An Algorithm for the Accurate Localization of Sounds

Justin A. MacDonald
Army Research Laboratory
Human Research and Engineering Directorate
Attn: AMSRD-ARL-HR-SD
Aberdeen Proving Ground, MD 21005
USA

jmacdonald@arl.army.mil

ABSTRACT

A computer-based algorithm that localizes sounds in near-real time has been developed. The algorithm takes input from two microphones and estimates the position of the sound source relative to the microphone array. The algorithm requires no a priori knowledge of the stimuli to be localized. The accuracy of the algorithm was tested using binaural recordings from a pair of microphones mounted in the ear canals of an acoustic mannequin. Sounds were played at 5 degree steps around the mannequin and the outputs were recorded at the entrance to each ear canal. These recordings were fed into the algorithm that estimated the location of the incoming sound on the horizontal plane. The algorithm utilizes a Head-Related Transfer Function (HRTF) to estimate the location of incoming sound stimuli. Although the HRTF of the acoustic mannequin was used, any HRTF can be inserted into the algorithm, allowing for predictions about individual performance differences. The results of this effort have been highly encouraging: the algorithm was able to identify accurately the location of a variety of sounds, committing an average of 2.9 degrees of unsigned localization error. Better than chance performance was found in noisy conditions of up to a -10 dB signal-to-noise ratio. The initial purpose of this algorithm is to predict the localization performance afforded by different types of combat helmets. Current and future encapsulating helmet designs are likely to impede localization performance, and an accurate localization model would be an invaluable tool in the helmet selection process. Adapting the model for use as a highly accurate machine-based localizer is an additional goal of this line of research. Applications for this technology include target tracking on unmanned vehicles, sniper detection, and machine-assisted sound localization.

1.0 INTRODUCTION

Many protective military ensembles include partially- or fully-encapsulating headgear which has a negative effect on sound detection, localization, and auditory distance estimation. This problem has prompted the Army to search for a model of human sound localization that can predict the effects of different headgear ensembles on the auditory localization ability of the soldier.

Constructing an accurate model of human sound localization is a complex task. When the listener estimates the location of an incoming sound many factors other than the stimulus itself can affect the judgment. Listeners are likely to incorporate visual information into their interpretation of the sound, which can alter the listener's perception of the auditory spatial information. Moreover, familiarity with the environment and prior knowledge of the type of sound and its amplitude and frequency characteristics can help to improve the

MacDonald, J.A. (2005) An Algorithm for the Accurate Localization of Sounds. In *New Directions for Improving Audio Effectiveness* (pp. 28-1 – 28-10). Meeting Proceedings RTO-MP-HFM-123, Paper 28. Neuilly-sur-Seine, France: RTO. Available from: http://www.rto.nato.int/abstracts.aps.

localization performance of the listener. In addition to a relatively accurate method to estimate the location of sound sources, any predictive model of human performance should incorporate both a learning mechanism as well as a component to model the interaction between visual and auditory information about the possible location of a sound source.

The initial phase of model construction has focused upon the development of an algorithm to produce location estimates from inputs to each of the ears. Development of this algorithm serves two purposes. First, the algorithm is intended to function as the basis for a more complete model of human sound localization that mirrors processes used by human listeners to produce location judgments. Second, the algorithm can be used as the basis of a computer-based "sound localizer" aiding the user in the accurate determination of sound source position. Such a system would have a number of commercial and military applications, such as inclusion into the navigation system for autonomous robotic vehicles or as an aid to the soldier to improve sound localization on the battlefield. Such a system would benefit from a fast and highly accurate localizing algorithm, ideally considerably more accurate than humans. Therefore, the initial goal of this project was to develop a localization algorithm that maximizes localization accuracy while minimizing computational requirements. The algorithm should maintain reasonable functionality in noisy environments, and provide accurate location estimates without utilizing any prior knowledge of the sound stimulus including its typical level and frequency characteristics.

1.1 Development of the Localization Algorithm

Several previous attempts have been made to develop sound localization algorithms. Berdugo, Doron, Rosenhouse, and Azhari [1] utilized an array of seven microphones to estimate the azimuth and elevation of a 20-second speech stimulus played from a loudspeaker in a quiet environment. Estimates were computed using the differences in time-of-arrival observed at each of the seven sensors. The mean unsigned localization error was found to be approximately 5°. While this system was found to be relatively accurate, it is less than ideal for the current application. First, testing of the algorithm was conducted using stimuli that were 20 seconds in duration, which should provide the algorithm with considerable data with which to make a localization estimate. Considering that humans are quite able to localize sounds that are less than 500 ms, any algorithm used as a basis for a model of human sound localization must generate accurate estimates using considerably less data. Secondly, the use of seven sensors in the localization system is clearly inappropriate for modelling a two-sensor system (the human). Viera and Almeida [2] constructed a localization system with two sensors that estimated the location of the sound based upon the time delay between the sensors. Testing required the system to localize sounds located at 0° elevation and between +60 and -60 degrees azimuth. The system averaged approximately 9° of unsigned localization error. While this algorithm utilizes only two sensors and is therefore appropriate as a model of human localization, it is less accurate than a human observer. Moreover, any two-sensor system that localizes sounds based solely upon the time delay will be susceptible to frequent front/back confusions. Any given time delay between the ears will restrict the subset of possible locations to the set of points that lie upon the surface of a cone extending outward from one of the sensors [3]. Even if source locations are restricted to those on the horizontal plane and the source distance is held constant, time delay information will not uniquely determine the location of the stimulus. Additional information about the frequency characteristics of the incoming sound must be utilized to resolve this ambiguity. Fortunately, the head and torso of the listener provide these cues: the effect of the head and torso upon an incoming sound wave depends upon the direction of the waveform. The catalogue of effects of the head and torso upon incoming sound sources makes up the Head-Related Transfer Function (HRTF; see [3], [4]) of the listener. These frequency differences can be used to eliminate the front/back ambiguity.

For this reason the localization algorithm developed by the author was based upon the HRTF, which incorporates both time delay and frequency cues to the location of a sound source. The algorithm was implemented using the HRTF of the Knowles Electronics Mannequin for Acoustics Research (KEMAR), although the HRTF of any listener could have been used. Two possible methodologies were considered during the development of the localization algorithm: the Inverse Localizer and the Cross-Channel Localizer. Both utilize the HRTF of the KEMAR and require no previous knowledge of the sound stimulus to be localized.

1.2 The Inverse Localizer

Consider a sound that originates directly in front of the right ear (at +90°) on the horizontal plane. The waveform that reaches each of the ears will be subject to the head and torso effects summarized as the HRTF when the sound originates at +90° azimuth and 0° elevation. Assume that the HRTF at +90° is represented in the time domain as a pair of Finite Impulse Response (FIR) digital filters, denoted by $H_{Left}^{(+90,0)}$ and $H_{Right}^{(+90,0)}$. Suppose that microphones were placed in each of the ears to record the incoming waveform. Let R_{Left} and R_{Right} be digital recordings of the sound obtained at the left and right ears, 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 $H_{Left}^{(-90,0)}$ and $H_{Right}^{(-90,0)}$, respectively, 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 direction θ and elevation ϕ in relation to a point *P*, where $-180° < \theta, \phi \leq +180°$. Imagine that the centre of the head of the listener 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:

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

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.

This procedure is not without its difficulties, however, most of which arise when computing the inverses of $H_{Left}^{(\hat{\theta},\hat{\phi})}$ and $H_{Right}^{(\hat{\theta},\hat{\phi})}$. These filters include both the time delay and frequency effects caused by the head. Time delays are implemented as differences in the onsets of the filters for the left and right ears, and inverting the filters requires compensating for this delay. More importantly, the inversion of an FIR filter can lead to unpredictable results. Constructing the minimum-phase inverse results in a filter whose magnitude is only an approximate inverse of the original [5]. In addition, the phase response must also be inverted; this can be accomplished using the procedure outlined in Greenfield & Hawksford [6]. This procedure yields an inverse filter that is roughly three times larger than the original. For these reasons, the development of the inverse localizer was postponed in favour of another methodology that does not require the calculation of inverse filters.

1.3 The Cross-Channel Localizer

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:

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

In words, the cross-channel localizer filters each of the recordings by the HRTF of the opposite ear 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 the recording from the left ear is convolved with the HRTF of the right ear, then

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

This 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 method is preferable to the inverse localizer for several reasons: first, the construction of inverse filters is unnecessary; only the original HRTF is required to localize sounds. Second, the FIR filter that represents the HRTF is approximately one-third the size of an accurate inverse filter. Minimizing the size of the filters used in convolution will have a tremendous impact on the computational resources required. Accordingly, an initial implementation of the cross-channel localizer was developed and subsequently tested in signal-to-noise (S/N) ratios between 40 and -40 dB to assess its performance in noisy environments. As this is only a preliminary test of the model, the positions of all sound sources were limited to the horizontal plane.

2.0 EXPERIMENT METHOD

2.1 Stimuli. Eight naturally-occurring sounds were chosen as the test signals. They included the sounds of breaking glass, a barking dog, a closing door, the insertion of an M-16 magazine, a cough, dripping water, a camera in operation, and machine gun fire. 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.

2.2 Stimulus presentation apparatus. Stimuli were presented using the Army Research Laboratory Human Research and Engineering Directorate's RoboArm 360 system (see Figure 1 below). This system consists of a speaker attached to a computer-controlled robotic arm. The stimuli were output through a Tucker-Davis Technologies (TDT) System II DD1 D/A converter, amplified using a TDT System 3 SA1 amplifier, and presented from a GF0876 loudspeaker (CUI, Inc.) at the end of the robotic arm. Stimuli were presented at approximately 75 dB (A) measured one meter from the loudspeaker. The arm positioned the loudspeaker at 5° intervals around the KEMAR (a total of 72 positions). The loudspeaker was located 1 meter from the centre of the head of the KEMAR and at 0° elevation for all stimulus presentations.

2.3 Recording apparatus. EM-125 miniature electret microphones (Primo Microphones, Inc.) were mounted in foam inserts in the artificial pinnae of the KEMAR. 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. A total of 576 recordings were made, one for each position/sound combination. These WAV files served as the sounds to be localized by the cross-channel localizer. Each recording was filtered to remove the frequency effects introduced by the recording system.

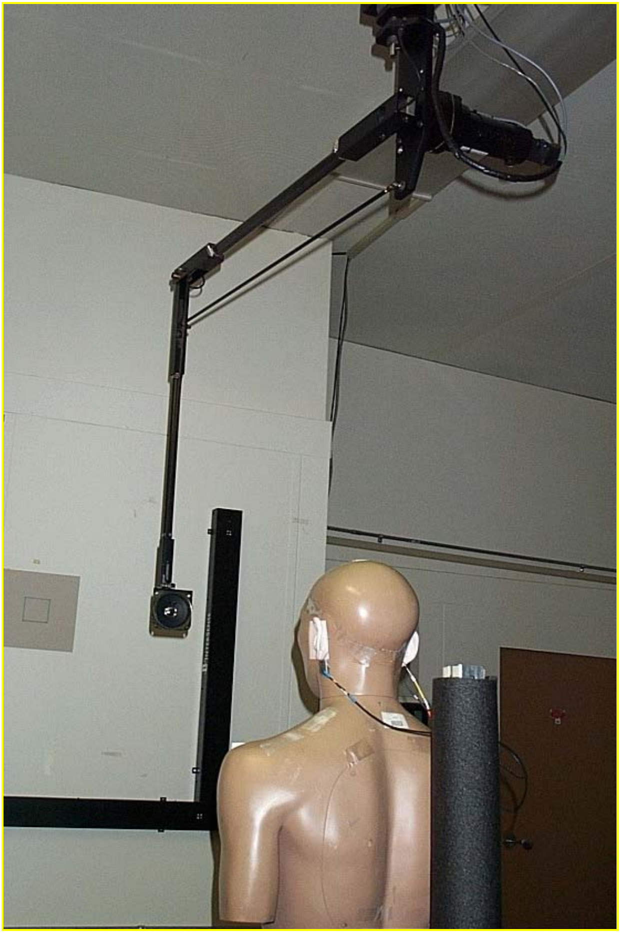

Figure 1 – The RoboArm 360 system. The computer-controlled robotic arm can place the loudspeaker to 1° precision.

2.4 HRTF measurement. The HRTF of the KEMAR was measured using the same presentation and recording apparatus detailed above. Maximum-Length Sequence (MLS; see [8]) stimuli were presented at 5° intervals around the head of the KEMAR and the resulting waveforms were recorded at each of the ears.

These recordings were digitized and processed to determine the HRTF of the KEMAR at each location. Each HRTF was stored as a 256-tap FIR filter and the frequency effects introduced by the recording system were removed.

2.5 Localization task. The cross-channel localizer applied the measured HRTF of the KEMAR to the binaural recordings to estimate the location of each sound to 5° precision. The algorithm estimated the location of the sound source using a version of Equation 2 that was implemented in MATLAB. A random sample of Gaussian noise was added to each recording to obtain signal-to-noise ratios ranging from 40 to -40 dB in 10 dB increments. The localizer was required to localize each recording 10 times; a different sample of Gaussian noise was added on each attempt. This resulted in a total of 51,840 localization attempts (9 S/N ratios X 576 recordings X 10 trials each).

3.0 RESULTS

The mean unsigned error for each S/N ratio is shown in Figure 2. The means were obtained by averaging over all stimuli and locations within a given S/N ratio. Not surprisingly, mean error increased as the S/N ratio decreased. The cross-channel localizer maintained an error rate of less than 5° for S/N ratios of 40, 30, and 20 dB, and even performed well above chance performance at -10 dB.

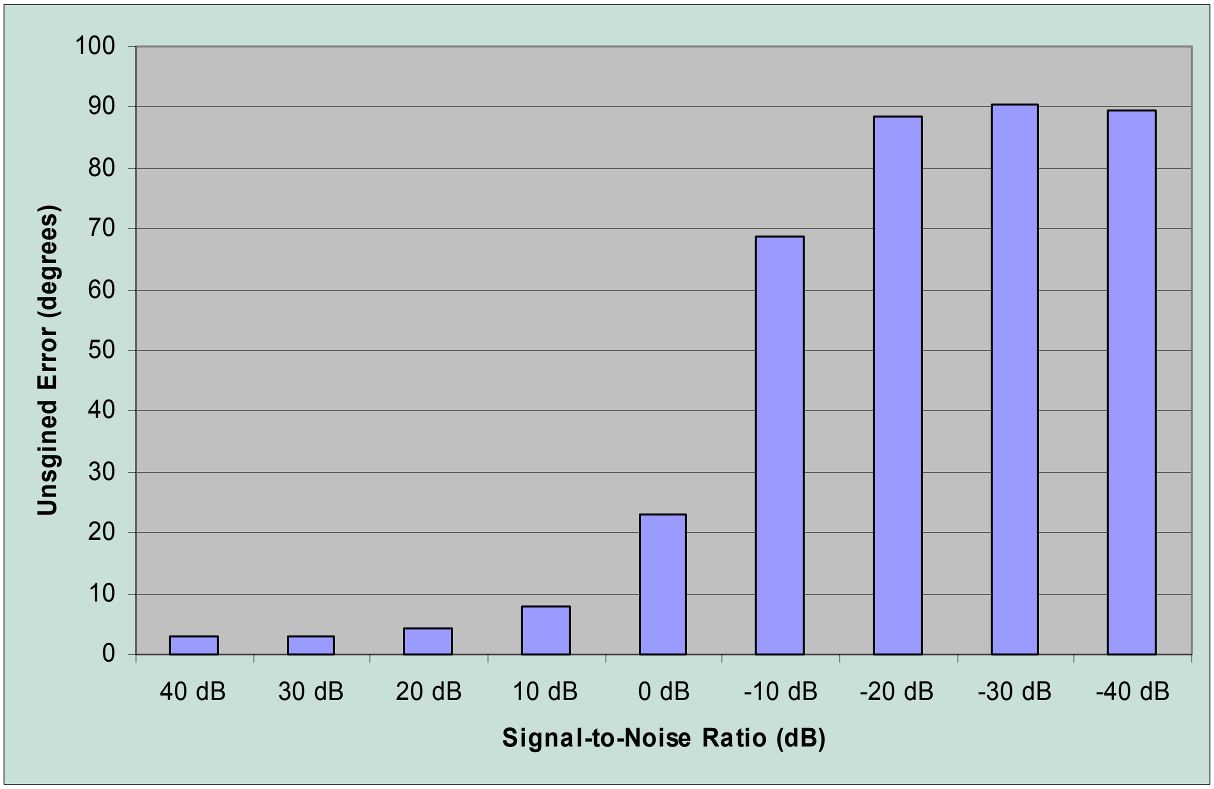

Figure 2 – Mean Unsigned Error by S/N Ratio

The proportions of front/back reversals are shown in Figure 3. For the purposes of this experiment, a reversal occurred when the estimated and actual locations of a sound were on opposite sides of the interaural axis. Sounds located at +90° and -90° were omitted from this analysis for obvious reasons. The localizer correctly resolved the front/back ambiguity for more than 95% of trials at 40 and 30 dB. Performance did not deteriorate to chance level until the S/N ratio reached -20 dB.

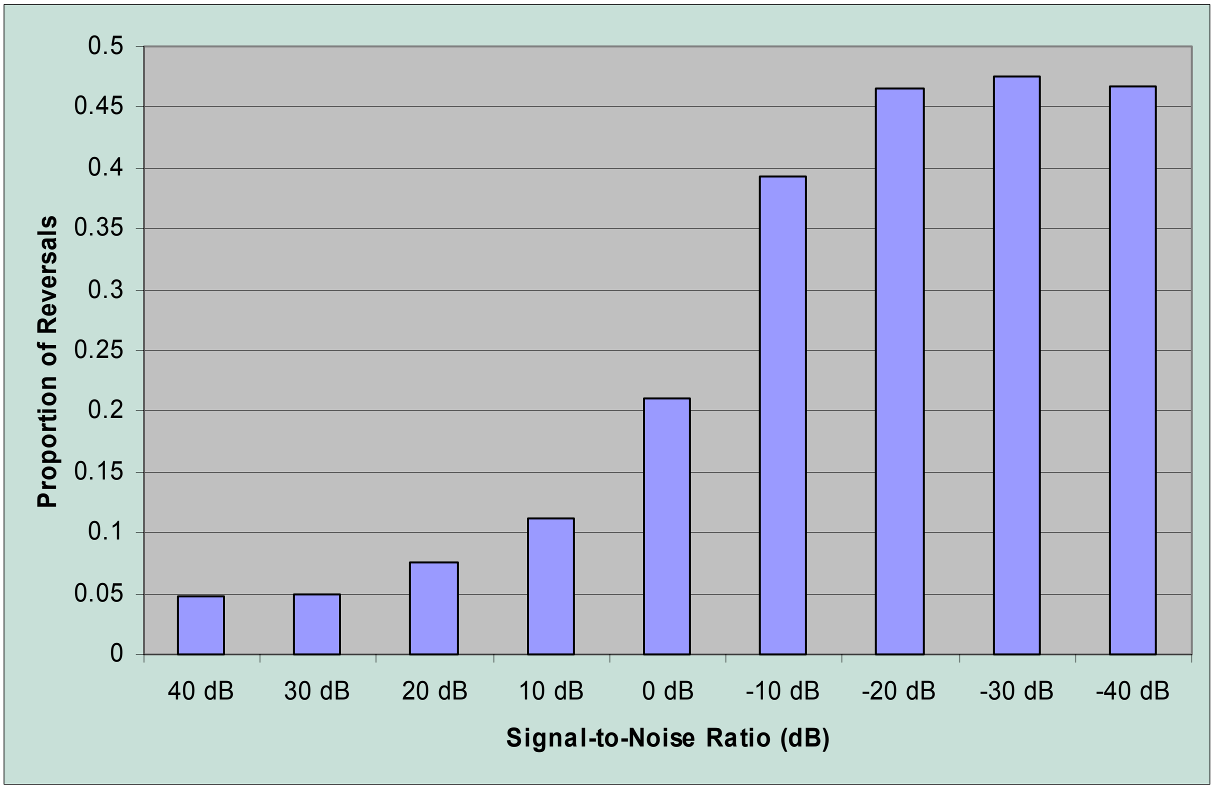

Figure 3 – Proportion of Reversals by S/N Ratio

Finally, the mean error corrected for front/back reversals is illustrated in Figure 4. If the localizer committed a reversal on a trial, the estimated location was flipped across the interaural axis and the unsigned error was computed using this corrected estimate. This is the performance that could be expected if the domain of possible sound source locations was restricted to those locations in the front hemisphere. The mean corrected error was below one degree for S/N ratios of 10 dB and higher, and fair performance was also obtained at -10 dB.

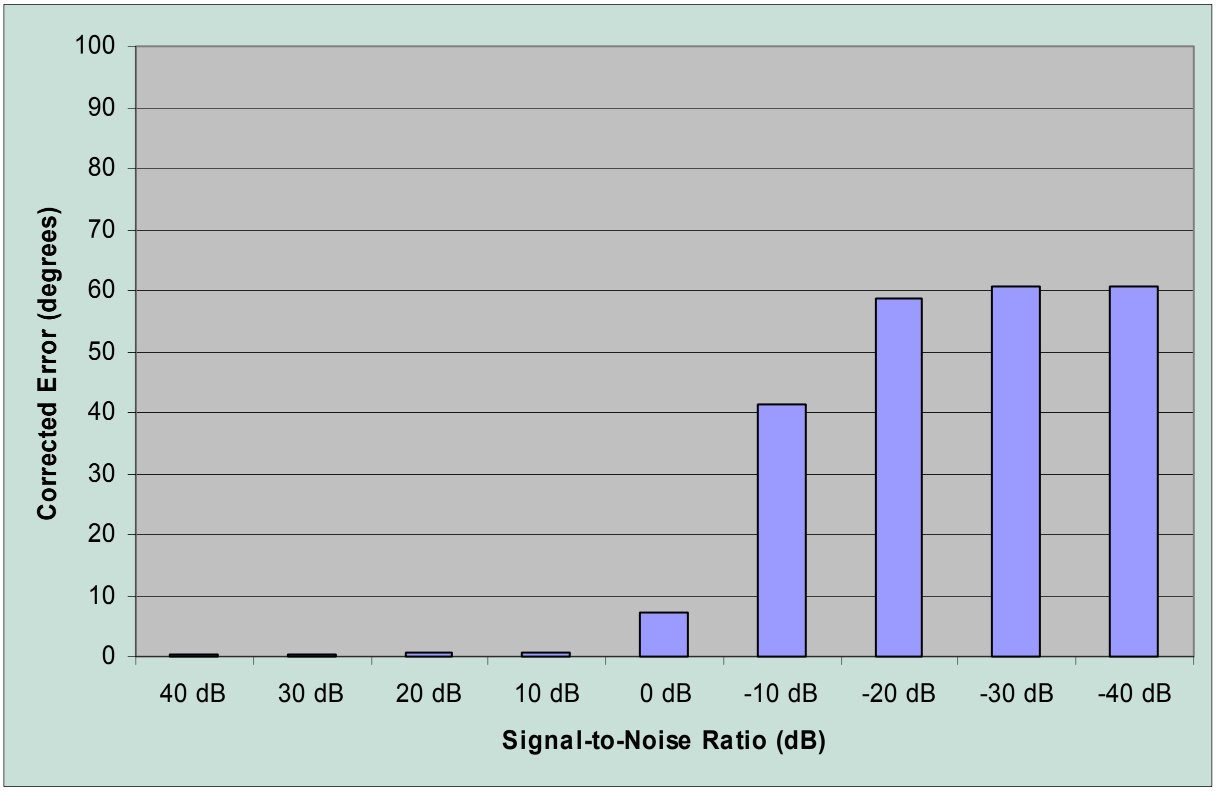

Figure 4 – Mean Corrected Error by S/N Ratio

4.0 DISCUSSION

The cross-channel localization algorithm clearly provides an effective means to estimate accurately the direction of a sound source. The localizer utilizes input from only two sensors and requires only moderate computational resources. The algorithm maintains accuracy under noisy conditions, and indeed performs at a better-than-chance level at S/N ratios of -10 dB and higher. This high level of performance is obtained without previous knowledge of the sound stimuli, instead utilizing relying solely upon HRTFs. The inclusion of both time delay information as well as frequency effects related to the head and torso serves to minimize the proportion of reversals. In fact, a localizer relying upon the time delay information alone would likely commit reversals on 50% of trials in all conditions. In contrast, the cross-channel localizer maintained reversal rates of less than 10% for S/N ratios 20 dB and higher.

Considered solely as part of a computer-based localization system, the cross-channel localizer shows considerable promise for military and commercial applications. The algorithm would be ideal for inclusion into a system to aid human sound localization. The algorithm could also be used for navigation or target tracking purposes, or possibly sniper detection. Future work will continue the testing of the algorithm with sounds at positive and negative elevations. In addition, the effect of adding input from a third microphone to the algorithm will be investigated. Three sensors may help to minimize front-back reversals as well as improve elevation discrimination.

The cross-channel localizer described in this paper should perform well as the basis of a model of human sound localization. Preliminary testing suggests that it performs similarly to humans when the accuracy of the localizer is reduced by adding noise to the stimulus. The algorithm is also well-suited to predict individual differences: performance can be compared when using the HRTFs of several different listeners. Future modelling work will focus upon predicting the localization performance afforded by different Army headgear designs, as well as the inclusion of a learning mechanism in the model. The eventual goal is a model that can predict performance in both sound localization and auditory distance estimation tasks in various environments and for variety of headgear combinations.

5.0 REFERENCES

[1] 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.

[2] Viera, J. & Almeida, L. (2003). A sound localizer robust to reverberation. *Proceedings of the 115th Convention of the Audio Engineers Society,* Paper 5973.

[3] Blauert, J. (1989). Spatial hearing. MIT Press, Cambridge, Massachusetts.

[4] 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.

[5] Oppenheim, A. V. Schaefer, R. W., & Buck, J. R. (1999). Discrete-time Signal Processing. Prentice Hall Publishers, Upper Saddle River, New Jersey.

[6] Greenfield, R. & Hawksford, M. O. (1991). Efficient filter design for loudspeaker equalization. *Journal of the Audio Engineering Society, 39*, 739 – 751.

[7] Rife, D. D. and Vanderkooy, J. (1989). Transfer-Function Measurement with Maximum-Length Sequences. *Journal of the Audio Engineering Society, 37,* 419-444.