LOCALISATION OF MULTIPLE SOUNDS AS A FUNCTION OF THE DURATION OF THE INTER-SOUND GAP

Russell L. Martin, Ken I. McAnally, and J.P. Watt

Defence Science and Technology Organisation
P.O. Box 4331, Melbourne, Australia, 3001
russell.martin@dsto.defence.gov.au

ABSTRACT

One issue that arises in relation to the use of three-dimensional audio displays as a means of conveying spatial information to human operators concerns the ability of operators to track the locations of multiple sounds. We have examined the effect of varying the gap between a target and five distracter sounds on the accuracy with which the target sound can be localised under conditions in which participants were informed of the identity of the target sound after stimulus presentation. Inter-sound gap durations of 50, 100, 200, 400, 800 and 1600 ms were employed. Localisation performance was observed to improve as gap duration increased but was generally poor in comparison to that associated with localisation of the target sounds in the absence of the distracter sounds.

1. INTRODUCTION

Three dimensional (3D) audio displays are being implemented in several work environments to provide a means of conveying spatial information to human operators. Individualised 3D audio displays have the potential to generate spatial percepts as accurate as those associated with free-field sounds [e.g., 1]. One issue that arises in relation to the use of these displays concerns the ability of operators to track the locations of multiple sounds that may or may not overlap in time.

A number of studies have examined the way in which the localisation of target sounds is affected by the presence of temporally concurrent masking sounds. Good, Gilkey and Ball [2] found that the impact of a masking sound depends on both the ratio of the target- and masking-sound levels (SNR) and the location of the masking sound. For some masking-sound locations good localisation performance was observed at SNRs a few dB above the detection threshold. For others, localisation performance was severely disrupted at relatively high SNRs (i.e., 8 dB above the detection threshold). Similar results have been described by Lorenzi, Gatehouse and Lever [3]. Good, Gilkey and Ball reported a tendency for the apparent location of target sounds to be shifted toward the location of a masking sound but Getzmann [4] and Best, van Schaik, Jin and Carlile [5] reported a localisation bias in the opposite direction (i.e., away from the location of a masking sound).

Langendijk, Kistler and Wightman [6] examined the effect on the localisation of target sounds of the presence of one or two distracter sounds that were interleaved with, but did not overlap with, the target sound in time. The target sound in their study was a train of noise bursts and the distracter sounds were trains of distinct complex tones. They found that localisation performance deteriorated substantially as the number of distracter sounds increased from 0 to 2. Martin, Flanagan, McAnally and Eberle [7] examined the effect on the localisation of target sounds of the presence of up to five distracter sounds that did not overlap with the target sound in time. Target and distracter sounds in their study were readily identifiable environmental sounds. Whereas participants in Langendijk et al.'s study always knew that the noise burst would be the target sound, participants in Martin et al.'s study were informed of the target sound's identity either before (their experiment two) or after (their experiment one) stimulus presentation. When Martin et al.'s participants were informed of the target sound's identity before stimulus presentation, the presence of distracter sounds had no significant effect on localisation performance. When they were informed of the target sound's identity after stimulus presentation, in contrast, localisation performance became steadily worse as the number of distracter sounds increased from 0 to 5 and was reduced substantially for 3 or more distracter sounds.

The essential difference between the tasks in Martin et al.'s [7] first and second experiments is that the identities and locations of all sounds had to be determined and remembered in one experiment (i.e., that in which the target's identity was provided after stimulus presentation) but not the other (i.e., that in which the target's identity was provided before stimulus presentation). The contrasting effects observed in these experiments suggest that listeners lack either the processing resources required to localise multiple sounds as accurately as one could be localised or the memory resources required to retain the identities and locations of multiple sounds as accurately as those of one could be retained.

In the study by Martin et al. [7] sounds were separated by a 200-ms gap. Given the above considerations, variation in the size of this gap could be expected to influence the extent to which the presence of distracter sounds affects the localisation of a target sound under conditions in which the identity of the target sound is provided after stimulus presentation. For example, increasing the duration of the gap could help listeners retain the identities and locations of multiple sounds by allowing them more time for rehearsal of this information between sound presentations. Alternatively, it could result in more accurate localisation of multiple sounds by allowing listeners' processing resources to be devoted to each sound for a longer period of time. The current study examined the effect on the localisation of target sounds of varying the gap between the target and five distracter sounds under conditions in which

participants were informed of the identity of the target sound after stimulus presentation. Inter-sound gap durations of 50, 100, 200, 400, 800 and 1600 ms were employed.

2. METHOD

2.1. Participants

Four males and two females ranging in age from 28 to 46 years participated in this study. All had taken part in previous sound localisation studies in our laboratory. Two were coauthors of this paper. Each participant's hearing was tested by measuring his or her absolute thresholds for 1-, 2-, 4-, 8-, 10-, 12-, 14- and 16-kHz tones using a two-interval forced-choice task combined with the two-down one-up adaptive procedure [see 8 for details]. For each participant, all thresholds were lower than the relevant age-specific norm [9, 10]. Each participant gave his or her informed consent before taking part in the study.

2.2. Measurement of Head-Related Impulse Responses (HRIRs)

A set of HRIRs comprising responses for 448 sound-source locations was generated for each participant using a "blocked ear-canal" measurement technique. Miniature microphones (Sennheiser, KE4-211-2) encased in swimmer's ear putty were placed in the participant's left and right ear canals. Care was taken to ensure that the microphones were positionally stable and their diaphragms were at least 1 mm inside the ear-canal entrances.

The participant was seated in a 3- x 3-m, sound-attenuated, anechoic chamber at the center of a 1-m radius hoop on which a loudspeaker (Bose, FreeSpace tweeter) was mounted. The participant placed his or her chin on a rest that helped to position the head at the center of rotation of the hoop. Head position and orientation were tracked magnetically via a receiver (Polhemus, 3Space Fastrak) attached to a plastic headband worn by the participant. The head's position and orientation were displayed on a bank of light emitting diodes (LEDs) mounted within the participant's field of view. HRIR measurements were not made unless the participant's head was positioned within 0.3 cm of the hoop center (with respect to each of the x, y and z axes) and oriented within 1^o of straight and level.

HRIRs were measured at lateral angles ranging from –90 to $+90^o$ in steps of 10^o and polar angles ranging from 0 to 350^o in steps of 360^o (+/- 90^o lateral angles), 30^o (+/- 80^o lateral angles), 20^o (+/- 70 and 60^o lateral angles) or 10^o (all other lateral angles), provided the location's elevation was within the range from –50 to $+80^o$. For each location, two 8192-point Golay codes were generated at a rate of 50 kHz (Tucker-Davis Technologies, System II), amplified and played at 75 dB (A-weighted) through the hoop-mounted loudspeaker. The signal from each microphone was low-pass filtered at 20 kHz and sampled at 50 kHz (Tucker-Davis Technologies, System II) for 327.7 ms following initiation of the Golay codes. An impulse response was derived from each sampled signal [11], truncated to 128 points and stored.

Immediately following HRIR measurement, the impulse responses of the two miniature microphones were determined together with those of the headphones (Sennheiser, HD520 II) that would be subsequently used to present virtual sound. The headphones were carefully placed on the participant's head and Golay codes were played through them while the responses of the microphones were sampled. An impulse response was derived from each sampled signal, truncated to 128 points and stored.

The impulse response of the hoop-mounted loudspeaker had been derived previously from its response to the Golay code stimulus as measured using a microphone with a flat frequency response (Brüel and Kjær, 4003). The loudspeaker impulse response was truncated to 128 points and deconvolved from each HRIR by division in the complex frequency domain. The impulse responses of the microphones and headphones combined were zero-padded to 370 points and inverted in the complex frequency domain.

2.3. Localisation Procedure

The participant was seated on a swiveling chair at the center of the loudspeaker hoop in the same anechoic chamber in which his or her HRIRs had been measured. The participant's view of the hoop and loudspeaker was obscured by an acoustically transparent, cloth sphere supported by thin fiberglass rods. The inside of the sphere was dimly lit to allow visual orientation. Participants wore a headband on which a magnetic-tracker receiver and a laser pointer were rigidly mounted. They also wore the headphones for which impulse responses had been measured during the HRIR measurement procedure.

At the beginning of each trial the participant placed his or her chin on the rest and fixated on an LED at 0^o azimuth and elevation. When ready, he or she pressed a hand-held button. An acoustic stimulus was then presented, provided the participant's head was stationary (its azimuth, elevation and roll did not vary by more than 0.2^o over three successive readings of the head tracker made at 20-ms intervals), positioned within 1 cm of the hoop center, and oriented within 3^o of straight and level. Participants were instructed to keep their heads stationary during stimulus presentation.

Each stimulus consisted of a sequence of six readily identifiable sounds separated by 50-, 100-, 200-, 400-, 800- or 1600-ms gaps and followed by a speech cue that indicated which of the six sounds was the target for the trial. Each participant was assigned a unique set of six sounds that was composed of sounds selected from a larger set of twelve (see Table 1). The twelve sounds had been tested for identifiability and localisability prior to their use in the study. All were 500 ms in duration, incorporating 10-ms cosine-shaped rise and fall times. They were allocated to six-sound sets such that each appeared in three of the sets and no set contained the two sounds comprising either of two potentially confusable pairs (duck quacking/rooster crowing and kazoo being played/party whistle being played). On each trial the six sounds were played in a pseudorandom order such that the sound designated as the target for the trial appeared in each of the six timeslots on an equal number of occasions across the 42 trials in each session. The target sound for each trial was also chosen pseudorandomly such that each of the six sounds was the target on an equal number of occasions. The speech cue indicating which of the six sounds was the target was presented 1100 ms after the offset of the sixth sound. Speech cues consisted of the words "Where was the" followed by the identifier for the target sound (see

Table 1), as generated by a text-to-speech synthesizer (AT&T Natural Voices). All speech cues were 1500 ms in duration, incorporating 10-ms cosine-shaped rise and fall times. The six sounds and the speech cue were presented at a clearly audible level of around 60 dB (A-weighted).

Sound	Identifier
Chainsaw being revved	Chainsaw
Dog barking	Dog
Duck quacking	Duck
Elephant trumpeting	Elephant
Galah screeching	Galah
Machine-gun being fired	Gun
Kazoo being played	Kazoo
Party-whistle being played	Party-whistle
Rooster crowing	Rooster
Woman screaming	Scream
Steamboat horn being blown	Steamboat
Glass being broken	Broken glass

Table 1: Sounds and their identifiers.

Following stimulus presentation, the participant turned his or her head (and body, if necessary) to direct the head-mounted laser pointer's beam at the point on the cloth sphere from which he or she perceived the target sound to come. The location and orientation of the laser pointer were measured by the magnetic tracker, and the point where the beam intersected the sphere was calculated geometrically. The true location of the source of the sound was calculated taking the position and orientation of the participant's head at the time of stimulus presentation into account.

Target-sound locations were chosen following a pseudorandom procedure from the 448 locations for which HRIRs had been measured. The part-sphere from –47.6 to +80^{o} elevation and 0 to 359.9^{o} azimuth was divided into 42 sectors of equal area. Each sector contained from 7 to 15 locations for which HRIRs had been measured. To ensure a reasonably even spread of target-sound locations in each session, one sector was selected randomly without replacement on each trial and a location within it was selected randomly. Sound-sound locations were chosen by randomly selecting five additional sectors (i.e., five sectors that were different from each other and that containing the target-sound location) on each trial and randomly selecting one location within each of them.

The duration of the inter-sound gap was held constant within each session. Each participant completed two sessions for each of the six inter-sound gap durations. These 12 sessions were organised in two blocks. The order in which the inter-sound gap durations were presented within each block was counterbalanced across participants following a Latin-square design. For each participant the order was reversed across blocks.

2.4. Data Analysis

A general difficulty associated with the analysis of data from sound-localisation studies results from the occasional occurrence of front-back confusions (i.e., localisations to the incorrect front-back hemifield). As front-back confusions are inherently different from other less-than-perfect localisation responses and commonly associated with particularly large localisation errors, their presence usually results in the overall distribution of localisation errors being bimodal. It is common practice, therefore, to separate front-back confusions from other localisation responses before proceeding with subsequent analysis.

In the case of the current study, data analysis is more complicated than usual because the presence of more than one source of sound leads to the possibility of additional types of localisation response. For example, a participant could have confused the location of a target sound with that of one of the distracter sounds, or even with that of the distracter sound's front-back reflection. As there were five distracter sounds in this study and each of them could have been located relatively close to the target sound, it would be difficult to separate localisation responses of this type from other localisation responses.

We addressed this problem by describing localisation responses using an alternative metric to the localisation error. This metric was whether or not the localisation error (which we defined as the angle subtended at the centre of the participant's head between the vectors pointing to the true and perceived target-sound locations) for the particular response was less than or equal to a reference localisation error. The reference localisation error was the participant's 75th percentile localisation error for the six sounds assigned to him or her in this study under conditions in which no distracter sounds were present. This localisation error was measured prior to the current study using procedures identical to those described here except for the absence of distracter sounds. The participant localised each of the six sounds on 21 occasions spread across three sessions of 42 trials each. The localisation errors associated with all 126 localisation responses were combined and the 75th percentile was determined. The 75th percentile localisation errors for the six participants ranged from 14.6 to 21.6^{o}.

3. RESULTS

The average percentage of trials on which the localisation error was less than or equal to the participant's 75th percentile error for localisation without distracter sounds is shown in Figure 1 for each of the inter-sound gap durations. A clear trend for localisation performance to improve as gap duration increases is apparent, although the extent of the improvement across the full range of durations is relatively small (i.e., from 34.3% for 50 ms to 44.4% for 1600 ms). Statistical analysis (trend analysis based on repeated-measures ANOVA) indicated that a significant linear trend was present across the gap durations employed ($F(1,5)=10.32$, $p=.0237$).

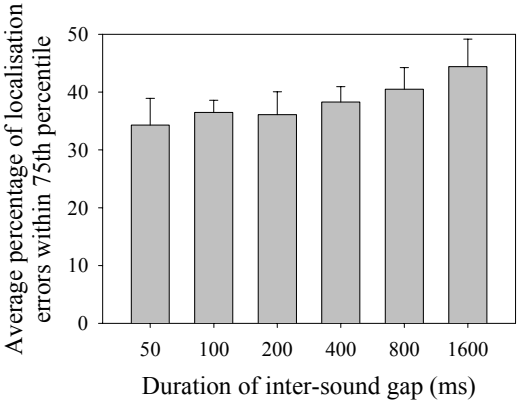

Figure 1. *Average percentage of trials on which the localisation error was less than or equal to the participant's 75th percentile error for localisation without distracter sounds for each of the inter-sound gap durations. Error bars show one standard error of the average.*

4. CONCLUSIONS

The results of this study are consistent with our expectation that increasing the gap between target and distracter sounds, under conditions in which participants were informed of the identity of the target sound after stimulus presentation, would enhance the accuracy with which the target sounds were localised. As noted earlier, such enhancement could result via several mechanisms. Increasing the duration of the inter-sound gap could help listeners retain the identities and locations of multiple sounds by allowing them more time to rehearse this information between sound presentations. It could also result in more accurate localisation of multiple sounds by allowing listeners' processing resources to be devoted to each sound for a longer period of time. At any rate, this study indicates that the enhancement obtained when the inter-sound gap is increased from 50 to 1600 ms is not particularly large, at least in the case of target sounds immersed among five distracter sounds.

That the highest average percentage of trials on which the localisation error was less than or equal to the participant's 75th percentile error for localisation without distracter sounds in this study was only 44.4% (for a 1600-ms gap) indicates that listeners experience considerable difficulty when required to identify and localise six sounds and retain that information for several seconds. (The expected value of this metric in the absence of an effect of distracter sounds is, of course, 75%.) This suggests that operators would find it difficult to track the location of six sounds in a 3D audio display. It should be noted, however, that the task in this study may be more difficult than that of an operator using a 3D audio display. In this study the location of sounds on any given trial was completely arbitrary. For operators using a 3D audio display, the location of sounds would be mapped to some data set. Knowledge concerning that data set, such as the spatial locations normally occupied by particular objects and the way the locations of those objects normally change across time, could help operators extract and retain information from the display.

5. REFERENCES

[1] R.L. Martin, K.I. McAnally, and M.A. Senova, "Free-field equivalent localization of virtual audio," *J. Audio Eng. Soc.*, vol. 49, no. 1/2, pp.14-22, Jan/Feb. 2001.

[2] M.D. Good, R.H. Gilkey, and J.M. Ball, "The relation between detection in noise and localization in noise in the free field," in *Binaural and Spatial Hearing in Real and Virtual Environments*, R.H. Gilkey and T.R. Anderson (eds.), Erlbaum, Mahwah, U.S.A., 1997, pp. 349-376.

[3] C. Lorenzi, S. Gatehouse, and C. Lever, "Sound localization in noise in normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 105, no. 3, pp. 1810-1820, Mar. 1999.

[4] S. Getzmann, "A comparison of the contrast effects in sound localization in the horizontal and vertical planes," *Exp. Psychol.*, vol. 50, no. 2, pp. 131-141, 2003.

[5] V. Best, A. van Schaik, C. Jin, and S. Carlile, "Sharing auditory space," in *International Workshop on Spatial and Binaural Hearing*, Utrecht, The Netherlands, June, 2003.

[6] E.H.A. Langendijk, D.J. Kistler, and F.L. Wightman, "Sound localization if the presence of one or two distracters," *J. Acoust. Soc. Am.*, vol. 109, no. 5, May 2001.

[7] R.L. Martin, P. Flanagan, K.I. McAnally, and G. Eberle, "Localisation of sequentially presented sounds," In preparation.

[8] D.B. Watson, R.L. Martin, K.I. McAnally, S.E. Smith, and D.L. Emonson, "Effect of normobaric hypoxia on auditory sensitivity," *Av. Space Environ. Med.*, vol. 71, no. 8, pp. 791-797, Aug. 2000.

[9] J.F. Corso, "Age and sex differences in pure-tone thresholds," Arch Otolaryngol., vol. 77, pp. 385-405, Apr. 1963.

[10] P.G. Stelmachowicz, K.A. Beauchaine, A. Kalberer, and W. Jesteadt, "Normative thresholds in the 8- to 20-kHz range as a function of age," J. Acoust. Soc. Am., vol. 86, no. 4, pp. 1384-1391, Oct. 1989.

[11] B. Zhou, D.M. Green, and J.C. Middlebrooks, "Characterization of external ear impulse responses using Golay codes," *J. Acoust. Soc. Am.*, vol. 92, no. 2, pp. 1169-1171, Aug. 1992.