Searching in Audio: The Utility of Transcripts, Dichotic Presentation, and Time-compression

Abhishek Ranjan[1], Ravin Balakrishnan[1], Mark Chignell[2]

[1]Department of Computer Science
University of Toronto
www.dgp.toronto.edu
aranjan| ravin @dgp.toronto.edu

[2]Dept. of Mechanical & Industrial Engineering
University of Toronto
www.mie.toronto.edu
chignel @mie.utoronto.ca

ABSTRACT

Searching audio data can potentially be facilitated by the use of automatic speech recognition (ASR) technology to generate text transcripts which can then be easily queried. However, since current ASR technology cannot reliably generate 100% accurate transcripts, additional techniques for fluid browsing and searching of the audio itself are required. We explore the impact of transcripts of various qualities, dichotic presentation, and time-compression on an audio search task. Results show that dichotic presentation and reasonably accurate transcripts can assist in the search process, but suggest that time-compression and low accuracy transcripts should be used carefully.

Author Keywords

Dichotic listening, transcripts, audio time-compression.

ACM Classification Keywords

H5.2 [User Interfaces]: Interaction styles, Auditory interfaces

INTRODUCTION

Despite the fact that speech is an integral part of human-human discourse and is typically the primary mode of communication in various types of meetings, outside of courts and political bodies (e.g., parliaments) it is seldom recorded or transcribed. Speech also conveys emotions and many subtle nuances that are lost after transcription to text. Along with these natural advantages, speech is relatively easy to capture, with low cost and effort.

While the expressiveness of speech makes audio archives a rich information source, the sequential nature of audio makes it time consuming to access, and as the size of archives increase it becomes impractical to search by playing it back as a single stream at normal speed. For example, if a student has to search within the recorded audio of a three hour lecture for the answer to a specific question, listening through the entire recording until the answer is encountered is clearly an onerous task. Advances in audio and speech processing techniques, such as dichotic and spatial presentation of audio, time-compression, and automated speech-recognition (ASR), provide opportunities for searching audio efficiently. ASR technology is of particular interest because it can convert audio signals to text transcripts, which have the following advantages: 1) it is typically faster to read the transcript than to listen to the audio sequentially; 2) text retrieval methods are available to jump into segments of text that match a query.

Unfortunately, current ASR technology does not work well in natural environments where there is no feedback between transcribing and speaking (this is in contrast to a dictation setting where the speaker is watching the output from ASR software and can adjust the input by, for example, slowing down or enunciating more clearly). The situation worsens when it is infeasible to provide speaker specific training data. This results in situations where transcription errors make the use of text search (querying) on the transcripts impractical. In cases where no matching text is found because the matching portion of the audio has been inaccurately transcribed, or where there are multiple hits which have to be differentiated, listening to the audio may be necessary to find the answer. Thus, despite the existence of ASR and text search technology, dealing with audio data continues to require techniques for facilitating easy browsing and skimming of the audio itself.

While several tools have been developed to facilitate audio browsing [3] [12-15], there exists little empirical research to guide the design of new audio browsing and searching techniques that leverage partially accurate transcripts, dichotic presentation, and time-compression of audio. In response to this deficiency, our research explores the following questions: Can partially accurate transcripts help in searching audio, or do the inaccuracies in the transcript make it essentially useless for this task? Does dichotic presentation help or hinder searching through audio? Does time-compression of audio help or hinder searching through audio? How do these different techniques for facilitating search in audio work in combination? In this paper, we report on an experiment that attempts to shed light on these questions, with the goal of providing insights that can help to guide the future design of audio search interface.

CHI 2006, April 22–27, 2006, Montréal, Québec, Canada.

RELATED WORK

Recent advances in ASR technology have made it possible to automatically generate partially accurate transcripts, and researchers have begun building systems for browsing audio using these transcripts [19, 21-24]. Whittaker et al. [23] designed a system, SCANMail, that leveraged partially accurate transcripts for browsing voicemail data. SCANMail displays the voicemail transcript with important phrases such as the caller's name and number highlighted. Users can read, search, and annotate these transcripts, as well as randomly access the underlying audio by simply selecting the relevant portions of the transcript. A user evaluation showed that this interface outperformed a conventional voice mail system for search, information extraction, and summarization tasks. Whittaker et al. [21] extended the use of the transcripts by developing an interface for semantic editing of audio via the corresponding transcripts. Their evaluations showed that this transcript based semantic editing of audio is more efficient than conventional acoustic signal based editing even when transcripts are only partially accurate. Stark et al. [17] studied the effect of transcript quality on user performance in retrieval tasks involving speech documents.

Apart from ASR, other insights to facilitate efficient audio browsing can be gleaned from auditory psychology and studies of human attention. Cherry and Taylor [7] found that when two ears receive different audio signals (in what is known as a "dichotic" listening task), people can listen to only one of the audio signals (i.e., through one ear) while tuning out the other. This selection ability has been termed the *cocktail party effect* [1]. To demonstrate this effect, Cherry conducted an experiment in which participants were presented two audio streams dichotically and were asked to attend to one stream and to concurrently recite aloud that stream. It was observed that participants could successfully recognize and verbalize every word, but often had little idea as to what the message was about [6]. Broadbent [5] further investigated the task of shadowing in a dichotic listening task and posited the existence of a selective filter that could attend to only one of the competing inputs.

Spieth et al. [16] performed a series of experiments investigating responses to one of two simultaneously presented messages. They found that larger separation between loudspeakers improves the response, suggesting that presenting two audio channels separately to each ear via headphones is probably the best configuration for a selective listening task. Webster et al. [20] studied human ability to respond to two simultaneously presented audio messages. They found that if participants were allowed to switch attention between two simultaneously presented audio messages, an average of 60% of each of the two messages was received or understood, resulting in a greater total information intake per unit time than messages presented one at a time. Stifleman [18] conducted an experiment to study the effects of simultaneous presentation of audio in tasks involving comprehension and target-monitoring. The performance of participants decreased in both of the tasks as the number of simultaneous audio channels increased. These results agree with Webster et al's [20] findings that the ability to rapidly shift attention between two sources does not help much in the information intake if the information content in each source is high.

Various systems have made use of these experimental findings in the design of audio browsing interfaces [11, 13-15]. Schmandt et al.'s AudioStreamer [15] used spatial audio in an audio browsing interface, enabling a user to simultaneously listen to three spatialized audio streams. Sawhney et al. [13] introduced audio and speech based interaction techniques to be used in nomadic environments. Kobayashi et al.'s Dynamic Soundscape [12] provided a spatial interface for temporal navigation of audio data, where the user could hear multiple portions of an audio clip simultaneously from speakers arranged in different spatial positions (in a circle around the user's head). Users could select a speaker, and hence jump to a new position inside the audio, by moving their heads in the speaker's direction.

In addition to dichotic and spatialized presentation, time-compression of audio can also be used as a tool in audio interface design. Previous studies have shown that single well-learned phonetically balanced words can remain intelligible at 10 times the normal speed and that connected speech remains comprehensible at twice the normal speed [2]. Arons's SpeechSkimmer [3] system introduced fast audio browsing techniques using time-compression of audio. The effect of combining transcripts with time-compressed speech was studied by Vemuri et al. [19]. Their experimental interface displayed the transcripts and enabled audio playback at different speeds. The results showed that as transcript quality degraded, speech comprehension was correspondingly negatively impacted. Furthermore, an increase in playback speed also led to degraded comprehension. However, they also found that audio presentation with error-laden transcripts resulted in better comprehension than without any transcript.

In summary, humans can leverage dichotic presentation or simultaneous spatial presentation of multiple audio streams, if the information intake required is not too high. Further, time-compression of audio and text transcripts can also assist in browsing and comprehension. However, it is unclear as to how dichotic presentation combined with transcripts will impact user performance. This will be examined in the study reported below, along with the impact of combining the three factors: dichotic presentation, transcripts, and time-compression. We focus on an auditory search task since research [20] has shown that simultaneous presentation of audio is most effective when the task's required information intake is not as uniformly high as in complete comprehension tasks. Search tasks tend to require quick skimming and keyword spotting rather than complete understanding of the audio, hence the required information intake is lower and thus more likely to benefit from dichotic presentation.

EXPERIMENT

Goals

The overall goal of the experiment was to test the effect of dichotic presentation, transcripts, and time-compression on a user's ability to perform search tasks. The following questions were asked:

- Can people find facts in two separate audio streams faster if they are presented dichotically as compared to sequentially?
- Does the availability of text transcripts help or hinder the task, and how does the quality of the transcript factor in?
- Does time-compression of audio further improve search speed when used in combination with transcripts and dichotic presentation?

Techniques

We tested two audio stream presentation techniques, and refer to this independent variable as *Stream*:

- Single stream (S1): The user can listen to the audio in a single stream and play the audio from any point of the stream. This represents the status quo.
- Dichotic stream (S2): The user can play two different parts of the audio clip simultaneously as two separate streams, one stream per ear.

Time-compression techniques shorten the playback time of audio, causing an increase in the playback speed (measured in terms of word per minute) [9]. We tested three compression levels for this *Speed* independent variable:

- Uncompressed (1x): Audio stream is not time compressed. The playback is done at normal speed.
- 1.5 times compressed (1.5x): Audio is compressed by a factor of 1.5. The playback is 1.5 times normal speed.
- 2 times compressed (2x): Audio is compressed by a factor of 2. The playback is 2 times normal speed.

This choice of speed levels is motivated by our intention to design an interface for fast search in audio and a fact discussed by Arons [2] that "connected speech remains comprehensible to a 50% compression (twice normal speed)". A further motivation is that we wanted to look at performance at both ends of the spectrum of possibilities (i.e., 1x to 2x compression) as well as at one intermediate point to obtain a sense of the bounds.

We measure the quality of a transcript as the percentage of the words in it which are uttered in the audio. We assume that the sequence of words matches in both the audio and transcript streams. We generated transcripts of different qualities using software which randomly 'garbled' a perfectly accurate manually generated transcript, with the word error rate (WER) and garbling pattern controlled by input parameters. For example, in order to generate a transcript with 50% WER, out of ten consecutive words the software would replace five randomly selected words by five random words from a dictionary. The quality levels were selected such that they approximately cover the entire spectrum of possible transcripts generated by ASR technology under various input conditions [8]. We tested five levels of this transcript *Quality* independent variable:

- No transcript (Q0)
- 25% accurate transcript (Q25)
- 50% accurate transcript (Q50)
- 75% accurate transcript (Q75)
- perfectly accurate transcript (Q100).

All combinations of different levels of Stream, Speed, and Quality were tested. We denote a combination by Si-nx-Qj where i = 1, 2; n = 1, 1.5, 2; and j = 0, 25, 50, 75, 100. For example, S1-1.5x-Q75 represents single stream audio (S1) at 1.5x speed with 75% accurate transcripts (Q75).

Interface

We built a simple interface for presenting one and two streams of audio with the associated transcript. The transcript corresponding to the audio is shown in a text browser with time-code on the left of the text display. The user can click at any position within the transcript in this browser to playback the audio beginning at that time-code. Once the playback starts, the portion of the transcript whose relevant audio is being played is highlighted in yellow. The user can also pause or play the audio by clicking anywhere inside of the rectangle at the bottom of the window. Figure 1 shows the single stream version of our interface.

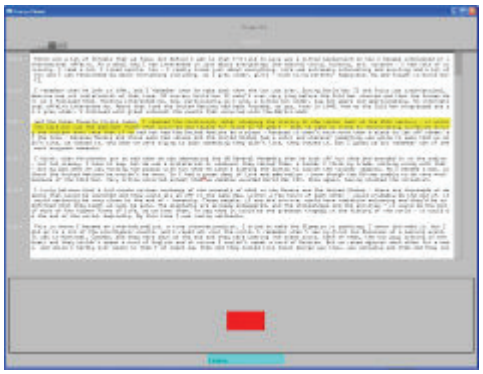

Figure 1. Single stream browsing interface. Text corresponding to audio currently being played is highlighted in yellow.

The dichotic stream version of the interface presents two transcripts in adjacent text browsers (Figure 2). Both streams are time synchronized and can only be played simultaneously and synchronously. When the user pauses one of the streams both streams pause. Similarly, if the user clicks on any point in either transcript, the system plays both audio streams starting from the same time instant. The

user can slightly adjust the volume of each stream separately, but can not completely mute any stream. This potentially introduces a small confound in the experiment in that different participants could use slightly different volume settings which could result in performance differences between participants. However, we included this feature because pilot studies indicated that participants had different subjective preferences for comfortable volume levels in each ear, and volume levels are typically not constant throughout most of the audio clips used. Furthermore, this design is representative of the volume adjustments that would be available in any real audio browsing interface, thus increasing the external validity of our experiment. However, in order to force the user to listen to two streams all the time in the dichotic presentation case (S2), the maximum allowed volume level difference between the two ears was set to be 0.3 on a scale of 0 to 1.

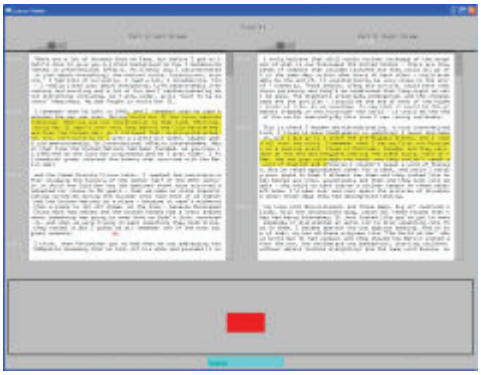

Figure 2. Dichotic stream browsing interface. The transcript browser on the left plays audio in the left ear and the browser on the right plays audio in the right ear. Text corresponding to audio currently being played is highlighted in yellow.

When no transcript is available (case Q0), we use the same interface as shown in Figure 1 and Figure 2 for single stream and dichotic stream audio respectively, except that the transcript text is replaced with dotted lines. The user can, however, click anywhere on the dotted lines to start playing the audio at that time-code. As such, the basic interaction is similar to when the transcript text is available. The primary difference is that the notion of selecting a portion of text obviously does not exist when there is no transcript, and yellow highlighting just denotes the current location along the time-code.

Task

Participants performed a search task in which they were asked to find the answer to a specific question within an audio clip. As discussed previously, we chose to use a search task rather than a comprehension task because previous research [20] indicates that simultaneous presentation of multiple audio streams is most effective when the information intake required is not too high. The quick skimming and keyword spotting nature of search tasks does not require a complete understanding of the entire audio stream and as such requires a lower information intake, which likely allows it to benefit from dichotic presentation.

The play-length of each audio clip, in uncompressed form, was approximately 4 minutes. This particular play-length was selected to simulate realistic audio search tasks, which usually tend to be lengthy, while still allowing for a manageable experiment lasting 60-90 minutes.

We used public domain audio clips of speeches or lectures in computer science, physics, chemistry, politics, business, literature and mythology. We chose this content because the recordings are of high quality, these famous speakers tend to speak with proper and easily understood enunciation, and high quality transcripts are readily available for them.

In order to prevent participants from getting too familiar with a particular audio clip, we only asked one question per audio clip, and repeated the task for several different clips. To formulate appropriate questions, we first determined the location where the answer to that question should lie in the clip and selected some unique fact from that location. Further, we ensured that the answer to a question could be found at only a single location in the audio. A question was then constructed in such a way that it contained some keywords that characterized the answer unambiguously. A sample question along with the corresponding transcript is shown in Appendix A. For the dichotic stream techniques (S2), each 4 minute audio clip was split into two halves. Subjects listened to the first half in the left ear and the second half in the right ear. In different clips the answers were located at different portions of the audio

Participants

13 participants (3 female and 10 male, all university students) volunteered for the experiment. All participants were fluent in English and had no trouble understanding the given audio clips and associated questions. Only one of the participants was somewhat familiar with non-pitch-adjusted time-compression (which sounds like chipmunks), and two had ever knowingly tried dichotic listening. Apart from refreshments during the experiment, they did not receive any compensation for their participation.

Design and Setup

All participants performed the tasks under all conditions. 2 Stream levels (S1, S2), 3 Speed levels (1x, 1.5x, 2x), and 5 Quality levels (Q0, Q25, Q50, Q75, Q100) resulted in a total of 2 x 3 x 5 = 30 conditions.

Each participant listened to total 30 clips, each clip under a different condition. 15 clips were presented in the single Stream (S1) condition, and the other 15 in the dichotic Stream (S2) condition. Participants were randomly divided into two groups of 7 and 6, and the presentation of S1 and S2 was counterbalanced between the groups (i.e., one group was presented with 15 clips using S1 followed by 15 clips using S2, and the second group was presented the clips in the reverse order: S2 followed by S1).

For each Stream condition, clips were presented at 3 Speed levels in subsets of 5 clips in the following sequence: 5 clips at 1X, 5 clips at 1.5X, and 5 clips at 2X. Each subset of 5 clips had 5 different kinds of transcript Quality (Q0, Q25, Q50, Q75, Q100) associated with it. We randomized the order in which these 5 clips with different transcript quality levels were presented. The experimental design is summarized in Figure 3 and resulted in a total of

13 participants x
2 Stream levels x
3 Speeds x
5 transcript Qualities
= 390 observations.

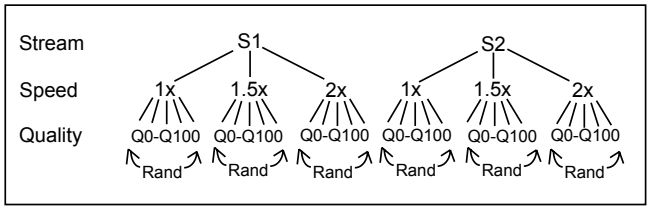

Figure 3. Design of the experiment. Various levels of Stream, Speed, and Quality are shown. Q0-Q100 represents all transcript quality levels. 'Rand' implies randomization in the associated variable.

We define another control variable, *Position*, indicating the position of the answer in the audio clip. Position is expressed in terms of percentage of a clip's total time length. In our experiment we considered two answer positions: answer located in the first half of the clip (Position 1) and second half of the clip (Position 2).

To ensure that the time to perform the task was balanced as far as answer location was concerned, answer positions were distributed with mean position being approximately equal to 50% of the audio clip length. Figure 4 shows the distribution of answer positions with respect to questions.

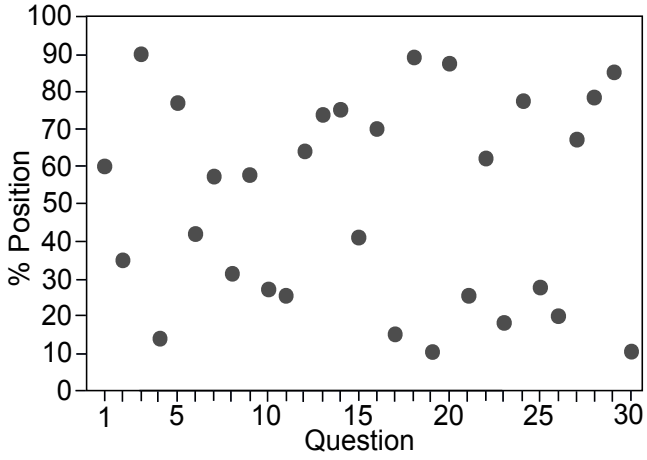

Figure 4. A scatter plot of answer position distribution with respect to question numbers.

Participants had to successfully answer each question before proceeding to the next question and audio clip. Consequently, they sometimes had to replay the audio or read the transcript several times in order to figure out the answer. In the few cases where participants already knew the answer, they were told to find the exact answers in the clip. We logged all the following data during the experiment: all mouse movements and clicks with time-stamps, audio play-pause instants and corresponding portions in the transcripts, audio stream information, and volume levels of both streams over time. The experimenter also observed the participants and took notes.

All participants underwent a short practice session before the experiment. They were asked to complete 2 trials using S1 and 2 trials using S2 in exactly same manner as they would in the experiment. These practice sessions helped familiarize them with the interface involving transcripts, dichotic presentation, and time-compression, decide strategies for searching in audio, and get some idea about the types of questions.

Several different strategies could be used for completing the search task and these differences could affect the validity of results. In order to minimize the effect of strategy difference across participants, all participants were asked to carefully look for keywords, digits, and proper nouns since they are easy to pick out in text as well as in audio. Participants were also cautioned that even inaccurate transcripts might contain keywords that might lead to the answer very quickly. Although, it is difficult to completely control the individual strategy for answer finding, these precautions were aimed at minimizing the major differences in strategies across participants. Furthermore, all participants were given a post-experiement questionnaire aimed at eliciting information about the strategies adopted to find answers, experiences during the experiment, and perceived difficulty of the task. We also requested feedback from them for ways to improve the system.

Hypotheses

From our literature review of the independent impact of transcripts, dichotic presentation, and time-compression on audio comprehension, we came up with five hypotheses with regards to the impact of these three factors on a user's ability to search in audio. We define the dependent variable 'search-time' as the time taken by the user to find the answer to a given question by listening to the audio data and/or reading the transcripts. The orderings of various techniques in these hypotheses are based on the order of the corresponding search-time dependent variable.

H1. Dichotic presentation is expected to decrease search-time. i.e., S1 > S2.

H2. Search-time should decrease as transcript quality gets better. i.e., Q100 < Q75 < Q50 < Q25 < Q0.

H3. Dichotic presentation should reduce search-time regardless of transcript quality. i.e., S1Qi > S2Qi, for i = 0, 25, 50, 75, and 100.

H4. Time-compression should decrease search-time. i.e., 2x < 1.5x < 1x.

H5. Search-time corresponding to various presentation techniques should be influenced by the position of the answer in the audio clip. More specifically, dichotic presentation is expected to significantly reduce search-time when the answer is in position 2.

Results

We performed analysis of variance using search-time as the dependent variable and the four aforementioned independent variables.

Testing Hypothesis H1

Stream had a significant main effect on search-time ($F_{1, 12}$ = 15.72, p = 0.002). Mean search times for S1 and S2 were 69 and 82 seconds respectively (Figure 5), and the difference between least square means was significant (p = 0.006). This result confirms Hypothesis H1 in that dichotic presentation significantly reduces the search time.

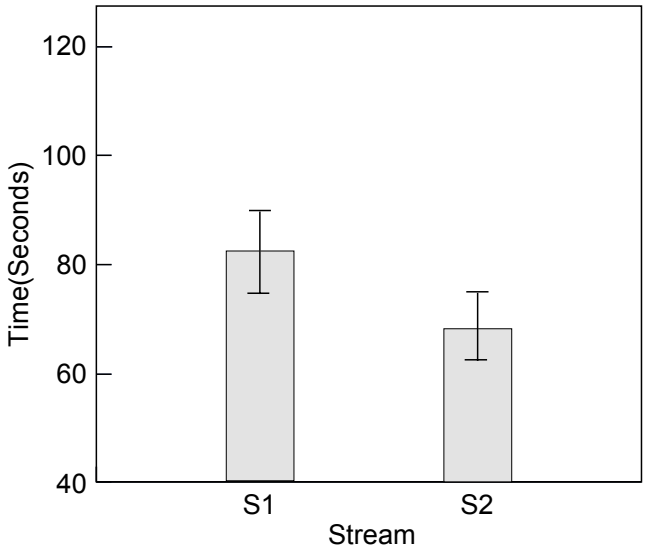

Figure 5. Mean search-time for the two stream techniques (S1, S2) for all participants, with 95% CI error bars.

Testing Hypothesis H2

There was a significant main effect for transcript quality (Q0 to Q100) on search time ($F_{4, 48}$ = 12.16, p < 0.001). Mean overall search-times were 94, 83, 77, 68, and 57 seconds for Q0 to Q100 respectively (Figure 6). Differences between least square means with Tukey-Kramer adjustment indicated that search-time corresponding to Q100 was significantly faster than Q0, Q25, Q50, but not Q75. Furthermore, differences between pairs Q0-Q25, Q0-Q50, Q25-Q50, and Q50-Q75 were not significant, but Q0-Q75 was significant. These results indicate that having low quality transcripts (Q0, Q25, and Q50) does not provide much benefit. This result could be explained by considering the perceived difference between various transcript qualities. While analyzing the questionnaire responses we found that 4 out of 13 participants explicitly mentioned that low quality transcripts were rather distracting and therefore, they relied completely on audio playback. This corroborates the ANOVA analysis which did not find Q25 and Q50 to result in significantly better results than having no transcripts at all (Q0). These results only partially confirm Hypothesis H2, but are in agreement with the results of various previous studies [17, 19, 24].

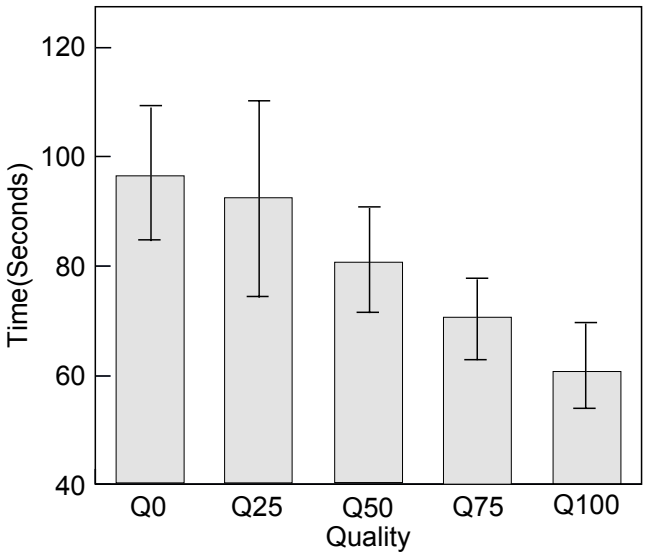

Figure 6. Mean search-time for the five transcript qualities (Q0-Q100) for all participants, with 95% CI error bars.

Testing Hypothesis H3

In Figure 7, we plot search-time for various transcript quality levels (Q0-Q100), separately for S1 and S2.

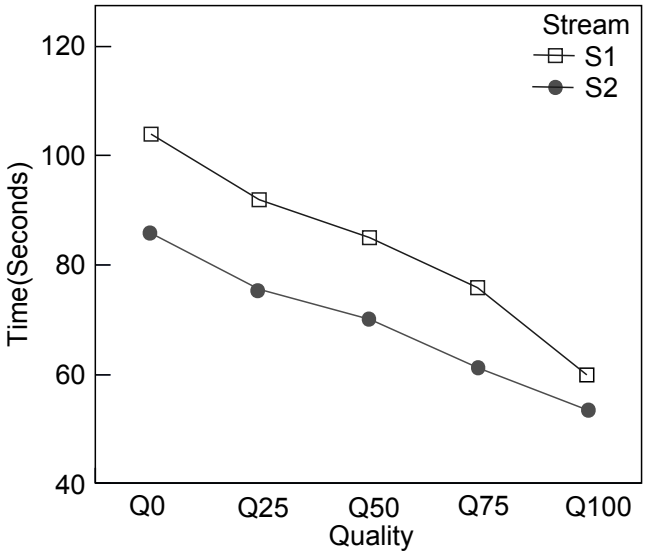

Figure 7. Mean search-time for all transcript quality levels (Q0-Q100) for the two stream techniques (S1, S2) for all participants.

Although we did not find a statistically significant Stream x Quality interaction ($F_{4, 47} = 0.30$, $p = 0.87$), from Figure 7, we see that for each transcript quality, mean search-time for S2 is slightly less than that for S1. Further analysis shows that as transcript quality varies from Q0 to Q100, the least square means difference, S1 – S2, follows a decreasing pattern (16, 17, 17, 14, and 6 seconds for Q0, Q25, Q50, Q75, and Q100, respectively). This partially confirms Hypothesis H3 in that dichotic presentation reduces search-time regardless of transcript quality. However, when transcript quality is high (e.g., Q100), dichotic presentation does not provide significant benefits, indicating that users mostly rely on transcripts when they are accurate.

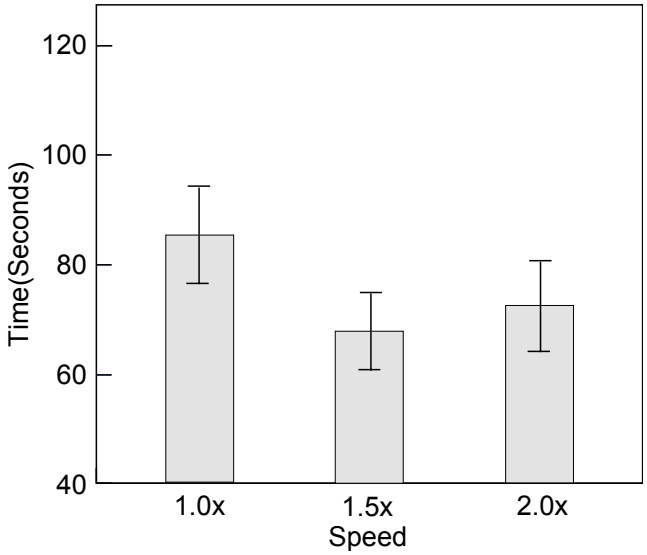

Figure 8. Mean search-time for different speed levels (1x, 1.5x, 2x), for all participants, with 95% CI error bars

Testing Hypothesis H4

There was a significant main effect for Speed on search-time ($F_{2, 24} = 7.60$, $p = 0.003$). Pair-wise least square means comparisons showed that search-times corresponding to speeds of 1.5x and 2x were significantly less than 1x, but the difference between 1.5x and 2x was not significant (Figure 8). This partially confirms Hypothesis H4. The ANOVA showed a Stream x Speed interaction at the p = 0.06 level ($F_{2, 24} = 3.16$, $p = 0.06$). A detailed analysis of search-time for various speed levels, grouped by stream technique, showed that for single stream presentation (S1), search time for 1.5x speed was significantly less than that for 1x speed, but on further increasing the speed to 2x search time increased (Figure 9). For S2 there was no significant change in search-time when speed goes from 1.5x to 2x. This could be attributed to higher cognitive load associated with the task of keyword searching in time-compressed playback. Analysis of user logs showed that participants using S1 technique repeated the playback in 6%, 8%, and 10% of trials for 1x, 1.5x and 2x speed levels respectively. Since repeated playback adds to the total search-time, this result could explain why participants took longer or did not improve upon search-time at 2x speed level. We note that Vemuri et al. [19] made a similar observation that comprehension of speech decreases with increasing speed in a comprehension task, while our results indicates the same trend in a search task.

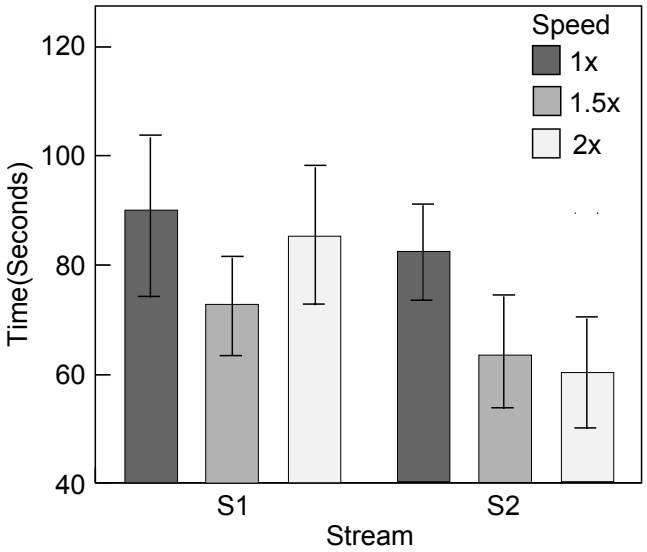

Figure 9. Mean search-time for the two stream techniques (S1, S2), grouped by Speed (1x, 1.5x, 2x), for all participants, with 95% CI error bars

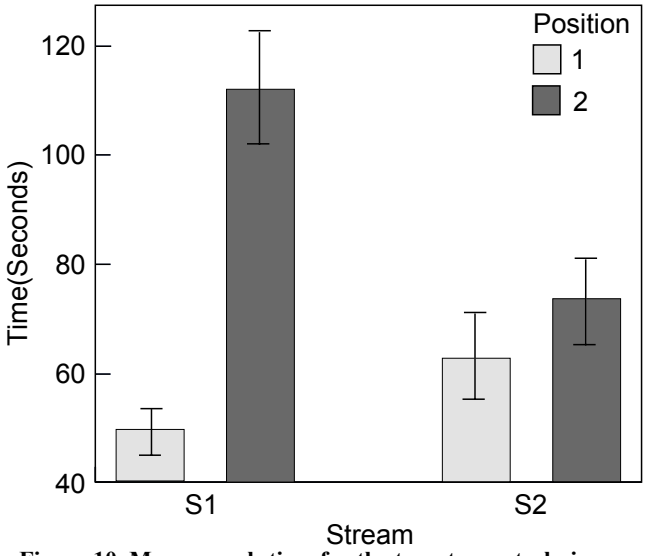

Figure 10. Mean search-time for the two stream techniques (S1, S2), grouped by answer position (1, 2), for all participants, with 95% CI error bars.

Testing Hypothesis H5

The answer to a question could lie in one of two positions: first half of the audio clip (Position 1) or second half (Position 2). We found a strong Position x Stream interaction ($F_{1, 12} = 48.73$, $p < .001$). While for S1 there was a significant difference between search-times associated with Position 1 and Position 2 ($p < 0.001$), for S2 this difference was insignificant ($p = 0.25$). This could be explained by the fact that for both stream techniques, the playback reached the answer at almost the same time if the answer was in position 1, but not if it was in position 2. The

dichotic playback (S2) reached the answer in position 2 earlier than S1 because it played both halves of the clip simultaneously and it thus took the user less time to find the answer. This confirms Hypothesis H5. Figure 10 illustrates.

DISCUSSION

During pilot studies participants indicated that finding keywords in dichotic streams was fairly manageable. They could also give a summary of the content, despite the fact that they could hardly remember much of the verbal audio content. This observation motivated us further to include playback at 1.5x and 2x speeds in the experiment to test the limits of human capability in switching attention between two channels.

In verifying Hypothesis H1, we found that dichotic presentation outperformed single stream presentation in the search task we studied. Previous studies [20] have shown that the human auditory system is capable of comprehending two streams of audio simultaneously when the information rate is low. Our results further reinforce those of Webster et al. [20] by showing that a search task can clearly benefit from dichotic presentation of two audio streams. All participants were able to determine the stream in which the answer was present using S2-1x technique in one pass, and then they would playback the portion where answer was located, focusing their attention only on that stream.

The application of dichotic presentation in search interfaces raises the issue of cognitive load experienced by the listener while trying to find answers in two streams. In our experiment, most of the participants were either amused or baffled when the experiment was explained to them. In the post-questionnaire the participants were asked to select the difficulty levels of the task (on a 3-point scale of 'okay', 'tiring', and 'very demanding') for techniques S2-1x, S2-1.5x, and S2-2x. Out of 13 participants, 7 found the dichotic stream at normal speed (S2-1x) to be 'okay', 5 found it to be 'tiring', and 1 found it to be 'very demanding'. In response to another question in the post-questionnaire, some participants said that they tried to comprehend the content of the audio in both the streams and were able to figure out which of them might contain the answer to the given question. However, regardless of the strategy used or the difficulty perceived, all the participants managed to find the answers to all of the questions using S2-1x.

Verification of Hypothesis H2 proved that transcripts clearly add a very useful stream of information to the search task. Whenever 100% accurate transcripts (Q100) were provided, participants found answers very quickly by skimming the text and relying less on the audio. However, our finding with regards to Hypothesis H2 that low quality transcripts (Q25, Q50) did not perform significantly better than having no transcript at all, suggests that providing such low quality transcripts is of limited value. Only 2 out of 13 participants used these inaccurate transcripts as guides for very quickly skimming the audio and found the answer location even when the answer was not present in the transcript. One participant made the following comment on keyword searching using inaccurate transcripts: "… *as long as neighboring words were similar it was okay*".

The partial confirmation of Hypothesis H3 demonstrates the effectiveness of combining transcripts with dichotic presentation for highly inaccurate transcripts (Q25, Q50). This suggests that people can pick up words of interest from dichotic streams while reading text. One of the participants commented on the dichotic presentation technique (S2) as: "*…with transcript, I tend to skim one part with my eyes while my ears listen to different part to quickly find keywords.*"

One participant followed an interesting strategy to deal with focusing attention in dichotic presentation: this participant turned down the volume of one channel and only read the transcript corresponding to that channel while focusing auditory attention to the other channel at higher volume. This strategy agrees with previous research on simultaneous presentation showing that difference in volume level helps to focus attention on one stream [4, 6]. It should be noted here that our dichotic stream presentation interface did not allow the user to completely mute one stream; it allowed only small deviations between the volume levels, therefore, participants had to listen to two streams all the time.

Hypothesis H4 indicates that in our experiment a speed of 1.5x was an optimal level at which participants could spot keywords quickly. Increasing the time-compression further resulted in either increase in search time (in case of S1) or no significant change (in case of S2). This suggests the possibility of finding an optimal time-compression level at which users can efficiently find keywords. However, any claim about complete understanding of underlying behavior or determination of the optimal compression level from this experiment would be premature.

Our original motive in introducing time-compression was to determine the limits of human ability to search for keywords in audio, and push these limits, if possible. The post-questionnaire responses indicate that most of the participants experienced high cognitive load while using dichotic presentation combined with time-compression. In the case of S2-1.5x or S2-2x techniques, 8 out of 13 participants found the search task to be 'very demanding', 3 participants found it 'tiring', and 2 participants found it 'okay'. However, despite the issues of cognitive load associated with this technique, the advantage of achieving significantly lower search time cannot be totally rejected. Therefore, we believe that further study is needed to explore in detail the interaction between dichotic audio and various time-compression levels.

DESIGN IMPLICATIONS

The analysis of results with respect to answer position and confirmation of Hypothesis H5 show that dichotic presentation outperforms the single stream audio when the

answers lie in the second half of the clip. This suggests that an interface that could split the audio and transcripts at any desired point and play the streams dichotically while displaying the corresponding transcripts could be beneficial. Most formal speeches, lectures, or even meetings, begin with an introduction, followed by details, discussion, and finally, conclusion. Splitting the audio archives of such events based on these broad topics could help the user quickly narrow down the search streams and then listen to these competing streams dichotically (similar to an earlier technique used in the Nomadic Radio system [13]). In addition to supporting the Nomadic Radio design with empirical data, our experiment results also suggest that the use of transcripts along with the audio can further improve performance.

While several participants expressed that low quality transcripts were unreadable and rather distracting, two participants actually made use of them to obtain contextual information. They reported that even though the answer was not present in the transcript, accurately transcribed neighboring words conveyed enough information to pinpoint the answer location. We notice here that while such transcripts would be hardly of any use for a user doing a comprehension task or a search engine doing a keyword finding task, they proved to be important in a search task for human readers since humans can efficiently integrate their command of language and domain knowledge with keyword search. These observations suggest an interface that removes known incorrect words and phrases [10] and shows only words transcribed with high confidence by the system. This could minimize the distraction caused and show some useful information at the same time.

While a careful use of audio time-compression in the interface could enhance the performance as demonstrated by Arons [2], our results caution that use of time-compression can drastically reduce the performance for certain tasks. However, an interface that allows the user to control dichotic presentation and time-compression levels would cater to the user's subjective preferences and could be very useful for fast keyword search in long audio tracks.

CONCLUSIONS AND FUTURE WORK

This paper has discussed approaches for improving a user's ability to browse audio, and has evaluated the combined impact of dichotic presentation, transcripts, and time-compression in the performance of a search task. While previous research has studied the impact of spatialization and dichotic presentation [6, 7, 11, 16, 20], transcripts [21, 23], and time-compression [2, 19] on audio tasks, and also the impact of transcripts and time-compression in combination [19], our work evaluates all three factors together and in various subset combinations.

Our results clearly demonstrate the value of accurate transcripts, but also illustrate the additional overhead users have to deal with when given a partially accurate transcript. A particularly interesting result is the value of dichotic presentation, particularly when transcripts are of low quality or do not exist at all. Given that 100% accurate transcripts cannot currently be generated automatically, and automatic searching within partially accurate transcripts does not provide reliable outcomes, this empirical result indicates that dichotic presentation is a valuable technique that should be exploited in interface designs for audio browsing and searching tools.

As improvements in speech recognition technology occur, the use of techniques such as those suggested here will need to be re-evaluated. One technique that might improve the value of partially accurate speech transcripts is to replace known incorrect words and phrases [10] with ellipses, thus reducing the need for the user to parse and discard erroneous transcriptions. Future research should evaluate the impact of a partially accurate transcript enhanced in this manner on user performance in search tasks, with and without dichotic presentation.

ACKNOWLEDGEMENTS

We thank our experiment participants, the CHI reviewers, and in particular the CHI meta-reviewer whose comments significantly improved the paper.

APPENDIX A

Following is a snippet of one of the 100% accurate transcripts used in our experiment, and is included here simply to illustrate the type of questions asked and transcripts used. The actual transcript was significantly longer and covered the full 4 minute long audio clip.

"Tonight, I chose to speak from the chamber of the Texas House of Representatives because it has been a home to bipartisan cooperation. Here in a place where Democrats have the majority, Republicans and Democrats have worked together to do what is right for the people we represent. We've had spirited disagreements. And in the end, we found constructive consensus. It is an experience I will always carry with me, an example I will always follow. I want to thank my friend, House Speaker Pete Laney, a Democrat, who introduced me today. I want to thank the legislators from both political parties with whom I've worked. Across the hall in our Texas capitol is the state Senate. And I cannot help but think of our mutual friend, the former Democrat lieutenant governor, Bob Bullock. His love for Texas and his ability to work in a bipartisan way continue to be a model for all of us."

An example question used for this transcript:

"The speaker gives thanks to the House Speaker. What is that House Speaker's name?"

Following is a snippet of a garbled (Q75) transcript:

"Every day, after a hard day's work, Jillian and I were allowed to walk out careers profile plains among the animals and you know, gazelle and giraffe recklessly zebra, one night jug rhino. One evening, two young male lions came and followed about twice microprocessors childhood of this room, which was a bit scary saddled speed exciting. And every morning when I woke up, I was in tangy dream. What magic. And that averting misgivings Louis Leakey realized, he says, that he'd diadem the person he'd been

looking for, for many years. Someone to Mallory and try and find out something about the behavior bugler our closest living relatives in their natural habitat. He gazer even know then how closely related to us feats subtractor but it was thought that they commune intractable very close and he believed that interdependence capacitances about their behavior would help him to have a better Dunn for how our own ancestors may have behaved. So off I went, recreating there were two serviceman Douglas overcome. First of all, how did he get money for a young, untrained girl?"

An example question used for this transcript:

"At what age did the speaker actually go to Africa?"

REFERENCES

1. Arons, B. (1992). A review of the cocktail party effect. *Journal of the American Voice I/O Society*, *12*. p. 35-50.
2. Arons, B. (1992). Techniques, perception, and applications of time-compressed speech. *American Voice I/O Society Conference*. p. 169-177.
3. Arons, B. (1997). SpeechSkimmer: a system for interactively skimming recorded speech. *ACM Transactions on Computer Human Interaction*, *4(1)*. p. 3-38.
4. Bregman, A. (1994). Auditory scene analysis: MIT Press, Cambridge, MA.
5. Broadbent, D. (1958). Perception and communication: Pergamon, New York.
6. Cherry, E. (1953). Some experiments of the recognition of speech, with one and with two ears. *Journal of the Acoustic Society of America*, *25*. p. 975-979.
7. Cherry, E. and Taylor, W. (1954). Some further experiments on the recognition of speech, with one and two ears. *Journal of the Acoustic Society of America*, *26*. p. 554-559.
8. Fiscus, J., Fisher, W., Martin, A., Przybocki, M., and Pallett, D. (2000). NIST Evaluation of conversational speech recognition over the telephone: English and Mandarin performance results. *DARPA Broadcast News Workshop*.
9. Henja, D. and Musicus, B. (1991). The SOLAFS time-scale modification algorithm. Technical Report, Bolt Beranek & Newman.
10. Inkpen, D. and Desilets, A. (2004). Extracting semantically-coherent keyphrases from speech. *Canadian Acoustics*, *32(3)*. p. 130-131.
11. Kilgore, R., Chignell, M., and Smith, P. (2003). Spatialized audioconferencing: what are the benefits. *IBM Center for Advanced Studies Conference (CASCON)*. p. 135-144.
12. Kobayashi, M. and Schmandt, C. (1997). Dynamic soundscape: mapping time to space for audio browsing. *ACM CHI Conference on Human Factors in Computing Systems*. p. 194-201.
13. Sawhney, N. and Schmandt, C. (2000). Nomadic radio: speech and audio interaction for contextual messaging in nomadic environments. *ACM Transactions on Computer-Human Interaction*, *(7)*. p. 353-383.
14. Schmandt, C. (1998). Audio hallway: a virtual acoustic environment for browsing. *ACM UIST Symposium on User Interface Software and Technology*. p. 163-170.
15. Schmandt, C. and Mullins, A. (1995). Audiostreamer: exploiting simultaneity for listening. *Extended Abstracts of the ACM CHI Conference on Human Factors in Computing Systems*. p. 218-219.
16. Spieth, W., Curtis, J., and Webster, J. (1954). Responding to one of two simultaneous messages. *Journal of the Acoustic Society of America*, *26(1)*. p. 391-396.
17. Stark, L., Whittaker, S., and Hirschberg, J. (2000). ASR satisficing: the effects of ASR accuracy on speech retrieval. *International Conference on Spoken Language Processing*.
18. Stifelman, L. (1994). The cocktail party effect in auditory interfaces: A study of simultaneous presentation. Technical Report, MIT Media Laboratory.
19. Vemuri, S., DeCamp, P., Bender, W., and Schmandt, C. (2004). Improving speech playback using time-compression and speech recognition. *ACM CHI Conference on Human Factors in Computing Systems*. p. 295-302.
20. Webster, J. and Thompson, P. (1954). Responding to both of two overlapping messages. *Journal of the Acoustic Society of America*, *26(1)*. p. 396-402.
21. Whittaker, S. and Amento, B. (2004). Semantic speech editing. *ACM CHI Conference on Human Factors in Computing Systems*. p. 527-534.
22. Whittaker, S. and Hirschberg, J. (2003). Look or listen: Discovering effective techniques for accessing speech data. *Proceedings of Human Computer Interaction*. p. 253-269.
23. Whittaker, S., Hirschberg, J., Amento, B., Stark, L., Bacchiani, M., Isenhour, P., Stead, L., Zamchick, G., and Rosenberg, A. (2002). Scanmail: a voicemail interface that makes speech browsable, readable and searchable. *ACM CHI Conference on Human Factors in Computing Systems*. p. 257-282.
24. Whittaker, S., J, H., Choi, J., Hindle, D., Pereira, F., and Singhal, A. (1999). SCAN: designing and evaluating user interfaces to support retrieval from speech archives. *ACM SIGIR Conference on Research and Development in Information Retrieval*. p. 26-33.