NEW DIRECT APPROACHES TO ROBUST SOUND SOURCE LOCALIZATION

Yong Rui and Dinei Florencio
Microsoft Research
One Microsoft Way, Redmond, WA 98052

ABSTRACT

When more than two microphones are used, the traditional time-delay-of-arrival (TDOA) based sound source localization (SSL) approach involves two steps. The first step computes TDOA for each microphone pair, and the second step combines these estimates. This two-step process discards relevant information in the first step, thus degrading the SSL accuracy and robustness. Although less used, one-step processes do exist. In this paper, we review these processes, create a unified framework, and introduce two new one-step algorithms. We compare our proposed approaches against existing 1 and 2-step approaches and demonstrate significantly better SSL performance.

1. INTRODUCTION

Using microphone arrays to do sound source localization (SSL) has been an active research topic since the early 1990's [2]. It has many important applications including video conferencing [1],[4],[7], surveillance, and speech recognition. There exist various approaches to SSL in the literature. So far, the most studied and widely used technique is the time delay of arrival (TDOA) based approach [2],[7],[9].

When using more than two microphones, the conventional TDOA SSL is a two-step process (referred to as 2-TDOA in this paper). In the first step, TDOA (or equivalently the bearing angle) is estimated for each pair of microphones. This step is performed in the cross correlation domain, and a weighting function is generally applied to enhance the quality of the estimate. In the second step, multiple TDOAs are intersected to obtain the final source location [2]. The 2-TDOA has two main advantages: it is a well studied area (e.g., good weighting functions have been investigated for a number of scenarios), and the computation of the second step is cheap [2]. The disadvantage is that it makes a premature decision on an intermediate TDOA in the first step, thus throwing away useful information. A better approach would use the principle of *least commitment* [1]: preserve and propagate all the intermediate information to the end and make an informed decision at the very last step. Because this approach solves the SSL problem in a single step, we call it *direct* approach in this paper. We investigate two direct approaches: one-step TDOA (referred to as 1-TDOA) SSL and steered beam (SB) SSL. Conceptually, these two approaches are similar – finding the point in the space which yields maximum energy. But they differ in theoretical merits and algorithm complexity.

During the past few years, with the ever increasing computing power, researchers started to focus more on the robustness of SSL while concerning less with computation cost [1][5][6]. However, they have not taken full advantage of the well studied weighting functions. New weighting functions, e.g.,[8], can simultaneously handle reverberation and ambient noise, achieving higher accuracy and robustness**.**

The rest of the paper is organized as follows: in Section 2 we analyze the theoretical merits and compare the computation complexity of the 1-TDOA SSL and SB SSL. In Section 3, we propose two new techniques, one based on 1-TDOA and the other based on SB. In Section 4, we conduct extensive experiments and compare the proposed approaches against existing ones. The results demonstrate superior performance of the proposed techniques. We give concluding remarks in Section 5.

2. SB SSL AND 1-TDOA SSL

The commonality between these two approaches is that they both localize the sound source through hypothesis testing -- pick as the sound source location the point in the space which produces the highest energy. Let M be the number of microphones in an array. The signal received at microphone m, where $m = 1, \ldots, M$, at time n is:

$$x_m(n) = h_m(n) * s(n) + n_m(n) \quad (1)$$

where $n_m(n)$ is additive noise, and $h_m(n)$ represents the room impulse response. Even if we disregard reverberation, the signal will arrive at each microphone at different times. SB SSL selects the location in space which maximizes the sum of the delayed received signals. To reduce computation cost, usually only a finite number of locations L are investigated. Let $P(l)$ and $E(l)$, $l = 1, \ldots, L$, be the location and energy of point l. Then the selected sound source location $P^*(l)$ is:

$$p^*(l) = \arg\max_l \{E(l)\} \quad (2)$$

$$E(l) = \left| \sum_{m=1}^{M} x_m(n - \tau_m) \right|^2 \quad (3)$$

where τ_m is the time that takes sound to travel from the source to microphone m. Equation (3) can also be expressed in the frequency domain:

$$E(l) = \left| \sum_{m=1}^{M} X_m(f) \exp(-j2\pi f \tau_m) \right|^2 \quad (4)$$

where $X_m(f)$ is the Fourier transform of $x_m(n)$. If we explicitly expand the terms in Equation (4), we have:

$$E(l) = \sum_{m=1}^{M} |X_m(f)|^2 + \sum_{r=1}^{M} \sum_{s \neq r}^{M} \left| X_r(f) X_s^*(f) e^{-j2\pi f(\tau_r - \tau_s)} \right|^2 \quad (5)$$

We note that the first term in Equation (5) is constant across all points in space, thus it can be eliminated for SSL purpose. Equation (5) then reduces to summations of the cross correlations of all the microphone pairs in the array. The cross correlations in Equation (5) are exactly the same as the cross correlations in the traditional 2-TDOA approaches. But instead of introducing an intermediate variable TDOA, Equation (5) retains all the useful information contained in the cross correlations. It solves the SSL problem *directly* by selecting the highest *E(l)*. We call this approach 1-TDOA.

Note further that Equations (4) and (5) are the same mathematically. 1-TDOA and SB, therefore, have the same origin. But they differ in theoretical merits and computation complexity, which we will investigate next.

2.1. Theoretical merits

Computing *E(l)* in frequency domain gives us flexibility to add weighting functions. Equations (4) and (5) then become:

$$E(l) = \left| \sum_{m=1}^{M} V_m(f) X_m(f) \exp(-j2\pi f \tau_m) \right|^2 \quad (6)$$

$$E'(l) = \sum_{r}^{M} \sum_{s \neq r}^{M} |W_{rs}(f) X_r(f) X_s^*(f) \exp(-j2\pi f(\tau_r - \tau_s))|^2 \quad (7)$$

where $V_m(f)$ and $W_{rs}(f)$ are the filters (weighting functions) for individual channels *m* and a pair of channels *r* and *s*.

Finding the optimal $V_m(f)$ for SSL is a challenging task. As pointed out in [5], it depends on the nature of source and noise, and on the geometry of the microphones. While heuristics can be used to obtain $V_m(f)$ (as will be discussed in Section 3), they may not be optimal. On the other hand, the weighting function $W_{rs}(f)$ is nothing but the same weighting function used in the traditional 2-TDOA SSL, which is a well studied area. In Section 3, we will introduce a new weighting function we developed recently which simultaneously handles ambient noise and room reverberation [8].

2.2. Computational complexity

The points in the 3D space that have the same time delay for a given pair of microphones form a hyperboloid. Different time delay values give origin to a family of hyperboloids centered at the midpoint of microphone pair. Therefore, any point in 3D space has its mapping to the 1D cross correlation curve of this pair of microphone. This observation allows us to efficiently compute *E'(l)* in (7). Given the cross correlation curves for all the microphone pairs, computing *E'(l)* is just a table-look-up and summation process.

We now compare the main steps and computation complexity between 1-TDOA SSL and SB SSL. For 1-TDOA SSL we have:

1. Compute the *N*-point FFT $X_m(f)$ for the *M* microphones: *O(MNlogN)*.
2. Let $Q = C_M^2$ be the number of the microphone pairs formed from the *M* microphones. For the *Q* pairs, compute $W_{rs}(f)X_r(f)X_s(f)^*$ according to Equation (7): *O(QN)*.
3. For the *Q* pairs, compute the inverse FFT to obtain the cross correlation curve: *O(QNlogN)*.
4. For the *L* points in the space, compute their energies by table look-up from the *Q* interpolated correlation curves: *O(LQ)*.

Therefore, the total computation cost for 1-TDOA SSL is *O(MNlogN + Q(N+NlogN+L))*.

The main algorithm steps for SB SSL are:

1. Compute *N*-point FFT $X_m(f)$ for the *M* microphones: *O(MNlogN)*.
2. For the *L* locations and *M* microphones, phase shift $X_m(f)$ by $2\pi f \tau_m$ and weight it by $V_m(f)$ according to Equation (6): *O(MLN)*.
3. For the L locations, compute the energy: *O(LN)*.

The total computation cost is therefore *O(MNlogN + L(MN+N))*. The dominant term in 1-TDOA SSL is *QNlogN* and the dominant term in BS-SSL is LMN. If *QlogN* is bigger than LM, then SB SSL is cheaper to compute. Furthermore, it is possible to do SB SSL in a hierarchical way, which can result in further savings. On the other hand, weighting functions for 1-TDOA are well studied, and may result in better performance.

2.3. Summarize it up

Based on the above analysis, we can provide a few general recommendations for selecting a SSL algorithm family. First, if using only 2 microphones, use TDOA-based SSL. Because of its well studied weighting functions, it will provide better results with no added complexity. Second, for multiple (>2) microphones, use direct algorithms for better accuracy. Only consider 2-TDOA if computational resources are extremely scarce, and source location is 2-D or 3-D. Third, if accuracy is important, prefer 1-TDOA over SB, because of its better studied weighting functions. Finally, if *QNlogN < LM*, use 1-TDOA SSL for lower computational cost and better performance.

3. PROPOSED APPROACHES

In the field of SSL, there are two branches of research being done in relative isolation. On one hand, various weighting functions have been proposed in 2-TDOA. But 2-TDOA is inherently less robust. On the other hand, 1-TDOA SSL and SB SSL are more robust but their weighting function choices are not well explored yet. In this section, we propose two new approaches based on our recent work on a new weighting function, which simultaneously handles ambient noise and reverberation [8].

3.1. A new 1-TDOA SSL approach

So far, existing 1-TDOA SSL approaches use either PHAT or ML as the weighting function, [1][5]:

$$W_{PHAT}(f) = \frac{1}{|X_1(f)||X_2(f)|} \quad (8)$$

$$W_{ML}(f) = \frac{|X_1(f)||X_2(f)|}{|N_2(f)|^2|X_1(f)|^2 + |N_1(f)|^2|X_2(f)|^2} \quad (9)$$

PHAT works well only when the ambient noise is low. Similarly, ML works well only when the reverberation is small. In [8], we developed the maximum likelihood estimator when both ambient noise and reverberation are present. The corresponding weighting function is:

$$W_{MLR}(f) = \frac{|X_1(f)||X_2(f)|}{2q|X_1(f)|^2|X_2(f)|^2 + (1-q)|N_2(f)|^2|X_1(f)|^2 + |N_1(f)|^2|X_2(f)|^2} \quad (10)$$

where q is a constant in [0,1]. The very successful PictureTel [9] weighting function is a special case of [8]. Substituting Equation (10) into (7), we obtain a new 1-TDOA approach.

3.2. A new SB SSL approach

There exists a rich literature on weighting functions for beam forming for speech enhancement [3]. But so far little research has been done in developing good weighting functions $V_m(f)$ for SB SSL. Weighting functions for enhancement and SSL have related but different objectives. For example, SSL does not care the quality of the captured audio, as long as the location estimation is accurate. Most of the existing SB SSL use no weighting functions, e.g., [6][10]. While it is challenging to find the optimal weights, we may obtain reasonably good solutions by using observations obtained from the new 1-TDOA SSL described above. If we make the following approximations

$$\begin{aligned} |X_1(f)X_2(f)| &\approx |X_1(f)|^2 \approx |X_2(f)|^2 \\ |N(f)|^2 &\approx |N_1(f)|^2 \approx |N_2(f)|^2 \end{aligned} \quad (11)$$

we can obtain an approximated weighting function to (10):

$$W_{AMLR}(f) = \frac{1}{q|X_1(f)||X_2(f)| + (1-q)|N_1(f)||N_2(f)|} \quad (12)$$

The benefit of this approximated weighting function is that it can be decomposed into two individual weighting functions for each microphone. A good choice for $V_m(f)$ is therefore:

$$V_m(f) = \frac{1}{q|X_m(f)| + (1-q)|N_m(f)|} \quad (13)$$

4. EXPERIMENTAL RESULTS

We have implemented a working SSL system based on our proposed approaches. It is developed in C++ on Windows DirectShow platform. No code optimization is attempted and the system runs comfortably in real time on a regular P4. This system is a component in our Distributed Meeting effort [4], whose goal is to facilitate effective local and tele-meetings.

In this section, we will focus on three sets of comparisons through extensive experiments: 1) the proposed new 1-TDOA approach against existing 1-TDOA ones; 2) the proposed new SB approach against existing SB ones; and 3) compare the 2-TDOA, 1-TDOA and SB SSL approaches in general.

4.1. Testing data description

We have tested our system both by putting it into the actual meeting room and by using synthesized data. Because it is easier to obtain the ground truth (e.g., source location, SNR and reverberation time) for the synthesized data, we report our experiments on this set of data. We take great care to generate realistic testing data. We use the imaging method to simulate room reverberation. To simulate ambient noise, we capture actual office fan noise and computer hard drive noise using a close-up microphone. The same room reverberation model is then used to add reverberation to these noise signals, which are then added to the reverberated desired signal. We make our testing data as difficult as, if not more difficult than, the real data obtained in our actual meeting room.

The testing data setup corresponds to a 6m×7m×2.5m room, with eight microphones arranged in a planar ring-shaped array, 1m from the floor and 2.5m from the 7m wall. The microphones are equally spaced, and the ring diameter is 15cm. Our proposed approaches work with 1D, 2D or 3D SSL. But due to page limitation, we focus on the 1D and 2D cases: the azimuth θ and elevation ϕ of the source with respect to the center of the microphone array. For θ, the whole 0º-360º range is quantized into 360º/4º = 90 levels. For ϕ, because of our tele-conferencing scenario, we are only interested in ϕ = [50º, 90º], i.e., if the array is put on a table, ϕ = [50º, 90º] cover the range of meeting participant's head position. It is quantized into (90º-50º)/5º = 8 levels. For the whole θ-ϕ 2D space, the number of cells L = 90*8 = 720.

We have designed three sets of data for the experiments:

Test A: Varies θ from 0º to 360º in 36º steps, with fixed ϕ = 65º, SNR = 10dB, and reverberation time T_{60} = 100ms;

Test R: Varies the reverberation time T_{60} from 0ms to 300ms in 50ms steps, with fixed θ = 108º, ϕ = 65º, and SNR = 10dB;

Test S: Varies the SNR from 0db to 30db in 5dB steps, with fixed θ = 108º, ϕ = 65º, and T_{60} = 100ms.

Sampling frequency is 44.1 KHz, and we use a 1024 samples (~23ms) frame. The raw signal is band-passed to 300Hz-4000Hz. Each configuration (e.g., a specific set of θ, ϕ, SNR and T_{60}) of the testing data is 60-second long (2584 frames) and about 700 frames are speech frames. The results reported in this section are from all of the 700 frames.

4.2. Experiment 1: 1-TDOA SSL

Table 1 compares the proposed 1-TDOA approach and the existing 1-TDOA. The left half of the table is for Test R and the right half is for Test S. The numbers in the table are the "wrong count", defined as the number of estimations that are more than 10º from the ground truth (i.e., higher is worse).

4.3. Experiment 2: SB SSL

The comparison between the proposed new SB approach against existing SB approaches is summarized in Table 2.

Table 1 - Comparison between 1-TDOA approaches

Wrong count		Reverberation time (ms)							SNR (db)						
		0	50	100	150	200	250	300	0	5	10	15	20	25	30
θ	New	0	4	7	17	27	53	82	47	13	7	4	4	4	4
	Phat	2	5	10	10	20	45	75	80	19	10	6	4	4	4
	ML	0	1	20	76	124	172	230	36	23	20	27	27	28	26

Table 2 - Comparison between SB approaches

Wrong count		Reverberation time (ms)							SNR (db)						
		0	50	100	150	200	250	300	0	5	10	15	20	25	30
θ	New	1	5	6	17	27	52	89	44	11	6	5	4	4	4
	Phat	2	5	9	10	21	50	75	78	19	9	6	5	4	4
	ML	0	1	20	79	122	172	226	33	22	20	29	28	28	27

Table 3 - Comparison between 2-TDOA, 1-TDOA and SB using tests R and S.

Wrong count		Reverberation time (ms)							SNR (db)						
		0	50	100	150	200	250	300	0	5	10	15	20	25	30
θ	2TDOA	4	4	12	25	49	80	140	46	18	12	8	7	8	8
	1TDOA	0	4	7	17	27	53	82	47	13	7	4	4	4	4
	SB	1	5	6	17	27	52	89	44	11	6	5	4	4	4
φ	2TDOA	4	7	27	151	295	409	504	83	37	27	25	23	19	21
	1TDOA	0	3	11	54	133	210	276	17	14	11	9	7	7	7
	SB	1	2	11	76	176	264	335	18	17	11	12	8	8	8

Table 4 - Comparing 2-TDOA, 1-TDOA and SB using test A

Wrong count		Different azimuth (degrees)									
		0	36	72	108	144	180	216	252	288	324
θ	2TDOA	3	11	3	12	4	1	6	9	6	10
	1TDOA	0	16	2	7	2	0	3	5	2	10
	SB	0	15	2	6	2	1	3	4	2	10
φ	2TDOA	65	287	14	27	23	33	24	29	21	304
	1TDOA	30	134	3	11	8	14	7	6	6	157
	SB	36	169	2	11	9	18	12	8	6	195

4.4. Experiment 3: 2-TDOA vs. 1-TDOA vs. SB

The comparison between the proposed new 1-TDOA and SB approaches against an existing 2-TDOA approach is summarized in Table 3. The 2-TDOA approach we use is the maximum likelihood estimator J_{TDOA} developed in [2], which is one of the best 2-TDOA algorithms. In addition to use Tests R and S, we further use Test A to see how they perform with respect to different source locations. The result is summarized in Table 4.

The following observations can be made based on Tables 1-4:

- From Table 1, the proposed new 1-TDOA outperforms the PHAT and ML based approaches. The PHAT approach works quite well in general, but performs poorly when the SNR is low. Tele-conferencing systems, e.g., [4], require prompt SSL, and the promptness often implies working with low SNR. PHAT is less desirable in this situation. A similar observation can be made from Table 2 for the SB SSL approaches.
- From Tables 3 and 4, both the new 1-TDOA and the new SB approaches perform better than the 2-TDOA approach, with the 1-TDOA slightly better than the SB approach, because of its good weighting functions. This result matches our analysis that 2-TDOA throws away useful information during the first step.
- Because our microphone array is a ring-shaped planar array, it has better estimates for θ than for φ (see Tables 3 and 4). This is the case for all the approaches.
- There are two destructive factors for SSL: the ambient noise and room reverberation. It is clear from the tables that when ambient noise is high (i.e., SNR is low) and /or when reverberation time is large, the performance of all the approaches degrades. But the degrees they degrade differ. Our proposed 1-TDOA is the most robust in destructive environment.

5. CONCLUSIONS

The main algorithms for multiple microphones SSL are the 2-TDOA, and two direct approaches (SB and 1-TDOA). We developed a unified framework including all three approaches, pointing out their similarities and differences. We analyzed and explained why direct approaches are more robust than the widely used 2-TDOA. We further proposed two new direct approaches. Experimental results demonstrate superior SSL performance of the proposed approaches over existing 2-step and direct approaches.

6. REFERENCES

[1]. S. Birchfield and D. Gillmor, Acoustic source direction by hemisphere sampling, *Proc. of ICASSP*, 2001.

[2]. M. Brandstein and H. Silverman, A practical methodology for speech localization with microphone arrays, Technical Report, Brown University, November 13, 1996.

[3]. M. Brandstein and D. Ward (Eds.), Microphone Arrays signal processing techniques and applications, Springer, 2001.

[4]. R. Cutler, Y. Rui, et. al., Distributed meetings: a meeting capture and broadcasting system, Proc. of ACM Multimedia, Dec. 2002, France.

[5]. J. DiBiase, A high-accuracy, low-latency technique for talker localization in reverberant environments, PhD thesis, Brown University, May 2000.

[6]. R. Duraiswami, D. Zotkin and L. Davis, Active speech source localization by a dual coarse-to-fine search. *Proc. ICASSP* 2001.

[7]. J. Kleban, Combined acoustic and visual processing for video conferencing systems, MS Thesis, The State University of New Jersey, Rutgers, 2000.

[8]. Y. Rui and D. Florencio, Time delay estimation in the presence of correlated noise and reverberation, Microsoft Research Tech Report, 2002. http://www.research.microsoft.com/~yongrui/ps/TR.pdf

[9]. H. Wang and P. Chu, Voice source localization for automatic camera pointing system in videoconferencing, *Proc. of ICASSP*, 1997.

[10]. D. Ward and R. Williamson, Particle filter beamforming for acoustic source localization in a reverberant environment, *Proc. of ICASSP*, 2002.