MIT Media Laboratory Technical Report. September, 1994.

The Cocktail Party Effect in Auditory Interfaces: A Study of Simultaneous Presentation

Lisa J. Stifelman

MIT Media Laboratory
lisa@media.mit.edu

Abstract

Researchers have proposed building auditory interfaces that make use of the "cocktail party effect"—the human's ability to selectively attend to a single talker or stream of audio among a cacophony of others—to reduce the amount of time required to listen [Arons 1992, Cohen 1992, Mullins 1993]. The premise is that by presenting multiple streams of audio simultaneously, the user will be able to focus on one, yet overhear interesting information in another and easily switch attention. In this study, subjects listen to multiple channels of audio and perform two tasks simultaneously—listening comprehension and target monitoring. While listening to one passage of speech, the subject must identify target words in one or two other passages played simultaneously. The results showed that the subject's performance, both listening comprehension and target monitoring, decreased significantly as the number of simultaneous audio channels increased.

Introduction

The use of speech and sound as a means of interacting with computers is rapidly increasing. This is due to a large extent to two major trends: the decrease in the size of computers and consumer devices, and the use of the telephone for remotely accessing information. The problem with audio interfaces (e.g., voice mail) is that they tend to be slow and serial, and do not allow random access. Researchers have begun to address these problems by providing the user with interactive control over audio playback and the ability to "skim" audio recordings [Arons 1993, Stifelman 1993]. Several researchers have proposed building audio interfaces that make use of the "cocktail party effect"—the human's ability to selectively attend to a single talker or stream of audio among a cacophony of others—to reduce the amount of time required to listen [Arons 1992, Cohen 1992, Mullins 1993]. Arons proposes a system that "simultaneously presents multiple streams of speech information such that a user can focus on one stream, yet easily shift attention to the others" (p. 35, [Arons 1992]). Mullins has built such a system for presenting multiple channels of audio news [Mullins 1993]. The premise is that by presenting multiple streams of audio simultaneously, the user will be able to focus on one, yet overhear interesting information in another and easily switch attention. The focus of this study is to evaluate this basic premise.

In this study, subjects listen to multiple channels of audio and perform two tasks simultaneously—listening comprehension and target monitoring. While listening to one passage of speech, the subject must identify target words in one or two other passages played simultaneously. The passages of speech are played over headphones to one of three locations—left, center, or right—using stereo panning. It is hypothesized that the subject's ability to simultaneously comprehend a primary passage and overhear targets in secondary ones will decrease as the number of channels increase. In addition to studying the user's performance on these tasks, another goal is to observe the user's response (i.e., overall reaction, likes, dislikes) to this type of simultaneous listening experience.

Background Research

Recall of "Rejected" Material

One of the earliest studies of selective attention and the cocktail party effect was performed by Cherry [Cherry 1953, Cherry 1954]. Cherry analyzed the listener's ability to focus on one of two speech messages when mixed and played to both ears (i.e., diotic), and when unmixed and played to different ears (i.e., dichotic). The subjects shadowed (repeated the speech out loud word by word) a primary message while rejecting a secondary one. Cherry found that when two audio messages were played dichotically, subjects could not report much about the message in the rejected ear. Cherry then attempted to determine what attributes of the rejected message are recognized. The results showed that subjects could not identify words or phrases from the rejected message, notice if the language changed or that the speech was reversed[1]. Subjects did recognize whenever a 400 Hz tone was played in the rejected ear and if the voice changed from male to female. Similarly, Treisman found that difference in voice (i.e., male versus female) allows more efficient rejection of the irrelevant signal when messages are presented diotically [Treisman 1964].

Following Cherry's results, Moray further studied recall of the rejected message in a dichotic listening task [Moray 1959]. While the subject shadowed a passage in one ear, lists of words were repeatedly played to the other. Subjects did not recall material in the rejected ear, even when explicitly instructed beforehand to try to remember as much as possible. Moray states that the subject's name or other material of "importance" to the subject is the only stimuli that can break through the selective attention "barrier" [Moray 1959, Moray 1970].

The difficulty with Cherry and Moray's findings about the lack of information recalled from the rejected ear is as Norman states, "how are we able to switch our attention to new voices or events when the occasion arises if we are unaware of the content of those other events" (p. 13, [Norman 1976]). Norman challenges Cherry's findings in two critical ways: the high cognitive load of the shadowing task, and the time elapsed between message playback and the request for the subject to recall the rejected material. Cherry himself reports that the subject may have "very little idea of what the message that he has repeated is all about" (p. 978 [Cherry 1953]). So in some sense, the message being shadowed is "rejected" as well [Norman 1976]. In addition, subjects spoke in a monotonous tone of voice with no emotional quality or stress when shadowing, another indication of the task's difficulty.

Both Cherry and Moray waited until after task completion to ask subjects to recall the rejected material. Studies by Norman [Norman 1969] and Glucksberg and Cohen [Glucksberg 1970] address the effect of elapsed time on the ability to recall unattended material. Norman interrupted subjects while shadowing words presented to one ear to ask them to recall digits just played to the other. He demonstrated that there is some recall if subjects are interrupted less than 20 seconds after the presentation of the material. He concluded that the unattended material does get into short-term memory but never gets transferred to long-term memory. Glucksberg and Cohen further determined that memory for the nonattended material continually decreases as the delay increases from 0 to 5 seconds and that no recall is apparent between 5 and 20 seconds. Norman [Norman 1976] refers to this as the "what did you say" phenomenon—the way a listener might ask this question when not paying full attention to the talker and then suddenly recall what was said before the talker has responded.

Despite these studies, determining the high cognitive load of shadowing and its interference with the transfer of nonattended material from short to long-term memory, researchers

[1] In the case of reversed speech played to the rejected ear, some subjects noticed there was "something queer about it" [Cherry 1953].

continue to use shadowing in cocktail-party-related experiments.[2] In the study presented in this paper, the goal is to use a more realistic task as a basis, and avoid the use of shadowing. Given a speech communication system that presents multiple channels of auditory information, it will be important for the user to be able to comprehend a foreground message while overhearing interesting information in background ones. For this reason, listening comprehension, rather than shadowing was used to evaluate the subject's performance for the foreground information.

Target Detection Experiments[3]

Cherry's experiments focused on the human's ability to attend to one speech signal in the presence of others in the background (i.e., selective attention). However, another aspect of the cocktail party effect is the ability to switch our focus when something in the background draws our attention. A number of experiments have been performed in which subjects do not simply "reject" a secondary speech signal but must detect target words contained in it.

Treisman and Geffen performed an experiment in which subjects listen to two messages dichotically and must shadow the primary one while trying to detect target words in either message [Treisman 1967]. The results showed that subjects detected a higher percentage of target words in the primary message (87%) than in the secondary one (9%). Target words in either passage were detected more readily if introduced in context rather than randomly. Subjects performed better when instructed to listen for a single word rather than a class of words (e.g., any digit). The secondary task of "tapping" upon hearing a target word was disruptive to the primary task of shadowing, but more so when targets were contained in the secondary message than in the primary one. Even though the subjects were instructed not to shift their attention away from the primary passage at any time, the results indicate otherwise. Over 30% of shadowing errors occurred when subjects were tapping to targets contained in the secondary message.

Lawson performed a similar experiment to Treisman, however, using tones instead of words as targets [Lawson 1966]. In contrast to Treisman's findings, there was no difference in detection of targets in the primary and secondary passages. Zelniker et. al. further studied the detection of tones presented to the unattended ear and found that tones were reliably detected even when played at near-threshold intensities [Zelniker 1974]. This is consistent with Cherry's findings that a 400 Hz tone played in the unattend ear is readily recognized.

Treisman and Lawson's experiments suffer from the same problems discussed previously due to the use of shadowing. A study by Dennis examines target monitoring performance while subjects are shadowing compared to while listening silently [Dennis 1977]. Subjects were also given a recognition test to determine if they "processed the meaning" of the passages they were shadowing. While Treisman used prose passages for both signals, Dennis used a prose passage for the primary source and word lists for the secondary or target source. One problem is that the prose is always presented to the subject's left ear, and the word lists to the right causing a possible bias in the results. In addition, the recognition tests are difficult since Dennis attempted to make all the alternative answers as plausible as possible. As expected, shadowing interfered less with visual than with auditory monitoring. Auditory monitoring performance was better when subjects listened silently then when shadowing. However, when subjects listened silently, although monitoring improved, recognition test scores decreased.

---

[2]Underwood has even studied the effect of practice on the task of shadowing using Moray as an expert shadower subject [Underwood 1974].

[3]This section discusses target detection experiments relevant to the study described in this paper. For more information in this area, see also [Bookbinder 1979, Stephens 1988].

Such a reliable decrement in recognition performance did not occur when the primary message was shadowed.

Like the Dennis study, the experiment described in this paper uses "silent listening"[4] rather than shadowing of the primary passage. However, the design of this study is based on the type of interaction that would be necessary given an interface for browsing multiple simultaneous channels of speech. Therefore, rather than word lists, this experiment uses prose passages for each channel, tests up to three simultaneous sources rather than just two, and uses comprehension rather than recognition tests. Another important distinction is the method for control over the subject's division of attention. Both the Treisman and Dennis used only instructions to control the subject's attention. Dennis instructed subjects to "allow any fall in performance to appear in monitoring rather than shadowing" (p. 440, [Dennis 1977]). However, for conditions in which subjects listened silently, "despite the fact that the passage was given explicit priority...subjects seem to have attended to it less effectively" (p. 447, [Dennis 1977]) then when shadowing. The problem is that instructions alone cannot control the subject's attention or ensure consistency across subjects. In this experiment, subjects are presented with an explicit pay-off matrix, indicating how many points each task was worth. In addition, subjects are given practice trials and immediate feedback of their score before the actual test conditions.

More Than Two Competing Speech Signals

Most research on selective attention has focused on dichotic listening, employing only two competing speech signals. However, there have been some studies employing more than two competing messages. One of the first was by Treisman [Treisman 1964] who studied the effect of irrelevant messages on selective attention. This study compared the subject's shadowing performance when 0-2 irrelevant speech signals were introduced. Treisman found that one irrelevant channel did not significantly decrease shadowing efficiency, but two did. This interference occurred when the irrelevant messages were distinguished by voice (male versus female) or by spatial location (left, center, or right). However, if both the voice and the spatial location of the irrelevant messages are the same, then performance improves. This result is due to what Bregman refers to as *auditory stream segregation*—when the two irrelevant messages have similar acoustic features (in this case spatial location and pitch) they form a single auditory stream, making them equivalent to a single interference signal [Bregman 1990].

Webster and Solomon [Webster 1955] found that increasing the number of competing signals (from one to two to three) decreased the subject's efficiency in responding to a primary message. As in Treisman's study, this occurred when the primary and competing signals differed in location or frequency-range.

A very recent study by Yost tested divided attention with up to three sound sources [Yost 1994]. The intention was to determine the contribution of binaural processing to cocktail party listening tasks. This study focused on divided rather than selective attention—subjects listened to simultaneous speech signals and had to respond to all of them. Multiple channels of words, letters, and numbers were played and subjects had to type in all the words they heard. The results showed that subjects in binaural listening conditions performed better than those listening monaurally. In addition, subject's performance decreased as the number of simultaneous sound sources increased from one to two to three.

[4]See also [Inoue 1981] for a dichotic listening study in which shadowing was compared with silent listening.

The work presented in this paper, like the Treisman study, looks at the effect background signals (zero, one, or two) on selective attention. While Treisman refers to these signals as "irrelevant", they are referred to here as "target" channels. This is an important distinction because all the information contained in these additional speech signals is *not* irrelevant. Given a multi-channel speech interface, some of the background material may be relevant to the subject's interests, and other parts of it may not; the point is to provide users with an interface that enables them to switch their attention when "important" or interesting information is presented in a secondary or background channel.

Enhancing Selective Attention

Many researchers have studied factors that can enhance selective attention to one speech signal among several others in the background. Some factors are: spatial separation, differential filtering, degree of synchrony, intensity differences, and pitch differences between competing messages.

Spieth et. al. report that horizontal separation and differential filtering (high-pass one signal, low-pass the other) enhanced subject performance in responding to one of two simultaneous messages [Spieth 1954]. Differences in the time onset of the messages also reduces error. However, the combination of the selective listening aids did not further enhance performance over the use of a single aid alone.

Webster and Thompson also report that responding to one or more of several simultaneous speech messages is enhanced by spatial separation (in this case, output from multiple loudspeakers as opposed to just one) [Webster 1954]. However, this advantage only occurred when subjects "pulled down" the primary message into an earphone, and was not as evident when messages were overlapping.

One problem with increasing the spatial separation of sound sources is that this increases the amount of time it takes to switch attention between signals. Rhodes reports that the time to localize a sound increases linearly for distances up to 90 degrees from the point of attention [Rhodes 1987]. This is cited as evidence of a spatial constraint for auditory attention that "shifts through an analogical of topographical representation" (p. 13, [Rhodes 1987]). This increase in response time does not continue past 90 degrees, possibly because beyond this distance change over a short period of time (< 1 sec), the second signal will be perceived as coming from a different source as opposed to a movement of the current one.

Divenyi and Oliver studied the amount of spatial separation needed to discriminate two simultaneous sound sources [Divenyi 1989]. Previous studies have found a minimum audible angle of about 2-5 degrees [Yost 1977, Blauert 1983] for discriminating sound sources in the horizontal plane. However, in Divenyi's study the sounds are simultaneous (as in a cocktail party) as opposed to sequential as in previous studies. Divenyi recommends that no more than six simultaneous sound sources be used since a minimum of 60 degrees of separation was required for subjects to localize the simultaneous sounds. However, it may still be possible to "discriminate" rather than "locate" more than six simultaneous sound sources, so this recommendation depends upon the given task.

Egan et. al. [Egan 1954] like Spieth, report an advantage for high-pass filtering of either the signal or irrelevant message, but in this case, both voices were the same (Spieth used two of four different male voices). Corbett et. al. confirms Egan's findings that differential filtering only enhances selective listening when the two speakers are the same. Egan also plotted the increase in recognition of the primary message as the signal to noise ratio of two synchronous messages increases. For synchronous messages presented dichotically, articulation scores rise much more rapidly than for monaural presentation (rising to over 90% at a 0 dB S/N ratio for

dichotic presentation and to only just over 50% for monaural). Intensity level and filtering interacted—the threshold of intensity was lower when one of the messages was high-pass filtered.

Irwin and Noble [Irwin 1973] challenge the results of Egan's intensity analysis, stating that the interfering signal's intensity should be varied rather than that of the primary signal. Modifying the intensity of the primary signal may confound the results by reducing its intelligibility. Similar to Egan, the results indicate an increase in intelligibility of the primary message as the S/N ratio increases, however, the amount of increase was much less (34% as opposed to 60%).

Brokx and Nooteboom studied the effect of pitch differences on selective attention and found improved intelligibility of the target speech signal in the presence of an interfering one by introducing a difference in pitch between the competing messages [Brokx 1982]. The difference in pitch between messages was approximately 110 Hz, with 110 Hz for the low pitch message and 220 Hz for the high pitch one.[5] The authors hypothesize that this effect is due to a decrease in the probability of "perceptual fusion" of the competing signals and switching attention to the wrong voice (p. 34 [Brokx 1982]). No systematic difference in pitch was examined, and so the minimum pitch increase necessary for improvement was not determined.

Lastly, bimodal presentation of the two competing signals will also enhance selective attention. For example, Dennis [Dennis 1977] found that shadowing of a primary message interfered less with a visual target monitoring task than with an auditory one.

Method

Subjects

Three pilot subjects and 12 test subjects participated in the experiment. All subjects were students from the Media Laboratory at the Massachusetts Institute of Technology.

Apparatus

The experiment was run on an Apple Macintosh Quadra 840AV. Subjects listened to passages of speech over Sony MDR-V600 stereo headphones.

Auditory Stimuli

Comprehension passages were selected from 6 TOEFL preparatory exams [Arco 1991]. These tests were explicitly designed to be given verbally (i.e., listening comprehension). A total of 16 passages were used in the experiment, 10 for the practice trials (see Figure 1). Of the 16 passages, 7 were used for the listening comprehension task and 9 for the target monitoring task.

These passages were modified in several ways. First, the length of all passages was normalized to 180 words. Next, the 9 passages selected for the target monitoring task were edited, and targets were inserted in context. Two versions of each target passage were created, one for the two channel conditions (composed of 1 listening and 1 target passage) and one for the three channel conditions (composed of 1 listening comprehension and 2 target passages). A total of 8 targets was always used, so for the two channel conditions, 8 targets were inserted into a

[5]These F0 values correspond to the approximate averages for male (approx. 132 Hz) and female (approx. 223 Hz) speakers (p. 48, [O'Shaughnessy 1990]).

single target passage, and for the three channel conditions, 4 targets were inserted into each of two target passages (see Figure 2). Targets were separated from one another by a minimum of 6 words and were never within the first or last 10 words of the passage (based on [Treisman 1967]). Lastly, for the comprehension passages, the number of questions was normalized to 5 each.[6]

Test Type	**# Comp Passages**	**# Target Passages**
Practice Trials	4	6
Test Trials	3	3

Figure 1: Number of listening comprehension and target monitoring passages used in the practice and test trials of the experiment.

# Channels	Type of Passage Comprehension	Target 1	Target 2
1 channel	—	—	—
2 channels	—	8	—
3 channels	—	4	4

Figure 2: Target distribution for 1, 2, and 3 channel conditions.

A single target word— the name "John" was used in each passage. This target word was selected for several reasons. In previous experiments, the subject's own name[7] was often spotted in a "rejected" (i.e., secondary) auditory channel (the classic cocktail party example). In addition, the target words needed to fit into the context of the passage, and this was easier to accomplish with a name. For example, in a passage about Muhammad Ali, the opponent fighter's name "Sonny Liston" was changed to "John" in the target version (see Figure 3). Lastly, this kind of target word is likely given a task like browsing through news stories [Mullins 1993]. For example, perhaps I am listening to a story about Nancy Kerrigan in one channel—will the mention of the name "Tonya" in another successfully draw my attention?

Each of the passages was then read aloud by 3 speakers, 2 female and 1 male and recorded on the Macintosh Quadra 840AV at 8 bits linear 22 kHz. Seven comprehension passages read by 3 speakers gives a total of 21 sound files. Two versions (two channel and three channel) of 9 target passages read by 3 speakers gives a total of 54 sound files. Therefore, a total of 75 sound files were created for use in this experiment.

Speakers reread each passage several times until they were relatively free of repairs and hesitations and were approximately one minute in length (resulting in a speaking rate of approximately 180 wpm). Each recording was then edited, removing any remaining hesitations. The comprehension passages were edited to exactly 60 seconds. The target passages were edited to 59 seconds to allow the comprehension passages to start playing one second before the target ones and still end at the same time. Note that before editing, each file was at least one minute long so that normalization was accomplished by shortening pauses but never adding any. Pauses at the beginning and end of the recording were always removed.

[6]Since TOEFL does not normalize the number of questions associated with each passage, sometimes a question had to be added or removed to achieve 5 questions per test.
[7]Note, however, that none of the subjects in this experiment were named "John".

Born Cassius Clay in Louisville, Kentucky, in 1942, Muhammad Ali retired from boxing in 1980. As an amateur, he won 100 out of 108 fights. Later, as a professional boxer, Ali was trained by Angelo Dundee, who pushed him to face champion boxer Sonny Liston in 1964; in this exciting match, Liston was unable to continue the fight after the sixth round.

After winning the championship fight against Liston, Ali announced that he had joined the Black Muslim religion. At the time, the Vietnam War was in progress. When Ali claimed to be a conscientious objector to the war, based on religious grounds, he was denied exemption from military service. Refusing to be inducted into the U.S. Army, he was stripped of his right to box and he was givn a five-year prison sentence, a conviction that was reversed three years later. Once again able to enter the boxing ring, Muhammad Ali became the first man to win the world's heavy-weight boxing title three times. Muhammad Ali remains a legendary figure in the sport of boxing.

Born Cassius Clay in Louisville, Kentucky, in 1942, Muhammad Ali retired from boxing in 1980. As an amateur, he won 100 out of 108 fights. Later, as a professional boxer, Ali was trained by Angelo Dundee, who pushed him to face champion boxer **John** in 1964; in this exciting match, **John** was unable to continue the fight after the sixth round.

After winning the championship fight against **John**, Ali announced that he had joined the Black Muslim religion. At the time, the Vietnam War was in progress. When Ali claimed to be a conscientious objector to the war, based on religious grounds, he was denied exemption from military service. Refusing to be inducted into the U.S. Army, he was stripped of his right to box and he was givn a five-year prison sentence, a conviction that was reversed three years later. Once again able to enter the boxing ring, Muhammad Ali became the first man since **John** to win the world's heavy-weight boxing title three times. Muhammad Ali remains a legendary figure in the sport of boxing.

Figure 3: A passage used in the experiment—the original is on the left, and the one with the target word "John" (inserted 4 times for the 3 channel condition) is on the right.

The gain of each sound file was normalized.[8] The sounds were normalized against the loudest recording such that sounds were never scaled down (i.e., a scale factor > 1.0). This was accomplished as follows:

1. For each sound file, create a histogram of the sample values.
2. Determine the peak value above a threshold—in this case the top 0.01% of the histogram was selected as the threshold.
3. Calculate the largest "peak" value across all sound files (call this LARGEST_PEAK).
4. Normalize all other sound files using the following scale factor:
   scaleFactor = LARGEST_PEAK / peakForSoundFile

A more robust result is obtained by using a histogram and calculating the peak value above a threshold. If an absolute peak was used, a single large sample value could cause the threshold to increase.

Experimental Design

A single factor within subjects design was used in this experiment. The subject's ability to listen to and comprehend one passage while monitoring for targets in others was tested. The number of target passages (0, 1, or 2) was varied. These three conditions will be referred to by the total number of channels presented to the listener (comprehension plus target)—one channel, two channel, and three channel conditions. The one channel condition was used as a control to get a baseline listening comprehension performance measure. The subject's overall performance on the listening comprehension and target monitoring tasks was used as the dependent measure. Performance scores were calculated according to the pay-off matrices shown in Figures 4–5.

[8] A C program written for normalizing the sound levels was run as a batch process on all of the sound files used in the experiment.

Ss Response	Stimulus: Target	Stimulus: No Target
Target	Hits = 12 pts	—
No Target	Misses = 0 pts	—

Figure 4: Listening comprehension scoring.

Ss Response	Stimulus: Target	Stimulus: No Target
Target	Hits = 5 pts	False Alarms = -7 pts
No Target	Misses = 0 pts	Correct Rejections = 0 pts

Figure 5: Target monitoring scoring.

Pay-off matrices were used to minimize intersubject variability and to influence the subject's division of attention. More points were awarded for listening comprehension questions, since this was considered the primary task and the target monitoring secondary. A 60/40 split was used—60 total possible points for listening comprehension (5 questions, 12 points each) and 40 points for target monitoring (8 targets, 5 points each). Points were subtracted for false alarms to discourage subjects from randomly pressing the target detection button.

Due to the difficult of the task (evaluated by running three pilot subjects), false alarms were judged leniently. After a target is played, if the subject does not press the detection button before the next target is presented, than this is considered a miss and no points are subtracted (i.e., the "target window" extends from the start of one target until the start of the next). However, if the target button is pressed more than once during a given "target window" than these extra presses are considered false alarms and points are subtracted accordingly. One possible problem with this scoring procedure is that a late response for one target could be considered as a response to the next one. However, since targets are separated from each other by a minimum of 6 words (approx. 3 seconds) this is unlikely.

Each subject participated in a total of 7 trials—4 practice and 3 test. Subjects received two practice trials of the two channel condition (one comprehension passage and one target passage played simultaneously) and two practice trials of the three channel condition (one comprehension passage and two target passages played simultaneously). The two channel practice tests were given first, followed by the three channel ones so that the subject could get accustomed to multi-channel presentation. The comprehension and target passages for each practice trial were selected at random and then randomly assigned one of three voices and one of three output locations (left, center, right).

The sound separation—left, center, right—was achieved using stereo panning and played over headphones. For two channel output, played to the left and right ears, buffers from each of two sound files are interleaved and handed off to the Macintosh Sound Manager. For three channel output, a center channel is created by playing the sound at equal gain to both ears. Buffers from the first and second files are each mixed with buffers from a third file and then interleaved. When three sounds are played simultaneously, the ones played to the left and right ears are scaled up by 3 dB to attain approximately equal sound levels between all channels. This is necessary because a sound played to both ears will sound louder than a sound played to one ear. Mixing, scaling, and interleaving are performed in real-time, to allow random assignment of voice and output location on-the-fly for each trial.

Note that because the sounds were not fully spatialized and were played over headphones, they were heard inside the user's head and not externalized[9].

Procedure

Subjects were instructed that they would hear two or three passages playing at once. They were told that the passage that began first was the comprehension one, and that they would be tested on it afterwards. They were also instructed that the other passage(s) would contain the target word "John" in several places, and to hit the [Enter] key as soon as they heard the target. The scoring procedure (i.e., pay-off matrix) was explained and a written break-down was given to the subjects (as shown in Figure 6). They were warned that they would lose points for random button presses (i.e., false alarms).

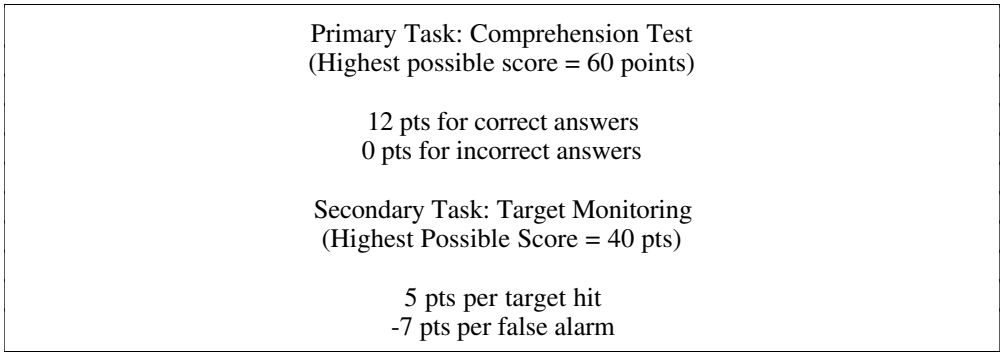

Figure 6: Scoring break-down given to subjects before the experiment.

Following each practice trial, the subject was given a written sheet of 5 multiple choice test questions to answer. The experimenter ran a program to calculate the subject's target detection score for the trial and added this result to the listening comprehension score. The goal was to give subjects immediate feedback about their performance so that they could adjust their strategy in accordance with the score. This is the reason that at least two practice trials were needed for each condition (i.e., so that the subject could adjust their strategy after the first practice if necessary and see the result in the second one).

Subjects were instructed when the actual test trials were beginning and told that either one, two, or three passages would be heard simultaneously. During the pilot tests it was noted that subjects often did not pay attention during the single channel control condition. Therefore, subjects in the actual experiment were warned that performance on the one channel test was also important and to pay close attention.

In between each test, the subject was asked if he/she recalled the content of the target passage(s) and if so to give a very brief (i.e., a few words at most) summary of what it was about. During a pilot test, after receiving this summary question on the first trial, the subject explicitly tried to listen for the "gist" of each passage in order to be able to answer this question. In the actual experiment, it was stressed that the subject did not receive any points for the summary question and that they should not modify their strategy to answer it. Between trials and at the end of the experiment, subjects were also asked to contribute any comments they wanted to make about the task (e.g., how difficult, enjoyable, etc.).

[9]Although these are not the only reasons that sounds fail to externalize [Durlach 1992].

Results

The mean performance scores for the three conditions—one, two, and three channels—were compared (Figures 7-8). The target monitoring and overall performance scores were compared across the two and three channel conditions only, since the one channel condition does not have an associated target monitoring score. Listening comprehension scores were compared across all three conditions.

Score	# Channels 1 Channel	2 Channels	3 Channels
Listening Comprehension			
Mean (max = 60 pts)	46.0	37.0	31.0
SD	10.0	14.9	10.8
Mean (%)	76.7	61.7	51.7
Target Monitoring			
Mean (max = 40 pts)	—	25.3	15.9
SD	—	10.9	7.2
Mean (%)	—	63.1	39.8
Overall Performance			
Mean (max = 100 pts)	—	62.3	48.6
SD	—	19.0	11.9

Figure 7: Mean subject performance scores and standard deviation for each experimental condition. The first row of means is based on the raw scores—out of 60 points for listening comprehension and 40 points for target monitoring. The second means (mean %) are calculated by averaging the percent scores—x/60 * 100 for listening comprehension and x/40 * 100 for target monitoring.

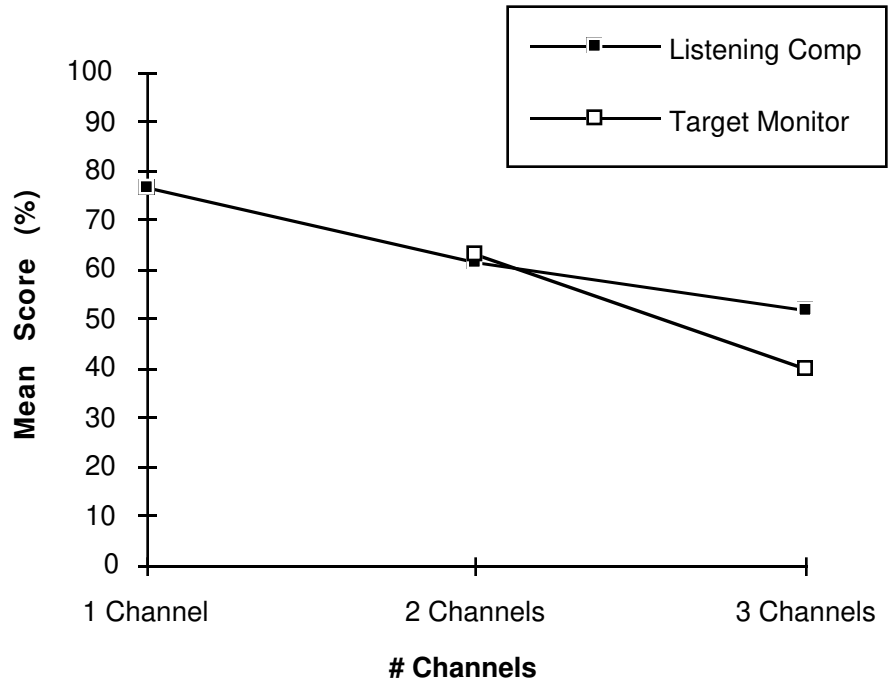

Figure 8: Performance decrement in listening comprehension and target monitoring scores as the number of channels increases. The scores are out of 100%—60/60 pts equals 100% for listening comprehension and 40/40 pts is 100% for target monitoring.

A one-way analysis of variance was run for the number of channels factor for each of the dependent measures—overall performance, listening comprehension, and target monitoring. The overall performance score significantly decreased as the number of channels increased from two to three ($F(1, 22) = 4.47$, $p < .05$).

Listening comprehension scores significantly decreased as the number of channels increased ($F(2, 33) = 4.68$, $p < .05$). A comparison of each pair—one vs. two, two vs. three, and one vs. three channels—indicates a significant difference for only the one vs. three channels comparison ($t(11) = 2.20$, $p < .05$). The difference between one and two channels was not significant.

For the target monitoring task, there was a significant decrease in the score when the number of channels was increased from two to three ($F(1, 22) = 6.15$, $p < .05$).

Interviews

Recall of Background Passages

Following each trial, subjects were asked if they could recall in a few words what each passage was about. Many of the subjects were able to briefly describe the passages when only two were presented simultaneously, however, only a few were able to do so when there were three simultaneous passages. One subject commented, "I've given up trying to follow the other two, if there's just one other, I can follow a little." The subject's recall of passages for the two channel conditions was not consistent however, and depended on the content in some cases. For example, most subjects were able to recall the passage about Muhammad Ali, since many commented that this was an interesting story. In fact, one subject attended to this passage primarily even though it was a background passage. The subject commented that he found it interesting and was drawn to listen to it. When asked what the passages were about, in several instances subjects commented, "too late it's gone" or "it's like a dream...I used to know." This corresponds to Norman's [Norman 1969] finding that unattended material can be reported within a brief period following presentation (< 20 seconds) but not after since it never gets transferred from short to long-term memory.

These results concerning the unattended or secondary speech signal differ from those cited by researchers like Cherry and Moray [Cherry 1953, Moray 1959]. This is most likely due to the use of comprehension tests as opposed to shadowing. Listening silently left over more attentional resources than shadowing for subjects to monitor the secondary channel(s). Similar to Moray's findings, however, subjects were particularly drawn to information contained in the secondary passage that they considered interesting or "important".

Task Difficulty: Two Versus Three Channels

Subjects found the three channel conditions much harder than two channels. Many subjects commented "wow that's hard...what a hard test...three is very hard" after experiencing the three channel tests. The main reason cited for three channels being more difficult than two was greater difficulty in focusing on the primary passage or "tuning out" the two secondary ones. With two simultaneous passages, subjects could time share, but with three they tended to "block out the other two completely."

Another problem with three passages was difficulty in switching attention between them—"you lose your place...it's harder to get back to the primary [passage]." With three passages, comprehension was difficult—subjects said "I had to give up trying to figure out the content of the other two" and "with two it's doable but with three it's not practical unless you're trying to scan all three and not comprehend any."

Several subjects described a practice effect—"after listening to three at once, two got easier...but three are still blowing by."

Some subjects complained that one voice was clearer than another—most preferred the female voices over the male, which was said to be less clear, but a few commented that the male voice was clearer. In addition, ear advantages, sometimes left and sometimes right, were often cited.

In general, whether two or three simultaneous passages were used, the tasks seem to require a lot of attention. Many subjects closed their eyes, some covering them in looks of deep concentration.

Conscious Versus Unconscious Target Detection

A more difficult question to test is whether users serendipitously overhear information of interest in a secondary channel while focusing on a primary one, or if they simply monitor each channel of information serially. If users only monitor the channels one at a time, then the benefit of simultaneous presentation is limited. However, if users are able to spontaneously overhear information of interest, then the use of simultaneous presentation may provide a new mechanism for efficiently listening to and browsing of audio.

In a study by Treisman and Geffen (previously described) subjects were instructed "not to shift their attention to the secondary message, since we were interested in seeing whether they heard it despite the fact that they were attending to something different." (p. 5, [Treisman 1967]) In addition, subjects "were asked after each passage in which they tapped to a word in the secondary message, whether they felt they had shifted their attention in order to hear the word or whether it had just come through" (p. 5, [Treisman 1967]).

As in the Treisman study, subjects were asked whether they consciously listened for the target words or if the words "just came through" without any conscious effort. Most subjects reported hearing targets in both manners, some through conscious monitoring of the secondary channel(s) and others unconsciously. When listening consciously, subjects would try to "swap out" of the primary channel when they thought they would not lose any important information (e.g., at a pause) and at times when they "predicted" a target was about to occur (e.g., due to context or a target hadn't been heard in a while). Other times, subjects reported "just hearing" the target word "out of the blue somehow" or "serendipitously." Only a couple of subjects reported detecting targets only consciously or only unconsciously, most experienced both ways of finding targets.

An interesting comment by one subject was that it was easier to monitor passages when the subject was repeated several times. In particular, the passages about Muhammad Ali and the Lucitania mentioned these names several times throughout the passage. Perhaps introducing a number of repetitions of a subject phrase into background passages could improve ease of use for a multi-channel browsing system.

How Desirable is Simultaneous Listening?

Many subjects felt that the task of listening to multiple channels, especially more than two at once, was too difficult, required "too much effort," or was "too distracting." One subject commented that "in a real conversation, if I didn't catch something I could ask." However, several subjects responded favorably to the idea of a multi-channel browsing system saying—"I'd love it, my favorite pastime is listening to conversations on the subway" and "it would be a way to pick up interesting facts."

Discussion

The goal of this experiment was to gain insight about the type of performance that could be expected in a multi-channel speech browsing system (as described in [Arons 1992] and [Mullins 1993]). In particular, this study attempts to determine how well subjects can detect "targets" (i.e., topics of interest)[10] in a secondary channel, while attending to a primary one. The results show a clear declination in the subject's performance on both tasks as the number of background channels increases. Given two simultaneous channels of speech, subjects listening comprehension performance did not significantly decline over the one channel condition, and they were able to detect 63.1% of the targets in the secondary channel. However, when the number of simultaneous channels increases to three (one primary and two target channels), subjects listening comprehension scores decreased significantly as did their target detection performance. These findings agree with the subject's perception of task difficulty.

Number of Simultaneous Channels

Given these results, it seems that increasing the number of channels beyond three would cause further decrements in performance. And yet the goal of a multi-channel browsing system is "to reduce the amount of time that is required to find [information] of interest to a listener" (p. 4, [Mullins 1993]). Therefore, the goal is to present as many channels simultaneously as possible—the more channels, the greater the benefit of simultaneous over serial presentation. The question then becomes, how do we go beyond two simultaneous channels without seriously impairing performance?

Attentional Enhancements

One answer is to enhance the segregation of the speech signals. "Synthesis of perceptual cues by a machine for a human listener might allow an application to perceptually nudge a user, making it easier to attend to a particular voice, or suggest that a new voice come into focus" (p. 36 [Arons 1992]). As described in the section—Enhancing Selective Attention—factors such as spatial separation, differential filtering, degree of synchrony, intensity differences, and pitch differences between competing messages can be manipulated to enhance separation of a single message in the presence of several others in the background. However, enhanced segregation of the speech signals can also result in increased interference. As Treisman states:

> "Any over-all feature which makes the two irrelevant messages distinguishable as sounds makes them more distracting than a single irrelevant message, but if there is no cue to differentiate them, they are easy to 'shut out' as a single one." (p. 540, [Treisman 1964]).

This corresponds to Treisman's findings that selective attention to a primary message declined when secondary channels differed along even a single dimension (e.g., voice or spatial location).

An important question is how to enhance the separability of multiple channels of speech to allow efficient switching of attention between them and yet allow focus on a single one. There are many contradictory variables, so the problem is not trivial. For example, presentation of a cue tone on a background channel can be used to draw a subject's attention to new topics [Mullins 1993], however, this will also distract them from their primary focus. Treisman [Treisman 1967] emphasizes the importance of "setting the [perceptual] filter"

[10]One difficulty, is that the targets in the study do not necessarily correspond to topics of interest or information that is important to the user as they would in an actual multi-channel browsing task.

before the messages are played (i.e., clearly indicating which is primary). Without knowing in advance what information is important to a user, this kind of pre-selection may not always be possible or 100% accurate. In addition, what role does the prosody of a speech message naturally play in drawing the subject's attention to emphasized material? Most of the target detection experiments use monotone speech[11] to avoid any confounding effects (e.g., stress placed on a target word) so this has not been studied.

Another contradiction concerns the role of spatial separation. While spatial separation of signals can enhance a subject's ability to respond to one of them [Spieth 1954, Webster 1954], the more separation, the greater the amount of time required to switch attention between them [Rhodes 1987].

Most of the studies on aids to selective attention deal primarily with attention to a main speech signal and consider the secondary channels as "irrelevant". However, in a multi-channel browsing system, the goal is to pick up important information from these background channels. Although this study has not addressed the role of "attentional enhancements", further studies like this one (i.e., which consider both primary and secondary channels) can be used to help determine the tradeoffs between these variables.

Cognitive Load and Density of Speech Information

Another question regards the cognitive load that this kind of multi-channel browsing system would impose. Clearly, listening silently requires much less attentional resources than shadowing, given the subject's ability to detect targets in the secondary channels. However, subjects complained that the task was difficult and required too much effort, especially when there are three simultaneous channels. Listening to speech, as opposed to viewing a display, has the advantage of keeping our hands and eyes free for other activities. And yet, it seems there would be very little cognitive capacity remaining for any additional activity—subjects' eyes were not open to look at other things, they were tightly shut in deep concentration.

One way to address this problem is to reduce the density of the information presented on each channel. Webster and Thompson report that the number correct responses was "significantly lower for material having equal density per channel than for the material having 70-5-20-5[12] message density." In the study described in this paper, each signal was a prose passage with approximately the same density of information. The AudioStreamer system [Mullins 1993] uses three channels of audio news, again, all with similar high density of information. Varying the amount of information presented on each channel may assist in reducing the cognitive load of the tasks while maintaining the efficiency benefit of simultaneous presentation.

An intuitive analogy can be made with a cocktail party. Imagine the first few guests arriving, a small intimate conversation begins...several others join the party and separate groups begin to form...at some point the conversation in the room builds to a clamor, making resolution of a single voice or conversational thread difficult. A multi-channel conversational speech system must be like a cocktail party, enabling the user to browse through a multitude of information while not allowing them to get lost in the clamor.

---

[11]Given that the degree to which a talker's intonation can be controlled by mearly instructing them to speak in a "monotone voice" is very limited.

[12]This represents the percent breakdown of content for four channels of material.

Acknowledgements

Thanks to Nat Durlach for his direction and support during this project. Jill Kliger and Earl Rennison kindly provided their voices for the passages. Thanks to Chris Schmandt, Barry Arons, and Atty Mullins for their support and encouragement with regard to this work.

References

[Arons 1992] B. Arons. A Review of the Cocktail Party Effect. Journal of the American Voice I/O Society, 12:35–50, 1992.

[Arons 1993] B. Arons. SpeechSkimmer: Interactively Skimming Recorded Speech. In Proceedings of the ACM Symposium on User Interface Software and Technology, pages 187-195. ACM, 1993.

[Arco 1991] E. H. Babin, C. V. Cordes and H. H. Nichols, Ed. (1991). ARCO TOEFL: Test of English as a Foreign Language. Prentice Hall.

[Blauert 1983] J. Blauert. Spatial Hearing. Cambridge, MIT Press, 1983.

[Bookbinder 1979] J. Bookbinder and E. Osman. Attentional Strategies in Dichotic Listening. *Memory & Cognition*, 7(6):511–520, 1979.

[Bregman 1990] A. S. Bregman. Auditory Scene Analysis: The Perceptual Organization of Sound. Cambridge, The MIT Press, 1990.

[Brokx 1982] J. P. L. Brokx and S. G. Nooteboom. Intonation and the Perceptual Separation of Simultaneous Voices. *Journal of Phonetics*, 10:23–36, 1982.

[Cherry 1953] E. C. Cherry. Some Experiments on the Recognition of Speech, with One and Two Ears. *Journal of the Acoustical Society of America*, 25:975-979, 1953.

[Cherry 1954] E. C. Cherry and W. K. Taylor. Some Further Experiments on the Recognition of Speech, with One and Two Ears. *Journal of the Acoustical Society of America*, 26:554-559, 1954.

[Cohen 1992] M. Cohen. Integrating Graphic and Audio Windows. *Presence*, 1(4):488-481, 1992.

[Dennis 1977] I. Dennis. Component Problems in Dichotic Listening. *Quarterly Journal of Experimental Psychology*, 29:437–450, 1977.

[Divenyi 1989] P. L. Divenyi and S. K. Oliver. Resolution of Steady-State Sounds In Simulated Auditory Space. *Journal of the Acoustical Society of America*, 85(5):2042–2052, 1989.

[Durlach 1992] N. I. Durlach, A. Rigopulos, X. D. Pang, W. S. Woods, A. Kulkarni, H. S. Colburn and E. M. Wenzel. On the Externalization of Auditory Images. *Presence*, 1(2):251-257, 1992.

[Egan 1954] J. P. Egan, E. C. Carterette and E. J. Thwing. Some Factors Affecting Multi-Channel Listening. *Journal of the Acoustical Society of America*, 26(1):774–782, 1954.

[Glucksberg 1970] S. Glucksberg and J. G. N. Cowen. Memory for Nonattended Auditory Material. *Cognitive Psychology*, 1:149–156, 1970.

[Inoue 1981] T. Inoue. Effects of Shadowing and Selective Attention In Dichotic Listening. *Psychologia*, 24:21–31, 1981.

[Irwin 1973] H. J. Irwin and W. G. Noble. The Role of Relative Intensity in Selective Listening. *Australian Journal of Psychology*, 25(1):23–27, 1973.

[Lawson 1966] E. A. Lawson. Decisions Concerning the Rejected Channel. *Quarterly Journal of Experimental Psychology*, 18:260–265, 1966.

[Moray 1959] N. Moray. Attention in Dichotic Listening: Affective Cues and the Influence of Instructions. Quarterly Journal of Experimental Psychology, 11:56–60, 1959.

[Moray 1970] N. Moray. Attention: Selective Processes in Vision and Hearing. New York, Academic Press, 1970.

[Mullins 1993] A. Mullins. Eavesdrop: A Spatial Display for Audio News. Masters Thesis Proposal. Massachusetts Institute of Technology, 1993.

[Norman 1969] D. A. Norman. Memory While Shadowing. *Quarterly Journal of Experimental Psychology*, 21:85–93, 1969.

[Norman 1976] D. A. Norman. Memory and Attention. John Wiley and Sons, 1976.

[O'Shaughnessy 1990] D. O'Shaughnessy. Speech Communication: Human and Machine. Addison-Wesley, 1990.

[Rhodes 1987] G. Rhodes. Auditory Attention and the Representation of Spatial Information. *Perception & Psychophysics*, 42(1):1–14, 1987.

[Spieth 1954] W. Spieth, J. F. Curtis and J. C. Webster. Responding to One of Two Simultaneous Messages. *Journal of the Acoustical Society of America*, 26(1):391–396, 1954.

[Stephens 1988] R. F. Stephens and J. L. Pate. Response Competition and Target Word Detection During Shadowing in a Dichotic Listening Task. *The Journal of General Psychology*, 115(3):285–292, 1988.

[Stifelman 1993] L. J. Stifelman, B. Arons, C. Schmandt and E. A. Hulteen. VoiceNotes: A Speech Interface for a Hand-Held Voice Notetaker. In Proceedings of INTERCHI '93, pages 179-186. ACM SIGCHI, 1993.

[Treisman 1964] A. Treisman. Verbal Cues, Language, and Meaning in Selective Attention. *American Journal of Psychology*, 77:206–219, 1964.

[Treisman 1967] A. Treisman and G. Geffen. Selective Attention: Perception or Response? *The Quarterly Journal of Experimental Psychology*, XIX(1):1–17, 1967.

[Treisman 1964] A. M. Treisman. The Effect of Irrelevant Material on the Efficiency of Selective Listening. *The American Journal of Psychology*, LXXVII(4):533–546, 1964.

[Underwood 1974] G. Underwood. Moray vs. the Rest: The Effects of Extended Shadowing Practice. *Quarterly Journal of Experimental Psychology*, 26:368–372, 1974.

[Webster 1955] J. C. Webster and L. N. Solomon. Effects of Response Complexity Upon Listening to Competing Messages. *Journal of the Acoustical Society of America*, 27:1199–1203, 1955.

[Webster 1954] J. C. Webster and P. O. Thompson. Responding to Both of Two Overlapping Messages. *Journal of the Acoustical Society of America*, 26(3):396–402, 1954.

[Yost 1977] W. Yost and D. Nielsen. Fundamentals of Hearing. Holt, Rinehart, and Winston, 1977.

[Yost 1994] W. A. Yost. Divided Attention With Up To Three Sound Sources: A Cocktail Party. Unpublished Manuscript. Parmly Hearing Institute, Loyola University, 1994.

[Zelniker 1974] T. Zelniker, J. Rattok and A. Medem. Selective Listening and Threshold for Tones Appearing On a Relevant and On an Irrelevant Input Channel. *Perception & Psychophysics*, 16(1):50–52, 1974.