Proceedings of the 57th Hawaii International Conference on System Sciences | 2024

# Influence of Audio Speech Rate and Source Text Difficulty on Health Information Comprehension and Retention

Arif Ahmed
Management Information Systems
University of Arizona
arifahmed@arizona.edu

Gondy Leroy
Management Information Systems
University of Arizona
gondyleroy@arizona.edu

Philip Harber
Public Health
University of Arizona
pharber@arizona.edu

Sumi Lee
East Asian Studies
University of Arizona
sumilee@arizona.edu

David Kauchak
Computer Science
Pomona College
david.kauchak@pomona.edu

Stephen A. Rains
Communication
University of Arizona
srains@email.arizona.edu

Prosanta Barai
Management Information Systems
University of Arizona
Prosantabarai@arizona.edu

## Abstract

*Health literacy is crucial for patients to make informed healthcare decisions. Although text has historically been the main form of health information dissemination, people rely increasingly on audio-delivered information, e.g., through smart speakers. In this study, we evaluate the effects of audio speech rate and source text difficulty on audio information comprehension and retention. We created audio snippets from easy and difficult text and conducted a study on Amazon Mechanical Turk (AMT). Audio speech rate and source text difficulty are the independent variables and perceived difficulty (measured with a Likert scale) and comprehension and retention (measured with AI-generated multiple-choice questions and free recall of information) are the dependent variables. Audio created from difficult source text was perceived as more difficult and comprehension was also lower than for audio from easy text. Speech rate also influenced information comprehension and retention of information: a higher speech rate (+60% faster audio speech rate) lowered the comprehension of health information by 38% compared to a moderate speech rate.*

**Keywords:** Health Literacy, Cognitive Processing, Audio Delivery, Text Difficulty, Audio Speech Rate, User Study, AI As a Tool.

## 1. Introduction

Effective communication using clear and understandable language is crucial in healthcare to promote health literacy. In the U.S., improving health literacy is a major national goal due to the significant costs associated with poor health literacy [1]. Limited comprehension of healthcare information can lead to poor decision-making and increased healthcare costs for patients, as even a small percentage increase in costs at the system level can result in thousands of dollars of additional costs at the patient level [1]. While text has been the primary medium for delivering healthcare information for decades due to its cost-effectiveness and efficiency, audio, interactive videos, and other multimedia tools are emerging as alternatives. Audio information delivery has gained popularity among patients in recent years, and the use of virtual assistants and smart speakers for health-related queries is increasing. By 2022, it was estimated that nearly 94.9 million smart speakers (e.g., Siri, Alexa) were used in the U.S., with an annual adoption rate of 30-40% [2, 3]. Patients can receive health information through audio formats such as via a smart

URI: https://hdl.handle.net/10125/106814
978-0-9981331-7-1

speaker or virtual assistant, provided the information is delivered in a clear and easy-to-understand manner [4]. Smart speakers are also being incorporated into hospital systems, and patients are using them to ask questions to clinicians and communicate with them [5]. Among all questions asked to a smart speaker in 2019, 16% were health-related [6]. Voice assistant use by American adults for healthcare increased dramatically from 19 million in 2019 to 51.3 million in 2020 and 54.4 million in 2021 [7]. Incorporating audio into existing health information delivery guidelines could be a valuable opportunity for improving health literacy.

Although recent efforts to simplify text have focused on syntactic and semantic analysis to identify textual features that reduce text difficulty [8], more research is needed to understand how text and audio characteristics impact the difficulty of health information. In this paper, we focus on audio speech rate and source text difficulty and their impact on the perceived and actual difficulty of health information when delivered using audio.

## 2. Background

### 2.1 LC4MP

Speech rate has been recognized as one of the most significant aspects that may affect the understanding of information [9]. The Limited Capacity Model of Motivated Mediated Message Processing (LC4MP) [10] explains how speech rate can affect comprehension. LC4MP explains the mediated message and human comprehension of information that focuses on the recall and attention processes [11]. It describes information processing with three subprocesses: encoding, storage, and retrieval. Receiving and encoding a message involves selecting relevant information from the stimulus and forming a mental representation. If this information is retained in working memory, it is stored for later retrieval. Thorough processing occurs when the data is encoded, stored, and retrievable in subsequent processes.

The extent of optimal processing relies on the allocation of resources to each subprocess, which can either be automatic or controlled [12]. Automatic resources are considered unconscious and require minimal attention, while controlled resources are conscious and require more attention [11]. The number of available resources an individual needs to process a message depends on the information density and the stimuli's structural or formal complexity [12]. Certain structural elements of the message can automatically allocate resources to speed up information processing. For instance, speech rate may influence the allocation of automatic resources and impact comprehension. A fast speech rate may grab the listener's attention but could also impede understanding by affecting the phonological loop [13]. The phonological loop is a working memory model's constituent that explicitly processes auditory information by temporarily holding the verbal information [14].

LC4MP proposes that the structural characteristics of a message exert a significant influence on the constrained capacity of the human cognitive system. These attributes prompt situating responses and thus lead to the automatic allocation of resources for message encoding [15]. Extensive research has been conducted for information delivered on the Internet (incorporating emotional images, animations etc.), or television (including sudden movements, camera adjustments, etc.), and radio (involving shifts in voice, sound effects, onsets of music, etc.). However, these inventories of stimuli remain incomplete [10] and the theory has not yet been tested for audio and processing involving alterations in prosodic elements fundamental to audio such as changes in speech rate. We hypothesized that listeners will be able to comprehend health information better at a moderate speech rate than a faster speech rate because according to the LC4MP model human cognitive system has limited capacity on the attention and recall process of mediated information. [10]

### 2.2 Audio speech rate

Most research about audio rate can be found in media psychology and advertising research. For example, in media psychology, studies of various forms of media evaluate how the design and content of mediated messages influence human information processing [16]. One such study using five levels of audio speech rate showed that a speech rate between 170 to 190 words per minute (wpm) generates the highest level of recognition for information. The recognition of information deteriorated with rates over 210 wpm [16].

Research in radio advertising indicated that a faster speech rate holds greater advantages compared to a slower one. In the context of time-compressed advertisements, a study recommended that advertisers should strive for a speech rate approximately 30 percent faster than regular speech, which amounts to roughly 160 words per minute. This implies that announcers should aim for a moderate pace of around 160 words per minute during regular speech, and when employing compression techniques, the final speech rate should not exceed 180 wpm [17].

**Table 1.** Text features (T-test, * = p < .05, ** = p < .01, ** = p < .001)

| Variables (Avg) | Easy source text (N=30) | Difficult source text (N=30) |
|---|---|---|
| **Total characters *** ** | 1368.20 | 1537.23 |
| **Word counts ** ** | 218.33 | 217.50 |
| **Sentence length * ** | 20.02 | 22.56 |
| **Percentage Nouns *** ** | 30.17 | 35.93 |
| **Verb's percentage *** ** | 17.47 | 13.07 |
| **Adverb's percentage** | 4.10 | 3.27 |
| **Adjective's percentage *** ** | 10.40 | 13.93 |
| **Function word percentage *** ** | 37.87 | 33.80 |
| **Google word frequency *** ** | 368871654 | 236026454 |
| **Number of Lexical Chains * ** | 11.57 | 13.57 |
| **Chain Length * ** | 3.29 | 3.03 |
| **Chain Span** | 102.24 | 109.68 |
| **Number of Cross Chains * ** | 11.57 | 13.47 |
| **Number of Half Document Length Chains** | 4.33 | 5.10 |

Multiple studies have demonstrated the impact of speech rate on both information recall and recognition[18, 19]. When the rate is increased, listeners require additional cognitive resources to process and encode the information. As a result, fewer resources remain available for information storage, potentially leading to a negative effect on the subsequent recognition of the information [20]. A marketing study found that as speech rate of 180 wpm achieved the highest level of comprehension and retention [21]. Researchers at the University of California found that students can comprehend information with rates up to 2x the speed of the regular rate. After 2x speed, comprehension starts to decline [22].

## 2.3 Source text and audio information difficulty

A text is difficult if it is not easily understandable to readers [23]. The analysis of text difficulty encompasses various quantitative and qualitative measures. Quantitatively, factors like word length, word frequency, sentence length and text cohesion are important. Qualitatively, aspects such as language structure, language conventions, levels of meaning, clarity, and the reader's knowledge need also be taken into account when assessing text difficulty [24]. For example, difficult texts have a higher percentage of nouns, a lower percentage of verbs, a lower percentage of function words, and a low Google word frequency. In contrast, the simpler texts have a higher percentage of verbs, a lower percentage of nouns, and high Google word frequency [25]. The number of topics and their distribution through a text can also be utilized to distinguish difficult and easy texts [26].

Even though there are ways to evaluate the sound or audio quality, there is currently no metric available for measuring the difficulty of information delivered over audio [27]. In most cases, audio is produced by capturing spoken words and saving them as an audio file. However, with new audio delivery methods, generating audio automatically is becoming an interesting new option. Platforms such as Amazon Web Services (AWS) and Microsoft Azure offer tools to adjust speech rate, choose between male or female voices, apply various accents, and incorporate pauses and emphasis into the generated audio. We focus here on speech rate.

## 2.4 Hypotheses

To our knowledge, we are the first to study the effects of source text difficulty and audio speech rate on audio-delivered health information. We have generated the audio snippets by leveraging automated audio generation using the collected text snippets.

We evaluate perceived difficulty using a Likert scale. We evaluate actual difficulty by measuring comprehension, using multiple choice and true-false questions, and retention, using free recall of information. We hypothesize that:

**H1:** Audio health information delivered using a moderate speech rate will be perceived as less difficult than when using increased speech rate.

**H2:** Audio health information delivered using a moderate speech rate will be result in better information comprehension and retention than when using increased speech rate.

**H3**: Audio health information generated from difficult source text will be perceived as more difficult than when generated from easy source text.

**H4:** Audio health information generated from difficult source text will result in lower information comprehension and retention than when generated from easy source text.

**Table 2. Worker Characteristics**

| | **Difficult Source Text** | | **Easy Source Text** | | **Total** |
|---|---|---|---|---|---|
| | **Default Speech Rate** | **Increased Speech Rate** | **Default Speech Rate** | **Increased Speech Rate** | |
| **Characteristic** | N (%) | N (%) | N (%) | N (%) | N (%) |
| **Count** | 21 | 14 | 35 | 14 | 84 |
| **Sex** | | | | | |
| Male | 9 (42.85) | 8 (57.14) | 15 (42.85) | 8 (57.14) | 40 (47.61) |
| Female | 12 (57.14) | 6 (42.85) | 20 (57.14) | 6 (42.85) | 44 (52.38) |
| Other | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) |
| **Age** | | | | | |
| Younger than 30 years old | 4 (19.04) | 6(42.85) | 6(17.14) | 9(64.28) | 25 (29.76) |
| 31 to 40 years old | 8 (38.09) | 2(14.28) | 11(31.42) | 1(7.14) | 22 (26.19) |
| 41 to 50 years old | 5 (23.8) | 3(21.42) | 11(31.42) | 2(14.28) | 21 (25) |
| 51 to 60 years old | 4 (19.04) | 2(14.28) | 6(17.14) | 1(7.14) | 13 (15.47) |
| 61 to 70 years old | 0 (0) | 1 (7.14) | 1 (2.85) | 1 (7.14) | 3 (3.57) |
| **Race** | | | | | |
| Asian | 2 (9.52) | 0 (0) | 2 (10.52) | 0 (0) | 4 (4.76) |
| American Indian/ Native Alaskan | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) |
| Black or African American | 2 (9.52) | 1 (5.26) | 4 (21.05) | 1 (5.26) | 8 (9.52) |
| Native Hawaiian or other Pacific Islander | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) |
| White | 16 (76.19) | 13 (68.42) | 28 (147.36) | 13 (68.42) | 70 (83.33) |
| Asian & White | 1 (4.76) | 0 (0) | 1 (5.26) | 0 (0) | 2 (2.38) |
| **Ethnicity** | | | | | |
| Hispanic or Latino | | 2 (10.28) | 3 (8.57) | 1 (7.14) | 6 (7.14) |
| Not Hispanic or Latino | | 12 (85.71) | 32 (91.42) | 13 (92.85) | 57 (67.85) |
| **Education** | | | | | |
| Less Than High School | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) |
| High School | 5 (23.8) | 3 (21.42) | 9 (25.71) | 1 (7.14) | 18 (21.42) |
| Associate's degree | 0 (0) | 0 (0) | 5 (14.28) | 0 (0) | 5 (5.95) |
| Bachelor's degree | 8 (38.09) | 7 (50.00) | 17 (48.57) | 11 (78.57) | 43 (51.19) |
| Master's Degree | 7 (33.33) | 4 (28.57) | 4 (11.42) | 2 (14.28) | 17 (20.23) |
| Doctorate Degree | 1 (4.76) | 0 (0) | 0 (0) | 0 (0) | 1 (1.19) |
| Other Professional Degree | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) |
| **English Speaking** | | | | | |
| Never English | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) |
| Rarely English | 1 (4.76) | 0 (0) | 0 (0) | 0 (0) | 1 (1.19) |
| Half English | 1 (4.76) | 0 (0) | 0 (0) | 0 (0) | 1 (1.19) |
| Mostly English | 2 (9.52) | 0 (0) | 3 (8.57) | 2 (14.28) | 7 (8.33) |
| Only English | 17 (80.95) | 14 (100.00) | 32 (91.42) | 12 (85.71) | 75 (89.28) |

# 3. Methods

## 3.1 Study design

We designed a 2x2 experiment using default and increased speech rate and easy and
difficult source text.

The dependent variables are perceived difficulty measured with a 5-point Likert scale, and actual difficulty measured with multiple-choice (MC), true-false (TF) and free-recall questions. Amazon Mechanical Turk (AMT) workers were recruited as study participants.

## 3.2 Audio creation

We collected health-related source text on various diseases from multiple websites and health-related journals based on the International Classification of Diseases-10 (ICD-10) codes disease list. ICD-10 is a medical coding system primarily developed by the World Health Organization (WHO). Its purpose is to categorize health conditions into groups of similar diseases, with specific conditions listed within those categories. This system helps in the mapping of detailed diseases to more general morbidities, enabling comprehensive cataloging of health conditions [28].

**Table 3. Averages of audio features per condition**

| | | Word Count | Audio Length (S) | Word Per Minute (WPM) |
|---|---|---|---|---|
| **Difficult Source Text** | **Default Rate** | 217.5 | 119.1 | 110.83 |
| | **Increased Rate** | 217.5 | 67.37 | 195.54 |
| **Easy Source Text** | **Default Rate** | 218.3 | 85.9 | 152.96 |
| | **Increased Rate** | 218.3 | 54.3 | 241.87 |

We selected diseases from the ICD-10 disease list and collected texts for those selected diseases from various health-related websites and journals. Then we have chosen 60 snippets from this set representing easy and difficult text (30 each) based on their origin and by verifying they their difficulty level using our easy and difficult text characteristics that we discovered in earlier work [25]. In table 1 we report the texts' features.

The snippets were 200-250 words long. Difficult texts were selected from Rheumatic Disease journal abstract (6 texts), Wikipedia (5 texts), PubMed abstract (16 texts), and Medscape (3 texts). Easy texts were selected from the Rheumatic Disease journal lay summary (15 texts) and Cochrane Plain Language Summary (15texts). Overall, the two groups differ significantly for metrics that are related to difficult. For example, the difficult texts contained a higher percentage of nouns, the words were less common (less frequent). From table 1 we can see that the percentage of noun is 35% for difficult texts and 30% for easy text and percentage of verb is 13% for the difficult texts and 17% for the easy texts.

We used Microsoft Azure's text-to-speech[1] service to generate the audio snippets. We used a US male voice and two audio rates: default and increased. When we generated the audio snippets using Azure's default rate, we have found that Azure's default audio produced audio in a range of 91 to 135 wpm for difficult texts and a range of 138 to 177 wpm for easy texts. For the increased audio rate, we used +60% of the default rate. For increased audio rate the produced audio had a range of 163 to 230 wpm for difficult texts and a range of 225 to 242 wpm for the easy texts. Table 3 contains the averages for each condition.

## 3.3 Dependent variables

To measure perceived difficulty of the audio information, we asked participants to evaluate difficulty using 5-point Likert scale labeled from very easy (1) to very difficult (5).

To measure comprehension of the information, we created multiple-choice and true-false questions using two AI systems: questgen.ai[2] and chatGPT[3]. For each AI model, we generated two multiple-choice and two true-false questions for a total of four questions for each of the 60 texts. We manually evaluated each question and answer to verify they focused on the content and contained appropriate multiple-choice and true-false questions.

To measure retention of information, the participants were requested to recall as much information as possible. To analyze the free recall, we use two notions of overlap with the original text: the percentage of exactly matching words and the

[1] https://azure.microsoft.com/en-us/products/cognitive-services/text-to-speech/

[2] https://www.questgen.ai/

[3] https://chat.openai.com/

percentage of words that are similar based on word embeddings. The latter allows for a more flexible and semantic notion of recall.

### 3.4 Recruitment of participants

We recruited study participant using AMT. AMT workers first completed demographic information questions. One Human Intelligence Task (HIT) consisted of an audio snippet followed by multiple questions and the request for free recall. The workers listen to the audio snippet and then answered the perceived difficulty question, the four multiple-choice questions, and the four true-false questions and a free recall question about the audio information they heard.

Each of the 60 texts were revaluated by at least three workers. Data for the four experimental conditions were collected separately in a one-week interval to avoid recall of information from previous participation. Each worker received $1.00 for a completed HIT. Although workers were allowed to complete multiple HITs, no worker evaluated a given text more than one time.

## 4. Results

### 4.1 Data cleaning

We cleaned the data to remove data from participants that did not participate appropriately. We eliminated answers of workers for inattentive responses in the free recall question. Those inattentive workers used audio transcription software to capture the audio information and used that for their retention response. We also checked the average time each worker took to complete a HIT. If the completion time of a HIT is less than the audio time length indicating that they didn't even listen to the entire audio clip, we removed that data. Following this process, we removed 52 total responses generated by nine workers. The final sample included in the analyses consisted of 308 total unique evaluations of the texts in our corpus. The demographic information on the remaining workers is included in Table 2.

### 4.2 Worker Characteristics

The majority of the workers were white (83%) and not Hispanic or Latino.[4] Slightly more than half of the workers were female (52%). The workers were mostly between 31-40 (26%) and between 41-50 (25%) years old. The highest level of education for the majority of the workers was a bachelor's degree (51%), followed by a high school degree (21%). The majority (89%) of the workers speaks only English at home. (see Table 2)

### 4.3 ANOVA

In this study, we performed a two-way ANOVA to determine the statistical significance of different conditions. However, we first verified key data assumptions before performing the ANOVA. For instance, the response should be independent and normally distributed for each group with equal variance [29, 30]. In our study, workers were randomly selected for each group (handled by AMT). We also plotted the Q-Q plot (see Figure 1 and Figure 2) for dependent variables, which resembles a straight line indicating the approximately normal distribution of the response variables [30]. Moreover, we also checked for non-constant variance across groups using the residual vs. fitted value graph, which did not find any pattern [30].

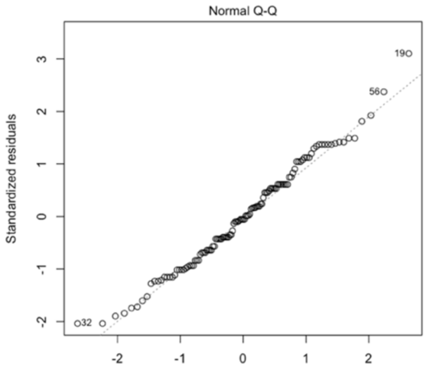

**Figure 1. Normal Q-Q plot for perceived difficulty responses.**

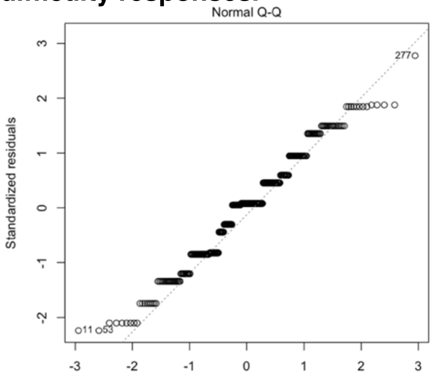

**Figure 2. Normal Q-Q plot for actual difficulty responses.**

[4] Due to a technical problem during data collection, we do not have ethnicity data for condition 1

### 4.4 Perceived difficulty

To analyze the results, we used a two-way ANOVA with perceived difficulty (1-5) as the dependent variable, and speech rate (default vs increased) and source text difficulty (easy vs difficult) as independent variables. There was a significant main effect of speech rate (F (1,304) =52.53, p<.001). We found that audio health information delivered over moderate rate was perceived as more difficult than the increased rate for both easy and difficult source text. (See table 4) That result does not support our hypothesis H1 that audio health information delivered using a moderate speech rate will be perceived as less difficult than increased speech rate.

**Table 4. Results of perceived difficulty (A lower number indicates easier text)**

| | Difficult Source Text | | | Easy Source Text | | |
|---|---|---|---|---|---|---|
| | Default Rate (StD. Dev.) | Increased Rate (StD. Dev.) | Both Rates (StD. Dev.) | Default Rate (StD. Dev.) | Increased Rate (StD. Dev.) | Both Rates (StD. Dev.) |
| **Perceived Difficulty** | 3.49 (1.12) | 3.34 (1.3) | 3.42 (1.21) | 2.94 (1.17) | 1.91 (0.78) | 2.49 (1.14) |

The results also showed that there was a significant main effect of source text difficulty (F (1,304) =22.20, p<.001) (see Figure 3). Difficult source text was perceived as more difficult than easy source text. (See table 4) That supports H3 that audio health information generated from difficult source text is perceived as more difficult than easy source text.

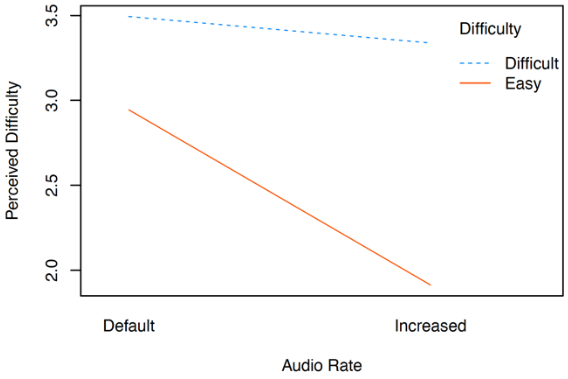

**Figure 3. Perceived difficulty**

The 2-way interaction between speech rate and text difficulty (F (1,304) =11.60, p<.001) was also significant. For easier source text and default speech rate the perceived difficulty score was lower than the difficult source text and default speech rate. And, for easier source text and increased speech rate the perceived difficulty score was lower than the difficult source text and increased speech rate. (see Table 4)

### 4.5 Actual difficulty: comprehension and retention

#### 4.5.1 Comprehension

To measure the actual difficulty or information comprehension, we first analyzed the Multiple-Choice (MC) and true-false (TF) responses. We conducted a two-way ANOVA with accuracy of MC and TF as the dependent variable and audio speech rate (default vs increased), and source text difficulty (easy vs difficult) as independent variables.

For the MC responses, there was a significant main effect of rate (F (1,931) = 7.17, p<.01) but no main effect for text difficulty (F (1,931) =0.20, p>.05). There was a significant interaction between rate and difficulty as well. For easy texts, rate had a significant effect on response accuracy (F (1,464) =42.43, p<.001), in contrast to difficult text where rate had no effect (F (1,467) =0.09, p>.05). (See figure 4) For default speech rate and difficult source text result showed 42% accuracy for MC responses and for increased rate and difficult source text the accuracy drops to 37%. For default speech rate and easy source text result showed 56% accuracy for MC responses and for increased rate and easy source text the accuracy drops to 29%. (see Table 5)

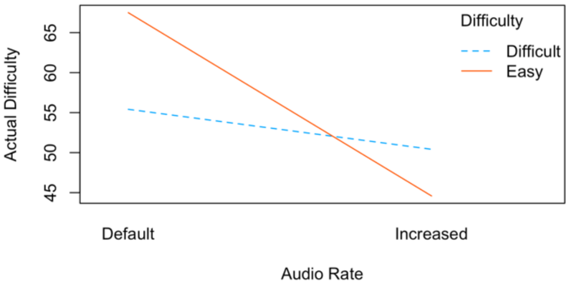

**Figure 4. Actual difficulty (MC responses)**

For the TF responses, there was a significant main effect of rate (F (1,931) =22.977, p<.01) but no main effect for text difficulty (F (1,931) =0.54, p>.05). Effect of text difficulty was only present in the default speech rate conditions (F (1,238) =11.78, p<.001), but not the increased speech rate condition. For default speech rate and difficult source text result showed 57% accuracy for TF responses and for increased rate and difficult source text the accuracy drops to 43%. For default speech rate and easy source text result showed 79% accuracy for TF responses and for increased rate and easy source text the accuracy drops to 39%. (see Table 5)

### 4.5.2 Retention

Retention was measured by asking subjects to recall the information from the audio. The free recall rate of four conditions were analyzed by a 2-way ANOVA with the percentage of similar words and percentage of matching words as dependent variables, and speech rate (default vs increased) and text difficulty (easy vs difficult) as independent variables.

**Table 5. Results of actual difficulty**

| | **Difficult Source Text** | | | **Easy Source Text** | | |
|---|---|---|---|---|---|---|
| **Comprehension** | **Default Rate (StD. Dev.)** | **Increased Rate (StD. Dev.)** | **Both Rate (StD. Dev.)** | **Default Rate (StD. Dev.)** | **Increased Rate (StD. Dev.)** | **Both Rate (StD. Dev.)** |
| **Multiple Choice Accuracy (%)** | 42.26 (39.82) | 37.2 (40.12) | 39.73 (40.16) | 56.02 (41.60) | 28.98 (41.10) | 42.73 (42.50) |
| **True False Accuracy (%)** | 56.98 (36.26) | 43.3 (38.54) | 50.14 (37.81) | 79.37 (27.18) | 38.88 (45.76) | 59.48 (40.30) |

There was a significant main effect of rate for percentage of similar word (F (1,304) =32.248, p<.001) but no main effect for source text difficulty (F (1,304) =0.388, p>.05). There was no significant interaction effect as well for the percentage of similar word (F (1,304) =0.193, p>.05). We have found highest percentage of similar word 21% for condition easy source text and increased Rate and lowest 11% for difficult source text and default rate. (see Table 6)

**Table 6. Results of retention**

| | **Difficult Source Text** | | | **Easy Source Text** | | |
|---|---|---|---|---|---|---|
| **Retention** | **Default Rate (StD. Dev.)** | **Increased Rate (StD. Dev.)** | **Both Rates (StD. Dev.)** | **Default Rate (StD. Dev.)** | **Increased Rate (StD. Dev.)** | **Both Rates (StD. Dev.)** |
| **Free recall (matching words) (%)** | 7.79 (7.04) | 16.8 (17.13) | 12.01 (13.45) | 7.74 (7.84) | 14.29 (21.64) | 10.42 (15.89) |
| **Free recall (similar words) (%)** | 10.9 (8.2) | 19.53 (18.33) | 14.94 (14.42) | 13.11 (9.93) | 21.07 (20.96) | 16.00 (16.46) |

For percentage of matching words, the main effect of rate was significant (F (1,304) =28.66, p<.001). But there was no significant main effect found for text difficulty (F (1,304) =0.971, p>.05). There was no significant interaction effect for the percentage of matching word (F (1,304) =0.106, p>.05). We have a higher free recall of matching word with the increased rate. The highest result for free recall (matching word) was 17% for the condition difficult text and increased rate, and the lowest 8% is for easy text and default rate. (see Table 6)

There was no interaction between speech rate and text difficulty for both free recall measures. This indicates that text difficulty did not affect the free recall rate of provided texts in comparison to speech rate which showed to significantly affect both difficult and easy texts.

From the results of comprehension and retention we can see that our hypothesis H2 is supported for comprehension but not for retention that audio health information delivered using a moderate speech rate will be result in better information comprehension and retention than health information delivered using increased speech rate.

In addition, when we consider the source text difficulty the percentage of similar word is higher in easy source text than difficult source text and the percentage of matching word is lower in easy source text than difficult source text. So, our hypothesis four is supported for comprehension but not for retention that audio health information generated from difficult source text results in a lower information comprehension and retention than audio generated from easy source text.

## 5. Conclusion and Discussion

We examined the effect of source text difficulty and audio speech rate on the perceived difficulty as well as the actual information comprehension and retention of health information. Perceived difficulty differed between easy and difficult source text: easier texts were also perceived as easier. We also found that information comprehension is higher in the default audio rate than in the increased audio rate. The findings support the LC4MP model. Cognitive processes struggle to manage information encoding, retrieval, and storage at increased speech rates than they do at moderate speed rate. Overall, information comprehension is also higher for easy source text than source difficult text. However, when we consider the audio rate only, information retention is higher with an increased audio rate than the default audio rate. The increased speech rate grabs listener's attention and this may be why retention is higher in increased rate than default rate. This effect may have a short duration and longer-term effects may differ.

Our study has several limitations. The first limitation is that the Azure text-to-speech tool generates audio snippets with variable speech rates even though we have selected the default and increased (+60%) rate. As the audio were created from the texts, the text's features, such as word length and syllables, might influence generated audio speech rate. In further work, we will look whether these variances influenced perceived and actual difficulty outcomes.

Since we do not know what type of environment and audio listening devices the workers used during working on the HITs, we may assume that they were in appropriate condition. Furthermore, AMT workers are representative of the general public, but findings may differ when patients with a vested interest listen to the audio.

## Acknowledgment

This research reported in this paper was supported by the National Library of Medicine (NLM) of the National Institutes of Health (NIH) under Award Number 2R01LM011975. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

## References

[1] Eichler, K., S. Wieser, and U. Brügger, *The costs of limited health literacy: a systematic review.* International Journal of Public Health, 2009. **54**: p. 313-324.

[2] B., K. *Global Smart Speaker Growth Cools in Q1 as Pandemic Leads to Declining China Sales, Amazon Retains Top Spot Says Strategy Analytics.* 2020 [cited 2023 04-29-2023]; Available from: https://voicebot.ai/2020/05/25/global-smart-speaker-growth-cools-in-q1-as-pandemic-leads-to-declining-china-sales-amazon-retains-top-spot-says-strategy-analytics/2020.

[3] Laricchia, F. *US: Smart speaker installed base 2018-2022*. 2022 [cited 2023 Apr 29]; Available from: https://www.statista.com/statistics/967402/united-states-smart-speakers-in-households/#:~:text=The%20installed%20base%20of%20smart,nearly%2091%20million%20in%202021.

[4] Leroy, G. and D. Kauchak, *A comparison of text versus audio for information comprehension with future uses for smart speakers.* Journal of the American Medical Informatics Association open, 2019. **2**(2): p. 254-260.

[5] Leibler, S. *Cedars-Sinai Taps Alexa for Smart Hospital Room Pilot*. 2019 [cited 2023 April 29]; Available from: https://www.cedars-sinai.org/newsroom/cedars-sinai-taps-alexa-for-smart-hospital-room-pilot/2019.

[6] Yoo, T.K., et al., *Deep learning-based smart speaker to confirm surgical sites for cataract surgeries: A pilot study.* Public Library of Science One, 2020. **15**(4): p. e0231322.

[7] Modev Staff Writers. *Voice Tech in Healthcare: Transformation and Growth*. 2022 6 June 2023]; Available from: https://www.modev.com/blog/voice-tech-in-healthcare-transformation-and-growth#:~:text=The%20report%20tells%20us%20that,2019%20to%2021%25%20in%202021.

[8] Sulem, E., O. Abend, and A. Rappoport, *Semantic structural evaluation for text simplification.* arXiv preprint arXiv:1810.05022, 2018.

[9] Schelten-Cornish, S., *The significance of speaking rate in speech treatment.* Die Sprachheilarbeit, 2007. **4**: p. 136-145.

[10] Lang, A., *The limited capacity model of mediated message processing.* Journal of Communication, 2000. **50**(1): p. 46-70.

[11] Lang, A., *Using the limited capacity model of motivated mediated message processing to design effective cancer communication messages.* Journal of Communication, 2006. **56**: p. S57-S80.

[12] Fox, J.R., B. Park, and A. Lang, *When available resources become negative resources: The effects of cognitive overload on memory sensitivity and*

*criterion bias.* Communication Research, 2007. **34**(3): p. 277-296.

[13] Baddeley, A., *Short-term Memory. Baddeley, AD, Eysenck, MW, and Michael C. Anderson, MC Memory.* Psy. Press. New York. Baddeley, A., Gathercole, S., & Papagno, C.(1998). The phonological loop as a language learning device. Psychological Review, 2015. **105**: p. 158-173.

[14] Fiez, J.A., *Chapter 68 - Neural Basis of Phonological Short-Term Memory*, in *Neurobiology of Language*, G. Hickok and S.L. Small, Editors. 2016, Academic Press: San Diego. p. 855-862.

[15] Potter, R. F., Lang, A., & Bolls, P. D. (2008). Identifying structural features of audio: Orienting responses during radio messages and their impact on recognition. *Journal of Media Psychology*, *20*(4), 168-177.

[16] Rodero, E., *Influence of Speech Rate and Information Density on Recognition: The Moderate Dynamic Mechanism.* Media Psychology, 2016. **19**: p. 224–242.

[17] LaBarbera, P. and J. MacLachlan, *Time-compressed speech in radio advertising.* Journal of Marketing, 1979. **43**(1): p. 30-36.

[18] Hudson, R.F., H.B. Lane, and P.C. Pullen, *Reading fluency assessment and instruction: What, why, and how?* The Reading Teacher, 2005. **58**(8): p. 702-714.

[19] Megehee, C.M., K. Dobie, and J. Grant, *Time Versus Pause Manipulation in communications directed to the young adult population: does it matter?* Journal of Advertising Research, 2003. **43**(3): p. 281-292.

[20] Schlinger, M.J.R., et al., *Effects of time compression on attitudes and information processing.* Journal of Marketing, 1983. **47**(1): p. 79-85.

[21] Rodero, E., *Do Your Ads Talk Too Fast To Your Audio Audience?: How Speech Rates of Audio Commercials Influence Cognitive and Physiological Outcomes.* Journal of Advertising Research, 2020. **60**(3): p. 337-349.

[22] Murphy, D.H., et al., Learning in double time: The effect of lecture video speed on immediate and delayed comprehension. Applied Cognitive Psychology, 2021. 36(1): p. 69-82.

[23] Fulcher, G., *Text difficulty and accessibility: Reading formulae and expert judgement.* System, 1997. **25**(4): p. 497-513.

[24] Davidson, M.M., *Reading comprehension in school-age children with autism spectrum disorder: Examining the many components that may contribute.* Language, Speech, and Hearing Services in Schools, 2021. **52**(1): p. 181-196.

[25] Kauchak, D., et al. *Text simplification tools: Using machine learning to discover features that identify difficult text*. in *2014 47th Hawaii international conference on system sciences*. 2014. IEEE.

[26] Mukherjee, P., G. Leroy, and D. Kauchak, *Using lexical chains to identify text difficulty: a corpus statistics and classification study.* IEEE journal of biomedical and health informatics, 2018. **23**(5): p. 2164-2173.

[27] Sun, W., et al. *A deep learning based no-reference quality assessment model for ugc videos*. in *Proceedings of the 30th ACM International Conference on Multimedia*. 2022.

[28] Brämer, G. R. (1988). International statistical classification of diseases and related health problems. Tenth revision. World health statistics quarterly. Rapport trimestriel de statistiques sanitaires mondiales, 41(1), 32-36.

[29] St, L. and Wold, S., 1989. Analysis of variance (ANOVA). *Chemometrics and Intelligent Laboratory Systems*, *6*(4), pp.259-272.

[30] Montgomery, Douglas C. *Design and analysis of experiments*. John Wiley & Sons, 2017.