NONNEGATIVE DECONVOLUTION FOR TIME OF ARRIVAL ESTIMATION

Yuanqing Lin[1]*, Daniel D. Lee*[1]*, and Lawrence K. Saul*[2]

[1]Department of Electrical and Systems Engineering
[2]Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104

ABSTRACT

The interaural time difference (ITD) of arrival is a primary cue for acoustic sound source localization. Traditional estimation techniques for ITD based upon cross-correlation are related to maximum-likelihood estimation of a simple generative model. We generalize the time difference estimation into a deconvolution problem with nonnegativity constraints. The resulting nonnegative least squares optimization can be efficiently solved using a novel iterative algorithm with guaranteed global convergence properties. We illustrate the utility of this algorithm using simulations and experimental results from a robot platform.

1. INTRODUCTION

Estimating the interaural time difference (ITD) of a sound source is critical for determining the location of a sound source. Time-delay estimation has been used for video conferences to track the active speaker, and in surveillance applications to locate people and vehicles [1]. With the emergence of low-cost, embedded sensor networks, algorithms for efficiently estimating the time delays among acoustic sensors in noisy, reverberant environments have attracted renewed interest [2].

The earliest time delay estimation algorithms used matched filters following the development of radar and sonar arrays. Most current algorithms rely upon first calculating the cross-correlation between the received signals at the different sensors [3]. With a few sensors, various methods have been proposed to optimize the temporal cross-correlation function to be more robust to noise in the sensors. With a large array of sensors, spectral-based algorithms such as MUSIC have proven successful in discriminating the signal subspace from the noise using an eigendecomposition of the correlation matrix [4].

This submission presents an alternative approach to ITD estimation based upon nonnegative deconvolution. We begin in Section 2 with a brief review of cross-correlation based methods as a form of maximum likelihood estimation within the context of a simple generative model. In Section 3, we generalize the maximum likelihood optimization to include the effects of echoes. For effective time difference estimation, we introduce the use of nonnegativity constraints in the optimization and describe a novel iterative algorithm with guaranteed convergence properties. In Section 4, we illustrate the advantages of this representation in simulations, and in Section 5 present its performance on data taken from an experimental robot head.

2. GENERATIVE MODEL

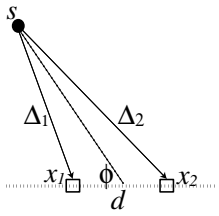

Fig. 1. Generative model for estimation of the ITD between two microphones.

In this submission, we focus on estimation of the time difference of arrival of a single acoustic source with a pair of microphones, although generalization to a larger acoustic sensing array with more sources is certainly possible. In this framework, sound originates from a source with waveform $s(t)$ and impinges on a pair of microphones as shown in Figure 1. The resulting signals measured on the two microphones are described in the time domain as:

$$x_i(t) = s(t) * h_i(t) + \eta_i(t),\ i = 1, 2 \qquad (1)$$

where $h_i(t)$ are the impulse responses between the source and microphones, and $\eta_i(t)$ are corrupting noise terms which are assumed to be spatially and temporally uncorrelated with the source.

In an ideal, free-space scenario, the impulse responses between a far-field source and the microphones are related

to the propagation delays Δ_i between the source and the microphones: $h_i(t) = \delta(t - \Delta_i)$. The azimuthal direction angle ϕ of the sound source can then be estimated from the time difference of arrival between the two microphones, $\tau = \Delta_2 - \Delta_1$, as shown in Fig. 1.

2.1. Cross-correlation based methods

By modelling only the direct-path propagation and assuming the noise terms in Eq. (1) are Gaussian distributed and white with equal variance, the maximum likelihood estimation of the propagation delays is given by the minimization:

$$\min_{\Delta_i, s(t)} \sum_i \int dt \, |x_i(t - \Delta_i) - s(t)|^2. \tag{2}$$

Solving for the source terms in Eq. (2) and transforming the optimization into the frequency domain, the estimated time difference of arrival between the two microphones τ is given by the optimization:

$$\hat{\tau} = \arg\min_{\tau} \frac{1}{2} |x_2(t) - x_1(t - \tau)|^2 \tag{3}$$

$$= \arg\max_{\tau} \sum_f X_1(f)^* X_2(f) e^{j2\pi f \tau} \tag{4}$$

where $X_i(f)$ is the Fourier transform of $x_i(t)$. Thus, the optimal maximum likelihood estimate is given by determining when the cross-correlation function:

$$C(\tau) = \sum_f X_1(f)^* X_2(f) e^{j2\pi f \tau} \tag{5}$$

obtains its maximal value. Note that by writing the parametric optimization in the frequency domain, the estimated time difference can be accurately resolved even though the Fourier estimates $X_i(f)$ are obtained from transforming discretely sampled time sequences. This is in contrast to the usual calculation of the cross-correlation in the time-domain $C[m] = \sum_n x_1[n] x_2[n + m]$ which only gives the cross-correlation values on discretely sampled time differences.

The cross-correlation in Eq. (5) can be modified by first weighting the signals $X_i(f)$ before calculating their temporal correlation. In one possible modification called the phase alignment transform (PHAT), the amplitude information in $X_i(f)$ is discarded by cross-correlating only the phases, $X_i'(f) = X_i(f)/|X_i(f)|$ [5]. In this transform, the amplitude normalization results in pseudo-wide-band signals which are then temporally correlated to estimate the time difference of arrival between the two signals. It has been shown that this normalization can help compensate for differences in the frequency responses obtained at the two microphones.

2.2. Deconvolution

In the previous section, we saw that the optimization of Eq. (3) over a single time difference variable was equivalent to maximizing the cross-correlation. Here, in order to explicitly model the multi-path reflections, we instead consider the following least squares optimization problem [6]:

$$\min_{\alpha_i} \frac{1}{2} |x_2(t) - \sum_i \alpha_i x_1(t - \tau_i)|^2 \tag{6}$$

where $\{\tau_i\}$ are a discrete set of possible time delays. The motivation for this optimization derives from adaptive filter techniques for echo cancellation [7]. In echo cancellation, a set of filter coefficients are found which best predict the resulting echoes from a given input signal. Since we expect the signal $x_2(t)$ to approximately be a time-shifted and possibly scaled version of $x_1(t)$, the minimization of Eq. (6) should yield a set of coefficients α_i which can be used to estimate the dominant time delay [8].

Eq. (6) may be regarded as a deconvolution since it decomposes the signal $x_2(t)$ as the convolution $x_1(t) * \sum_i \alpha_i \delta(t - \tau_i)$. This optimization can be rewritten as a quadratic function over the coefficients $\vec{\alpha} = \{\alpha_i\}$:

$$\min_{\alpha_i} \frac{1}{2} \sum_{ij} \alpha_i K_{ij} \alpha_j + \sum_i b_i \alpha_i \tag{7}$$

where the linear coefficients are given by $b_i = -C(\tau_i)$, with $C(\tau)$ being the cross-correlation function given in Eq. 5. The quadratic coupling terms derived from Eq. (6) are functions of $|X_1(f)|^2$; however, we can symmeterize the optimization by taking

$$K_{ij} = \frac{1}{2} \sum_f \left[|X_1(f)|^2 + |X_2(f)|^2 \right] \cos(2\pi f(\tau_j - \tau_i)). \tag{8}$$

This is equivalent to performing the minimization for the symmetric deconvolution:

$$\min_{\alpha_i} \frac{1}{2} |x_2(t) - \sum_i \alpha_i x_1(t - \tau_i)|^2 + \frac{1}{2} |x_1(t) - \sum_i \alpha_i x_2(t + \tau_i)|^2. \tag{9}$$

With no constraints on the coefficients $\vec{\alpha}$, the minimum of Eq. (7) can be solved exactly and yields: $\vec{\alpha} = -K^{-1}\vec{b}$.

3. NONNEGATIVE DECONVOLUTION

Unfortunately, the matrix K in Eq. 7 can be badly conditioned and the resulting linear solution for $\vec{\alpha}$ is very susceptible to noise. To alleviate this problem in the deconvolution, we introduce the use of nonnegativity constraints, $\alpha_i \geq 0$, in the optimization. The use of nonnegativity is physically motivated since echoes should only attenuate and

delay the signal. Nonnegativity constraints for deconvolution have also been used before in image deconvolution problems [9], as well as in learning features [10].

The optimization of Eq. 7 with nonnegativity constraints $\alpha_i \geq 0$ is convex, and thus there is a guaranteed global minimum. However, there is no known analytical solution and so we use the following iterative solution [11]. First, the matrix K is written in terms of its positive and negative components: $K = K^+ - K^-$ where $K^+_{ij} \geq 0$ and $K^-_{ij} \geq 0$ so that both K^+ and K^- are *nonnegative* matrices.

In terms of these nonnegative matrices, the estimate of $\vec{\alpha}$ is iteratively updated using the following rule:

$$\alpha_i \longleftarrow \alpha_i \left[\frac{-b_i + \sqrt{b_i^2 + 4(K^+\vec{\alpha})_i (K^-\vec{\alpha})_i}}{2(K^+\vec{\alpha})_i} \right]. \qquad (10)$$

These iterative updates are simple to implement, can be computed in real-time, and do not require the adjustment of any rate parameters that are needed for gradient-based algorithms. They prescribe a multiplicative update for the nonnegative coefficients α_i using a nonnegative factor in the right hand side of Eq. (10). A rigorous proof of global convergence for these updates can be proved using an auxiliary function [11]. The form of the updates can also be motivated by showing that these updates have fixed points at the global minima of the objective function. This can be seen by considering the gradient of Eq. (7): $K^+\vec{\alpha} - K^-\vec{\alpha} + \vec{b}$. The three terms in the gradient can be used to define multiplicative factors r_i in terms of a quadratic equation:

$$(K^+\vec{\alpha})_i r_i^2 + b_i r_i - (K^-\vec{\alpha})_i = 0. \qquad (11)$$

The general solution for r_i is given as the multiplicative factors in the update Eq. (10). In the case when the gradient vanishes, the solution is $r_i = 1$ in Eq. (11). Thus, it is clear that the multiplicative update rule exhibits a fixed point when the quadratic function achieves its minimum value. The delay τ_i associated with the maximal coefficient of α_i at this optimum is then given as an estimate for the ITD. If the possible delays are not known in practice, they may be estimated by first optimizing Eq. (6) over a uniformly sampled set of delays. This set may then be refined by iteratively adding more values of τ_i in the vicinity of the largest coefficients α_i, and recomputing the optimization. This procedure can be repeated until the time delay estimate is of the desired resolution.

It should be noted that when the K matrix is equal to a scaled identity matrix, the estimate of the ITD is equivalent to the estimate from maximizing the cross-correlation. This would be the case if the source signal is temporally uncorrelated and if there is no reverberation in the environment. The difference between nonnegative deconvolution and cross-correlation estimates arises when there are temporal correlations present in the signals $x_i(t)$. The nonnegative deconvolution explicitly models possible time delays in these correlations, and iteratively estimates nonnegative coefficents α_i that describe how these correlations could have arisen from possible echoes.

4. SIMULATION

To illustrate the differences between the cross-correlation based algorithms and nonnegative deconvolution, the various algorithms are used to perform ITD estimation on the following simulated source signal:

$$s(t) = e^{-\frac{1}{2}(t/T)^2} \sin 2\pi f_0 t \qquad (12)$$

In this simulation, we chose the carrier frequency to be $f_0 = 1000$ Hz, and T was chosen to give a bandwidth around 700 Hz. The source signal was delayed by a relative time shift $\tau = 125$ μs and discretely sampled at 10 kHz to generate two signals. To investigate the effect of noise and reverberation in these algorithms, white noise and a simulated secondary echo was also added to the signals $x_1(t)$ and $x_2(t)$. The level of the Gaussian white noise was chosen with a signal to noise ratio equal to 50 dB, while the simulated echo was designed with a delay time of 725 μs and a relative amplitude of -3 dB.

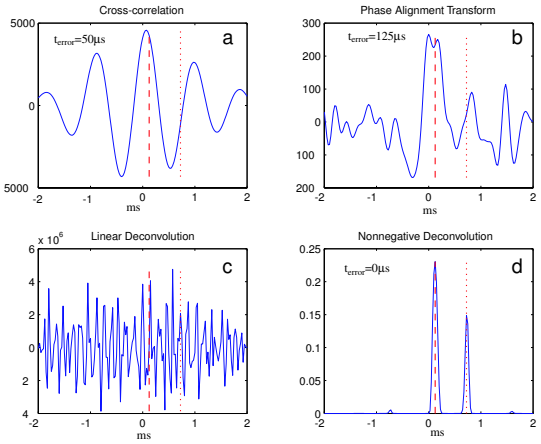

Fig. 2. Time-delay estimation on simulated data using: (a) cross-correlation, (b) phase alignment transform, (c) linear deconvolution, (d) nonnegative deconvolution. The true time delay between the signals is located at .125 ms, indicated by the dashed line. A secondary echo occurs at .725 ms, indicated by the dotted line.

Fig. 2 shows the results of applying various time delay estimation algorithms to this simulated data. The cross-correlation function is contaminated by the presence of the secondary echo, and the maximal value of the cross-correlation is shifted by 50 μs from the true time delay. On the other hand, the phase alignment transform (PHAT) is more affected by the white noise present in the frequency

bands outside the bandwidth, and its peak value shows much more variability (shifts as large as 125 μs). The linear deconvolution solution is also very susceptible to noise, due to the ill-conditioned matrix K in the optimization of Eq. (7). In fact, it is almost impossible to tell from the linear deconvolution coefficients anything about the time delays present in the signals.

In constrast, the deconvolution with nonnegativity constraints exhibits quite different behavior than the linear deconvolution. Not only is it able to accurately predict the ITD, it also is able to predict the scale and time shift of the secondary echo as well. In this case, the nonnegativity constraints prevent the deconvolution from amplifying the noise present in the signals, yet still are able to robustly model the contamination from echoes due to reverberation.

5. EXPERIMENTAL RESULTS

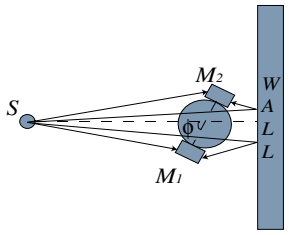

Fig. 3. Experimental robot platform for ITD estimation in a reverberant room. S: source source; M1 and M2: microphones.

To test the performance of the algorithms in a real acoustic environment, we performed ITD estimation using a robot platform as shown in Fig. 3. A small mobile robot was used to record the sound signals from a pair of omnidirectional microphones, with a microphone separation of 7 cm. A speaker playing recorded speech was used as a sound source and placed approximately 150 cm away from the robot. The acoustic signals were digitized at a sample rate of 16 kHz, and digitally recorded over a wireless network onto a central server where the data was analyzed. The experiment was conducted in a noisy lab environment, and the robot was placed approximately 40 cm from a concrete wall.

The robot head was oriented at several different azimuthal angles, and measurements were taken with window sizes of 512 samples from the two microphones. At each azimuthal angle, 30 measurements were taken, and the different algorithms were used to estimate the ITD. The true ITD was calculated using a longer wide-band white noise sound signal in which all the algorithms agreed on the time delay estimate.

The results of the ITD estimates are shown in Table 1. The nonnegative deconvolution algorithm shows a smaller bias in the ITD estimates than from cross-correlation, and smaller variability than the phase alignment transform. Thus, in this particular noisy environment with reverberation, there is some preliminary evidence that nonnegative deconvolution may be advantageous in estimating ITD for source localization.

Azimuthal angle	35^0		60^0		85^0	
True ITD (μs)	150		241		307	
	avg	std	avg	std	avg	std
Deconvolution	163	18	262	18	316	14
Cross-correlation	174	15	272	17	321	18
PHAT	158	11	235	68	298	48

Table 1. Experimental ITD estimation by the different algorithms, showing the average and standard deviation of the time difference estimates.

To summarize, we have presented nonnegative deconvolution as an alternative for ITD estimation. A computationally efficient algorithm with global convergence properties exists for computing the deconvolution, and results in estimates that may be more robust in noisy, reverberant environments. Current work involves extending the algorithm to also estimate the true underlying source signal, along with the nonnegative filter coefficients of the multipath reflections. Finally, we acknowledge the ARO and NSF for financial support, and Fei Sha for useful discussions.

6. REFERENCES

[1] J.D. de Jesus, J.J.V. Calvo, and A.I. Fuente, *IEEE Aerospace and Electronic Systems Magazine*, vol. 15, no. 2, pp. 9–16, 2000.

[2] Joe Chen, Kung Yao, and Ralph E. Hudson, *IEEE Signal Processing Magazine*, vol. 19, no. 2, pp. 30–39, 2002.

[3] H. Krim and M. Viberg, *IEEE Signal Processing Magazine*, vol. 13, no. 4, pp. 67–94, 1996.

[4] P. Stoica and R. L. Moses, *Introduction to Spectral Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1997.

[5] C. H. Knapp and G. C. Carter., *IEEE Transactions on ASSP*, vol. 24, no. 4, pp. 320–327, 1976.

[6] J.J. Fuchs, *IEEE Transactions on Signal Processing*, vol. 47, pp. 237–243, 1999.

[7] Jacob Benesty, Tomas Gansler, Dennis R. Morgan, M. Mohan Sondhi, and Steven L. Gay, Springer-Verlag, 2001.

[8] Y. T. Chan, J. M. Riley, and J. B. Plant, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, pp. 8–16, 1980.

[9] L.B. Lucy, *Astronomical Journal*, vol. 79, pp. 745–754, 1974.

[10] Daniel D. Lee and H. Sebastian Seung, *Nature*, vol. 401, pp. 788–791, 1999.

[11] Fei Sha, Lawrence K. Saul, and Daniel D. Lee, in *Advances in Neural Information Processing Systems*, Sebastian Thrun Suzanna Becker and Klaus Obermayer, Eds. 2002, vol. 15, The MIT Press.