Improved TDOA Disambiguation Techniques for Sound Source Localization in Reverberant Environments

Cecilia Maria Zannini, Albenzio Cirillo, Raffaele Parisi, Aurelio Uncini
INFOCOM Dept.
University of Rome "La Sapienza"
Via Eudossiana,18
00184 Rome, Italy
Email: cmzannini@infocom.uniroma1.it

***Abstract*—Given a single sound source in a non reverberant environment, an estimate of the *Time Difference Of Arrival* (TDOA) between microphones can be obtained by observing the time value at which the cross correlation of the two microphone signals displays a maximum. In the presence of reverberation, however, the cross correlation function displays a great number of local maxima caused by reflected signals blending unpredictably. This results in TDOA estimation ambiguity, making correct source localization impossile. The aim of this work is to present an improved disambiguation algorithm and compare it to existing algorithms in terms of estimation accuracy and processing time.**

***Index Terms*—Time Difference of Arrival Estimation, Sound Source Localization, Reverberation**

I. Introduction

The ambiguity inherent in TDOA estimation in reverberant environments through *Generalized Cross Correlation* (GCC) [1] maxima search can be attributed to the formation of an unknown number of signal paths from source to microphones. Each reflected sound wave can be modeled as a new source which not only determines its maximum in the GCC function based on its position with respect to the microphones, but also interacts unpredictably with every other reflected signal present in the room. In order to highlight the room's effect on the emergence of this ambiguity, the following time-discrete signal model can be used:

$$x_k(n) = h_k(n) * s(n) \tag{1}$$

where $x_k(n)$, the output of the k-th microphone, is the result of the convolution integral of the source $s(n)$ and the room's impulse response $h_k(n)$ from the source to the k-th microphone. The estimated *Generalized Cross Correlation Phase Transform* (GCC-PHAT) of the signals from microphones k and l will therefore be [1]:

$$\hat{R}_{kl}(\tau) = \int_{-\infty}^{\infty} \frac{\hat{G}_{kl}(f)}{|G_{kl}(f)|} e^{j2\pi f\tau}\, \mathrm{d}f \tag{2}$$

where $\hat{G}_{kl}$ is the estimated cross power spectral density of microphones k and l. The selection of one GCC-PHAT maximum will yield an estimated TDOA. The geometrical relationship between the intra-microphone distance, the TDOA, and the speed of sound will provide an estimated direction of arrival of the sound source with respect to the microphones. This relationship can be summarized by the following formula [2]:

$$\theta = \frac{c \cdot TDOA}{d} \tag{3}$$

where c is the speed of sound and d is the distance between the two microphones to which the TDOA refers.

The GCC-PHAT of two microphone signals placed in a reverberant room displays a number of peaks increasing with the reverberation time. In order to make source localization possible, it is necessary to develop a set of *a priori* criteria to single out the peak occuring at the moment most likely corresponding to the TDOA between the direct paths from source to microphones.

II. Algorithms

The starting point of our research is an algorithm which proposes to solve this ambiguity by the *Disambiguation of TDOA Estimates in Multi-path Multi-Source Environments* (DATEMM) algorithm [3]. In a best case scenario, this algorithm returns an ordered graph structure containing the best estimate of the time delay relationships between all microphones and all sources in a reverberant room. Given a one-source multi-microphone setup, each node in the graph represents a microphone and each branch connecting two nodes will be labeled with the TDOA between the two respective microphones [4].

The TDOA estimate most likely associated to the direct path between source and microphone is obtained by selecting the GCC-PHAT maximum possessing the highest *quality function* value. The quality function value of each maximum is set equal to the maximum's value.

Once each maximum of each GCC-PHAT function has been ranked according to its quality value, all possible subgraphs of three microphones are assembled. Each of these microphone-triplets is the result of a specific combination of peaks from

978-1-4244-5309-2/10/$26.00 ©2010 IEEE

three different GCC-PHAT functions. Given that the sum of actual TDOAs within a closed graph must equal zero for geometrical reasons [3], only a subset of all microphone triplets, defined as *consistent* [4], needs to be considered. Furthermore, the quality function definition can be extended to the whole triplet. The triplet quality takes into account both the quality of the single peaks composing the triplet as well as how closely the consistency condition is met.

The construction of the set of complete graphs, each containing all microphones and all TDOAs, begins with the selection of the highest quality consistent triplet available. A fourth microphone is added by selecting a consistent triplet which has two branches in common with the starting triplet. Adding a fourth microphone via consistent triplets does not guarantee that the resulting 4-microphone-graph is consistent; this method however reduces the number of consistency checks needing to be performed [5], [6].

The sequential assembly of subgraphs will yield a set of completely connected graphs which can be ordered according to more than one factor. A series of descriptive parameters can be calculated for these final graphs: total quality, resulting in the sum of the quality of all triplets used to build the graph, a measure of how closely the zero cyclic sum condition is met by the graph as a whole, and total cyclic sum value. Ideally one and only one of the synthesized completed graphs contains direct path TDOAs exclusively.

The first variation of this algorithm is what we have named *No Raster Matching DATEMM* (NRM-DATEMM), in accordance with the terms introduced in [3]. The first N peaks from the GCC-PHAT are selected. All consistent triplets are formed using these peaks. Subsequent microphones are added recursively until a completely connected graph is achieved. Repeated frame processing can give a measure of the cost/benefit ratio of this course of action with regards to the capability of the algorithm to reject non direct path TDOAs.

A second proposed method, named *Simplified DATEMM* (S-DATEMM), aims at attaining a fixed processing time, while maintaining an acceptable level of correct direct path TDOA estimate selection. Compared to the previous algorithms, a smaller number of GCC-PHAT maxima are selected. All possible combinations of these peaks are used to construct all attainable complete graphs regardless of their consistency. A total quality value for each of these graphs is defined as the sum of the component triplets quality function values. The highest quality value graphs are most likely to contain the highest number of direct path TDOAs.

III. Proposed Methods

Five omnidirectional microphones and one sound source were placed in a simulated 4m $\times$ 2m $\times$ 4m room with variable reverb times at the coordinates indicated in Table I. The sound source was a Gaussian noise signal for the first run and a male voice speech sample for the second run. Each microphone-source impulsive response was calculated using the image method [7]. Processing was performed on 8192 sample frames using a 44100Hz sampling frequency. Each microphone signal

TABLE I
Microphone and Source Coordinates

Element	x [m]	y [m]	z [m]
Source	1.86	0.75	1.68
Microphone 1	3.71	0.49	1.59
Microphone 2	3.69	1.21	1.57
Microphone 3	2.61	0.84	0.17
Microphone 4	2.72	0.65	1.25
Microphone 5	1.85	1.63	1.25

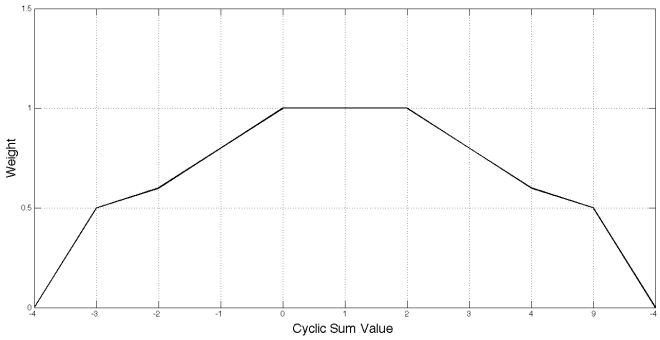

Fig. 1. Zero Cyclic Sum Tolerance Weighing Function

resulting from the convolution of the source with the source-microphone impulse response was cropped so as to include time instants within the $[-d/c, d/c]$ range, where d was the intra-microphone distance and c the speed of sound.

The performance of each algorithm was compared to that of a method which makes no attempt to compensate for the ambiguity arising from higher reverb times. This is the conventional GCC-PHAT method, which produces a TDOA estimate selecting only the highest peak value for each of the ten cross correlation functions.

A. NRM-DATEMM

Given the set of five microphones, ten GCC functions were obtained using Welch's averaged, modified periodogram method of spectral estimation [8]. This method was applied to the entire 8192-sample frame using 4096-sample subframes with a 50% overlap. This was deemed necessary due to the discrete-time nature of the sampled signals. The PHAT weighing was modified to avoid division by numbers close to zero when the signal level was low. An empirical limiting parameter named Γ_{PHAT} was introduced; its value was set to the number of samples in the processing frame multiplied by 10^{-3}, as in [5]. The weighing function used was the maximum between Γ_{PHAT} and the absolute value of the cross power spectrum density estimate. An inverse FFT yielded the GCC-modified-PHAT functions in the time domain.

A maximum of fifteen peaks were selected from each GCC-PHAT function. The quality function of each of these peaks was set equal to the maxima's value. All possible 3-microphone-subgraphs were assembled combining all avail-

able peaks. A consistency check was performed on each triplet, discarding those whose absolute value of cyclic sum was greater than four samples. The tolerance function in Fig. 1 was used during the consistency checks in order to measure how precisely each triplet fulfilled the zero cyclic sum condition. This function provided a weighing factor used to scale down the triplets quality value proportionally to the value of its cyclic sum. The final quality function value for each consistent triplet was defined as the sum of the component peaks quality function value weighed by the cyclic sum function.

The triplet with the highest quality value was chosen as the starting block for the synthesis of the final complete graphs. A fourth microphone was added to the starting triplet by selecting all consistent triplets with two branches overlapping the initial triplet. The fifth microphone was added to the initial triplet in the same way. This resulted in 2 separate sets of 4-microphone-fully-connected-subgraphs, one containing the initial triplet plus the fourth microphone, and the other containing the initial triplet plus the fifth microphone. Merging each couple of 4-microphone-subgraph produced 5-microphone-graphs missing a fourth and fifth connecting branch. The final step in the algorithm was reached by selecting a consistent triplet connecting the fourth and fifth microphones and having the remaining two branches in common with the graph they were completing.

Three different descriptive parameters were calculated for all resulting graphs. The first parameter, named *total cyclic sum*, was calculated by adding the cyclic sums of all ten component triplets. The second parameter was obtained by taking the total cyclic sum and using it to calculate the corresponding cyclic sum tolerance function value. This parameter was therefore a *total cyclic sum weight*. The last parameter, named *total graph quality*, was defined as the sum of the quality value of each triplet present in the graph.

B. S-DATEMM

The S-DATEMM algorithm required the selection of the first three maxima of each GCC-PHAT function. All combinations of peaks were used to assemble a total of 3^{10} 5-microphone-graphs without any consistency check that could discard triplets. A *total quality parameter* was calculated for each of the 59049 graphs following the definition given for the same parameter in the previous two algorithms. According to the algorithms final step, the graphs with the highest total quality function were selected.

The selection of the highest valued GCC-PHAT maximum as best TDOA estimate was taken as the touchstone of algorithm performance. This method yields one estimate per microphone pair, thus producing a single graph for each processed signal frame.

C. Error Definition

According to the simulated setup used, each complete graph yielded ten TDOA estimates. Exact TDOA values relative to the direct path between microphones and source were calculated and used as reference against which to compare estimated values. Each time an estimated TDOA was not within 20 samples of the exact TDOA value, the estimate was considered unacceptable. Each algorithm's *best performance graph* was the one possessing the lowest number of unacceptable TDOA estimates. The number of unacceptable TDOA estimates returned by the best performance graph of each algorithm for each frame processed was stored and averaged over 36 signal frames.

All algorithms but the reference one returned more than one final complete graph. In these cases, the graphs with the least number of unacceptable TDOA estimates were selected to represent the algorithms performance; this assured a best performance comparison with respect to the number of viable TDOA estimates.

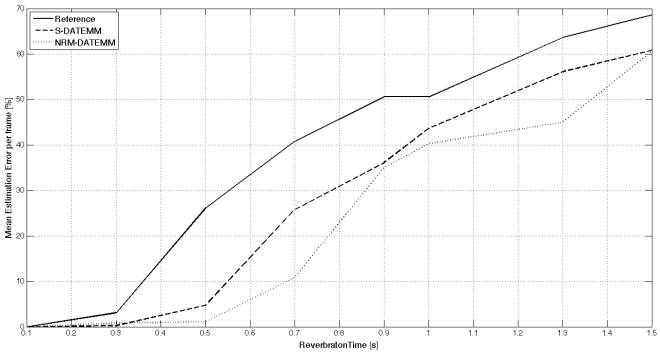

Fig. 2. Mean Estimation Error of all algorithms calculated for a run of 36 frames with Gaussian noise source

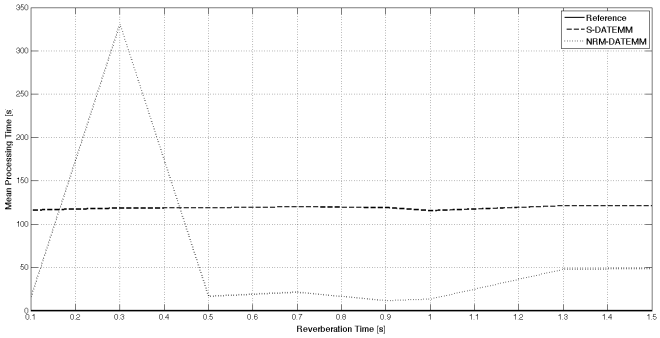

Fig. 3. Mean Processing time for all algorithms calculated for a run of 36 frames with Gaussian noise source

IV. Results

A. Gaussian noise source

The algorithm performance with respect to mean TDOA estimate error using the Gaussian noise source is illustrated in Fig.2. The NRM-DATEMM shows the smallest mean error of all algorithms for reverberation times of 0.5s or higher. This improvement in estimation precision with respect to both the reference algorithm as well as the S-DATEMM corresponds to an increase in processing time. As can be seen in Fig.3, the NRM-DATEMM algorithm requires less time than S-DATEMM to execute apart from an anomaly at 0.3s reverb time.

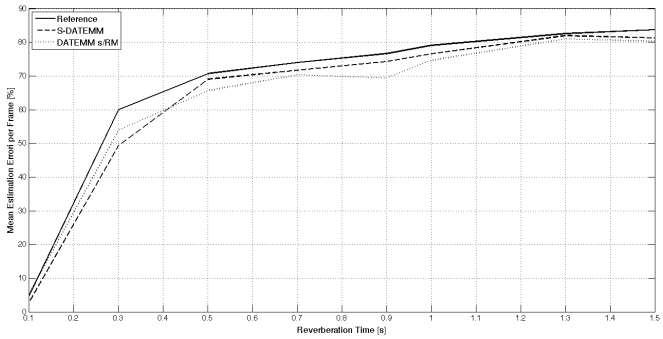

Fig. 4. Mean Estimation Error of all algorithms calculated for a run of 36 frames with male voice speech source

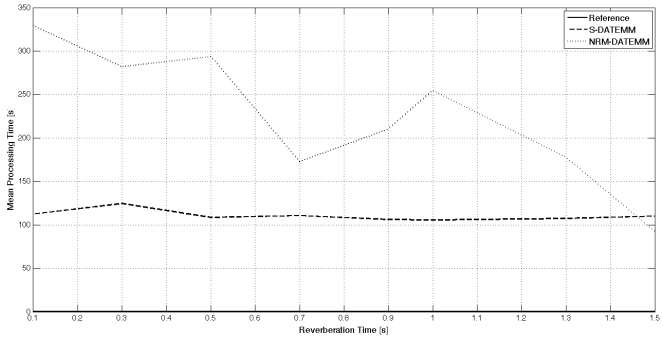

Fig. 5. Mean Processing time for all algorithms calculated for a run of 36 frames with male voice speech source

B. Speech source

Switching to a speech source does not alter the relative performance of the algorithms in terms of TDOA errors; however, the improvement in terms of mean error percentage displayed by the NRM-DATEMM is less striking as can be seen in Fig. 4. A possible explanation for this behavior is that the pseudo-periodic nature of the vocal signal introduces additional ambiguities which the algorithms cannot solve completely. With this source, S-DATEMM mean processing time is very low compared to that of NRM-DATEMM (ref. Fig. 5). This is probably caused by the greater number of peaks present in the GCC-PHAT of a speech signal which generates a greater number of consistent triplets. This in turn leads to a greater number of completely connected graphs which need to be ranked.

V. Conclusion

The results obtained by our simulations have shown that the source type changes the performance of the proposed algorithms. Significant improvement in TDOA estimation quality with respect to the reference algorithm was obtained for both algorithms. If the sound source is Gaussian noise, NRM-DATEMM delivers the least number of wrong TDOA estimates in the least time for reverberation times of 0.5s or higher. For a speech signal this is no longer true; while NRM-DATEMM still displays the smallest mean estimation error, it is not the fastest. Furthermore, the difference in mean error percentage between the algorithms is small; for this reason the S-DATEMM algorithm is a better choice if processing speed is a key factor. Computational complexity of the S-DATEMM algorithm as well as the cost of the *raster matching* step present in the original DATEMM algorithm are determined by the number of microphones and extracted cross correlation maxima. On the other hand, computational costs of both DATEMM and NRM-DATEMM are ruled by the number of consistent triplets found in each processed frame. This number depends on the presence of non direct source microphone paths, which in turn vary with reverb time and microphone placement. This observation implies that NRM-DATEMM performance can be improved by decreasing the number of consistent triplets available for graph assembly. As reverb time increases, however, ambiguity caused by indirect source-microphone paths forces the analysis of a greater number of cross correlation peaks. In order to limit the risk of missed detection, the number of triplets to be labelled *consistent*, determined by the tolerance function shape, should increase with reverb time. The optimal compromise between tolerance function shape and algorithm performance can be addressed in future work. The three suggested parameters used to order the final graphs failed to provide a unique best algorithm selection criterium. A final graph selection technique capable of isolating the single least error ridden final graph amongst those provided by our proposed algorithm in a non experimental scenario where exact TDOA values are unknown is the next logical step in our research.

References

[1] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 24, no. 4, pp. 320–327, 1976.

[2] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *Speech and Audio Processing, IEEE Transactions on*, vol. 5, no. 3, pp. 288–292, May 1997.

[3] J. Scheuing and B. Yang, "Disambiguation of tdoa estimates in multi-path multi-source environments (datemm)," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 4, May 2006, pp. IV–IV.

[4] ——, "Efficient synthesis of approximately consistent graphs for acoustic multi-source localization," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, April 2007, pp. IV–501–IV–504.

[5] ——, *Speech and Audio Processing in Adverse Environments*. Hänsler, Eberhard and Schmidt, Gerhard, Springer Publishing Company, Incorporated, 2008, ch. Correlation-Based TDOA-Estimation for Multiple Sources in Reverberant Environments, pp. 381–416.

[6] ——, "Disambiguation of tdoa estimation for multiple sources in reverberant environments," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 8, pp. 1479–1489, Nov. 2008.

[7] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Acoustical Society of America Journal*, vol. 65, pp. 943–950, Apr. 1979.

[8] P. Welch, "The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *Audio and Electroacoustics, IEEE Transactions on*, vol. 15, no. 2, pp. 70–73, Jun 1967.