

A METHOD FOR LOCATING MULTIPLE SOURCES FROM A FRAME OF A LARGE-APERTURE MICROPHONE ARRAY DATA WITHOUT TRACKING

Hoang Do, Harvey F. Silverman.

LEMS

Division of Engineering

Box D, Brown University, Providence, RI 02912

ABSTRACT

In this paper we present a new method for locating multiple sound sources using only a local segment of data from a large-aperture microphone array. The result of this work may be used directly or as an open-loop input to a tracking algorithm. The proposed method employs the proven-robust steered response power using the phase transform as a functional, agglomerative clustering, and low-cost global optimization (stochastic region contraction). Testing on real data from five talkers in a noisy environment, we show that, for each frame, our method finds correct locations of active sources under high noise and reverberation conditions without *a priori* knowledge of the number of sources.

Index Terms – Acoustic radiators, microphones, arrays, acoustic position measurement

1. INTRODUCTION

Locating multiple talkers using microphone arrays has many applications, such as: teleconferencing, speech data acquisition, and voice capture in adverse environments. There are two main approaches to solve this problem. The first approach uses a beamformer to find multiple peaks of an energy-based functional, such as the steered response power [1, 2]. The second one finds the time-differences of arrival (TDOA's), or directions of arrival (DOA's) from a multitude of microphone pairs and then estimates the source locations by using proper clustering techniques [3, 4]. Under high noise and reverberant conditions, strong reflections of the source signals severely affect the TDOA estimates, which create large errors in source-location estimation [5, 6]. Hence, the first approach will be more robust under such conditions.

In general, localization of multiple sources is done using a ‘closed loop’ tracking algorithm, which employs knowledge of prior source locations. Tracking can be done using particle filtering [1, 7] or Kalman filtering [8]. On the other hand, an ‘open loop’ problem is where multiple source locations are estimated based on current information (a single frame) only, without tracking. These ‘open loop’ estimates need to

be as accurate as possible to be useful as data into a tracking system.

In this paper, we propose a novel ‘open loop’ location method for multiple sources. This method, which belongs to the first approach, uses the proven-robust steered response power using the phase transform functional (SRP-PHAT) with agglomerative clustering (AC)[9], and the low-cost global optimization algorithm, stochastic region contraction (SRC)[10].

Assume a set of K point sources are active in data frame n at spatially separated locations $\vec{Q}_n(k)$, $k \in [1, K]$. $P_n(\vec{x})$ is the real-valued SRP-PHAT functional for the 3-D spatial vector \vec{x} obtained by *steering* a delay-and-sum beamformer. It has been fully described in [6, 10]. A typical slice for $P_n(\vec{x})$ at a fixed height $y = \text{constant}$ is shown in Fig. 1 for 5 talkers. The hypothesis is that we can isolate exactly K spatially separated peaks of $P_n(\vec{x})$ at locations $\vec{\lambda}_n(k)$ such that the set $\vec{\lambda}$ is the same as the true source location set \vec{Q} . The basic components of our algorithm for determining $\vec{\lambda}_n(k)$ are:

1. Evaluate $P_n(\vec{x})$ on a large set of R randomly selected points, keeping the highest N of them.
2. Agglomerative cluster these N points, obtaining an estimate of K and the K cluster volumes.
3. Apply stochastic region contraction on each volume to find $\vec{\lambda}_n(k)$, $1 \leq k \leq K$.

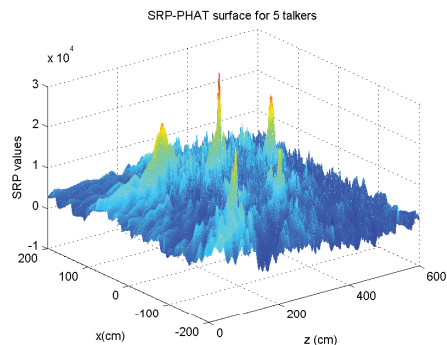


Fig. 1. SRP-PHAT 3D illustration for 5 talkers

2. AGGLOMERATIVE CLUSTERING (AC)

AC is chosen over the widely-used k -means clustering because it does not require a *priori* knowledge of the number of clusters, i.e., number of sources, K . It is also efficient for the small data set sizes that we use in this problem.

Denote i as the iteration index. For iteration i , cluster $C^{(i)}(k)$ has $|C^{(i)}(k)|$ points, where k is the cluster index, and $|\cdot|$ denotes the cluster cardinality. An assigned point from the vector space of this cluster is denoted as $\vec{p}_k^{(i)}(u)$, where $1 \leq u \leq |C^{(i)}(k)|$. Also, $\|\cdot\|$ denotes the Euclidean distance, and d_t is the Euclidean threshold distance that is chosen a *priori*. In our algorithm, d_t is set to 50 cm, the typical minimum distance that separates two human sources in a real life situation. The AC algorithm for an N -point data set is:

1. **Initialize:** $i = 0$. Start with N clusters, one for each point: $C^{(0)}(k)$, $k = 1, \dots, N$. Select the linkage parameter L ('average', 'simple' or 'complete').
2. **Calculate:** distance $d_{mn}^{(i)}$ between all pairs of clusters $C^{(i)}(m)$ and $C^{(i)}(n)$.
IF $L = \text{'average'}$:

$$d_{mn}^{(i)} = \text{mean}_{u,v} \|\vec{p}_m^{(i)}(u) - \vec{p}_n^{(i)}(v)\| \forall m, n$$

IF $L = \text{'simple'}$:

$$d_{mn}^{(i)} = \min_{u,v} \|\vec{p}_m^{(i)}(u) - \vec{p}_n^{(i)}(v)\| \forall m, n$$

IF $L = \text{'complete'}$:

$$d_{mn}^{(i)} = \max_{u,v} \|\vec{p}_m^{(i)}(u) - \vec{p}_n^{(i)}(v)\| \forall m, n$$

3. **Test:** IF $d_{mn}^{(i)} \geq d_t \forall m, n$:
STOP. KEEP RESULT.
4. **Merge:** $C^{(i)}(k_1)$ and $C^{(i)}(k_2)$ such that:

$$d_{k_1 k_2}^{(i)} = \min_{m,n} (d_{mn}^{(i)})$$

5. **Iterate:** $i = i + 1$. GO TO STEP 2.

3. AN ALGORITHM FOR MULTIPLE SOURCE LOCATION

Let \mathbf{V}_0 be the boundary vector of the rectangular search region with volume V_{room} containing the sources. SRC's parameters depend considerably on the environment's conditions, such as the room dimensions. Thus the algorithm's parameters, i.e., $R = 15000$ and $N = 500$ are determined empirically and shown in Sec.4. The algorithm is:

1. **Evaluate:** R random points in \mathbf{V}_0 .
2. **Select:** The best N points.

3. **Cluster:** N points into M clusters using AC with $L = \text{'average'}$.
4. **Determine:** M centroids: $\vec{c}_j \equiv \text{mean}(\vec{p}_j(u))$,
for all $\vec{p}_j(u) \in C(j)$, $j = 1, \dots, M$
5. **Calculate:** $M \times M$ Mahalanobis distances, μ_{ij} , between every \vec{c}_i and cluster $C(j)$.
6. **Test:** WHILE $\mu_{ij} \leq \mu_{\text{thres}} \forall i \neq j$ and $|C(i)| \neq 0$:
Merge $C(j)$ to $C(i)$; Set $|C(j)| = 0$.
7. **Apply SRC:** on each $C(k)$, $1 \leq k \leq P$ to achieve P estimates.
8. **Cluster:** P estimates using AC with $L = \text{'simple'}$.
Achieve Q clusters.
9. **Select:** The highest energy point \vec{h}_u in each cluster $Q(u) \subset Q$ that has $E_{\vec{h}_u} \geq E_{\text{noise}}$. The set $\{\vec{h}_u\}$ are the source location estimates.

Notes:

- $|\cdot|$ denotes cardinality of the set, and E denotes the energy or SRP-PHAT value.
- In Step 5, it is required that $C(j)$ has at least 2 points in order for the Mahalanobis distance to make sense, hence μ_{ij} of all $C(j)$ such that $|C(j)| = 1$ are set to infinity. In Step 6, $\mu_{\text{thres}} = 6$ (standard deviations) indicates the threshold that a point assuredly belongs to the cluster.
- In Step 7, the rectangular boundary of the volume containing cluster $C(i)$ for which SRC is applied is defined as follows:
 $\vec{B}_{\text{lower}} \equiv [x_{\min}(\vec{p}_i(n)) \ y_{\min}(\vec{p}_i(n)) \ z_{\min}(\vec{p}_i(n))]$,
 $\vec{B}_{\text{upper}} \equiv [x_{\max}(\vec{p}_i(n)) \ y_{\max}(\vec{p}_i(n)) \ z_{\max}(\vec{p}_i(n))]$
 $\forall \vec{p}_i(n) \in C(i)$.
- The parameters for SRC[10] used in Step 7 are: $J_0 = 1000$, $n = 100$.

The use of Mahalanobis distance in this algorithm is explained as follows. The clusters of high energy points appear to be spreading in an elliptical shape whose principle axis elongates along the direct paths from the sources to the microphones that we use, see Figure 2. Hence, Mahalanobis distance describes the correlation among data points in the clusters better than Euclidean distance. However, the Mahalanobis distance only makes sense when considering the relationship between a group of points with a single point or with another group, therefore we need some initial clusters to start with, before using Mahalanobis distance. AC in Step 3 provides an efficient preliminary clustering for that purpose.

Figure 3 shows the final clusters after merging AC clusters in Fig. 2 that have Mahalanobis distance less than μ_{thres} . SRC is then applied on each of these clusters to give the global maxima.

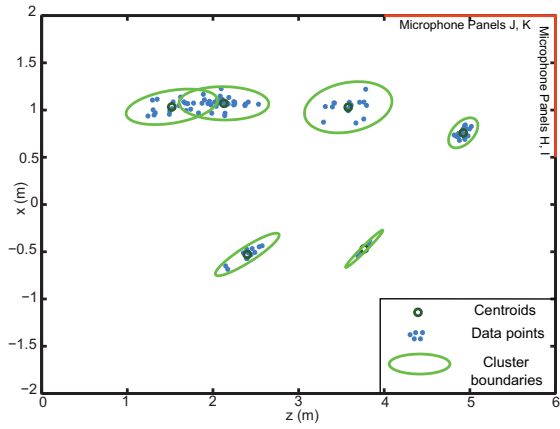


Fig. 2. Elliptical clusters for 5 talkers have principal axis directions defined by eigenvectors and lengths defined by eigenvalues of the covariance matrices of the data sets.

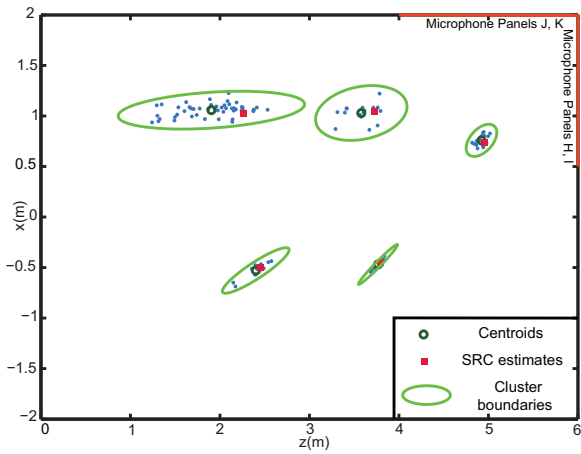


Fig. 3. Clusters after merging using Mahalanobis distance (Step 6) and applying SRC on them (Step 7).

4. EXPERIMENTS

4.1. Experimental conditions

The system, and room with a $T_{60} = 0.45s$ and a focal volume, $V_{\text{room}} = 4m \times 1m \times 6m$ that we used in our experiments has been described in [6]. 10 second recordings (wav files) of five native American English talkers (1 female and 4 males) carried on a conversation were played on Adobe Audition through five Advent AV009 speakers, approximately facing the 24 locator microphones as shown in Figure 4 with the average distances and SNR's indicated. Frames of 102.4ms, advancing each 25.6ms within the speech, and a sampling rate of 20 KHz were the conditions for testing. An estimate was considered an error if it were either off by more than 5cm in x or z or 10cm in y , the vertical dimension.

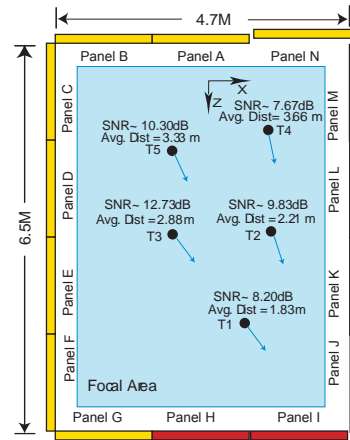


Fig. 4. Top View of the Array, showing Source Locations and Panels (Locator uses Microphones on Panels H, I, J, K). The arrows indicate the orientation of the talkers and the SNR's are for background noise only.

4.2. Determination of parameters

A preliminary experiment using the conditions described above was used to calculate the parameters R and N . From the data, we determined that $\frac{V_{\text{peak}}}{V_{\text{room}}} \approx 5 \times 10^{-4}$. Hence, from Table 1 in [10], $R = 15000$ will err by missing the peak volume less than 0.1% of the time. Also, an $N = 500$ gives us sufficient and efficient data points for clusters at source locations, and eliminates a large amount of noise outliers, i.e. see Fig. 5. These values of R , N provide correct cluster information as compared to the 2D energy map given by SRP-PHAT grid-search results.

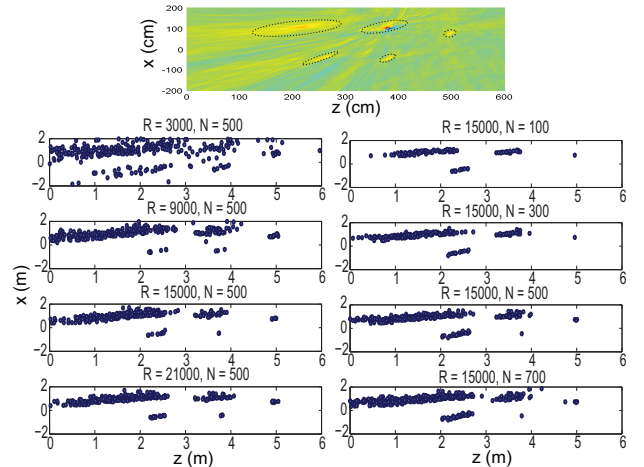


Fig. 5. 2D energy map given by grid-search (top) and different values of R and N for the 5 talker case

4.3. Experiments

The following experiments are conducted to evaluate the performance of the algorithm. The first task is to determine how

many talkers and which ones are active at every frame as the performance baseline. To do this, first we make recordings (label them as A,B,C,D and E) of five individual talkers, with a 100-ms long, 1KHz to 5KHz chirp inserted at 200ms into the beginning (silence portion) of the five playing WAV files (clean speech of A,B,C,D and E). From the correlation between the chirp segment of the clean speech data and the recorded data, we can determine the time $t_i, i = 1, \dots, 5$, of where the chirp starts in the recorded ones. Knowing that the chirps in the five clean speech data start at the same time, we now can synchronize A,B,C,D and E to start at t_i . We also align the recording of simultaneous five talkers (labeled as S) with the individual recorded data (A,B,C,D and E). Once these recorded data are synchronized in time, we calculate the SNR of each talker at each frame from A,B,C,D and E. If an SNR threshold of Θ dB is selected: At every frame, talkers who have $\text{SNR} \geq \Theta$ dB will be marked as active, thus creating a baseline for the evaluation of our proposed method, which processes on S. Figure 6 illustrates which talkers are active (the baseline) and the corresponding estimates given by the proposed algorithm for each frame, from frame 50 to 100 at $\Theta=20$ dB.

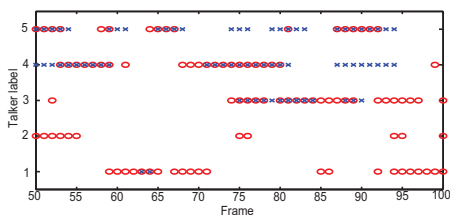


Fig. 6. $\Theta=20$ dB: Estimates of the baseline ('x' marks) and the proposed algorithm ('o' marks) from frame 50 to 100 for 5 talkers.

For different values of Θ , we calculate the percent matching between the correct estimates given by the proposed algorithm and the baseline. The result is shown in Figure 7. As we see from Fig. 7, $\Theta=26$ dB gives us 99% matching be-

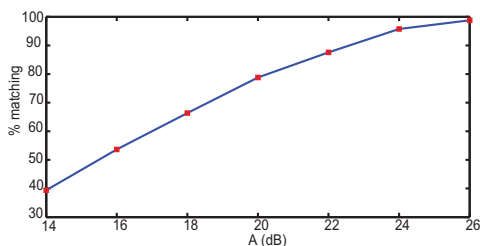


Fig. 7. Percent matching between the correct estimates given by the proposed algorithm and the baseline for different SNR values.

tween the correct estimates given by the proposed algorithm and the baseline. However, lower matching percent for lower values of SNR does not necessarily mean the proposed algo-

rithm is not correct. Note that the baseline is created from individual recorded data, while the proposed algorithm is run on recorded data of 5 talkers altogether. Hence, the baseline does not take the effects created by the interference among simultaneous active talkers into account as the proposed algorithm does. When the SNR decreases, these effects seem to increase as talkers are less distinguished, and they mingle with each other. Therefore, the percent matching for low SNR's will decrease.

5. CONCLUSION

We have presented a novel open-loop location estimate method for multiple talkers using SRP-PHAT with stochastic region contraction (SRC) and agglomerative clustering (AC). The new method correctly locates all talkers active in each frame, even under the high reverberant and noisy experimental conditions. Instead of using SRC, this method can also use coarse-to-fine region contraction (CFRC) that has been studied in [11] to find the maxima.

6. REFERENCES

- [1] J. Valin, F. Michaud, and J. Rouat, "Robust 3d localization and tracking of sound sources using beamforming and particle filtering," in *Proc. of ICASSP 2006*, Toulouse, France, May 2006, vol. 4, pp. 841–844.
- [2] D. N. Zotkin and R. Duraiswami, "Accelerated speech source localization via a hierarchical search of steered response power," *IEEE Trans. Speech, Audio Process.*, vol. 12, no. 5, pp. 499–508, Sept. 2004.
- [3] E. D. Di Claudio, R. Parisi, and G. Orlandi, "Multi-source localization in reverberant environments by root-music and clustering," in *Proc. of ICASSP 2000*, Istanbul, Turkey, June 2000, vol. 2, pp. 921–924.
- [4] T. Nishiura, T. Yamada, S. Nakamura, and K. Shikano, "Localization of multiple sound sources based on a csp analysis with a microphone array," in *Proc. of ICASSP 2000*, Istanbul, Turkey, June 2000, vol. 2, pp. 1053–1056.
- [5] J. M. Peterson and C. Kyriakakis, "Hybrid algorithm for robust, real-time source localization in reverberant environments," in *Proc. of ICASSP 2005*, Philadelphia, PA, Mar. 2005, vol. 4, pp. 1053–1056.
- [6] H. F. Silverman, Y. Yu, J. M. Sachar, and W. R. Patterson, "Performance of real-time source-location estimators for a large-aperture microphone array," *IEEE Trans. Speech, Audio Process.*, vol. 4, no. 13, pp. 593–606, July 2005.
- [7] S. Xu, M. Bugallo, and P. Djuric, "Maneuvering target tracking with simplified cost reference particle filters," in *Proc. of ICASSP 2006*, Toulouse, France, May 2006, vol. 4, pp. 937–940.
- [8] D. E. Sturim, H. F. Silverman, and M. S. Brandstein, "Tracking multiple talkers using microphone-array measurements," in *Proc. of ICASSP 1997*, Munich, Germany, Apr. 1997, vol. 1, pp. 371–374.
- [9] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, New York: John Wiley and Sons, 1990.
- [10] H. Do, H. F. Silverman, and Y. Yu, "A real-time srp-phat source location implementation using stochastic region contraction (src) on a large-aperture microphone array," in *Proc. of ICASSP 2007*, Honolulu, Hawaii, Apr. 2007, vol. 1, pp. 121–124.
- [11] H. Do and H. F. Silverman, "A fast microphone array srp-phat source location implementation using coarse-to-fine region contraction (cfrc)," in *Proc. of WASPAA 2007*, New Paltz, NY, Oct. 2007, To appear.