INTERSPEECH
2009 BRIGHTON

Audio spatialisation strategies for multitasking during teleconferences

Stuart N. Wrigley[1], *Simon Tucker*[2], *Guy J. Brown*[1], *Steve Whittaker*[2]

[1]Dept. of Computer Science, [2]Dept. of Information Studies
University of Sheffield, Regent Court, 211 Portobello, Sheffield S1 4DP, United Kingdom
{s.wrigley, g.brown}@dcs.shef.ac.uk, {s.tucker, s.whittaker}@sheffield.ac.uk

Abstract

Multitasking during teleconferences is becoming increasingly common: participants continue their work whilst monitoring the audio for topics of interest. Our previous work has established the benefit of spatialised audio presentation on improving multitasking performance. In this study, we investigate the different spatialisation strategies employed by subjects in order to aid their multitasking performance and improve their user experience. Subjects were given the freedom to place each participant at a different location in the acoustic space both in terms of azimuth and distance. Their strategies were based upon cues regarding keywords and which participant will utter them. Our findings suggest that subjects employ consistent strategies with regard to the location of target and distracter talkers. Furthermore, manipulation of the acoustic space plays an important role in multitasking performance and the user experience.

Index Terms: multitasking, spatialisation, teleconference

1. Introduction

Businesses of all sizes are coming under increasing pressure to reduce costs and improve efficiency. Since employees spend significant amounts of time in meetings [1], this is one area in which companies are investing in IT and communication technologies to ease the financial burden. Meetings, especially ones which involve some travel component, can be very costly. Consequently, organisations are achieving cost reductions by adopting 'virtual meetings' [2].

However, despite increasing meeting commitments, employees are still expected to meet their productivity goals as normal. In order to achieve this, it is becoming increasingly common for participants — generally located at their office desk — to multitask during virtual meetings [3]. Since virtual meeting participants are more susceptible to confusion due to the unavailability of non-verbal communication [4], it is important that the technology used to present the meeting to the participant does so in a manner that allows them to multitask with greatest efficiency.

Our previous study [5] examined three different techniques for presenting the audio from a virtual meeting to the listener: *mono*, *dichotic* and *spatialised*. When presented in *mono*, the audio presentation was equivalent to a conventional teleconferencing system in which all audio streams are mixed together in equal proportions and presented via a single audio channel. The *dichotic* style was similar to everyday stereo in which there are two independent audio signal channels (one for the left ear and one for the right ear). When presenting two talkers, one would appear in the left ear and the other in the right ear. If there were more than two talkers, one or more would be mixed in equals proportions and presented to the left ear whilst the remaining talkers would also be mixed but presented to the right ear. The final technique — *spatialised* — involved creating an audio-based 'virtual reality' in which each talker was presented in such a way as to give the impression that they were located at a particular position around the listener's head. In this situation, it was possible to simulate what the listener would have heard had they been present in the room when the recording took place. Furthermore, it was possible to simulate *any* spatial configuration of the recorded talkers.

The experimental subjects were given the task of listening for a keyword to be uttered in the audio whilst performing a screen-based text manipulation task using the mouse. This scenario closely matches the situation in which the virtual meeting participant is simultaneously present in the virtual meeting as well as performing a standard office activity. In such scenarios it is common for the virtual meeting participant to multitask [3].

The study established that the use of spatialised audio significantly increased multitasking efficiency [5]. Interestingly, in the spatialised audio presentations, we expected listeners to prefer the keyword to appear from directly ahead. However, despite the analysis showing no performance advantage associated with direction, all subjects who indicated a preference to keyword location stated left or right: none preferred straight ahead.

The purpose of the current study is to investigate how subjects themselves arrange meeting participants given knowledge of the talker who will utter the keyword. Specifically, we are interested in the positioning of the target talkers relative to the interfering talkers; subjects' strategies for reducing the influence of the interfering talkers; and also what effect these strategies have on their multitasking performance. The experiment was conducted in a similar fashion to that of [5] with the exception that subjects were given control over azimuthal location of the talkers and also, in some scenarios, the simulated 'distance' of each talker relative to the subject.

The rest of the paper describes the experimental protocols, followed by the presentation and analysis of results. The paper concludes with a discussion of our findings.

2. Experiments

To simulate the type of multitasking in question, the experiment consisted of an audio-based task and a text-based task which were performed concurrently. Subjects used a computer mouse to find as many occurrences of the letter 'e' as possible from a section of text. For each letter 'e', subjects clicked on the character using the mouse; the time of occurrence and the actual letter clicked were logged allowing the computation of e-spotting rate (e's per second). In each scenario, a different section of text was presented.

All scenarios were accompanied by an audio playback of a meeting recording consisting of three participants. For the concurrent audio-based task, subjects were asked to listen for

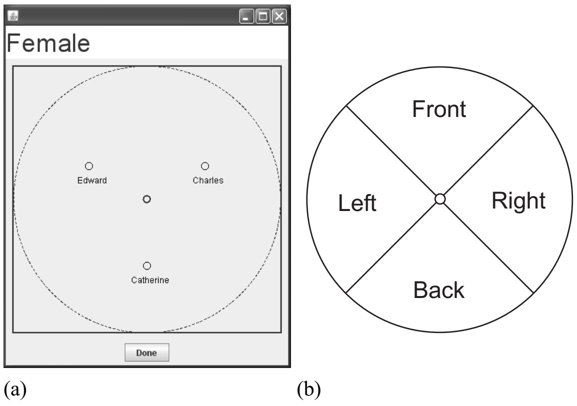


Figure 1: *(a) The interface used to position the three meeting participants within the acoustic space. The small central circle represents the subject and the three captioned circles represent the participants. The large dotted circle represent the furthest distance a participant may be moved to. Note the gender cue in the top left. (b) The split of the acoustic space into four quadrants with the subject represented as the central circle.*

a particular word (the 'keyword') in addition to performing the e-finding task. When they heard the keyword, they were instructed to click a button on the interface. The scenario ended when the keyword was detected or 60 seconds had elapsed.

Before each presentation, subjects were given 30 seconds to arrange the spatial locations of the three talkers based upon knowledge of who would say the keyword. Subjects were told the name or gender of the talker who would utter the keyword: this would allow the subject to either identify the individual talker (*single target*) or narrow the selection to two out of the three talkers present (*dual target*). Positioning was achieved using the interface shown in Fig. 1(a). Thirty seconds of continual speech was created for each talker by concatenating randomly selected sentences uttered by that talker from throughout the meeting. The continual speech recordings were played concurrently during the spatial positioning phase in order for the subjects to hear, in realtime, the effect of moving any of the three talkers. Once the subject was happy with the position of each participant (or 30 seconds had elapsed), the multitasking presentation commenced.

There were two 'levels' of flexibility with regard to the arrangement of talkers available to subjects. In half the presentations, subjects were able to move participants in a fixed circle around their head. This had the effect of altering the azimuth of the participant but not the distance. In the other half of the presentations, subjects could also alter the distance of the participant relative to themselves as well as the azimuth. If the participant was moved to the outer dotted circle, shown in Fig. 1(a), the amplitude of the participant's recording was at a minimum. If the participant was moved toward the circle representing the subject, the amplitude was maximal. The absolute amplitudes of the two extremes were set empirically; it ought to be noted, however, that at the minimum amplitude, the participant recording was still just audible.

2.1. Stimuli

The audio data used in the experiments was taken from a number of meetings within the AMI corpus [6]. In this corpus, each participant is recorded using a separate microphone (channel). The word-level transcripts were used to remove crosstalk from each channel and replace it with silence; this ensured each channel contained only the audio from the participant wearing the microphone. Each channel was amplitude normalised to ensure the RMS values of the speech portions were equal. To homogenise the speech and silence sections, low-amplitude white noise was added to simulate natural recording 'hiss'. Channels were positioned in the virtual acoustic space using the OpenAL (Open Audio Library) audio API (http://www.openal.org).

To identify the audio segments we used the manual transcripts from the AMI corpus and selected a pool of suitable meeting segments. Segments were chosen to be 60 seconds in duration and the start of each segment was aligned with the beginning of an utterance. The segment was chosen to feature the required number and gender of talkers for the given experimental condition. We then analysed each segment, identified any words which occurred once in the duration of that segment and scored the uniqueness of each using a measure of TF*IDF [7].

From this pool of segments and associated unique keywords we then chose the final selection of meeting segments ensuring that keywords had a sufficiently high TF*IDF score and that the keyword occurred at least 20 seconds after the clip started and at most 10 seconds before the clip ended. We also ensured that, for each experiment, the keyword start times were evenly distributed between these two limits.

The text for the e-spotting task was extracted from *The Metamorphosis* by Franz Kafka.

2.2. Procedure

15 subjects were used, namely 8 males and 7 females. All were native English speaking graduates of our university and had some experience with psychophysical experiments. None of the subjects reported hearing difficulties. Subjects received a small reward for participating.

Subjects sat in a single walled sound-attenuating booth (IAC 402-A Audiometric Booth). The audio was presented to a pair of Sennheiser HD 25 SP headphones. The amplitude of the stimuli was set to a comfortable listening level (no direct SPL measurements were taken).

The experiment lasted approximately 25 minutes and subjects had the opportunity to take a break between each presentation if desired.

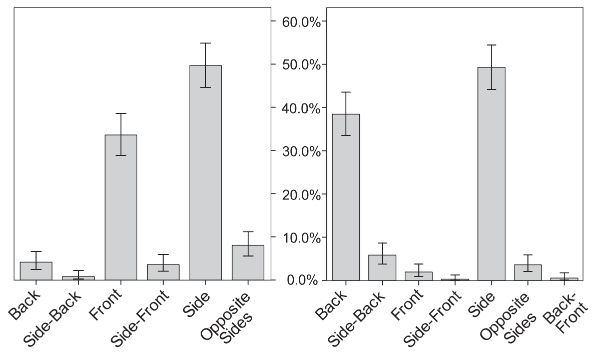
At the end of the experiment, subjects were asked to complete a brief questionnaire to evaluate various aspects of the presentation styles as well as rate the difficulty, or otherwise, of the multitasking scenarios used in the experiments.

3. Hypotheses

We evaluated a number of different hypotheses in this study:
H1: There will be across- and within-subject preferences for talker arrangements.
H2: Subjects will place multiple targets and distracters at similar angular displacements
H3: Subjects will place multiple targets and distracters at similar distances from the themselves.
H4: Subjects will spot e's faster when listening to single targets than when listening to multiple targets.

4. Results

Recall that subjects were cued to the name or gender of the talker who would utter the keyword. Therefore, it was possi-



(a) Target positions (b) Distracter positions

Figure 2: *Frequency (95% CI) with which targets (a) and distracters (b) were placed in quadrants of the acoustic space.*

ble to count the number of e's spotted when the subject was listening to either target or non-target audio.

4.1. H1a: Across-subject talker arrangement preference

To investigate which arrangements were employed by subjects, we split the acoustic space into four areas: front, back, left and right as in Fig. 1(b). Targets were commonly placed in different areas to the distracters (97% of trials exhibited this behaviour). By treating either side location as a single category we can further show that the most common location for target talkers was at the side (49% of trials) or in front of the subject (33% of trials); see Fig. 2(a). For the distracter locations it can be seen that the most common location was either to the side (49% of trials) or behind the subject (38% of trials); see Fig. 2(b). In 55% of trials the target talkers were placed to the side directly opposite the distracters.

In cases when targets were placed to the side of the subject the left side was favoured over the right (73% of relevant trials) and the opposite was true for distracters (66% of relevant trials). In a MANOVA, we found no effect of the overall target location on the target e-spotting rate ($F_{(5,351)} = 1.53, p > 0.1$) nor of the overall distracter location on the non-target e-spotting rate ($F_{(6,349)} = 0.14, p > 0.9$).

4.2. H1b: Within-subject talker arrangement preference

As stated above the most common arrangement was for the target talker to be placed on the opposite side to the distracters. Eight of the 15 subjects chose this as their most frequent strategy; a further 3 subjects preferred to place the targets in front of them with the distracters at the back. Restricting our analysis to the cases where the two most common strategies were chosen we found that there was no effect of the strategy on the e-spotting rate. MANOVA analysis suggests that there was subjective preference for angular separation ($F_{(14,90)} = 4.36, p < 0.05$ for targets and $F_{(14,211)} = 6.89, p < 0.05$ for distracters) and for distance ($F_{(14,150)} = 26.47, p < 0.05$ for targets and $F_{(14,150)} = 17.67, p < 0.05$ for distracters).

4.3. H2: Angular separation

We investigated the angular separation between dual targets and dual distracters. The average separation of target talkers was 56.40° although there were several cases where target talkers were placed in different areas; for example, some subjects placed targets so they appeared in different ears. If we focus on targets that have been placed in the same area, the mean angular separation was 16.60° with a standard deviation of 20.81° (the median separation was 7.59°). A similar analysis can be done for the separation of dual distracters: the average angular separation overall was 63.93°. Again removing the cases where distracters have been placed in different areas results in a mean of 9.34° with a standard deviation of 15.69° (the median separation was 3.05°).

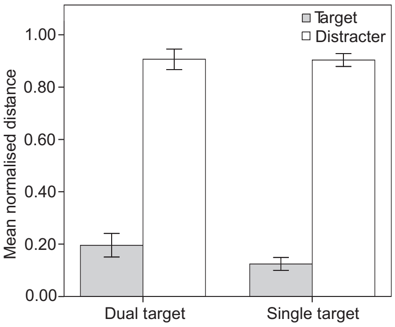


Figure 3: *Normalised target and distracter distances (95% CI).*

These results suggest that subjects generally placed multiple targets or multiple distracters in similar locations although multiple targets were placed further apart (mean difference was 7.26, $p < 0.05$). This suggests that subjects preferred to place distracters at the same location but felt that target talkers should be spatially distinguishable. However, we also found a negative correlation between the angular separation of multiple targets and the resultant target e-spotting rate ($r = -0.27, p < 0.01$) and a weak negative correlation between the angular separation of multiple distracters and the non-target e-spotting rate ($r = -0.12, p < 0.08$).

4.4. H3: Talker distance

Distance measurements within the acoustic space were normalised such that a distance of 0 denoted a talker placed as close as possible to the subject and a distance of 1 denoted a talker situated anywhere on the boundary circle (the dotted line in Fig. 1(a)). Targets were generally placed closer to the subject than distracters (mean distance to target was 0.14 and mean distance to distracters was 0.90; see Fig. 3). In a MANOVA, there was no effect of the overall arrangement on the distance of the distracters ($F_{(6,173)} = 0.68, p > 0.6$) and the difference in distances between two distracters was consistently small (mean of 0.012 with a standard deviation of 0.036). In the dual target scenarios, the talkers were generally placed at similar distances although we found a main effect for the arrangement on the distance from the subject ($F_{(4,175)} = 8.76, p < 0.05$). Tukey post-hoc tests indicated that this was caused by subjects placing targets further away if the targets were on opposite sides (in each case $p < 0.05$). Again the difference in distances was small (mean of 0.022 with a standard deviation of 0.058).

4.5. H4: E-Spotting rate

To investigate the effect of the number of targets on e-spotting rate we carried out a MANOVA with target and non-target e-rates as dependent variables and the number of target talkers and whether the talker distances were fixed as independent vari-

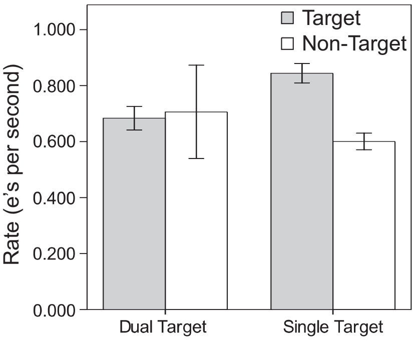


Figure 4: *E-spotting rates (95% CI).*

ables. The results showed that there was a strong effect on the target e-rate ($F_{(1,358)} = 31.05, p < 0.05$) and a weaker effect on the non-target rate ($F_{(1,358)} = 2.79, p < 0.1$); see Fig. 4. In addition to this we found no effect of fixing the distance of the talkers on the resultant e-spotting rates ($F_{(1,356)} = 0.01, p > 0.9$ for target e-rates and $F_{(1,356)} = 0.56, p > 0.4$ for non-target e-rates).

5. Summary and discussion

This study has investigated the audio spatialisation strategies employed by subjects who must perform an audio-visual multitasking exercise. They were given the freedom to move three teleconference talkers within a virtual acoustic space with the goal of maximising their multitasking efficiency. The exercise involved listening for a keyword to be uttered whilst concurrently performing a text-based task with a computer mouse. In some scenarios, they could only adjust the azimuth of each talker; in others, they also had control over the 'distance' of the talker (and hence amplitude). Subjects were cued to the talker who would utter the keyword allowing subjects to identify one or two talkers from the three present.

Our previous study [5] found that subjects exhibited a preference for side presentation as opposed to straight ahead. The present study has confirmed this by showing that a common strategy employed by subjects is to place the target talker(s) on one side and the distracter(s) directly opposite. In the less common approach of placing the target talker(s) ahead, the distracter(s) tended to be placed behind.

Analysis showed that in half of the trials, subjects placed the target talker to one side. This finding agrees with proxemics research (how people use physical space in interpersonal interaction); in two-talker situations, a separation of 90° occurs most often in natural conversation, followed by the two talkers facing each other [8]. We also found that each individual's strategy was largely consistent throughout the experiment. Since nearly three quarters of subjects placed distracters opposite the target talkers this suggests a general strategy of placing targets as prominently as possibly in a single ear and placing the distracters far away and preferably, in the opposite ear.

In cases where there were dual targets, subjects had a tendency to separate the target talkers more than they would separate dual distracters; we found that this increased separation had a detrimental effect on the ability of subjects to spot e's. In addition this is a micro rather than a macro effect: we found no overall effect of the arrangement on the target e-spotting rate, although subjects generally placed dual targets in close proximity to each other. We found similar results for placement of distracters although in this case subjects tended to place distracters into a single 'trash' location.

As expected, subjects made full use of the ability to alter the distance of talkers in the acoustic space. Target talker(s) were consistently moved closer to the subject while distracter talker(s) were moved further away. However, there was no effect of distance on e-spotting rates: there was no difference in performance between the cases where the subject could not alter the distance of the talkers and those where talkers could be placed at any distance.

It was found that the target e-spotting rate was reduced in the dual target scenarios relative to the single target scenarios. This is likely to be caused by the increased uncertainty in the audio monitoring task creating a higher cognitive load. Conversely, the number of distracters did not have any influence over the non-target e-rate: one talker is as easy to ignore as two.

In summary, we have shown that listeners employ consistent and effective strategies to aid their multitasking performance. Our results show that large spatial separations between targets and distracters is preferred and that subjects tend to prefer placing them on opposite sides of their head. Furthermore, the ability to create a realistic acoustic space by using distance cues improves the multitasking experience. The findings presented in this paper raise important points for the design of future teleconference presentation approaches: spatialised audio improves multitasking performance and the ability of the user to create their own acoustic space improves the user experience. We have also found that some of the strategies chosen by subjects were not necessarily optimal from a performance perspective. Subjects preferred to separate target talkers but performed better when targets were closer in acoustic space. Thus, while spatialisation does improve the performance of multi-tasking workers [5], the results of this study suggests that some limits should be imposed on how the spatialisation should be implemented.

6. Acknowledgements

This work was supported by the European Union 6th FWP IST Integrated Project AMIDA (Augmented Multi-party Interaction with Distant Access, FP6-0033812).

7. References

[1] S. G. Rogelberg, C. Scott, and J. Kello, "The science and fiction of meetings," *MIT Sloan Manage. Rev.*, pp. 18–21, Winter 2007.

[2] F. Cairncross, *The death of distance : how the communications revolution will change our lives*. London: Orion Business, 1998.

[3] C. Wasson, "Multitasking during virtual meetings," *Human Resource Planning*, vol. 27, pp. 47–60, 2004.

[4] L. F. Thompson and M. D. Coovert, "Teamwork online: The effects of computer conferencing on perceived confusion, satisfaction, and postdiscussion accuracy," *Group Dynamics: Theory, Research, and Practice*, vol. 7, no. 2, pp. 135–151, 2003.

[5] S. N. Wrigley, S. Tucker, G. J. Brown, and S. Whittaker, "The influence of audio presentation style on multitasking during teleconferences," in *Interspeech 2008*, 2008, pp. 801–804.

[6] I. McCowan et al., "The AMI meeting corpus," in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, 2005.

[7] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *J. Doc.*, vol. 28, pp. 11–21, 1972.

[8] T. M. Ciolek and A. Kendon, "Environment and the spatial arrangement of conversational encounters," *Sociological Inquiry*, vol. 50, no. 3–4, pp. 237–271, 1980.