Improving Speech Playback Using Time-Compression and Speech Recognition

Sunil Vemuri, Philip DeCamp, Walter Bender, Chris Schmandt
MIT Media Lab
20 Ames St.
Cambridge, MA 02139 USA
{vemuri, decamp, walter, geek}@media.mit.edu

ABSTRACT
Despite the ready availability of digital recording technology and the continually decreasing cost of digital storage, browsing audio recordings remains a tedious task. This paper presents evidence in support of a system designed to assist with information comprehension and retrieval tasks from a large collection of recorded speech. Two techniques are employed to assist users with these tasks. First, a speech recognizer creates necessarily error-laden transcripts of the recorded speech. Second, audio playback is time-compressed using the SOLAFS technique. When used together, subjects are able to perform comprehension tasks with more speed and accuracy.

Author Keywords
Speech recognition, time-compressed audio, information retrieval.

ACM Classification Keywords
H.5.2. Information interfaces and presentation (e.g., HCI): User Interfaces, Voice I/O; I.2.7. Natural Language Processing: Speech recognition and synthesis

INTRODUCTION

Browsing, searching, and retrieving information stored in textual format has been a well-studied area for many years [12]. As multimedia collections have become more widespread, there is an increasing need to browse and search non-textual data. This paper focuses on audio.

Retrieving information from audio collections has applicability to several areas. Examples include reviewing recorded lecture material, recorded meetings, searching the web for audio recordings, and retrieving information from one's personal recordings. Furthermore, projects aiming to store one's life experiences for later analysis and retrieval have been gaining momentum [6,7,9,10]. Audio is among the proposed data types recorded by such systems and our attempt to build an audio-based personal memory aid motivates the desire to create improved audio-retrieval systems. Despite the current interest in personal data accrual, less attention has been paid to what to do with these data once collected. This paper examines two technologies in support of searching and browsing collections of audio recordings: automatic large-vocabulary speech recognition and audio time-compression, in regard to their interaction.

Audio presents unique challenges. The average speech rate of an English speaker is 180 words per minute while the reading rate is 400 words per minute [14]. This large disparity suggests that automatically transcribing audio and then accessing it as a written document would be most effective for information retrieval tasks. However, in reading a transcript, the prosodic cues, which make speech rich in meaning and subtlety, are lost. Additionally, automatic transcription of natural speech remains extremely difficult. Computer speech recognizers attempt to transform speech to the corresponding text. When generated, such transcripts almost always suffer from poor recognition accuracy, are difficult to read, can confound readers, and can waste their time [22].

Despite these shortcomings, speech recognizers have been making incremental improvements to the point where it is no longer uncommon to encounter commercial versions in daily life. However, only a limited set of applications are currently viable due to poor recognition accuracy. Successful systems tend to achieve better accuracy when limiting vocabularies, speakers, speech style, and acoustic conditions. These constraints are slowly loosening as the technology improves. Although we are still far from the panacea of high-accuracy, speaker-independent, large-vocabulary recognition systems that would enable a vast array of speech applications, the state of speech recognition is nearing the point in which a limited set of new applications would benefit from speech recognition even with the limited accuracy found in today's recognition systems.

An example of both the utility as well as limitations of speech-recognizer-generated transcripts can be seen in studies of voicemail transcription [22]. Research-lab-quality

speech recognition (which typically perform better than commercial systems) using the known identity of the caller to improve recognition accuracy was required to achieve acceptable transcripts of voicemails. Despite this effort for recognition accuracy, some recipients of the transcribed voicemails tended to commit more errors in summarizing and information extraction tasks when relying too heavily on these error-laden speech-recognizer-generated transcripts.

In a related user evaluation, subjects performed question-answer tasks using error-laden speech-recognizer-generated transcripts with simultaneous audio playback. Results suggest that higher-quality transcripts lead to a reduction in solution time, less recorded-speech played, and less time spent reading [17]. However, there was no evidence that higher-quality transcripts produced better answers. Finally, subjects tended to abandoned lower-quality transcripts more quickly.

Techniques to improve information-retrieval performance for recorded-speech collections have been studied in detail as part of the TREC Spoken Document Retrieval (SDR) track [5]. TREC SDR has ceased since researchers claim success at the task of using a large-vocabulary speech recognizer for audio broadcast-news information-retrieval tasks. In fact, [23] suggests no degradation in information retrieval tasks even with 25% word error rate (WER) and a linear performance decay as WER increases. It is important to understand that despite examples such as these, high-quality transcription (a task different from information retrieval) of recorded audio is not possible today, and remains a difficult problem.

Another approach to improving user experience and performance when browsing and searching collections of audio recordings is to reduce the time needed to listen to the audio. Audio time-compression techniques attempt to play recorded speech in less time while maintaining intelligibility. There are many approaches to this [1] and one such technique, SOLAFS, presents audio at higher rates without modifying the pitch [8]. High-rate *non*-pitch-adjusted speech is sometimes described as sounding like chipmunks because the pitch of the speaker increases as playback rate increases. Since SOLAFS time-compression maintains the pitch of the original speaker, listeners are able to comprehend speech played at higher rates compared to without time-compression [4]. Furthermore, once accustomed to time-compressed speech, people prefer it over uncompressed speech [3]. Another demonstrated success for time-compression includes the following example: when presented with audio of teaching materials, subjects who listened to a time-compressed recording twice at twice-normal rate performed better than counterparts who listed to the same recording once a normal rate [16].

Projects employing time-compression to achieve shorter audio playback times include SpeechSkimmer [2]. SpechSkimmer employed time-compression as well as other audio summarization and navigation techniques, but did not provide any corresponding visual representation of the audio. Audio Notebook [15] offered additional cues by linking the recording with marks that a listener made on paper while hearing a lecture for the first time (this technique assumes the listener was present). Additionally, Audio Notebook analyzed acoustical cues in the recorded speech to attempt to identify new topics introduced in the recording; these were used to assist with skimming the recording by, for example, playing introductory snippets rapidly as part of a search strategy.

Experiences with these projects suggest a strategy of using both transcript and time-compression together in the listening user interface, if a screen is available. Indeed, SCANMail [22] provided such a visual interface, in which audio was correlated with the text transcript, but users rarely employed time-compression of the audio playback [21].

In this study, we consider the interaction between audio time-compression and the error-laden transcripts generated by a commercially-available speech recognizer. We wish to determine whether and how effectively transcripts displayed in synchrony with time-compressed audio playback improves the utility of playback at higher and higher speeds, as measured by playback rate versus comprehension. The remainder of the paper describes experiments conducted in which subjects were tested on their ability to understand time-compressed speech combined with error-laden speech-recognizer-generated transcripts.

DESIGNING THE USER INTERFACE

The present experiment was designed to test if the combination of time-compression and speech-recognition can reduce the time it takes to listen to recorded speech without sacrificing the listener's ability to understand what was said. To conduct the experiments, a computer program was constructed that allowed playback of time-compressed audio while visually presenting an error-laden speech-recognizer-generated transcript of the same recording.

For the experiment, recordings from a series of conference talks were collected. An off-the-shelf version of IBM's ViaVoice speech-recognition software [18] was used to convert recorded speech to text. Along with each recognized word, ViaVoice reports a "phrase score," which is documented as follows: "[it] is not a confidence… it is an average acoustic score per second. The acoustic score depends on the quality of the match and the length of the speech aligned with the word." [19] To better understand the meaning of phrase score in relation to the speech recordings used in the evaluation, the recordings were hand-transcribed. These transcripts were then compared with the speech-recognizer-generated ones. Figure 1

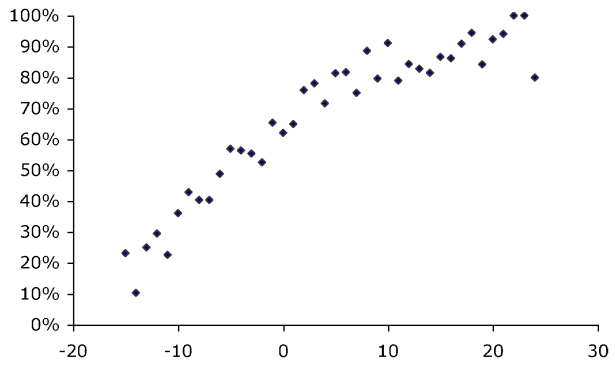

Figure 1: Percent of words recognized correctly at each recognizer-assigned "phrase score" (~2,300 words, minimum 10 words per score).

illustrates the correlation between phrase score and recognition rate (r_s=0.9385, p<0.0001).

The program plays SOLAFS time-compressed audio at arbitrary speeds while displaying a transcript of that audio. The transcript appears as 18-pt. white text over a black background. While the audio plays, the program draws a line through words that have been played and highlights the current word in green. Similar to [13], the brightness of each word in the speech-recognizer-generated transcripts are rendered proportional to its phrase score. Figure 2 shows the interface.

To test hypotheses pertaining to the subjects' comprehension of time-compressed audio with associated transcripts, five different transcript presentation styles are used:

C1: Human-constructed "perfect" transcript with uniform word brightness.

C2: Speech-recognizer-generated transcript with word brightness proportional to phrase score.

C3: Speech-recognizer-generated transcript with uniform word brightness.

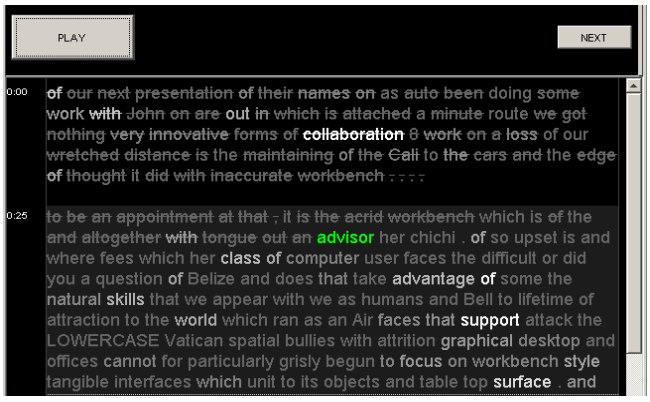

Figure 2: User interface showing brightness of individual words proportional to its phrase score.

C4: Completely incorrect transcript with uniform word brightness.

C5: No transcript. Audio only.

It should be noted that style C4 transcripts are not random words. Instead, speech-recognizer-generated transcripts from sections of audio *not* corresponding to the recording are used. Next, when style C5 is presented, the program displays a string of dots whose length is proportional to the length of the audio recording and the program shows progress of audio-playback with these dots.

Word error rate (WER) for speech-recognition systems is defined as the sum of insertion, deletion, and substitution errors divided by the number of words in the perfect transcript. For the present recordings, the speech recognizer was not trained to the speakers' voices. The speech-recognizer-generated transcripts in the present data set have WERs ranging between 16% and 67% with a mean of 42% and σ = 15%. Despite the wide range and fairly uniform distribution of sample WER, it was decided not to "adjust" transcripts to a narrower band or fixed WER since it was not clear what strategy to employ to either perturb a good transcription or to correct a bad one. Furthermore, this variability seems to be an intrinsic property of large-vocabulary speech-recognition systems.

HYPOTHESES

The experiment presented in this paper is designed to test the effectiveness of combining speech-recognizer-generated transcripts in conjunction with pitch-normalized time-compressed speech. In particular, the following hypotheses are examined:

H1. Variation in comprehension is expected when time-compressed speech is presented in conjunction with each of the different transcript styles (C1–C5). Specifically, the transcript styles, in decreasing order of expected comprehension are C1, C2, C3, C5, and C4.

H2. The comprehension of speech played in conjunction with speech-recognizer-generated transcripts is expected to be inversely proportional to the WER of that transcript.

H3. Comprehension of SOLAFS time-compressed audio is expected to be inversely proportional to the overall speech rate expressed as words per minute (WPM).

H4. Native speakers of English are expected to be able to comprehend time-compressed audio at higher speech rates compared to non-native speakers.

The comprehension of the speech is chosen as the metric to assess these hypotheses. In the study of time-compressed audio, "'comprehension' refers to the understanding of the content of the material."[1] Both objective and subjective measures are used to estimate this. First, a subject's subjective assessment of when they can understand a speaker under different transcript styles and time-compression rates is

measured. Second, a more objective question-answering task in which subjects are tested on the contents of speech under different styles and compression-factors is performed. The next section describes this in more detail.

EXPERIMENTAL SETUP

The experiment has two phases. In Phase 1, subjects are presented with three different audio samples, each taken from a single conference talk given by a single speaker. Each sample is associated with transcript style C1, C2, or C5. The order in which the samples are presented is randomized between subjects. The speech rate for all three samples averages 148 words per minute.

Subjects are presented with an interface similar to the one shown in Figure 2. When the subject presses the "PLAY" button, the transcript appears (or no transcript with style C5) and the audio begins playing at normal speed. The speed incrementally increases over time by increasing the SOLAFS time-compression factor. Subjects were instructed to press a "TOO FAST" button whenever they felt the playback speed was too fast to "generally understand" what was being said. This exact phrase was used so subjects would not stop simply because they missed an individual word, but would wait until the speech, in general, could not be understood. When the "TOO FAST" button is pressed, the time-compression factor is immediately reduced by 0.5 and then begins to slowly increase again. After the subject presses the button three times, playback is stopped. The software records the time-compression-factor every time the subject presses the "TOO FAST" button and averages the results.

One of the purposes of Phase 1 is to acclimate subjects to time-compressed audio in preparation for Phase 2. Previous studies suggest naïve listeners can understand pitch-normalized time-compressed audio up to a compression-factor of 2.0 and this ability improves with more exposure [11]. Subjects typically completed Phase 1 in 10–15 minutes, which is far short of the 8–10 hours prescribed by [11].

For Phase 2, subjects are presented with a series of 38 short clips of recorded speech and were tested on their understanding of those clips. To quantify subject comprehension, fill-in-the-blank style questions are asked. This provided a more objective metric compared to the self-reported comprehension assessment of the subjects in Phase 1.

The clips, when played at normal speed, have a mean duration of 20.6 seconds with $\sigma = 5.8$. Longer clips were avoided in order to minimize primacy and recency effects. As mentioned earlier, the clips were collected from a series of conference talks spanning a wide range of speakers; speakers who enunciated clearly and whose recording-quality was good were preferred. The content of the talks is mostly academic research and computer technology. The specific audio samples were selected such that there was little to no domain-specific language, jargon, and no prior knowledge was needed to understand them.

The 38 clips were presented in random order and with a random transcript style among C1 to C5. Each sample was played at a fixed time-compression-factor. Audio playback speed is expressed as a time-compression factor. For example, audio played at compression 2.0 will complete in half the time of the original recording, a factor of 3.0 will complete in one-third time, etc. The first three samples were presented at factor 1.0 (i.e. original speed), the next three samples at 1.5, and in sequentially increasing factors, four samples at 1.75, 2.0, 2.25, 2.5, 2.75, 3.0, 3.25 and 3.5. Figure 3 shows an example distribution of the 38 sample/transcript-style pairs that might be given to a subject. Samples were presented in increasing compression-factors in order to minimize effects related to the subjects' limited exposure to time-compressed audio.

	1.0	1.5	1.75	2.0	2.25	2.5	2.75	3.0	3.25	3.5
C1	x	x		x	x	x	x		x	x
C2		x	x		x	x	x	x		x
C3		x	x	x		x	x	x	x	
C4	x		x	x	x		x	x	x	x
C5	x		x	x	x	x		x	x	x

Figure 3: Example of sample distribution for a single subject .

Phase 2 was limited to 38 samples since pilot studies suggested subjects could complete a set of this size within our desired subject time-commitment limit. Fewer questions were assigned to the 1.0 and 1.5 compression-factors primarily due to previous results suggesting naïve listeners can understand time-compressed speech up to factor 2.0 [11].

The interface seen in Figure 2 was used. When the subject presses the "PLAY" button, the transcript appears (or a string of dots if transcript style C5) and the audio begins playing. When the sample finishes playing, the transcript disappears and is replaced by one to three questions about that sample. The questions ask simple, unambiguous facts about what the speaker said and do not require any interpretation or specialized knowledge. Each subject is given two practice samples and corresponding questions before the test.

Speech-rate variation among speakers suggests that time-compression factors should be normalized by a more standard speech-rate metric: words per minute (WPM). Specifically, when played at their original speeds, the audio samples in the present collection were spoken between 120 to 230 WPM with a mean of 174 and $\sigma = 29$.

RESULTS

Two out of 34 subjects who participated stated they had previous exposure to time-compressed audio similar to SOLAFS. Four others said they had experience with high-speed audio, but cited examples were limited to the fast-forward feature of an analog audio-tape player, fast speech in television commercials, and some videos airing on the "MTV" television channel. Eleven subjects stated they had

previous experience with speech-recognition technology, seven said they had a little experience, and 15 subjects correctly recognized the identity of at least one speaker among the recorded speakers.

Phase 1 examined subjects' self-reported maximum time-compression factor for three of the transcript styles. Figure 4 shows the average maximum time-compression-factor for each transcript style. Using a repeated measures, one-way ANOVA, the mean time-compression factors for all Phase 1 transcript styles were found to be different and, more precisely, C1 > C2 > C5 ($p<0.01$ for each relation). This suggests that, using the Phase 1 subjective comprehension metric, part of Hypothesis H1—which posits differences in subject comprehension among transcription styles—is confirmed.

Seven of the 34 subjects were non-native speakers of English. Across all Phase 1 transcript styles, non-native speakers averaged a maximum compression-factor of 2.47 while native speakers achieved 2.88. This difference was found to be significant ($p=0.015$) and confirms the subjective aspect of native versus non-native comprehension difference (Hypothesis H4).

In the Phase 2 question-answering task, answers were judged to be correct if they indicated a subject's basic understanding of the sample's content, and incorrect otherwise. Table 1 shows a summary of the aggregate data for all subjects in Phase 2. The scores indicate the percentage of questions answered correctly. At compression-factors 1.0 and 1.5 each cell represents 20 to 21 data points. At all higher compression-factors, each cell represents 27 or 28 data points. These numbers do not apply to the totals, which contain the aggregate data of an entire row, column, or, in the case of the lowest-rightmost box, all 1290 data points. For samples that had more than one question, only the question subjects attempted to answer the most is included in the data. The average WPM for each cell in Table 1 was computed (after accounting for rate increases due to time-compression) and Figure 5 shows subjects' question-answering accuracy for each transcript style when normalized by WPM.

Adjusting for speech-rate increases due to time-compression, the range for all samples actually played to all subjects during Phase 2 was 120 to 810 WPM. Figure 6 shows the fraction of all questions answered correctly across all transcript styles at each WPM decile. Significant correlation was found between WPM and subjects' question-answering accuracy ($r=-0.429$, $p<0.0001$). Hence, Hypothesis H3, which posits degradation of subject comprehension with increasing speech rate, is confirmed.

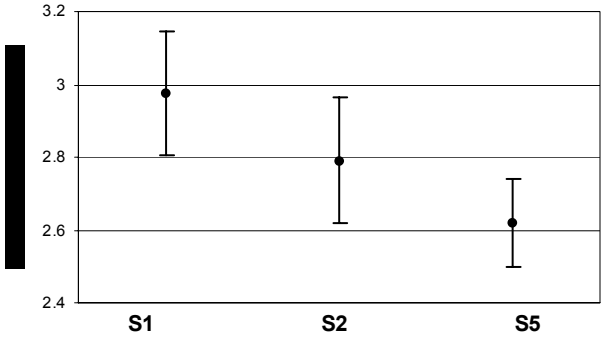

Figure 4: Phase 1 subject self-reported maximum time-compression-factors for transcript styles C1, C2, and C5 with 95% confidence intervals.

	C1	C2	C3	C4	C5	Total
1.0	80	85	75	90	85	83
1.5	95	90	67	81	90	84
1.75	89	79	74	81	67	78
2.0	78	78	78	74	61	73
2.25	61	81	59	59	59	64
2.5	67	59	63	46	63	60
2.75	63	55	54	37	41	50
3.0	70	54	48	15	18	41
3.25	48	22	26	11	4	22
3.5	36	26	37	7	7	23
Total	68	62	57	48	47	56

Table 1: Percentage of questions answered correctly at each style for each time-compression factor in Phase 2.

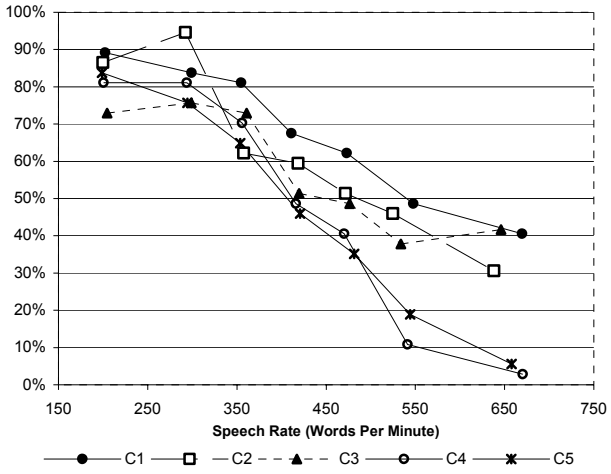

Figure 5: Subjects' question-answering accuracy at varying speech rates.

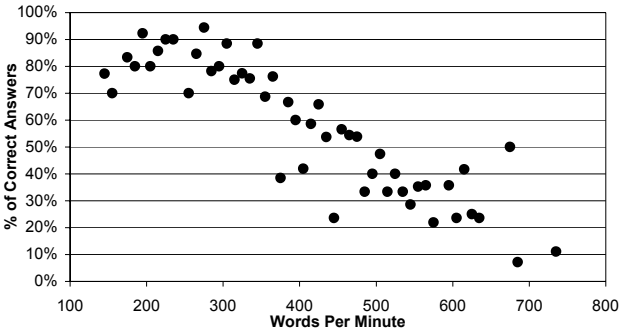

Figure 6: Percentage of questions answered correctly at each decile of words per minute (minimum 10 samples per decile).

	C1	C2	C3	C4	C5
C1	-	ns	< 0.01	< 0.001	< 0.001
C2		-	ns	< 0.001	< 0.001
C3			-	< 0.05	< 0.05
C4				-	ns
C5					-

Table 2: p-values associated with each pairwise comparison between transcript styles for Phase 2 question-answering task.

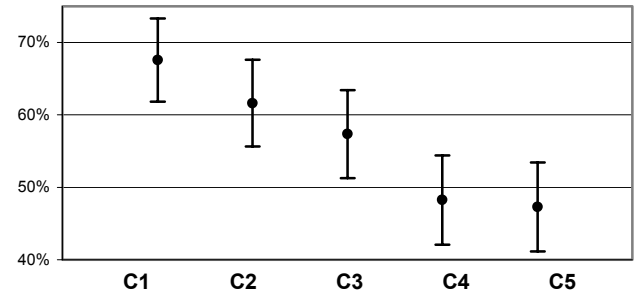

Figure 7: Percentage of questions answered correctly for each transcript style averaging across all time-compression factors.

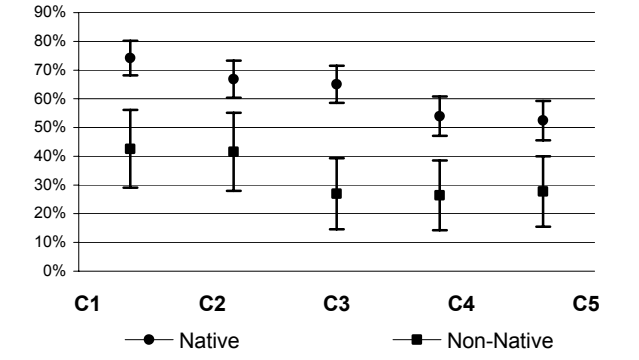

Figure 8: Comparison of question-answering performance for native versus non-native English speakers with 95% confidence intervals.

Using the data from the Phase 2, question-answering task, a two-way ANOVA was conducted in which transcript style and WPM were used as independent variables and percentage of questions answered correctly was used as the dependent variable. Both transcript style and WPM showed significant variation ($p<0.0001$ for both), while the interaction between them did not ($p=0.373$). A one-way ANOVA was conducted using just transcript style as the independent variable and percentage of questions answered correctly as the dependent variable ($p<0.0001$). The data for this test was paired by subject. Each subject had five measures corresponding to the average number of correctly answered questions under a given transcript style. The questions for each measure were not perfectly distributed by speed and some questions were more difficult to answer than others. However, normalizing the data for speed and difficulty had a negligible effect on the overall results and such normalization has been left out of this analysis. Table 2 shows the p-values obtained by comparing the comprehension scores under each transcript style with a Student-Newman-Keuls post test. Tukey and Bonferroni tests did not find a significant difference between C3/C4 and C3/C5, but otherwise yielded similar results. Figure 7 displays the means and 95% confidence intervals for the percentage of questions answered correctly under each of the transcripts styles.

While these results do not confirm every aspect of Hypothesis H1, they do confirm several subcomponents. Specifically, subject comprehension of audio presented with a perfect transcript (C1) was found to be better than C3, C4, and C5 and the comprehension of C2 and C3 was found to be better than C4 and C5. No significant difference was found between completely wrong transcripts (C4) and no transcript (C5). C4 scored 0.96% higher than C5, which translates to about three questions out of 258.

In order to evaluate hypothesis H2, correlation tests were performed comparing the WER of a given audio sample to the percentage of times subjects answered the associated question correctly across all speeds. As previously mentioned, the WER distribution across all samples was fairly uniform. Correlations were found for transcript styles C2 ($r=-0.44$, $p=0.01$) and C3 ($r=-0.34$, $p=0.04$). To ensure there were no effects related to the quality of the recordings, correlation tests were performed with styles C1 (perfect transcript, $r=-0.06$, $p=0.73$) and C5 (no transcript, $r=-0.04$, $p=0.81$). Surprisingly, a correlation was found with the C4 transcript style (wrong transcript, $r=-0.39$, $p=0.02$).

Finally, with respect to the differences between native and non-native English speakers (Hypothesis H4) in the Phase 2 question-answering task, Figure 8 shows the percentage of questions answered correctly by each group under each transcript style with $p<0.005$ for all native versus non-native comparisons at each style. These results suggest confirmation of Hypothesis H4.

	Phase 1: Subjective	Phase 2: Objective
H1	Confirmed for transcript styles C1, C2, and C5	Partially confirmed
H2	Not tested	Confirmed for C2, C3 and C4
H3	Not tested	Confirmed
H4	Confirmed	Confirmed

Table 3: Summary of hypothesis testing

Table 3 summarizes hypothesis testing for both the Phase 1 subjective tests and Phase 2 question-answering tests.

DISCUSSION

The perfect transcript style (C1) is tantamount to reading and, not-surprisingly, results from both Phase 1 (self-reported maximum) and Phase 2 (question-answering task) suggest this style is the best supplement to improving comprehension of speech playback. However, generating such transcripts is costly, time-consuming, and must be done manually. Using a computer speech-recognizer to generate lower-quality transcripts, like C2 and C3, can be done cheaply, quickly, and in an automated fashion.

To date, the poor transcript quality of large-vocabulary, speaker-independent recognizers has hindered more wide-scale adoption of this technology. Despite this shortcoming, the present experiment provides evidence suggesting comprehension improvements when using speech-recognizer-generated transcripts, even when laden with errors, and especially when rendered in the C2 transcript style. Specifically, comprehension of transcript style C2 was found to be better than both audio alone (C5) and a completely wrong transcript (C4). Differences between Style C2 and C3 were not confirmed to be significant, so it is not yet clear how much Style C2's confidence-based text-brightness-rendering contributed to this, if at all.

In a worst-case scenario, a speech-recognizer may generate a completely-incorrect transcript (C4). Part of Hypothesis H1 posits speech presented in conjunction with a style C4 transcript is expected to reduce comprehension compared to no transcript (C5). The supposition is that a transcript with many errors will tend to distract subjects and result in fewer correct answers. However, Phase 2 results could not confirm any significant difference between styles C4 and C5. Consequently, no evidence was found suggesting a completely-wrong transcript would worsen comprehension compared to audio only. One possible explanation is that subjects ignored bad transcripts. Similar to the low-quality transcript abandonment results found in [17], some subjects in the present experiment stated that they would read a transcript for a few seconds, and elect whether or not to continue reading it based on its quality. In fact, several subjects looked away from the computer display and stated they did so to avoid the distraction of a faulty transcript.

Unexpectedly, the difference between the perfect transcript style (C1) and the brightness-coded speech-recognition style (C2) was not found to be significant in the Phase 2 objective question-answering task (though a significant difference was found in the Phase 1 subjective task). In Phase 2, a significant difference was found between C1 and the uniform-brightness speech-recognition style (C3). While it is premature to conclude that style C2 is better than C3, the evidence suggests there is some utility to visualizing text in this manner, but further investigation is needed to understand the role of brightness-coded text.

Hypothesis H1 posits comprehension variation among all transcript styles. While some aspects of this were confirmed (as detailed in Table 2), the trend suggests some of the un-confirmed H1 parts (specifically, C1 vs. C2 and C2 vs. C3) may achieve statistically significant variation with additional subjects.

Hypothesis H2 posits that comprehension of audio presented with a transcript will increase as the WER of the transcript decreases. This correlation was observed with the two transcript styles that had variable WER, C2 and C3. Surprisingly, comprehension of audio played with the completely wrong transcript style (C4) was correlated to the WER of the corresponding speech-recognizer-generated transcript of that audio. This non-intuitive result cannot be explained. The fact that style C1 (perfect transcript) and style C5 (no transcript) showed no correlation with WER suggest audio quality across samples was even. Results for this hypothesis remain inconclusive and more work is needed to understand the nature of the relationship between WER and comprehension.

Evidence for Hypothesis H3, which posits that comprehension decreases with increasing speech rate, was clearer and in agreement with [4]. Differences between native and non-native speakers (Hypothesis H4) were also found.

Collectively, these results paint an optimistic picture. Despite the fact that comprehension of time-compressed speech decreases as compression-factors increase [4], speech-recognizer-generated transcripts used in conjunction with such speech improve comprehension. In effect, the results suggest people can either save time or improve their understanding when reading error-laden speech-recognizer-generated transcripts in synchrony with time-compressed speech. The cost to provide speech-recognizer-generated transcripts is low and since very bad transcripts do not seem to confuse users, there is no apparent downside.

Searching large collections of personal audio recordings intended as a personal memory aid is of particular interest. The authors are cautiously optimistic that, based on the results described herein, users of an audio-based memory aid will better utilize archives of their recorded past.

CONCLUSIONS

Results of an experiment comparing comprehension of time-compressed speech presented in synchrony with transcripts

of varying qualities and presentation styles were presented. Motivating this experiment is the desire to construct improved audio browsing and searching tools by minimizing the time needed to review time-compressed audio and improving the comprehension of audio presented in this manner. Results suggest a speech-recognizer-generated transcript, despite having errors, aids in improving comprehension of time-compressed speech. Similarly, comprehension can be maintained at slightly higher time-compression factors (i.e., faster) when accompanied with a speech-recognizer-generated transcript.

No evidence was found suggesting a completely-wrong transcript has different comprehension compared to no transcript. Consequently, it does not seem harmful to provide a poor speech-recognizer-generated transcript. Rendering transcripts with word-brightness proportional to the speech-recognizer-assigned "phrase score" improved comprehension compared to no transcript, but it is not clear how important the brightness-rendering contributes to this improvement.

In confirmation of previous studies [4], comprehension decreased with speech rate. This study presents evidence that comprehension also decreases with increasing word error rate of speech recognizer-generated transcripts, although this hypothesis is not yet confirmed.

The authors are optimistic on the use of these methods to improve browsing and searching of personal recordings intended as part of an audio-based personal memory aid.

ACKNOWLEDGMENTS

The authors wish to thank Brad Lassey for his contribution, Chalapathi Neti and Edward Epstein for their guidance on the speech recognition system, and all of our experimental subjects. This work was sponsored in part by the "information:organized" research consortium at the MIT Media Lab.

REFERENCES

1. Arons, B. Techniques, Perception, and Applications of Time-Compressed Speech. *Proc. 1992 Conference, American Voice I/O Society*, 169–177. (1992)
2. Arons, B. SpeechSkimmer: Interactively Skimming Recorded Speech. *Proc. UIST 1993*, 187–196. (November 1993).
3. Beasley, D.S. and Maki, J.E. Time- and Frequency-Altered Speech. In N.J. Lass, editor, *Contemporary Issues in Experimental Phonetics*, Academic Press, 419–458. (1976).
4. Foulke, W., and Sticht, T.G. Review of research on the intelligibility and comprehension of accelerated speech. *Psychological Bulletin*, **72**, 50–62, (1969).
5. Garofolo, J., Auzanne, C., and Voorhees, E. The TREC Spoken Document Retrieval Track: A Success Story. *Proc. TREC 8*. 107–130. (1999).
6. Gemmell, J., Bell, G., Lueder, R., Drucker, S., and Wong, C., MyLifeBits: Fulfilling the Memex Vision, *Proc. ACM Multimedia '02*, Juan-les-Pins, France, 235–238. (December 2002).
7. Gerasimov, V., *Every Sign of Life*, MIT Ph.D. Thesis, Media Arts & Sciences. (January 2003).
8. Henja, D. and Musicus, B.R., The SOLAFS Time-Scale Modification Algorithm, BBN Technical Report, (July 1991).
9. LifeLog, http://www.darpa.mil/baa/baa03-30.htm
10. Lin, W., Hauptmann, A.G. A Wearable Digital Library of Personal Conversations. In *JCDL 2002*: 277–278 (2002).
11. Orr, D.B., Friedman, H.L., and Williams, J.C., Trainability of Listening comprehension of Speeded Discourse." *Journal of Educational Psychology*, **56**, 148–156 (1965).
12. Salton, G., *Automatic Text Processing: the transfomation, analysis, and retrieval of information by computer*. Addison-Wesley (1989).
13. Schmandt, C. The Intelligent Ear: An Interface to Digital Audio. *Proc. IEEE International on Cybernetics and Society*, IEEE, Atlanta, GA (1981).
14. Schmandt, C. *Voice Communication with Computers: Conversational Systems*. Van Nostrand Reinhold, New York. (1994).
15. Stifelman, L., Arons, B., and Schmandt, C. The audio notebook: paper and pen interaction with structured speech. *Proc. CHI 2001*, 182–189, (2001).
16. Sticht, T.G. Comprehension of repeated time-compression recordings. *The Journal of Experimental Education*, **37**(4), (Summer 1969)
17. Stark, L., Whittaker, S., and Hirschberg, J. ASR satisficing: the effects of ASR accuracy on speech retrieval. *Proc. International Conference on Spoken Language Processing*. (2000).
18. ViaVoice, http://www-3.ibm.com/software/speech/
19. ViaVoice, Frequently Asked Questions http://www.wizzardsoftware.com/voice/voicetools/dictationforsdkfaq.htm#What%20is%20Phrase%20Score.
20. Wactlar H.D., Hauptman A.G., and Witbrock M.J., Informedia News-On Demand: Using Speech Recognition to Create a Digital Video Library. Tech Report. CMU-CS-98-109 (March 1998).
21. Whittaker, S., pers. comm.. (December 2003).
22. Whittaker, S., Hirschberg, J., Amento, B., Stark, L., Bacchiani, M., Isenhour, P., Stead, L., Zamchick G., & Rosenberg, A. SCANMail: a voicemail interface that makes speech browsable, readable and searchable. *Proc. CHI 2002*. 275–82 (2002).
23. Witbrock, M., http://infonortics.com/searchengines/boston1999/witbrock/index.htm, Lycos (1999).