A Joint Audio-Visual Approach to Audio Localization

Jensen, Jesper Rindom; Christensen, Mads Græsbøll

Published in:
I E E E International Conference on Acoustics, Speech and Signal Processing. Proceedings

DOI (link to publication from Publisher):
10.1109/ICASSP.2015.7178010

Publication date:
2015

Document Version
Accepted author manuscript, peer reviewed version

A JOINT AUDIO-VISUAL APPROACH TO AUDIO LOCALIZATION

Jesper Rindom Jensen and Mads Græsbøll Christensen

Audio Analysis Lab, AD:MT, Aalborg University, Denmark, {jrj,mgc}@create.aau.dk

ABSTRACT

Localization of audio sources is an important research problem, e.g., to facilitate noise reduction. In the recent years, the problem has been tackled using distributed microphone arrays (DMA). A common approach is to apply direction-of-arrival (DOA) estimation on each array (denoted as *nodes*), and then map the DOA estimates to a location. In practice, however, the individual nodes contain few microphones, limiting the DOA estimation accuracy and, thereby, also the localization performance. We investigate a new approach, where range estimates are also obtained and utilized from each node, e.g., using time-of-flight cameras. Moreover, we propose an optimal method for weighting such DOA and range information for audio localization. Our experiments on both synthetic and real data show that there is a clear, potential advantage of using the joint audio-visual localization framework.

***Index Terms*—** Localization, DOA, range, optimal weighting, distributed microphone arrays, time-of-flight camera.

1. INTRODUCTION

In the current "age of big data", the number of sensors such as microphones and cameras is rapidly increasing in our electronic devices. This increase in sensor quantities and thereby amounts of available data facilitates new applications that were previously unfeasible, and ease certain signal processing tasks [1]. Several examples of this can be found within the domain of microphone array processing [2–4]. Localization of audio sources (e.g, as input to beamforming, separation, and steering methods) using microphone arrays is a well-established example for the case where we have a single microphone array. However, in the recent years, more focus have been on localization using several of such arrays, forming a so-called (wireless) sensor network [5]. These arrays can then be distributed at different locations, increasing the probability of having an array with good noise conditions, ultimately increasing the potential localization performance.

In theory, such distributed arrays can, of course, be considered as one big array. However, since these arrays typically belong to different devices (e.g., different smart phones [6]), and to reduce the amount of data to transmitted between devices, the arrays are considered as individual sensor nodes. A popular approach to localization using acoustic sensor networks is therefore to let each node in the network estimate a direction-of-arrival (DOA) between the node itself and the audio source. Then, each of these DOA estimates are transmitted to a central node and combined into a single location estimate. This approach were considered in, e.g., [7] were a least squares (LS) estimate of the location is found from the DOA estimates from each sensor node. Clearly, this approach is suboptimal if the noise conditions are different across the sensors nodes. This problem has since been tackled in [4, 8], by considering different methods for detecting and removing outliers among the DOA estimates. Moreover, methods have been proposed that use spectral, biologically inspired features together with DOA estimates to conduct the localization in probabilistic frameworks [9]. One general limitation of the angle-based approach to localization, and, thereby, all of the aforementioned methods, is that they rely only on angle estimates from each node, since range (distance from source to node) estimates are difficult to obtain using closely spaced microphones only [10].

In this paper, we therefore consider a novel approach to audio localization, relying on both angle and range information. As in the traditional approach, the angle information is obtained using the microphone recordings, whereas the range information is obtained using time-of-flight (TOF) cameras. Such cameras, as well as microphones, are found in consumer (e.g., Microsoft Kinect), and industrial grade (e.g., SoftKinetic DS325) products, which facilitate this approach in practice. Compared to when using small microphone arrays only, range information can be extracted with a much higher accuracy using TOF cameras. For example, the SoftKinetic DS311 camera has a range accuracy below 3 cm at 3 meters [11]. Using this approach, we propose a localization method, where the individual location (i.e., angle and range) estimates from each of the nodes are weighted according to their (estimated) noise variances. Finally, we give an example of how to find the optimal weights in practice. We note that the proposed method as well as the traditional, angle-based methods require the node and sensor positions to be known. Methods for estimating these positions have been proposed recently [6, 12, 13], however, if this information is unknown in practice. Moreover, we note that joint audio-visual localization is not a new idea [14–20], but existing approaches typically utilize regular digital cameras to get additional angular source information and not range information as considered herein.

The remainder of the paper is organized in the following way: in Section 2, we introduce the localization problem, and present a typical least squares method for localization using angles only. Then, we propose the optimal weighting method for localization using both angles and ranges in Section 3. The weights for the localization methods can, e.g., be estimated as proposed in Section 3.1. Finally, experimental results and conclusions are presented in Section 4 and 5, respectively.

2. DOA-BASED LOCALIZATION

We consider a setup containing K acoustic nodes in some enclosure. Each acoustic node consists of multiple microphones forming a microphone array. Given an angle, θ_k, and a range, c_k, between node k and the source, the source position can be written as

$$\mathbf{s} = \mathbf{m}_k + c_k \mathbf{b}_k, \tag{1}$$

This work was supported by the Danish Council for Independent Research, grant ID: DFF 1337-00084.

where $\mathbf{m}_k$ is the position of node k, and $\mathbf{b}_k = [\cos\theta_k \ \sin\theta_k]^T$. Note that we consider localization in two dimensions which is sufficient as a proof of our concept, but the results can be generalized to three dimensions. In practice, estimates of the angles (and possibly also the ranges) is obtained, e.g., using a microphone array in each node. The task is then to estimate the acoustic source location, using these estimates. If we only have angle estimates, we replace the DOA in node k by an estimate $\widehat{\theta}_k$, such that $\widehat{\mathbf{b}}_k = [\cos\widehat{\theta}_k \ \sin\widehat{\theta}_k]^T$. Then, we can, e.g., introduce an error function J_{MSE} being the mean squared error (MSE) between the true source position and its estimates $\widehat{\mathbf{s}}_k$ [7]:

$$J_{\text{MSE}} = \frac{1}{K}\sum_{k=1}^{K} \|\mathbf{s} - \widehat{\mathbf{s}}_k\|^2 \tag{2}$$

$$= \frac{1}{K}\sum_{k=1}^{K} (\mathbf{s} - \mathbf{m}_k + c_k\widehat{\mathbf{b}}_k)^T(\mathbf{s} - \mathbf{m}_k + c_k\widehat{\mathbf{b}}_k). \tag{3}$$

Differentiating with respect to the unknown ranges, c_k, setting to zero, and solving for the ranges yields

$$\widehat{c}_k = (\mathbf{s} - \mathbf{m}_k)^T\widehat{\mathbf{b}}_k, \quad \text{for } k = 1, \ldots, K. \tag{4}$$

Doing the same with respect to the unknown source position yields

$$\widehat{\mathbf{s}} = K^{-1}\sum_{k=1}^{K} \mathbf{m}_k + c_k\widehat{\mathbf{b}}_k. \tag{5}$$

Inserting the range estimate in (4) in the above expression, and solving for the unknown source position then finally yields

$$\widehat{\mathbf{s}} = \left(\sum_{k=1}^{K} \frac{\mathbf{I}_2 - \widehat{\mathbf{b}}_k\widehat{\mathbf{b}}_k^T}{K}\right)^{-1} \sum_{k=1}^{K} \frac{\mathbf{I}_2 - \widehat{\mathbf{b}}_k\widehat{\mathbf{b}}_k^T}{K}\mathbf{m}_k, \tag{6}$$

where $\mathbf{I}_N \in \mathbb{R}^{N\times N}$ is the identity matrix.

3. EXPLOITING RANGE INFORMATION

If we also have estimates, $\widehat{c}_k$, of the ranges, c_k, we explicitly have K estimates of the audio source location, each given by

$$\widehat{\mathbf{s}}_k = \mathbf{m}_k + \widehat{c}_k\widehat{\mathbf{b}}_k. \tag{7}$$

This range information could be estimated from the audio data [21], or, possibly with even higher accuracy, from visual data obtained using a TOF camera.

When the latter is the case, the main task in source localization is to weigh these K location estimates appropriately (e.g., according to the EXIP principle [22]). One approach, which is considered here, is to formulate a linear model for the observations for which the minimum variance (MVU) estimator of the unknown location is well-known. The observed data is the individual location estimates in this case. A linear model is readily obtained by stacking the location estimates, i.e.,

$$\underline{\widehat{\mathbf{s}}} = \left[\widehat{\mathbf{s}}_1^T \cdots \widehat{\mathbf{s}}_K^T\right]^T = \mathbf{H}\mathbf{s} + \underline{\mathbf{e}}, \tag{8}$$

where $\mathbf{H} \in \mathbb{R}^{2K\times 2}$ is defined as $\mathbf{H} = [\mathbf{I}_2 \cdots \mathbf{I}_2]^T$, $\underline{\mathbf{e}} = [\mathbf{e}_1^T \ \cdots \ \mathbf{e}_K^T]^T$, and $\mathbf{e}_k = \mathbf{s} - \widehat{\mathbf{s}}_k$ is the error associated with the k'th location estimate. Let us assume that the localization errors are zero mean and follow a Gaussian distribution, i.e., $\underline{\mathbf{e}} \sim \mathcal{N}(\mathbf{0}_{2K}, \mathbf{C})$, where $\mathbf{0}_N \in \mathbb{R}^N$ is the vector of zeros and $\mathbf{C}$ is the error covariance matrix defined as $\mathbf{C} = \mathrm{E}\{\underline{\mathbf{e}}\underline{\mathbf{e}}^T\}$. This is often a good assumption, according to the asymptotic properties of maximum likelihood estimators [23]. Under this assumption, the MVU estimator of the source location is [24]

$$\widehat{\mathbf{s}} = \left(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{C}^{-1}\underline{\widehat{\mathbf{s}}}. \tag{9}$$

In practice, we do not know the error covariance matrix, but it is described in the next section how we can estimate it.

3.1. Finding the Optimal Weights

To find the optimal weights for the different location estimates obtained in Section 3, we need to know the covariance matrices of the localization errors. This section is dedicated to explaining how to estimate these in practice.

First, let us write the location estimate obtained using node k as

$$\widehat{\mathbf{s}}_k = \widehat{c}_k\widehat{\mathbf{b}}_k + \mathbf{m}_k. \tag{10}$$

Furthermore, let us model the range and DOA estimates from each node as

$$\widehat{c}_k = c_k + \delta_k, \quad \text{and} \quad \widehat{\theta}_k = \theta_k + \xi_k, \tag{11}$$

with δ_k and ξ_k being the range and DOA estimation errors, respectively. Using these additive noise models, we can rewrite the k'th source location estimate as

$$\widehat{\mathbf{s}}_k = (c_k + \delta_k)\begin{bmatrix}\cos(\theta_k + \xi_k)\\ \sin(\theta_k + \xi_k)\end{bmatrix} + \mathbf{m}_k. \tag{12}$$

We seek an additive noise model, to be able to model the noise covariance matrix model. The first step in obtaining such model is to rewrite the above expression using trigonometric identities. This leads us to

$$\widehat{\mathbf{s}}_k = (c_k + \delta_k)\mathbf{R}(\theta_k)\begin{bmatrix}\cos\xi_k\\ \sin\xi_k\end{bmatrix} + \mathbf{m}_k. \tag{13}$$

If we assume that the DOA estimation error is small, we can apply small angle approximations, i.e., $\cos\xi_k \approx 1$ and $\sin\xi_k \approx \xi_k$. Using these approximations, we get the following expression for the location estimates:

$$\widehat{\mathbf{s}}_k \approx \mathbf{s} + \delta_k\mathbf{b}_k + (c_k\xi_k + \delta_k\xi_k)\mathbf{b}_k'. \tag{14}$$

That is, the error vector associated with the k'th location estimate can be approximated by

$$\mathbf{e}_k = \mathbf{s} - \widehat{\mathbf{s}}_k \approx \delta_k\mathbf{b}_k + (c_k\xi_k + \delta_k\xi_k)\mathbf{b}_k', \tag{15}$$

The error covariance is needed for the optimal weighting, and it is given by $\mathbf{R}_\mathbf{e} = \mathrm{E}\{\mathbf{e}_k\mathbf{e}_k^T\}$ where $\mathrm{E}\{\cdot\}$ is the mathematical expectation operator. If we assume that the DOA and range errors are uncorrelated, we get that $\mathbf{R}_\mathbf{e} \approx \mathbf{Q}_k\mathbf{R}_\mathbf{v}\mathbf{Q}_k^T$, where $\mathbf{Q}_k = [\mathbf{b}_k \ c_k\mathbf{b}_k']$,

$$\mathbf{R}_\mathbf{v} = \mathrm{E}\left\{\mathbf{v}_k\mathbf{v}_k^T\right\} \approx \begin{bmatrix}\sigma_{\delta_k}^2 & 0\\ 0 & \sigma_{\xi_k}^2\end{bmatrix}, \tag{16}$$

with $\mathbf{v}_k = \begin{bmatrix}\delta_k & \xi_k\end{bmatrix}^T$, and $\sigma_{\delta_k}^2$ and $\sigma_{\xi_k}^2$ are the variances of δ_k and ξ_k, respectively. The error variance matrix for the stacked location estimates needed in (9) can then be formed as

$$\mathbf{C} = \begin{bmatrix}\mathbf{C}_1 & & \mathbf{0}\\ & \ddots & \\ \mathbf{0} & & \mathbf{C}_K\end{bmatrix}. \tag{17}$$

The remaining task is then to estimate the variances of the DOA and range estimates.

3.2. Estimating DOA and range variances

To estimate the DOA variance, we can, e.g., assume a model for the source to be localized. Many parts of audio signals, such as voiced speech and musical instrument recordings, are quasi-periodic, and a well known model for such signals is the harmonic model. Let us then consider a scenario where a node consists of P microphones organized in a uniform linear array structure. This example is considered for illustration purposes, but can be generalized to hold for other array structures. If we preprocess our real audio recordings using the Hilbert transform and assume the microphones are closely spaced compared to the source-to-node distance, we can model our observations using sensor p at time instance n as

$$x_p(n) = \sum_{l=1}^{L} \alpha_l e^{jl\omega_0 n} e^{-jl\omega_0 f_s p \frac{d \sin\theta}{v}} + w(n), \qquad (18)$$

where L is the harmonic model order, α_l is the complex amplitude of the l'th harmonic, ω_0 is the fundamental frequency, f_s is the sampling frequency, d is the spacing between two adjacent microphones, θ is the DOA, v is the speed of sound, and $w(n)$ is observation noise (microphone self-noise, interfering sources, reverberation, etc.). If we have N observations in time using P microphones, and the model in (18) holds with the additional assumption that the observation noise is white Gaussian with equal variance on each microphone, it can be shown that the asymptotic Cramér-Rao bound (CRB) on the DOA estimation variance is [25]

$$\mathrm{CRB}(\theta) = \left[\left(\frac{c}{\omega_0 f_s d \cos\theta} \right)^2 \frac{6}{NP^3} + \left(\frac{\tan\theta}{\omega_0} \right)^2 \frac{6}{N^3 P} \right] \mathrm{PSNR}^{-1}, \qquad (19)$$

where $\mathrm{PSNR} = \sigma_w^{-2} \sum_{l=1}^{L} l^2 |\alpha_l|^2$, with σ_w^2 being the variance of the noise, $w(n)$. If we utilize a DOA estimator that attains the CRB, we can replace $\sigma_{\xi_k}^2$ by $\mathrm{CRB}(\theta_k)$. Some of the parameters needed to calculate the CRB of θ_k are unknown in practice and has to be estimated. These are θ_k itself, ω_0, L, $|\alpha_l|^2$, and σ_w^2. The DOA can be replaced by its estimate, which is input to the localization methods. In this paper, we use the nonlinear least squares (NLS) method in [25] for DOA estimation in each node. This method exploits the harmonic model. The pitch and model order are estimated jointly in each node, but using the NLS estimator in [26]. Note that the pitch and DOA could be estimated jointly using the NLS method in [25], but we estimate the parameters in two stages to reduce the computational complexity. Finally, the amplitudes and noise variances are estimated using maximum likelihood estimators [27] on each microphone signal, and averaged within each node, i.e.,

$$\widehat{\boldsymbol{\alpha}}_{p,k} = (\mathbf{Z}^H \mathbf{Z})^{-1} \mathbf{Z}^H \mathbf{x}_{p,k}, \qquad (20)$$

$$\widehat{\boldsymbol{\alpha}}_k = \frac{1}{P} \sum_{p=1}^{P-1} \widehat{\boldsymbol{\alpha}}_{p,k}, \quad \text{and} \quad |\widehat{\alpha}_{l,k}| = [\widehat{\boldsymbol{\alpha}}_{p,k}]_l, \qquad (21)$$

$$\widehat{\sigma}_{w_k}^2 = \sum_{p=1}^{P} \| \mathbf{x}_{p,k} - \mathbf{Z}(\widehat{\omega}_{0,k}) \widehat{\boldsymbol{\alpha}}_{p,k} \|^2. \qquad (22)$$

In the above equations, $\widehat{\boldsymbol{\alpha}}_{p,k}$ and $\widehat{\boldsymbol{\alpha}}_k$ are vectors of estimates of the harmonic amplitudes in node k and microphone p of node k, respectively, $\widehat{\alpha}_{l,k}$ is an estimate of the l'th harmonic in node k, and $[\cdot]_n$ denotes the n'th element of a vector. Furthermore, $\widehat{\sigma}_{w_k}^2$ is an estimate of the noise variance in node k, $\mathbf{x}_{p,k} \in \mathbb{R}^N$ is a vector of time-consecutive samples from microphone p in node k, $\widehat{\omega}_{0,k}$ is the pitch estimate obtained in node k, and $\mathbf{Z}(\widehat{\omega}_{0,k}) = \left[\mathbf{z}(\widehat{\omega}_{0,k}) \cdots \mathbf{z}(L\widehat{\omega}_{0,k}) \right]$, $\mathbf{z}(l\widehat{\omega}_{0,k}) = \left[1 \; e^{jl\widehat{\omega}_{0,k}} \cdots e^{jl\widehat{\omega}_{0,k}(N-1)} \right]^T$.

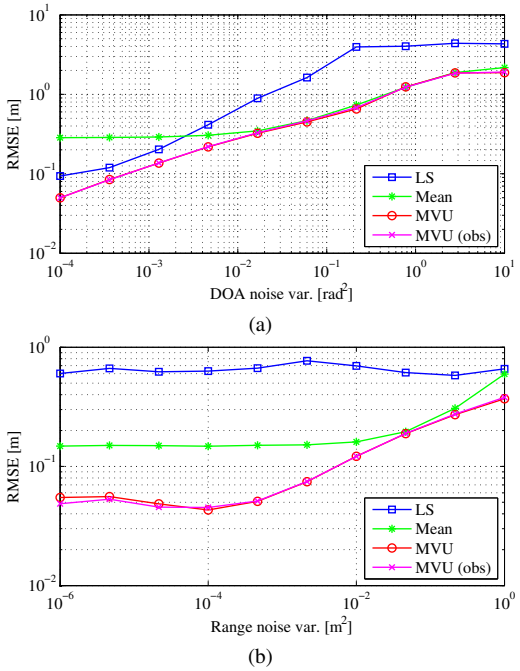

Fig. 1. RMSEs of location estimates obtained from synthetic data using the LS and MVU estimators for varying (a) DOA noise variance, and (b) range noise variance.

The distribution of the range errors depends on the type algorithm used for range estimation in the TOF cameras. In [28], for example, the range distribution was derived for cameras using correlation of amplitude-modulated continuous-wave signals. They showed that the computed range follows an offset normal distribution, and has a variance that can be related to the measured amplitude of the modulation signal. In other words, the range variance will depend on distance and reflectivity. Existing TOF cameras typically provide amplitude estimates [29] that can be used to estimate the range variance. To give an idea of the achievable range estimation accuracy with a TOF camera, the SoftKinetic DS311 yields range errors below 3 cm at 3 m [11], and the more expensive SR4000 camera typically has the same accuracy, but up to a distance of 10 m [29].

4. EXPERIMENTAL RESULTS

In our experiments, we shed light on the potential gain of using both angle and range information for localization using distributed microphone arrays, where the range information is thought to be obtained using TOF cameras. This was achieved by comparing the LS method presented in Section 2, and the minimum variance unbiased (MVU) estimator proposed in Section 3. Moreover, to further evaluate the benefit of our proposed optimal weighting in some of the experiments, we included a mean estimator which just takes the mean of the location estimates in (3.1).

The first experiment, was using synthetic data to verify our model and method. A series of Monte-Carlo simulations were conducted, where the source position was sampled uniformly within a circle with center $[0, 0]$ m, and a radius of 1 m. The positions of three distributed arrays were sampled uniformly in the region

between two circles with center $[0, 0]$ m and radii of 2 m and 3 m, respectively. Then, white Gaussian noise was added to the true angles and ranges between the array positions and the source position. With this setup, we first fixed the range noise variance to $(2 \cdot 0.03 \text{ m})^2$ in each node[1] while the noise variance on the DOAs in radians were varied but identical in each node. For each DOA noise variance, 10000 Monte-Carlo simulations were conducted, yielding the results in Figure 1a in terms of root MSEs (RMSEs). For all DOA noise variances, the proposed MVU estimator outperforms the angle-only based LS estimator. Moreover, we observe that there is no visible difference between using the true (MVU) and observed (MVU (obs)) angles and ranges when forming $\mathbf{Q}_k$. Finally, we note that above a DOA noise variance of 10^{-2} rad^2, there is no difference between the mean and MVU estimators. However, the difference is expected to be larger, if the noise on the DOAs and ranges vary across the nodes. We then conducted a second series of Monte-Carlo simulations, where the DOA noise variance was fixed to 0.01 rad^2, while the range variance was varied. These results are depicted in Figure 1b. In all cases, the proposed MVU estimator clearly outperforms the angle-based only approach. The potential of using range information (e.g., obtained using TOF cameras) is largest for low range noise variances. In summary, the results in Figures 1a and 1b cleary show that localization can be conducted more accurately if range information is available and exploited.

In the second experiment, we applied the proposed method for localization of a real speech signal. The signal was single channel and contained a female speaker uttering the sentence "Why were you away a year, Roy?" two times. To generate a multichannel signal, we used a online available room impulse response (RIR) generator [30]. The simulation scenario was as follows: three nodes were used to localize the speaker in a room with dimensions $(5 \times 4 \times 3)$ m. Each node consisted of a uniform linear array (ULA) with three omnidirectional microphones with 4 cm spacing. During the scenario, the speaker was moving from the left side of the room to the right. The node placements and source movement is depicted in Figure 2a. First, we then considered the case with no reverberation. To estimate the DOAs needed in the localization, we first extracted time-consecutive blocks of 200 samples from each microphone in each node. Then, on the blocks from the first microphone in each node, we applied the NLS method in [26] to estimate the pitch and harmonic model order of the speech signal jointly. The pitch and model order estimates were then given as input to the method in [25] for DOA estimation in each node. The range estimate in each node was generated synthetically by assuming a range standard deviation of 3 cm, and adding white Gaussian noise with this standard deviation to the true ranges[2]. Using these DOA and range observation, we then conducted the localization using the LS and MVU estimators, in Section 2 and 3, respectively, where the optimal weights were found as described in Section 3.1. The localization errors over time for this experiment are provided in Figure 2b. From the results, we see that there is a clear, potential benefit of using the proposed method as opposed to using an angle-only estimator. Using the same setup, except that reverberation with a T_{60} of 0.3 s was added, we conducted another simulation. With reverberation, the errors in Figure 2c were obtained. As expected, the errors are generally higher with reverberation, but again the proposed MVU estimator clearly has a potentially better localization performance compared to when using the angle-only based LS estimator.

[1]This corresponds to setting the standard deviation to two times the typical error of the SoftKinetic DS311 up to 3 m.

[2]This roughly corresponds to the accuracy of the SoftKinetic DS325 camera.

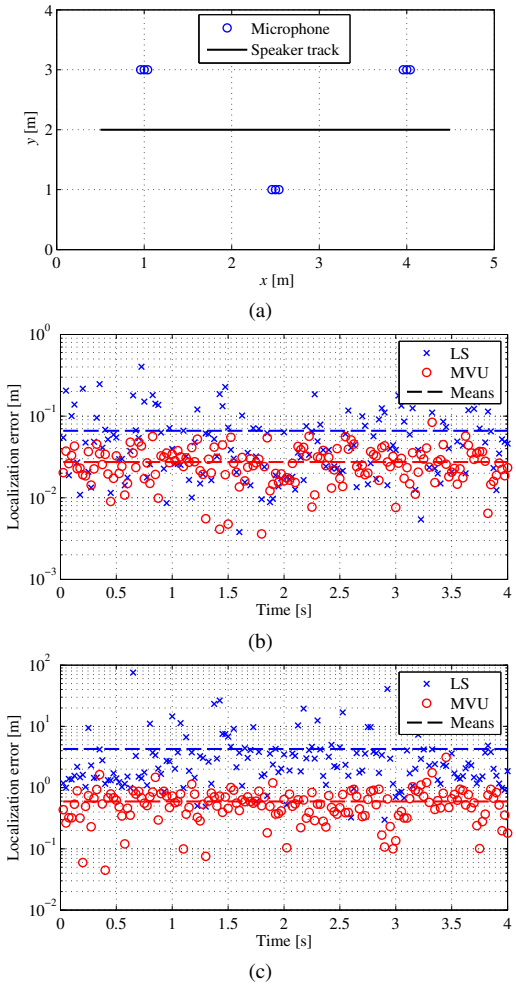

Fig. 2. Localization results [(a) without and (b) with reverberation] obtained with the LS and MVU estimators in the scenario depicted in (a).

5. CONCLUSION

In this paper, we have considered the problem of audio localization using distributed microphone arrays. Traditionally, this problem has been tackled by estimating the DOA of the audio source in each node of a sensor network, where each node consist of a microphone array. The number of microphones in each array, however, is typically low in practice, limiting the DOA estimation accuracy and therefore also the accuracy of the location estimate. To improve on this, we therefore propose a new approach, where the range of the audio source is also estimated in each node. The range can, for example, be estimated using time-of-flight cameras (e.g., one in each node) at a very high accuracy. Moreover, we proposed a method for optimally weighting the DOA and range estimates from the different sensor nodes, to obtain a location estimate of an audio source, and showed how the optimally weights can be found. In our experiments, we have showed that there is a significant potential of using additional range information in localization of audio by using both microphone and cameras to capture the source. This was shown on both synthetic and real speech data. In future work, the method proposed herein will be applied on real, measured data.

6. REFERENCES

[1] D. Lahat, T. Adali, and C. Jutten, "Challenges in multimodal data fusion," in *Proc. European Signal Processing Conf.*, Sep. 2014, p. 5.

[2] I. Himawan, I. McCowan, and S. Sridharan, "Clustered blind beamforming from ad-hoc microphone arrays," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 4, pp. 661–676, May 2011.

[3] S. Markovich-Golan, A. Bertrand, M. Moonen, and S. Gannot, "Optimal distributed minimum-variance beamforming approaches for speech enhancement in wireless acoustic sensor networks," *Signal Process.*, 2014, accepted.

[4] A. Griffin and A. Mouchtaris, "Localizing multiple audio sources from DOA estimates in a wireless acoustic sensor network," in *Proc. IEEE Workshop Appl. of Signal Process. to Aud. and Acoust.*, Oct. 2013, pp. 1–4.

[5] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *Proc. IEEE Symp. Commun. Veh. Technol.*, Nov. 2011, pp. 1–6.

[6] P. Pertila, M. S. Hamalainen, and M. Mieskolainen, "Passive temporal offset estimation of multichannel recordings of an ad-hoc microphone array," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 11, pp. 2393–2402, Nov. 2013.

[7] H. Wang, C. E. Chen, A. Ali, S. Asgari, R. E. Hudson, K. Yao, D. Estrin, and C. Taylor, "Acoustic sensor networks for woodpecker localization," in *Proc. of SPIE Conf. Advanced Signal Process. Algorithms, Architectures, and Implementations*, Aug. 2005, vol. 5910.

[8] A. Ledeczi, G. Kiss, B. Feher, P. Völgyesi, and G. Balogh, "Acoustic source localization fusing sparse direction of arrival estiamtes," in *Proc. of Int. Workshop on Intelligent Solutions in Embedded Systems*, Jun. 2006.

[9] A. Plinge and G. A. Fink, "Multi-speaker tracking using multiple distributed microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2014, pp. 614–618.

[10] M. N. El Korso, R. Boyer, A. Renaux, and S. Marcos, "Non-matrix closed-form expressions of the Cramér-Rao bounds for near-field localization parameters," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2009, pp. 3277–3280.

[11] SoftKinetic®, *DS311 Datasheet – Far & Close Interaction Time-of-Flight 3D Depth-Sensing Camera*.

[12] N. D. Gaubitch, W. B. Kleijn, and R. Heusdens, "Auto-localization in ad-hoc microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2013, pp. 106–110.

[13] A. Plinge and G. A. Fink, "Geometry calibration of distributed microphone array exploiting audio-visual correspondences," in *Proc. European Signal Processing Conf.*, Sep. 2014, p. 5.

[14] F. Talantzis, A. Pnevmatikakis, and A. G. Constantinides, "Audio-visual active speaker tracking in cluttered indoors environments," *IEEE Trans. Syst., Man, Cybern. B*, vol. 38, no. 3, pp. 799–807, Jun. 2008.

[15] N. Strobel, S. Spors, and R. Rabenstein, "Joint audio-video object localization and tracking," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 22–31, Jan. 2001.

[16] Z. Li, T. Herfet, M. Grochulla, and T. Thormahlen, "Multiple active speaker localization based on audio-visual fusion in two stages," in *Proc. IEEE Conf. Multisensor Fusion and Integration for Intell. Syst.*, Sep. 2012, pp. 1–7.

[17] V. Khalidov, F. Forbes, M. Hansard, E. Arnaud, and R. Horaud, "Detection and localization of 3d audio-visual objects using unsupervised clustering," in *Proc. Int. Conf. Multimodal Interfaces*, Oct. 2008, pp. 217–224.

[18] H. Zhou, M. Taj, and A. Cavallaro, "Target detection and tracking with heterogeneous sensors," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 4, pp. 503–513, Aug. 2008.

[19] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan, "Multimodal multispeaker probabilistic tracking in meetings," in *Proc. Int. Conf. Multimodal Interfaces*, Oct. 2005, pp. 183–190.

[20] P. Pérez, J. Vermaak, and A. Blake, "Data fusion for visual tracking with particles," *Proc. IEEE*, vol. 92, no. 3, pp. 495–513, Mar. 2004.

[21] J. R. Jensen and M. G. Christensen, "Near-field localization of audio: a maximum likelihood approach," in *Proc. European Signal Processing Conf.*, Sep. 2014, p. 5.

[22] P. Stoica and T. Söderström, "On reparametrization of loss functions used in estimation and the invariance principle," *Signal Process.*, vol. 17, no. 4, pp. 383–387, Aug. 1989.

[23] P. Stoica, R. L. Moses, B. Friedlander, and T. Söderström, "Maximum likelihood estimation of the parameters of multiple sinusoids from noisy measurements," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 3, pp. 378–392, Mar. 1989.

[24] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice Hall, Inc., 1993.

[25] J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Nonlinear least squares methods for joint DOA and pitch estimation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 5, pp. 923–933, May 2013.

[26] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech and Audio Processing*, vol. 5, no. 1, pp. 1–160, 2009.

[27] M. G. Christensen, "Multi-channel maximum likelihood pitch estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2012, pp. 409–412.

[28] M. Frank, M. Plaue, H. Rapp, U. Koethe, B. Jähne, and F. A. Hamprecht, "Theoretical and experimental error analysis of continuous-wave time-of-flight range cameras," *Opt. Eng.*, vol. 48, no. 1, pp. 16, Jan. 2009.

[29] MESA Imaging, *SR4000 Data Sheet*.

[30] E. A. P. Habets, "Room impulse response generator," Tech. Rep., Technische Universiteit Eindhoven, 2010, Ver. 2.0.20100920.