SOUND SOURCE LOCALIZATION FOR CIRCULAR ARRAYS OF DIRECTIONAL MICROPHONES

Yong Rui, Dinei Florêncio, Warren Lam, and Jinyan Su

Microsoft Corporation
One Microsoft Way, Redmond, WA, USA

ABSTRACT

Previous research in sound source localization has helped increase the robustness of estimates to noise and reverberation. Circular arrays are of particular interest for a number of scenarios, particularly because they can be placed in the center of the sources. First, that improves the sound capture due to the reduced distance. Second, it helps on the direction estimation, not only because of the reduced distance, but also because it increases the angle differences. Nevertheless, most research on circular arrays focused on the case of omni-directional microphones. In this paper we present a new algorithm for sound source localization developed specifically for directional microphones. Results obtained from real meeting room setups show a typical error of less than 3 degrees.

1. INTRODUCTION

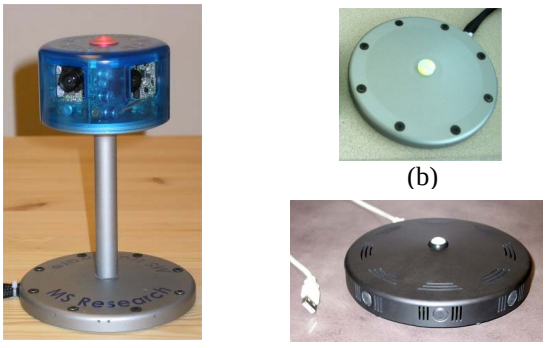
Sound source localization (SSL) has important application in a number of scenarios, including meeting recording and real-time audio-visual conferencing. In these scenarios, SSL can be used for directing a pan-tilt-zoom camera towards the speaker such that the viewing experience is more interesting and/or network bandwidth is used more efficiently [2]. The traditional configuration, with the camera at the end of the table, has been used for years. This configuration is convenient because it does not place significant constrains on the equipment size. Nevertheless, it interferes with the experience of distant participants, as most people are looking away from the camera.

Recently, with the reduced size of camera sensors, placing a camera in the center of the table has become practical, and provides a much better user experience. In the RingCam system [2], shown in Figure 1-a, the 360º camera array and the microphone array sit at the center of the meeting table, with meeting participants sitting around the table, as in a standard meeting. Thus, due to the positioning of the participants, we need the SSL algorithm to cover a 360º range.

Previous research on sound source localization has focused mostly on linear arrays[3][7], and can only resolve sound location within a small range, e.g., 150º, of angles. For 360º SSL, circular arrays are more appropriate. There are a few papers on circular arrays but they mostly focused on omni-directional microphones[1][5]. Nevertheless, other factors may influence the microphone selection. In particular, the microphones are often shared between the SSL and sound capture, as the case in the RingCam [2]. Since arrays using directional microphones provide significantly superior sound quality [4], that alone may determine the microphone selection. Accordingly, our new microphone array replaces the previous omni-directional microphones pointing upwards (see Fig 1-b) with directional microphones pointing outwards (see Figure 1-c). Among other advantages, this helps significantly reduce the noise from a projector, often placed directly above the microphone array, on the room ceiling.

While the change from omni-directional to uni-directional microphones may seem a small detail, one particular characteristic of uni-directional microphones makes SSL particularly challenging: their phase response varies significantly with frequency, direction, and even from microphone to microphone. Thus, in this paper we investigate a new SSL algorithm, developed specifically for circular uni-directional microphone arrays.

The remainder of this paper is organized as follows. In Section 2 we review the two basic techniques that serve as basis for the proposed algorithm. Section 3 derives the separable weighting function and addresses the phase variability issue. It further gives a block diagram of the complete algorithm. Section 4 presents results from real-world recordings, showing good results, with typical errors below 3 degrees. Section 5 concludes the paper.



(a) (b) (c)

Figure 1. Circular arrays. *(a-b) the original RingCam design with omni-directional microphones pointing upwards, (c) the new design with unidirectional microphones pointing outwards.*

2. REVIEW OF BASIC ALGORITHMS

The proposed algorithm is a hybrid between steered beam SSL (SB-SSL) and the one-step time-delay-of-arrival (1-TDOA) SSL[5]. Let us now review how these algorithms work for omni-directional microphone arrays.

2.1. Reviewing SB-SSL

SB-SSL localizes the sound source through hypothesis testing – pick as the sound source location the point in the space producing the highest energy after adding the delayed signals.

More formally, let *M* be the number of microphones in an array. We can model the signal received at microphone *m*, where *m = 1, …, M*, at time *n* as:

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

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. More precisely, if p is a point in the 3-D space, then the selected source location p* is:

$$p^* = \arg\max_l \left(\sum_{m=1}^{M} x_m(t - \tau_m^l) \right)^2 \quad (2)$$

where τ_m^l is the time it takes sound to travel from a source at location *l* to microphone *m*. To reduce complexity, usually only a finite number of points *L* are investigated. Note that SB-SSL does not account for noise or reverberation.

2.2. Reviewing 1-TDOA

The 1-TDOA [5] finds the sound source location by maximizing the generalized cross-correlation between the several microphones, i.e.,:

$$p^* = \arg\max_l \left\{ \sum_{r=1}^{M} \sum_{s \neq r}^{M} \left| W_{rs}(f) X_r(f) X_s^*(f) e^{-j2\pi f(\tau_r - \tau_s)} \right|^2 \right\} \quad (3)$$

where *r* and *s* are indexes for a pair of microphones, and *X(f)* is the Fourier transform of *x(n)*. The weighting function $W_{rs}(f)$ has the objective of minimizing the effects of reverberation and noise. In [6] we derived the maximum likelihood estimator when both noise and reverberation are present. The corresponding weighting function W_{MLR} is:

$$W_{MLR} = \frac{|X_r||X_s|}{2q|X_r|^2|X_s|^2 + (1-q)|N_r|^2|X_r|^2 + |N_s|^2|X_s|^2} \quad (4)$$

where the dependency of f in X and N has been omitted for compactness. In [5], we proved that if no weighting function is used, 1-TDOA and SB-SSL are the same mathematically. However, 1-TDOA's strength is that it can derive a pair-wise optimal weighting function, i.e., W_{MLR} in Eq. (4). We next discuss how this 1-TDOA weighting function can be made separable, and be used in SB-SSL, hence the hybrid SSL.

3. THE PROPOSED ALGORITHM

There are two unique features in the proposed algorithm: using 1-TDOA's weighting function in SB-SSL, and the handling of circular uni-directional microphone array's phase issue. We now discuss these two topics in detail.

3.1. The separable weighting function

The W_{MLR} weighting function in Eq. (4) is pair-wise specific, as are most weighting functions for 1-TDOA. These functions are appropriate for use in Eq. (3), for example, where microphones always appear in pairs. In contrast, SB-SSL does not usually use a weighting function. Nevertheless, if one were designed for use directly in Eq. (2), the weighting function would need to be mic specific, instead of a different function for each mic pair (this is mathematically equivalent to requiring the pair-wise function to be separable). This can be seen as a hybrid approach between SB-SSL and 1-TDOA, since we are using a weighting function – typical of 1-TDOA – but at the same time forcing it to fit the separability requirement of SB-SSL. Such a separable function can be used in Eq. (3), and – since it is separable – incorporated by simply pre-multiplying the Fourier transform of each mic signal. This significantly reduces the overall computational complexity in a hypothesis testing implementation of Eq. (3). This is exactly the approach we take: we modify Eq. (3) by including a separable weighting function:

$$p^* = \arg\max_l \left\{ \sum_{r=1}^{M} \sum_{s \neq r}^{M} \left| W_r X_r W_s^* X_s^* e^{-j2\pi f(\tau_r - \tau_s)} \right|^2 \right\} \quad (5)$$

and adopt a weighting function obtained by assuming that the reverberation and noise are constant across microphones, i.e.,:

$$W_m(f) = \frac{1}{q|X_m(f)| + (1-q)|N_m(f)|} \quad (6)$$

This choice of weighting function has most of the advantages of more sophisticated functions, but can be computed much faster in a hypothesis testing algorithm[5].

3.2. The phase problem

One of the problems of using directional microphones for SSL is the "unruly" phase characteristics of these microphones. For typical unidirectional microphones the phase has reasonably flat response for lower angles, but it becomes highly variable as the angle of arrival increases. In fact, for hypercardiod microphones, the phase may actually flip by 180º for sources coming from the back. To make things even worse, this phase response is frequency dependent. This phase variation is not necessarily a

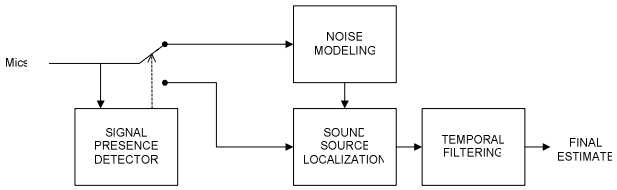


Figure 2 - *Block diagram of the overall SSL algorithm.*

problem for *linear* arrays – since the angle of arrival is roughly the same for all microphones. However, they become a strong problem for *circular* arrays: with the traditional configuration of microphones pointing outward, sound arrives at the microphones from directions essentially spaced uniformly over the 360º span. Indeed, our simulations showed that ignoring the phase behavior of the microphones increased significantly the variance and bias in the estimates (see Table 1 in Section 4). Similarly, trying to compensate for the phase shift by incorporating a model of the phase behavior does not help either, as mic-to-mic variations are too strong.

To handle the phase problem, we decided to only select the microphones that are facing the sound source. Specifically, we measure the flat region of the particular microphone type, and only include in the computation those that fall in the predictable phase region. This is possible in our algorithm because we use the hypothesis testing system. For the microphones we use (ISL CM9752), the predictable phase region extends beyond 95º to each side from the main direction (190º total). So, we only include in the computation the microphones whose angle is less then the cutoff angle of 95º. This has also the additional advantage of reducing computation by a factor of two. We call this approach cutoff angle approach.

It is worth noting that the proposed approach is very different from trying to compensate the phase/gain shift directly. In the compensation approach, it will only work if the phase/gain shift measurement is perfect, which is not possible in reality. In Section 4, we will show that when the measurement is less than perfect, trying to compensate will only hurt the performance.

3.3. The overall algorithm flow chart

Figure 2 presents a block diagram of the overall algorithm. The signals are received from the microphones on a frame basis (20ms). The first step is a speech (signal) presence detector. If no signal is detected in the frame, the signal is simply passed to the noise modeling module to update the model. If signal is present, then an estimate for the direction on the particular frame is made, based on the hybrid algorithm described in the previous section. Finally, this estimate goes through a temporal filtering step, which removes spurious estimates. This temporal filter accumulates multiple (we use 40 in our current system) frame-level estimates. If the multiple estimates yield a significant trend for a particular direction, that direction is declared as the sound source; otherwise, i.e., no significant trend, the filter decreases the confidence of estimate and does not report a sound source.

4. EXPERIMENTS AND RESULTS

4.1. Test data description

Our microphone array is planar, ring-shaped, and has six uni-directional microphones, each pointing outwards (see Figure 1-c). The microphones are equally spaced, with a 14cm radius. During data collection sessions, the microphone array is placed roughly at the center of a conference table, and people are sitting/standing around the table.

We have collected two sets of data. The first set consists of 30 6-channel recordings of a single sound source and the second set consists of 18 6-channel recordings of multiple sound sources. Each recording is between 20 and 200 seconds long. All the 48 recordings are collected from real-world environment. The conference rooms that we use range from moderately quiet to very noisy. Some have projector fans and others have whiteboards with strong acoustic reflection. The room sizes range from 3.6m×6m to 5.4m×12m.

4.2. Handling the phase problem

As discussed in Section 4, uni-directional microphones arranged in a circular fashion impose great challenges to SSL. Table I shows the comparison between no phase compensation, with phase compensation and use cutoff angles.

We can make the following observations. Because of the big variations among different microphones and different frequencies, modeling and compensating for the phase shift is not easy. We used various approaches to estimate the phase shift. Still, the phase-compensated SSL gives the worse result, i.e., present a large bias and large standard deviation (std) and a small number of

Table I. Performance comparison between different approaches. Bias *is the average difference from the ground truth.* Std *is the standard deviation of the estimates, and* #Fs *is the number of frames that each algorithm reports (the larger the number, the more responsive the algorithm is).*

	No Compensation			With Compensation			Use cutoff angle		
	Bias	Std	#Fs	Bias	Std	#Fs	Bias	Std	#Fs
T1	-1.9	1.2	249	29.8	79.8	195	-3.9	0.6	338
T2	0.0	0.0	298	156.9	21.2	54	0.0	0.0	320
T3	0.7	0.2	205	-146.3	94.6	229	0.8	0.0	282
T4	-0.6	0.6	112	-168.2	5.7	34	-2.6	0.2	429
T5	0.1	0.2	153	159.6	16.5	271	0.0	0.0	422
T6	7.0	29.9	66	160.0	4.0	426	2.5	0.1	419
T7	24.3	52.6	43	114.7	49.3	76	-2.7	0.4	351
T8	-0.2	0.4	161	-151.3	0.7	3	-1.0	0.4	333
T9	-3.6	0.6	92	-132.7	89.7	154	-3.8	0.4	450

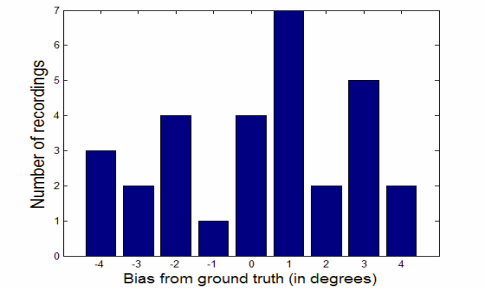


Figure 3 – *Bias histogram. On the test set, all biases were within 4º from the ground truth.*

successfully classified frames (#Fs). Comparing no compensation and the proposed cutoff angle approach, we can see that while the former sometimes gives lower bias, e.g., T1 and T4, other times it gives very large bias and std, e.g., T6 and T7. In addition, the no compensation approach gives a small number of frames, which indicates it is not responsive. In contrast, the proposed cutoff angle approach consistently gives low bias, low std and high number of frames. In addition, it cuts the computation cost by half.

4.3. Real-world single source cases

Figure 3 shows the histogram of the 30 recordings in terms of their bias from ground truth. They have an average bias of less than 4 degrees. The histogram peak occurs when the bias is less than 1 degree. We do not have space to show the histogram for the std. But out of the 30 recordings, 28 give std that is less than 1 degree. Combining the bias and std statistics, we can see that the proposed approach not only gives accurate estimates, but also gives consistent estimates.

4.4. Real-world multi-source cases

In this test set, we have 18 recordings, and we want to test how the proposed approach will behave in spontaneous environment. Because of space limitations, we only show the result of one example recording (see Figure 4). The dark (blue) curve is the ground truth, and the light (pink) curve is the estimate. It can be seen that the proposed approach works in most of the cases. There are, however, a few false positive estimates towards the end of the recording. Listening to the audio, those places correspond to paper shuffling and other non-speech sound. This indicates that in order to have good overall performance, other modules (see the block diagram in Figure 2), e.g., pre-processing and speech/non-speech classification, are as important as the SSL algorithm itself. But overall, our system gives very good performance in the spontaneous multi-source test cases.

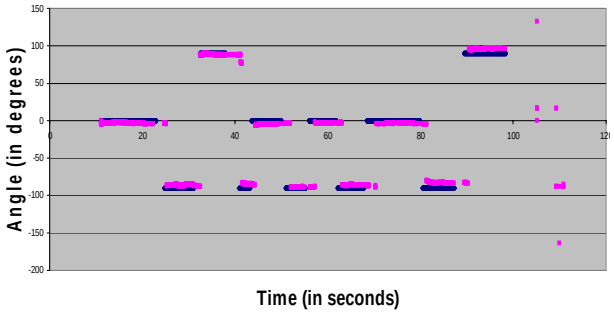


Figure 4 – *sample results on a multi-speaker test. The sources are at 0º, -90º, and +90º.*

5. CONCLUSIONS

A new scenario emerged recently, where a circular unidirectional mic array is place at the center of a conference table to conduct both sound capture and SSL. Where this particular mic array configuration supports good sound capture, its phase variations (across different frequencies, directions and microphones) poses challenges to SSL. In this paper, we have proposed a new SSL algorithm to address this issue. Specifically, two features make the proposed algorithm unique. First, instead of using all the microphones whose phase patterns are very difficult to estimate/compensate, the algorithm selects the right subset of the microphones such that not only the SSL is more robust, but also cuts the computation cost. Second, the proposed algorithm decomposes a robust 1-TDOA pair-wise weighting function into a separable mic-wise weighting function for SB-SSL. Our experiments on 48 single- and multi-sources cases show that the proposed algorithm is both robust and accurate and gives a typical error around 3 degrees.

ACKNOWLEDGEMENTS

The authors would like to thank Ivan Tashev for building the microphone arrays, Ross Cutler and Henrique Malvar for interesting discussions**.**

REFERENCES

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

[2] Cutler, R., et al. "Distributed Meetings: A Meeting Capture and Broadcasting System", *Proc. of ACM Multimedia 2002*, Dec. Juan-les-Pins, France

[3] Kleban, J., Combined acoustic and visual processing for video conferencing systems, MS Thesis, Rutgers University, 2000

[4] Tashev, I. and Malvar, H., "A new beamformer design algorithm for microphone arryas", Proc. of *ICASSP*, 2005

[5] Rui, Y. and Florencio, D., "New direct approaches to robust sound source localization," Proc. of *ICME*, 2003

[6] Rui, Y. and Florencio, D., "Time Delay estimation in the presence of correlated noise and reverberation," Proc. of *ICASSP*, 2003

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