A ROBUST METHOD FOR SPEECH SIGNAL TIME-DELAY ESTIMATION IN REVERBERANT ROOMS

Michael S. Brandstein[1] *Harvey F. Silverman*[2] *

[1] Division of Engineering and Applied Sciences
Harvard University
Cambridge, MA 02138

[2]Laboratory for Engineering Man/Machine Systems
Division of Engineering
Brown University
Providence, RI 02912

ABSTRACT

Conventional time-delay estimators exhibit dramatic performance degradations in the presence of multipath signals. This limits their application in reverberant enclosures, particularly when the signal of interest is speech and it may not possible to estimate and compensate for channel effects prior to time-delay estimation. This paper details an alternative approach which reformulates the problem as a linear regression of phase data and then estimates the time-delay through minimization of a robust statistical error measure. The technique is shown to be less susceptible to room reverberation effects. Simulations are performed across a range of source placements and room conditions to illustrate the utility of the proposed time-delay estimation method relative to conventional methods.

1. INTRODUCTION

The relative time-delay associated with a signal source and a pair of spatially separated sensors forms the basis of many microphone-array algorithms; passive-source localization and beamformer steering are typical. The prevalent technique for time-delay estimation (TDE) is based upon Generalized Cross-Correlation (GCC) in which the delay estimate is obtained as the time-lag which maximizes the cross-correlation between filtered versions of the received signals [1]. A variation on the GCC TDE involves the minimization of a weighted least-squares (LS) criterion in the spectral phase domain [2]. In the presence of single-path propagation, maximum likelihood (ML) versions of the GCC- and LS-based TDE's have been well studied and shown to be practical and to obtain theoretical bounds. However, these methods exhibit dramatic performance degradations in the presence of simple multipath channels (a few echoes) [3] as well as the more complex scenario of a reverberant room (a very large number of closely spaced echoes, the equivalent of nearly-flat multiplicative noise in the spectral domain) [4]. These shortcomings limit their applicability for time-delay estimation in realistic enclosures.

Several methods are available for improving TDE performance in multipath environments. These include incorporating sub-optimal but more robust filters into the methods [1] and applying cepstral prefilters to deconvolve the signals prior to TDE estimation [5]. The use of sub-optimal filters will be examined here in a limited context and shown to provide improvement over the ML estimators. However, the deconvolution approach is not appropriate for this application. Because the signal content is unknown, non-stationary speech, and the sources themselves may be moving, it is typically not possible to estimate the characteristics of the reverberant channel as required for these prefiltering techniques.

Motivated by the shortcomings of these conventional TDE algorithms under blind multipath conditions, this paper develops an alternative method of time-delay estimation based upon a natural extension of the LS TDE. By reformulating the problem as a linear regression of phase data, robust statistical methods may be applied to the estimation procedure. Instead of minimizing a weighted least-squares criterion, alternative error measures capable of deemphasizing the outlier effects created under realistic environments may then be employed. This alternative approach is shown to be less susceptible to room reverberation effects. The next section outlines the conventional TDE's addressed here and discusses the proposed robust method. In Section 3 simulations are performed across a range of source placements and room conditions to illustrate the utility of the proposed time-delay estimator. Section 4 presents some interpretations and conclusions.

2. METHODS FOR TIME-DELAY ESTIMATION

The signals received at the two microphones, $x_1(t)$ and $x_2(t)$, may be modeled as:

$$\begin{aligned} x_1(t) &= h_1(t) * s(t) + n_1(t) \qquad 0 \le t \le T \\ x_2(t) &= h_2(t) * s(t-\tau) + n_2(t) \end{aligned}$$

where τ is the relative signal delay of interest, $h_1(t)$ and $h_2(t)$ represent the impulse responses of the reverberant channels, $s(t)$ is the speech signal, $n_1(t)$ and $n_2(t)$ correspond to uncorrelated noise, and $*$ denotes linear convolution. Following [6], the analysis time interval, T, will be limited to a $20ms - 30ms$ range corresponding to the stationarity interval of speech.

TDE based upon Generalized Cross-Correlation

The GCC TDE may be expressed as:

$$\begin{aligned} \hat{\tau}_{GCC} &= \arg\max_{\tau} R_{x_1x_2}(\tau) \\ R_{x_1x_2}(\tau) &= \int_{-\infty}^{\infty} W(\omega) X_1(\omega) X_2'(\omega) e^{j\omega\tau} d\omega \end{aligned} \quad (1)$$

The generalized cross correlation function for a given time-lag, $R_{x_1x_2}(\tau)$, is calculated as the inverse Fourier transform of the received signal cross-spectrum, $X_1(\omega)X_2'(\omega)$, scaled by a weighting function, $W(\omega)$. Two distinct GCC TDE's will be considered here. The first is the maximum likelihood TDE, τ_{GCC-ML}, detailed in [1]. Using an ideal weighting function, $W_{ML}(\omega)$, derived from magnitude squared signal coherence and roughly equivalent to a frequency-dependent

*This work funded by NSF grants MIP-9314625 and MIP-9509505

SNR, the estimator is asymptotically unbiased and efficient for uncorrelated, stationary Gaussian signal and noises and no multipath. In practice the required coherence function is unavailable *a priori* and must be estimated from the given data. This is typically done via a temporal averaging technique, such as in [7]. The coherence-estimation process can prove to be problematic for any non-stationary signal. This issue is addressed in [6] where a single-frame, ML-type weighting approximation, $\hat{W}_{ML}$, is shown to offer advantageous results with speech signals. The approximated ML weighting, which will be used in the simulations to follow, is roughly equivalent to the signal SNR evaluated from a single frame of observed data and given by:

$$\hat{W}_{ML}(\omega) = \frac{|X_1(\omega)||X_2(\omega)|}{|N_1(\omega)|^2|X_2(\omega)|^2 + |N_2(\omega)|^2|X_1(\omega)|^2}$$

The noise power spectra, $|N_1(\omega)|^2$ and $|N_2(\omega)|^2$, are assumed to be available or estimatable during silence intervals.

A second GCC-based TDE known as the Phase Transform, $\tau_{GCC-PHAT}$, uses only the phase information at each frequency to derive the weighting function:

$$W_{PHAT}(\omega) = |X_1(\omega)X_2'(\omega)|^{-1}$$

By placing equal emphasis on each frequency, the W_{PHAT} weighting is sub-optimal under ideal conditions, but tends to be less susceptible to anomalous conditions, particularly reverberation [1]. As a result of these features, the Phase Transform has been investigated as a means for speech signal TDE in realistic environments [8]. Several other practical weighting schemes which compromise theoretical optimality for robust performance are available for use with the GCC TDE [1]. However, only these two relatively extreme weightings will be evaluated here.

TDE based upon Linear Regression

The TDE problem may be reformulated in terms of a linear regression problem by noting that the phase of the cross-spectrum, $\theta(\omega)$, varies linearly with angular frequency, the constant of proportionality being τ, i.e.:

$$\theta(\omega) = arg(X_1(\omega)X_2'(\omega)) = \omega\tau + \epsilon(\omega)$$

where $\epsilon(\omega)$ is a noise term. The time delay is now found by 'fitting' a line to the phase data. The traditional approach involves the minimization of a weighted least-squares (LS) cost function:

$$\hat{\tau}_{LS} = \arg\min_{\tau} \int_{-\infty}^{\infty} \psi(\omega)(\theta(\omega) - \omega\tau)^2 d\omega \qquad (2)$$

In [2] it is shown that under similar ideal signal conditions to the GCC-ML estimate, the $\epsilon(\omega)$ terms are zero-mean, uncorrelated, and Gaussian. Under such circumstances, the LS TDE generates the ML estimate, τ_{LS-ML}. The ideal phase weighting, $\psi_{ML}(\omega)$, is derived from the magnitude squared coherence function in a manner similar to $W_{ML}(\omega)$.

The GCC and LS-based approaches to time-delay estimation are nearly equivalent in terms of their mathematical development. The expression in (2) represents a first-order approximation to (1) and may be derived directly from the GCC TDE criteria.

The GCC and LS based approaches to the TDE problem are derived under single-path conditions and each produces theoretically optimal results only under ideal noise conditions. When reverberations are introduced, the cross-spectrum phase noise terms may be significantly biased, rendering either ML weighting ($W_{ML}(\omega)$ or $\psi_{ML}(\omega)$) inappropriate and producing significant error in the corresponding estimators. The modification of the weighting by use of the Phase-Transform dissipates some of these effects. Given the nature of the phase errors encountered under these conditions, some benefit may be derived by putting aside the explicit weighting functions and addressing the shortcomings of the LS norm as a cost function; it tends to overemphasize biased phase data. A potentially favorable approach is to consider alternative regression cost functions which are robust to the outlier phases due to reverberation and still allow for some means of weighting phases associated with high SNR conditions. Specifically, the linear regression problem may be generalized [9] in terms of a generic cost function $\rho(x)$:

$$\hat{\tau}_R = \arg\min_{\tau} \int_{-\infty}^{\infty} \rho\left(\frac{\theta(\omega) - \omega\tau}{S(\omega)}\right) d\omega \qquad (3)$$

The weighted LS criterion is a special case of the generalized regression problem with $\rho(x) = x^2$ and $S(w) = 1/\sqrt{\psi(\omega)}$. A popular and effective cost function for robust regression is Tukey's Biweight [10, 11], given by:

$$\rho_{BI}(x) = \begin{cases} -(1-x^2)^3/6 & |x| \leq 1 \\ 0 & |x| > 1 \end{cases} \qquad (4)$$

The Biweight corresponds to the negative of a smooth, symmetric, unimodal probability density. It assigns a maximal error value to any scaled absolute residual > 1 thereby diminishing the effect of outliers in skewing the cost function. A scaling term, $S(\omega)$, has been included in the cost function. In general, the scale factor affects the nature of the error space. Uniformly small $S(\omega)$ terms produce many local minima and discontinuities in the error surface, while large scaling terms generate a more continuous, but less sensitive error regions. The scaling factor may also be used to place some emphasis on those frequencies expected to offer an SNR advantage. In this respect it is analogous to an inverse variance weighting.

Figure 1 offers an informal comparison of the features associated with the time-delay estimation criteria discussed. The criteria are plotted as functions of potential delay for a pair of 25.6ms Hanning windowed, 20kHz sampled speech segments generated under a .1s room reverberation condition (The simulation process will be detailed in the next section.) with a .55ms delay and uncorrelated, white noise added to each channel giving a 25dB SNR. For the GCC-based methods, plots (A) and (B), the delay is estimated from the peak of the appropriately filtered cross-correlation function. Two Biweight error criteria are presented. Plot (C) incorporates a small scaling factor ($S(\omega) = \pi/3$) while plot (D) uses a larger scale ($S(\omega) = 3\pi/4$). The corresponding delay estimates are found from the minimum of their respective criteria. In each case, the true delay (10.99 samples) is plotted with a solid line while the corresponding estimate location is indicated by a dashed line. Each of the criteria functions possesses a global extreme at the true delay. In each case the spurious extrema are present to some degree. The GCC-ML criterion appears most sensitive to these anomalous peaks; the GCC-PHAT is less so, producing a more pronounced global maximum. The Biweight cost functions possess a desirable, distinct valley at the true delay and

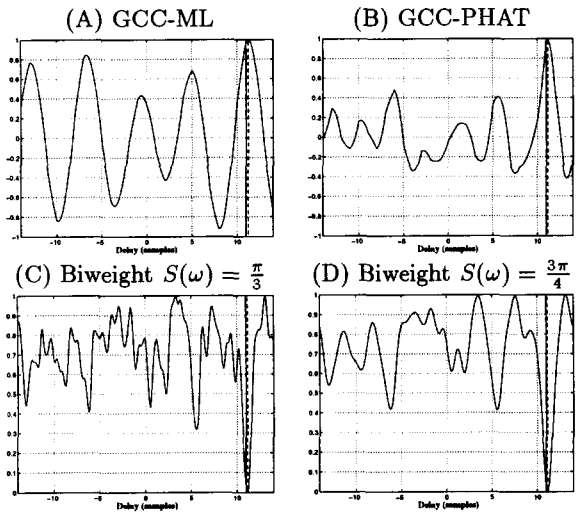

Figure 1. Sample Error Criteria associated with each of the TDE's: For the GCC-based methods, (A) and (B), the delay is estimated from the peak of the GCC function. For Biweight methods, (C) and (D), the estimate is found from the error minimum. In each case, the true delay is plotted with a solid line while the corresponding estimate is shown with a dashed line.

appear to provide the best estimate resolution of the set. The varying scale factors exhibit a general tradeoff between estimate resolution and spurious minima. The error space associated with the small scale factor, plot (C), possesses a well defined minima at the true location. The coarser Biweight scaling, plot (D), produces a smoother error surface, compromising resolution for rejection of local extrema.

3. SIMULATIONS

The relative performance of the time-delay estimators was evaluated through a series of Monte Carlo trials in a simulated $(4m \times 7m \times 2.75m)$ rectangular room (illustrated in Figure 2) with plane reflective surfaces and uniform, frequency-independent reflection coefficients. Room impulse responses were generated with the image model technique [12] using intra-sample interpolation, up to sixth order reflections, and cardioid microphone patterns. Room reverberation times, T, ranged from $0s$ to $0.4s$. The corresponding reflection ratio, β, used by the image model was calculated via Eyring's formula:

$$\beta = \exp(-13.82/[c(L_x^{-1} + L_y^{-1} + L_z^{-1})T])$$

where L_x, L_y, and L_z are the room dimensions and c is the speed of sound in air ($\approx 342m/s$). Three different source locations were considered corresponding to small, moderate, and large bearing angles relative to the microphone pair along with a microphone separation of 0.3m. Details of the simulation parameters are listed in Figure 2.

For each combination of parameters, 250 segments of 20kHz sampled speech were convolved with the appropriate channel impulse response. White, zero-mean, Gaussian noise with a fixed energy level was added to the segments which were then truncated to 25.6ms analysis frames using a Hanning window. The resulting signal-to-noise ratios varied from 20-35dB depending on the content of the speech signal. Knowledge of the background-noise variance was assumed to be available and incorporated into the applicable weighting functions.

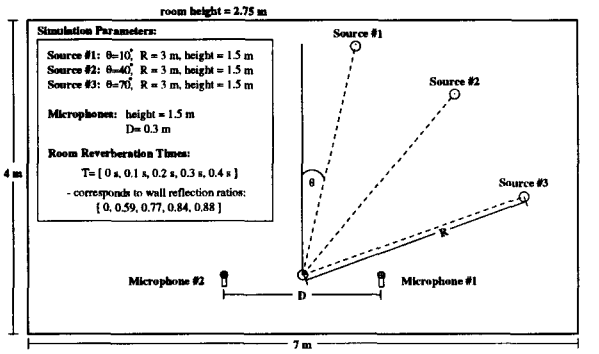

Figure 2. Overhead View of the Room Set-Up and Simulation Parameters

The GCC-based TDE's were calculated from discrete-time versions of (1) through conventional processing techniques. A sampled version of the GCC function was computed with the FFT and the integer sample delay corresponding to the maximum of the GCC function was found. In practice, the domain of potential time-delays was limited to possible values given the microphone separation. This coarse delay estimate was refined through quadratic interpolation and then used as the initial condition of a limited Newton-Raphson iterative search of the sinc interpolated GCC function. While computationally more intensive, this multi-stage approach achieves sub-sample TDE resolution with accuracy beyond that of a parabolic fit while avoiding convergence to a local maximum [3].

The linear regression-based delay estimator, τ_R, was calculated via a global search over the continuous range of possible delay values. A widely available, simplex-based non-linear optimization routine (the MATLAB™ 'fmins' function) was employed over a regularly spaced-grid of initial search locations. The evaluation of discrete-time version of (3) includes the unwrapping modulo 2π of the phase data relative to the phase of the hypothesized line fit prior to calculation of the error measure in question.

Four different TDE's were evaluated: GCC-ML, GCC-PHAT, Biweight with scale $\pi/3$ (BI1), and Biweight with scale $3\pi/4$ (BI2). Bias, variance, root-mean square error (rmse), and % anomaly statistics were calculated over the 250 speech segments with each of the sources and reverberation times. Figures 3 presents % anomaly and rmse statistics obtained using the three source locations. The % anomaly figures represent the percentage of estimates outside a 10° absolute error threshold. The rmse value incorporates the tradeoff between bias and variance into a single statistic. It was calculated using the non-anomalous time-delays and then converted to a direction-of-arrival (DOA) in degrees. The GCC-based TDE's have been plotted with dashed lines while the results obtained with the Biweight method are delineated by a solid lines.

With regard to estimate anomalies, the TDE's performed comparable under most combinations of source bearing and reverberation times. Exceptions to this pattern occurred for Source#1 (the mild bearing angle) with the GCC-ML and BI1 producing markedly higher % anomalies as the reverberation time was incremented. This behavior is consistent with the interpretations of Figure 1 where the GCC-ML and the small-scale Biweight criteria exhibit inferior rejection of spurious extrema and are therefore more susceptible to anomalous estimates. As expected, the GCC-ML yields superior rmse performance in the noise only case, but degrades dramatically as reverberations are

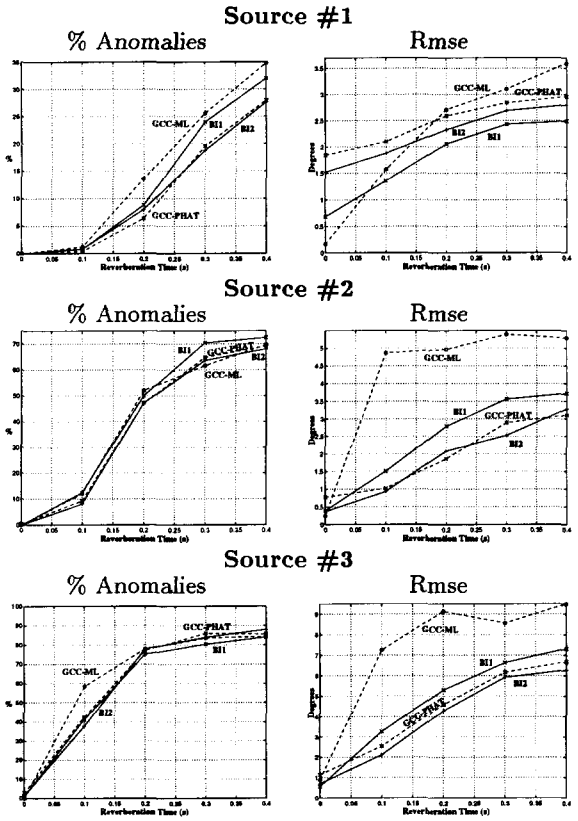

Figure 3. TDE simulation results: Anomalies and rmse trial statistics for different source locations and four TDE's: GCC-ML, GCC-PHAT, small-scale Biweight (BI1), and large-scale Biweight (BI2).

introduced. The GCC-PHAT and Biweight TDE's outperform the GCC-ML when reverberations are present. With Source#1, the small-scale Biweight (BI1) achieves the lowest rmse figures. This is consistent with the anomaly robustness versus estimate resolution tradeoff associated with the Biweight scaling size. As the source bearing angle is increased, Sources #2 and #3, the smoothing benefits of the larger scale size appear to outweigh the decreased resolution. The BI2 TDE consistently achieves lower rmse scores than BI1 and mildly outperforms the GCC-PHAT TDE.

4. DISCUSSION

As the results of the preceding section clearly illustrate, the popular method for time-delay estimation, the Generalized Cross-Correlator, is extremely susceptible to environmental reverberation effects. In the presence of even mild reverberations, the GCC TDE benefits considerably by eliminating the optimal filter weights.

The major goal of this work has been to introduce an effective means for doing time-delay estimation in the presence of unknown reverberant channels. The TDE presented employs linear regression using Tukey's Biweight as the relevant error criterion. The motivation behind this approach comes from the interpretation of the time-delay estimation problem as a line fit in the phase domain. The effects of reverberation on the observed phases are analogous to outliers in the linear regression problem. The Biweight measure assigns maximum weight to errors beyond a certain threshold thereby discounting the effects of outliers in the estimation process. (The traditional least-squares fit makes no account for aberrant data and tends to be easily led astray by outliers.) The utility of this strategy has been shown through a series of simulations in reverberant enclosures. In each of the conditions investigated, the Biweight TDE outperformed estimators based upon the conventional Generalized Cross-Correlator.

The Biweight criterion, while popular and seemingly effective for this application, represents only an initial approach to this genre of TDE estimator. Many possible means exist for performing robust linear regression in the presence of outliers. Other error measures motivated by this basic strategy are open to investigation as well as schemes for adjusting the kernel scaling to incorporate prior knowledge of the phase attributes. While not detailed here, dynamic adjustment of the criterion parameters as a function of source bearing, frequency-dependent SNR, and environmental conditions also shows considerable promise for improving the performance of the proposed time-delay estimator.

REFERENCES

[1] C. H. Knapp and G. C. Carter. The generalized correlation method for estimation of time delay. *IEEE Trans. Acoust. Speech Signal Process.*, ASSP-24(4):320–327, August 1976.

[2] Y. Chan, R. Hattin, and J. Plant. The least squares estimation of time delay and its use in signal detection. *IEEE Trans. Acoust., Speech, Signal Processing*, 26(3):217–222, June 1978.

[3] J. Hassab. *Underwater Signal and Data Processing.* CRC Press, Boca Raton, FL, 1989.

[4] B. Champagne, S. Bèdard, and A. Stèphenne. Performance of time-delay estimation in the presence of room reverberation. *IEEE Trans. Speech Audio Proc.*, 4(2):148–152, March 1996.

[5] A. Stèphenne and B. Champagne. Cepstral prefiltering for time delay estimation in reverberant environments. In *Proceedings of ICASSP95*, pages 3055–3058. IEEE, 1995.

[6] M. Brandstein, J. Adcock, and H. Silverman. A practical time-delay estimator for localizing speech sources with a microphone array. *Computer, Speech, and Language*, 9:153–169, April 1995.

[7] G. C. Carter, C. H. Knapp, and A. H. Nuttall. Estimation of the magnitude-squared coherence function via overlapped fast fourier transform processing. *IEEE Transactions Audio and Electroacoustics*, AU-21(4):337–344, August 1973.

[8] M. Omologo and P. Svaizer. Acoustic source location in noisy and reverberant environment using csp analysis. In *Proceedings of ICASSP96*, pages 921–924. IEEE, 1996.

[9] P. Huber. *Robust Statistics.* John Wiley and Sons, New York, 1981.

[10] A. Beaton and J. Tukey. The fitting of power series, measuring polynomials, illustrated on band-spectrographic data. *Techometrics*, 16:147–185, 1974.

[11] S. Morgenthaler. Fitting redescending m-estimators in regression. In K. Lawrence and J. Arthur, editors, *Robust Regression: Analysis and Applications*, pages 105–128. Marcel Dekker, Inc., New York, 1991.

[12] J. B. Allen and D. A. Berkley. Image method for efficiently simulating small room acoustics. *J. Acoust. Soc. Am.*, 65(4):943–950, April 1979.