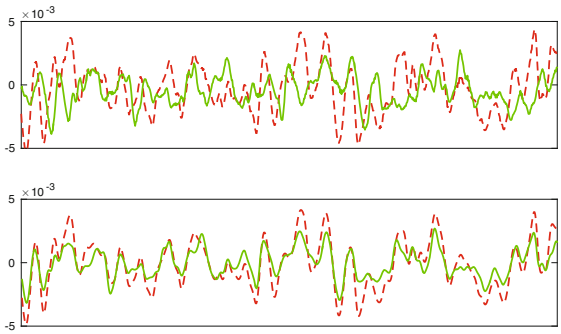

Fig. 7. The received signal patches at a certain time – the first signal in dashed (- -) and the second as solid (—). The top plot shows the signals as they were received. In the lower plot the same patches have been modified using the optimal parameters h and α.

The method presented in this paper is developed to estimate small translations, s.t. $h \in [-10, 10]$ samples. However, in the experiments the delays were larger than that. Therefore we began by pre-estimating an integer delay $\tilde{h}^{(i)}$ using GCC-PHAT. The GCC-PHAT method is described in [11]. After that we did a

subsample refinement of the translation and estimated the Doppler parameter using our method. This was done by minimization of the intergral

$$\int_t (V^{(i)}(t) - \bar{V}^{(i)}(\alpha^{(i)}t + \tilde{h}^{(i)} + h^{(i)}))^2 \, \mathrm{d}t. \tag{45}$$

Here, the optimization was over $h^{(i)}$ and $\alpha^{(i)}$, while $\tilde{h}^{(i)}$ should be seen as a constant.

The results after applying the optimized parameters to one of the signal patches can be seen in Fig. 7. The optimization was carried out for all different patches.

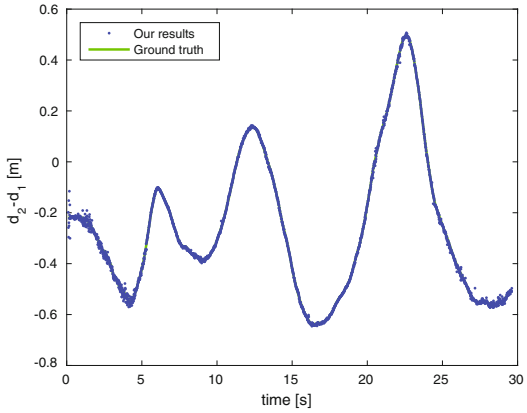

Fig. 8. The figure shows the difference between the distances from receiver 1 to the sender (d_1) and receiver 2 to the sender (d_2) over time. The ground truth $\Delta d^{(i)}$ is plotted as a solid line (—) and the values $\Delta \bar{d}^{(i)}$ obtained from time-difference estimates as dots (•). Each dot represents the value for one signal patch. It is hard to distinguish the line representing ground truth since the estimations agree well with this. The plot is similar to Fig. 7 in [10], but has been generated using the updated and more independent method which is presented in this paper.

Comparison with Ground Truth. The distances $d_1^{(i)}$ and $d_2^{(i)}$ from the microphones to the loudspeaker were computed from the ground truth receiver and sender positions (r_1, r_2 and $s^{(i)}$) according to

$$d_1^{(i)} = |r_1 - s^{(i)}|, \qquad d_2^{(i)} = |r_2 - s^{(i)}|. \tag{46}$$

The difference of these distances,

$$\Delta d^{(i)} = d_2^{(i)} - d_1^{(i)} \tag{47}$$

has a connection to our estimated translation $h^{(i)}$ and the time difference of arrival. However, $\Delta d^{(i)}$ is measured in meters, while we compute $h^{(i)}$ in samples.

To be able to compare these two, we multiplied $h^{(i)}$ with a scaling factor c/f. The recording frequency was $f = 96$ kHz and $c = 340$ m/s is the speed of sound. From this we could obtain an estimation of $\Delta d^{(i)}$,

$$\Delta \bar{d}^{(i)} = \frac{c}{f} \cdot h^{(i)}. \tag{48}$$

Thereafter we could compare our estimated values $\Delta \bar{d}^{(i)}$ to the ground truth values $\Delta d^{(i)}$. The ground truth is plotted together with our estimations in Fig. 8. The plot shows the results over time, for all different patches. It is clear that the two agree well.

The Doppler parameter measures how the distance differences changes, i.e.

$$\frac{\partial \Delta d}{\partial t} = \frac{\partial d_2}{\partial t} - \frac{\partial d_1}{\partial t}. \tag{49}$$

Here, the distances over time are denoted d_1 and d_2 respectively. The derivative of $d_1(t) = |r_1 - s(t)|$ is

$$\frac{\partial d_1}{\partial t} = \frac{r_1 - s}{|r_1 - s|} \cdot \frac{\partial s}{\partial t}, \tag{50}$$

where $\cdot$ denotes the scalar product between the two time dependent vectors. The derivative of d_2 can be found correspondingly. If $n_1^{(i)}$ and $n_2^{(i)}$ are unit vectors in the direction from $s^{(i)}$ to r_1 and r_2 respectively, i.e.

$$n_1^{(i)} = \frac{r_1 - s^{(i)}}{|r_1 - s^{(i)}|}, \qquad n_2^{(i)} = \frac{r_2 - s^{(i)}}{|r_2 - s^{(i)}|}, \tag{51}$$

the derivatives can be expressed as

$$\frac{\partial d_1^{(i)}}{\partial t} = n_1^{(i)} \cdot \frac{\partial s^{(i)}}{\partial t}, \qquad \frac{\partial d_2^{(i)}}{\partial t} = n_2^{(i)} \cdot \frac{\partial s^{(i)}}{\partial t}. \tag{52}$$

Thus

$$\frac{\partial \Delta d^{(i)}}{\partial t} = n_2^{(i)} \cdot \frac{\partial s^{(i)}}{\partial t} - n_1^{(i)} \cdot \frac{\partial s^{(i)}}{\partial t}. \tag{53}$$

These ground truth Doppler values can be interpreted as how much Δd changes each second. However, our estimated Doppler factor α is a unit-less constant. We can express the relation between the two values as

$$\frac{\partial \Delta d}{\partial t} = (\alpha - 1) \cdot c, \tag{54}$$

where c still denotes the speed of sound. In Fig. 9 the ground truth is plotted as a solid line together with our estimations marked with dots. The similarities are easily distinguishable even if the estimations are noisy.

It is clear from the plots that the estimations contain relevant information. However, there is quite some noise in the estimates in Figs. 8 and 9. This can be reduced further by computation of a moving average. We have computed a

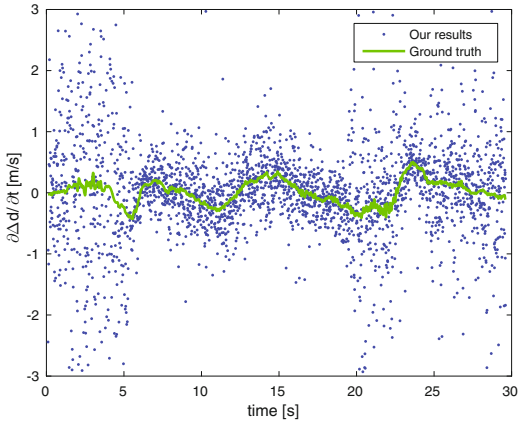

Fig. 9. The derivative of the distance differences Δd plotted over time. The dots (•) are our estimations and the solid line (—) is computed from ground truth. We see that even though the estimations are noisy the pattern agree with ground truth. The plot is similar to Fig. 8 in [10], but has been generated using the updated and more independent method which is presented in this paper.

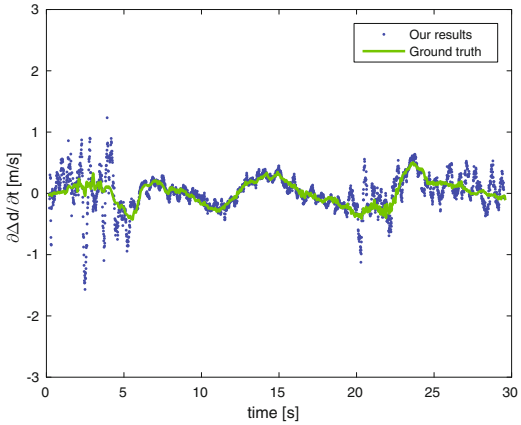

Fig. 10. This plot shows essentially the same thing as Fig. 9, i.e. $\partial \Delta d / \partial t$, but with a 20-patches moving average over the estimations. The averaging substantially reduces the noise. The plot is similar to Fig. 10 in [10], but has been generated using the updated and more independent method which is presented in this paper.

moving average over 20 patches – approximately 0.2 s – for the distance difference derivative and plotted the result in Fig. 10. The plot can be compared to Fig. 9, where no averaging has been done. We see that the moving average substantially reduces the noise in the estimates.

Even in Fig. 10, the estimates in the beginning are noisy. This is due to the character of the song that was played, where the sound is not persistent until

after 5–6 s. In the beginning there are just intermittent drumbeats and silence between these. Then the information is not sufficient to make good estimates. Thus, it is more fair to the algorithm to review the results from 5–6 s and forward.

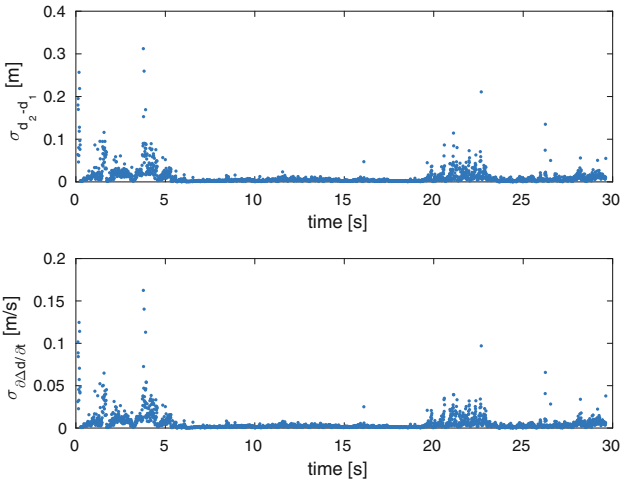

Fig. 11. The standard deviation for parameters that was estimated for the real data. The upper plot shows the standard deviation of the distance difference in Fig. 8 over time and the lower plot shows the standard deviation of the derivative of the distance difference in Fig. 9.

Estimating the Standard Deviation of the Parameters. We have also computed the standard deviations of the parameters in accordance to Sect. 3.1. These are plotted over time in Fig. 11. We can see that the estimations are more uncertain in the beginning of the song, in consistence with when the signal is not persistent. However, just by looking at the estimated Doppler factor this seems to be more uncertain than the theoretical standard deviation suggests.

We also estimated the standard deviations empirically. This was done using the results in Figs. 8 and 9. The empirical standard deviation was computed for the difference between our estimations and ground truth, for a certain time window, namely $t \in [10, 15]$.

The different standard deviations are displayed in Table 2. For the theoretical values we have computed the mean and median, both for all signal and for $t \in [10, 15]$ for comparison with the empirical values.

We can see that the theoretical and empirical values agree quite well for the translation. The reason that the mean of the theoretical standard deviation is higher for all signal is due to the parts of the signal that are more uncertain. However, in the chosen time window the values agree well.

For the Doppler factor the theoretical standard deviation is lower compared to the empirical estimates. This is interesting and there can be several reasons.

Table 2. The mean and the median for the standard deviation of the estimated distance difference (Fig. 8) and the Doppler factor (Fig. 9) for the two received signals.

		Translation, $d_2 - d_1$	Doppler factor, $\partial \Delta d / \partial t$
Theoretical, all signal	Mean of std	$1.03 \cdot 10^{-2}$	$5.25 \cdot 10^{-3}$
	Median of std	$4.71 \cdot 10^{-3}$	$2.39 \cdot 10^{-3}$
Theoretical, $t \in [10, 15]$	Mean of std	$4.58 \cdot 10^{-3}$	$2.29 \cdot 10^{-3}$
	Median of std	$4.12 \cdot 10^{-3}$	$2.08 \cdot 10^{-3}$
Empirical, $t \in [10, 15]$		$3.88 \cdot 10^{-3}$	$4.43 \cdot 10^{-1}$

To begin with, we made some assumptions for the received signals when we derived the equations in Sect. 3.1, which are probably not true for our data. E.g. in our experiments we estimated the noise in the signals as the difference between the two signals after modification. In the bottom plot of Fig. 7 we see that there is still an amplitude difference between the two signals. This means that our estimated noise will not be w.s.s., as was assumed in the derivations. Furthermore, the noise will thus be overestimated. Actually, it turned out the SNR was below 4.7.

Except from this, our method is developed to work with one signal with constant parameters and does not take into account that the patches in our real data actually constitutes one long signal. Also, we might have forgotten to take some important factor into account in out derivations for the standard deviation of the Doppler factor. It might be that the problem cannot be modeled as linear. Regardless, an interesting point for future focus is to investigate this.

4.3 Expanding the Model for Real Data

As mentioned in Sect. 3.2 it is in practice not much harder to estimate three model parameters. Therefore, to get a more precise solution (see Sect. 4.1 and the end of the previous section), we have also made experiments on the same data using (42) as model for the signals. The computations are made in the same manner as in the previous section but the error function (45) is replaced by

$$\int_t (V^{(i)}(t) - \gamma^{(i)} \bar{V}^{(i)}(\alpha^{(i)} t + \tilde{h}^{(i)} + h^{(i)}))^2 \, \mathrm{d}t, \tag{55}$$

and the optimization is performed over all three parameters, the subsample translation $h^{(i)}$, the Doppler factor $\alpha^{(i)}$ and the amplitude factor $\gamma^{(i)}$.

The results from using this model for the same signal patch as in Fig. 7 can be seen in Fig. 12. After moving the signals according to the estimated parameters the norm of the difference between the signals (bottom plot in the figures) has decreased with 20% when we included the amplitude factor compared to when we did not.

The plots for the translation parameter and the Doppler factor look similar to the plots in Figs. 8 and 9. However, we can now make a comparison to ground truth for the amplitude factor γ as well.

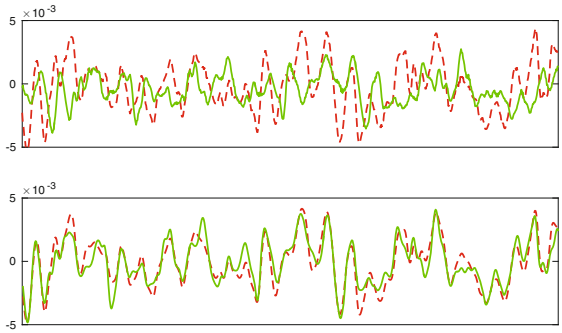

Fig. 12. The plot shows the same signal patches as in Fig. 7. The difference is that a larger model, namely (42), has been used here and thus an amplitude has been estimated as well. The bottom image shows the same signals after modifications using the optimal parameters.

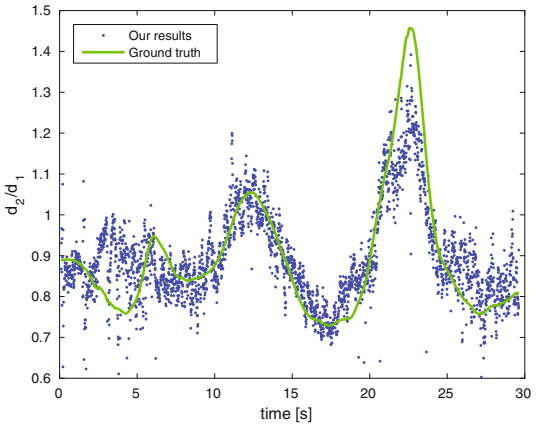

Fig. 13. The distance quotient d_2/d_1 plotted over time. The solid line (—) represents the ground truth and each dot (•) is the estimation for a certain patch. While the estimations are somewhat noisy there is no doubt that the pattern is the same. The plot is similar to Fig. 9 in [10], but has been generated using the updated and more independent method which is presented in this paper.

The amplitude difference of the two received signals can be compared to $d_1^{(i)}$ and $d_2^{(i)}$. The amplitude estimate $\gamma^{(i)}$ is related to the quotient of the distances, $d_2^{(i)}/d_1^{(i)}$. Since the sound spreads as on the surface of a sphere, the distance quotient is proportional to the square root of the amplitude $\gamma^{(i)}$,

$$\frac{d_2^{(i)}}{d_1^{(i)}} = C \cdot \sqrt{\gamma^{(i)}}. \tag{56}$$

The unknown constant C depends on the gains of the two recording channels. For the experiment, the estimated proportionality constant was $C = 1.3$.

The distance quotient is plotted over time in Fig. 13 – our estimations as dots and ground truth as a solid line. Again we see that they clearly follow the same pattern.

5 Conclusions

In this paper we have studied how to estimate three parameters – time-differences, amplitude changes and minute Doppler effects – from two audio signals. The study also contains a stochastic analysis for these estimated parameters and a comparison between different signal models. The results are important both for simultaneous determination of sender and receiver positions, but also for localization, beam-forming and diarization. In the paper we have built on previous results on stochastic analysis of interpolation and smoothing in order to give explicit formulas for the covariance matrix of the estimated parameters. In the paper it is shown that the approximations that are introduced in the theory are valid as long as the smoothing is at least 0.55 sample points and as long as the signal-to-noise ratio is greater than 4.7. Furthermore, we show using experiments on both simulated and real data that these estimates provide useful information for subsequent analysis.

Acknowledgements. This work is supported by the strategic research projects ELLIIT and eSSENCE, Swedish Foundation for Strategic Research project "Semantic Mapping and Visual Navigation for Smart Robots" (grant no. RIT15-0038) and Wallenberg Autonomous Systems and Software Program (WASP).

References

1. Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., Vinyals, O.: Speaker diarization: a review of recent research. IEEE Trans. Audio Speech Lang. Process. **20**(2), 356–370 (2012)
2. Anguera, X., Wooters, C., Hernando, J.: Acoustic beamforming for speaker diarization of meetings. IEEE Trans. Audio Speech Lang. Process. **15**(7), 2011–2022 (2007)
3. Åström, K., Heyden, A.: Stochastic analysis of image acquisition, interpolation and scale-space smoothing. In: Advances in Applied Probability, vol. 31, no. 4, pp. 855–894 (1999)
4. Batstone, K., Oskarsson, M., Åström, K.: Robust time-of-arrival self calibration and indoor localization using wi-fi round-trip time measurements. In: Proceedings of International Conference on Communication (2016)
5. Brandstein, M., Adcock, J., Silverman, H.: A closed-form location estimator for use with room environment microphone arrays. IEEE Trans. Speech Audio Process. **5**(1), 45–50 (1997)
6. Cirillo, A., Parisi, R., Uncini, A.: Sound mapping in reverberant rooms by a robust direct method. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 285–288, April 2008

7. Cobos, M., Marti, A., Lopez, J.: A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling. IEEE Signal Process. Lett. **18**(1), 71–74 (2011)
8. Crocco, M., Del Bue, A., Bustreo, M., Murino, V.: A closed form solution to the microphone position self-calibration problem. In: ICASSP, March 2012
9. Do, H., Silverman, H., Yu, Y.: A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array. In: ICASSP 2007, vol. 1, pp. 121–124, April 2007
10. Flood, G., Heyden, A., Åström, K.: Estimating uncertainty in time-difference and doppler estimates. In: 7th International Conference on Pattern Recognition Applications and Methods (2018)
11. Knapp, C., Carter, G.: The generalized correlation method for estimation of time delay. IEEE Trans. Acoust. Speech Signal Process. **24**(4), 320–327 (1976)
12. Kuang, Y., Åström, K.: Stratified sensor network self-calibration from tdoa measurements. In: EUSIPCO (2013)
13. Kuang, Y., Burgess, S., Torstensson, A., Åström, K.: A complete characterization and solution to the microphone position self-calibration problem. In: ICASSP (2013)
14. Kuang, Y., Åström, K.: Stratified sensor network self-calibration from tdoa measurements. In: 21st European Signal Processing Conference 2013 (2013)
15. Lindeberg, T.: Scale-space theory: a basic tool for analyzing structures at different scales. J. Appl. Stat. **21**(1–2), 225–270 (1994)
16. Lindgren, G., Rootzén, H., Sandsten, M.: Stationary Stochastic Processes for Scientists and Engineers. CRC Press, New York (2013)
17. Petersen, K.B., Pedersen, M.S., et al.: The matrix cookbook. Tech. Univ. Den. **7**(15), 510 (2008)
18. Plinge, A., Jacob, F., Haeb-Umbach, R., Fink, G.A.: Acoustic microphone geometry calibration: an overview and experimental evaluation of state-of-the-art algorithms. IEEE Signal Process. Mag. **33**(4), 14–29 (2016)
19. Pollefeys, M., Nister, D.: Direct computation of sound and microphone locations from time-difference-of-arrival data. In: Proceedings of ICASSP (2008)
20. Shannon, C.E.: Communication in the presence of noise. Proc. IRE **37**(1), 10–21 (1949)
21. Zhayida, S., Segerblom Rex, S., Kuang, Y., Andersson, F., Åström, K.: An Automatic System for Acoustic Microphone Geometry Calibration based on Minimal Solvers. ArXiv e-prints, October 2016
22. Zhayida, S., Andersson, F., Kuang, Y., Åström, K.: An automatic system for microphone self-localization using ambient sound. In: 22st European Signal Processing Conference (2014)
23. Zhayida, S., Åström, K.: Time difference estimation with sub-sample interpolation. J. Signal Process. **20**(6), 275–282 (2016)