COMPREHENSION OF SPEECH PRESENTED AT SYNTHETICALLY ACCELERATED RATES: EVALUATING TRAINING AND PRACTICE EFFECTS

Christina Wasylyshyn, Brian McClimens, and Derek Brock

U.S. Naval Research Laboratory
Washington, DC 20375
christina.wasylyshyn@nrl.navy.mil

ABSTRACT

The ability to monitor multiple sources of concurrent auditory information is an integral component of Navy watchstanding operations. However, this leads to attentionally demanding environments. The present study tested the utility of a potential solution to listening to multiple speech communications in an auditory display environment: presenting speech serially at synthetically accelerated rates. Comprehension performance of short auditory narratives was compared at seven accelerated speech rates. Practice effects and training effects were examined. An optimum acceleration rate for comprehension performance was determined, and training was found to be an effective method when synthetic speech was presented at slow to moderately accelerated rates.

1. INTRODUCTION

It is expected that future Naval forces will be defined by their agility and their capacity for coping with high-stakes, uncertain environments [1]. The individual Naval watchstander might be responsible for the concurrent monitoring of numerous radio communications channels, along with actively monitoring and responding to events on multiple visual displays. Such attentionally demanding environments have motivated various HCI solutions to help warfighters deal with the vast number of information sources needing to be monitored in order to perform their duties successfully.

However, a critical consideration when designing potential solutions in auditory display research is to take into account the limitations in listeners' abilities to attend to multiple competing communications channels. For example, communications performance has been shown to decline significantly as watchstanders were asked to handle additional radio circuits [2]. Likewise, comprehension in multi-talker speech displays decreased when listening to concurrent speakers [3]. Both of these studies [2], [3] examined performance in conditions with up to four concurrent speakers. These results were replicated and extended by comparing comprehension performance in four different conditions: two concurrent speakers, four concurrent speakers, four serial speakers with normal speech rates, and four serial speakers with accelerated speech rates (accelerated 75% faster) [4]. Major findings included better comprehension performance in the accelerated speech condition compared to the concurrent speech conditions (for both two and four concurrent speakers). Participants performed better in the normal speech rate condition than the accelerated speech rate condition, suggesting that while listening to accelerated speech is more difficult than normal speech, there is a significant improvement in one's ability to make sentence comprehension judgments when listening to accelerated speech compared to listening to concurrent speech.

Presenting listeners with messages that are serialized and accelerated, therefore, may be one potential solution to concurrent monitoring of communications channels. However, building an effective system of synthetically accelerated voice communications will require new information paradigms that directly address the strengths and limitations of human operators. This paper reports on a work in progress that examines listeners' abilities to adapt to synthetically accelerated speech in an auditory display environment through the use of practice and training.

Unlike reading, in which the information input rate can be controlled by one's eye movements, comprehension of speech is often dependent on a transient acoustic signal whose information input rate is largely controlled by the talker, not the listener. The information input rate, thus, is determined by the environment, and previous information is often not reviewable. In order to comprehend auditory information effectively, input must be analyzed, segmented, and processed for structure and meaning, all of which must occur even as new auditory information continues to arrive. When auditory input is rapid, listeners will have even less time to carry out these integrative processes, and successful comprehension will require greater effort at accelerated rates of speech.

Accelerated speech is marked by an increased word rate, so that more information can be transmitted per unit of time. Even when the pitch and prosody of the original speech is preserved, loss of information occurs, most notably from the loss of processing time that the listener would typically use to integrate the auditory information [5]. However, with practice, subjects have shown increased recall of information that was presented at an accelerated rate [6], and mere exposure to accelerated speech has been shown to generalize to increased speech comprehension of other accelerated speech, even if the subject is exposed to accelerated speech in a foreign language with similar phonemes [7].

While these studies suggest that there are detectable performance differences in accelerated speech comprehension,

it isn't clear to what extent training participants to listen to synthetically accelerated speech will be helpful. The amount of practice and/or exposure to accelerated speech needed to produce benefits in comprehensibility differs across studies, as does the ability to distinguish practice effects from training effects.

The present study tested the utility of listening to synthetically accelerated speech by comparing comprehension performance of information presented at seven accelerated speech rates. Speech rates were blocked so that three short narratives were presented at each rate. This approach allowed for the testing of practice effects across the three narratives for each speech rate. Furthermore, the blocked narratives were either presented at incrementally faster rates (the "training" group) or in a random manner (the "random" group). This allowed us to determine whether presenting accelerated speech in a systematic way from slow to fast speech rates was beneficial to comprehension performance or whether participants were only able to integrate accelerated auditory information up to a certain speed-related processing threshold. The study sought to answer three primary questions:

1. Do practice effects occur, i.e., is there systematic improvement in comprehension performance after listening to multiple auditory excerpts presented at the same accelerated speech rate?

2.. Does comprehension performance vary by presentation method, i.e., can listening to accelerated speech be trained or do participants simply have a natural threshold for listening to and processing accelerated speech content?

3. What is the optimum acceleration rate for comprehension performance, i.e., what is the fastest rate at which speech can be presented so that comprehension performance does not differ from comprehension performance of speech presented at a normal rate?

2. METHOD

2.1 Participants

Twenty NRL employees participated (11 males, 9 females). All participants were native English speakers and claimed to have normal (i.e., non-corrected) hearing. Participants were randomly assigned to the "training" (5 males, 5 females, mean age = 40.9, SD = 11.1) or to the "random" (6 males, 4 females, mean age = 39.6, SD = 10.8) presentation group. There are no significant differences in participant characteristics between groups.

All participants were presented with a baseline listening comprehension task, that is, all participants listened to two narratives at a normal speech rate and completed comprehension questions. The training group (mean = 0.76, SD = 0.11) performed equally well as the random group (mean = 0.78, SD = 0.10), $t(18) = -0.34$, $p = .54$, so that there were no differences between the groups for baseline comprehension performance.

2.2 Task and Apparatus

The main battery was composed of brief auditory narratives and comprehension questions [8]. Each narrative described an event in a person's life. These narratives were approximately 300 words in length (range = 298 – 308 words); they were equated for number of ideational propositions and content difficulty. Each narrative was recorded in a female voice at a speaking rate of approximately 180 words/min (normal speaking rates are anywhere between 130-200 words/min).

After listening to each narrative, participants were asked to evaluate statements about ideas represented (or not) in the narrative. These consisted of 24 statements that included both main ideas and specific details about the narrative. Three different types of statements were included for comprehension evaluation:

1. *True* statements represented ideas that were included in the narrative
2. *False* statements represented ideas that were inconsistent with those told in the narrative
3. *Distractor* statements represented ideas that were consistent with the narrative, but were not actually part of it.

The narratives were synthetically accelerated at rates ranging in 15% increments from 50% to 140% faster-than-normal. The "training" presentation group listened to the accelerated narratives at incrementally faster rates from 50% to 140% faster-than-normal. The "random" presentation group listened to the narratives at accelerated speeds presented in a random fashion. For both presentation groups, the narratives were presented in triads at each speed to test for practice effects within speeds. For example, a participant in the "training" group would have heard three narratives at 50% faster-than-normal, followed by three narratives at 65% faster-than-normal, followed by three narratives at 80% faster-than-normal, and so on, up to 140% faster-than-normal.

In order to create the accelerated test battery, the narratives that were first recorded at a normal speaking rate were subjected to a patented NRL speech-rate compression algorithm [9], known as "pitch synchronous segmentation" (PSS). PSS retains the fundamental frequency of speech signals and preserves a high degree of intelligibility. This high degree of intelligibility remains because the PSS method does not try to generate an electric analog of the human speech production mechanism. Instead, PSS represents the speech waveform by individual pitch cycle waveforms. The output speech sounds more natural because it is constructed from raw speech and because pitch interference is absent in the speech representation.

The visual part of the study was displayed on a large flat-panel monitor and the auditory component was rendered binanrally in Sony MDR-600 headphones. Brief auditory examples of what participants heard at each accelerated speech rate are given in the following sound files:

Speed50%	[SPEED50.WAV]
Speed65%	[SPEED65.WAV]
Speed80%	[SPEED80.WAV]
Speed95%	[SPEED95.WAV]
Speed110%	[SPEED110.WAV]
Speed125%	[SPEED125.WAV]
Speed140%	[SPEED140.WAV]

2.2 Procedure

Participants were randomly assigned to either the "training" or the "random' presentation group. After providing informed consent, participants completed a short practice exercise that resembled the format of the experimental task and a baseline comprehension measure. Immediately, after listening to each narrative, participants were visually presented with 24 statements (8 true, 8 false, and 8 distractor) and asked to evaluate whether or not the statement identified ideas heard in the narrative.

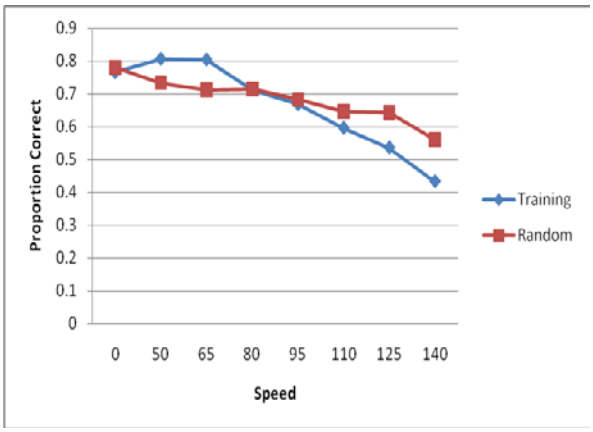

Figure 1: Mean proportion of correctly identified comprehension statements by speed for each presentation method group (training and random).

3. RESULTS

The proportion of correctly identified comprehension statements served as the dependent measure. The dependent measure was submitted to an 8(speed: normal, 50%, 65%, 80%, 95%, 110%, 125%, 140%) x 3 (practice: narrative 1, 2, 3 within each triad) x 2 (presentation method: training, random) mixed ANOVA. Speed and practice were repeated-measures variables, and presentation method was a between-groups variable. The main effect of speed was significant, $F(7, 126) = 19.88$, $p < .0001$. Regardless of presentation method and practice, participants correctly identified more comprehension statements when narratives were presented at the slower speeds (i.e., normal, 50%, and 65%). There was no main effect of practice, $F(2, 36) = 0.63$, $p = 0.54$. Across presentation method and speed, the mean comprehension scores for narratives presented first in each triad was 0.68, second in each triad was 0.67, and third in each triad was 0.67. There was also no main effect of presentation method, $F(1, 18) = 0.13$, $p = 0.72$. Across speed and practice, the average comprehension score in the training presentation group was 0.67 and the average comprehension score in the random presentation group was 0.68.

However, the speed by presentation method interaction was significant, $F(7, 126) = 3.01$, $p = 0.006$. Figure 1 displays the mean proportion of correctly identified comprehension statements by speed (across practice) for each presentation method. Note that the values marked Speed 0 indicate each group's average comprehension score at baseline (i.e., normal speed speech). As can be seen in Figure 1, both the training and the random presentation groups performed equally well at Speed 0. Planned contrasts indicated that the optimum acceleration rate for comprehension performance was 65% faster-than-normal, that is, 65% faster-than-normal was the synthetic speech rate at which comprehension performance did not differ from comprehension performance of speech presented at a normal rate. Planned contrasts also indicated that participants in the training group correctly identified more of the comprehension statements at Speeds 50 and 65 (mean proportion correct = 0.81 and 0.80, respectively) compared to the participants in the random group (mean proportion correct = 0.73 and 0.71 for Speeds 50 and 65, respectively). This suggests that training participants may be an effective method, but only at slower speeds. Figure 1 seems to suggest that the random presentation method is more effective at higher speeds (e.g., Speeds 125 and 140) than the training presentation method, however, there were no significant differences between the presentation groups at the higher speeds. The way in which the narratives were presented to the training group (i.e., at incrementally faster rates) may have induced fatigue over the course of the experimental session, further supporting the notion that training may only be effective at slower speeds.

In summary, participants seemed to adapt quickly to comprehending the synthetically accelerated speech. Training was effective at slower accelerated speech rates, however, systematic training to higher accelerated speech rates led to fatigue. Practice (i.e., performance across the three narratives within each speech rate) did not seem to aid comprehension performance.

4. DISCUSSION

This present study reports results from a work in progress that examines listeners' abilities to adapt to synthetically accelerated speech in an auditory display environment through the use of practice and training. Previous research conducted at NRL [4] demonstrated that comprehension performance can benefit from accelerated and serialized audio communications channels, compared to comprehension performance when listening to concurrent speech on two and/or four channels. However, participants in this previous study did not perform as well when listening to synthetically accelerated speech rates at 75% faster-than-normal as when listening to normal speech rates [4]. The present study extends those previous results. We tested a larger scale of synthetic speech rates ranging in 15% increments from 50% to 140% faster-than-normal. We found that the optimum acceleration rate for comprehension performance was 65% faster-than-normal. This was the fastest rate at which synthetically accelerated speech could be presented where comprehension performance did not differ from comprehension performance of speech presented at a normal rate.

The main analysis compared participants who listened to the narratives at incrementally faster rates from 50% to 140% (the training group) to those participants who listened to the narratives at speeds presented in a random fashion (the random group). As expected, comprehension performance declined as

speech rate increased. At faster synthetic speech rates, participants were not able to integrate the structure and meaning of the narratives as well as they were able to at slower speech rates. The training presentation method was found to be effective for comprehension performance compared to the random presentation method, but only at the slower synthetic speech rates (i.e., 50% and 65% faster-than-normal). What was not expected, however, was how quickly listeners adapted to the synthetically accelerated speech. This can be seen by the lack of practice effects within speeds; on average, participants tended to perform equally across the three narratives of each triad.

That being said, it should also be noted that the highest comprehension accuracies were between 78-81%. Participants were particularly good at distinguishing the true and false statements, but performed significantly worse when presented with the distractor statements. Again, distractor statements represented ideas that were consistent with the narrative, but were not actually part of it. Determining ways to improve listeners' abilities to distinguish between distracting and true information is especially relevant to building effective systems of synthetically accelerated voice communications that can be used in attentionally demanding environments.

The current results may have future applications for coordinating the numerous communications between various disaster relief organizations and municipal services, for managing air traffic control centers, and for organizing communications in Naval combat information centers. Once we know the limits of human operators' abilities to listen to synthetically accelerated speech, we can begin to design auditory display environments that capitalize upon strengths and minimize weaknesses. The present study addresses two critical areas of concern: the trainability of listening to synthetically accelerated speech and the optimum acceleration rate for comprehension performance. Future research seeks to enhance the auditory display environment by presenting information in a way that approximates how listeners more naturally perceive it, that is, by employing auditory cues to specify communications channels that are rendered in a virtual listening space.

5. ACKNOWLEDGMENT

This work was supported by the Office of Naval Research.

6. REFERENCES

[1] V. Clark, "Sea power 21 series – part I: Projecting decisive joint capabilities," *Naval Institute Proceedings Magazine*, 2002.

[2] D. Wallace, C. Schlicting, and U. Goff, *Report on the Communications Research Initiatives in Support of Integrated Command Environment (ICE) Systems*, Naval Surface Warfare Center, Dahlgren Division, TR-02/30, Jan. 2002.

[3] D. S. Brungart, M. A. Ericson, and B. D. Simpson, "Design considerations for improving the effectiveness of multitalker speech displays," *Proceedings of the 2002 International Conference on Auditory Display*, Kyoto, Japan, 2002.

[4] D. Brock, B. McClimens, G. Trafton, M. McCurry, and D. Perzanowski, "Evaluating listeners' attention to and comprehension of spatialized concurrent and serial talkers at normal and a synthetically faster rate of speech," *Proceedings of the 14th International Conference on Auditory Display*, Paris, France, 2008.

[5] A. Wingfield, P. A. Tun, C. K. Koh, and M. J. Rosen, "Regaining lost time: Adult aging and the effect of time restoration on recall of time-compressed speech," *Psychology and Aging*, vol. 14, no 3, pp. 380-389, 1999.

[6] G. T. M. Altmann, and D. Young, "Factors affecting adaptation to time-compressed speech," *Eurospeech 9*, Berlin, Germany.

[7] N. Sebastian-Galles, E. Dupoux, A. Costa, and J. Mehler, "Adaptation to time-compressed speech: Phonological determinants," *Perception & Psychophysics*, vol. 62, pp. 834-842, 2000.

[8] R. A. Dixon, and C. M. de Frias, "The Victoria Longitudinal Study: From characterizing cognitive aging to illustrating changes in memory compensation," *Aging Neuropsychology, and Cognition*, vol. 11, pp. 346-376, 2004.

[9] G. S. Kang, and L. J. Fransen, *Speech Analysis and Synthesis Based on Pitch-Synchronous Segmentation of the Speech Waveform*, Naval Research Laboratory, TR-9743, Nov. 1994.