Robust Multichannel Acoustic Time Delay Estimation in Reverberant Environments

Marius H. Hennecke, Gernot A. Fink

Robotics Research Institute, TU Dortmund University, Germany

Abstract

Time delay estimation is a basic building block of an acoustic source localization system. Such a system must cope with adverse acoustic conditions to be generally applicable, i.e. high reverberation times and low signal-to-noise ratios. In general, better results can be achieved if multiple channels are considered. We propose an extension of a mutual information based delay estimator to the multichannel case. Additionally, we give a unified description of the proposed estimator, the generalized cross-correlation and its multichannel counterpart. Using simulated and real recordings, a comparative performance evaluation shows the robustness of the proposed method under adverse acoustic conditions.

1 Introduction

Essential premise for any intelligent system in complex real world scenarios is to determine the focus of attention [10]. Acoustic source localization (ASL) is a means to accomplish this task for an audio-based perception system. A video-conferencing application with one or more active cameras for example could use the location of the active speaker for selecting or steering a camera, such that the speaker is optimally visible. A common building block in ASL systems is time delay estimation (TDE).

The goal of TDE is to measure the time difference of arrival (TDOA) of a source signal between two or more receivers. Employing a spatially distributed sensor array, one can deduce the location of the source from multiple TDOAs with a variety of techniques, e.g. [4, ch. 8]. Still, the main challenge is the robustness of the TDE against reverberation and background noise inherent in real environments. In this paper, we focus on a robust information theoretical multichannel TDE as a basis for ASL; however, the presented results are not constrained to the domain of acoustics alone. A comparative performance evaluation of the proposed method shows its robustness in terms of noise and reverberation, using artificially reverberated signals as well as recordings made in a typical conference room.

2 Related Work

A number of TDE methods exist, using two channels only as well as techniques which exploit the redundancy among multiple microphone channels. A very common approach for determining the TDOA for the two-channel case is the generalized cross-correlation (GCC) [7]. It is based on the assumption of an ideal direct path signal propagation. Thus, the maximum of the cross-correlation of a channel pair marks the TDOA. Applying the GCC to reverberant signals, this is not necessarily the case. Reflections lead to local maxima and could even mask the true TDOA. In order to shape the cross-correlation for improved TDOA estimation, different pre-filters were proposed in the literature [7]. Most notably the phase transformation (PHAT), which is a whitening of the cross-spectrum, is motivated by the fact that a pure time delay results in a phase shift and leaves the signal's amplitude unchanged. Despite of its solely heuristic nature, PHAT as a pre-filter has shown robustness under mild reverberation and noise.

Unlike the GCC method adaptive eigenvalue decomposition (AED) [2] models reverberation explicitly. Using the covariance matrix of the signals AED blindly estimates the channel impulse responses. For delay estimation only the direct path component in a reverberated signal is important, which is marked by the maximum of an impulse response. A TDOA estimate using AED is thus given by the distance of impulse response peaks of corresponding channel pairs. An overview of extensions of the AED method to frequency-domain block-processing and the multichannel case can be found in [5].

Employing more than two sensors, the multichannel cross-correlation (MCCC) [5] can be considered as a generalized version of the GCC. The MCCC algorithm estimates only one TDOA for an array in such a way that the correlation among all channels is maximized. Similar to the PHAT pre-filter, a prewhitening could be incorporated in the MCCC algorithm [5] which is then equivalent to the GCC for a two channel array.

A more general formulation of multichannel TDE using entropy as a statistical measure for uncertainty is given by Benesty et al. [3]. The joint entropy of all array channels is evaluated for all realizable TDOAs and the delay that leads to the smallest joint entropy is used as a TDOA estimate. If the multivariate random variable modeling the channel signals is assumed to follow a Gaussian distribution, the minimum joint entropy approach is equivalent to the MCCC algorithm. The most prominent source signal in the context of ASL, however, is speech, which follows a Laplace distribution. Integrating this assumption into the entropy based approach, a better TDOA estimation could be achieved [3].

Modeling reverberation without an explicit estimate of the impulse response is the goal of Moddemeijer's delay estimator [8, 11]. In the anechoic case, a sample in one channel corresponds only to a delayed version of this sample in a second channel. This exact correspondence does not hold in a reverberant environment. Due to reflections neighbouring samples contain information about the time delay as well. Moddemeijer incorporates this insight into the delay estimation using a minimum mutual information (MI) formulation. An extension, however, to the multichannel case is missing, which is the main contribution of this paper.

3 Robust Time Delay Estimation

The acoustic path between a source and a receiver in a reverberant environment is typically modeled as a linear time invariant system. Such a system is described by an impulse response which contains the direct path component, i.e. the strongest peak, between sender and receiver and all reflections induced by the environment. The resulting signal $x_i[n] = h_i * s[n] + w_i[n]$ for channel i is composed of the

ISBN 978-3-8007-3300-2 © VDE VERLAG GMBH · Berlin · Offenbach

source signal s convolved with the corresponding impulse response h_i and additive noise w_i. The impulse response depends on the source position, environmental conditions such as temperature and humidity and even includes changes of the room layout, e.g., furniture and people moving around. We assume that w_i is a zero-mean Gaussian random process and that it is uncorrelated with s.

In the following we will give a unified description of the minimum joint entropy approach [3], Moddemeijer's MI delay estimator [8, 11] and our proposed extension of the latter to the multichannel case.

First, we define the vector $\boldsymbol{x}_i(n,m)$ comprised of M future samples of channel i starting at time index n as

$$\boldsymbol{x}_i(n,m) = (x_i[n+f_i(m)], x_i[n+f_i(m)+1], \ldots, x_i[n+f_i(m)+M])^{\mathrm{T}} \quad . \quad (1)$$

It is delayed by $f_i(m)$ samples with respect to the first channel. Hence, $f_1(m) = 0$. In general, the delay function $f_i(m)$ depends on the array geometry and could even account for two or three TDOAs in the near-field case [3]. For simplicity and in order to focus on the proposed method, we assume that the sources lie in the far-field and that a uniform linear array (ULA) is used. The general applicability of the proposed method is not constrained by these assumptions. Combined with the fact that—as we are working with real signals—only integer delays are realizable, the delay function simplifies to $f_i(m) = \lceil (i-1)m \rfloor$, with $\lceil \cdot \rfloor$ denoting rounding.

Concatenating all $\boldsymbol{x}_i(n,m)$ results in the column vector

$$\boldsymbol{x}(n,m) = \left(\boldsymbol{x}_1^{\mathrm{T}}(n,m), \boldsymbol{x}_2^{\mathrm{T}}(n,m), \ldots, \boldsymbol{x}_N^{\mathrm{T}}(n,m)\right)^{\mathrm{T}} \quad (2)$$

of length $N(M+1)$. Considering $\boldsymbol{x}(n,m)$ as the realization of a multivariate random variable with a zero-mean Gaussian density, the joint entropy [6]

$$H(\boldsymbol{x}(n,m)) = \frac{1}{2} \ln \left((2\pi \mathrm{e})^{N(M+1)} \det \boldsymbol{R}_n(m) \right) \quad (3)$$

mainly depends on the covariance matrix of the concatenated channel blocks. For signals with time-varying characteristics such as speech, the covariance matrix

$$\boldsymbol{R}_n(m) = \mathrm{E}\left\{ \boldsymbol{x}(n,m) \boldsymbol{x}^{\mathrm{T}}(n,m) \right\} \quad (4)$$

is commonly estimated on short signal blocks where $\mathrm{E}\{\cdot\}$ represents mathematical expectation.

To estimate a TDOA we seek for the delay $\hat{\tau}_n$ that leads to the least uncertainty, i.e., the minimum joint entropy

$$\hat{\tau}_n = \arg\min_m H(\boldsymbol{x}(n,m)) = \arg\min_m \ln \det \boldsymbol{R}_n(m) \quad . \quad (5)$$

The possible values for m are constrained to realizable integer delays $m \in \mathbb{N} \cap [-d f_s/c, d f_s/c]$. Here, f_s is the sampling frequency, d is the smallest Euclidean distance over all sensor pairs and c is the speed of sound.

The described algorithm leads to the MCCC method [5] and the equivalent minimum joint entropy approach [3] without using future samples, i.e. $M = 0$. Additionally, if a pre-whitening of the signal is performed, it is similar to GCC-PHAT for $M = 0$ and the two-channel case ($N = 2$). Incorporating future samples ($M > 0$) and two channels ($N = 2$), the proposed algorithm is equivalent to Moddemeijer's information theoretic delay estimator [8]. Our proposed method extends the latter to the multichannel case, i.e. $N > 2$.

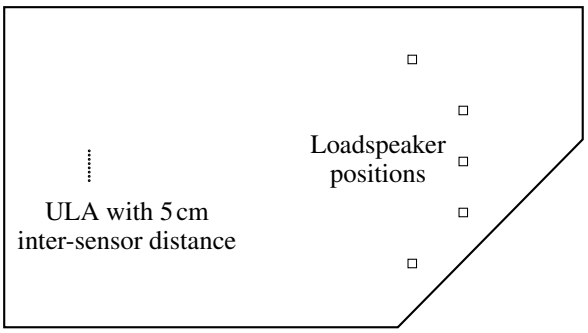


Figure 1: Recording setup in the FINCA

4 Experiments

In order to show the robustness of the proposed multichannel TDE method in the presence of noise and reverberation the following section includes the results in comparison to Moddemeijer's MI estimator GCC-PHAT and MCCC. Recordings have been made in a smart conference room, the FINCA [9], with a reverberation time of approximately 600 ms. To allow for a variation of the reverberation time and the signal-to-noise ratio (SNR), an analysis using artificially reverberated signals is included in this section. We employ the image method [1] for synthesizing reverberated signals with parameters (i.e. room geometry, acoustic damping factors) matching the FINCA as closely as possible.

The FINCA as depicted in Fig. 1 has a size of 6.8 m × 3.8 m with a height of 2.6 m. The reverberation time (T_{60}) in the simulation is ranging from 200 ms up to 800 ms and the signal-to-noise ratio is varied from 20 dB to −10 dB. A uniform linear array (ULA), composed of eight omnidirectional microphones with an inter-sensor distance of 5 cm is used throughout the experiments. Its spatial location during recording stays the same in simulation. Five different positions with a minimum distance from the ULA of 3.3 m—ensuring the far-field assumption holds—are taken as source positions for an utterance with a duration of 8 s played back by a loudspeaker facing the ULA. The sampling rate is fixed to $f_s = 48$ kHz in order to achieve a high TDOA resolution. The identical utterance is used to generate the artificially reverberated signals. White noise is added to achieve the desired range of SNRs.

For a fair comparison of the two-channel only GCC-PHAT and Moddemeijer's MI estimator with our proposed multichannel method, a minimum least squares (MLS) combination of all pairwise TDOA estimates is used [4, chap. 8]. The MLS estimate is

$$\hat{\tau}_n = \arg\min_m \sum_{i=1}^{P} (\hat{\tau}_n^{(i)} - f_{d_i/d}(m))^2 \quad , \quad (6)$$

where d_i is the microphone distance of pair i and d is the ULAs inter-microphone distance. Performance of the compared TDE methods is measured as the root mean square error (RMSE) in degrees over all positions and estimates. For sub-sample resolution of the TDE estimates, a quadratic interpolation is employed.

The influence of the number of future samples M is evaluated first. Figure 2 shows the RMSE for the TDE task on the simulated signals with our proposed approach for a decreasing SNR and increasing M. The reverberation time is $T_{60} = 500$ ms and for this experiment the four inner

ISBN 978-3-8007-3300-2 © VDE VERLAG GMBH · Berlin · Offenbach

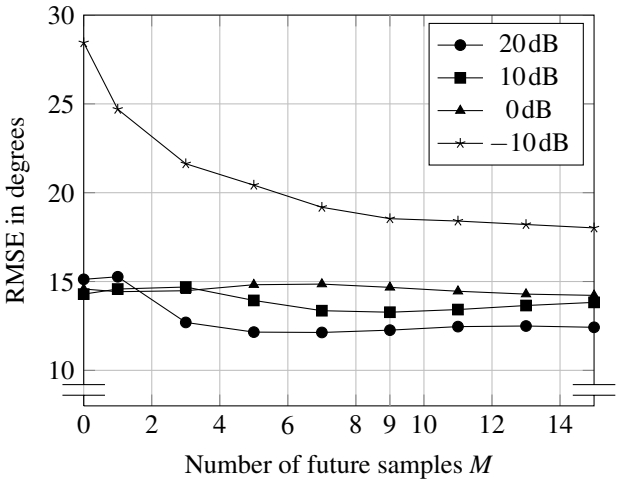


Figure 2: RMSE for the proposed method for an increasing number of future samples M, four channels and different SNRs using reverberated signals with $T_{60} = 500\,\text{ms}$.

N	GCC-PHAT	MI	MCCC	Proposed
2	19.6	10.6	29.8	10.6
4	11.5	9.9	13.9	9.2
6	10.8	9.6	11.8	7.5
8	8.1	8.9	5.8	5.9

(a) SNR=0 dB

N	GCC-PHAT	MI	MCCC	Proposed
2	23.6	11.9	25.7	11.9
4	17.1	11.6	20.3	10.4
6	15.0	11.6	12.1	9.0
8	13.1	11.2	10.6	8.7

(b) SNR=−10 dB

Table 1: RMSE in degrees using recorded signals with an approx. SNR of −10 dB for increasing channels N. Number of neighbouring samples used is fixed to $M = 9$.

channels of the ULA are used. For different reverberation times the results are similar but, as expected, with a general increase in the RMSE for higher reverberation times. Incorporating more future samples leads to better RMSE results. The improvement is especially visible for the worst case examined with an SNR of −10 dB. Based on these findings $M = 9$ is chosen for the following experiments. It is a good trade-off between RMSE performance improvement and computational costs.

The results of a comparative performance evaluation for increasing reverberation time between Moddemeijer's mutual information approach, MCCC and the proposed method are shown in Fig. 3. Due to its poor performance with the simulated reverberant signals under these extremely low SNR conditions, results for GCC-PHAT are not included in this comparison. In all experiments the covariance matrix (4) is estimated on Hamming windowed blocks with a length of 300 ms and a 50 % overlap. For the proposed approach, the number of future samples is chosen as $M = 9$ in accordance with the aforementioned result. The algorithms show similar performance in terms of RMSE down to an SNR of 0 dB. Even adding more channels gives no significant improvement for an SNR of 0 dB. This is shown for the four channel case in Fig. 3a in a direct comparison to the eight channel case in Fig. 3b. Decreasing the SNR down to −10 dB even further (see Fig. 3c and Fig. 3d), the proposed approach shows superior results over the whole range of examined reverberation times in contrast to the other methods. In general, the results show a significant improvement from the four channel to the eight channel case. Exemplarily, for a reverberation time of $T_{60} = 0.3\,\text{s}$ the RMSE for the proposed approach drops from 15° to 11.6°. The promising results on simulated reverberant signals under very high noise and reverberation conditions give an indication, that our approach leads to a robust TDE estimator.

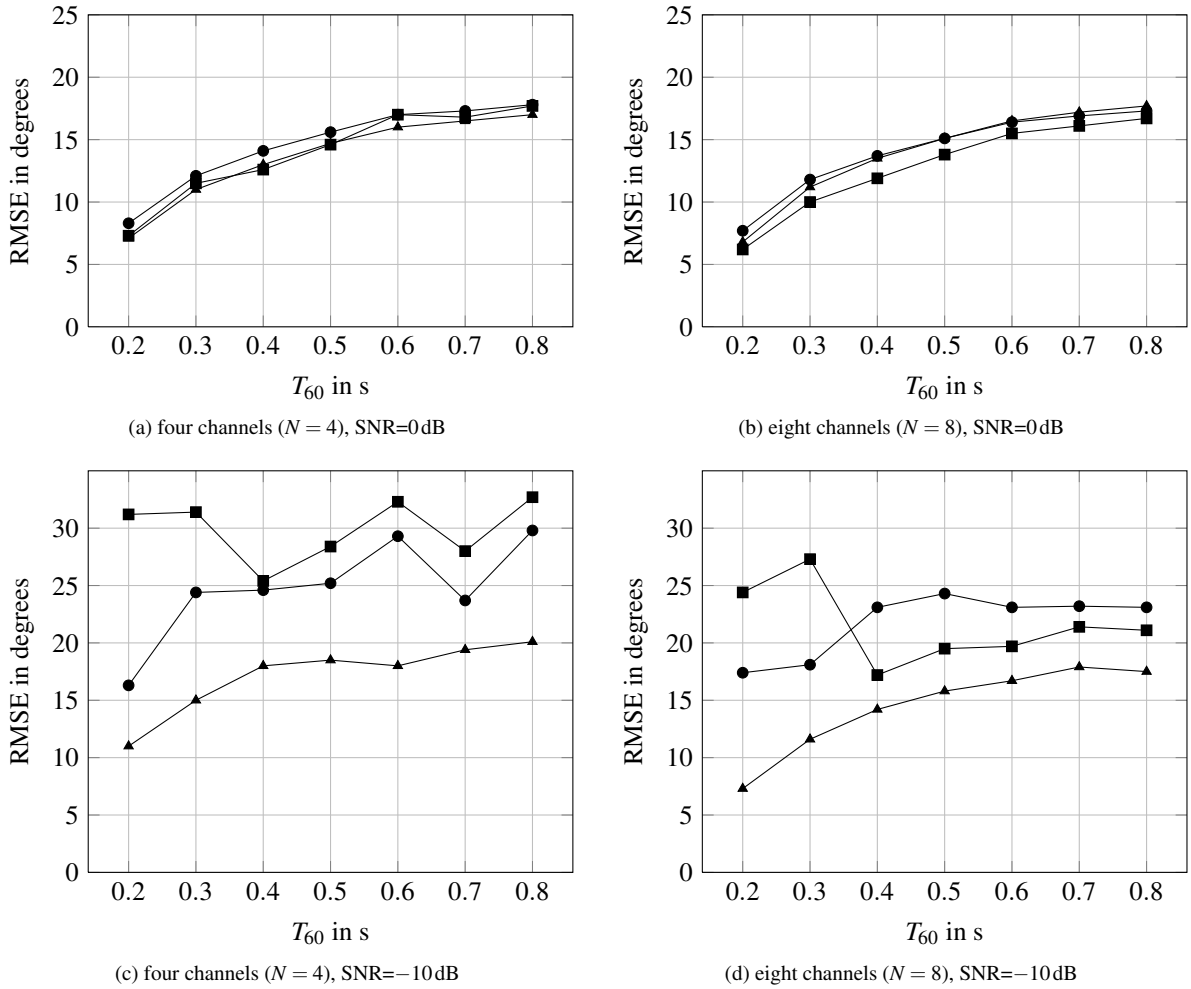
In the following, experiments with recordings made in a reverberant conference room ($T_{60} \approx 600\,\text{ms}$) will show the performance of the different methods under similar adverse acoustic conditions. The results are shown in Tab. 1 for an increasing number of channels N and a fixed number of neighbouring samples $M = 9$ for Moddemeijer's MI approach and our proposed multichannel extension. Qualitatively, all methods behave similar, i.e. the RMSE decreases if more channels are used and increases for worse SNR conditions. In contrast to the simulated signals the results for GCC-PHAT are included. The good results on the recorded signals could be attributed to the non i.i.d noise, as is the case in the simulation. The main source of noise in the FINCA (see Fig. 1) are fans of electrical equipment in the intermediate ceiling of the room which leads to an approximately diffuse noise field. Quantitatively, the results are approximately twice as good as the simulation predicts, e.g., for $N = 8$ and an SNR of −10 dB the proposed approach leads to an RMSE of 8.7° (Tab. 1) in contrast to 16.7° (Fig. 3d) in the simulation. The authors consider the discrepancy in the results between the simulation and the real conference room recordings as an artifact due to the assumed shoe-box model in the image method [1], which is not met by the FINCA. Still, the proposed method performs better in all experiments, in simulation as well as with real room recordings.

No explicit knowledge about the noise characteristics is needed in our approach. This is demonstrated by the results obtained with the recordings for which the white noise assumption used in the derivation of the proposed delay estimator (5) does not hold but gives still good results. Furthermore, the room reverberation is not modelled explicitly. In contrast to GCC no pre-filter of any kind is applied nor needed. Especially the equal weighting of all frequencies of GCC-PHAT, which could boost insignificant frequency bins, is avoided. Consequently, our proposed approach is a robust TDOA estimator under adverse acoustic conditions.

5 Conclusion

A robust TDE is an important component of an ASL system. In this paper, we presented an extension of a mutual information based two-channel estimator to the multichannel case. Additionally, we gave a unified description of the MCCC, Moddemeijer's MI estimator and our proposed approach. A comparative performance evaluation showed the robustness of the latter in simulation and with recordings made in a real conference room under adverse acoustic conditions. Additionally, no pre-filtering is needed in our approach.

ISBN 978-3-8007-3300-2 © VDE VERLAG GMBH · Berlin · Offenbach



(a) four channels ($N = 4$), SNR=0 dB

(b) eight channels ($N = 8$), SNR=0 dB

(c) four channels ($N = 4$), SNR=−10 dB

(d) eight channels ($N = 8$), SNR=−10 dB

Figure 3: RMSE in degrees for two different SNRs, increasing reverberation time and channel count using simulated reverberant signals. LS MI (—●—), MCCC (—■—) and proposed method with $M = 9$ (—▲—).

References

[1] J. B. Allen and D. A. Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979.

[2] J. Benesty. Adaptive eigenvalue decomposition algorithm for passive acoustic source localization. *The Journal of the Acoustical Society of America*, 107(1):384–391, Jan. 2000.

[3] J. Benesty, Y. Huang, and J. Chen. Time delay estimation via minimum entropy. *IEEE Signal Process. Lett.*, 14(3):157–160, Mar. 2007.

[4] M. S. Brandstein and D. Ward, editors. *Microphone Arrays – Signal Processing Techniques and Applications*. Digital Signal Processing. Springer, 2001.

[5] J. Chen, J. Benesty, and Y. Huang. Time delay estimation in room acoustic environments: An overview. *EURASIP Journal on Applied Signal Processing*, page 19, 2006.

[6] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, New York, July 2006.

[7] C. H. Knapp and G. C. Carter. The generalized correlation method for estimation of time delay. *IEEE Trans. Acoust., Speech, Signal Process.*, 24(4):320–327, Aug. 1976.

[8] R. Moddemeijer. An information theoretical delay estimator. In *Ninth Symposium on Information Theory in the Benelux*, pages 121–128, Enschede (NL), 1988.

[9] T. Plötz. The FINCA: A Flexible, Intelligent eNvironment with Computational Augmentation, 2007. www.finca.irf.de.

[10] B. Schauerte, T. Plötz, and G. A. Fink. A multi-modal attention system for smart environments. In *Computer Vision Systems*, volume 5815 of *LNCS*, pages 73–83. Springer Berlin / Heidelberg, 2009.

[11] F. Talantzis, A. G. Constantinides, and L. C. Polymenakos. Estimation of direction of arrival using information theory. *IEEE Signal Process. Lett.*, 12(8):561–564, Aug. 2005.

ISBN 978-3-8007-3300-2 © VDE VERLAG GMBH · Berlin · Offenbach