interspeech 2008 incorporating SST 08

# The influence of audio presentation style on 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

Teleconference participants often multitask: they work on a text-based 'foreground' task whilst listening in the 'background' for an item of interest to appear. Audio material should therefore be presented in a manner that has the smallest possible impact on the foreground task without affecting topic detection. Here, we ask whether dichotic or spatialised audio presentation of a meeting is less disruptive than the single-channel mixture of talkers normally used in teleconference audio. A number of talker location configurations are used, and we examine how these impact upon a text-based foreground task: finding all letter 'e' occurrences in a block of text. Additionally, we examine the effect of cueing the listener to direction or gender and record listener preferences for audio presentation style. Our results suggest that spatialised audio disrupts the foreground task less than single-channel audio when direction or gender is cued.
**Index Terms**: multitasking, spatialisation, teleconference

## 1. Introduction

It has been estimated that the average American employee spends approximately 6 hours per week in scheduled meetings with workers higher in the organisation spending more than 20 hours per week in meetings [1]. The desire to reduce the financial burden of meetings (e.g., travel costs, etc.) coupled with rapid developments of IT and communication technologies have lead many organisations to adopt '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]. Participants in virtual meetings are simultaneously present in two (or more) *interactional spaces*. These interactional spaces are the 'virtual meeting space', the 'local space' and 'other virtual spaces'. In the local space, the participant is alone and possibly engaging in solitary tasks such as reading documents; in the other virtual space the participant may interact with other people (e.g., via email, IM, etc.); in the virtual meeting space the participant adopts either listener or talker roles within a range of interactions (e.g., presentations, discussions, etc.) [3, p. 53].

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.

This study examines three different techniques for presenting the audio from a virtual meeting to the listener. We term these techniques *mono*, *dichotic* and *spatialised*. When presented in 'mono', the audio signals from each talker are mixed in equal proportions and presented to the listener. This is equivalent to standard teleconference approaches. The 'dichotic' technique involves presenting one or more talkers to the left ear and the remaining talkers to the right ear. Finally, the 'spatialised' technique simulates a full 3D sound environment in which each talker in the meeting can be placed at any position around the listener's head. Specific details are described below.

The experimental subjects are 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 described above in which the virtual meeting participant is simultaneously present in the 'virtual meeting space' and the 'local space'. The degree to which each audio presentation technique is conducive to efficient multitasking is evaluated by examining both the subjects' keyword spotting ability and text-based processing performance.

The motivation for investigating these three audio presentation types lies in how listeners process sound environments. It has been proposed that a listener creates a mental representation of a sound environment by subjecting it to a form of Auditory Scene Analysis [5] (c.f., visual scene analysis). This process separates each sound source into a different 'stream' on the basis of a number of cues such as common spatial location. In order to listen to a particular part of a sound environment (e.g., a particular person) selective attention brings that stream to the fore. However, if the listener must be aware of the contents of more than one stream, attention will be divided between them, hence reducing the amount of processing available to each. Indeed, further processing limitations arise when attending to tasks in different modalities such as in this study (see [6] for a review). It is unclear whether monitoring a single stream containing multiple talkers for the occurrence of a keyword involves a higher cognitive load than extracting a spatially distinct stream and hence attending to a particular talker. In this study we investigate how the allocation of attention to one or more streams affects the subject's ability to multitask. Furthermore, we investigate how cuing the subject to a particular location or participant gender can aid the attentional selection process and hence improve their multitasking ability.

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 and an assessment of the subjects' opinions on each presentation technique.

## 2. Experiments

In order to investigate the effect of audio presentation style on a subject's multitasking ability, we developed three experiments


Accepted after peer review of full paper

which all had a common basis. For all experiments, the subject sat at a computer performing a task which involved finding as many occurrences of the letter 'e' as possible from a section of text and clicking on them using the mouse. For each mouse click, the time of occurrence and the actual letter clicked were logged allowing the computation of e-spotting rate (e's per second).

The experiments were split into a number of scenarios lasting 60 seconds. In each scenario, a different section of text was presented. Some (but not all) scenarios were also accompanied by an audio playback of a meeting recording consisting of between three and four participants. When audio was present, the subject was asked to listen for 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.

In the first experiment we asked whether mono and spatialised speech were equally disruptive to the subject's multitasking performance. In the following two experiments we investigated whether there was any benefit in also informing the subject about either the direction from which the keyword would be said or the gender of the participant who utters it.

### 2.1. Stimuli

The audio data used in the experiments was taken from a number of meetings within the AMI corpus [7]. 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. The audio channels were upsampled from the original 16 kHz to 48 kHz to ensure sufficient spatial resolution when spatialised. 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'.

Spatialised signals were created by convolving the original mono audio stream with head related transfer functions (HRTFs) measured from a KEMAR artificial head in an anechoic environment [8] to generate a stereo signal. This allows the talker's audio to be placed at any arbitrary horizontal azimuth relative to the listener's head.

The 'mono' signal used in the experiments below was created by spatialising each meeting participant's microphone recording to $0^\circ$ (straight ahead) and adding all the left channels from the subsequent stereo signals. The resultant signal was mono. Note that spatialisation was used to create the 'mono' condition to ensure any frequency filtering introduced by the spatialisation process was the same across 'mono', 'dichotic' and 'spatialised' conditions.

The 'dichotic' signal was created in a similar fashion to the 'mono' signal. However, here the resultant signal was stereo. One or more meeting participants were placed in the left channel while the remaining meeting participant(s) were place in the right channel.

In the 'spatialised' condition, each meeting participant's audio was spatialised such that it was perceived as originating from a different position around the head (see Figs. 1(c) and 1(d)).

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 speakers for the given experimental condition (e.g. for the direction cued experiments the segments were chosen to feature only three speakers of the same gender). We then analysed each of these segments, identified any words which occurred once in the duration of that segment and scored the uniqueness of each using a measure of TF*IDF [9]:

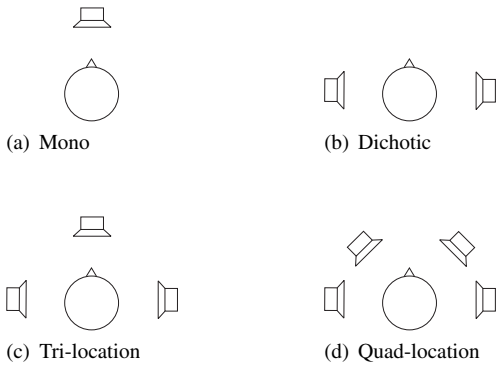


Figure 1: *The simulated location of the audio streams.*

$$imp_{td} = \frac{log(count_{td} + 1)}{log(length_d)} \times log(\frac{N}{N_t}), \qquad (1)$$

where $count_{td}$ is the frequency with which term $t$ appears in document $d$, $length_d$ is the number of unique terms in document $d$, $N$ is the number of documents and $N_t$ is the number of documents which contain the term $t$. Before computing TF*IDF we first exclude stop words (such as *the*, *of*, *and*) using a standard stopword list. The document set used here was the complete set of manual transcripts in the AMI corpus.

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 (1.0 was empirically selected as a minimum 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.

Segments were balanced between subjects ensuring that the same number of subjects heard each keyword under each audio condition. We also ensured that no keyword was repeated for a single subject and that no subject experienced the same audio condition consecutively.

The text for the e-spotting task was extracted from *The Metamorphosis* by Franz Kafka. In each presentation a different, randomly selected, portion was used.

### 2.2. Keyword-only cued experiment

In this experiment, the subject experienced 30 presentations; the presentations were split evenly across three conditions: 'silence', 'mono' and 'spatialised'.

When audio was present, it was drawn from portions of meetings that had up to four participants active over the 60 second duration. Half of the audio presentations were mono (Fig. 1(a)) and the remaining audio presentations were fully spatialised (Fig. 1(d)).

### 2.3. Keyword and direction cued experiment

Subjects experienced 39 presentations of which 3 had no audio and acted as controls. The remaining 36 presentations were split across 'mono' (6 conditions), 'dichotic' (6 conditions each for keyword left and keyword right) and 'spatialised' (6 conditions each for keyword left, front and right).

When audio was present, it was drawn from portions of meetings that had exactly three participants active over the 60 second duration. Three participants were chosen in this experiment (and the gender cued experiment below) to simplify the subjects' task of listening to a particular location. Half the presentations were all male and half were all female participants.

### 2.4. Keyword and gender cued experiment

Subjects experienced 69 presentations of which 3 had no audio and acted as controls. The remaining 66 presentations were split evenly across two categories: 'single' and 'dual'. In each category, there were either one male and two female participants or vice versa. Here, 'single' refers to cueing a gender for which there was only one participant and 'dual' refers to cueing a gender for which there were two participants. For example, 'single' would refer to cueing 'female' in a 'female, male, male' condition. Within each category, presentations were either 'mono' (6 conditions) or 'spatialised' (6 conditions each for keyword left, front and right). Note that the direction of the keyword was not cued.

### 2.5. Procedure

12 subjects were used, namely 6 males and 6 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 HD250 linear II headphones. The amplitude of the stimuli was set to a comfortable listening level (no direct SPL measurements were taken).

Each experiment was conducted in a separate session with the exception of the gender-cued experiment which was split across two 30-minute sessions. 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 tested a number of hypotheses in the current study:

**H1:** Subjects will perform better when listening to spatialised audio.

**H2:** Subjects will perform better when cued as to how the target word is to be presented.

**H3:** There will be no effect of speaker gender on e-spotting performance.

**H4:** There will be no effect of direction (regardless of cue type) on e-spotting performance.

## 4. Results

### 4.1. General observations

Subjects were generally good at the keyword spotting task, only failing to hear the target word in 10 % of the experimental conditions and indicating keyword occurrence in a median time of 4.29 seconds after the start of the target word (the mean word duration was 0.6 seconds). Carrying out two MANOVAs investigating the effects of audio condition, gender and direction on the propensity to miss the target word and the time taken to switch to the meeting we found no main effects ($p > 0.1$ in all cases). We therefore concentrated our subsequent analysis on the subjects' e-spotting ability.

### 4.2. H1: Effect of audio type

To investigate the effect of the audio condition in the keyword-only experiment we carried out an ANOVA with audio condition (mono, silence or spatialised) as the independent variable and e-rate as the dependant variable. The analysis showed a main effect for audio condition ($F_{(2,357)} = 7.886, p < 0.01$) with Bonferonni post hoc tests indicating that subjects were faster in the silence condition ($p < 0.01$) but that there was no difference between the mono and spatialised condition in terms of the e-rates ($p > 0.96$). See Fig. 2(a).

To investigate the effect of audio condition on the ability of subjects to carry out the primary task in the direction-cued and gender-cued experiments we carried out an ANOVA with audio condition (mono, dichotic, spatialised) as the independent variable and normalised e-rate as the dependant variable. The e-rates were normalised by the mean e-rate of the subject in the silent conditions for the given experimental condition. Here we found a main effect for audio condition ($F_{(2,1221)} = 14.773, p < 0.01$) with Bonferonni post hoc tests showing that subjects were faster in the spatialised condition than in either the mono or dichotic conditions ($p < 0.01$ in both cases).

We then counted the number of e's spotted when the subject was listening to portions of audio consistent with the cue and when they were not. Thus we were able to quantify the subject's e-rate when listening to either target or non-target audio. A MANOVA with audio condition (mono, spatialised, dichotic) as the independent variable and target and non-target normalised e-rates as the dependant variable again showed a main effect for audio ($F_{(2,1219)} = 8.724, p < 0.01$ and $F_{(2,1219)} = 5.408, p < 0.01$ for target and non-target respectively). Bonferonni post hoc tests indicated that when the subject was listening to the target audio the spatialised audio was superior ($p < 0.01$ in both cases; Fig. 2(b)) but that when the subject was listening to non-target audio there was no difference between dichotic and spatialised ($p > 0.9$) and that spatialised audio was superior to mono ($p < 0.01$; Fig. 2(c)).

### 4.3. H2: Effectiveness of each cue on e-spotting rates

Analysis of the results confirms H2. We carried out a paired samples t-test in the direction-cued and gender-cued experiments which showed a significant difference between the e-spotting rates ($p < 0.01$). We also examined the effect of the cueing type on the e-rate by carrying out an ANOVA with cueing type (direction, gender) as independent variables and normalised e-rate as the dependant variable. The results show a main effect ($F_{(1,1220)} = 68.268, p < 0.01$) indicating that subjects performed better when cued to the gender of the target speaker than the direction of the target speaker.

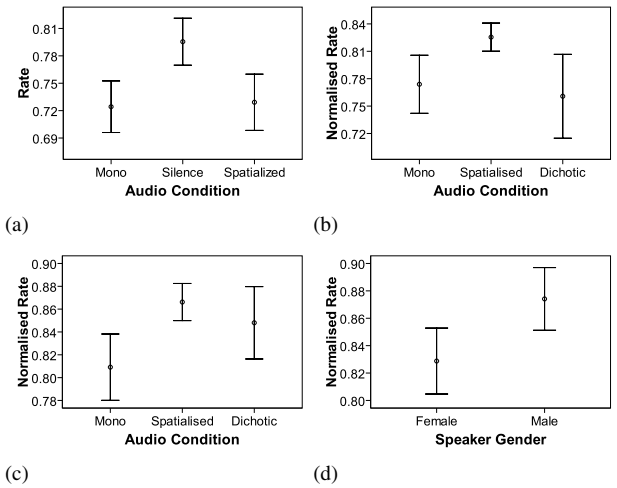


Figure 2: *(a) E-spotting rates (e's per second) organised by audio condition for the keyword-only experiment. Normalised e-spotting rates for the keyword and direction/gender cued experiments for (b) target portions and (c) non-target portions. (d) Normalised e-spotting rates for the non-target portions of the direction-cued experiment organised by speaker gender.*

### 4.4. H3: Preference of speaker gender

To analyse the effects of gender we looked at each of the cueing experiments separately since the mix of participant genders was different in each. To examine the effect of speaker gender in the direction cued experiments we carried out a 2 (gender of subject) $\times$ 2 (gender of speakers) MANOVA with normalised e-rate and target and non-target e-rates as dependant variables. This analysis found a main effect for gender for e-rates ($F_{(1,428)} = 5.167, p < 0.05$) which the results showed to be limited to the non-target e-rates ($F_{(1,428)} = 7.426, p < 0.05$; $p > 0.3$ for target e-rates). Further analysis showed that subjects were able to spot e's faster when listening to male speakers than when listening to females (Fig. 2(d)). We carried out the same analysis in the gender cued experiments and again found an effect of gender on the non-target e-rates ($F_{(1,786)} = 3.804, p < 0.05$) and a further analysis shows that subjects were able to spot e's faster when listening to male speakers. H3 is therefore refuted.

### 4.5. H4: Effect of direction of presentation

To investigate the effect of direction on e-rate we removed the mono conditions from the direction-cued and gender-cued experimental results and carried out a 2 (gender or direction cued) by 3 (90, 0, -90 degrees) ANOVA with normalised e-rate as the target variable. The results showed that there was no effect of the direction on the e-rate ($F_{(2,1000)} = 0.711, p > 0.49$) nor any interaction between the direction and the cueing type ($F_{(2,1000)} = 0.382, p > 0.68$). H4 is confirmed.

### 4.6. Questionnaire results

The post experiment questionnaire indicated that subjects were split as to whether they preferred direction or gender as a means of cueing (50 % in both cases). If the preferred directions are recoded into three categories (cued either left or right, cued straight ahead and no preference) then the majority of subjects preferred to be cued either left or right (58 %) but none stated a preference for being cued in the forward direction. The majority of subjects did not state a preference for gender cueing. In line with the results presented above, the majority of subjects felt they could increase their e-spotting rate when listening to non-target audio segments (83 %).

## 5. Summary and discussion

This study has investigated whether different audio presentation styles can have an effect on multitasking efficiency in teleconferences. Our first experiment demonstrated that when a listener is cued only by keyword, spatialised audio presentation provides no improvement over mono (Fig. 2(a)). However, in the more realistic scenarios in which the listener has more information about who will utter the keyword, significant differences were observed. The most important of these is the increased multitasking efficiency when listening to spatialised audio (Figs. 2(b) and 2(c)). This suggests that extracting a spatially distinct stream and subsequently attending to it involves a lower cognitive load than simply attending to a single stream containing multiple talkers.

Unexpectedly, listeners were found to spot e's faster when listening to non-target male speakers than when listening to non-target females (Fig. 2(d)); i.e., males voices were easier to ignore. We also expected listeners to prefer the keyword to appear from directly ahead. However, despite the experimental results showing no performance advantage associated with direction (H4), all subjects who indicated a preference stated left or right: none preferred straight ahead.

In future experiments we will investigate the interaction between personal preference and multitasking performance further by allowing the subjects to position the participants in a virtual auditory space. In addition to spatialisation, we intend to incorporate a distance metaphor by allowing subjects to place participants of less relevance to their interests further away (i.e., lower amplitude and increased reverberation).

## 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] A. S. Bregman, *Auditory Scene Analysis. The Perceptual Organization of Sound*. MIT Press, 1990.

[6] J. Duncan, "Brain mechanisms of attention," *Q. J. Exp. Psychol.*, vol. 59, no. 1, pp. 2–27, 2006.

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

[8] W. G. Gardner and K. D. Martin, "HRTF measurements of a KEMAR," *J. Acoust. Soc. Am.*, vol. 97, no. 6, pp. 3907–3908, 1995.

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