A localization algorithm based on head-related transfer functions[a)]

Justin A. MacDonald[b)]
U.S. Army Research Laboratory, Human Research and Engineering Directorate, Aberdeen Proving Ground, Maryland 21005

(Received 27 September 2007; revised 22 March 2008; accepted 24 March 2008)

Two sound localization algorithms based on the head-related transfer function were developed. Each of them uses the interaural time delay, interaural level difference, and monaural spectral cues to estimate the location of a sound source. Given that most localization algorithms will be required to function in background noise, the localization performance of one of the algorithms was tested at signal-to-noise ratios (SNRs) from 40 to −40 dB. Stimuli included ten real-world, broadband sounds located at 5° intervals in azimuth and at 0° elevation. Both two- and four-microphone versions of the algorithm were implemented to localize sounds to 5° precision. The two-microphone version of the algorithm exhibited less than 2° mean localization error at SNRs of 20 dB and greater, and the four-microphone version committed approximately 1° mean error at SNRs of 10 dB or greater. Potential enhancements and applications of the algorithm are discussed.
 [DOI: 10.1121/1.2909566]

PACS number(s): 43.60.Jn, 43.66.Pn, 43.66.Qp [EJS] Pages: 4290–4296

I. INTRODUCTION

An ongoing project at the U.S. Army Research Laboratory has been to develop a biologically inspired algorithm to localize sounds in noisy environments in near real time. The motivations for this project are twofold: first, the algorithm could be implemented in autonomous robots or other unmanned vehicles to allow for accurate navigation and environment monitoring. The human listener provides an excellent example of an autonomous system that provides accurate location estimates in a wide variety of suboptimal environments. Second, a biologically inspired sound localization algorithm could be integrated into a computational auditory scene analysis (CASA) framework to segregate concurrent sounds based on the spatial locations of the sound sources. Location-based CASA approaches rely on localization algorithms to estimate sound source positions.

Machine-based sound localization systems take the input from two or more microphones to estimate the azimuth and elevation of a sound source. The localization algorithm must somehow extract location cues from the inputs to the sensors and determine the sound source location most likely to have produced the observed cues. Development of a new localization algorithm requires identification of the cues to be extracted as well as the decision process to be used to produce a location estimate. Which cues and decision processes are chosen depends on the goal of the algorithm designer. The human listener achieves accurate localization performance by using three types of location cues: the interaural time difference (ITD), the interaural level difference (ILD), and monaural spectral cues resulting from the irregular shape of the head and torso of the listener. Many sound localization algorithms utilize only the time delay cue to estimate the location of the sound source presumably because it is the easiest cue to extract from an incoming signal. For example, Calmes *et al.* (2007) constructed a neurally inspired model to detect ITDs to localize pure tones and wideband stimuli. The model performed well for wideband stimuli when the domain of potential locations was restricted to the front hemisphere. Viera and Almeida (2003) constructed a two-sensor system that localized sound sources between +60° and −60° azimuth to a precision of 9°. The restriction of possible locations to a single hemisphere is common to all localization algorithms that exclusively rely on the time delay between two microphones to estimate the location of the sound (e.g., Lotz *et al.*, 1989; Halupka *et al.*, 2005). The performance of these algorithms will suffer if sounds located in the rear hemisphere are included because any two-sensor system that exclusively relies on 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 on 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 Calmes *et al.* (2007), Viera and Almeida (2003), and others] or by using additional cues to estimate the source location. Chung *et al.* (2000) sought to resolve the ambiguity of the time delay cue by including monaural spectral cues in the localization algorithm. The model exhibited approximately 10° of localization error when broadband stimuli were presented from either hemisphere. Zakarauskas and Cynader (1993) developed an algorithm that compared the frequency spectrum of the incoming stimulus to the head-related transfer function (HRTF) (see Wightman and Kistler, 1989) in the frequency

[a)]Portions of this work were presented in "A sound localization algorithm for use in unmanned vehicles," Papers from the American Association for Artificial Intelligence Symposium on Aurally Informed Performance, Technical Report No. FS-01-01, Arlington, VA, October 2006.

[b)]Present address: New Mexico State University, P.O. Box 30001/MSC 3452, Las Cruces, NM 88003. Electronic mail: jmacd@nmsu.edu

domain, which is equivalent to using the ILD and monaural spectral cues to localize the sound. The mean error ranged from 0.29° to 25.4° depending on the stimulus being localized. Several other models have relied on ILD and monaural spectral cues to localize stimuli as well [Neti *et al.*, 1992; Middlebrooks, 1992; Chau and Duda, 1996]. Lim and Duda (1995) constructed a localization algorithm that utilized ITD, ILD, and monaural spectral cues to accurately estimate the location of an impulse source in an anechoic environment. The algorithm functioned by estimating the ITD and ILD from an incoming signal and comparing the estimates to a set of ITDs and ILDs from known source locations. The algorithm was able to perform quite accurately, exhibiting 0.8° of azimuth error in an anechoic environment.

The real-world performance of these algorithms is difficult to determine, however, given that nearly all of them were tested in a quiet environment. Most localization systems will be required to operate in a noisy environment, and the type of stimuli to be localized could considerably vary from the ideal. To this end, this paper details the design and subsequent testing of a biologically inspired sound localization algorithm that uses ITD, ILD, and monaural spectral cues to estimate the locations of real-world sounds. The human listener also takes advantage of other location cues during real-world sound localization tasks, including cues based on head movement, knowledge of the environment, or previous exposure to the stimulus being localized. Ideally, these location cues would have been included in the algorithm as well. Given the difficulty in extracting these cues, however, the localization algorithm was constructed to take advantage of only those cues available to a stationary, naive listener. The performance of the algorithm was measured across a range of signal-to-noise ratios (SNRs) to estimate the performance under suboptimal conditions.

The algorithm was designed to function using two or more microphones mounted in nearly arbitrary locations. The number of microphones included in the array depends on the application of the algorithm. If the algorithm is to be part of a system in which biological plausibility is required (in CASA applications or human localization modeling, for example), only two microphones are appropriate. If one is not subject to this restriction, however, then the number of microphones included in the array will be determined by the accuracy required and the computational resources that are available. Increasing the number of microphones from two to four, for example, is likely to improve the performance by reducing the number of front/back confusions exhibited by the localization system. This will be accomplished at the cost of increased computational requirements of the algorithm. Given that one of the eventual goals of this effort is to develop a location-based approach to CASA, the microphones were mounted to the Knowles electronics mannequin for acoustic research (KEMAR). The KEMAR is a human model that mimics the effects of the head and torso on an incoming sound wave. Both two- and four-microphone implementations of the algorithm were tested to determine the increased performance gained when another pair of microphones is added to the array.

II. LOCALIZATION ALGORITHMS

Consider an array of m microphones mounted at arbitrary locations whose center is at point P. Imagine a sound that originates from azimuth θ and elevation ϕ relative to P. The task of any localization algorithm is to process each of the m microphone inputs $\{I_1, \ldots, I_m\}$ to generate azimuth and elevation estimates $\hat{\theta}$ and $\hat{\phi}$, respectively. Ideally, the algorithm should utilize all available location cues to maximize accuracy. Differences in times of arrival between the microphones will vary with the location of the sound source and can therefore be utilized to generate location estimates. Additional location cues are available if the frequency content of the microphone inputs varies with the location of the sound source. This can be achieved by inserting an object centered at P into the listening environment so that the filtering properties of the object will vary with the orientation of the sound source.

For illustrative purposes, consider the situation in which $m=2$ microphones are mounted at the opening to each ear canal of a KEMAR. Let the center of the head of the KEMAR be located at P. Consider a sound that originates at azimuth θ and elevation ϕ relative to P. The sound is altered by the head and torso of the KEMAR before it arrives at the microphones. If I_j is a digital recording of the input to the jth microphone, then

$$I_j = O * F_j^{(\theta,\phi)}, \tag{1}$$

where O is the sound that would arrive at point P if the KEMAR were absent, $*$ is the convolution operator, and $F_j^{(\theta,\phi)}$ is the head-related impulse response (HRIR) for microphone j when a sound originates from (θ,ϕ). The HRIR is a representation of the HRTF in the time domain rather than the frequency domain and can therefore include both the time- and frequency-based filtering effects of the head and torso.

Consider the result when I_1 is convolved with $[F_1^{(\theta,\phi)}]^{-1}$, which is the inverse of the HRIR associated with (θ,ϕ) at microphone 1. In this case,

$$I_1 * [F_1^{(\theta,\phi)}]^{-1} = (O * F_1^{(\theta,\phi)}) * [F_1^{(\theta,\phi)}]^{-1} = O \tag{2}$$

due to the associativity of the convolution operator. In other words, if the effects of the head and torso of the KEMAR are removed from the recordings, the stimulus that would have arrived at P if the KEMAR was absent is the result. Similarly,

$$I_2 * [F_2^{(\theta,\phi)}]^{-1} = (O * F_2^{(\theta,\phi)}) * [F_2^{(\theta,\phi)}]^{-1} = O. \tag{3}$$

In both cases, if the inverse of the HRIR associated with the actual location of the sound source is chosen, then the original unaltered stimulus is the result. However, if the inverse of the HRIR associated with some other location (θ',ϕ') is convolved with the microphone inputs, then

$$I_1 * [F_1^{(\theta',\phi')}]^{-1} = (O * F_1^{(\theta,\phi)}) * [F_1^{(\theta',\phi')}]^{-1} \tag{4}$$

and

$$I_2 * [F_2^{(\theta',\phi')}]^{-1} = (O * F_2^{(\theta,\phi)}) * [F_2^{(\theta',\phi')}]^{-1}. \quad (5)$$

In this case, the convolution does not lead to the same result for I_1 and I_2. This suggests a method for determining the location of the sound source (θ, ϕ) from the microphone inputs I_1 and I_2: choose $(\hat{\theta}, \hat{\phi})$ to maximize the similarity between $I_1 * [F_1^{(\hat{\theta},\hat{\phi})}]^{-1}$ and $I_2 * [F_2^{(\hat{\theta},\hat{\phi})}]^{-1}$. Of course, a wide variety of similarity metrics are available; a moderate amount of testing suggested that the Pearson correlation maximized the accuracy and reliability of the "inverse" localization algorithm. Formally, the inverse algorithm chooses $(\hat{\theta}, \hat{\phi})$ according to the following equation:

$$\max_{(\hat{\theta},\hat{\phi})} r(I_1 * [F_1^{(\hat{\theta},\hat{\phi})}]^{-1}, I_2 * [F_2^{(\hat{\theta},\hat{\phi})}]^{-1}). \quad (6)$$

Another localization algorithm that does not require inverse filters was also developed. Continuing the example of two microphones mounted at the openings of the ear canals of the KEMAR, each input is convolved with the HRIR associated with the opposite microphone. If the HRIR associated with the correct location (θ, ϕ) is used, then

$$I_1 * F_2^{(\theta,\phi)} = (O * F_1^{(\theta,\phi)}) * F_2^{(\theta,\phi)} = O * F_1^{(\theta,\phi)} * F_2^{(\theta,\phi)} \quad (7)$$

and

$$I_2 * F_1^{(\theta,\phi)} = (O * F_2^{(\theta,\phi)}) * F_1^{(\theta,\phi)} = O * F_2^{(\theta,\phi)} * F_1^{(\theta,\phi)} = O * F_1^{(\theta,\phi)} * F_2^{(\theta,\phi)}. \quad (8)$$

This follows from the commutativity and associativity of the convolution operator. As with the inverse algorithm, if the correct location is chosen, then the operation will lead to the same result for both microphone inputs. If the HRIR associated with some other location (θ', ϕ') is chosen, however, then the results will differ:

$$I_1 * F_2^{(\theta',\phi')} = (O * F_1^{(\theta,\phi)}) * F_2^{(\theta',\phi')} = O * F_1^{(\theta,\phi)} * F_2^{(\theta',\phi')} \quad (9)$$

and

$$I_2 * F_1^{(\theta',\phi')} = (O * F_2^{(\theta,\phi)}) * F_1^{(\theta',\phi')} = O * F_2^{(\theta,\phi)} * F_1^{(\theta',\phi')} = O * F_1^{(\theta',\phi')} * F_2^{(\theta,\phi)}. \quad (10)$$

As before, the "cross-channel" algorithm uses the Pearson correlation coefficient as the similarity metric, choosing $(\hat{\theta}, \hat{\phi})$ as follows:

$$\max_{(\hat{\theta},\hat{\phi})} r(I_1 * F_2^{(\hat{\theta},\hat{\phi})}, I_2 * F_1^{(\hat{\theta},\hat{\phi})}). \quad (11)$$

The generalization of these algorithms to more than two microphones is relatively straightforward. A microphone array with $2N$ microphones is arbitrarily partitioned into N microphone pairs. Let the first and second microphones in the kth pair be denoted by k_1 and k_2, respectively. The associated microphone inputs will be denoted by I_{k_1} and I_{k_2} by using this notation. For the inverse algorithm, the first microphone input in each pair is convolved with the associated inverse impulse response, as in Eq. (2). The results of the N convolutions are then concatenated:

$$I_{1_1} * [F_{1_1}^{(\hat{\theta},\hat{\phi})}]^{-1} \,\&\, \cdots \,\&\, I_{N_1} * [F_{N_1}^{(\hat{\theta},\hat{\phi})}]^{-1}, \quad (12)$$

where & is the concatenation operator. Similarly, the second microphone input in each pair is convolved with the associated inverse impulse response, and the results of the N convolutions are concatenated:

$$I_{1_2} * [F_{1_2}^{(\hat{\theta},\hat{\phi})}]^{-1} \,\&\, \cdots \,\&\, I_{N_2} * [F_{N_2}^{(\hat{\theta},\hat{\phi})}]^{-1}. \quad (13)$$

The concatenated results are then correlated to determine $(\hat{\theta}, \hat{\phi})$:

$$\max_{(\hat{\theta},\hat{\phi})} r(I_{1_1} * [F_{1_1}^{(\hat{\theta},\hat{\phi})}]^{-1} \,\&\, \cdots \,\&\, I_{N_1} * [F_{N_1}^{(\hat{\theta},\hat{\phi})}]^{-1},$$

$$I_{1_2} * [F_{1_2}^{(\hat{\theta},\hat{\phi})}]^{-1} \,\&\, \cdots \,\&\, I_{N_2} * [F_{N_2}^{(\hat{\theta},\hat{\phi})}]^{-1}). \quad (14)$$

Note that Eq. (14) simplifies to Eq. (6) when only one microphone pair is used. The multichannel implementation of the cross-channel algorithm is structured in a similar fashion. In this case, $(\hat{\theta}, \hat{\phi})$ is chosen as follows:

$$\max_{(\hat{\theta},\hat{\phi})} r(I_{1_1} * F_{1_2}^{(\hat{\theta},\hat{\phi})} \,\&\, \cdots \,\&\, I_{N_1} * F_{N_2}^{(\hat{\theta},\hat{\phi})},$$

$$I_{1_2} * F_{1_1}^{(\hat{\theta},\hat{\phi})} \,\&\, \cdots \,\&\, I_{N_2} * F_{N_1}^{(\hat{\theta},\hat{\phi})}). \quad (15)$$

Equation (15) reduces to Eq. (11) when only one pair of microphones is used.

Of the two algorithms presented here, it is likely that the cross-channel algorithm will be preferred for most applications. The inverse algorithm is likely to require greater computational resources than the cross-channel algorithm. Appropriate inverse filters that account for the magnitude and phase portions of the HRIR are typically of greater complexity than the original filter. For example, by using the method detailed by Greenfield and Hawksford (1991), an inverse filter that accounts for both magnitude and phase response will be approximately three times longer than the original HRIR. Considering that these inverted HRIRs are convolved with the microphone inputs, the computational requirements of the inverse algorithm will be substantial. In addition, methods to compute inverse filters produce filters that are only an approximate inverse of the original (Rife and Vanderkooy, 1989). Because the accuracy of the inverse filter increases with its length, one must consider the trade-off between the accuracy of the inverse filter and computational requirements of the algorithm. Fortunately, the cross-channel algorithm does not suffer from these drawbacks: inverse filters are not required. For this reason, the cross-channel algorithm was chosen for further testing.

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 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: the 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 confusions occurred in approximately 5% of the trials at the higher SNRs.

These promising initial results prompted a larger-scale test using both two- and four-microphone versions of the algorithm. Accordingly, two additional microphones were mounted on the front and rear of the head of the KEMAR. The additional microphones should allow for a reduced number of front/back confusions and an increased localization accuracy at the expense of an increased computation time.

III. SIMULATION METHOD

A. Stimuli

Ten naturally occurring sounds were chosen as the test signals: the sounds of breaking glass, a speech stimulus, the insertion of an M-16 magazine, a camera shutter release sound, machine gun fire, a cough, a dog bark, a door being slammed, a water dripping noise, and the sound of a heavy object being dropped into a body of water. 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.

B. Stimulus recording apparatus

Stimuli were presented using the Army Research Laboratory Human Research and Engineering Directorate's RoboArm 360 system. 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 digital to analog converter, which was 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 1 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 m from the center of the head of the KEMAR and at 0° elevation for all stimulus presentations.

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 microphones were placed at 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 same elevation at the rearmost part of the head. Inputs to the microphones were amplified by a TDT System 3 MA3 microphone amplifier before being sent to a TDT System II DD1 analog to digital converter. The digital output of the DD1 was sent to a computer for storage in a 44.1 kHz, 16 bit Microsoft WAV format. By combining across recording sessions, a total of 720 four-channel recordings were made, one for each position/sound combination.

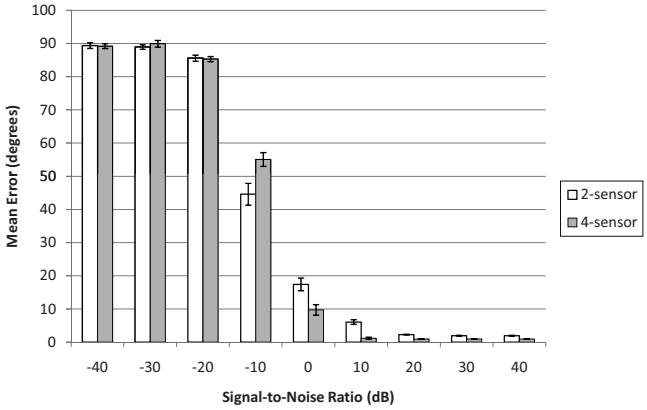

FIG. 1. Mean localization error in each SNR condition. The bars indicate the standard error associated with each mean. The location estimates were not corrected for front/back confusions. Chance performance in this localization task corresponds to a 90° mean error.

C. HRIR measurement

The HRIR of the KEMAR was measured using the same presentation and recording apparatus detailed above. The maximum-length sequence (see Rife and Vanderkooy, 1989) stimuli were presented at 5° intervals around the head of the KEMAR and the signals recorded at the microphones determined the HRIR of the KEMAR at each location. As with the stimulus recordings, the front/back and left/right impulse responses were separately estimated. Each HRIR was stored as a 256-tap finite impulse response digital filter.

D. Procedure

Simulations were conducted using a script written in MATLAB (The Mathworks, Natick, MA) 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 HRIRs associated with the left and right microphones to process the recordings made at those locations, and estimates were produced using Eq. (11). The four-sensor simulation used Eq. (15) to apply the four-channel HRIRs to the four-channel recordings. Locations were estimated with 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 −40 dB in 10 dB increments. The SNR for each trial was calculated based on the signal channel with the greatest root-mean-squared amplitude. The algorithm was required to localize each recording ten times; a different sample of Gaussian noise was added on each attempt. This resulted in a total of 64 800 localization attempts for each of the simulations (9 SNRs × 720 recordings × 10 trials each).

IV. RESULTS

All location estimates in the simulation were left uncorrected: estimates were not reflected across the interaural axis when a front/back confusion occurred. The absolute error (the absolute value of the angular distance between the estimated and actual sound locations, in degrees) was used as the error measure. The mean error observed at each SNR (collapsed across the ten sound stimuli) is shown in Fig. 1.

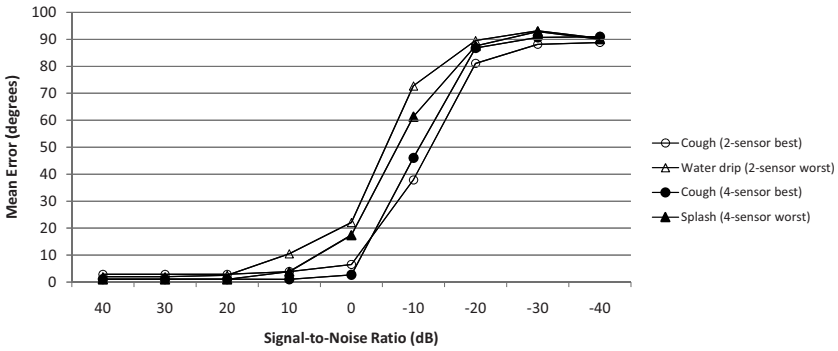

FIG. 2. Errors associated with the best- and worst-localized sounds in each SNR condition. The performance of both algorithms tended to increase with the bandwidth of the stimulus.

The two-microphone implementation exhibited approximately 2° localization error when the SNR was greater than 10 dB and performed well above chance levels to −20 dB. The four-microphone implementation exhibited an even greater accuracy, maintaining a mean error of approximately 1° in SNRs of 10 dB and greater and performing above chance to −20 dB. The performance of the algorithm varied somewhat across stimuli; the error bars in the figure indicate the standard error of the mean calculated across the stimulus set. The effect of the stimulus on the performance of the algorithm is illustrated in greater detail in Fig. 2: the errors associated with the best- and worst-localized sounds are shown for the two- and four-sensor versions of the algorithm. In general, performance increased with the bandwidth of the stimulus.

The mean localization error observed in the 10 dB SNR condition at each location is shown is Fig. 3. The 10 dB condition was chosen so that a sufficient number of errors could be included in the figure. The two-sensor implementation of the algorithm exhibited systematic errors at higher noise levels when sounds were located just behind the interaural axis. The large majority of these errors were back-to-front confusions. There is a slight asymmetry in the two-sensor error pattern; this is likely due to the acoustic properties of the room in which the sound recordings were made. The four-sensor implementation was much less susceptible to these errors, illustrating the benefits of including an additional two sensors in the array.

The proportion of front/back confusions in each SNR condition is shown in Fig. 4. A front/back confusion occurred when the estimated and actual locations of the sound source were on opposite sides of the interaural axis. Two-microphone systems that exclusively rely on time-of-arrival differences will exhibit a 50% confusion rate. The inclusion of the ILD and monaural cues in the cross-channel algorithm led to a significant reduction in the number of confusions: fewer than 5% of the trials resulted in confusions in the 40, 30, and 20 dB SNR conditions, and performance was well above chance to −20 dB. As expected, the addition of two microphones in the four-sensor implementation led to an increased performance: confusions were reduced to a trivial level (0.28%) at 10 dB and were entirely eliminated in the 20, 30, and 40 dB conditions.

V. 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 2° despite the addition of a moderate amount of Gaussian noise. The accuracy of the algorithm is especially impressive considering that sounds were allowed to originate from the rear hemisphere, thereby allow-

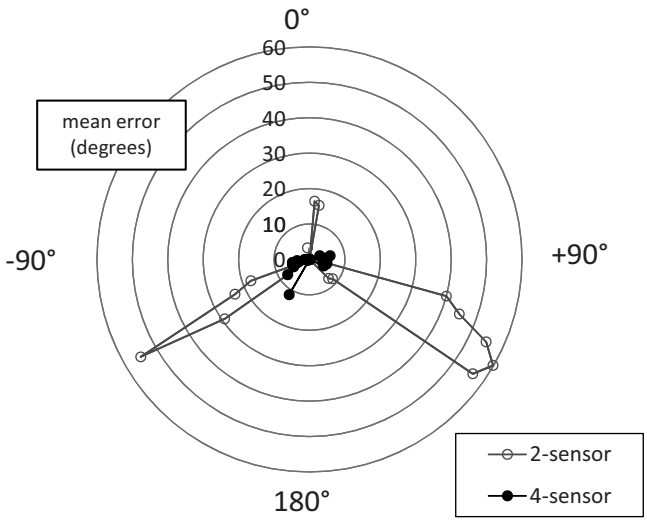

FIG. 3. Mean error for each location in the 10 dB SNR condition. The two-sensor version of the algorithm exhibited frequent back-to-front confusions at moderate and high noise levels for sounds located just behind the interaural axis. These errors were not observed in the four-sensor version of the algorithm.

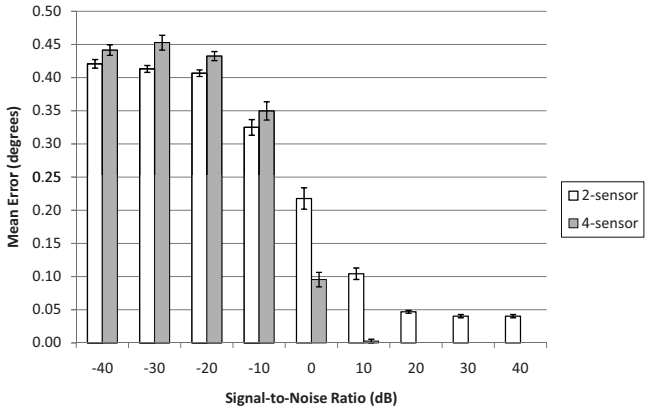

FIG. 4. Proportion of front/back confusions in each SNR condition. A confusion occurred when the estimated and actual source locations were on opposite sides of the interaural axis. Chance performance corresponds to a 50% confusion rate. No confusions (out of 21 600 trials) were observed for the four-sensor implementation of the algorithm at SNRs of 40, 30, and 20 dB.

ing for the possibility of front/back confusions. The inclusion of frequency-based location cues allowed for a severe reduction in the number of front/back confusions. As expected, the four-microphone implementation of the algorithm exhibited an even better performance, committing almost no reversals in all SNRs greater than 0 dB.

It is worth noting that the performance of the algorithm depends on the transfer function of the structure to which the microphones are mounted. Asymmetrical structures with maximally separated microphone mounting points should possess transfer functions that exhibit considerable variance across sound source locations, thereby increasing the performance of the algorithm. The KEMAR is likely to be a suboptimal choice in this regard: it is relatively symmetric with respect to both the medial and interaural axes and there is only a short distance between the mounting points for the microphones. Despite this handicap, the KEMAR-based implementation compares favorably to the large majority of other localization algorithms, exhibiting a mean localization error of 1.9 degrees in the 40 dB SNR condition when localizing ten different real-world sounds. In comparison, Berdugo *et al.* (1999) reported errors of approximately five degrees in quiet when using an array of seven microphones to localize a 20 s speech signal. Viera and Almeida (2003) reported a mean localization error of approximately nine degrees when source locations were restricted to the front hemisphere. Schauer and Gross (2001) observed a mean localization error of approximately ten degrees under the same restriction. Zakarauskas and Cynader (1993) measured localization errors between 0.29 and 25.4 degrees depending on the stimulus being localized. Neti *et al.* (1992) reported a mean localization error of 6.3° when source locations were restricted to the range between −30° and +30°. The strongest performance was reported by Lim and Duda, who observed a 0.8° mean error in azimuth when localizing an ideal broadband stimulus (an impulse) in anechoic conditions. The performance of the algorithm in suboptimal (noisy) conditions was not reported.

An analysis of the results of the two-microphone implementation of the algorithm can provide insight into the performance of the human listener. By assuming that the KEMAR is an accurate model of the human head and torso, the results of the simulation indicate that a highly accurate localization performance is possible using information that is available at the entrance to the ear canal. Accurate localization is quite possible in noisy environments without previous exposure to the stimulus. The inferior performance of the human listener in these conditions must arise from the following: either the information available at the ear canal is not available in the central nervous system where the location estimate is made, or the decision process used to produce the location estimate is suboptimal, or (most likely) both. The former possibility can be partially tested by filtering the recording through a model of the auditory periphery and by using the cross-channel algorithm to localize sounds based on the output of the model.

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. An examination of the accuracy of the cross-channel algorithm across elevations is currently underway. In addition, localization accuracy in a multisound environment must be investigated especially if the localization algorithm is to be integrated into a CASA algorithm.

Both the inverse and cross-channel algorithms could be altered in a variety of ways to determine if the accuracy of the algorithm can be improved. For example, the Pearson correlation is only one of the many possible similarity metrics that could be used in the inverse and cross-channel algorithms. Several other metrics were considered during the initial testing of the algorithms, including using the sum of the squared deviations rather than the Pearson correlation. Of the metrics considered, however, the Pearson correlation led to the best localization performance in an initial test and was therefore chosen for use in the subsequent full-scale evaluation. In addition, the computational requirements of the algorithm could be reduced using shortcuts to eliminate potential source locations. In a quiet environment, for example, all locations in the right hemisphere could be eliminated as potential source locations if the system determined that the sound arrived at the left microphone before the right. Many possible compromises between the two- and four-microphone algorithm implementations are worth investigating as well. For example, the algorithm could use the left and right channels to generate location estimates that are modified based on the relative intensity of the input to the front and back microphones. In addition, the locations of the microphones were somewhat arbitrarily chosen; it is quite possible that other locations will lead to better performance. Finally, it is likely that other mounting structures could be found that introduce greater variation in the HRTF across sound source locations, thereby increasing the performance of the algorithm. Refinements such as these will be explored in future work.

ACKNOWLEDGMENTS

The author wishes to thank Phuong Tran for her help in making the four-channel recordings used in the testing of the algorithm.

Berdugo, B., Doron, M. A., Rosenhouse, J., and Azhari, H. (**1999**). "On direction finding of an emitting source from time delays," J. Acoust. Soc. Am. **105**, 3355–3363.

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

Calmes, L., Lakemeyer, G., and Wagner, H. (**2007**). "Azimuthal sound localization using coincidence of timing across frequency on a robotic platform," J. Acoust. Soc. Am. **121**, 2034–2048.

Chau, W., and Duda, R. O. (**1996**). "Combined monaural and binaural localization of sound sources," IEEE Proceedings of the 29th Asilomar Conference on Signals, Systems, and Computers, pp. 1281–1285.

Chung, W., Carlile, S., and Leong, P. (**2000**). "A performance adequate computational model for auditory localization," J. Acoust. Soc. Am. **107**, 432–445.

Greenfield, R., and Hawksford, M. O. (**1991**). "Efficient filter design for loudspeaker equalization," J. Audio Eng. Soc. **39**, 739–751.

Halupka, D., Mathai, N. J., Aarabi, P., and Sheikholeslami, A. (**2005**). "Ro-

bust sound localization in 0.18 μm CMOS," IEEE Trans. Signal Process. **53**, 2243–2250.
Lim, C., and Duda, R. O. (**1995**). "Estimating the azimuth and elevation of a sound source from the output of a cochlear model," IEEE Proceedings of the 28th Asilomar Conference on Signals, Systems, and Computers, pp. 399–403.
Lotz, K., Bölöni, L., Roska, T., and Hámori, J. (**1999**). "Hyperacuity in time: A CNN model of a time-coding pathway of sound localization," IEEE Trans. Circuits Syst., I: Fundam. Theory Appl. **46**, 994–1002.
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 no. 28. Available: http://www.rta.nato.int/pubs/rdp.asp?RDP=RTO-MP-HFM-123.
Middlebrooks, J. C. (**1992**). "Narrow-band sound localization related to external ear acoustics," J. Acoust. Soc. Am. **92**, 2607–2624.
Neti, C., Young, E. D., and Schenider, M. H. (**1992**). "Neural network models of sound localization based on directional filtering by the pinna," J. Acoust. Soc. Am. **92**, 3140–3156.
Rife, D. D., and Vanderkooy, J. (**1989**). "Transfer-function measurement with maximum-length sequences," J. Audio Eng. Soc. **37**, 419–444.
Schauer, C., and Gross, H. M. (**2001**). "Model and application of a binaural 360° sound localization system," IEEE Proceedings of the International Joint Conference on Neural Networks, Vol. **2**, pp. 1132–1137.
Viera, J., and Almeida, L. (**2003**). "A sound localizer robust to reverberation," Proceedings of the 115th Convention of the Audio Engineers Society, Paper No. 5973.
Wightman, F. L., and Kistler, D. J. (**1989**). "Headphone simulation of free-field listening. I: Stimulus synthesis," J. Acoust. Soc. Am. **85**, 858–867.
Zakarauskas, P., and Cynader, M. S. (**1993**). "A computational theory of spectral cue localization," J. Acoust. Soc. Am. **94**, 1323–1331.