ONSET ASYNCHRONY IN SPOKEN MENUS

Alistair F. Hinde, Michael Evans

BBC Research and Development,
5 Dock House, MediaCity UK,
Salford, M50 2LH
United Kingdom
afh508@york.ac.uk,
michael.evans@bbc.co.uk

Anthony I. Tew, David M. Howard

Audio Lab, Dept. Electronics,
University of York,
York, YO10 5DD
United Kingdom
tony.tew@york.ac.uk,
david.howard@york.ac.uk

ABSTRACT

The menu is an important interface component, which appears unlikely to be completely superseded by modern search-based approaches. For someone who is unable to attend a screen visually, however, alternative non-visual menu formats are often problematic. A display is developed in which multiple concurrent words are presented with different amounts of onset asynchrony. The effect of different amounts of asynchrony and word length on task durations, accuracy and workload are explored. It is found that total task duration is significantly affected by both onset asynchrony and word duration. Error rates are significantly affected by both onset asynchrony, word length and their interaction, whilst subjective workload scores are only significantly affected by onset asynchrony. Overall, the results appear to suggest that the best compromise between accuracy, workload and speed may be achieved through presenting shorter or temporally-compressed words with a short inter-stimuli interval.

1. INTRODUCTION

The menu is a common feature deployed in user interfaces to allow users to navigate and find content of interest. With the development of ever more sophisticated search algorithms, it may seem as though the menu's role in user interfaces is soon to be confined to legacy software. Search functions, however, work best with well defined target items. In interfaces for entertainment systems, such as the electronic-programme-guide (EPG) on televisions, users may have only loose criteria governing their search (e.g. they may be wanting to view a comedy or plan their evening's viewing [1]). Within browsing scenarios such as this, search functions may, at best, reduce the size of the menu structure that must be traversed, as it is highly likely that users will still have to navigate lists of possible matches. It is also worth noting that these browsing activities additionally expose users to information about alternative, or new content.

This work was conducted by BBC Research & Development and the University of York, funded by an EPSRC Industrial CASE award.

Menu structures can be large and confusing to interact with and they pose a particular problem for people who are unable to attend the screen visually. Due to these complexities, non-visual representation of menus is an area of auditory display which has attracted a great deal of research activity. Traditionally, menus are represented using text-to-speech (TTS) rendering, as is found in commercially available screen readers or in telecommunications. This approach is effective as it allows large amounts of complex and novel information to be displayed. Whilst speech is undoubtedly a logical representation for textual information, temporal redundancy means that there are many situations in which it is not necessarily the quickest method of communicating information. Researchers looking for speed improvements have typically turned to non-speech methods, which aim to represent the information through instrumental motifs [2], ecological sounds [3] or sped-up speech [4]. These approaches have proven to be effective to varying degrees (compared in [4]), but where information is regularly updated and often unfamiliar, as in the case of an EPG, or where a large amount of information needs to be represented, the usability of such systems is likely to be severely affected and any advantages in terms of navigational speed could be lost. In such systems, the redundancy in speech could be a distinct advantage for reducing confusion.

The problem of how to make speech interfaces faster to use is well known and a popular solution is simply to increase the rate of the speech in the system, as regular users of screen readers are commonly reported as doing (e.g. [5]). However, an alternative solution is to increase the amount of information available to the user by using several talkers at once. This form of display, referred to as a concurrent-speech, or multi-talker display, is more analogous to the manner in which visual displays are used, in that a large amount of information is displayed to the user, who then chooses to attend to the item which they are interested in. To facilitate this behaviour, however, the concurrent speech must be separable as individual perceptual streams by the auditory system [6].

As the amount of information presented to a user at any moment increases, one would expect a user to have to work harder to focus on the desired stream of speech. Much of the work on the use of multiple streams of concurrent speech has been in the field of military communications (e.g. [7, 8, 9]), in which high workloads may be necessary to ensure that time critical information is received. By contrast, consumer menu interfaces are likely to have a lower threshold for what constitutes an acceptable amount of effort. Nevertheless, several authors have proposed auditory displays using concurrent speech for menu navigation in consumer

HCI applications.

Frauenberger and Stockman proposed a design using concurrent speech to navigate auditory menus using the idea of a virtual horizontal dial with items located around its perimeter [10]. The display used a virtual room with the centre of the dial positioned outside so that a maximum of three items from the menu would be inside the room, and therefore audible, at any time, along with two additional 'preview sources' if the selected item was a submenu [11]. The user navigated the menu by rotating the virtual ring using a game pad dial until the desired item was directly in front. The display made use of different voice identities and talking styles (i.e. voiced or whispered) to reduce between-stream confusions. When compared with the performance of a traditional screen reader, navigation times were found initially to be faster but in the later trials participants became faster with the screen reader. This was attributed to fatigue effects caused by the repetitive presentation within the prototype display.

Ikei *et al.* [12] proposed the use of multiplexed speech, where delays were included between the onset of successive spoken menu items. Speech sources were spatialised and a range of onset intervals, between zero and 500 ms, were assessed. Trials consisted of between two and four sources with different source orderings (i.e. from left to right or alternating between sources from either hemisphere) and either with or without a linear increase in attenuation applied over the course of each word. The work examined the impact of these display variations on participants' ability to identify the temporal or spatial location of a target word within the mixture. It was found that with three or more voices high accuracy ($\geq 99.7\%$) could be achieved with onset delays of 200 ms, however, this increased to 300 ms if adjacent sources were used and no attenuation applied. As noted by Ikei *et al.* [12], the optimal asynchrony without attenuation (300 ms) is about half of the duration of the stimuli (530 - 600 ms). This is important when considering the playback of more than two voices because when the onset asynchrony is less than half the duration of the stimuli, all three items overlap, but when the onset asynchrony is half the duration or greater, a maximum of two items are presented at any one time (Fig. 1). This simplifies the task in terms of both attentional load and signal-to-noise ratio. If the optimal asynchrony was purely due to this effect, a similar behaviour would not be expected within the two-talker condition. The two-talker results appear to indicate improving performance with increasing onset asynchrony, but ceiling effects in the participants' performances make it difficult to ascertain the strength of this effect.

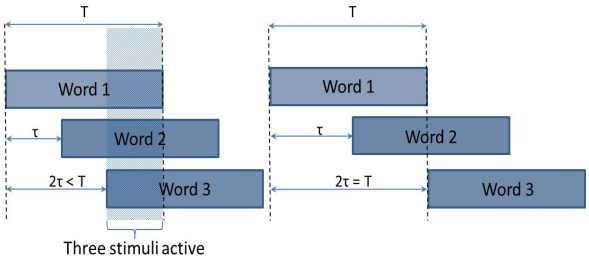


Figure 1: The reduction in the number of concurrent sources when an onset asynchrony of 50% or greater is used (adapted from [12]).

In his thesis, Parente proposed and tested an auditory display system using spatialised concurrent speech for computer-based GUI tasks [13]. The display consisted of five speech sources with differing vocal characteristics (accent, sex, identity), each responsible for reporting different types of information. In addition to this, speech was presented with a 200 ms onset asynchrony between streams. Testing, which compared task performance of the system to that with a conventional screen reader, showed that participants were able to navigate with reasonable accuracy. Interpretation of task durations was complicated due to the different interaction capabilities of the different displays (i.e. availability of search functions) and therefore it is difficult to determine what degree of advantage was provided through the use of concurrent speech.

In scenarios in which content is displayed to the user, who then elects to attend a particular stream (e.g. [14]), it would appear indisputable that concurrency has the potential to reduce presentation times. When users are expected to interact with concurrent speech displays, however, a reduction in task time is not a foregone conclusion. Users may take longer to respond due to the additional processing required to disentangle the contents of the display, or may be more likely to make mistakes due to reduced intelligibility. Therefore, there is still some uncertainty over the amount of time saving which these displays are able to provide. The usage of onset asynchrony alongside pitch and spatial separation of speech streams appears to be promising for reducing navigation times. Due to the inherent trade-off between improving response accuracy [12] and increasing overall presentation time, onset asynchrony is a factor which needs further consideration regarding its impact on overall navigation times. In addition to this, it is still unclear whether a display designer interested in the use of multi-talker display should specify onset asynchrony or overlap. The aim of this paper is to provide deeper insight into the effects of overlap and onset asynchrony in multi-talker menu displays with a particular focus on navigation time.

2. DISPLAY DESIGN

With a concurrent auditory presentation of speech, it is clear that users find it harder to identify and detect accurately the displayed words as the number of concurrent speech streams is increased [15, 9]. Ikei *et al.* [12] found that, when onset asynchrony is introduced into the display, very high accuracy can be maintained for greater numbers of talkers. The greater the number of concurrent talkers, the larger the potential saving in terms of navigation times. To investigate these issues a three-talker design was used in this study.

Many previous auditory displays using concurrent speech have made use of binaural processing to spatialise the audio sources (e.g. [10, 12]). For two reasons, it was decided to use intensity panning to lateralise the stimuli in this study. Firstly, when only two or three concurrent speech streams are used, it is not clear what advantage binaural spatialisation provides over intensity-panning-based lateralisation, as with intensity panning it is possible to confine sources entirely to one channel and therefore remove the energetic masking caused by sound arriving at the contralateral ear. Secondly, binaural spatialisation introduces perceptual factors which tend to vary between participants (e.g. externalisation and front-back confusions), whereas intensity-panning does not.

The display was designed to minimise the amount of interactions required. It is based on the idea of allowing the user to select between three items at a time or move on to the next set of three

items (referred to henceforth as a *triplet*) (see Fig. 2). This method was chosen as it reduces the number of interactions required compared to a display in which the target must be at a specific location for selection, effectively forcing the user to move through the list one item-at-a-time, as in [10].

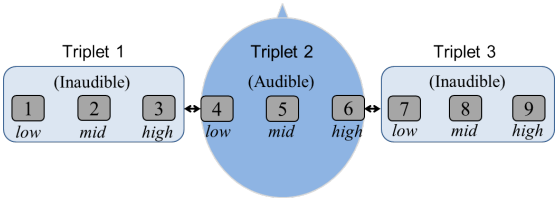


Figure 2: Illustration of display design concept (italicised writing refers to stimuli pitch).

The sources within a triplet were presented from maximally separated lateral positions using intensity panning, such that one source would appear on either side of the head through being presented to only the ipsilateral ear, whilst the third source was presented at the same level in both channels so as to appear in the centre of the head. Within a triplet, sources were also distinguished by a pitch difference such that the stimulus on the left was lowest and the stimulus on the right was highest. The order of presentation was kept constant and ran from left to right to correspond with normal reading direction. Whilst other studies have found advantages in modifying the talker's apparent sex through vocal tract length modification [16], this was not attempted in this study to avoid making the voices sound excessively unnatural.

The system was controlled using five of the drum pads and one rotary dial on a USB MIDI controller. Two rows of pads on the device were assigned to different functions. Playback controls (start playback and navigation to the next triplet) were provided using two pads on the top row and selection of the three items in the currently selected triplet was performed with three pads on the bottom row. An earlier iteration of the prototype also allowed the user to navigate back in the list to a previously heard triplet, and to repeat the current triplet. This functionality was subsequently removed to minimise variance in the navigation time data by ensuring that users took the most direct root to the target. As a result of this, more errors were expected to occur than if these functions had been left in.

The prototype system was developed in Pure Data (Pd-extended 0.43.4) [17] and run on Ubuntu 12.04 LTS with a low latency kernel. To reduce computational load and the possibility of error the triplets were processed in advance, as discussed in Section 3.1. Therefore, the patch was mainly responsible for receiving the MIDI messages from the controller, handling trial configurations, file playback and recording responses.

3. EXPERIMENT

Sixteen participants were recruited from amongst the BBC Future Media department staff and its visitors. Volunteers who reported hearing impairments were not included in the study but no audiometric testing was performed on the participants. No attempt was made to recruit participants with vision disabilities, since the utility of a non-visual display is not solely restricted to people with visual impairments or who are blind.

During the experiment two participants experienced a fault in the software. For one of the participants this fault affected the experimental trials; therefore, this participant was excluded from the study and an additional participant was recruited to fill the space.

3.1. Stimuli

The wordlists from the Modified Rhyme Test (MRT) [18] were used as the source of the words for the experiment. The MRT wordlists consist of two sets, each of which is made up of 25 lists of 6 words. Stimuli were selected from the first set, within which each list of words shares the same first consonant-vowel pair but differs in the final consonant (e.g. *page, pale*), or in some cases has no final consonant (e.g. *ray*). For this experiment 22 of the lists were chosen and for each of the lists one word was removed. Removals were mostly made because of the americanised pronunciations required for some words to share common vowel sounds (e.g. *pass, pat*), the lack of a final consonant, or because some words were deemed too unusual or inappropriate for the experiment.

The 110 words were recorded being spoken by a male talker, who was asked to enounce with minimal intonation and variation in word duration. Recordings were captured as 24-bit audio files with a sample rate of 44.1 kHz. Words were cropped to remove silences before and after they were spoken. Where possible, the crop was made at a zero crossing. In a few cases, however, no suitable zero crossing was available and the crop introduced a small discontinuity into the waveform. In such cases, the word was auditioned to ensure that no click could be heard. For some fricative consonants, low-level sounds at the start or end of the word were removed if they were considered not to be contributing to the intelligibility of the word. The durations of the stimuli were found to vary quite considerably, ranging from 301 to 709 ms with an average of 486 ms.

The stimuli were manipulated to have the same constant pitches and durations (either 360 or 600 ms) in Praat [19]. The pitch values of words in the centre of a triplet were adjusted to correspond to the average pitch of all of the stimuli. The pitches of the words on the right or the left in a triplet were adjusted approximately to plus or minus one ERB [20], respectively, compared to the pitch of the centre word. While other authors have reported benefits from much larger pitch differences [16], a comparatively small pitch difference was used here, as there was concern that further modification would have jeopardised intelligibility and may have led to users becoming distracted by the unnatural character of the voices. The durations were chosen to ensure that no word would be stretched to more than twice, or shortened to less than half, of its original length.

The stimuli were processed such that the words appeared to be approximately the same loudness in all onset asynchrony conditions. The onset asynchrony varies the amount of overlap between the words and this varies the overall loudness of the presentation. While it would have been possible to normalise the loudness of all presentations, this would have altered the levels of the words between presentation conditions. It was therefore decided that it would be more consistent to preserve the variations in overall presentation loudness. To achieve this effect the stimuli were mixed to equal loudness by ear.

The stimuli were combined into triplets and onset asynchronies were adjusted using MATLAB to ensure that they were as accurate as possible. Each triplet presented in the experiment met

the conditions that all words had to be from the same list, with the same word length and each word could only appear once within that triplet. This resulted in the creation of 10,560 triplets. The highest peak amplitude in the set of triplets was found and used to calculate the scaling factor necessary to bring this peak to an amplitude of (+/-) 0.9999 so as to maximise signal-to-noise ratio whilst avoiding any clipping distortion. This scaling factor was applied to all of the stimuli to ensure the relative loudness was not altered. The triplets were then exported as 44.1 kHz, 16-bit WAV files.

3.2. Procedure

The independent variables were onset asynchrony and word length. These variables had four [180, 280, 380, 480 ms] and two [360, 600 ms] levels respectively. The experiment was structured as a within-subjects design, with all participants experiencing all presentation conditions. Experimental trials were split into sessions of fixed word length. Each session contained four blocks in which the onset asynchrony was kept constant. This structure was imposed on the trials so that NASA TLX subjective workload assessments could be performed on each of the word length/onset asynchrony combinations. On completion of each block, a computer based version of the evaluation [21] was undertaken by each participant.

Each trial started with a target word being displayed on a screen and then, when the user was ready, they pressed the 'start' pad which immediately played the first triplet in the list. The user then would navigate until they found the target, whereupon they would select it by pressing the appropriate pad. In some conditions the target word was not present in the list, in which case the correct response was to navigate onwards from the final triplet.

Each list in the experiment was nine words long, with the target words, if present, only presented once at one location. In trials in which the target was present the lists were constrained such that each triplet had at least one stimulus that was not in the previous triplet; all words in a triplet were different; the target only appeared at the target location; all non-target words had to appear twice and never in the same lateral location. These constraints were put in place to ensure that there was variation between the triplets in a list and therefore avoid participants being able to rely on simply detecting that the triplet had changed. This led to 960 possible list combinations for each word at each location within the list. When the target was not included in the list only one item could appear three times, whilst all others would appear twice and words would never appear in the same lateral positions. These criteria led to 576 possible lists for each word.

The experiment was split into three sessions with twenty-minute breaks between them to reduce the effects of any fatigue. In the first (training) session participants completed an informed consent form and then were introduced to the system, an initial playback level was set (participants had the ability to adjust this throughout the experiment) and they were given practice tasks. During the training, each participant performed 40 tasks which consisted of 4 blocks of 10 trials, one block for each onset asynchrony. Each block consisted of all target locations and these were pseudorandomly allocated a word length condition such that 50% of each block was of each condition. Target locations were pseudorandomly ordered so that for each participant each target location could appear once for each trial number within a block. For each of the blocks the participant completed a NASA TLX questionnaire. Participants were given additional guidance when it appeared they had not fully understood how to use the setup or were unclear on how to respond to the TLX questions.

The second two sessions consisted of the experimental trials (limited to a maximum of 20 minutes each), with each session containing one of the word-length conditions. Half of the participants were presented with the short words first and the other half heard the long words first. To reduce the influence of ordering on the onset asynchrony conditions over the training and experimental sessions, three counterbalanced Latin squares were used to vary the orderings. For each instance of the Latin square the dummy values were substituted pseudorandomly for onset asynchrony conditions, such that no dummy value represented the same onset asynchrony condition twice. A row from each of the Latin squares was then used for each of the sessions. The order in which the Latin squares were used for the sessions was varied for every four participants to produce variations for all 16 participants.

Within each block, target location order was varied pseudorandomly, with the restriction that for each participant each target location could appear no more than twice at each trial index in the training and experimental sessions and not at the same trial index as in the previous session. Each trial's list was randomly chosen from the 22 possibilities such that the same list was never used in two consecutive trials. The target word was randomly chosen from the list. The experiment list was then randomly chosen from all possible lists for which the target was at the specified location.

To ensure that data points were captured for all target locations, trial accuracy information was output from Pure Data and read by a Python script. This script analysed the configuration of the original trials and generated new repeat trials when a participant failed to select the target. These repeat trials were then added to the list of trials being read by Pure Data. Repeat trials were reordered and modified so that they used a different one of the 22 wordlists to both the preceding trial and the trial to be repeated. This ensured that the target identity and list were also different so that a participant's previous exposure to the same target location would have minimal effect on their performance. Two additional dummy trials were added using target locations for which the user had already registered an accurate response. These served as a buffer zone in the event that the participant's response to the final trial in a block was incorrect. As the Python script was editing the input to the Pure Data program while the participant was using it, the target location of the last output trial was used to decide where the repeated trials should be added. The repeats were added to the end of the original 10 trials until the last completed trial was beyond the eighth trial, at which point repeats were added after two trials. This ensured that no trials were altered after the user had already begun them and that a repeated trial was always separated from the original by at least two trials. For the repeats the list was randomly selected from the 960 (or 576 if for a 'no-target' trial) possibilities, making the chances of a trial sharing the same list as another trial negligible.

Following the experimental tasks participants were asked a series of questions regarding their experience with the interface (details of which are beyond the scope of this paper) and were then debriefed.

4. RESULTS

During the running of the experiment, on four occasions it was clear that the participant made several attempts to navigate to the

next triplet but had not pressed the pad with sufficient force, causing a significant delay in their navigation time. The experimenter flagged these trials during the experiment and repeats were generated, as if they had been incorrectly answered. Of the four affected trials, one had been a dummy trial. The original affected trials were removed from all subsequent analyses, with the data from the repeated trials being used in their place.

All statistical analysis was performed using SPSS. Further information on the statistical tests used can be found in [22].

4.1. Total task duration

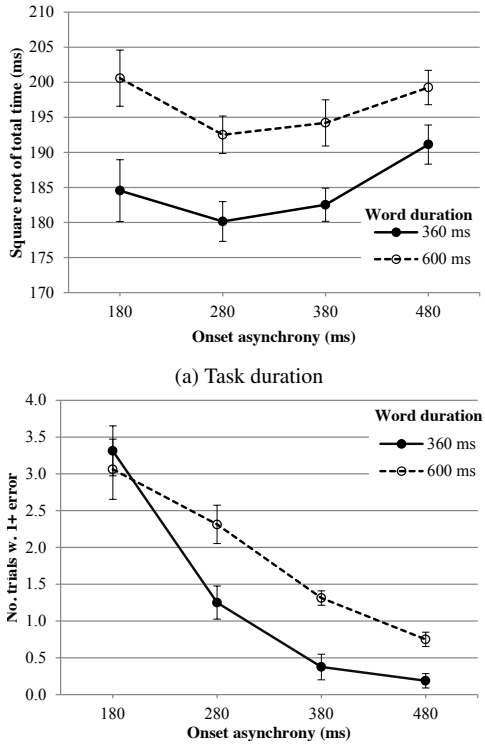
The duration of a trial (i.e. the 'task') was taken as the time from the playback of the first triplet, following the user pressing the 'start' button, to the time when a selection was registered by Pure Data. The task durations of all scoring trials were then summed over all target locations (including when the target was not present) within each onset asynchrony/word duration block for each participant. Trials in which the participant responded incorrectly or which were added as dummy trials were excluded from this sum. This effectively removed the nuisance variable 'target location' from the analysis, leaving each participant one total task duration for each experimental block.

A positive skew at one onset asynchrony/word duration combination was observed. Since this violated the normality assumption required for parametric analysis, the square root of the aggregated task duration data was used. A Shapiro-Wilk test confirmed that the transformed data was not significantly different from normal ($p > .05$). A two-way repeated measures ANOVA (rm-ANOVA) was performed with onset asynchrony and word duration as independent variables.

Mauchly's test indicated that the sphericity assumption was violated for the onset asynchrony condition ($\chi^2(5) = 13.5, p < .05$) and therefore it was decided to use the Greenhouse-Geisser correction ($\epsilon = .60$). Sphericity was met for the word duration (2 levels) and the onset asynchrony $\times$ word duration interaction ($p > .05$). The results of the rm-ANOVA indicated significant effects for onset asynchrony ($F(1.81, 27.1) = 8.79, p = .002, h_p^2 = .369$) and word duration ($F(1, 15) = 25.3, p < .001, h_p^2 = .627$), while the interaction was found to be non-significant ($F(3, 45) = 1.19, p = .323, h_p^2 = .074$) (see Figure 3a). *Post-hoc* pairwise tests were performed for onset asynchrony using a Bonferroni correction, which indicated that the 280 ms onset asynchrony conditions led to significantly shorter total task durations than the 180 ms condition ($p = .038$) and the 480 ms condition ($p < .001$). The total task durations were also found to be significantly shorter in the 380 ms condition than the 480 ms condition ($p < 0.001$). All other comparisons were found to be non-significant ($p > .05$).

4.2. Error rate

The error rates were taken as the number of target locations which required one or more repeats per block (including the not-in-list option). Statistical analysis was performed using generalised estimating equations (GEE) [23]. GEE analysis was chosen because the observed error rates violate the assumptions of normality required for traditional ANOVA-based methods. As the dependent variable was count data, the model was constructed using a Poisson distribution and a log-link function. The working correlation matrix was specified as auto-regressive (AR(1)) because error rates were likely to be more correlated with neighbouring onset asynchrony/word duration conditions. Convergence criteria were set as an absolute difference between iterations of less than 10^{-6}.



(a) Task duration

(b) Error-rate

Figure 3: (a) Marginal means of the square root transformed total task durations (b) Marginal means (original scale) for the number of trials requiring one or more repeats during each block of 10 target locations. The 360 and 600 ms word durations are represented by the solid line with filled markers and the dotted line with hollow markers respectively. (Error bars = $\pm 1 S.E.$)

The model fit values were 118 and 122 (to 3 s.f.) for the quasi likelihood under independence model criterion (QIC) and the corrected quasi likelihood under independence model criterion (QICC) respectively. Results of the model indicated that the effects of onset asynchrony ($Wald\ \chi^2(3) = 113, p < .001$), word length ($Wald\ \chi^2(1) = 26.7, p < .001$) and their interaction ($Wald\ \chi^2(3) = 37.0, p < .001$) were significant (Fig. 3b). *Post hoc* Bonferroni-corrected pairwise comparison of the interaction indicated that for the 360 ms stimuli all onset asynchronies were significantly different from each other with the exception of the 380 and 480 ms conditions. For the 600 ms word duration stimuli no adjacent onset asynchronies were found to provide significant improvements, although each condition was found to be significantly different from all others. Word durations were significantly different for the same asynchrony only in the 380 and 480 ms asynchrony conditions.

4.3. Workload

The unweighted scores produced by the TLX software [21] were used, which took the mean of the sub-scale scores for each participant to one decimal place. Shapiro-Wilk and Mauchly tests indicated that the normality and sphericity assumptions were met ($p > .05$). Results from an rm-ANOVA (onset asynchrony $\times$ word duration) indicated a significant main effect ($F(3, 45) = 36.3, p < .001, h_p^2 = .708$) for onset asynchrony but no significant effect from word length ($F(1, 15) = 3.43, p = .084, h_p^2 = .186$) or the interaction ($F(3, 45) = .617, p = .608, h_p^2 = .040$). *Post-hoc* Bonferroni-corrected pairwise comparisons for onset asynchrony indicated that all of the treatments were significantly different from each other, with the exception of the 380 and 480 ms conditions.

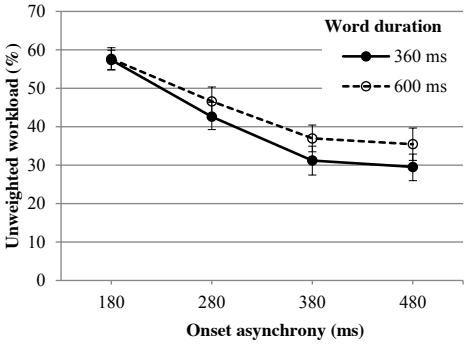


Figure 4: Marginal means for the unweighted TLX scores for the 360 (solid line with filled marker) and 600 ms (dashed line with hollow marker) word durations. (Error bars = $\pm 1 S.E.$)

5. DISCUSSION

5.1. Task duration

Due to the implicit effect of shortened words on the time taken to present information, it is of little surprise that the word duration factor exhibited a large significant effect on task durations. The *post hoc* analysis of the effect of the onset asynchrony on the total task duration indicates an optimum asynchrony of around 280 ms, despite this representing considerably different durations of overlap between the two word duration conditions.

The lack of a significant interaction between the word duration and onset asynchrony conditions suggests that the degree of asynchrony, as opposed to the proportion of overlapping stimuli, was most important in determining the time taken on each task. For the onset asynchrony conditions above 280 ms it is possible that reaction speed advantage is present due to the words being more easily identifiable. As the task durations increase, however, any improvement in time taken to detect the target is less than the increase in presentation time when the asynchrony is at its maximum of 480 ms.

5.2. Error rates

The observed interaction in the error rates appears to be due to diverging error rates for the two word duration conditions as the asynchrony increased, with the difference becoming significant at the 380 and 480 ms asynchrony conditions. At these asynchronies the shorter stimuli are no longer overlapping, whereas the longer stimuli still overlap with the following word. This fact is particularly pertinent when it is recognised that the overlaps involve the endings of two words in each triplet, which in this task can be seen as the critical section for distinguishing between the maskers and the target word. This contrast in acoustic conditions was evidently more significant than between varying degrees of overlap in the smaller asynchrony conditions. It is, however, notable that the difference between the word durations in the 280 ms asynchrony conditions is considerable and it is speculated that an increased sample size might have led to significance.

The fall in error rate data over onset asynchrony appears generally in agreement with results from other studies on onset asynchrony [24, 12]. However, error rates appear to be higher than those found by [12] in equivalent conditions. Whilst it is possible that the use of intensity panning rather binaural spatialisation may have contributed to the decreased accuracy, it seems unlikely that this difference alone would cause such a large discrepancy. It is also possible that modification of the word duration and pitch may have affected word intelligibility and inflated error values. However, it can be seen for both word durations that the accuracy approached 100% as the words became temporally distinct, implying that the processing of the words was not a major factor in itself. However, it is likely that the difference is predominantly due to the choice of experimental stimuli within this study. Whilst stimuli in this trial were distinguishable through only the final vowel-consonant transition, the words used by Ikei *et al.* [12] were more phonetically varied. This will have made the tasks considerably easier, as the increased phonetic variation will have provided the participants with more cues by which to distinguish the target word from the maskers.

It is possible that the increase in error rate observed here is responsible for an apparent disparity between the trend in error rate found by Ikei *et al.* [12] and the one found in this study. Whilst the results in [12] appeared to show an optimum onset asynchrony of 300 ms (for three voices and no attenuation), the results of this experiment appear to show further reduction in error rates up to the 480 ms condition for the longer stimuli. It is thought that this inconsistency is a by-product of the inflated error rate present in this study, and therefore the optimum asynchrony suggested in [12] is the product of a floor effect on error rates. Whilst [12] indicates that further accuracy could be achieved through the addition of 'cross-ordering' (presenting each word on the contralateral hemisphere to the preceding word) and applying an attenuation over the course of the word, neither of these methods were included in the design of the present study. Cross-ordering would not have been applicable due to the use of only three overlapping sources. It is feasible that through improving the audibility of word onsets, attenuation processing could have improved stream formation. In scenarios where the critical information is at the end of the word, however, the reduced signal-to-noise ratio is hard to justify.

Research into backwards recognition masking (BRM) indicates that vowel recognition performance plateaus when vowel onsets are separated by 200-250 ms or greater [25]. The range of asynchronies in the present study therefore suggests that BRM is unlikely to have been an influential factor for any asynchronies other than the 180 ms treatment. Due to the non-stationary nature of the speech signals used here, it could be that BRM impacted the stream formation and therefore made the location of the target more challenging to resolve.

It is notable that the constant location of critical information at the word ending may have led participants to listen only for the ending of the words and then use ordinal, spatial and, depending on the voicedness of the word ending, pitch information to derive which of the three locations the target had originated from.

5.3. Workload

The results of analysing the workload scores indicates that onset asynchrony was the only factor that influenced the participants' perception of task difficulty. In fact the workload scores appear to exhibit a divergent behaviour similar to error rate, though this difference was not large enough to be significant. Interestingly, this implies that the additional overlap associated with the longer stimuli did not significantly contribute to participants subjective workload in the two largest asynchronies, despite significantly increasing their error rate.

5.4. Overlap or onset asynchrony

It would appear that onset asynchrony describes observed trends for task duration and workload better than the amount of overlap. The error rate, however, displays a more complex interaction between the onset asynchrony and word length. The divergence between word durations with increasing asynchrony implies that both asynchrony and overlap are influential on performance. It is acknowledged that the difference between word durations was comparatively small due to the nature of the stimuli chosen and, therefore, based on this study it is not possible to come to any conclusion regarding situations in which the amount of overlap is considerably larger.

5.5. Asynchrony in menu display

Considering the effects of asynchrony on navigational speed, accuracy and subjective workload, it would appear that, of the treatments measured, the onset asynchrony of 380 ms provides the best compromise across all performance measures. Considering task durations alone, the lack of a significant difference between the 280 and 380 ms asynchrony conditions implies that an optimum exists between the two measured treatments. If one considers the additional time that would be incurred due to the higher error rates associated with the 280 ms condition, it seems likely that in practice this optimum is closer to the 380 ms condition. This conclusion is supported further through the workload scores, which show a significant reduction in workload from 280 to 380 ms onset asynchrony, suggesting that users felt that this condition made the interface significantly easier to use. The lack of overlap for this asynchrony condition for the shorter words, and its effect on error rate and navigational speed, is particularly pertinent, as it suggests that a more efficient solution would be to temporally compress the stimuli and present them with a short inter-stimuli interval.

Were a non-overlapping display to be used, a question is raised over whether the grouping of stimuli into triplets is advantageous. Grouping would seem likely to increase speed, as the number of physical interactions with the interface are reduced. Previous work comparing grouped and individual presentations of temporally distinct spoken items, however, indicates that participants are able to navigate to target locations faster when words are presented one at a time [26]. The grouped display in [26] imposed 200 ms inter-stimuli delays, whereas the present study, when using the shorter stimuli and the 380 and 480 ms asynchronies creates inter-stimuli delays of 20 and 120 ms, respectively. This suggests that faster navigation may have been possible by reducing the size of the inter-stimuli delay with minimal impact on workload or error rate. Further investigation is recommended to ascertain the effect of grouped displays with lower inter-stimuli delays on performance to inform the future design of spoken auditory displays.

The methodology presented here primes the user with a visual representation of the target word and therefore simulates only a user with a very clear idea of the item which they are looking for. In such circumstances, a search-based navigation is likely to prove more efficient. The methodology also implies a selective attention task in which the user need only listen out for one word within the list and ignore all others. This is distinct from what is required in a browsing task where a user would be expected to have to listen to a set of possible selections before making a choice. The methodology in the present study was adopted to reduce response variation due to possible target identity confusion and therefore represents the ideal scenario in terms of target knowledge.

This paper has focused on the experimental investigation into the effects of onset asynchrony in spoken menus. Cognitive and perceptual theories that surround the use of concurrent or serial speech within user interfaces have not been discussed as they are beyond the scope of this paper.

It is worth noting that the stimuli used within this study were quite short, which may have restricted the degree of stream formation that could occur, causing critical information to be missed, or its location/order/pitch to be unresolved. With longer, less informationally dense content, as in [14], users may have been able to orientate their attention more effectively towards a desired stream of speech. However, it is at present still unclear whether this would offer a significant advantage in terms of both time saved and accuracy.

6. CONCLUSION

The problem of providing users with non-visual menus capable of facilitating browsing behaviour is a considerable design challenge for auditory display. Due to the limitations of non-speech methods regarding the representation of dynamic, novel content, it would appear that speech-based methods are most appropriate. This work has sought to explore the feasibility of using asynchronous, overlapping speech for menu representation and to determine what effect this has on the speed of navigation.

An experiment in which participants attempted to find a target word within a list of words was performed so that task duration, accuracy and subjective workload could be assessed for different onset asynchronies and word durations. The results of this experiment indicate that though some speed advantage may be present, it appears to be small and not significantly better than using shorter or temporally-compressed words with some grouping. This approach appears to have the added advantage of improving accuracy and perceived workload.

7. ACKNOWLEDGEMENTS

The authors would like to thank those who volunteered to participate in the experiment and the colleagues who have been recorded speaking the stimuli as part of this project. The authors would also like to thank the reviewers for their comments and suggestions.

8. REFERENCES

[1] D. Elsweiler, S. Mandl, and B. Kirkegaard Lunn, "Understanding casual-leisure information needs: a diary study in the context of television viewing," in *Proc. 3rd Symp. Inform. Interaction in Context (IIiX)*, New Brunswick, NJ, Aug. 2010, pp. 25–34.

[2] M. M. Blattner, D. Sumikawa, and R. Greenberg, "Earcons and icons: their structure and common design principles," *Human-Comp. Interaction*, vol. 4, no. 1, pp. 11–44, Mar. 1989.

[3] W. Gaver, "Auditory icons: using sound in computer interfaces," *Human-Comp. Interaction*, vol. 2, no. 2, pp. 167–177, Jun. 1986.

[4] B. N. Walker, J. Lindsay, A. Nance, Y. Nakano, D. K. Palladino, T. Dingler, and M. Jeon, "Spearcons (speech-based earcons) improve navigation performance in advanced auditory menus," *Human Factors: J. Human Factors and Ergonom. Soc.*, vol. 55, no. 1, pp. 157 –182, Feb. 2013.

[5] Y. Borodin, J. P. Bigham, G. Dausch, and I. V. Ramakrishnan, "More than meets the eye: a survey of screen-reader browsing strategies," in *Proc. 2010 Int. Cross-Disciplinary Conf. Web Accessibility (W4A)*, Raleigh, NC, Apr. 2010, Article 13.

[6] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: MIT Press, 1994.

[7] J. C. Webster and P. O. Thompson, "Responding to both of two overlapping messages," *J. Acoustical Soc. Am.*, vol. 26, no. 3, pp. 396–402, May 1954.

[8] M. A. Ericson, D. S. Brungart, and B. D. Simpson, "Factors That Influence Intelligibility in Multitalker Speech Displays," *Int. J. Aviation Psychology*, vol. 14, no. 3, pp. 313 – 334, 2004.

[9] W. T. Nelson, R. S. Bolia, M. A. Ericson, and R. L. McKinley, "Spatial audio displays for speech communications: a comparison of free field and virtual acoustic environments," *Proc. Human Factors and Ergonomics Society Annu. Meeting*, vol. 43, no. 22, pp. 1202–1205, Sep. 1999.

[10] C. Frauenberger and T. Stockman, "Patterns in auditory menu design," in *Proc. 12th Int. Conf. Auditory Display (ICAD)*, T. Stockman, L. V. Nickerson, C. Frauenberger, A. D. N. Edwards, and D. Brock, Eds., London, UK, Jun. 2006, pp. 141–147.

[11] C. Frauenberger, Personal correspondence, 2013.

[12] Y. Ikei, H. Yamazaki, K. Hirota, and M. Hirose, "vCocktail: multiplexed-voice menu presentation method for wearable computers," in *Proc. IEEE Virtual Reality Conf.*, Alexandria, VA, Mar. 2006, pp. 183–190.

[13] P. Parente, "Clique: perceptually based, task oriented auditory display for GUI applications," PhD Thesis, University of North Carolina, 2008.

[14] J. Guerreiro and D. Gonçalves, "Text-to-speeches: evaluating the perception of concurrent speech by blind people," in *Proc. ASSETS'14*, Rochester, NY, Oct. 2014, pp. 169–176.

[15] V. Shafiro and B. Gygi, "Perceiving the speech of multiple concurrent talkers in a combined divided and selective attention task." *J. Acoustical Soc. Am.*, vol. 122, no. 6, pp. EL229–EL235, Dec. 2007.

[16] C. J. Darwin, D. S. Brungart, and B. D. Simpson, "Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers," *J. Acoustical Soc. Am.*, vol. 114, no. 5, pp. 2913 – 2922, Nov. 2003.

[17] "Pd-extended." [Online]. Available: http://puredata.info/downloads/pd-extended

[18] A. S. House, C. E. Williams, M. H. L. Hecker, and K. D. Kryter, "Articulation-testing methods: consonantal differentiation with a closed-response set," *J. Acoustical Soc. Am.*, vol. 37, no. 1, pp. 158 – 166, Jul. 1965.

[19] P. Boersma and D. Weenink, "Praat: doing Phonetics by Computer." [Online]. Available: www.praat.org

[20] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1-2, pp. 103–138, Aug. 1990.

[21] A. Cao, K. K. Chintamani, A. K. Pandya, and R. D. Ellis, "NASA TLX: software for assessing subjective mental workload." *Behavior Research Methods*, vol. 41, no. 1, pp. 113–117, Feb. 2009.

[22] IBM, "IBM Knowledge Center," Jan. [Online]. Available: http://www-01.ibm.com/support/knowledgecenter/SSLVMB_21.0.0/com.ibm.spss.statistics_21.kc.doc/pv_welcome.html

[23] S. L. Zeger and K.-Y. Liang, "Longitudinal data analysis for discrete and continuous outcomes," *Biometrics*, vol. 42, no. 1, pp. 121 – 130, Mar. 1986.

[24] J. H. Lee and L. E. Humes, "Effect of fundamental-frequency and sentence-onset differences on speech-identification performance of young and older adults in a competing-talker background." *J. Acoustical Soc. Am.*, vol. 132, no. 3, pp. 1700–1717, Sep. 2012.

[25] D. W. Massaro, "Perceptual units in speech recognition," *J. Experimental Psychology*, vol. 102, no. 2, pp. 199–208, Feb. 1974.

[26] J. Sodnik, G. Jakus, and S. Tomažič, "Multiple spatial sounds in hierarchical menu navigation for visually impaired computer users," *Int. J. Human-Comp. Stud.*, vol. 69, no. 1-2, pp. 100–112, Jan. 2011.