

Robust probabilistic TDOA estimation in reverberant environments

Steven J. Rennie

PSI-TR-2005-011, February 2005 *

Abstract

In this paper, a novel Expectation-Maximization algorithm for estimating the time-delay-of-arrival of multiple non-stationary sound sources in non-stationary reverberative acoustic environments, is presented. Motivated by the success of the phase-transform/histogram based approaches of Aarabi [2, 3], the algorithm operates by learning a probabilistic relationship between the latent TDOAs and the observed microphone phase over a small collection of short-time DFTs, and in the process automatically estimates the TDOA posterior over the collection of DFTs, of each individual DFT, and also provides a measure of the frequency content of each sound source. Experimental results demonstrate that the algorithm performs as well as the Histogram techniques of Aarabi [2, 3], which have demonstrated until now unmatched results for the problem of acoustic TDOA estimation in natural environments. The model is generative and parametric and thus can potentially be seamlessly fused with probabilistic descriptions of speech production and mixing such as defined in [10], to achieve enhanced speech separation capability.

1 Introduction

The problem of estimating the time-differences of arrival (TDOAs) of one or more underlying sources at a pair of transduction points is a fundamental one, lying at the heart of many problems in the areas of communication, control, and tracking, to name but a few.

Here we consider one of the most challenging manifestations of the TDOA estimation problem; the problem of recovering the TDOAs of multiple, potentially simultaneously active, non-stationary, acoustic sources, situated in a highly, non-stationary, reverberative environment. This is the general acoustic source TDOA estimation problem in its most general, and most typical form.

*based upon a course project completed in December 2003

The acoustic source TDOA estimation problem is fundamentally difficult and open one, and represents a severe bottleneck to be overcome in the realization of robust sound localization engines, which, in addition to applications where the position of one or more sound sources (speakers) is directly desired, are a critical component of a growing family of speech enhancement and separation techniques that utilize information about speaker position to achieve enhanced performance [10, 12, 1, 9, 11, 4].

So, what makes this problem so difficult to solve? The first major difficulty is environmental acoustic reflection. For a given sound source at a given position and orientation, the relationship between the source signal and an transduction of it are approximately related by a non-trivial linear convolution. The further away, the transduction point, the higher the relative intensity of the multipath. Unfortunately the relating convolution is a highly non-stationary function of the relative position and orientation of the sound source (Brandstien has shown that changes in source position of less than 2 inches, and orientation of less than 10 degrees, affect the dominant peaks of the relating convolution dramatically [7]) and so cannot be reasonably calibrated or learned.

Another major difficulty is that multiple sound sources are generally simultaneously active, making the effective noise level in the estimation of the TDOA of a given source extremely high. In addition, most sound sources of interest (speakers) are generally spatially dynamic, and have non-stationary output (in particular, they are not always speaking), and so short analysis windows must be used, and the TDOA of a given source can generally not be 'tracked'. A final major difficulty is that generally little or no information about the configuration of the sound sources of any kind is available.

The call for a reverberation tolerant, multi-source capable TDOA estimation algorithm that is applicable given the above difficulties is a tall one, but progress has been made. Aarabi's phase-transform/histogram-based TDOA estimation algorithm embodies both past historical advances [8, 4] and the current state of the art [2, 3, 6, 5] in TDOA estimation, and will be described briefly here.

Aarabi's technique is foundationed on applying the phase transform in the frequency domain over short (e.g. 20ms) segments of the microphone readings, to obtain a point estimate of the underlying TDOA of each frame, and then histogramming the point estimates over a window of segments (eg. 20) to obtain an estimate of the posterior over TDOA space.

The phase transform-based estimate of the TDOA of a given segment is given by:

$$\tau^* = \underset{\tau}{\operatorname{argmax}} \sum_w \cos(w\tau - \phi_w) \quad (1)$$

where ϕ_w is the phase difference between the DFT of the two microphone observations at frequency w , and the maximum is generally taken over a discrete set $\{\tau\}$. In reverberative environments, the simple phase transform is the best known estimator of the MAP TDOA of a segment of speech, operating in the frequency

domain to make the dominant TDOA more discernable, and giving each frequency equal weight. The phase transform in reverberative environments actually does *better* than a maximum likelihood weighted version of the above, even when the SNR at each frequency is (artificially assumed) to be known [2].

Histogramming over a short window of (active) segments, an estimate of the posterior over TDOA space is obtained. The key here is the utilization of short segments to take advantage of diversity in the dominance of the underlying sound sources over time to identify the underlying TDOAs, and then histogram the result. This algorithm seems exceedingly simple, but it is the best TDOA estimation engine developed to date, and it operates at real-time speed. Note that it is the recovery of a probability distribution over TDOA space that makes the technique so powerful. Because environmental conditions are generally so severe that a consistently good point estimate of a given source TDOA is not possible, it is the fusability of the obtained TDOA estimations over an array of microphones that is critical to obtaining a good estimates of the location of all underlying sources. In particular, a fusion of greedy TDOA estimates from a given set of microphone pairs will not generally be consistent.

In this paper, a novel Expectation-Maximization algorithm for estimating the time-delay-of-arrival of multiple non-stationary sound sources in non-stationary reverberative acoustic environments, is presented. Motivated by the success of the phase-transform/histogram based approaches of Aarabi [2, 3], the algorithm operates by learning a probabilistic relationship between the latent TDOAs and the observed microphone phase over a small collection of short-time DFTs, and in the process automatically estimates the TDOA posterior over the collection of DFTs, of each individual DFT, and also provides a measure of the frequency content of each sound source. Experimental results demonstrate that the algorithm performs as well as the Histogram techniques of Aarabi [2, 3], which have demonstrated until now unmatched results for the problem of acoustic TDOA estimation in natural environments. The model is generative and parametric and thus can potentially be seamlessly fused with probabilistic descriptions of speech production and mixing such as defined in [10], to achieve enhanced speech separation capability.

2 An Expectation-Maximization Algorithm for TDOA estimation

In this section we utilize the ideas of the Aarabi's TDOA estimation algorithm and formulate a generative probability model that relates the latent TDOA space to an observed collection (window) of phase difference vectors $\{\phi\}$ defined in the frequency domain.

Here we model the TDOA space as discrete (as Aarabi did), and model the generation of each observed vector ϕ , as a two stage process:

- 1) The selection of a TDOA from the distribution $p(\tau) = \pi_\tau$

2) Conditioned on τ sampling from the distribution $p(\phi|\tau) = N(\mathbf{w}\tau, \mathbf{\Sigma})$

where $\mathbf{\Sigma}$ is taken to be diagonal, and so the probability of the data $p(\phi)$ is being modeled by a mixture of diagonal covariance gaussians. The joint probability of the observed collection of phase difference vectors under this model is given by:

$$p(\{\phi\}) = \prod_t \sum_{\tau} \pi_{\tau} N(\mathbf{w}\tau, \mathbf{\Sigma}) \quad (2)$$

And so the estimation of the posterior of τ is transformed into a maximum likelihood parameter estimation problem in π_{τ} and $\mathbf{\Sigma}$.

The following EM updates may be iterated until convergence to a fixed point, to identify π_{τ} , $\mathbf{\Sigma}$:

E Step: Estimate the posterior of τ for each phase difference vector based on the current settings of π_{τ} and $\mathbf{\Sigma}$:

$$p(\tau_t|\phi_t)' = \frac{\pi_{\tau} N(\mathbf{w}\tau, \mathbf{\Sigma})}{\sum_{\tau_t'} \pi_{\tau} N(\mathbf{w}\tau, \mathbf{\Sigma})} \quad (3)$$

M Step: Update the parameter estimates, based on new posterior for τ :

$$\pi_{\tau} = \frac{1}{T} \sum_t p(\tau_t|\phi_t)' \quad (4)$$

$$\mathbf{\Sigma}_{w,w} = \frac{\sum_t p(\tau_t|\phi_t)' (w\tau - \phi_{w,t})^2}{\sum_t p(\tau_t|\phi_t)'} \quad (5)$$

Note that the difference operator in the last equation must calculate the unwrapped difference between $w\tau$ and $\phi_{w,t}$.

The estimation process, once converged, yields a posterior for τ over the window, a posterior for τ for each frame, and the conditional variability of ϕ given τ , $\mathbf{\Sigma}$, which can be used as an indicator of the frequency activity of a postulated sound source with TDOA τ . The algorithm can be viewed as a probabilistic version of Aarabi's histogramming technique, where instead a *soft* estimate of the TDOA of each segment is taken into consideration in forming an estimate of the posterior of τ over the window.

3 Experimental Results

It is not immediately clear how to assess the performance of given TDOA algorithm in isolation under severe conditions, because, as discussed in the introduction, it is the fusability of the TDOA information provided by a given algorithm that is most fundamental, as the estimation conditions are generally very severe. One

could quantify the fidelity of the posterior estimate of τ via the Kullback-Liebler divergence of the estimate from the true posterior, but in this case, this is not a feasible measure of fusability; as the true posterior is zero almost everywhere. Instead we will quantify fusability by comparing the average posterior for τ over all estimation windows, to the true posterior for stationary sound sources.

Figure 1 depicts a plot of the average posterior estimate of τ over 500 estimation windows, for both Aarabi’s histogram method, and the EM algorithm developed here, for the case of 1 stationary speaker situated at approximately 1.3m from the center of a microphone pair with intra-distance 0.2m, in a reverberent room (reverberation time approx. 200ms), at a delay of -8 samples. Here we can see that both estimates of the posterior have large probability alias at -3 samples, and the histogram approach has large probability alias at the endpoints of the TDOA range.

Figure 2 depicts a plot of the average posterior estimate of τ over 500 estimation windows, for both Aarabi’s histogram method, and the EM algorithm developed here, for the case of 2 stationary speakers situated at approximately 1.3m from the center of a microphone pair with intra-distance 0.2m, with each microphone corrupted by 0dB IID gaussian noise, with the speakers at TDOAs of 4 and -3 samples. Here we can see that with no reverberative noise corruption, both techniques are able to recover good estimates of the true TDOA density.

Figure 3 depicts a plot of the average posterior estimate of τ over 500 estimation windows, for both Aarabi’s histogram method, and the EM algorithm developed here, for the case of 2 continuously speaking, stationary speakers situated at approximately 1.3m from the center of a microphone pair with intra-distance 0.2m, in a reverberent room (reverberation time approx. 200ms), with the speakers at TDOAs of 4 and -3 samples. Here we can see that both TDOA estimation algorithms perform very poorly, and the cumulative multipath of both sources at delay -1 samples is dominating the estimate of the posterior under both methodologies. Obviously this information is no longer useful for fusion.

The results show that the derived EM algorithm for TDOA estimation performs on par with Aarabi’s histogram- based approach, which as previously mentioned, defines the current state-of-the-art with regard to TDOA density estimation for subsequent information fusion. As mentioned previously, the EM algorithm developed here additionally provides a posterior for τ for each processing frame, and also a measure of the frequency content of each source over the processing window. The model is generative and parametric and thus can potentially be seamlessly fused with probabilistic descriptions of speech production and mixing such as defined in [10], to achieve enhanced speech separation capability. Note that in these experiments the speakers were spatially stationary, so as to make averaging the posterior over all windows of algorithm application meaningful. In all the presented experiments, the processing window was defined over 20 20ms processing frames. Both algorithms are hence applicable to the estimation of the TDOAs of non-stationary sound sources whose spatial evolution over this length of processing window is essentially negligible.

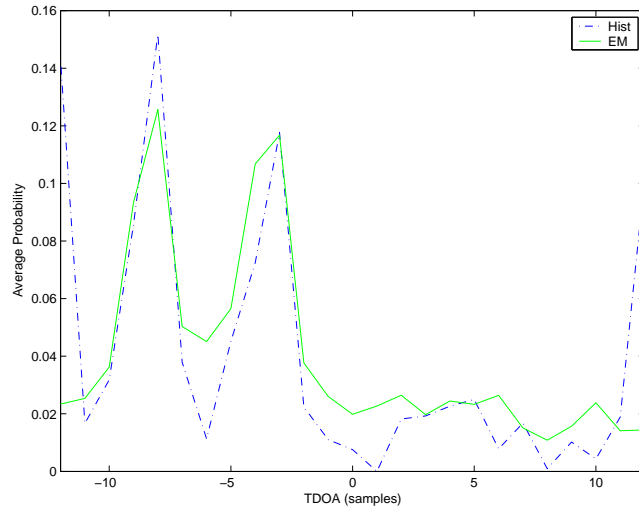


Figure 1: Plot of the average posterior estimate of τ over 500 estimation windows, for both Aarabi's histogram method, and the EM algorithm developed here, for the case of 1 stationary speaker situated at approximately 1.3m from the center of a microphone pair with intra-distance 0.2m, in a reverberent room (reverberation time approx. 200ms), at a delay of -8 samples ($F_s = 20$ kHz)

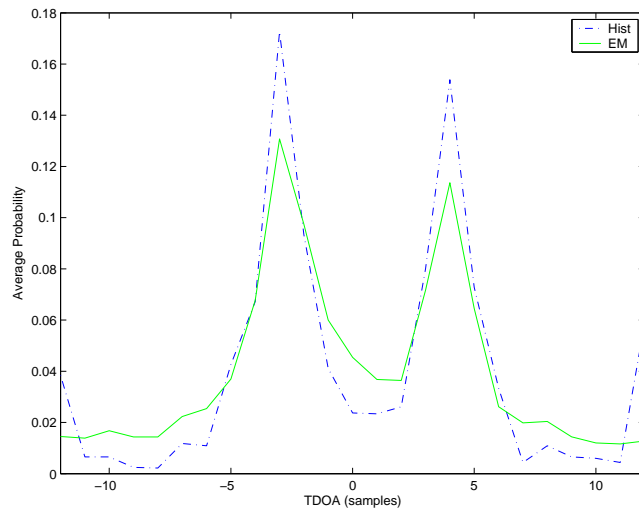


Figure 2: Plot of the average posterior estimate of τ over 500 estimation windows, for both Aarabi's histogram method, and the EM algorithm developed here, for the case of 2 stationary speakers situated at approximately 1.3m from the center of a microphone pair with intra-distance 0.2m, with each microphone corrupted by 0dB IID gaussian noise, with the speakers at TDOAs of 4 and -3 samples ($F_s = 20$ kHz)

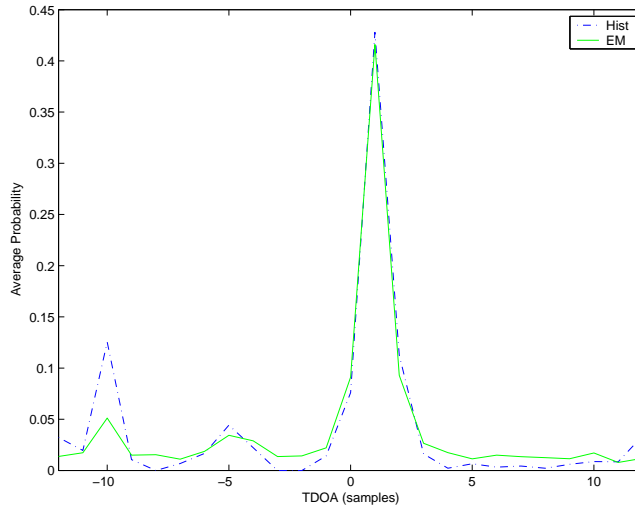


Figure 3: Plot of the average posterior estimate of τ over 500 estimation windows, for both Aarabi’s histogram method, and the EM algorithm developed here, for the case of 2 stationary speakers situated at approximately 1.3m from the center of a microphone pair with intra-distance 0.2m, in a reverberent room (reverberation time approx. 200ms), with the speakers at TDOAs of 4 and -3 samples ($F_s = 20$ kHz)

4 Concluding Remarks

In this paper, a novel Expectation-Maximization algorithm for estimating the time-delay-of-arrival of multiple non-stationary sound sources in non-stationary reverberative acoustic environments, was presented. Experimental results demonstrate that the algorithm performs as well as the histogram-based techniques of Aarabi [2], which have demonstrated until now unmatched results for the problem of acoustic TDOA estimation in natural environments. The model is generative and parametric and thus can potentially be seamlessly fused with probabilistic descriptions of speech production and mixing such as defined in [10], to achieve enhanced speech separation capability. Possible directions of future work include the modification of the probability model to model the emission of multiple sound sources in a given processing frame, and the incorporation of the presented work into a probabilistic sound localization engine.

References

- [1] P. Aarabi. The application of spatial likelihood functions to multi-camera object localization. In *Proceedings of the 5th SPIE Conference on Sensor Fusion*, April 2001.

- [2] P. Aarabi. The fusion of distributed microphone arrays for sound localization. *EURASIP Journal of Applied Signal Processing (Special Issue on Sensor Networks)*, 2003 No. 4:338:347, March 2003.
- [3] P. Aarabi and S. Zaky. Iterative spatial probability based sound localization. In *Proceedings of the 4th World Multiconference on Circuits, Systems, Computers, and Communications*, July 2000.
- [4] M. Brandstein and D. Ward. *Microphone arrays: Signal processing techniques and applications*. Springer-Verlag, 2001.
- [5] M.S. Brandstein. *A Framework for Speech Source Localization Using Sensor Arrays*. PhD thesis, Brown University, May 1995.
- [6] M.S. Brandstein and H. Silverman. A robust method for speech signal time-delay estimation in reverberant rooms. In *Proceedings of ICASSP*, May 1997.
- [7] J. DiBiase, H. Silverman, and M. Brandstein. Robust localization in reverberant rooms. *M.S. Brandstein and D.B. Ward (eds.), Microphone Arrays: Signal Processing Techniques and Applications*, 2001.
- [8] C. H. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-24(4):320–327, August 1976.
- [9] A. Pentland. Smart rooms. *Scientific American*, 274:4:68–76, April 1996.
- [10] S. Rennie, P. Aarabi, T. Kristjansson, B. Frey, and K. Achan. Robust variational speech separation using fewer microphones than speakers. In *Proceedings of the 2003 IEEE Conference on Acoustics, Speech, and Signal Processing*, April 2003.
- [11] H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura. Speech enhancement using nonlinear microphone array based on complementary beamforming. *IEICE Trans. Fundamentals*, E82-A(8), 1999.
- [12] G. Shi, P. Aarabi, and N. Ladic. Adaptive time-frequency data fusion for speech enhancement. In *Proceedings of the 6th International Conference on Information Fusion*, July 2003.