Real-time Loudspeaker Distance Estimation with Stereo Audio

Nielsen, Jesper Kjær; Gaubitch, Nikolay; Heusdens, Richard; Martinez, Jorge; Jensen, Tobias Lindstrøm; Jensen, Søren Holdt

Published in:
23rd European Signal Processing Conference (EUSIPCO), 2015

DOI (link to publication from Publisher):
10.1109/EUSIPCO.2015.7362383

Publication date:
2015

Document Version
Accepted author manuscript, peer reviewed version

REAL-TIME LOUDSPEAKER DISTANCE ESTIMATION WITH STEREO AUDIO

Jesper Kjær Nielsen[12], *Nikolay D. Gaubitch*[3], *Richard Heusdens*[3], *Jorge Martinez*[3], *Tobias Lindstrøm Jensen*[1], *and Søren Holdt Jensen*[1]

[1]Aalborg University, Denmark
Dept. of Electronic Systems
{jkn,tlj,shj}@es.aau.dk

[2]Bang & Olufsen A/S
Struer, Denmark

[3]Delft University of Technology, The Netherlands
Signal and Information Processing Lab
{n.d.gaubitch,r.heusdens,
j.a.martinezcastaneda}@tudelft.nl

ABSTRACT

Knowledge on how a number of loudspeakers are positioned relative to a listening position can be used to enhance the listening experience. Usually, these loudspeaker positions are estimated using calibration signals, either audible or psycho-acoustically hidden inside the desired audio signal. In this paper, we propose to use the desired audio signal instead. Specifically, we treat the case of estimating the distance between two loudspeakers playing back a stereo music or speech signal. In this connection, we develop a real-time maximum likelihood estimator and demonstrate that it has a variance in the millimetre range in a real environment for even a modest sampling frequency.

Index Terms— Loudspeaker localisation, distance estimation, time-of-arrival estimation

1. INTRODUCTION

The distribution of a number of loudspeakers relative to the listening position has a large impact on the listening experience and the perceived spaciousness of sound [1–3]. Often, however, the loudspeakers are not placed in the optimal position since other interior design considerations take higher priority or the desired listening position moves. This can to some extent be compensated for by preprocessing the loudspeaker signals. However, in order to apply the correct preprocessing, the location of the loudspeakers relative to the listening position must be known.

Existing approaches to solving this loudspeaker localisation problem can roughly be dichotomised into two groups. In the first group, synthetic test signals such as sinusoidal sweeps or maximum length sequences (MLS) are used as calibration signals [4–6]. This has the advantage of high estimation accuracy, but also requires the user to actively start the calibration sequence every time, e.g, the listening position or the loudspeaker locations change. This is solved in the second group of methods by adding a calibration signal to the desired audio signal [7, 8]. The calibration signal is shaped psycho-acoustically and hidden inside the audio signal so that it is inaudible to the listener. Consequently, the energy of the calibration signal is low compared to the energy of the audio signal. This is a problem since the audio signal is considered to be "noise" in the source localisation algorithm, and this affects the estimation accuracy [8].

So why not use the audio signal for source localisation? To our surprise, we have not been able to find any previous work on this. We believe that the main reason for this is that audio signals are much more difficult to work with since they are heavily correlated in both time and in between the loudspeaker channels and have an unknown frequency content. Consequently, it is hard to estimate impulse responses, and the simple cross-correlation methods for loudspeaker localisation fail. Synthetic calibration signals, on the other hand, can be designed to be uncorrelated and to have a desirable frequency content. Thus, the simple cross-correlation methods and impulse response peak picking can be used to compute the distances and/or direction of arrivals (DOAs) between the loudspeakers and/or to the listening position.

In this paper, we take a first step in the direction of loudspeaker localisation using only the desired audio signals. Specifically, we focus on the case where we have to estimate the distance between two loudspeakers playing back a stereo music signal. Distances between all the loudspeaker pairs in a set of loudspeakers can be used to form an Euclidean distance matrix to which the positions of the loudspeakers can be fitted using, e.g, the multidimensional scaling (MDS) algorithm [9] or the algorithm by Crocco et al. [10]. The latter method has the advantage that the loudspeakers and microphones do not have to be co-located. Here, however, we assume that a microphone is mounted on every loudspeaker, which we will refer to as a transceiver, so that they are approximately co-located. This assumption is used in the proposed estimator of the distance to take into account that both transceivers in a transceiver pair should measure the same distance. This increases the robustness of the estimator. Inspired by the recent work in [11], we also formulate the signal model so that the estimator produces estimates from a continuous set without resorting to any heuristic interpolation method. This is in contrast to many of the proposed localisation methods whose resolution is bounded to the sampling grid (see, e.g., [4,5,7]).

2. ESTIMATING THE DISTANCE BETWEEN TWO TRANSCEIVERS

As alluded to in the introduction, we focus on estimating the distance between two transceivers playing back stereo music or a speech signal. In this paper, a transceiver is a loudspeaker with a microphone mounted close to the diaphragm of the loudspeaker. The developed estimator is not only limited in scope to this special case, but can also be used for the problem where the direct distance should be estimated from a loudspeaker to a microphone, e.g, placed at the listening position, and for the problem where the distance to a reflector should be estimated using just one transceiver. These special cases are obtained by appropriately selecting the source and sensor signals.

2.1. The Signal Model

We assume that the two transceivers record N samples each, and we model these as

$$x_1(n) = q_{11}(n) + q_{21}(n) + e_1(n) \tag{1}$$
$$x_2(n) = q_{22}(n) + q_{12}(n) + e_2(n) \tag{2}$$

where $e_i(n)$ and $q_{ki}(n)$ are the noise recorded by transceiver i and the signal recorded by transceiver i from transceiver k, respectively. Thus, $q_{ii}(n)$ is the part of the microphone signal $x_i(n)$ which originates from transceiver i. This signal is not of interest as it does not contain any information on the distance between the transceivers, and we therefore wish to suppress it as much as possible. To do that, we model $q_{ii}(n)$ as

$$q_{ii}(n) = \sum_{m=0}^{M-1} h_i(m) s_i(n-m) \tag{3}$$

where $s_i(n)$ and $h_i(m)$ are a source signal sample of transceiver i and an FIR filter coefficient of the ith M-length transceiver filter, respectively. Thus, a transceiver filter models the acoustic impulse response between the loudspeaker and microphone on a transceiver. We assume that the loudspeakers and microphones are all connected to the same system so that the source signals are known. On the other hand, the transceiver filters are assumed unknown since these might be slowly time-varying due to, e.g., temperature changes. These transceiver filters are very important in order to attenuate the contribution of $s_i(n)$ in $x_i(n)$ since only $q_{ki}(n)$ for $k \neq i$ contains information about the distance between the transceivers. Therefore, $q_{ki}(n)$ is modelled explicitly in terms of the delay parameter (in samples) $\eta \in [M, K]$ with $M < K < N$, which we are interested in estimating, and the gain $\beta \geq 0$ as

$$q_{ki}(n) = \beta s_k(n-\eta), \qquad \text{for } i \neq k\,. \tag{4}$$

This model describes the sound propagation of the direct path. Note that the reverberation is later modelled as part of the noise (see Sec. 2.1.2) and that β and η are not indexed since we assume that they are the same for both $q_{12}(n)$ and $q_{21}(n)$. Moreover, the delay η is related to the distance between the loudspeakers d via $d = \eta c/f_\mathrm{s}$ where c is the speed of sound and f_s the sampling frequency. If we define the vectors

$$\boldsymbol{x}_i = \begin{bmatrix} x_i(0) & x_i(1) & \cdots & x_i(N-1) \end{bmatrix}^T \tag{5}$$
$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_1^T & \boldsymbol{x}_2^T \end{bmatrix}^T \tag{6}$$
$$\boldsymbol{s}_i(\eta) = \begin{bmatrix} s_i(-\eta) & s_i(1-\eta) & \cdots & s_i(N-1-\eta) \end{bmatrix}^T \tag{7}$$
$$\boldsymbol{e}_i = \begin{bmatrix} e_i(0) & e_i(1) & \cdots & e_i(N-1) \end{bmatrix}^T \tag{8}$$
$$\boldsymbol{e} = \begin{bmatrix} \boldsymbol{e}_1^T & \boldsymbol{e}_2^T \end{bmatrix}^T \tag{9}$$
$$\boldsymbol{h}_i = \begin{bmatrix} h_i(0) & h_i(1) & \cdots & h_i(M-1) \end{bmatrix}^T\,, \tag{10}$$

it follows that the signal model can be written as

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{B}_1 & \boldsymbol{0} & \boldsymbol{s}_2(\eta) \\ \boldsymbol{0} & \boldsymbol{B}_2 & \boldsymbol{s}_1(\eta) \end{bmatrix} \begin{bmatrix} \boldsymbol{h}_1 \\ \boldsymbol{h}_2 \\ \beta \end{bmatrix} + \boldsymbol{e} \tag{11}$$
$$= \boldsymbol{B}\boldsymbol{h} + \boldsymbol{s}(\eta)\beta + \boldsymbol{e} \tag{12}$$

where the definitions of $\boldsymbol{B}$, $\boldsymbol{h}$, and $\boldsymbol{s}(\eta)$ are obvious and

$$\boldsymbol{B}_i = \begin{bmatrix} \boldsymbol{s}_i(0) & \boldsymbol{s}_i(1) & \cdots & \boldsymbol{s}_i(M-1) \end{bmatrix} \tag{13}$$

is a convolution matrix. To summarise, we have so far assumed a signal model which is linear in the unknown transceiver filters $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ and the gain β and is non-linear in the delay η. The main reason for using this signal model is that, as we show in Sec. 2.2, the linear parameters can easily be separated out of the problem leaving us with the single non-linear parameter η which we are interested in estimating. Before deriving the estimator for η, however, we make a number of assumptions about the source signal and the noise which enables sub-sample delay estimation accuracy, drastically reduces the computational complexity, and increases the robustness of the resulting estimator.

2.1.1. The Source Signals

Most scientific literature on time of arrival (TOA), time difference of arrival (TDOA), and DOA estimation formulates these problems in the frequency domain since a delay in the time domain corresponds to a phase-shift in the frequency domain. Consequently, the delay parameter can be separated out analytically from the source signal and modelled as a continuous parameter. For finite length signals, however, a delay in the time domain only corresponds to a phase shift in the frequency domain if the signal is periodic with fundamental frequency $2\pi/N$ radians per sample (or an integer multiple thereof) [11]. Since we here consider very long segments compared to the delay, we wish to estimate, we do not make an inappropriate error by assuming that the source signals are periodic. Thus, we have that

$$\boldsymbol{s}_i(\eta) = \boldsymbol{Z}\boldsymbol{A}_i\boldsymbol{d}(\eta) \tag{14}$$
$$\boldsymbol{B}_i = \boldsymbol{Z}\boldsymbol{A}_i\boldsymbol{F} \tag{15}$$

where we have defined

$$\boldsymbol{z}(\omega) = \begin{bmatrix} 1 & \exp(j\omega) & \cdots & \exp(j\omega(N-1)) \end{bmatrix}^T \quad (16)$$

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{z}(-2\pi L/N) & \cdots & \mathbf{1} & \cdots & \boldsymbol{z}(2\pi L/N) \end{bmatrix} \quad (17)$$

$$\boldsymbol{d}(\eta) = \begin{bmatrix} \exp(j2\pi\eta L/N) & \cdots & 1 & \cdots & \exp(-j2\pi\eta L/N) \end{bmatrix}^T \quad (18)$$

$$\boldsymbol{A}_i = N^{-1}\,\mathrm{diag}(\boldsymbol{Z}^H \boldsymbol{s}_i(0)) \quad (19)$$

$$\boldsymbol{F} = \begin{bmatrix} \boldsymbol{d}(0) & \boldsymbol{d}(1) & \cdots & \boldsymbol{d}(M-1) \end{bmatrix} . \quad (20)$$

Note that the time indices are symmetric around zero from $-L$ to L where $L = \lfloor N/2 \rfloor$. This is necessary to ensure that the decomposition of $\boldsymbol{s}_i(\eta)$ is real-valued for non-integer values of η [11].

2.1.2. The Noise

We assume that the noise is Gaussian and consists of two parts

$$\boldsymbol{e}_i = \boldsymbol{w}_i + \boldsymbol{v}_i \quad (21)$$

where the first part is due to reverberation and the second part is measurement noise. These two are assumed to be independent, and the measurement noise is modelled as white Gaussian noise with variance σ^2. We model $\boldsymbol{w}_i$ as a delayed and weighted sum of the two source signals so that

$$\boldsymbol{w}_i = \sum_{m=2}^{M} \left(\boldsymbol{s}_1(\eta_{1i,m})\beta_{1i,m} + \boldsymbol{s}_2(\eta_{2i,m})\beta_{2i,m} \right) \quad (22)$$

where $\eta_{1i,m}$ and $\beta_{1i,m}$ are the m'th reflection and gain from transceiver 1 to transceiver i. The summation index is running from $m = 2$ to indicate that the first component is already included in the model via (4). We now make the critical assumption that all reflections are uncorrelated so that

$$E[\boldsymbol{w}_i \boldsymbol{w}_k^H] \approx \mathbf{0} \quad (23)$$

$$E[\boldsymbol{w}_i \boldsymbol{w}_i^H] \approx \sum_{m=2}^{M} E\Big[\boldsymbol{s}_1(\eta_{1i,m})\beta_{1i,m}^2 \boldsymbol{s}_1^H(\eta_{1i,m}) + \boldsymbol{s}_2(\eta_{2i,m})\beta_{2i,m}^2 \boldsymbol{s}_2^H(\eta_{2i,m}) \Big] \quad (24)$$

$$\approx \gamma\sigma^2 \boldsymbol{Z}(\boldsymbol{A}_1\boldsymbol{A}_1^H + \boldsymbol{A}_2\boldsymbol{A}_2^H)\boldsymbol{Z}^H \quad (25)$$

where γ is an uninteresting scale parameter and the last expression follows from the decomposition in (14) and from

$$E\left[\sum_{m=2}^{M} \boldsymbol{d}(\eta_{i,m})\beta_{i,m}^2 \boldsymbol{d}^H(\eta_{i,m}) \right] \approx \gamma\sigma^2 \boldsymbol{I}_N . \quad (26)$$

These assumptions are hard to justify theoretically, but have been demonstrated to work well in practice [12, 13]. Under these assumptions, the covariance matrix of the noise can be written as

$$\boldsymbol{C} = E[\boldsymbol{e}\boldsymbol{e}^H] \approx \begin{bmatrix} \boldsymbol{C}_1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{C}_2 \end{bmatrix} \quad (27)$$

$$\boldsymbol{C}_i \approx \gamma\sigma^2 \left[\boldsymbol{Z}(\boldsymbol{A}_1\boldsymbol{A}_1^H + \boldsymbol{A}_2\boldsymbol{A}_2^H)\boldsymbol{Z}^H + \gamma^{-1}\boldsymbol{I}_N \right] . \quad (28)$$

Applying the matrix inversion lemma to $\boldsymbol{C}_i^{-1}$, we obtain that

$$\boldsymbol{C}_i^{-1} = \sigma^{-2}\left[\boldsymbol{I}_N - N^{-1}\boldsymbol{Z}\boldsymbol{Z}^H + (N^2\gamma)^{-1}\boldsymbol{Z}\boldsymbol{Q}\boldsymbol{Z}^H \right] \quad (29)$$

where we have defined

$$\boldsymbol{Q} = \left(\boldsymbol{A}_1\boldsymbol{A}_1^H + \boldsymbol{A}_2\boldsymbol{A}_2^H + (N\gamma)^{-1}\boldsymbol{I}_N \right)^{-1} . \quad (30)$$

With these, we obtain

$$\boldsymbol{Z}^H\boldsymbol{C}_i^{-1} = (\sigma^2 N\gamma)^{-1}\boldsymbol{Q}\boldsymbol{Z}^H \quad (31)$$

$$\boldsymbol{Z}^H\boldsymbol{C}_i^{-1}\boldsymbol{Z} = (\sigma^2\gamma)^{-1}\boldsymbol{Q} \quad (32)$$

which proves to be useful later.

2.2. A Maximum Likelihood Estimator

The log-likelihood function pertaining to the model in (12) is given by

$$l(\boldsymbol{h}_1, \boldsymbol{h}_2, \beta, \eta, \sigma^2, \gamma) = -\frac{1}{2}\Big[\ln|\boldsymbol{C}| + (\boldsymbol{x} - \boldsymbol{B}\boldsymbol{h} - \boldsymbol{s}(\eta)\beta)^H \boldsymbol{C}^{-1}(\boldsymbol{x} - \boldsymbol{B}\boldsymbol{h} - \boldsymbol{s}(\eta)\beta) \Big] \quad (33)$$

where all terms which do not depend on the unknown parameters have been ignored. Whereas the linear parameters $\boldsymbol{h}$ and β and the noise variance σ^2 can be separated out of the likelihood function, the scale factor γ cannot. Since γ is only a nuisance parameter, we assume that it is known and large. That is, we assume that the reverberation energy is much larger than that of the measurement noise. We have found that this works very well in practice. As seen from (30), this means that $(N\gamma)^{-1}$ acts as a regularisation parameter.

To derive the maximum likelihood (ML) estimator for the delay η, we perform the following steps. Given η and β, the ML-estimate of the transceiver filters is given by

$$\hat{\boldsymbol{h}} = \left(\boldsymbol{B}^H\boldsymbol{C}^{-1}\boldsymbol{B} \right)^{-1} \boldsymbol{B}^H\boldsymbol{C}^{-1}(\boldsymbol{x} - \boldsymbol{s}(\eta)\beta) , \quad (34)$$

Inserting this estimate back into the log-likelihood function in (33) and only keeping the terms which depend on η and β give the optimisation problem[1]

$$\hat{\beta}, \hat{\eta} = \underset{\beta \geq 0, \eta \in [M,K]}{\mathrm{argmin}} \; (\boldsymbol{x} - \boldsymbol{s}(\eta)\beta)^H \boldsymbol{C}^{-1}\boldsymbol{R}(\boldsymbol{x} - \boldsymbol{s}(\eta)\beta) \quad (35)$$

where $\boldsymbol{R} = \mathrm{diag}(\boldsymbol{R}_1, \boldsymbol{R}_2)$ is a block diagonal matrix with $\boldsymbol{R}_i = \boldsymbol{I}_N - \boldsymbol{B}_i(\boldsymbol{B}_i^H\boldsymbol{C}_i^{-1}\boldsymbol{B}_i)^{-1}\boldsymbol{B}_i^H\boldsymbol{C}_i^{-1}$. Despite the non-negative constraint on the gain β, it can still be separated out of the optimisation problem by solving a KKT system of equations. The final 1D optimisation problem for the delay is

$$\hat{\eta} = \underset{\eta \in [M,K]}{\mathrm{argmax}} \max(J(\eta), 0) \quad (36)$$

where the cost function is given by

$$J(\eta) = \frac{\boldsymbol{s}_2^H(\eta)\boldsymbol{C}_1^{-1}\boldsymbol{R}_1\boldsymbol{x}_1 + \boldsymbol{s}_1^H(\eta)\boldsymbol{C}_2^{-1}\boldsymbol{R}_2\boldsymbol{x}_2}{\sqrt{\boldsymbol{s}_2^H(\eta)\boldsymbol{C}_1^{-1}\boldsymbol{R}_1\boldsymbol{s}_2(\eta) + \boldsymbol{s}_1^H(\eta)\boldsymbol{C}_2^{-1}\boldsymbol{R}_2\boldsymbol{s}_1(\eta)}} .$$

This cost function is highly non-linear in η so we propose to find $\hat{\eta}$ using a two step procedure. First, a coarse value for $\hat{\eta}$ is computed from a search over $J(\eta)$ on a uniform grid. Second, the coarse estimate is refined using a line searching method such as a Fibonacci search [14, pp. 85–92].

[1] Note that $\boldsymbol{R}^H\boldsymbol{C}^{-1}\boldsymbol{R} = \boldsymbol{C}^{-1}\boldsymbol{R}$.

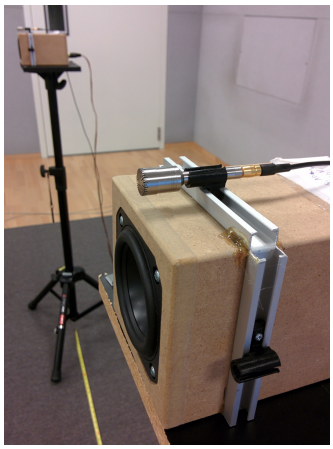

Fig. 1. A picture of the stereo setup.

Type	WGN	Music	Speech
Simulation	0.28	0.78	0.18
Measurement	0.61	1.42	1.13

Table 1. Standard deviation in mm of the estimated distance for three source signals in a simulated and real environment.

2.2.1. Efficient Implementation

The cost function $J(\eta)$ can be evaluated efficiently by using the intermediate results in (14), (15), (31), and (32), and by computing the economy size singular value decomposition (SVD) $\boldsymbol{Q}^{1/2}\boldsymbol{A}_i\boldsymbol{F} = \boldsymbol{U}_i\boldsymbol{S}_i\boldsymbol{V}_i$ so that

$$\boldsymbol{Z}^H\boldsymbol{C}_i^{-1}\boldsymbol{R}_i = (\sigma^2 N\gamma)^{-1}\boldsymbol{Q}^{1/2}(\boldsymbol{I}_N - \boldsymbol{U}_i\boldsymbol{U}_i^H)\boldsymbol{Q}^{1/2}\boldsymbol{Z}^H \ .$$

These results allow us to write the cost function as

$$J(\eta) = \frac{\boldsymbol{d}^H(\eta)\,(\boldsymbol{y}_1 + \boldsymbol{y}_2)}{\sqrt{2L + 1 - \boldsymbol{d}^H(\eta)\,(\boldsymbol{K}_1 + \boldsymbol{K}_2)\,\boldsymbol{d}(\eta)}} \tag{37}$$

where (for $k \neq i$)

$$\boldsymbol{y}_i = \boldsymbol{A}_k^H\boldsymbol{Q}^{1/2}(\boldsymbol{I}_N - \boldsymbol{U}_i\boldsymbol{U}_i^H)\boldsymbol{Q}^{1/2}\boldsymbol{Z}^H\boldsymbol{x}_i \tag{38}$$

$$\boldsymbol{K}_i = \boldsymbol{A}_i^H\boldsymbol{Q}^{1/2}\boldsymbol{U}_i\boldsymbol{U}_i^H\boldsymbol{Q}^{1/2}\boldsymbol{A}_i \ . \tag{39}$$

Note that $\boldsymbol{Z}^H\boldsymbol{x}_i$ and all elements of the diagonal matrices $\boldsymbol{A}_i$ and $\boldsymbol{Q}$ can be computed using an FFT algorithm. Moreover, $\boldsymbol{d}^H(\eta)\boldsymbol{K}_i\boldsymbol{d}(\eta)$ is approximately zero and depends only weakly on η since $\boldsymbol{d}(\eta)$ is asymptotically orthogonal to the columns of $\boldsymbol{F}$ for $\eta \geq M$. Therefore, we have in practice found that only the numerator in the cost function is sufficient to find the coarse estimate of η. On the Fourier grid, the numerator can be computed using a single FFT whereas the denominator requires $2M$ FFTs.

3. RESULTS

In this section, we demonstrate the applicability of the proposed method in both a simulated and a real environment. The former is necessary to be able to compare the produced estimates to a ground truth which is unknown and not well defined in a real environment. Specifically, we evaluated the estimator for three different source signals: (1) a white Gaussian noise signal, (2) a stereo music signal (track 61 on the EBU SQAM cd [15]), and (3) a stereo speech signal (track 4 on the Archimedes CD [16]). All signals were played back and recorded at a sampling rate of 44.1 kHz. The source signals to the loudspeakers were also recorded to remove internal delays in the PC and the sound card. Data frames of four seconds were obtained with a 75 % of overlap between the successive frames. These data were down-sampled by a factor of four to 11025 Hz since the 3" loudspeakers used in the measurements and shown in Fig. 1 have a very non-linear response at the higher frequencies. A MATLAB implementation of the proposed algorithm can process this amount of data in real-time on a standard desktop PC. For this sampling frequency and a speed of sound of 343 m/s, the sampling grid corresponds to a resolution of 3.1 cm. The code for running the simulation and making the measurements is available from `http://kom.aau.dk/~jkn/publications/publications.php`.

Fig. 2 shows an excerpt of the results of the simulation where the sources were assumed to be point sources and artificial reverberation [17] was added with a reverberation time of 0.5 seconds. From the figure and Table 1, we see that we got sub-millimetre accuracy for all source signals. From Fig. 3 and Table 1, we see that the variation of the estimates increased in the real environment despite that the loudspeakers were closer together. The main reason for this is that loudspeakers are not omnidirectional point sources. Instead, especially the higher frequencies are attenuated from one loudspeaker to the other when the loudspeakers are configured in a stereo setup as in Fig. 1, i.e., they are not pointed towards each other. Moreover, the acoustic centre of the loudspeaker is typically in front of the loudspeaker and frequency dependent [18].

Although not shown here, outliers in the estimated distances occur occasionally. These happen typically in very silent parts of the music/speech and can be removed by using a sound activity detector or by post-processing the computed estimates using a smoothing algorithm. However, even without these heuristics, we are able to estimate the transceiver distance to a millimetre precision for even a modest sampling frequency.

4. CONCLUSION

We have here taken a first step in the direction of using music or speech signals for the localisation of a number of loudspeakers. Specifically, we have considered the simplest case where we have to estimate the distance between two loudspeakers, each equipped with a single microphone. We derived an ML estimator for this problem and demonstrated that it could be used to obtain real-time distance estimates to

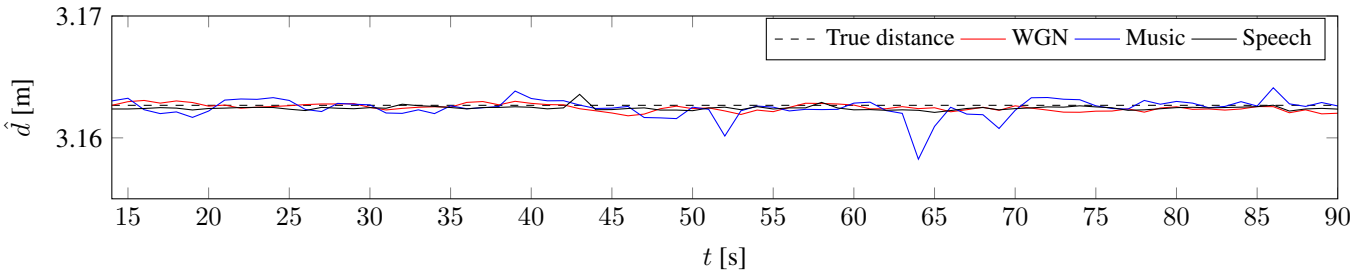

Fig. 2. The estimated distance for three source signals in a simulated environment.

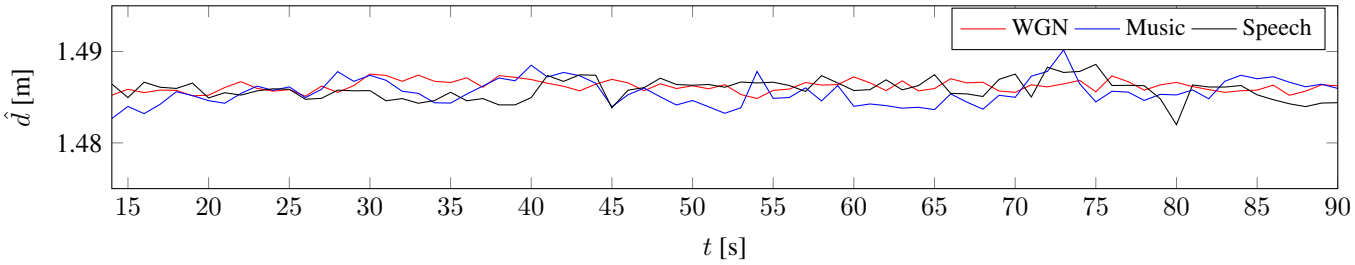

Fig. 3. The estimated distance for three source signals in a real environment.

within an accuracy of one millimetre for even a low sampling frequency. Only frame-by-frame processing was considered in this paper, but outliers can be removed and higher accuracy can be achieved by smoothing the computed estimates.

5. REFERENCES

[1] F. Rumsey, D. Griesinger, T. Holman, M. Sawaguchi, G. Steinke, G. Theile, and T. Wakatuki, "Multichannel surround sound systems and operations," The Audio Engineering Society, Tech. Rep. AESTD1001.0.01-05, 2001.

[2] C. Kyriakakis, "Fundamental and technological limitations of immersive audio systems," *Proc. IEEE*, vol. 86, no. 5, pp. 941–951, 1998.

[3] B. Gunel, "Loudspeaker localization using B-format recordings," in *Proc. IEEE Workshop on Appl. of Signal Process. to Aud. and Acoust.* IEEE, 2003, pp. 59–62.

[4] B. Atkinson, T. Blank, M. Isard, J. D. Johnston, and K. Olynyk, "An internet protocol (IP) sound system," in *Proc. AES Convention*. Audio Engineering Society, 2004.

[5] S. Choisel, G. G. Martin, and M. Hlatky, "Loudspeaker position estimation," US Patent 8 279 709 B2, 2012.

[6] G. Del Galdo, M. Lang, J. A. Pineda Pardo, A. Silzle, and O. Thiergart, "Acoustic measurement system for 3-D loudspeaker set-ups," in *Audio Eng. Soc.* Audio Engineering Society, 2010.

[7] A. J. V. Leest and D. W. Schobben, "Method of and system for determining distances between loudspeakers," US Patent 7 864 631 B2, 2011.

[8] F. Kolbeck, G. Del Galdo, I. Sobieraj, and T. Bliem, "Loudspeaker localization based on audio watermarking," in *Proc. AES Convention*. Audio Engineering Society, 2012.

[9] W. S. Torgerson, "Multidimensional scaling: I. theory and method," *Psychometrika*, vol. 17, no. 4, pp. 401–419, 1952.

[10] M. Crocco, A. Del Bue, and V. Murino, "A bilinear approach to the position self-calibration of multiple sensors," *IEEE Trans. Signal Process.*, vol. 60, no. 2, pp. 660–673, 2012.

[11] J. R. Jensen, J. K. Nielsen, M. G. Christensen, and S. H. Jensen, "On frequency domain models for TDOA estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015.

[12] C. Zhang, D. Florencio, and Z. Zhang, "Why does PHAT work well in low noise, reverberative environments?" in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2008, pp. 2565–2568.

[13] Y. Rui and D. Florencio, "Time delay estimation in the presence of correlated noise and reverberation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2. IEEE, 2004, pp. 133–136.

[14] A. Antoniou and W.-S. Lu, *Practical Optimization: Algorithms and Engineering Applications*. Springer, Mar. 2007.

[15] E. B. Union, "Sound quality assessment material recordings for subjective tests: Users' handbook for the EBU SQAM CD," European Broadcasting Union, Tech. Rep. EBU – TECH 3253, 2008.

[16] Bang & Olufsen. (1992) Music for archimedes. Compact disc CD B&O 101.

[17] E. A. P. Habets. (2010) Room impulse response generator. Ver. 2.0.20100920. [Online]. Available: http://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator

[18] F. Jacobsen, S. B. Figueroa, and K. Rasmussen, "A note on the concept of acoustic center," *J. Acoust. Soc. Am.*, vol. 115, no. 4, pp. 1468–1473, 2004.