ROBUST SPEAKER LOCALIZATION IN MEETING ROOM DOMAIN

P. Pertilä, M. Parviainen

{Pasi.Pertila, Mikko.P.Parviainen}@tut.fi
Tampere University of Technology, Institute of Signal Processing
P.O.Box 553 FI-33101, Tampere, Finland

ABSTRACT

Speech signal is perturbed in real environments due to room acoustics and noise sources. The speech wavefront is received by a microphone array, which is used to determine the wavefront direction of arrival (DOA). DOA estimates from several spatially separated arrays are used to locate the speaker. Array channels may sometimes be corrupted by noise. In these cases, a DOA estimate may differ from the actual direction of the speaker. Using such estimates in localization could deteriorate the position estimate.

A criterion is presented for DOA exclusion. A subset of DOA estimates is chosen that minimizes this criterion. The method is numerically shown to improve localization robustness. Recordings from a real meeting room are studied.

Index Terms— Acoustic position measurement, Acoustic tracking, Delay estimation, Acoustic arrays, Reliability

1. INTRODUCTION

The goal of acoustic localization is to determine the position of a sound source using measured audio signals. The applications of acoustic localization range from large and medium scale outdoor surveillance [1][2][3][4] to indoor speaker localization [5][6][7] used in e.g. automated camera management [8]. Speaker localization is also a basic part of smart meeting room infrastructure [9]. This work focuses on single speaker localization in a real meeting room environment with multiple microphone arrays.

Numerous methods exists for estimating the location of a sound source in free space based on the observed signals and microphone array geometry. Time delay estimation (TDE) between microphone pairs is a popular method used in locating sound wavefront origin. The source distance and array geometry determine whether the source is in *near field* or *far field*. Near field methods measure the wavefront curvature and locate the source directly from the time delays [6][7].

In far field, time delays can be used to calculate DOA of a sound source instead of location, because the observed wavefront is approximated to be planar [10]. The intersection point of DOA measurement lines from spatially separated arrays is the source location [2][3]. However, it is not realistic to assume that the lines along the estimated directions intersect in a single point in three dimensional space. In practice, some localization criterion is minimized over the space, e.g., the shortest distance from the observed direction lines to the source [3] or angle deviation between measured directions and array-to-source directions [4].

In this work, errors in array geometry, location, and orientation are omitted. It is also assumed that the effects of possible source movement during the time window required by DOA estimation are insignificant. For simplicity, it is further assumed that channels inside an array and between the arrays are synchronized. The precision of DOA-based localization is assumed to depend only on the accuracy of individual DOA estimates.

DOA-based localization assumes that there exists a line of sight from each array to the source. If the arrays do not observe the same sound source or the sound is reflected, the location estimate may deteriorate. Without prior knowledge about the surroundings and propagation conditions, these direction measurements are considered outliers [11].

To some extent, DOA reliability can be measured at individual array level. A method exists to estimate the reliability of a TDE-based DOA estimate. The method is based on the assumption that the time delays are defined by a plane wave. If the estimates disagree with this assumption, it can be argued that they are not produced by a far field sound source [12]. It is worth pointing out that the observed wavefront direction is not necessary the true source direction. Detecting the difference at the array level may not be trivial.

Due to the reasons presented above, faulty DOA estimates should be excluded before localization in order to avoid gross errors. The exclusion method should be independent of the actual DOA estimation procedure. This would make the method more suitable for heterogenous environments and multimodal sensors. In GPS and navigation, receiver autonomous integrity monitoring (RAIM) is used in fault detection and exclusion, based on overdetermined solutions. Several criteria, such as Least-Squares-Residuals can be used as a basis of observation exclusion [13]. In this work a distance-based criterion for DOA exclusion is presented. The method utilizes the redundancy of DOA measurements. An exclusion algorithm is tested with simulated DOA data and real data using spatially separated arrays in a real meeting room.

The outline of this work is as follows. Section 2 presents an overview of the localization system. The distance-based criterion and an exclusion algorithm are presented in Section 3. Section 4 discusses the experiments as listed next. A description of an existing performance evaluation metric is given. Then simulations are used to characterize the behavior of the exclusion algorithm compared to the normal least squares approach. The section then presents real data measurements. The results for the real data measurements are given again for both methods. Section 5 discusses the presented DOA exclusion method. A summary is given in Section 6.

2. LOCALIZATION SYSTEM

The baseline for the localization system used in this work was originally developed for the NIST Spring 2005 evaluation [14] and was improved for the CLEAR 2006 evaluation [15]. A short system description is given here, details are presented in [15]. The localization system is designed for locating a single speaker in a meeting room environment with multiple microphone arrays. The system block

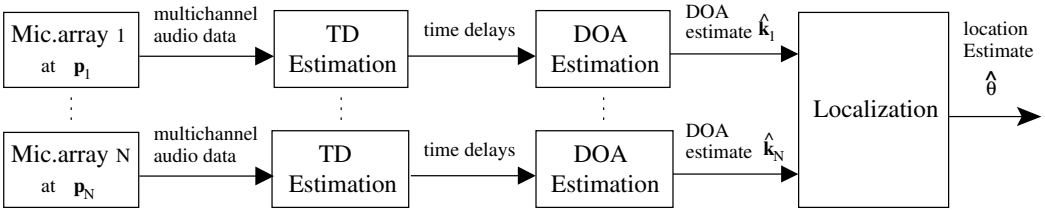


Fig. 1. The speaker localization system block diagram. Speaker direction is measured based on time differences of arrival measurements at each microphone array $i \in 1, ..., \mathrm{N}$. The direction estimates $\hat{\mathbf{k}}_i$ are combined in the localization step using Algorithm 1.

diagram is given in Fig. 1.

First, time delays of the planar sound wave between microphones inside each array i are estimated using a weighted cross-correlation method GCC-PHAT [16]. Then for each array, a least squares (LS) method is used to estimate a DOA vector $\hat{\mathbf{k}}_i$ utilizing the measured time delays [10]. The DOA vectors $\mathbf{k}_i$ are 3D Cartesian vectors $\mathbf{k}_i = [k_{i,x}, k_{i,y}, k_{i,z}]^{\mathrm{T}}$ where $i \in 1, ..., \mathrm{N}$ and N is the number of arrays. The arrays are located at coordinates $\mathbf{p}_i = [p_{i,x}, p_{i,y}, p_{i,z}]^{\mathrm{T}}$. Finally, a least squares method is used to combine the DOA estimates to produce a source location estimate $\hat{\boldsymbol{\theta}}$ [3][1]

$$\hat{\boldsymbol{\theta}} = (\mathrm{N}I_{(3)} - \hat{K}\hat{K}^{\mathrm{T}})^{-1} A \,\mathrm{diag}(I_{(\mathrm{N})}), \tag{1}$$

where $\hat{K} = [\hat{\mathbf{k}}_1, ..., \hat{\mathbf{k}}_\mathrm{N}]$, $I_{(\mathrm{N})}$ is an $\mathrm{N} \times \mathrm{N}$ identity matrix, and

$$A = [(I_{(3)} - \hat{\mathbf{k}}_1\hat{\mathbf{k}}_1^{\mathrm{T}})\mathbf{p}_1, ..., (I_{(3)} - \hat{\mathbf{k}}_\mathrm{N}\hat{\mathbf{k}}_\mathrm{N}^{\mathrm{T}})\mathbf{p}_\mathrm{N}]. \tag{2}$$

Figure 2 illustrates the localization geometry.

3. DISTANCE-BASED DOA ESTIMATE EXCLUSION

The distance from DOA measurement line to the estimate can be calculated. The distance d_i for each array i is [3]

$$d_i = \|\mathbf{p}_i + \mathrm{Proj}_{\hat{\mathbf{k}}_i}\mathbf{k}_i - \hat{\boldsymbol{\theta}}\|, \tag{3}$$

where $\mathrm{Proj}_{\hat{\mathbf{k}}_i}\mathbf{k}_i$ is defined as

$$\mathrm{Proj}_{\hat{\mathbf{k}}_i}\mathbf{k}_i = \frac{\hat{\mathbf{k}}_i \cdot \mathbf{k}_i}{\|\hat{\mathbf{k}}_i\|^2}\hat{\mathbf{k}}_i. \tag{4}$$

The LS solution (1) minimizes the squares of the distances (3) over all arrays. See Fig. 2. It is intuitive that the average sum of distances (3), where $i \in 1, ..., \mathrm{N}$ correlates with the error of the location solution (1). This reasoning is used to exclude DOA estimates that contribute highly to the distance and thus increase the localization error.

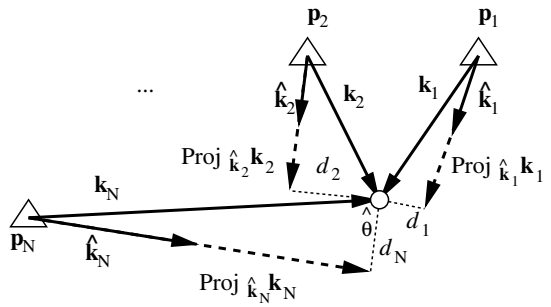


Fig. 2. Localization with multiple microphone arrays at $\mathbf{p}_i$. The distance between the measurement line and the location estimate $\hat{\boldsymbol{\theta}}$ is termed distance criterion d_i.

[1]The weights of the DOA estimates used in [3] are omitted.

The objective of the DOA exclusion method is to find a DOA subset with the smallest average distance criterion. The minimum subset size is set to three arrays to guarantee some redundancy, even though two arrays is the minimum requirement for (1).

For each possible subset of three or more arrays Ω_n calculate the subset's LS location estimate $\hat{\boldsymbol{\theta}}_n$ using (1) and (2). Then assign the subset an average distance criterion value, ADC(n):

$$\mathrm{ADC}(n) = \frac{1}{|\Omega_n|}\sum_{k\in\Omega_n} d_{k,n}, \tag{5}$$

where n is the subset index, $n = 1, ..., \sum_{s=3}^{N}\binom{N}{s}$, $|\Omega_n|$ is the number of arrays in the subset n, and $d_{k,n}$ the distance criterion contributed by the kth array in the subset n, see Eq. (3). Finally, choose ADC location estimate $\hat{\boldsymbol{\theta}}_{\mathrm{ADC}}$ to be the LS location estimate $\hat{\boldsymbol{\theta}}_{\tilde{n}}$ of DOA subset $\tilde{n}$

$$\tilde{n} = \underset{n}{\mathrm{argmin}}\ \mathrm{ADC}(n), \quad \hat{\boldsymbol{\theta}}_{\mathrm{ADC}} = \hat{\boldsymbol{\theta}}_{\tilde{n}}, \tag{6}$$

that minimizes the average distance criterion. The presented method can be regarded as a special case of observation subset testing. Typically a maximal subset with the smallest acceptable test statistics is chosen with respect to a predefined threshold [13]. Here, the threshold is omitted. The presented method is summarized in Algorithm 1.

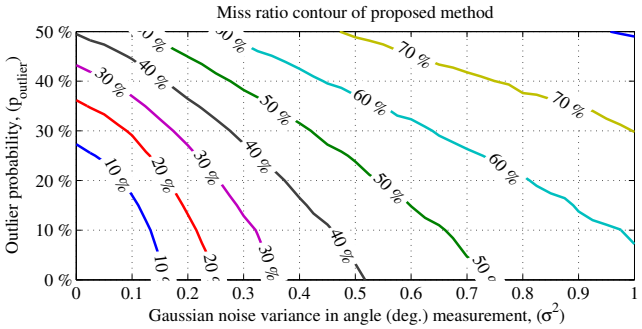
4. EXPERIMENTS

4.1. Localization performance metrics

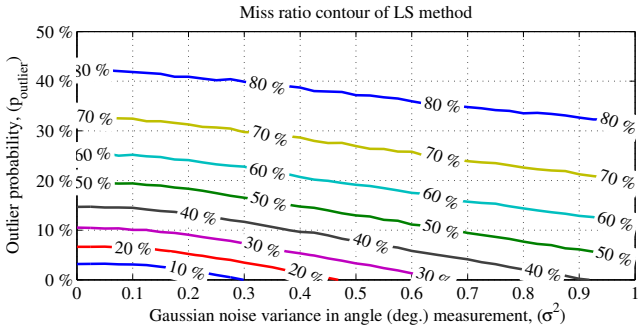
The localization performance is evaluated with two metrics, the miss ratio and the average estimate error (AEE) [9]. A location estimate is

Algorithm 1: ADC method for Speaker localization

for n = *1 to #array combinations* **do**
  ADC(n) $\leftarrow$ 0;
  $\hat{\boldsymbol{\theta}}_n \leftarrow$ calculate the LS solution of set Ω_n using (1);
  for *array* k = *1 to* $|\Omega_n|$ **do**
    $\mathbf{k}_k \leftarrow (\mathbf{p}_k - \hat{\boldsymbol{\theta}}_n)/\|\mathbf{p}_k - \hat{\boldsymbol{\theta}}_n\|$; (array-to-source vector)
    $\hat{\mathbf{k}}_k \leftarrow$ retrieve DOA measurement from array k;
    $\mathrm{Proj}_{\hat{\mathbf{k}}_k}\mathbf{k}_k \leftarrow$ calculate using (4);
    $d_{k,n} \leftarrow$ calculate using (3);
    ADC(n) $\leftarrow$ ADC(n) + $d_{k,n}$;
  end
  ADC(n) $\leftarrow$ ADC(n) / $|\Omega_n|$; (i.e. average distance)
end
Choose estimate $\hat{\boldsymbol{\theta}}_{\mathrm{ADC}}$ with minimum ADC(n), using (6);



(a) Results for the localization miss ratio of the ADC method.



(b) Results for the localization miss ratio of the LS method.

Fig. 3. Localization results in terms of miss ratio is given for the simulation data. The simulation geometry is described in Table 1. The x-axis is the variance of Gaussian noise present in a single DOA estimate σ^2. The y-axis is the probability that a single DOA is an outlier. It is noted that the surface area under the same miss ratio is larger in the average distance criterion (ADC) method than in the least squares (LS) method. See Subsection 4.2 for a detailed description of the results.

counted as a miss, if the distance of the estimate to the ground truth position is more than 50 cm. Miss ratio therefore describes what is the percentage of gross errors. AEE is calculated only for the non-miss estimates. It is the average distance between an estimate and the ground truth position. This metric describes estimate accuracy. In the simulations the AEE results are omitted.

4.2. Simulations

Simulations are used to compare the localization performance of the ADC (Algorithm 1) and LS (1) methods. The DOA estimates are simulated with varying amounts of two different types of disturbances. The simulation scenario consists of a continuous sound source and five microphone arrays. The arrays are located on the walls of a room of dimension 6 × 6 meters. The exact simulation setup is described in Table 1. The simulations are performed in 2D for simplicity. The DOA measurement model for each array is given in polar coordinates (α, r):

$$\alpha_i(m) = \begin{cases} \alpha_{i,\boldsymbol{\theta}} + \mathcal{N}(0, \sigma^2) & p > p_{outlier} \\ \mathcal{U}(0, 2\pi) & p \leq p_{outlier} \end{cases}, r_i = 1; \quad (7)$$

where m is a repetition number $m = 1, ..., 50000$, and $\alpha_{i,\boldsymbol{\theta}}$ is the ground truth direction at array i, and range r_i is unity. The disturbances are i.i.d Gaussian noise with zero mean and variance σ^2. The probability $p_{outlier}$ determines how probable it is that a DOA estimate is an outlier. An outlier is modeled as an uniformly distributed random direction estimate. In the simulation the degree of disturbances was varied to find out the typical behavior of the methods. The results for both methods are given in Fig. 3.

The ADC method used the same data set as the LS method. It is seen from Fig. 3 that the advantage of the ADC method is clear when the outlier probability increases. The miss ratio is better in the ADC method when a certain amount of outliers are present in the direction estimates. For example, if the outlier probability $p_{outlier} = 25$ % and Gaussian angle variance $\sigma^2 = 0.2$ the ADC method has a miss ratio of less than 30 % and the LS miss ratio is close to 60 %.

Table 1. Simulation geometry with the five microphone array locations $\mathbf{p}_1, \ldots, \mathbf{p}_5$ and the source location $\boldsymbol{\theta}$.

	$\mathbf{p}_1$	$\mathbf{p}_2$	$\mathbf{p}_3$	$\mathbf{p}_4$	$\mathbf{p}_5$	$\boldsymbol{\theta}$
x-coordinate [m]	0	2	6	6	2	1.5
y-coordinate [m]	3	0	0	6	6	1.5

4.3. Real data measurements

The ADC and LS methods are tested with a data set used in the CLEAR 2006 evaluation [9]. This database includes audio and video data, and is labeled into training and evaluation sets with respective durations of 3.6 and 3.2 hours. Note that the methods discussed in this paper do not require training. The recordings were performed in actual meeting room environments with a lecturer and an audience present. The data was recorded at several sites. In this work, only the data recorded at University of Karlsruhe is used. This data consists of 157 and 120 minute training and evaluation sets, respectively.

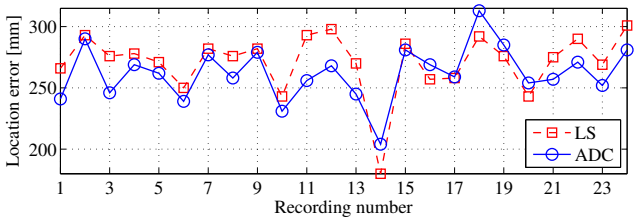
The recording room dimensions are 5.9 × 7.1 × 3.0 meters. Four microphone arrays are located on the walls of the room. Each array consists of four microphones. The arrays include three microphones on a horizontal line spaced 20 cm apart. The fourth microphone is 30 cm above the center microphone. Recordings were performed with a 44.1 kHz sampling rate with a resolution of 24 bits per sample. Other microphones and video cameras were also present, but are not considered by the described system.

The ground truth data consists of the active speaker's mouth coordinates, given in 3D Cartesian coordinates with one second intervals. If no speaker is active, no ground truth coordinates exists for that time instant. These non-active time instants are omitted in the performance scoring. For a complete description of the recordings and evaluation metrics, see [9].

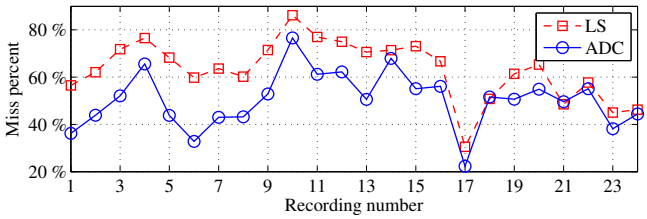
4.4. Real data results

The results are given in the metrics described in Subsection 4.1. The results for average estimate error (AEE) and miss ratio are presented in Fig. 4. In the results, only the evaluation data set was used. The evaluation data set consists of 24 different recording sessions. The average AEE scores for the LS and the ADC methods are 275 mm and 263 mm, respectively. Similarly, the average miss ratios are 62 % and 48 %.

The relatively large improvement of 14 percentage units in the average miss ratio and modest increase of 12 mm in the accuracy give reason to assume that the errors in the DOA data consist of variance and outliers or false estimates due to echoes and other noise sources. The results could be further improved by introducing a



(a) Average errors of non-miss estimates: LS 275 mm and ADC 263 mm



(b) Average miss ratios: LS 62 % and ADC 48 %

Fig. 4. Real data results for speaker localization. Results are given in average estimate error (AEE) [mm] and miss ratio [%] for each of the 24 recording segments. The average values of the Least Squares (LS) and Average Distance Criterion (ADC) methods are weighted averages. The weights are obtained from the number of available ground truth data points.

Monte Carlo-based recursive Bayesian filter to track the location estimate [15], which is, however, outside of the scope of this paper.

5. DISCUSSION

The novelty of this work is the exclusion criterion and its application in speaker localization. The method excludes DOA measurements that contribute the most to the average distance criterion. This improves the miss ratio. However, unnecessary removal of a measurement increases the location estimate variance. The method could be set to accept the maximal DOA measurement subset that fulfills a threshold value of the criterion to counter the unnecessary exclusion of estimates. Additionally, the criterion could be used as a test value of a more sophisticated method such as the Iterative-Reweight-Estimation for fault detection and exclusion [13].

6. SUMMARY

This paper discusses a speaker localization system. The system is based on directional measurements made at spatially separated microphone arrays. Speaker direction measurements are then combined to locate the speaker. However, if the direction estimate is an outlier, the location estimate can deteriorate. This work presents a novel criterion that can be used as a basis of excluding direction estimates. A method that chooses a subset with the smallest average value of the criterion is presented. The method is tested with simulations and real data measurements. The results show that the method reduces the number of large localization errors.

7. REFERENCES

[1] H. E. Bass et al., "Infrasound," *Acoustics Today*, vol. 2, no. 1, pp. 9–19, 2006.

[2] R. Blumrich and J. Altmann, "Medium-range localisation of aircraft via triangulation," *Applied Acoustics*, vol. 61, no. 1, pp. 65–82, 2000.

[3] M. Hawkes and A. Nehorai, "Wideband Source Localization Using a Distributed Acoustic Vector-Sensor Array," *IEEE Transactions on Signal Processing*, vol. 51, no. 6, pp. 1479–1491, 2003.

[4] P. Pertilä, M. Parviainen, T. Korhonen, and A. Visa, "Moving sound source localization in large areas," in *2005 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS 2005)*, 2005, pp. 745–748.

[5] M. Brandstein and D. Ward, Eds., *Microphone Arrays*, Springer-Verlag, 2001.

[6] P. Aarabi, "The Fusion of Distributed Microphone Arrays for Sound Localization," *EURASIP Journal on Applied Signal Processing*, vol. 4, pp. 338–347, 2003.

[7] H.F. Silverman, Yu Ying, J.M. Sachar, and W.R. II Patterson, "Performance of real-time source-location estimators for a large-aperture microphone array," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 593–606, 2005.

[8] Y. Rui, L. He, A. Gupta, and Q. Liu, "Building an intelligent camera management system," in *International Multimedia Conference, Proceedings of the ninth ACM international conference on Multimedia*, 2001, vol. 9, pp. 2–11.

[9] D. Mostefa et al., "Clear evaluation plan v.1.1," http://isl.ira.uka.de/clear06/downloads/chil-clear-v1.1-2006-02-21.pdf, Feb 2006.

[10] J. Yli-Hietanen, K. Kalliojärvi, and J. Astola, "Robust time-delay based angle of arrival estimation," in *Proceedings of the 1996 IEEE Nordic Signal Processing Symposium (NORSIG 96)*, 1996, pp. 219–222.

[11] D.G. Albert, L. Liu, and M.L. Moran, "Time reversal processing for source location in an urban environment.," *Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 616–619, 2005.

[12] T. Pirinen, J. Yli-Hietanen, P. Pertilä, and A. Visa, "Detection and compensation of sensor malfunction in time delay based direction of arrival estimation," in *Proceedings of 2004 IEEE International Symposium on Circuits and Systems (ISCAS'04)*, 2004, vol. 4, pp. 872 – 875.

[13] H. Kuusniemi, *User-Level Reliability and Quality Monitoring in Satellite-Based Personal Navigation*, Ph.D. thesis, Tampere University of Technology, 2005.

[14] T. W. Pirinen, P. Pertilä, and M. Parviainen, "The TUT 2005 Source Localization System," in *Proceedings of the Rich Transcription 2005 Spring Meeting Recognition Evaluation*, Royal College of Physicians, Edinburgh, UK, July 2005, pp. 93–99.

[15] P. Pertilä, T. Korhonen, T. Pirinen, and M. Parviainen, "TUT Acoustic Source Tracking System 2006," in *Proceedings of the CLEAR'06 Evaluation Campaign and Workshop*, Southampton, UK, April 2006, Springer-Verlag, (accepted for publication).

[16] C. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 4, no. 4, pp. 320–327, 1976.