Robust TDOA-Based Sound Source Localization

Felipe Barboza da Silva and Wallace Alves Martins.

***Abstract*— This paper proposes a new technique for solving sound source localization problems using time-difference of arrival (TDOA). The key aspect of the proposal relies on both the use of a median-based cost function, which allows for outlier filtering of misestimated TDOAs, and a TDOA discarding process. Simulation results indicate that the proposed method achieves centimeter-level localization accuracy even when dealing with high reverberation and low signal-to-noise ratio (SNR). Indeed, for an SNR of 10 dB and a 5.2 m × 7.5 m × 2.6 m room with reverberation time of 800 ms, our method achieves median errors under 10 cm and mean errors about 20 cm, outperforming closed-form least-squares solutions and classic maximum likelihood-based TDOA techniques.**

***Keywords*— Sound source localization, time-difference of arrival, generalized cross-correlation, reverberation.**

I. Introduction

Sound source localization (SSL) using microphone array finds many applications in industry and for military purposes. Gas leakage in a pipe may have its position estimated using this localization technique, as well as gun shots inside a military barrack [1], [2]. This technology may also be applied to speech enhancement in, for instance, audio conferences and entertainment — Kinect® is a case in point [3], [4].

For most SSL algorithms, the main underlying challenge is to estimate correctly the time-differences of arrivals (TDOAs) of the acquired signals associated with each pair of microphones. Futhermore, TDOA estimation is substantially affected by reverberation effects and acoustic noise. Thus, in this paper our focus is on proposing a more robust method to TDOA-based SSL purposes and compare it to other standard TDOA-based methods of the literature [5], [6].

II. Notation

The number of microphones is denoted as $M \in \mathbb{N}$, and the index of a given microphone pair is denoted by $p \in \{1, \cdots, P\}$, where $P = \frac{M(M-1)}{2}$ is the number of distinct pairs of microphones. The 3-D positions of the microphones of the pth pair are $\mathbf{m}_p^{(1)}, \mathbf{m}_p^{(2)} \in \mathbb{R}^3$. Sometimes, the mth microphone position, with $m \in \{1, \cdots, M\}$, will be denoted as $\mathbf{m}_m \in \mathbb{R}^3$ as well. The sound source 3-D position is denoted by $\mathbf{s} \in \mathbb{R}^3$. The sample mean operator is denoted as $\mathbb{E}_P$ and the sample median operator is denoted as $\mathbb{M}_P$.

III. TDOA Definition

As mentioned before, the TDOA quantifies the difference between the times the wavesound takes to travel from the source position to each microphone of the pair. Thus, once the wavesound propagation speed $c \in \mathbb{R}_+$ is known, one can determine the pth TDOA τ_p as:

$$\tau_p(\mathbf{s}) \triangleq \frac{\|\mathbf{m}_p^{(1)} - \mathbf{s}\| - \|\mathbf{m}_p^{(2)} - \mathbf{s}\|}{c}. \tag{1}$$

IV. TDOA Estimation using GCC-PHAT

There are many methods to estimate TDOAs, most of them based on cross-correlations of acquired signals [7]. In our case, the generalized cross-correlation with phase transform (GCC-PHAT) technique [7] was applied to find the best match between the projection of a captured signal from a microphone into delayed versions of the captured signal of the other microphone of the pair. Thus, the pth TDOA is estimated by finding the time-lag $\hat{\tau}_p \in \mathbb{R}$ which maximizes the GCC-PHAT function between the acquired signals of the pth microphone pair. Further details about TDOA evaluation and GCC-PHAT can be found in, for example, [7].

V. TDOA-based Sound Source Localization

A. Unconstrained LS-TDOA

The first classical method we employed to estimate the source position is the unconstrained least-squares (LS) TDOA technique [5]. This method estimates the sound source position through the following expression:

$$\hat{\mathbf{s}} = (\mathbf{\Phi}^T \mathbf{\Phi})^{-1} \mathbf{\Phi}^T \mathbf{b}, \tag{2}$$

where $\{\cdot\}^T$ denotes the transpose operation. Assuming that the Mth microphone plays the role of a reference microphone, then the mth row of $\mathbf{\Phi}$, with $m \in \{1, \cdots, M-1\}$, can be written as $[c\hat{\tau}_{m,M} \quad (\mathbf{m}_m - \mathbf{m}_M)^T]$, where $\hat{\tau}_{m,M}$ denotes the estimated TDOA between the mth and the Mth microphones, whereas the mth element of $\mathbf{b}$ is $\left(\|\mathbf{m}_m - \mathbf{m}_M\|^2 - c^2 \hat{\tau}_{m,M}^2\right)/2$. More details about this method can be found in [5].

B. ML-TDOA

The authors in [6] describe a way to find a point in space that best fits the estimated TDOAs. Mathematically, if one defines the error $e_p(\mathbf{g}) = \tau_p(\mathbf{g}) - \hat{\tau}_p$ for each 3-D point $\mathbf{g}$ of a predefined grid of points $\mathcal{G} \subset \mathbb{R}^3$ that comprise the search space, then one has the following maximum-likelihood (ML) TDOA estimate:

$$\hat{\mathbf{s}} = \underset{\mathbf{g} \in \mathcal{G}}{\operatorname{argmin}} \left\{ \mathbb{E}_P \left[e_p^2(\mathbf{g}) \right] \right\}, \tag{3}$$

where $\mathbb{E}_P[e_p^2] = \frac{e_1^2 + e_2^2 + \cdots + e_P^2}{P}$.

VI. Main Contribution: LMedS-TDOA

The LS-TDOA technique does not account for possible TDOA misestimation, while the ML-TDOA tries to mitigate possible TDOA estimation errors through an averaging process of the resulting squared errors e_p^2. Nonetheless, the existence of a few large TDOA estimation errors may be sufficient to impair the localization performance of mean-based techniques, such as the ML-TDOA scheme.

We propose employing the median of the squared TDOA errors, e_p^2, as cost-function instead of their mean in order to increase robustness to possible TDOA misestimation. The goal here is to give less focus to large errors, smoothing their effects on the source localization. The proposed method is therefore denominated least-median-of-squares TDOA (LMedS-TDOA).

F. B. da Silva and W. A. Martins are with Federal University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil (emails: felipe.silva@smt.ufrj.br, wallace.martins@smt.ufrj.br). The authors are grateful for the financial support provided by CNPq, CAPES, and FAPERJ, Brazilian research councils.

TABLE I
LOCALIZATION ERROR (CM) FOR SNR OF 30 dB.

		Median			Mean			Standard-Deviation		
SNR (dB)	RT60 (ms)	LS	ML	LMedS	LS	ML	LMedS	LS	ML	LMedS
30	200	5.07	5.03	5.03	5.15	6.94	7.13	4.37	4.46	4.77
	400	5.07	5.03	5.03	9.32	7.86	7.38	39.93	10.45	5.14
	600	10.47	9.40	5.03	106.9	53.04	7.38	189.3	116.3	5.14
	800	166.5	38.35	5.03	—	117.6	7.13	—	165.5	4.77

TABLE II
LOCALIZATION ERROR (CM) FOR SNR OF 10 dB.

		Median			Mean			Standard-Deviation		
SNR (dB)	RT60 (ms)	LS	ML	LMedS	LS	ML	LMedS	LS	ML	LMedS
10	200	24.46	9.40	5.03	—	32.63	9.13	—	78.21	9.51
	400	122.0	77.56	9.26	—	135.1	11.65	—	—	16.23
	600	—	—	9.40	—	—	15.79	—	162.1	21.35
	800	—	—	9.40	—	—	23.24	—	153.7	37.33

The method is divided into two steps, denoted as step A and step B. The first one evaluates $\hat{\mathbf{s}}^{(\mathrm{A})} \in \mathbb{R}^3$ as follows:

$$\hat{\mathbf{s}}^{(\mathrm{A})} = \underset{\mathbf{g}\in\mathcal{G}}{\operatorname{argmin}} \left\{ \mathbb{M}_P \left[e_p^2(\mathbf{g}) \right] \right\}, \tag{4}$$

where $\mathbb{M}_P[e_p^2]$ computes the median of the samples $e_1^2, e_2^2, \cdots, e_P^2$. The second step discards some presumably misestimated TDOAs according to a given threshold $\gamma \in \mathbb{R}_+$: if $|e_p(\hat{\mathbf{s}}^{(\mathrm{A})})| \geq \gamma$, then $\hat{\tau}_p$ is discarded. Then, once again an exaustive search takes place as in (4), but now using a reduced set containing $\mathbb{N} \ni \overline{P} \leq P$ error samples corresponding to the remaining TDOA estimates, i.e.:

$$\hat{\mathbf{s}}^{(\mathrm{B})} = \underset{\mathbf{g}\in\mathcal{G}}{\operatorname{argmin}} \left\{ \mathbb{M}_{\overline{P}} \left[e_p^2(\mathbf{g}) \right] \right\}. \tag{5}$$

VII. PERFORMANCE EVALUATION

In this section, we will compare the performance of the classic methods LS-TDOA and ML-TDOA against the proposed LMedS-TDOA method.

A. Simulation Procedure

The simulations considered that all sources and microphones were inside a room $\mathcal{R} \subset \mathbb{R}^3$ with dimensions $5.2\,\mathrm{m} \times 7.5\,\mathrm{m} \times 2.6\,\mathrm{m}$. Five different source positions were used and the array was composed by 16 microphones. Both source and microphone positions were based on the simulation setup described in [8]. All microphone signals were corrupted by independent additive white Gaussian noise (AWGN). During the simulations, a voice-activity detector (VAD) was employed before playing back the sound source signal, which was a female speech with 4.5-s duration divided into 60 blocks of 100 ms with 25 ms of overlapping at a sampling rate of 48 kHz. The grid $\mathcal{G} \subset \mathcal{R}$ is comprised of a regular grid of points with smallest distance between adjacent points of 10 cm. It is worth mentioning that any grid point 10 cm next to the walls was discarded as well as any point 30 cm next to the floor and 60 cm next to the ceiling. Considering $c = 340$ m/s, we assumed $\gamma = 1$ ms since it yields acceptable localization errors.

TABLES I and II show the simulation results for signal-to-noise ratios (SNRs) of 30 dB and 10 dB, respectively, and for 4 different levels of reverberation time (RT60) [9]. The columns of the tables show the statistics of the localization errors considering all signal blocks and source positions ($60 \times 5 = 300$ localization errors in total). It was assumed that when any error above 2 meters occurred, the respective method failed, which is indicated in the tables by a dash mark.

B. Simulation Results

As can be gathered from TABLE I, the proposed method yields the smallest localization errors for moderate to high reverberant environments. As for the results of TABLE II, even in a more noisy environment, the proposed method achieves errors much smaller than the LS and ML techniques, thus indicating its robustness to both reverberation and noise effects. For RT60 = 800 ms, the LMedS method discarded 10 and 40 TDOAs for SNR of 30 and 10 dB, respectively.

VIII. CONCLUDING REMARKS

This paper proposed a new TDOA-based sound source localization technique, inspired by the classical ML-TDOA scheme. The main contribution was to use a median-based cost-function, coupled to a discarding TDOA criteria. The simulation results indicated that the algorithm is quite robust to both reverberation and noise effects. As future work, the proposed method should be evaluated using real-world signals recorded under reverberant and noisy conditions. Futhermore, more efficient exaustive search algorithms will be conceived in order to increase the speed of the method, thus allowing the use of finer grids.

REFERENCES

[1] Lv Xiaoling, Z. Minglu, Y. Guangming, C. Qiang and Z. Haixian, “Robot sound source search strategy based on multi-blackboard model,” *in Proc. 2010 IEEE International Conference on Robotics and Biomemetics (ROBIO)*, pp. 633–638, Dec. 2010.

[2] A. M. C. R. Borzino, J. A. Apolinario and M. L. R. de Campos, “Estimating direction of arrival of long range gunshot signals,” *in Proc. 2014 International Telecommunications Symposium (ITS)*, pp. 1–5, Aug. 2014.

[3] H. K. Maganti, D. Gatica-Perez and I. McCowan, “Speech enhancement and recognition in meetings with an audio-visual sensor array,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 2257–2269, Nov. 2007.

[4] M. R. P. Thomas, J. Ahrens and I. Tashev, “Optimal 3D beamforming using measured microphone directivity patterns,” *in Proc. 2012 International WorkMaganti, shop on Acoustic Signal Enhancement (IWAENC)*, pp. 1–4, Sep. 2012.

[5] P. Stoica and J. Li, “Lecture notes - source localization from range-difference measurements,” *IEEE Signal Processing Magazine*, vol. 23, pp. 63–66, Nov. 2006.

[6] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications.* Digital Signal Processing - Springer-Verlag, Springer, 2001.

[7] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Springer Topics in Signal Processing, Springer Berlin Heidelberg, 2008.

[8] L. O. Nunes, W. A. Martins, M. V. S. Lima, L. W. P. Biscainho, M. V. M. Costa, F. M. Goncalves, A. Said and B. Lee, “A steered-response power algorithm employing hierarchical search for acoustic source localization using microphone arrays,” *IEEE Transactions on Signal Processing*, vol. 62, pp. 5171–5183, Oct. 2014.

[9] B. Dumortier and E. Vincent, “Blind RT60 estimation robust across room sizes and source distances,” *in Proc. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5187–5191, May 2014.