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
{hdo,hfs}@lems.brown.edu

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). We test the algorithm with five simultaneous "talkers" under very difficult conditions in a real room but using electrical speakers instead of human talkers to have a controlled experiment. Results are presented and discussed.

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 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 [1, 2]. The second one uses a beamformer to find multiple peaks of an energy-based functional, such as the steered response power [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 second approach is more robust under such conditions, and is now more feasible computationally given today's more powerful processors and substantive algorithm improvements [7, 8].

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 [3, 9] or Kalman filtering [10]. 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. If of sufficient quality, these 'open loop' estimates may be used directly or as initialization for tracking systems.

In this paper, we propose a novel 'open loop' location method for multiple sources. This method uses the proven-robust steered response power using the phase transform functional (SRP-PHAT) with agglomerative clustering (AC)[11], and the low-cost global optimization algorithm, stochastic region contraction (SRC)[7]. SRP-PHAT has been used for estimating two [3] or more simultaneous sources [12] using a small circular array with no range estimates, and averaging over frames. Our method, which has been developed for a large-aperture microphone array, does range estimation, uses a single frame of data, and works for more than 2 sources.

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, 7]. A typical slice for $P_n(\vec{x})$ at a fixed height $y = constant$ 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 $\vec{\lambda}_n(k)$, $1 \leq k \leq K$.

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.

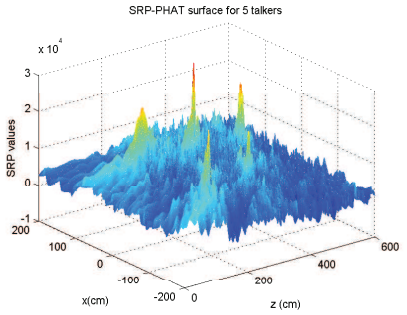

Fig. 1. SRP-PHAT 3D illustration for 5 talkers

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 $|.|$ 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, $||.||$ 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: $C^{(0)}(k), k = 1, .., N$. 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 $L =$*'average'*:
   $$d_{mn}^{(i)} = \underset{u,v}{\text{mean}} ||\vec{p}_m^{(i)}(u) - \vec{p}_n^{(i)}(v)|| \forall m, n$$
   IF $L =$*'simple'*:
   $$d_{mn}^{(i)} = \min_{u,v} ||\vec{p}_m^{(i)}(u) - \vec{p}_n^{(i)}(v)|| \forall m, n$$
   IF $L =$*'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 $N << R$ points.

3. **Cluster:** N points into P_0 clusters using AC with $L =$*'average'*.

4. **Determine:** P_0 centroids: $\vec{c}_j \equiv \text{mean}(\vec{p}_j(u))$, for all $\vec{p}_j(u) \in C(j), j = 1, ..., P_0$

5. **Calculate:** $P_0 \times P_0$ 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 $|C(j)| = 0$. Achieve $P_1 \leq P_0$ clusters.

7. **Apply SRC:** on each $C(k), 1 \leq k \leq P_1$, to achieve location estimates $\vec{x}_k$ with SRP values E_k. These estimates form the set $P_2 \leq P_1$ as non-converging clusters are terminated early.

8. **Cluster:** P_2 estimates using AC with $L =$*'simple'*. Achieve $P_3 \leq P_2$ clusters.

9. **Select:** The highest energy point $\vec{x}_k^*$ in each cluster of the P_3 clusters. Keep only those for which $E_{\vec{x}_k^*} \geq L_{\text{conf}}$, where L_{conf} is a sensible SRP-PHAT threshold value.

10. **Final output:** the set $P_4 = \{\vec{x}_k^*, E_{\vec{x}_k^*}\}$, $P_4 \leq P_3$.

Notes:

- $|.|$ 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 used in Step 7 are: $J_0 = 1000, n = 100$ [7].

We select the Mahalanobis distance because the clusters of high energy points appear to be spreading in an elliptical shape. Their principle axes (eigenvectors of the covariance matrices of the data) are along the direct paths from the sources to the microphones that we use. In Figure 2, we illustrate the clusters for 5 talkers using a 2D-plot in which height in the room has been projected onto a plane. The Mahalanobis distance

describes the correlation among data points in the clusters better than the 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 before using Mahalanobis distance. AC with the Euclidean distances shown in Fig. 2 (Step 3) provides the efficient preliminary clustering.

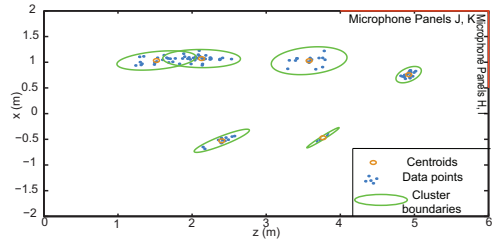

Fig. 2. Clusters for 5 talkers shown in 2D in which height(y) dimension has been projected onto a plane.

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

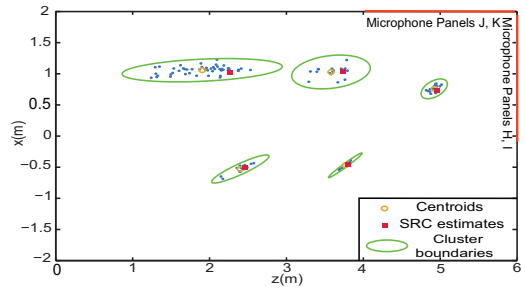

Fig. 3. Clusters after merging using the Mahalanobis distance (Step 6) and applying SRC(Step 7).

4. EXPERIMENTS

4.1. Experimental conditions

The system, room with a $T_{60} = 0.45s$, and a focal volume, V_{room} =4m $\times$ 1m $\times$ 6m that we used in our experiments has been described in [6]. Ten-second recordings (wav files) of five native American English talkers (1 female and 4 males) were individually recorded with close-talking microphones. For the testing, these individual recordings were played simultaneously by Adobe Audition through five Advent AV009 speakers, each approximately facing the 24 locator microphones as shown in Figure 4 with the average distances and SNR's indicated.

Note that the SNR's are for background noise only and just indicate the difficulty of the environment. In reality, the direct signal is also corrupted by intense reverberation and interference from other talkers and their reverberation as well. Frames of 102.4ms, advancing each 25.6ms, and a sampling

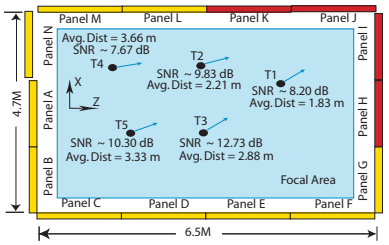

Fig. 4. Top view of the array, showing source locations and panels. This experiment used 24 microphones of the 128 on panels H, I, J, K. The arrows indicate the orientation of the talkers and the SNR's are for background noise only.

rate of 20 KHz were the conditions for testing. A location 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.

4.2. Determination of parameters

A preliminary experiment 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 [7], $R = 15000$ will err by missing the peak volume less than 0.1% of the time. Also, an $N = 500$ gave a sufficient number of data points for clusters at source locations, and eliminated a several "noise" outliers. In Fig. 5, we show the spread of points for the full grid-search (a) and for various choices of R and N (b $\rightarrow$ i). A value of $R < 15000$ (Fig. 5b–c) resulted in a spread of data points over a large area, which means very poor cluster information. A value of $N < 500$ (Fig. 5f–g) could lead to the case of missing data points (for T3). Hence, $R = 15000$ and $N = 500$ were the minimum values to provide useful cluster information when compared to the 2D energy map given by the SRP-PHAT straight grid-search.

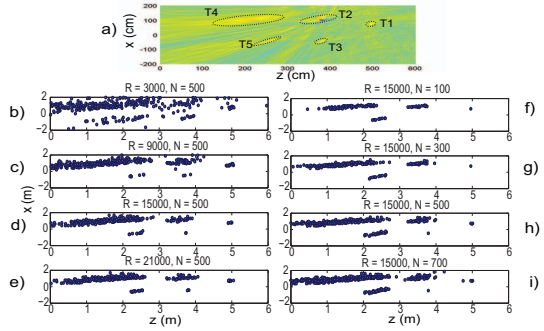

Fig. 5. 2D energy map given by grid-search (a) and for different values of R and N for the 5 talker case (b $\rightarrow$ i)

4.3. Experiments

To set a "truth" baseline, we computed the energy in each frame for the individual-talker, clean recordings. The histograms are shown in Fig. 6. Silence frames were all at about 67dB, and the loudest frames were 45dB above silence. We expected talker-frames having relatively low energy to be the most difficult for the algorithm, so we were able to use these

energy-data values to discriminate talker-frames by selecting a threshold, Θ, on absolute energy.

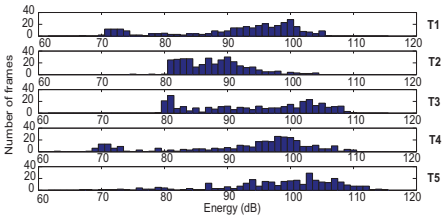

Fig. 6. Absolute energy (dB) histogram of 5 talkers over 300 frames.

As the experiment, we compared the estimates given by the algorithm on the **composite data** against the "truth" baseline with various settings of Θ. We also used the SRP-PHAT functional values as the confidence level, L_{conf}. An estimate was made only when its SRP value was greater than or equal to the preset L_{conf}. The variation in the number of estimates made as a function of conf is shown in Fig. 7. Clearly, we expect the percentage of correct estimates to increase as L_{conf} is larger.

In every frame, an estimate was counted as "correct" if it matched a "truth" baseline location, and as "extra" if it did not. A "missed estimate" is counted if there was any talker in the baseline not detected by the algorithm. Fig. 8 shows these statistics for different values of Θ at $L_{\text{conf}} = 6$, and Fig. 9 shows them for $L_{\text{conf}} = 8$. As the confidence level, L_{conf}, decreased, more estimates were made and, hence, both the percent correct and extra increased while the percent missed decreased. On the other hand, when L_{conf} was set higher, fewer estimates were compared, and these expectations are observed in the measured performance. Also, when the energy level of the speech was high ($\geq$ 100dB), the algorithm performed better.

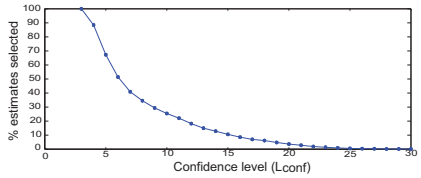

Fig. 7. Percent of estimates made vs. L_{conf}

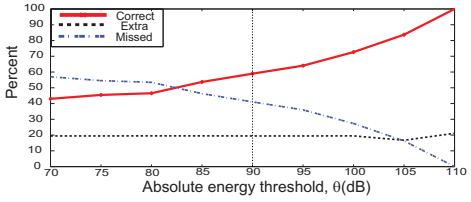

Fig. 8. Percent correct, extra, and missed for different values of Θ and $L_{\text{conf}} = 6$ over all frames. A performance line at 90dB is highlighted.

5. CONCLUSION

We have presented an open-loop, simultaneous-talker, location-estimation method using SRP-PHAT with SRC and AC. We have tested the new method under extremely adverse, real conditions of reverberation, background noise and, especially, loud, interfering sources. The experiments show, given we count only talkers in frames having energy 23dB above the background using the close-talking, original clean-speech data, over 60% of the talkers are found correctly, and about 20 % "extra" estimates are made (see Fig. 8). While there is still much to be done, we believe the performance at this level would work very well as input to a complete tracking algorithm.

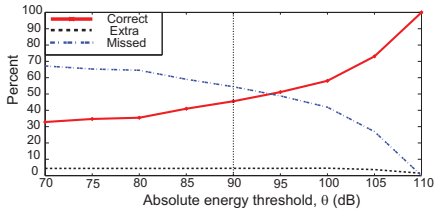

Fig. 9. Percent correct, extra, and missed for $L_{\text{conf}} = 8$.

6. REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[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] 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.

[8] J. Dmochowski, J. Benesty, and S. Affes, "A generalized steered response power method for computationally viable source localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2510–2526, Nov. 2007.

[9] 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.

[10] 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.

[11] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, New York: John Wiley and Sons, 1990.

[12] G. Lathoud and M. Magimai-Doss, "A sector-based, frequency-domain approach to detection and localization of multiple speakers," in *Proc. of ICASSP 2005*, Philadelphia, PA, Mar. 2005, vol. 3, pp. 265–268.