A Sound Localization Algorithm for Use in Unmanned Vehicles

Justin A. MacDonald & Phuong K. Tran

Human Research and Engineering Directorate
U. S. Army Research Laboratory
Aberdeen Proving Ground, MD 21005
jmacdonald@arl.army.mil
ptran@arl.army.mil

Abstract

This paper details the results of an effort to develop a computer-based algorithm for sound localization for use in unmanned vehicles. The algorithm takes input from two or more microphones and estimates the position of the sound source relative to the microphone array. A priori knowledge of the stimulus is not required. The algorithm takes advantage of time-of-arrival and frequency cues to estimate the location of the sound source. The performance of two- and four-microphone implementations of the algorithm was measured using recordings from microphones mounted at several points around the head of an acoustic mannequin. Sounds were played at 5 degree intervals around the mannequin and the outputs were recorded. These recordings were fed into the algorithm that estimated the location of the sound source. Both algorithm implementations were able to identify accurately the location of a variety of real-world, broadband sounds, committing less than 2 degrees of unsigned localization error in moderately noisy environments. The four-microphone implementation of the algorithm was shown to be more resistant to background noise, committing less than 3 degrees of unsigned error when the signal-to-noise ratio was -10 dB or better. Future directions for algorithm development as well as potential applications are discussed.

Introduction

The development of a sound localization algorithm in our laboratory was motivated by the need for a model of human sound localization to predict the effects of protective headgear upon localization accuracy. The effort to develop a model of human performance first required the development of an algorithm that produces location estimates based upon the sounds received by the ears. As human listeners are quite able to localize unfamiliar sounds, the algorithm must provide accurate location estimates without any previous exposure to the stimulus. Humans make use of the Interaural Time Delay (ITD), Interaural Level Difference (ILD), and pinna and torso effects to estimate the location of a sound source. Preferably, the localization algorithm should make use of these cues as well.

Unfortunately, a search of the literature failed to uncover a suitable localization algorithm. Most rely exclusively upon time-of-arrival differences between the sensors to estimate location. For example, Berdugo et al. (1999) estimated the location of a 20-s speech stimulus using an array of seven microphones. They reported a mean unsigned error of approximately 5 degrees. While the accuracy of this system is adequate (at least with stimuli that are long in duration), a large number of sensors are required, and location cues other than time-of-arrival differences between sensors in the array are not utilized. Two-sensor algorithms do exist, although the sole reliance on time delay information remains. Viera and Almeida (2003) constructed a two-sensor system that localized sound sources between +60 and -60 degrees azimuth and zero degrees elevation. Mean unsigned error was reported as 9 degrees despite the restricted domain of potential locations. Performance would likely have been considerably worse had locations in the rear hemisphere been included in the testing of the algorithm. Any two-sensor system that relies exclusively upon time delay cues will suffer from frequent front/back confusions. A time delay measured between the sensors will identify a subset of potential locations that lie upon the surface of a cone extending outward from one of the sensors (Blauert, 1989). The subset of potential locations can be further reduced in two ways: by restricting the range of potential locations (the approach taken by Viera and Almeida) or by collecting additional information. The former approach is less desirable, as restricting the domain of potential sound source locations limits the applicability of the algorithm. However, collecting additional information can maximize the accuracy and applicability of the algorithm.

Rather than increasing the number of sensors to disambiguate the time delay cues to location, we chose to make use of frequency-based location cues, much like the human listener. Of course, this requires that the frequency content of sounds that reach the sensors vary systematically with the location of the sound source. This can be accomplished by inserting an obstruction between the sensors. In our initial effort, we chose to mount the microphones in the ear canals of the Knowles Electronics Mannequin for Acoustic Research (KEMAR). The KEMAR is a human model that mimics the effects of the

head and torso upon an incoming sound wave, and therefore was appropriate for use in our effort to develop a model of human sound localization. The catalogue of effects of the head and torso upon incoming sounds makes up the Head-Related Transfer Function (HRTF; see [5]) of the KEMAR, and includes both time delay and frequency cues to location. These frequency differences can be used to eliminate the front/back ambiguity inherent in a time-delay-based two-sensor system.

Of course, there is nothing particular about a time- and frequency-based localization algorithm that necessitates its use as part of a model of human sound localization. Based on initial tests that indicated the potential of such an algorithm, we chose to focus on algorithm development independent of our human performance modeling efforts, instead working to maximize the accuracy of the algorithm. We have previously reported the results of a series of simulations using a two-sensor version of the algorithm to estimate the locations of real-world stimuli in a variety of noise environments (MacDonald, 2005). The current research involves an investigation of the utility of expanding the algorithm to four sensors. Although the KEMAR was used to introduce frequency cues to location in our simulations, it is not integral to the performance of the algorithm. In fact, any object that introduces differences in frequency content between sensors is sufficient for the algorithm to function.

Localization Algorithms

The Inverse Algorithm

Consider a KEMAR with microphones mounted at the entrance to each ear canal. Imagine that a sound originates directly adjacent to the right ear of the KEMAR (at +90°) on the horizontal plane. The waveform that reaches each of the microphones will be subject to the transfer function (the HRTF) of the KEMAR. These effects are specific to the location of the sound source (+90° azimuth and 0° elevation, in this example) and can be represented as a pair of Finite Impulse Response (FIR) digital filters $H_{Left}^{(+90,0)}$ and $H_{Right}^{(+90,0)}$, one for each microphone. Let R_{Left} and R_{Right} be digital recordings of the sound obtained at the left and right microphones, respectively. R_{Right} could be filtered by the inverse of $H_{Right}^{(+90,0)}$ to obtain the original sound before its alteration by the head and torso. Filtering R_{Left} by the inverse of $H_{Left}^{(+90,0)}$ would result in a copy of the original unaltered stimulus identical to that obtained from the right ear. However, if R_{Left} and R_{Right} were not filtered by the above functions but instead by the inverse of the FIR filter associated with some other location, the original stimulus would not result in either case. In fact, this operation would lead to considerably different waveforms.

This simple idea suggests a method by which the inverse of the HRTF could be used to estimate the location of sound sources. Consider a sound that originates from azimuth θ and elevation ϕ in relation to a point *P*, where $-180° < \theta, \phi \le +180°$. Imagine that the center of the head of the KEMAR is at *P*. The task of the inverse localizer is to provide the estimates $\hat{\theta}$ and $\hat{\phi}$ based upon the recordings R_{Left} and R_{Right}. In this case, the inverse localizer estimates the location of the sound source as follows:

$$\min_{\hat{\theta},\hat{\phi}} \sum \left(R_{Left} * \left[H_{Left}^{(\hat{\theta},\hat{\phi})} \right]^{-1} - R_{Right} * \left[H_{Right}^{(\hat{\theta},\hat{\phi})} \right]^{-1} \right)^2, \qquad (1)$$

where $*$ is the convolution operator and $\left[H_{Left}^{(\hat{\theta},\hat{\phi})} \right]^{-1}$ and $\left[H_{Right}^{(\hat{\theta},\hat{\phi})} \right]^{-1}$ are the inverses of $H_{Left}^{(\hat{\theta},\hat{\phi})}$ and $H_{Right}^{(\hat{\theta},\hat{\phi})}$, respectively.

In practice, this algorithm has proved difficult to implement for several reasons. First, constructing an inverse FIR filter can be problematic. FIR filters implement both time delay and frequency-related effects, and both must be accounted for when constructing an appropriate inverse. Second, appropriate inverse filters that account for the magnitude and phase portions of the transfer function typically require considerable computational resources to implement. Third, methods to compute inverse filters produce filters that are only an approximate inverse of the original (Rife & Vanderkooy, 1989). Note that the accuracy of the inverse increases with its length. None of these problems are insurmountable, of course, but our initial efforts were focused upon the testing of a variant of the inverse algorithm.

The Cross-Channel Algorithm

As before, the task of the localization algorithm is to provide the estimates $\hat{\theta}$ and $\hat{\phi}$ based upon the recordings R_{Left} and R_{Right}. The cross-channel localizer chooses $\left(\hat{\theta},\hat{\phi}\right)$ as follows:

$$\min_{\hat{\theta},\hat{\phi}} \sum \left(R_{Left} * H_{Right}^{(\hat{\theta},\hat{\phi})} - R_{Right} * H_{Left}^{(\hat{\theta},\hat{\phi})} \right)^2 \qquad (2)$$

In words, the cross-channel localizer filters each of the recordings by the transfer function associated with the opposite microphone and compares the resulting waveforms. To understand the reasoning behind this method, let $O^{(\theta,\phi)}$ be a digital recording of the sound originating from (θ,ϕ) recorded at *P* with the listener absent. Then $R_{Left} \approx O^{(\theta,\phi)} * H_{Left}^{(\theta,\phi)}$, and

$R_{Right} \approx O^{(\theta,\phi)} * H_{Right}^{(\theta,\phi)}$. If R_{Left} is convolved with the transfer function associated with the right microphone, then

$$\begin{aligned} R_{Left} * H_{Right}^{(\theta,\phi)} &\approx \left(O^{(\theta,\phi)} * H_{Left}^{(\theta,\phi)}\right) * H_{Right}^{(\theta,\phi)} \\ &= \left(O^{(\theta,\phi)} * H_{Right}^{(\theta,\phi)}\right) * H_{Left}^{(\theta,\phi)} \approx R_{Right} * H_{Left}^{(\theta,\phi)} \end{aligned} \quad (3)$$

This substitution follows from the commutability and transitivity of the convolution operator. The cross-channel localizer takes advantage of this relation by choosing the values of $\left(\hat{\theta},\hat{\phi}\right)$ so that the squared differences between the leftmost and rightmost terms in Equation 3 are minimized.

This algorithm is easily generalizable to more than two microphones. The algorithm requires *2N* microphones that can be arbitrarily divided into *N* pairs. The multi-channel version of Equation 2 becomes:

$$\min_{\hat{\theta},\hat{\phi}} \sum_{i=1}^{N} \sum \left(R_{i_1} * H_{i_2}^{(\hat{\theta},\hat{\phi})} - R_{i_2} * H_{i_1}^{(\hat{\theta},\hat{\phi})}\right)^2, \quad (4)$$

where i_1 and i_2 are the elements of pair *i.*

The cross-channel algorithm has a large advantage over the inverse algorithm: inverse filters are not required. Adequately accurate inverse filters are approximately three times longer than the FIR filters utilized in the cross-channel algorithm (Greenfield & Hawksford, 1991). Thus the cross-channel algorithm permits considerable reduction in the computational resources required. These practical reasons rather than any theoretical considerations prompted us to pursue the cross-channel algorithm.

Our initial test of the cross-channel algorithm examined the performance of a two-microphone implementation (MacDonald, 2005). Real-world, broadband sounds were recorded at 5° intervals around the head of the KEMAR. Noise was added to each recording to obtain signal-to-noise ratios (SNRs) from 40 to -40 dB, and the cross-channel algorithm estimated the location of the sound source from the noisy recordings. The algorithm performed well beyond expectations: localization error in quiet was measured at 2.9° using only two microphones, and above-chance performance was observed at greater than or equal to -10 dB SNRs. Front/back reversals occurred in approximately 5% of trials at the higher SNRs.

The promising results of the two-sensor implementation of the algorithm prompted us to conduct additional tests using four sensors. Accordingly, two additional microphones were mounted on the front and rear of the head of the KEMAR, and additional tests were conducted using four microphones. The additional microphones should allow for a reduced number of front/back confusions and increased localization accuracy at the expense of increased computation time. Results from both the two- and four-sensor algorithms were obtained to determine the increased accuracy gained from the addition of two microphones.

Simulation Method

Stimuli

Four naturally-occurring sounds were chosen as the test signals. They included the sounds of breaking glass, the insertion of an M-16 magazine, a speech stimulus, and a camera shutter release sound. Sounds ranged from 400 to 600 ms in duration and were stored in a 16-bit Microsoft WAV format with a sampling rate of 44.1 kHz.

Stimulus Recording Apparatus

Stimuli were presented using a Tucker-Davis Technologies (TDT) RoboArm 360 system. This system consists of a speaker attached to a computer-controlled robotic arm. The stimuli were output through a TDT System II DD1 D/A converter and amplified using a TDT System 3 SA1 amplifier. Stimuli were presented through a GF0876 loudspeaker (CUI, Inc.) mounted at the end of the robotic arm and positioned 1m from the center of the head of the KEMAR. Stimuli were presented at approximately 75 dB (A) measured at this point with the KEMAR absent. The arm positioned the loudspeaker at 5° intervals around the KEMAR (a total of 72 positions). All sounds were located at 0° elevation.

Two EM-125 miniature electret microphones (Primo Microphones, Inc.) were used to record the stimulus presentations. Recordings were made in two sessions. In the first, the pair of microphones was mounted in foam inserts at the entrance of the ear canals of the KEMAR. In the second, the same microphones were moved to the front and rear of the head of the KEMAR. The front microphone was attached to the center of the forehead just above the bridge of the nose, and the rear microphone was attached at the center of the back of the head at the same elevation as the front. Inputs to the microphones were amplified by a TDT System 3 MA3 microphone amplifier before being sent to a TDT System II DD1 A/D converter. The digital output of the DD1 was sent to a computer for storage in a 44.1 kHz, 16-bit Microsoft WAV format. Combining across recording sessions, a total of 288 four-channel recordings were made, one for each position/sound combination.

Transfer Function Measurement

The HRTF of the KEMAR was measured using the same presentation and recording apparatus detailed above. Maximum-Length Sequence (see Rife & Vanderkooy, 1989) stimuli were presented at 5° intervals around the head of the KEMAR and the resulting waveforms determined the HRTF of the KEMAR at each location. As with the stimulus recordings, the front/back and left/right transfer functions were recorded separately. Each HRTF was stored as a 256-tap FIR filter.

Simulation Task

Simulations were conducted to estimate the performance of both the two- and four-sensor versions of the cross-channel algorithm. In the two-sensor simulation, the algorithm utilized the HRTFs associated with the left and right microphones to process the recordings made at those locations. Estimates were produced using a version of Equation 2 that was implemented in Microsoft Visual C++. The four-sensor simulation used a version of Equation 4 to apply the four-channel HRTFs to the four-channel recordings. Locations were estimated to 5° precision in both simulations. A random sample of Gaussian noise was added to each channel of each recording to obtain SNRs ranging from 40 to -50 dB in 10 dB increments. The algorithm was required to localize each recording five times; a different sample of Gaussian noise was added on each attempt. This resulted in a total of 14,440 localization attempts for each of the simulations (10 S/N ratios X 288 recordings X 5 trials each).

Results

The mean localization errors observed at each SNR are shown in Figure 1. Chance-level performance is indicated by the dotted line located at 90 degrees. The two-microphone implementation performed as expected, exhibiting less than two degrees localization error when the SNR was greater than 10 dB, and performing well above chance levels to -10 dB. The four-microphone implementation exhibited even greater accuracy, maintaining a mean error of less than three degrees in SNRs of -10 dB and greater.

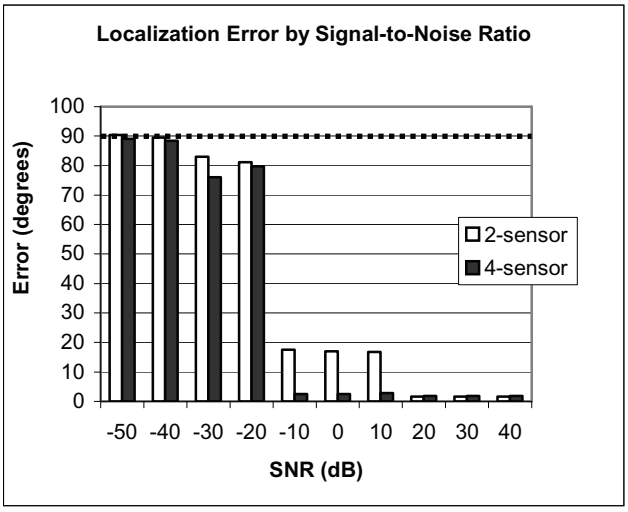

Figure 1

The proportion of front/back reversals in each SNR condition is shown in Figure 2. Chance-level performance is indicated by the dotted line located at 0.5. A front/back reversal occurred when the estimated and actual locations of the sound source were on opposite sides of the left/right axis. Two-microphone systems that rely exclusively on time-of-arrival differences will exhibit a 50% reversal rate. As expected, the cross-channel algorithm exhibited a minimal proportion of reversals. In fact, the overall reversal rate was very low in both implementations of the algorithm: performance was considerably above chance levels for all SNRs greater than -20 dB. As expected, the addition of two additional microphones in the four-sensor implementation led to increased performance: reversals were reduced to trivial levels in the 10, 0, and -10 dB conditions.

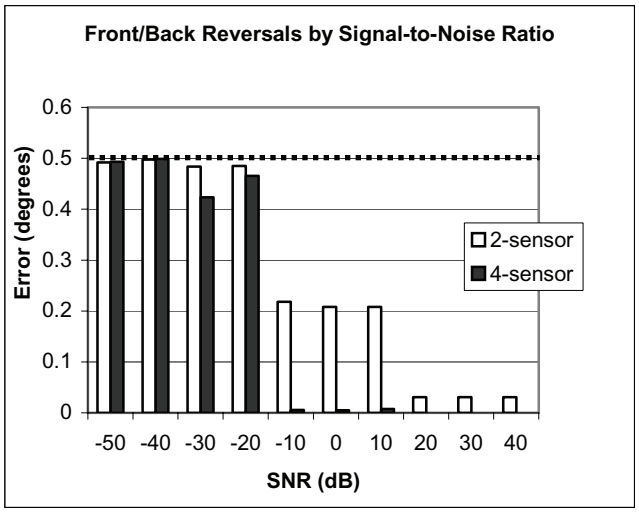

Figure 2

Discussion

These simulations demonstrate the extremely high accuracy that can be achieved with the cross-channel algorithm. The two-microphone implementation exhibited a mean localization error of less than two degrees despite the addition of a moderate amount of Gaussian noise. The performance of this algorithm in noise is superior to that of many other algorithms (e.g., Berdugo et al., 1999; Viera & Almeida, 2003) in quiet environments. The accuracy of the algorithm is especially impressive considering that sounds were allowed to originate from the rear hemisphere, thereby allowing for the possibility of front/back reversals. The inclusion of frequency-based location cues allows for a severe reduction in the number of front/back reversals.

As expected, the four-microphone implementation of the algorithm exhibited even better performance, committing almost no reversals in all SNRs greater than -20 dB. An evaluation of the four-microphone algorithm must consider the tradeoff between improved accuracy and increased computational requirements. The two- and four-microphone implementations performed similarly at SNRs of greater than 10 dB. The improved performance of the four-channel algorithm is not readily apparent until the SNR is decreased to 10 dB or lower. Which algorithm is chosen for a given application, therefore, depends on the

level of noise likely to be encountered. The two-channel algorithm functions admirably in relatively quiet environments, while the four channel version performs well even in moderately noisy environments. Many possible compromises between the two- and four-microphone algorithm implementations are worth investigating as well. For example, the two-channel algorithm could generate location estimates that are modified based upon the relative intensity of the input to the front and back microphones. Variations such as these will be explored in future work.

It is clear that several questions remain to be answered about the performance of the algorithm. As with all localization algorithms, performance is likely to decrease in reverberant environments. In addition, the performance of the algorithm is unknown when the elevation of the sound source is allowed to vary. It seems likely that the accuracy of elevation judgments would improve with the four-microphone version of the algorithm, but that remains to be investigated. In addition, the location of the front and back microphones may need to be optimized.

Despite these unanswered questions, the potential applications of this highly accurate algorithm are many. The algorithm could be implemented as a human localization aid: its performance is far superior to that of human listeners, especially in noisy environments. The algorithm is also ideal for use in unmanned ground vehicles or autonomous robots, as it can localize very accurately in the presence of background noise. It would function well in tandem with an auditory streaming algorithm to generate an accurate representation of the robot's surrounding environment. Microphones could be mounted at various points around the robot, and the computations are easily handled by a laptop computer. In fact, the algorithm is likely to benefit from increased distances between the microphones, as the time- and frequency-based cues to location will be enhanced. Future work will demonstrate the utility of the cross-channel algorithm by integrating it into a small, autonomous tracking vehicle.

References

Berdugo, B., Doron, M. A., Rosenhouse, J., & Azhari, H. 1999. On direction finding of an emitting source from time delays. *Journal of the Acoustical Society of America, 105,* 3355-3363.

Blauert, J. 1989. *Spatial Hearing.* MIT Press, Cambridge, MA.

Greenfield, R. and Hawksford, M. O. 1991. Efficient filter design for loudspeaker equalization. *Journal of the Audio Engineering Society, 39,* 739-751.

MacDonald, J. A. 2005. An algorithm for the accurate localization of sounds. *Proceedings of the NATO HFM-123 Symposium on New Directions for Improving Audio Effectiveness,* Paper P-28.

Rife, D. D. and Vanderkooy, J. 1989. Transfer-function measurement with Maximum-Length Sequences. *Journal of the Audio Engineering Society, 37,* 419-444.

Viera, J. and Almeida, L. 2003. A sound localizer robust to reverberation. *Proceedings of the 115th Convention of the Audio Engineers Society,* Paper 5973.

Wightman, F. L. & Kistler, D. J. 1989. Headphone simulation of free-field listening. I: Stimulus synthesis. *Journal of the Acoustical Society of America, 85,* 858-867.