Cost Function for Sound Source Localization with Arbitrary Microphone Arrays

Ivan J. Tashev
Microsoft Research Labs
Redmond, WA 98051, USA
ivantash@microsoft.com

Long Le
Dept. of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign,
IL 61820, USA
longle1@illinois.edu

Vani Gopalakrishna, Andrew Lovitt
Microsoft Corporation
Redmond, WA 98051, USA
{vanig, anlovi}@microsoft.com

***Abstract*—The sound source localizer is an important part of any microphone array processing block. Its major purpose is to determine the direction of arrival of the sound source and let a beamformer aim its beam towards this direction. In addition, the direction of arrival can be used for meetings diarization, pointing a camera, sound source separation. Multiple algorithms and approaches exist, targeting different settings and microphone arrays. In this paper we treat the sound source localizer as a classifier and use as features the phase differences and magnitude proportions in the microphone channels. To determine the proper mix, we propose a novel cost function to measure the localization capability. The resulting algorithm is fast and suitable for real–time implementations. It works well with different microphone array geometries with both omnidirectional and unidirectional microphones.**

***Index Terms*—sound source localization, phase differences, magnitude proportions, cost function, various geometries.**

I. Introduction

Localization of sound sources is a part of any microphone array which uses beamsteering to point the listening beam towards the direction of the sound source. The output of the localizer is also used by post-filters to additionally suppress unwanted sound sources [1], [2]. The idea of spatial noise suppression was proposed in [3] and further developed in [4], where a *de facto* sound source localizer per frequency bin is used to suppress sound sources coming from undesired directions for each frequency bin separately. The probability of the sound source, coming from a given direction, is estimated based on the phase differences only. A similar approach, adapted for a very small microphone array with directional microphones, was proposed in [5]. It uses both magnitudes and phases to distinguish desired from undesired sound sources. Therefore, localization of sounds with microphone arrays is a well–studied area with multiple algorithms defined over the years. The overall architecture of a sound source localizer is described in [6]. Typically, a sound source localizer (SSL) works in the frequency domain and consists of a voice activity detector, a per-frame sound source localizer (when an activity is detected), and a sound source tracker (across multiple frames). In this paper we will discuss algorithms for the per-frame SSL only. It is assumed that the microphone array geometry is known in advance: position, type and orientation for each of the microphones.

There are two major approaches for the sound source direction of arrival (DOA) estimation with microphone arrays: delay estimation and steered response power (SRP). The first approach splits the microphone array on pairs and estimates the time difference of arrival, usually using Generalized Cross-correlation function (GCC) [7]. The DOA for a pair can be determined based on the estimated time difference and the distance between two microphones. Averaging the estimated DOAs across all pairs provides poor results, as some of the pairs may have detected stronger reflections. In [8] the DOA is determined by combining the interpolated values of the GCCs for given delays corresponding to a hypothetical DOA. By scanning the space of DOAs, the direction with highest combined correlation can be determined. Since this process is computationally expensive, a coarse-to-fine scanning is proposed in [9]. SRP approaches vary from simply trying a set of beams and determining the direction from where the response power is highest, to a more sophisticated weighting of frequency bins based on the noise model [10]. One of the most commonly used and precise SSL algorithms is MUltiple SIgnal Classification (MUSIC) [11]. It is also one of the most computationally expensive. A good overview of the classic sound source localization algorithms can be found in Chapter 8 of [12].

In this paper we address two problems. The first is that most of the classic SSL algorithms are computationally expensive (GCC, FFTs, interpolations), which makes it difficult to be implemented in real-time systems. The second problem is that in many cases the SSL algorithms are tailored for a given microphone array geometry (linear, circular; large, small; omnidirectional or unidirectional microphones) and their performance degrades for different geometries. To address this degradation of performance, especially for microphone arrays with directional microphones, pointing in different directions, we propose to utilize the magnitudes as a feature. Practically all SSL algorithms assume omnidirectional microphones and for far-field sound sources there is no substantial difference in magnitudes across the channels. We assume that the same algorithm will be used for handling different microphone array geometries, and that the geometry is known before the processing starts, which allows faster execution during runtime. The general idea is to extract a set of features (differences in phases and proportion in the magnitudes for each microphone pair in each frequency bin) from the current frame, and to combine it into a cost function. This cost function of a hypothetic DOA is expected to have a sharp maximum at the sound source direction.

In section II we provide the modeling equations. Section III defines the preparation of the runtime tables and the cost function, in section IV is described the runtime of the SSL, while section V provides the experimental results. We draw some conclusions in Section VI.

II. Modelling

Given a microphone array of M microphones with known positions $p_m = (x_m, y_m, z_m) : m = 1, 2, \cdots, M$; the sensors have known directivity pattern $U_m(f, c)$, where $c = \{\varphi, \theta, \rho\}$ represents the coordinates of the sound source in a radial coordinate system and f denotes the signal frequency. After framing, weighting, and converting to frequency domain each microphone receives:

$$X_m(f, p_m) = D_m(f, c)\,.S(f) + N_m(f), \tag{1}$$

where the first term in the right-hand side,

$$D_m(f, c) = \frac{e^{-j2\pi f \frac{\|c-p_m\|}{\nu}}}{\|c - p_m\|}.U_m(f, c), \tag{2}$$

represents the delay and attenuation from the sound source to the microphone in non-echoic environment, ν is the speed of sound, $S(f)$ is the source signal and the last term, $N_m(f)$ is the captured noise. For digital processing is assumed that the audio frame has K frequency bins and we substitute further the frequency f with the discrete value k.

Let $c_n, n = 1, 2, \cdots, N$ be a set of N points in the space, evenly covering the expected locations of the sound sources. For example, 72 points every 5^o covering evenly the full circle around the microphone array in the horizontal plane. Then we can define a set of features for each frequency bin, microphone pair $l = \{i, j\}$, and direction c_n. The expected phase differences and magnitude proportions feature sets are:

$$\begin{aligned} \delta_\theta\left(l,k,n\right) &\triangleq ang\left(D_i\left(k,c_n\right)\right) - ang\left(D_j\left(k,c_n\right)\right) \\ \delta_M\left(l,k,n\right) &\triangleq \log\left(\frac{|D_j(k,c_n)|}{|D_i(k,c_n)|}\right). \end{aligned} \tag{3}$$

Note that microphone pairs can be formed with one reference channel (for example $i = 1 = const$), resulting in $L = M - 1$ pairs, or we can have a full set of unique pairs with a total number of $L = M\left(M-1\right)/2$ pairs. This difference tensor is of dimension $L \times K \times N$. While the values of the angle differences are naturally limited, the logarithm of the magnitudes proportions should be limited to certain minimal and maximal values. We can approximate the log() function with linearized one that's faster to compute:

$$\log(x) \approx \begin{cases} g_1\left(x-1\right) & x \geq 1 \\ g_2\left(1-x\right) & \text{otherwise} \end{cases} \tag{4}$$

where g_1 and g_2 are optimized to minimize the error in the given interval. We will continue to use log() further, but all results are obtained using this faster interpolation.

III. Preparation of the Runtime Tables

For each frequency bin we have a hyper-curve (defined in N points) that lies in an L-dimensional space. Each point from the real space has image in this L-dimensional space, but the opposite is not correct. We can compute the square of the Euclidian distance between each point n_{ref} to the rest of them:

$$\begin{aligned} \Delta_\theta\left(n,k,n_{ref}\right) &= \sum_{l=1}^{L}\left|\delta_\theta\left(l,k,n\right) - \delta_\theta\left(l,k,n_{ref}\right)\right|^2 \\ \Delta_M\left(n,k,n_{ref}\right) &= \sum_{l=1}^{L}\left|\delta_M\left(l,k,n\right) - \delta_M\left(l,k,n_{ref}\right)\right|^2 \end{aligned} \tag{5}$$

The two feature sets can be combined into one giving different weight to phases and magnitudes:

$$\Delta\left(\alpha,n,k,n_{ref}\right) = \alpha\tilde{\Delta}_\theta\left(n,k,n_{ref}\right) + \left(1-\alpha\right)\tilde{\Delta}_M\left(n,k,n_{ref}\right) \tag{6}$$

in a way to maximize the ability for sound source localization. For some geometries and frequencies, the phase differences bring more information, while it is the magnitudes for other geometries and frequencies. For each frequency bin we have $N \times N$ matrix. We define the following cost function as a selectability measure of the combined feature set:

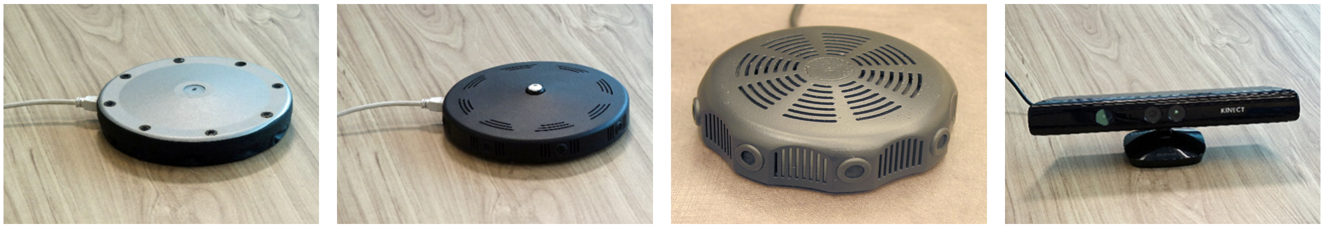
$$\begin{aligned} Q\left(\alpha,k\right) = \quad &\left(\tfrac{1}{NN}\sum_{n_1=1}^{N}\sum_{n_2=1}^{N}\left|\Delta\left(\alpha,n_1,k,n_2\right)\right|\right) - \\ &\left(\tfrac{1}{N}\sum_{n=1}^{N}\left|\Delta\left(\alpha,n,k,n\right)\right|\right) \end{aligned} \tag{7}$$

Simply put we want a weight α which maximizes the difference between $\|\ell\|_1$ (average distance to all directions) and the average diagonal (the distance to the hypothetic direction). By the definition in (5), the diagonal values of $\mathbf{\Delta_k}$ are zeros, so the second half of (7) is always zero. Then we can compute the optimal α for each frequency bin:

$$\alpha\left(k\right) = \arg\max_\alpha\left(Q\left(\alpha,k\right)\right) \tag{8}$$

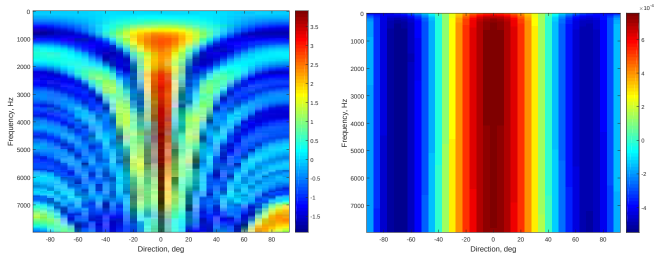
The last step in the preparation is to find the best way to combine the data from all frequency bins into one cost function for the entire frame. In general, we should not consider the lower frequency bins where the phase differences are small and smeared by the noise. Also, above certain frequency the spatial aliasing is decreasing the ability to localize the sound source. We will combine the per-bin localization functions into one per-frame by averaging them within a frequency band:

$$\Delta\left(n,n_{ref}\right) = \frac{1}{k_{end} - k_{beg} + 1}\sum_{k=k_{beg}}^{k_{end}}\Delta\left(n,k,n_{ref}\right) \tag{9}$$



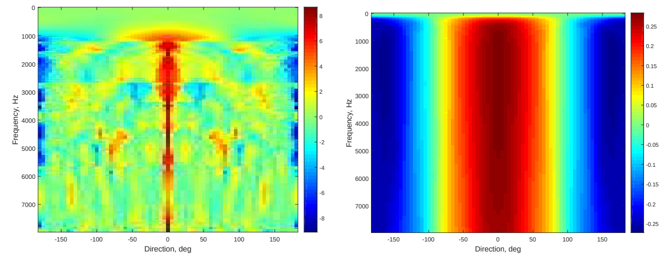
(a) 140 mm, omnidirectional (b) 140 mm, cardioid (c) 50 mm, cardioid (d) 225 mm, cardioid

Fig. 1. Circular and linear microphone arrays



(a) Phases (b) Magnitudes

Fig. 2. Kinect: Normalized distance criteria as function of the frequency and hypothesis angle, all pairs.



(a) Phases (b) Magnitudes

Fig. 3. Cardioid 140 mm: Normalized distance criteria as function of the frequency and hypothesis angle, all pairs.

where k_{beg} and k_{end} are selected in a way to maximize the selectability criterion, defined in (7).

At this point, based only on the geometry of the microphone array, we have calculated the expected differences functions δ_θ and δ_M, the combining weights vector α and the usable bandwidth $[k_{beg}, k_{end}]$.

IV. Runtime Localization

At runtime, after detecting sound activity in the current frame, the sound source localizer receives the complex matrix of size $M \times K$, containing the DFT coefficients of the M microphones. A classifier uses this input to find the direction. The first step is to compute the phase difference and magnitude proportion matrices:

$$\begin{aligned} \hat{\delta}_\theta(l,k) &= ang\left(X(i,k)\right) - ang\left(X(j,k)\right) \\ \hat{\delta}_M(l,k) &= \log\left(|X(i,k)|/|X(j,k)|\right) \end{aligned} \tag{10}$$

where the differences matrices are of size $L \times K$. Then we can compute the feature set, which is the squared Euclidian distance between the observation and the model for each hypothetic DOA:

TABLE I
Microphone Arrays

Geometry	Size, mm	Mics	Mic type and orientation
Circular	140	8	Omnidirectional
Circular	140	8	Cardioid, outward
Circular	50	8	Cardioid, outward
Linear	225	4	Cardioid, front - Kinect
Endfire	6	2	Subcardioid, pointing opposite

$$\begin{aligned} \hat{\Delta}_\theta(n,k) &= \sum_{l=1}^{L}\left|\hat{\delta}_\theta(l,k) - \delta_\theta(l,k,n)\right|^2 \\ \hat{\Delta}_M(n,k) &= \sum_{l=1}^{L}\left|\hat{\delta}_M(l,k) - \delta_M(l,k,n)\right|^2 \end{aligned} \tag{11}$$

and combine these features according to (6) using the pre-computed frequency dependent weight $\alpha(k)$. Now we have an $N \times K$ matrix $\hat{\Delta}$ from which the selectability criterion can be computed as defined in (7). To maintain the same magnitudes range for various microphone array geometries, we normalize the matrix as follows:

$$\varphi(n,k) = \frac{\bar{\Delta}(k) - \hat{\Delta}(n,k)}{\bar{\Delta}(k)} \tag{12}$$

where $\bar{\Delta}(k)$ is the mean across all hypothetic directions. The values of this selectability criterion vary between zero and one, where a higher value indicates features values closer to the hypothetic DOA. Now we can reduce to a vector of length N by summing the rows from k_{beg} to k_{end}:

$$\mathbf{\Phi}(n) = \frac{1}{k_{end} - k_{beg} + 1}\sum_{k=k_{beg}}^{k_{end}} \varphi(n,k) \tag{13}$$

We can compute the selectability criterion for the entire frame $\left(\max_n(\mathbf{\Phi}) - \bar{\mathbf{\Phi}}\right) \Big/ \max_n(\mathbf{\Phi})$ and if it is above certain threshold η decide that we have a reliable sound source localization. The estimated DOA for the current audio frame is where $\mathbf{\Phi}$ has a maximum.

V. Experimental Results

For evaluation of the proposed cost-function and classification we selected several microphone array geometries, shown in

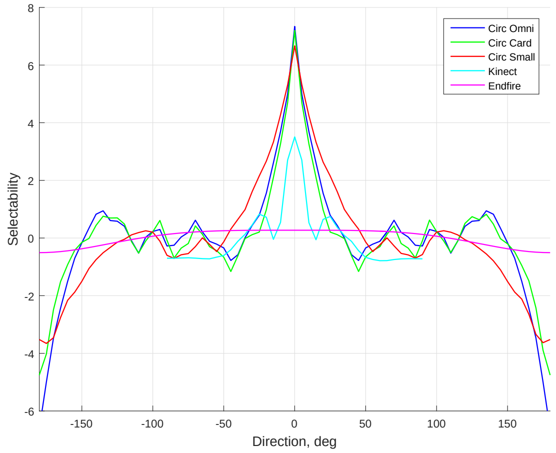


Fig. 4. Combined distance measures per frame for sound source at zero degrees, all pairs.

TABLE II
RESULTS PRE-RUNTIME

Geometry	Pairs	Q	α	beg F, Hz	end F, Hz
Circ 140 omni	Ref	6.994	1.00	1156	3906
	Unique	7.341	1.00	1156	3906
Circ 140 cardioid	Ref	7.191	1.00	1281	4219
	Unique	7.191	1.00	1281	4344
Circ 50 cardioid	Ref	5.711	0.98	1281	7969
	Unique	5.699	0.98	1281	7969
Linear 225	Ref	3.548	1.00	750	2625
	Unique	3.391	1.00	1281	7281
Endfire	Ref	0.428	0.89	1281	7969
	Unique	0.428	0.89	1281	7969

Table I, pictures of some of them are shown in Fig. 1. The sampling rate was set to 16 kHz, the hypotheses grid was set as points every 5^o, in the horizontal plane from -180^o to $+180^o$ for the circular arrays, from -90^o to $+90^o$ for the linear array, and at $[-180°, \pm 90°, 0°]$ for the endfire array. Theoretical directivity patterns were derived using the scripts provided in [6]. We evaluated as features phases only, magnitudes only, and both phases and magnitudes. Using one reference and all-unique-pairs was also a subject for evaluation. Some of the resulting distance measures are shown in Fig. 2 for Kinect and in Fig. 3 for 140 mm circular array with cardioid microphones. From the plots on the right it is visible that magnitudes do not provide noticeable selectability for the linear array and the contribution to the selectability for the circular array is negligible. In both cases, the phase differences feature set provides clear and well defined maximum.

TABLE III
LOCALIZATION ERRORS

Geometry	Pairs/alg	$\varepsilon, \%$	$\varepsilon \pm 1, \%$	Time, ms
Circular 140 mm omni	Ref	17	8	2.1
	Unique	13	4	7.3
	MUSIC	11	3	23.5
Circular 140 mm cardioid	Ref	31	24	2.1
	Unique	23	17	7.3
	MUSIC	34	29	23.5

The result after preparation of the run-time tables are shown in Table II. The table also provides the value of the cost function and the average of the phases weight α for the bandwidth from k_{beg} to k_{end}. The results in the table show the ability of the proposed approach to select the proper combination of the features (phases, magnitudes) based only on the microphone array geometry. Utilizing all pairs provides certain improvement in the large circular array with omnidirectional microphones, while it is less significant with the other arrays. In the case of the linear array utilizing all pairs actually worsens the selectability. The optimization procedure selected using mostly the phases for all microphone array geometries, except the endfire microphone array, which consists of two back to back subcardioid microphones.

The distance measures for the entire frame, according to equation (7), for all discussed geometries are plotted in Fig. 4. All discussed geometries provide well defined peak at the hypothetic DOA of the sound source, except the endfire array, which has only 6 mm distance between the microphones.

An evaluation with real audio recordings was done on two circular arrays with 140 mm diameter. The classification error is selected as evaluation criterion, defined as the percentage of the frames when the VAD triggered and the SSL did not estimate the correct direction. We added an additional criterion, the percentage of frames when the estimated direction is not in the correct and two neighboring directions. The recordings were completed in a conference room with sound source placed 2.0 meters from the center of the microphone array, normal noise (≈ 50 dBA SPL), and reverberation conditions (T_{60}=320 ms). The sound was produced by a head-and-torso-simulator, playing utterances from TIMIT database [13]. The sound source was placed in several different directions around the microphone array, with ten recorded files for each geometry. As a reference algorithm we used MUSIC, the overall implementation was done in Matlab according to the equations and the sample scripts in [6]. Besides measuring of the localization error, we recorded the localization execution time using the Matlab performance counters. The results are shown in Table III. The localization errors confirm the advantages provided by using all pairs. The proposed approach performs comparable to the reference MUSIC algorithm, while using significantly less computational time. Our approach uses only the four arithmetic operations, it does not contain square roots, logarithms, exponents. This allows very fast implementation, and even implementation using integer arithmetic.

VI. CONCLUSIONS

In this paper we proposed a generic algorithm for sound source localization using microphone arrays. It can work with a wide range of microphone array geometries, which are expected to be known in advance. After preparation of a set of tables, the run-time part of the algorithm is computationally efficient and allows fast implementation. Precision-wise, the proposed algorithm performs comparable to the MUSIC algorithm, which is much more computationally expensive.

References

[1] R. Zelinski, “A microphone array with adaptive post-filtering for noise reduction in reverberant rooms,” in *Proceedings of ICASSP*, 1988, vol. 5, pp. 2578–2581.

[2] I. McCowan and H. Bourlard, “Microphone array post-filter based on noise field coherence,” *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 709–716, November 2003.

[3] Ivan Tashev, Michael Seltzer, and Alex Acero, “Microphone array for headset with spatial noise suppressor,” in *Proceedings of Ninth International Workshop on Acoustic, Echo and Noise Control IWAENC*, Eindhoven, The Netherlands, September 2005.

[4] Ivan Tashev and Alex Acero, “Microphone array post-processor using instantaneous direction of arrival,” in *Proceedings of International Workshop on Acoustic, Echo and Noise Control IWAENC*, Paris, France, September 2006.

[5] Ivan Tashev, Slavy Mihov, Tyler Gleghorn, and Alex Acero, “Sound capture system and spatial filter for small devices,” in *Proceedings of Interspeech*, Brisbane, Australia, September 2008, International Speech Communication Association.

[6] Ivan J. Tashev, *Sound Capture and Processing: Practical Approaches*, Wiley, July 2009.

[7] C. Knapp and G. Carter, “The generalized correlation method for estimation of time delay,” *IEEE Transactions on Acoustic, Speech, and Signal Processing*, vol. ASSP-24, no. 4, pp. 320–327, August 1976.

[8] S. Birchfield and D. Gillmor, “Acoustic source direction by hemisphere sampling,” in *Proceedings of International Conference of Acoustic, Speech and Signal Processing ICASSP*, 2001.

[9] R. Duraiswami, D. Zotkin, and L. Davis, “Active speech source localization by a dual coarse-to-fine search,” in *Proceedings of International Conference of Acoustic, Speech and Signal Processing ICASSP*, Salt Lake City, Utah, USA, 2001.

[10] Y. Rui and D. Florencio, “New direct approaches to robust sound source localization,” in *Proceeding of IEEE International Conference on Multimedia and Expo (ICME)*, Baltimore, MD, USA, July 6-9 2003, pp. 737–740.

[11] R. Schmidt, “Multiple emitter location and signal parameter estimation,” *IEEE Transactions on Antennas and Propagation*, vol. AP-34, no. 3, pp. 276–280, March 1986.

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

[13] John S. Garofolo and et al., “TIMIT acoustic-phonetic continuous speech corpus,” Philadelphia, 1993, Linguistic Data Consortium.