EFFECTS OF ROOM REVERBERATION ON TIME-DELAY ESTIMATION PERFORMANCE †

Stéphane Bédard ‡, *Benoit Champagne*
and Alex Stéphenne

INRS-Télécommunications, Université du Québec, 16 Place du Commerce
Verdun, Québec, Canada H3E 1H6

ABSTRACT

Recently, time-delay estimation (TDE) between two or more microphones has been proposed as a means for the passive localization of a talker in an audio-conference room. In this paper, we present a simulation study of the effects of room reverberation on the performance of the maximum likelihood (ML) estimator of the time delay, which is obtained by maximizing the output of a generalized cross-correlator. To this end, synthetic microphone signals are generated with the image model technique and are used to evaluate the bias, variance and probability of anomaly of the ML estimator as a function of the room reflection coefficients and other external parameters of interest. The results clearly demonstrate the adverse effects of room reverberation on MLTDE performance. Qualitative and quantitative explanations of these effects are provided.

1. INTRODUCTION

Microphone arrays are used increasingly for sound transduction in audio-conference rooms due to their ability to remove (at least partially) reverberation and other directional interferences from the desired signal. In this context, a major problem consists of localizing the dominant talker so as to continuously steer the array in his/her direction. Recently, time delay estimation (TDE) between the direct path signals received by two or more microphones has been proposed as a means for the passive localization of the dominant talker [1].

One of the most popular methods for TDE is that of maximum likelihood (ML). In this method, the delay estimate between a pair of receivers is obtained as the time lag which maximizes the cross-correlation between filtered versions of the received signals [2]. The popularity of MLTDE stems from its relative simplicity of implementation and its optimality under appropriate conditions. Indeed, for uncorrelated Gaussian signal and noises and single path propagation (i.e. no reverberation), the ML estimator of time delay is asymptotically unbiased and efficient in the limit of long observation intervals.

† Support for this work has been provided by NSERC operating grant OGP0105533 and by FCAR grant 93-ER-1577.

‡ S. Bédard is now with Bell SYGMA Telecom Solutions, Montreal, Quebec, Canada.

The statistical performance of MLTDE under single path propagation has been extensively studied in the past and is generally well understood [2]-[3]. Some studies have also addressed the behavior of MLTDE in the presence of a few echoes [4]. However, these results can not be used directly to predict the performance of the MLTDE method when the latter is applied to microphone signals captured in a reverberant room. Indeed, reverberation consists in the superposition of a very large number of echoes, which are characterized by temporal and spatial correlation.

In this paper, the effects of room reverberation on the performance of MLTDE are investigated under controlled conditions via Monte Carlo simulations on a digital computer. Synthetic microphone signals corresponding to different levels of reverberation are generated with the image model technique [5]. These signals are used to study the bias, variance and probability of anomaly of the ML estimator of time delay as a function of the room reflection coefficients and other external parameters of interest.

The simulation results clearly demonstrate the adverse effects of room reverberation on MLTDE performance. Simple interpretations of these results are proposed. In particular, the quantitative behavior of the estimator variance for small to moderate levels of reverberation can be explained naturally in terms of an equivalent signal-to-noise ratio (SNR_{eq}) which treats the reverberant energy at the microphone output as undesirable noise. For high levels of reverberation (i.e. low values of SNR_{eq}) the behavior of the ML estimator is dominated by large errors.

2. REVIEW OF MLTDE

The classical TDE problem consists in estimating the time delay between noisy versions of a common source signal observed at the outputs of two spatially separated receivers (e.g., microphones). Typically, the unknown time delay is equal to the difference in travel time of a wavefront propagating from the source to the two receivers. Hence, estimation of the time delay provides information about the location of the source.

A simple and widely used signal model for the classical TDE problem is the following. Let $x_i(t)$, $i = 1, 2$, denote the output signal of the ith receiver. Then,

$$\begin{aligned} x_1(t) &= s(t) + n_1(t), \qquad 0 \le t \le T, \\ x_2(t) &= s(t+\tau) + n_2(t), \end{aligned} \tag{1}$$

0-7803-1775-0/94 $3.00

where $s(t)$ is the unknown source signal, τ is a free parameter representing the unknown delay, $n_i(t)$ is the additive noise at the ith receiver and T denotes the duration of the observation interval. It is further assumed that $s(t)$, $n_1(t)$ and $n_2(t)$ are (real) zero-mean, uncorrelated, stationary Gaussian random processes. This model corresponds to an ideal situation in which the signal propagation from the source to each receiver occurs along a single path, without attenuation in a non-dispersive medium. Moreover, the signal and noise spectra as well as the unknown time delay are assumed constant over time.

By definition, the ML estimator of the time delay is the value of τ which maximizes the likelihood function of the observed data, i.e.

$$\hat{\tau}_{ML} = \arg\max_{\tau \in \mathcal{D}} L(\mathbf{x}; \tau) \tag{2}$$

where $\mathbf{x}$ denotes the observed data (the signals $x_i(t)$, for $i = 1, 2$ and $0 \leq t \leq T$), $\mathcal{D}$ denotes the set of *a priori* delay values and $L(\mathbf{x}; \tau)$ is the likelihood function, defined as the joint probability density function of the data for a particular value of τ. The choice of the set $\mathcal{D}$ used for the search is generally based on physical considerations.

Using (1) together with the statistical assumptions made on the signal and noises, the following expressions can be obtained for the ML estimator of τ [2]:

$$\hat{\tau}_{ML} = \arg\max_{\tau \in \mathcal{D}} R_{ML}(\tau), \tag{3}$$

$$R_{ML}(\tau) = \int_{-\infty}^{\infty} \psi_{ML}(f) X_1(f) X_2^*(f) e^{j2\pi f \tau} df, \tag{4}$$

$$\psi_{ML}(f) = \frac{S(f)}{N_1(f)N_2(f)} \{1 + \frac{S(f)}{N_1(f)} + \frac{S(f)}{N_2(f)}\}^{-1}, \tag{5}$$

where $X_i(f)$ denotes the Fourier transform of $x_i(t)$ over the interval $0 \leq t \leq T$, * denotes the complex conjugate and $S(f)$, $N_1(f)$ and $N_2(f)$ denote the power spectral densities of $s(t)$, $n_1(t)$ and $n_2(t)$, respectively.

It is useful to interpret the function $R_{ML}(\tau)$ (4) as the cross-correlation between filtered versions of the observed signals $x_1(t)$ and $x_2(t)$, with filter transfer functions given by $H_1(f) = H_2(f) = \sqrt{\psi_{ML}(f)}$, respectively. The function $R_{ML}(\tau)$ is therefore a particular form of the generalized cross-correlation, which allows for the use of arbitrary filters $H_i(f)$. The role of the frequency weighting $\psi_{ML}(f)$ in (4) is to improve the accuracy of the delay estimate by attenuating the signals fed into the correlator in spectral regions where the signal-to-noise ratio is the lowest. In practice, the various spectral densities entering $\psi_{ML}(f)$ (5) are often unknown. In this case, a commonly used approach by which the ML estimator can be approximated consists of estimating those quantities from the observed data and substituting the estimates in (5).

For the classical TDE model (1), it can be shown that the ML estimator of the time delay is asymptotically unbiased and efficient in the limit $T \to \infty$. By efficient, we mean that the variance of $\hat{\tau}_{ML}$ reaches the Cramer-Rao lower bound, which is given by [2]:

$$\sigma_{cr}^2 = \{8\pi^2 T \int_0^{\infty} \psi_{ML}(f) S(f) f^2 df\}^{-1}. \tag{6}$$

For a fixed observation interval T, the performance of MLTDE depends on the signal-to-noise ratio (SNR). Although a mathematical analysis of the performance is usually quite involved for finite T, the overall behavior can usually be characterized in terms of a few fundamental parameters [3]. For instance, in the case of baseband signals, which is of particular interest here, the SNR domain can be partitioned into three disjoined regions, with distinct behavior of the ML estimator in each region. Above a so-called threshold SNR, denoted SNR_{th}, $\hat{\tau}_{ML}$ is generally unbiased and efficient. For intermediate SNR, i.e. just below SNR_{th}, the variance of the estimator suddenly departs from the Cramer-Rao lower bound. This behavior is due to an increase in the percentage of large errors or anomalous estimates, which correspond to erroneous peaks of $R_{ML}(\tau)$ (4) caused by the noise. For the signal model (1) with uncorrelated signal and noises, these peaks are uniformly distributed in the *a priori* search region $\mathcal{D}$. Finally, at very low SNR, the estimation process is entirely dominated by large errors and $\hat{\tau}_{ML}$ is therefore uniformly distributed in the interval $\mathcal{D}$.

3. SIMULATION DETAILS

The signal model (1) is inappropriate to represent the transmission of an acoustic signal between a source and two microphones in a small room. In this case, due to multiple reflections of the sound waves on the walls of the room, several delayed and attenuated versions of the source signal are received by each microphone in addition to the direct path signal. This perceivable phenomenon, known as reverberation, cannot be directly accounted for in model (1). Indeed, even if the received echoes were included in the noise terms, i.e. $n_i(t)$, the assumption of uncorrelated signal and noises would then be violated. Hence, without further understanding of the problem, the theoretical performance results reported in Section 2 for the classical MLTDE problem can not be applied when the received signals are subjected to reverberation. In this section, we describe computer simulation experiments precisely aimed at studying the effects of room reverberation on the performance of MLTDE.

The simulation scenario consists of a rectangular room with plane reflecting walls characterized by a unique reflection coefficient β ($0 \leq \beta \leq 1$) which is independent of the frequency and the angle of incidence of the acoustic rays. A rectangular coordinate system $Oxyz$, with origin in one corner of the room and axes parallel to the walls (the z-axis represents the vertical) is used to reference points in this room. The dimensions of the room along the x, y and z-axes are $L_x = 10.0m$, $L_y = 6.6m$ and $L_z = 3.0m$, respectively. The room contains a single omnidirectional point acoustic source with fixed position vector $\mathbf{r}_s$. The source signal is monitored by two point receivers (microphones), whose directivity patterns are not restricted to be omnidirectional. The microphones are

fixed with position vectors given by $\mathbf{r}_{m,1} = (6.5, 2.8, 1.8)m$ and $\mathbf{r}_{m,2} = (6.5, 3.8, 1.8)m$, respectively. Except for these ideal source and receivers, the room is empty.

Assuming that the acoustic space is time-invariant and linear, the microphone output signals, $x_i(t)$, can be expressed as follows:

$$\begin{aligned} x_1(t) &= [h_1 * s](t) + n_1(t), \qquad 0 \le t \le T, \\ x_2(t) &= [h_2 * s](t) + n_2(t), \end{aligned} \tag{7}$$

where $s(t)$ is the source signal, $n_i(t)$ is the additive noise, $*$ denotes the operation of convolution and $h_i(t)$ is the acoustic impulse response between the source and the output of the ith microphone. The same assumptions as in Section 2 are made on $s(t)$ and the $n_i(t)$. The latter are used to model uncorrelated interferences other than reverberation. The acoustic impulse response $h_i(t)$ completely characterizes the transmission of the signal $s(t)$ from its source to the (output of the) ith microphone. Hence, the presence of reverberation in the received signal $x_i(t)$ is entirely accounted for by $h_i(t)$.

For the purpose of digital simulations, we assume that the microphone outputs $x_i(t)$ are passed through identical low-pass filters with a cutoff frequency $f_{max} =$ 5000Hz and then sampled synchronously at the Nyquist rate $f_s = 2f_{max} =$ 10kHz (sampling interval $T_s = 1/f_s = 10^{-4}s$). To generate the low-pass, sampled versions of the acoustic impulse responses $h_i(t)$ in (7), we use Allen and Berkley's implementation of the image model technique [5] with Peterson's modification. The method has also been generalized to allow for microphones with arbitrary (although frequency independent) directivity patterns, which give the amplitude attenuation at the microphone output as a function of the direction of incidence. In the simulations, we use identical microphones with a common cardioid directivity pattern given by $f(\theta) = (1 + \cos\theta)/2$, where θ is the angle formed by the incident ray and the direction vector (-1,0,0).

Using the above method, a pair of digital impulse responses is generated for each values of β and $\mathbf{r}_s$ considered. Although $h_i(t)$ in (7) theoretically extends to infinity, the digital responses used in the simulations are truncated to about 3000 samples ($0.3s$). Even for the worst scenario considered, i.e. $\beta = 0.9$, the truncated tail of the response is approximately 30dB below the main peak corresponding to the direct path signal.

The sampled version of the source signal $s(t)$ is obtained by passing a Gaussian white noise sequence with zero-mean and unit variance through a band-pass filter with the desired spectral characteristics. For the results reported in the next section, the lower and upper cut-off frequencies (-3dB) of the filter are $f_l = 450$ and $f_u = 3475$, respectively. This corresponds approximately to the telephone transmission bandwidth. The source signal is then convolved with the room impulse responses calculated above. Finally, two independent Gaussian white noise sequences, properly scaled to obtain a desired value of the in-band SNR at $\beta = 0$, are added to each channel.

Following the initial transients, the synthetic microphone signals are partitioned into contiguous frames of $K = 1024$ samples for processing. This corresponds to an integration time of $T = 102.4ms$. For each frame, a time delay estimate is obtained via a digital (software) implementation of the ML equations (3)-(5).

Based on *a priori* knowledge of the signal and noise spectra, $\psi_{ML}(f)$ (5) is set to 1 for $f_l \le f \le f_u$ (signal passband) and to 0 otherwise. A sampled version of $R_{ML}(\tau)$ (4), corresponding to the lag values $\tau = kT_s$, for $|k| = 0, 1, ..., K/2$, is calculated using conventional processing techniques based on the fast Fourier transform (FFT). The time delay estimate is then obtained in two steps. First, $R_{ML}(kT_s)$ is maximized over an *a priori* search grid defined by $|k| \le k_{max}$ where $k_{max} = 30$. This choice is based on geometrical considerations. Quadratic interpolation is then used to refine the estimate obtained in the first step.

For each set of synthetic microphone signals, 500 frames are processed, resulting in 500 independent time delay estimates. Estimates for which the magnitude of the estimation error exceeds $3T_s$ are identified as anomalies. This definition of an anomaly is based on the shape of the autocorrelation of the signal. In practice, large errors can be avoided by using a tracking algorithm, provided their percentage of occurrence is sufficiently low. Using the 500 time delay estimates, the following statistical performance measures are calculated: the percentage of anomalous estimates and the sample bias and standard deviation of the non-anomalous estimates.

4. RESULTS AND INTERPRETATION

Typical results corresponding to a source position vector $\mathbf{r}_s = (2.4835, 2.0, 1.8,)m$ (true delay $= -9T_s$) and in-band SNR=30dB are presented in Fig. 1 to 3, where the percentage of anomalous time delay estimates and the sample bias and standard deviation of the non-anomalous estimates, respectively, are plotted as a function of the reflection coefficient β. The vertical bar superimposed on each data point represents the 95 percent confidence interval for that measurement.

As can be seen in Fig. 1, an abrupt transition of the percentage of anomalies, from 0% to about 90%, occurs as β increases from 0.6 to 0.8. This sudden and severe deterioration in the performance of MLTDE is similar to the threshold effect observed in the single path scenario (Section 2) when the SNR goes below SNR_{th}. It is due to the presence of erroneous peaks in the output of the ML cross-correlator (4). In this case, however, the peaks are not caused by the background noise but result from the correlation existing between pairs of echoes on the different channels. As β incrases and the echoes become stronger, the number and the amplitudes of these erroneous peaks increase, eventually making the ML estimator unreliable.

Fig. 2 and 3 show a continual deterioration of the absolute bias and the standard deviation as β increases from 0 to the threshold value of 0.6. We also note that the relative effect of reverberation is worst for small values of β. The increase of the bias with β is due to the correlation existing between certain pairs of echoes received by the two microphones for lag values close to the true delay. The net effect is an apparent modification in the shape of the cross-correlator output (not shown due to lack of

space). The increase of the standard deviation with β is also caused by the echoes. However, as explained below, the mechanism involved appears to be practically independent of the particular structure of the echoes.

To justify this statement, let us introduce an equivalent signal-to-noise ratio as follows:

$$\mathrm{SNR}_{eq} = \frac{\int |H_i(f;0)|^2 S(f) df}{\int [N_i(f) + |H_i(f;\beta) - H_i(f;0)|^2 S(f)] df} \quad (8)$$

where $H_i(f;\beta)$ is the frequency response of $h_i(t)$ for a given β. SNR_{eq} provides a measure of the direct-path signal power, relative to the total interference power, where the latter is interpreted to include both the background noise power and the reverberation power (excluding the direct path). In the simulations, SNR_{eq} is calculated using the synthetic impulse responses $h_i(t)$ and is almost identical for both microphones.

In the single path scenario,the variance of the ML estimator in the small error regime is closely predicted by the Cramer-Rao lower bound (6). For the flat signal and noise spectra considered here, this bound is a function of the in-band SNR. In an attempt to predict the small error performance of the ML estimator in the presence of reverberation, we have calculated an equivalent Cramer-Rao lower bound, obtained from (6) upon replacement of SNR by SNR_{eq} (8). The resulting expression, denoted $\sigma^2_{eq}(\beta)$, is shown as a continuous line in Fig. 3. For $\beta \leq 0.6$, the simulation results closely match this curve. This indicates that in the small error regime, the effect of reverberation on the variance of the ML estimator is equivalent to that of uncorrelated white noise.

The above results remain valid for other geometrical configurations, except for the two case described below. The first case, which is not very practical, involves a perfectly symmetrical echo structure in which the direct path signal and its echoes reache the two microphones simultaneously. Here, the performance of the ML estimator improves as β increases. In the second case, the source or one of the microphones is close to a wall. As a result, the first few echoes are relatively strong and the deterioration of MLTDE performance occurs at smaller values of β. This situation can usually be avoided in practice.

REFERENCES

[1] H. F. Silverman and K. J. Doerr, "An algorithm for determining talker location using a linear microphone array and optimal hyperbolic fit," Brown University, Division of Engineering, Technical Report LEMS-77, July 1990.

[2] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 24, pp. 320-327, Aug. 1976.

[3] J. P. Ianniello, "Time delay estimation via cross-correlation in the presence of large estimation errors," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 30, pp. 998-1003, Dec. 1982.

[4] Y. T. Chan, "Time delay estimation in the presence of multipath propagation," *Proc. 1984 NATO Advanced Study Institute.*

[5] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics", *J. Acoust. Soc. Am.*, vol. 65, pp. 943-950, April 1979.

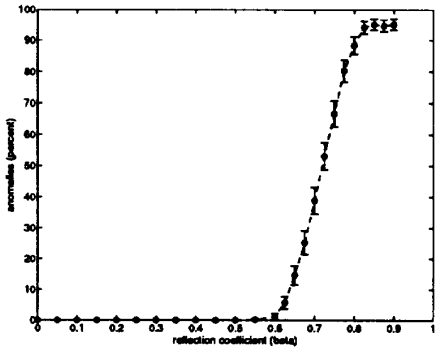

Fig. 1. Percentage of anomalous time delay estimates versus β (SNR = $30dB$).

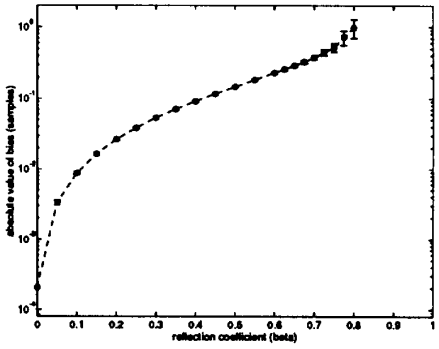

Fig. 2. Bias of non-anomalous estimates versus β (SNR = $30dB$).

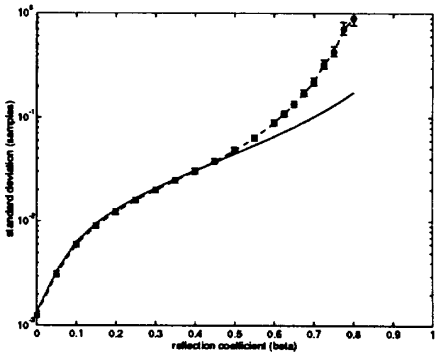

Fig. 3. Standard deviation of non-anomalous estimates versus β (SNR = $30dB$; σ_{eq} given by continuous line).