TIME DELAY ESTIMATION IN THE PRESENCE OF CORRELATED NOISE AND REVERBERATION

Yong Rui and Dinei Florencio
Microsoft Research
One Microsoft Way, Redmond, WA 98052

ABSTRACT

We propose a new two-stage framework for time delay estimation in the presence of correlated noise and reverberation. The new framework allows us to develop a set of new approaches as well as to unify existing ones. We further develop the maximum likelihood estimation when reverberation is present. The corresponding weighting function is a more accurate form of the weighting function proposed in [10], one of the best existing techniques. We compare our new algorithms with the existing ones and report superior performance.

1. INTRODUCTION

Using microphone arrays to locate sound source has been an active research topic since the early 1990's [2]. It has many important applications including video conferencing [1][5][10], video surveillance, and speech recognition [8]. In general, there are three categories of techniques for sound source localization, i.e. steered-beamformer based, high-resolution spectral estimation based, and time delay of arrival (TDOA) based [2]. So far, the most studied and widely used technique is the TDOA based approach. Various TDOA algorithms have been developed at Brown University [2], PictureTel [10], Rutgers [6], University of Maryland [12], USC [3], UCSD [4], and UIUC [8]. This is by no means a complete list. Instead, it is used to illustrate how much effort researchers have put into this problem.

While researchers are making good progress on various aspects of TDOA, there is still no good solution in real-life environment where two destructive noise sources exist: 1. spatially correlated noise, e.g., computer fans; and 2. room reverberation. With a few exceptions, most of the existing algorithms either assume uncorrelated noise or ignore room reverberation. Based on our own experience, testing on data with uncorrelated noise and no reverberation will almost always give perfect results. But the algorithm will not work well in real-world situations. In this paper, we explore various noise removal techniques to handle issue 1, and different weighting functions to deal with issue 2. The focus of this paper is on improving "single-frame" estimates. Multiple-frame techniques, e.g., temporal filtering [11], are outside the scope of this paper, but can always be used to further improve the "single-frame" results. On the other hand, better single frame estimates should also improve algorithms based on multiple frames.

The rest of the paper is organized as follows. In Section 2, we briefly review the general TDOA framework and various existing approaches. In Section 3, we look at the TDOA framework from a new two-stage perspective. The new perspective allows us to develop a set of new approaches as well as to unify existing ones. In Section 4, we give detailed comparisons between the set of proposed new approaches and the existing ones. The results show better performance of the proposed techniques. We give concluding remarks in Section 5.

2. TDOA FRAMEWORK

The general framework for TDOA is to choose the highest peak from the cross correlation curve of two microphones. Let $s(n)$ be the source signal, and $x_1(n)$ and $x_2(n)$ be the signals received by the two microphones:

$$\begin{aligned} x_1(n) &= s_1(n) + h_1(n) * s(n) + n_1(n) \\ &= a_1 s(n-D) + h_1(n) * s(n) + n_1(n) \\ x_2(n) &= s_2(n) + h_2(n) * s(n) + n_2(n) \\ &= a_2 s(n) + h_2(n) * s(n) + n_2(n) \end{aligned} \quad (1)$$

where D is the TDOA, a_1 and a_2 are signal attenuations, $n_1(n)$ and $n_2(n)$ are the additive noise, and $h_1(n)*s(n)$ and $h_2(n)*s(n)$ represent the reverberation. If one can recover the cross correlation between $s_1(n)$ and $s_2(n)$, $\hat{R}_{s_1 s_2}(\tau)$, or equivalently its Fourier transform $\hat{G}_{s_1 s_2}(\omega)$, then D can be estimated. In the most simplified case [3][8], the following assumptions are made:

1. signal and noise are uncorrelated
2. noises at the two microphones are uncorrelated
3. there is no reverberation

With the above assumptions, $\hat{G}_{s_1 s_2}(\omega)$ can be approximated by $\hat{G}_{x_1 x_2}(\omega)$, and D can be estimated as follows:

$$\begin{aligned} D &= \arg\max_{\tau} \hat{R}_{s_1 s_2}(\tau) \\ \hat{R}_{s_1 s_2}(\tau) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{G}_{s_1 s_2}(\omega) e^{j\omega\tau} d\omega \approx \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{G}_{x_1 x_2}(\omega) e^{j\omega\tau} d\omega \end{aligned} \quad (2)$$

While the first assumption is valid most of the time, the other two are not. Estimating D based on (2) therefore can easily break down in real-world situations. To deal with this

issue, various frequency weighting functions have been proposed, and the resulting framework is called *generalized* cross correlation:

$$D = \arg\max_{\tau} \hat{R}_{s_1 s_2}(\tau)$$
$$\hat{R}_{s_1 s_2}(\tau) \approx \frac{1}{2\pi}\int_{-\pi}^{\pi} W(\omega)\hat{G}_{x_1 x_2}(\omega)e^{j\omega\tau}d\omega \quad (3)$$

where $W(w)$ is the frequency weighting function.

In practice, choosing the right weighting function is of great significance. Early research on weighting functions can be traced back to the 1970's [6]. As can be seen from (1), there are two types of noise in the system, i.e., the ambient noise $n_1(n)$ and $n_2(n)$ and reverberation $h_1(n)*s(n)$ and $h_2(n)*s(n)$. Previous research [2][6] suggests that the traditional maximum likelihood (TML) weighting function is robust to ambient noise and phase transformation (PHAT) weighting function is better dealing with reverberation:

$$W_{TML}(\omega) = \frac{|X_1(\omega)||X_2(\omega)|}{|N_2(\omega)|^2|X_1(\omega)|^2 + |N_1(\omega)|^2|X_2(\omega)|^2} \quad (4)$$

$$W_{PHAT}(\omega) = \frac{1}{|\hat{G}_{x_1 x_2}(\omega)|} \quad (5)$$

where $X_i(w)$ and $|N_i(w)|^2$, $i = 1,2$, are the Fourier transform of the signal and the noise power spectrum, respectively. It is interesting to note that while $W_{TML}(w)$ can be mathematically derived [6], $W_{PHAT}(w)$ is purely heuristics based. Most of the existing work [2][3][6][8][12] use either $W_{TML}(w)$ or $W_{PHAT}(w)$.

3. A TWO-STAGE PERSPECTIVE

In this section, we look at the TDOA estimation problem as a two-stage process: remove the correlated noise and try to minimize the reverberation effect.

3.1. Correlated noise removal

In offices and conference rooms, there exist noise sources, e.g., ceiling fan, computer fan and computer hard drive. These noises will be heard by both microphones. It is therefore unrealistic to assume $n_1(n)$ and $n_2(n)$ as uncorrelated. They are, however, stationary or short-time stationary, such that it is possible to estimate the noise spectrum over time. We discuss three techniques to remove correlated noise. While the first one appeared in the literature [10], the other two did not appear explicitly.

3.1.1. Gnn subtraction (GS)

If $n_1(n)$ and $n_2(n)$ are correlated, then $\hat{G}_{x_1 x_2}(\omega) = \hat{G}_{s_1 s_2}(\omega) + \hat{G}_{n_1 n_2}(\omega)$. We therefore can obtain a better estimate of $\hat{G}_{s_1 s_2}(\omega)$ as:

$$\hat{G}^{GS}_{s_1 s_2}(\omega) = \hat{G}_{x_1 x_2}(\omega) - \hat{G}_{n_1 n_2}(\omega) \quad (6)$$

where $\hat{G}_{n_1 n_2}(\omega)$ is estimated when there is no speech.

3.1.2. Wiener filtering (WF)

Wiener filtering reduces stationary noise. If we pass each microphone's signal through a Wiener filter, we expect to see less amount of correlated noise in $\hat{G}_{x_1 x_2}(\omega)$.

$$\hat{G}^{WF}_{s_1 s_2}(\omega) = W_1(\omega)W_2(\omega)\hat{G}_{x_1 x_2}(\omega)$$
$$W_i(\omega) = (|X_i(\omega)|^2 - |N_i(\omega)|^2)/|X_i(\omega)|^2 \quad (7)$$
$$i = 1,2$$

where $|N_i(w)|^2$ is estimated when there is no speech.

3.1.3. Wiener filtering and Gnn subtraction (WG)

Wiener filtering will not completely remove the stationary noise. The residual can further be removed by using GS:

$$\hat{G}^{WG}_{s_1 s_2}(\omega) = W_1(\omega)W_2(\omega)(\hat{G}_{x_1 x_2}(\omega) - \hat{G}_{n_1 n_2}(\omega)) \quad (8)$$

3.2. Alleviate reverberation effects

While there exist reasonably good techniques to remove correlated noise as discussed above, no effective technique is available to remove reverberation. But it is possible to alleviate the reverberation effect to a certain extent. We next derive the maximum likelihood weighting function when reverberation presents.

If we consider reverberation as another type of noise, we have

$$|N_i^T(\omega)|^2 = |H_i(\omega)|^2|S(\omega)|^2 + |N_i(\omega)|^2 \quad (9)$$

where $|N_i^T(w)|^2$ represents the total noise. If we assume that the phase of $H_i(\omega)$ is random and independent of $S(\omega)$, then $E\{S(\omega)H_i(\omega)S^*(\omega)\}=0$, and, from (1), we have the following energy equation

$$|X_i(\omega)|^2 = a|S(\omega)|^2 + |H_i(\omega)|^2|S(\omega)|^2 + |N_i(\omega)|^2 \quad (10)$$

Both the reverberant signal and the direct-path signal are caused by the same source. The reverberant energy is therefore proportional to the direct-path energy, by a constant p:

$$|X_i(\omega)|^2 = a|S(\omega)|^2 + p|S(\omega)|^2 + |N_i(\omega)|^2$$
$$\Rightarrow \quad p|S(\omega)|^2 = p/(a+p)\times(|X_i(\omega)|^2 - |N_i(\omega)|^2) \quad (11)$$

The total noise is therefore:

$$|N_i^T(\omega)|^2 = p/(a+p)\times(|X_i(\omega)|^2 - |N_i(\omega)|^2) + |N_i(\omega)|^2$$
$$= q|X_i(\omega)|^2 + (1-q)|N_i(\omega)|^2 \quad (12)$$

where $q = p/(a+p)$. If we substitute (12) into (4), we have the ML weighting function for reverberant situation:

$$W_{MLR}(\omega) = \frac{|X_1(\omega)||X_2(\omega)|}{2q|X_1(\omega)|^2|X_2(\omega)|^2 + (1-q)|N_2(\omega)|^2|X_1(\omega)|^2 + |N_1(\omega)|^2|X_2(\omega)|^2} \quad (13)$$

To see the relationship between our derived $W_{MLR}(w)$ and the PictureTel one proposed in [10], we list the following approximations:

$$|\hat{G}_{x_1 x_2}(\omega)| = |X_1(\omega)|^2 = |X_2(\omega)|^2$$
$$|N(\omega)|^2 = |N_1(\omega)|^2 = |N_2(\omega)|^2 \quad (14)$$

With the above approximations, the PictureTel approach $W_{AMLR}(w)$ [10] approximates our proposed $W_{MLR}(w)$:

$$W_{AMLR}(\omega) = \frac{1}{q|\hat{G}_{x_1 x_2}(\omega)| + (1-q)|N(\omega)|^2} \quad (15)$$

There are several observations can be made based on $W_{MLR}(w)$ and $W_{AMLR}(w)$:

1. When the ambient noise dominates, they reduce to the traditional ML solution without reverberation $W_{TML}(w)$ (see (4)).

2. When the reverberation noise dominates, they reduce to $W_{PHAT}(w)$(see (5)). This agrees with the previous research that PHAT is robust to reverberation when there is no ambient noise [2].
3. Given the nature of $W_{TML}(w)$ (robust to ambient noise) and $W_{PHAT}(w)$ (robust to reverberation), $W_{MLR}(w)$ and $W_{AMLR}(w)$ can also be obtained by simply linearly combining the two basic weighting functions, hoping to obtain the benefits from the both worlds:

$$\frac{1}{W_{MLR}(\omega)} = q\frac{1}{W_{PHAT}(\omega)} + (1-q)\frac{1}{W_{TML}(\omega)} \quad (16)$$

We therefore can view $W_{MLR}(w)$ and $W_{AMLR}(w)$ as designed to simultaneously combat ambient noise and reverberation.

In practice, a precise estimation of q may be hard to obtain. Fortunately, the above observations allow us to design another weighting function heuristically, which performs almost as well as the optimum solution. Specifically, when the signal to noise ratio (SNR) is high, we choose $W_{PHAT}(w)$ and when SNR is low we choose $W_{TML}(w)$. We call this weighting function $W_{SWITCH}(w)$:

$$W_{SWITCH}(\omega) = \begin{cases} W_{PHAT}(\omega), & SNR > SNR_0 \\ W_{TML}(\omega), & SNR \le SNR_0 \end{cases} \quad (17)$$

where SNR_0 is a predetermined threshold, e.g., 15dB.

3.3. Putting the two stages together

If we put the various correlated noise removal techniques and different weighting functions in a 2D grid, we have the following table. It illustrates some of existing algorithms, as well as two of the proposed algorithms. Note that some of the existing algorithms also include further improvements, but fall generally in the category indicated.

Table 1. Different noise removal techniques and weighting functions.

	NR	GS	WF	WG
$W_{BASE}(w)$	[8]			
$W_{PHAT}(w)$	[2][3] [6]			
$W_{TML}(w)$	[2][7][12]			
$W_{SWITCH}(w)$				*
$W_{MLR}(w)$				*
$W_{AMLR}(w)$		[10]		

In Table 1, NR means no noise removal, and columns 3-5 correspond to the three techniques discussed in 3.1.1 to 3.1.3. $W_{BASE}(w)$ means the weighting function is a constant, i.e., $W_{BASE}(w) = 1$ for all frequencies. The symbol * represents proposed combinations that we observed can perform better than existing approaches, as shown in the next section.

4. EXPERIMENTAL RESULTS

We have done experiments on all the major combinations listed in Table 1. Furthermore, for the test data, we cover a wide range of sound source angles from -80 to +80 degrees. Detailed simulations results are available at our web site [13]. But due to limited space, here we report only three sets of experiments designed to compare different techniques on the following aspects:

1. For a uniform weighting function, which noise removal techniques is the best?
2. If we turn off the noise removal technique, which weighting function performs the best?
3. Overall, which algorithm (e.g., a particular cell in Table 1) is the best?

4.1. Test data description

We take into account both correlated noise and reverberation into account when generating our test data. We generated a plenitude of data using the imaging method [9]. The setup corresponds to a 6m×7m×2.5m room, with two microphones 15cm apart, 1m from the floor and 1m from the 6m wall (in relation to which they are centered). The absorption coefficient of the wall was computed to produce several reverberation times, but results are presented here only for T_{60} = 50ms. Furthermore, two noise sources were included: fan noise in the center of room ceiling, and computer noise in the left corner opposite to the microphones, at 50cm from the floor. The same room reverberation model was used to add reverberation to these noise signals, which were then added to the already reverberated desired signal. For more realistic results, fan noise and computer noise were actually acquired from a ceiling fan and from a computer. The desired signal is 60-second of normal speech, captured with a close talking microphone.

The sound source is generated for 4 different angles: 10, 30, 50, and 70 degrees, viewed from the center of the two microphones. The 4 sources are all 3m away from the microphone center. The SNRs are 0dB when both ambient noise and reverberation noise are considered. The sampling frequency is 44.1KHz, and frame size is 1024 samples (~23ms). We band pass the raw signal to 800Hz-4000Hz. Each of the 4 angle testing data is 60-second long. Out of the 60-second data, i.e., 2584 frames, about 500 frames are speech frames. The results reported in this section are obtained by using all the 500 frames.

There are 4 groups in each of the Figures 1-3, corresponding to ground truth angles at 10, 30, 50 and 70 degrees. Within each group, there are several vertical bars representing different techniques to be compared. The vertical axis in figures is error in degrees. The center of each bar represents the average estimated angle over the 500 frames. Close to zero means small estimation bias. The height of each bar represents 2x the standard deviation of the 500 estimates. Short bars indicate low variance. Note also that the fact that results are better for smaller angle is expected and intrinsic to the geometry of the problem.

4.2. Experiment 1: Correlated noise removal

Here, we fix the weighting function as $W_{BASE}(w)$ and compare the following four noise removal techniques : No Removal (NR), Gnn Subtraction (GS), Wiener Filtering (WF), and both WF and GS (WG). The results are summarized in Figure 1, and the following observations can be made:

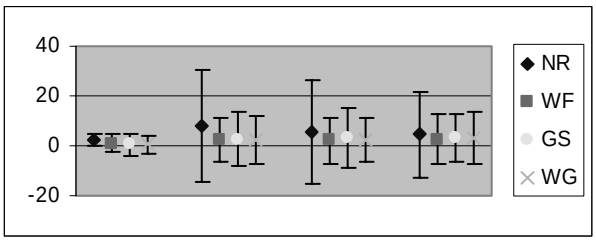

Figure 1. Compare NR, GS, WF and WG.

1. All the three correlated noise removal techniques are better than NR. They have smaller bias and smaller variance.
2. WG is slightly better than the other two techniques. This is especially true when the source angle is small.

4.3. Experiment 2: Alleviating reverberation effects

Here, we turn off the noise removal condition (i.e., NR in Table 1), and then compare the following 4 weighting functions: $W_{PHAT}(w)$, $W_{TML}(w)$, $W_{MLR}(w)$ ($q=0.3$), and $W_{SWITCH}(w)$. The results are summarized in Figure 2, and the following observations can be made:

1. Because the test data contains both correlated ambient noise and reverberation noise, the condition for $W_{PHAT}(w)$ is not satisfied. It therefore gives poor results, e.g., high bias at 10 degrees and high variance at 70 degrees.
2. Similarly, the condition for $W_{TML}(w)$ is not satisfied either, and it has high bias especially when the source angle is large.
3. Both $W_{MLR}(w)$ and $W_{SWITCH}(w)$ perform well, as they simultaneously model ambient noise and reverberation.

4.3. Experiment 3: Overall performance

Here, we are interested in the overall performance. Due to limited space, we report only two most promising techniques and compare them against the PictureTel approach [10], one of the best available. From the techniques involved, it is clear that $W_{MLR}(w)$-WG and $W_{SWITCH}(w)$-WG are the best candidates. The PictureTel technique [10] is $W_{AMLR}(w)$-GS when use our terminology (see Table 1). The results are summarized in Figure 3. The following observations can be made:

1. All the three algorithms perform well in general – all have small bias and small variance.
2. $W_{MLR}(w)$-WG seems to be the overall winning algorithm. It is more consistent than the other two. For example, $W_{SWITCH}(w)$-WG has big bias at 70 degrees and

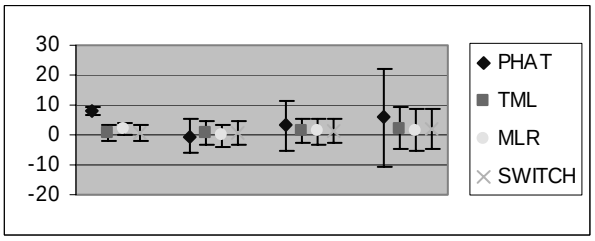

Figure 2. Compare $W_{PHAT}(w)$, $W_{TML}(w)$, $W_{MLR}(w)$, and $W_{SWITCH}(w)$.

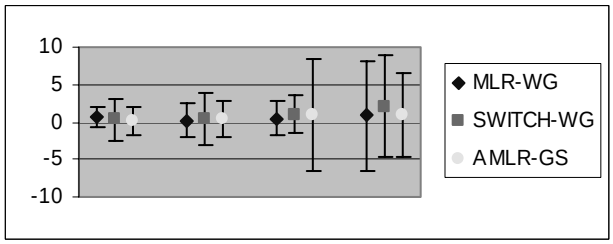

Figure 3. Compare $W_{MLR}(w)$-WG, $W_{SWITCH}(w)$-WG and $W_{AMLR}(w)$-GS.

$W_{AMLR}(w)$-GS has big variance at 50 degrees.

5. CONCLUSIONS

In this paper, we proposed a new two-stage perspective for estimating TDOA for real-world situations. The first stage concerns with correlated noise removal and the second stage tries to alleviate the reverberation effect. The new perspective allows us to develop a set of new approaches as well as to unify the existing ones. We have investigated a number of new combinations, and detailed experimental results are available at [13]. Two of the most promising ones are $W_{MLR}(w)$-WG and $W_{SWITCH}(w)$-WG. We also derived the ML weighting function for reverberant situation $W_{MLR}(w)$. It has nice physical interpretations as discussed in Section 3.2. The very successful PictureTel approach $W_{AMLR}(w)$ [10] is an approximation to our $W_{MLR}(w)$. We showed better performance of the new algorithms on realistically generated test data.

6. REFERENCES

[1]. S. Birchfield and D. Gillmor, Acoustic source direction by hemisphere sampling, *Proc. of ICASSP*, 2001.

[2]. M. Brandstein and H. Silverman, A practical methodology for speech localization with microphone arrays, Technical Report, Brown University, November 13, 1996

[3]. P. Georgiou, C. Kyriakakis and P. Tsakalides, Robust time delay estimation for sound source localization in noisy environments, *Proc. of WASPAA*, 1997

[4]. T. Gustafsson, B. Rao and M. Trivedi, Source localization in reverberant environments: performance bounds and ML estimation, *Proc. of ICASSP*, 2001.

[5]. Y. Huang, J. Benesty, and G. Elko, Passive acoustic source location for video camera steering, *Proc. of ICASSP*, 2000.

[6]. J. Kleban, Combined acoustic and visual processing for video conferencing systems, MS Thesis, The State University of New Jersey, Rutgers, 2000

[7]. C. Knapp and G. Carter, The generalized correlation method for estimation of time delay, *IEEE Trans. on ASSP*, Vol. 24, No. 4, Aug, 1976

[8]. D. Li and S. Levinson, Adaptive sound source localization by two microphones, *Proc. of Int. Conf. on Robotics and Automation*, Washington DC, May 2002

[9]. P.M.Peterson , Simulating the response of multiple microphones to a single acoustic source in a reverberant room," *J. Acoust. Soc. Amer.*, vol. 80, pp1527-1529, Nov. 1986.

[10]. H. Wang and P. Chu, Voice source localization for automatic camera pointing system in videoconferencing, *Proc. of ICASSP*, 1997

[11]. D. Ward and R. Williamson, Particle filter beamforming for acoustic source localization in a reverberant environment, *Proc. of ICASSP*, 2002.

[12]. D. Zotkin, R. Duraiswami, L. Davis, and I. Haritaoglu, An audio-video front-end for multimedia applications, *Proc. SMC* , Nashville, TN, 2000

[13]. http://www.research.microsoft.com/~yongrui/html/TDOA.html