Expanding a Large Inclusive Study of Human Listening Rates

DANIELLE BRAGG, Microsoft Research, USA
KATHARINA REINECKE and RICHARD E. LADNER, University of Washington, USA

12

As conversational agents and digital assistants become increasingly pervasive, understanding their synthetic speech becomes increasingly important. Simultaneously, speech synthesis is becoming more sophisticated and manipulable, providing the opportunity to optimize speech rate to save users time. However, little is known about people's abilities to understand fast speech. In this work, we provide an extension of the first large-scale study on human listening rates, enlarging the prior study run with 453 participants to 1,409 participants and adding new analyses on this larger group. Run on LabintheWild, it used volunteer participants, was screen reader accessible, and measured listening rate by accuracy at answering questions spoken by a screen reader at various rates. Our results show that people who are visually impaired, who often rely on audio cues and access text aurally, generally have higher listening rates than sighted people. The findings also suggest a need to expand the range of rates available on personal devices. These results demonstrate the potential for users to learn to listen to faster rates, expanding the possibilities for human-conversational agent interaction.

CCS Concepts: • **Human-centered computing** → **Auditory feedback**; **Empirical studies in accessibility;**

Additional Key Words and Phrases: Synthetic speech, listening rate, human abilities, accessibility, blind, low-vision, visually impaired, crowdsourcing

ACM Reference format:
Danielle Bragg, Katharina Reinecke, and Richard E. Ladner. 2021. Expanding a Large Inclusive Study of Human Listening Rates. *ACM Trans. Access. Comput.* 14, 3, Article 12 (July 2021), 26 pages.
https://doi.org/10.1145/3461700

1 INTRODUCTION

Conversational agents and digital assistants are only beginning to integrate into our lives. Designed to save people time and aggravation by answering questions and accomplishing tasks, they are typically voice-activated and return information through synthetic speech. With advances in natural language processing and big data, conversational agents will only become more powerful, useful, and pervasive. For example, recent studies have explored conversational agents in health care [22] and education [28]. Despite popular focus on the artificial intelligence powering these

We thank Daniel Snitkovskiy for development work on the study, Cynthia Bennett for collaboration on the initial study, and note that NSF grants IIS-1651487, IIS-1702751, and CNS-1539179 supported our work.
Authors' addresses: D. Bragg, Microsoft Research, 1 Memorial Drive, Cambridge, MA; email: danielle.bragg@microsoft.com; K. Reinecke and R. E. Ladner, University of Washington, 185 E. Stevens Way NE, Seattle, WA; emails: {reinecke, ladner}@cs.washington.edu.

1936-7228/2021/07-ART12 $15.00
https://doi.org/10.1145/3461700

agents, the opportunity to optimize speaking rate to maximize efficiency has largely been ignored. We argue that creating conversational agents that maximize saved time requires understanding the intelligibility of fast, synthetic speech.

Optimizing the speaking rate of conversational agents and text-to-speech software can save time for a growing group of users. Conversational agents are transforming the way we receive information, replacing text to be read with spoken words. Given the large amount of material people read, even a small increase in reading rate can amount to many hours of saved time over a lifetime. Consequently, people invest in learning to read faster, enrolling in speed-reading courses and practicing reading faster. As we receive more information aurally, optimizing speech rate becomes similarly valuable.

A better understanding of people's listening abilities could also support enriched interactions with conversational agents. Today's agents typically use a fixed rate of speech, which could instead dynamically adapt to the user, content, and surroundings. Consider that a person reading has dynamic control over the rate at which they receive information. A conversational agent that understands the user's listening abilities could provide similarly efficient delivery, slowing down and speeding up as needed. The agent could even adapt to context, perhaps slowing down in noisy environments. Alternatively, an agent could allow users to manipulate the speed directly. However, to for this strategy to be effective, such agents must provide an appropriate range of speeds from which to choose, which is currently not well understood.

While synthetic speech is new to many people using conversational agents, people who are blind or have low-vision have a long history of accessing text with audio. The National Library Service has been recording and distributing recorded books to visually impaired Americans since the 1930s [52], long before audio books became mainstream. Text-to-speech software is used to access other text aurally, and screen readers, which read interface content, help navigate computerized devices. To maximize efficiency, many people set their device speaking rates very high [12]. Because people who are **visually impaired (VI)** have experience with fast, synthetic speech, their abilities provide insight into human capacities to learn to process such speech.

Despite the potential informative power of visually impaired people's abilities, it is difficult to run an inclusive, large-scale study on listening rates. Traditional in-lab experiments compensate participants monetarily, which limits overall study size. Monetary compensation, scheduling during work hours, and fixed location also impact geographic, cultural, and socio-economic diversity [57]. In particular, by requiring participants to travel to the study location, in-lab experiments often exclude people who are visually impaired and other disabilities, due to inaccessibility of study locations.

In this article, we present extended results from the first large-scale study on human listening rates, with attention to how people who are visually impaired compare with sighted people in their listening rates [13]. The study was originally run in a two-month deployment with 453 participants [13], and in this work we present results from a longer 12-month deployment with 1,409 participants, and with expanded angles of analysis. This expanded journal paper allows us to add many additional analyses, for example adding an entire qualitative evaluation section as well as delving deeper into factors that may have impacted the results. The larger sample size also allows for increased reliability of findings, and for increased ability to study outlier populations.

The study's design as an online, screen reader accessible, volunteer-based study removed some participation barriers faced by previous studies. Participants listened to a series of clips read by synthetic speech and answered a variety of questions about what they heard. Our results show that visually impaired listeners had higher listening rates, likely attributable to early exposure to fast, synthetic speech. We position these results to motivate future research expanding possibilities for human-conversational agent interaction to consider not just interaction at speeds that a human

speaks, but to explore ways to make these interactions more efficient and productive by teaching users to understand faster speeds.

Our main contributions are as follows:

- We conduct the first large, inclusive, online study on human listening rates with 1,409 volunteer participants, demonstrating the feasibility of attaining volunteer crowdworkers for audio tasks, including people with disabilities. We extend an initial two-month deployment [13] to a 12-month deployment, and expand our analysis.
- Using the data gathered, we analyzed the intelligibility of fast, synthetic speech, developing models of people's listening rates, and assessing the impact of text complexity.
- Our results show that synthetic speech is intelligible to many people at rates much faster than typical human speaking rates, suggesting that there is room to increase and optimize conversational agent speaking rates to save users time.
- The superior performance of young participants who are visually impaired suggests that early exposure to synthetic speech increases ability to process fast synthetic speech, which could benefit everyone if fast listening is part of our future.

2 RELATED WORK

Our online study on listening abilities is informed by an understanding of how the human brain processes spoken language, developments in synthetic speech generation, past (smaller) studies on listening abilities, and the potential of online studies to study perceptual phenomena. Our work supports previous findings that people who are visually impaired typically outperform sighted people at listening tasks, and provides a model for validating prior in-lab listening studies by reaching a larger, more diverse population.

2.1 Psychoacoustics of Speech Perception

The process of converting speech to words with meanings is complex, spanning the fields of biology, psychology, physics, electronic engineering, chemistry, and computer science. Speech perception begins with an acoustic stimulus hitting the ear. At the inner ear, it vibrates the organ of Corti, which causes hair cells there to send signals to the auditory nerve. These impulses travel to the primary auditory cortex, where phonemes, individual sounds comprising words, are recognized. They also travel to Wernicke's area and other brain regions, which identify words and retrieve associated meanings. The exact roles of different brain regions in this process is an open area of research [56].

Several psychophysical models exist for how the brain converts audio signals to words [2]. Some models center around segmenting sounds into words (e.g., References [15, 44]). In such models, words are recognized as the word utterance finishes. However, these models do not account for accurate recognition of word sequences with ambiguous word boundaries. Other models account for this ability by assuming that the brain computes multiple sets of words and word parts that plausibly match the incoming audio (e.g., revised cohort [43] and TRACE [45] models). More recent research entirely rejects that speech is processed sequentially, instead assuming that future sounds impact interpretation of past sounds and words (e.g., Reference [17]). While our understanding of speech processing has advanced significantly, psychoacoustics is still an active research area.

2.2 Speech Synthesis

Speech synthesis is used by computers to produce human-like speech. During speech synthesis, the text is first broken down into sound units. In *concatenative synthesis*, these units are translated into audio by piecing together pre-recorded units of real human speech (e.g., References [11, 51, 65]).

When the domain is limited, entire words can be stored, but typically word parts are needed for greater flexibility. In *formant synthesis*, the text is translated into audio entirely synthetically using a model of speech generation (e.g., Reference [37]) or speech acoustics (e.g., Reference [75]). These two classes of synthesis exist in a wider landscape of methods, including methods more generally based on acoustic properties of the human vocal tract and articulation system, termed *articulatory synthesis* [67, 68], and methods built around popular artificial intelligence methods such as Hidden Markov Models [74, 75, 82] and more recently deep learning [70, 83].

Making intelligible, natural-sounding synthetic speech is difficult. Concatenative synthesis can distort speech, due to difficulties matching the text to a collection of recordings. Formant synthesis does not suffer from these distortion problems, but can sound unnatural, as it is difficult to model human speech. Pronunciation sometimes depends on context, but understanding natural language in real-time is not solved. For example, systems must handle words with identical spelling but different pronunciations (e.g., "wind" as a noun vs. verb). Conveying emotion in synthesized speech is also a challenge, with various methods proposed [4, 36, 64]. How emotion is conveyed can also impact understanding. This research is driven by industry as well as academia, with the emergence of digital assistants (e.g., Alexa, Siri, Cortana, and Google Assistant), and other speech-driven apps (e.g., text-to-speech, and GPS systems). However, companies do not always publish their speech synthesis methods, so determining the exact methods used by a given screen reader can be difficult.

The visually impaired community has a longer experience with synthetic speech. Screen readers, which emit synthetic speech, are this group's most popular assistive technology [40]. A screen reader is software that converts interfaces and digital text into spoken text, allowing users to navigate interfaces and access text without sight. Popular screen readers include ChromeVox [24], JAWS [66], NVDA [1], TalkBack [25], VoiceOver [5], and Window-Eyes [48] (recently discontinued). Screen readers typically allow users to choose a voice and speed. Newly blind people prefer voices and speeds resembling human speech (concatenative synthesis), while experienced screen reader users prefer voices more resilient to distortion at high speeds (formant synthesis), and save time by setting them to high speeds, even reaching 500 words per minute [12], compared to a normal speaking rate of 120–180 words per minute [49].

2.3 Listening Abilities of People Who Are Blind or Have Low Vision

People who are blind or have low-vision, and in particular those who are blind from birth, often outperform their sighted peers on a variety of auditory tasks. In terms of musical abilities, blind people are generally better at identifying relative pitch (e.g., Reference [26]), and are more likely to have perfect pitch, the ability to identify absolute sound frequencies (e.g., Reference [30, 63]). Blind people are also typically better at sound localization [62, 78], and process auditory stimuli faster [60]. Some blind people also use echolocation to understand their surroundings [38]. Experts can even ride bikes without hitting obstacles [47] and achieve spatial resolution comparable to peripheral vision [72]. Blind people excel at high-level cognitive functions as well, including processing words and sentences faster (e.g., Reference [58]), and remembering auditory stimuli (e.g., Reference [35]).

It is possible that these blind "superabilities" result from blind people's brains processing information differently from sighted people's brains [73]. Much of this evidence comes from brain scans taken while people perform tasks or are exposed to stimuli. Studies have shown that blind people use the visual cortex, a region traditionally thought to be reserved for processing visual stimuli, for other cognitive processes [61, 73, 80]. Such work provides evidence that our brains have a degree of plasticity, and that regions previously thought to be used exclusively for specific functions, and in particular sensory input, can be used for other purposes [3, 33].

One source of controversy is whether the onset of blindness affects people's auditory abilities, and if so, how much. Some studies suggest that the age of onset for blindness determines whether a person will have heightened auditory abilities (e.g., Reference [79]). These studies align with the fact that early childhood is a major time of cerebral growth and development, and suggest that the brain adapts more effectively during that time. However, other studies provide evidence that people can adapt both behaviorally and neurologically later in life (e.g., Reference [59]). Such conflicting results highlight the need for larger studies on the relation between age, visual impairment, and listening abilities, which we provide in our online study.

2.4 Listening Rate Studies

Past studies on human listening rates are small[1] and have not always included visually impaired listeners (see Reference [21]). More recently a push has been made to include people with visual impairments, given their extensive use of text-to-speech (e.g., Reference [6]). Since then, studies have compared sighted and visually impaired listeners (e.g., Reference [8]). Studies have also compared the intelligibility of speech produced by different mechanisms, including natural speech, formant synthesis, and concatenative synthesis (e.g., Reference [50]), and compared efficiency of single vs. multi-track speech (e.g., Reference [29]). One study of listening rates for blind users suggests that normal speaking rates are preferred when interacting with a conversational agents [14]. In our study, we focus on understanding how well blind and sighted users can understand faster synthetic speaking rates. Our larger sample size also allows for more reliable statistical analysis.

These studies have employed diverse methods for assessing listening rate. Comprehension questions (e.g., Reference [54]), word identification tasks (e.g., Reference [8]), and transcription or repetition tasks (e.g., References [6, 71]) have been used. Some studies also use subjective metrics (e.g., References [6, 50, 76]). Choice of test materials and questions is important, as using even different lengths of text can lead to different conclusions [21]. In our study, we use three types of test questions to help account for this disparity.

Past study results sometimes conflict, even when using similar tests. Many studies conclude that visually impaired people can comprehend speech at faster rates (e.g., References [50, 71, 76]). However, other studies have found no significant difference between these groups (e.g., Reference [54]). Evidence that other factors, such as age and practice, impacts listening abilities has also emerged (e.g., Reference [71]). Conflicting study results and the complexity of factors impacting listening rate suggest the need for a large-scale study on listening rates, such as ours.

2.5 Online Perceptual Studies

Crowdsourcing is a powerful tool for running large-scale experiments. Researchers have demonstrated the validity of online experiments by replicating in-lab results using online participants (e.g., References [23, 32, 53]). Past studies have often focused on visual perception, for example evaluating shape and color similarity [18] or visualization techniques [31]. More recently, the development of crowdsourced transcription systems demonstrates that audio tasks can also be effectively crowdsourced (e.g., Legion:Scribe [39] and Respeak [77]). While visually impaired workers could be valuable for auditory tasks, to the best of our knowledge, such tasks have not been made accessible to people with visual impairments, until now.

For our study, we used LabintheWild [57], a platform that motivates participation through self-discovery, providing information about the participant's performance compared to peers at the end of each study. Volunteer-based platforms like LabintheWild have been shown to reach larger, more diverse populations than crowdsourcing platforms with monetary compensation [41, 57]. It also

[1]Max participants: visually impaired 36 [71], sighted 65 [54].

has been shown to reach people with disabilities who participate in online studies to learn more about themselves and compare their abilities with others [41]. Importantly, experiments conducted on LabintheWild have been shown to accurately replicate the results of controlled laboratory studies [41, 42, 57]. In this work, we extend the space of volunteer-based crowdsourced experiments to include studies with auditory tasks.

3 STUDY

To help inform the optimization of speaking rates for conversational agents, we conducted a 5–10 min online study on LabintheWild to evaluate the intelligibility of fast, synthetic speech. We chose LabintheWild as an experiment platform because its participants are very diverse in terms of demographic backgrounds and abilities [41, 57] and its volunteer participants have been found to provide reliable data and exert themselves more than participants recruited from Mechanical Turk [7, 81]. LabintheWild also allowed us to reach a larger participant pool than common lab experiment, and to facilitate participation by visually impaired participants. The online study was made fully accessible to include people who are visually impaired, and other disabilities, who often face barriers when attempting to participate in crowdsourcing platforms like Mechanical Turk [84].

The study was designed to answer three main questions:

(1) What synthetic speaking rates are typically intelligible?
(2) How do demographic factors, including visual impairment and age, impact listening rate?
(3) Can people who are visually impaired process higher synthetic speaking rates than sighted people, and if so, to what extent does practice with screen readers account for this superior ability?

3.1 Question Types

The study employed three types of questions to evaluate participants' listening rate. They measure three different aspects of speech intelligibility: individual word recognition, sentence comprehension, and sentence recognition.

(1) Rhyme test: measures word recognition by playing a single recorded word, and asking the participant to identify it from a list of six rhyming options (e.g., went, sent, bent, dent, tent, rent). We used 50 sets of rhyming words (300 words total), taken from the Modified Rhyme Test [34], a standard test used to evaluate auditory comprehension.
(2) Yes/no questions: measures sentence comprehension by playing a recorded question with a yes/no answer, and asking if the answer is "yes" or "no" (e.g., Do all animals speak fluent French?). We used 200 questions (100 "yes" and 100 "no") chosen randomly from MindPixel [46], a large dataset of crowdsourced questions.
(3) Transcription: measures sentence recognition by playing a recorded simple sentence, and asking for a transcription. To create the statements, we converted 100 "yes"-answered MindPixel questions into statements. (ex: "Do bananas grow on trees?" became "Bananas grow on trees.")

3.2 Procedure

The study was designed as a single-page web application. It consisted of three main parts: (1) basic demographic questions and questions about participants' vision status (whether visually impaired, and if so whether blind, low-vision, or other), and experience with text-to-speech software, (2) a set of listening questions where recordings of synthetic speech were played at various speeds, and the participant answered questions about that text, and (3) feedback on the participant's listening rate in comparison to others.

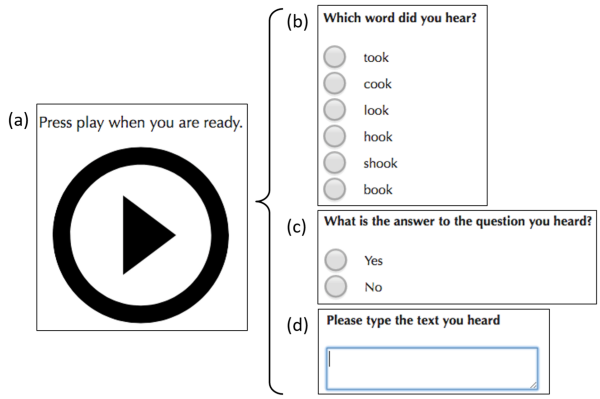

Fig. 1. Screen shots of the listening question interface. (a) The prompt for playing a listening question. ((b)–(d)) The subsequent question asked about the audio played, (b) for a rhyme test, (c) for a yes/no question, and (d) for transcription.

The listening questions comprised the main part of the study. These questions were presented one at a time, as shown in Figure 1. The page presented an audio clip, and instructed the participant to play it (Figure 1(a)). The recording could only be played once. Once the recording finished, they were given a question about the audio they just heard (Figure 1(b)–(d)). Participants answered three practice questions, one for each question type, followed by 18 questions used to measure listening rate. The set of 18 was divided into three groups of six questions. Each group of six comprised two random questions from each question type, all randomly ordered. After each group of six, the participant was instructed to take a break as needed.

The listening question speed was dynamically adapted using binary search, so that participants who did well progressed to faster speeds and those who struggled moved to slower speeds. Each set of six questions had a fixed speed, so that each person was tested with exactly three speeds. To determine correctness at each speed, we used a weighted sum that gave harder questions more weight. If the sum exceeded a threshold meaning that all six were answered correctly, with minor transcription errors allowed, then they advanced to a faster speed. Specifically, all users started with six questions at the middle speed (57). Users who passed the correctness threshold for this batch progressed to the midpoint of the speeds higher than 57 (86), and those who did not progressed to the midpoint of the slower speeds (29). The next six questions occurred at this speed, and their performance again determined whether the speed of the subsequent six would be faster or slower. The final speed was calculated as the midpoint between the two last speeds they experienced.

To compute the weighted sum, yes/no questions and rhyming tests were weighted by the probability of guessing incorrectly at random, and transcription was weighted by accuracy. The weights were: yes/no: 1/2 if correct, 0 else; rhyming: 5/6 if correct, 0 else; transcription: $\max(0, 1 - dist(S,T)/length(S))$ where S is the spoken text, T is the transcription, and $dist(a,b)$ is the edit distance between strings a and b. Our edit distance was Levenshtein distance, and punctuation was removed and capitalization ignored during computation. Intuitively, the transcription metric approximates the fraction of audio that was transcribed correctly. The threshold for advancing was 4.17 (out of $2(1/2+5/6+1) = 4.\overline{6}$), meaning all six questions were correct, except possibly minor transcription errors.

After completing the listening questions, participants received information on their performance. They were shown their final speed and percentile relative to other participants. To provide

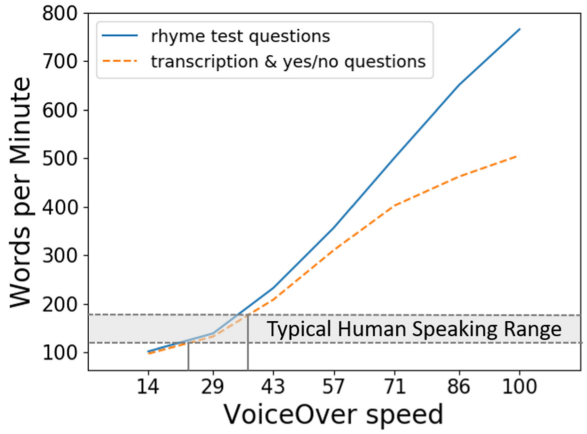

Fig. 2. VoiceOver speeds translated into words per minute, for the rhyme test questions (words) and for the transcription and yes/no questions (sentences). Typical human speaking rate 120–180 WPM corresponds to VoiceOver range 24–38.

comparison points for early participants, members of the research team (both BLV and sighted) took the study themselves. To help participants interpret their results, we provided audio samples of their listening rate, the average participant listening rate, and the fastest participant listening rate. To increase awareness among sighted people, we also explained what screen readers are, and described fast listening abilities of people who are visually impaired. This feedback provided education and self-awareness, which served as motivation and compensation for participation.

3.3 Digital Audio Recordings

The audio recordings used in the study were created using VoiceOver, Apple's screen reader. The default voice, Alex, was used, at Pitch 50, Volume 100, and Intonation 50. While this voice is commonly used, we note that basing our study on it may have impacted results, for example due to impact of speaker gender (male) or other qualities such as prosody. To convert question text to audio, we used AppleScript, an operating system-level scripting language, to make VoiceOver read the desired text, and trigger WavTap,[2] a program that pipes the system's audio to an audio file, to save the recording. We repeated the process for every question, at every speed.

We used seven equally spaced speeds spanning the full VoiceOver range (1–100): 14, 29, 43, 57, 71, 86, 100. We chose seven speeds so that the procedure's binary search would terminate quickly, with each participant answering questions at three speeds. To facilitate interpretation, we converted VoiceOver speeds to **words per minute (WPM)**, a more standard metric of speaking rate (Figure 2). Because this conversion is not publicly available, we computed it empirically by timing VoiceOver reading our test questions. To normalize word length, we used total letters divided by five, the average English word length, as word number: WPM = #letters / (5 × time(min)).

Because VoiceOver pauses between sentences, we computed WPM separately for the rhyme tests, which are individual words, and for the transcription and yes/no questions, which are sentences. The growing difference between the two corpuses shows that pause length does not scale proportionally with the VoiceOver rate, and begins to dominate WPM at high VoiceOver speeds. We use WPM for full sentences (transcription and yes/no questions) to interpret results, for

[2] http://download.cnet.com/WavTap/3000-2140_4-75810854.html.

applicability to interactions with conversational agents and text-to-speech software speaking full sentences.

3.4 Accessibility

To ensure that all participants had as similar an experience as possible, we created a single interface made to be universally accessible. The site design was minimalistic, with no unnecessary visuals or interactions. To support non-visual navigation, we made the site compatible with screen readers, as described in the following paragraph. To facilitate clicking on targets, which can be difficult for people with motor impairments or low-vision, all targets were large (as shown in Figure 1). To the best of our knowledge, the study is fully accessible to people with vision and motor impairments; we did not account for accessibility for people with hearing impairments as they were not eligible for this study.

To provide accessibility for visually impaired participants, all visual information was made available to screen readers. The page structure was made accessible by adding headings (e.g., <h1></h1>, etc.). Visual elements were made accessible by adding labels, aria-labels, and alternative text. To help ensure accessibility for different screen readers, we encoded visual information "redundantly" in multiple attributes, and tested the study with various screen readers and browsers.

To prevent output from the participant's screen reader from overlapping with a listening question, we incorporated a brief (1 second) pause at the beginning of each recording. The concern was that screen readers might announce that they are playing the audio at the same time the audio was playing, interfering with the study. This pause was created programmatically during the generation of the question recordings.

3.5 Measures

Our main performance metric is Listening Rate, which we define as the participant's fastest intelligible VoiceOver rate, as computed by our binary search procedure. Specifically, we compute whether the final speed they heard was too slow (i.e., if they "passed" our weighted cutoff), and compute the subsequent speed at which binary search would arrive. For example, if the last speed they heard was 71, and they answered all six questions at that speed correctly, their Listening Rate is 78.5, halfway between 71 and 86 (which they previously failed). We created our own measure, because measures from previous studies (e.g., Reference [6]), which advance participants through a range of speeds and provide statistics over the full range, do not apply; binary search tailors the study speeds to the participant, invalidating such comparisons.

We also measured question response time, which we used to eliminate outliers who took many standard deviations more time to answer questions than other participants. Participants who are visually impaired were typically slower than sighted participants at answering, likely because they had to navigate the study using a screen reader to read aloud all answer choices and interface options, rather than by sight. This navigation process may have required higher cognitive load, likely contributing to slower speed as well.

3.6 Participants

The study was launched on LabintheWild with IRB approval. The deployment ran for 12 months, during which 1,409 participants completed the study. This 12-month deployment is an extention of a smaller initial deployment, which ran for two months with 453 participants [13]. The completion rate was 77% (vs. 74% for the shorter deployment).

Recruitment occurred through the LabintheWild site, Facebook posts, relevant email lists targeting screen reader users, and word-of-mouth. Basic participant demographics are presented in Table 1. For our initial two-month study, we targeted blind and low-vision community members

Table 1. Participant Demographics Comparing the Current Work to Prior Published Work on a Subset of Results

	CHI 2019 (2-month deployment) [13]	Current Work (12-month deployment)
Number	453	1409
Age	8-80, m=34, sd=15	8-98, m=30, sd=14
Gender	257 (57%) female, 194 (43%) male, 2 (<1%) other	788 (56%) female, 589 (42%) male, 32 (2%) other
Vision Status	143 (32%) VI, 310 (68%) sighted	268 (19%) VI, 1141 (81%) sighted
VI Condition	101 (71%) blind, 23 (16%) low-vision, 9 (6%) other, 10 (7%) undisclosed	108 (40%) blind, 90 (34%) low-vision, 56 (21%) other, 14 (5%) undisclosed
First Language	354 (78%) English, 99 (22%) other	1013 (72%) English, 396 (28%) other
Primary Language	328 (72%) English, 125 (28%) other	1161 (82%) English, 248 (18%) other
Retakes	27 (6%) retakes, 426 (94%) new	77 (5%) retakes, 1332 (95%) new

during recruitment, in addition to the population at large. In contrast, our 12-month deployment did not leverage any targeted recruitment. Correspondingly, we see a higher incidence of sighted people in our longer deployment, which is more reflective of the population at large (estimated at about 0.58% of the global population being blind, and 4.2% blind or low vision in 2010 [55]). Specifically, the percentage of participants who self-reported as blind decreased from 22% (101 of 453) to 19% (268 of 1,409), and the percent who self-reported as having a visual impairment more generally from 32% (143 of 453) to 19% (268 of 1,409). The percent of participants who reported having previously participated (i.e., "Retakes" in Table 1) remained fairly unchanged.

4 RESULTS

To answer the three questions guiding our study design, we (1) computed the overall Listening Rate distribution to determine which synthetic speech rates are typically intelligible, (2) computed a linear regression analysis for the entire population to determine which demographic factors impact Listening Rate, and (3) computed a linear regression analysis for the visually impaired subpopulation to determine if and how experience with screen readers impacts ability to interpret fast, synthetic speech. We also examine extremely high performers to gain insight on outstanding listeners and examine the impact of text complexity on intelligibility.

4.1 Listening Rate Distribution

To determine which synthetic speaking rates are typically intelligible, we computed the distribution of Listening Rates, shown in Figure 3. The distribution resembles a skewed-right Gaussian distribution, and peaks at rates 57–71, with 31.23% of participants falling in this range. The mean Listening Rate was 55.3, which corresponds to 298 WPM. Given that people typically speak at a rate of 120–180 WPM, these results suggest that many people, if not most, can understand speech significantly faster than today's conversational agents with typical human speaking rates.

4.2 Factors Impacting Listening Rate—Overall Population

Our method for analyzing which factors impact Listening Rate replicated the procedure for the smaller dataset in Reference [13]. Specifically, to analyze which factors impact Listening Rate, we ran a linear regression analysis. We conducted a series of multiple regressions, and compared models using the **Akaike information criterion (AIC)** to determine which factors to include. The factors explored were identical to those in the prior publication: age, visual impairment, years of screen reader use, whether they use a screen reader in their daily lives, native language, and education level. We included the interaction between age and visual impairment as a covariate, because young people who are visually impaired have the opportunity to use technology and screen readers from a young age, unlike older generations, which could impact

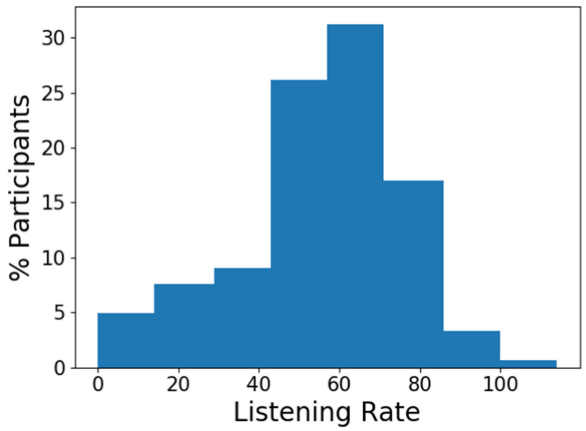

Fig. 3. Histogram of Listening Rates for all participants.

Table 2. Linear Regression Predicting Listening Rate for All Participants from Demographic Variables

Variable	Est.	SE	t	Pr(>|t|)	
(Intercept)	54.85	1.54	35.69	<0.001	***
VI [yes]	7.36	3.09	2.38	0.017	*
Age	−0.21	0.04	−4.82	<0.001	***
Age × VI [yes]	−0.21	0.09	−2.35	0.019	*
Native English [yes]	9.17	1.20	7.64	<0.001	***

Abbreviations: VI visually impaired, Est. estimate, SE standard error.
Significance codes: ***<0.001, **<0.01, *<0.05.

Listening Rate. Table 2 provides the results of the linear regression model that minimized information loss. This model explains 6% of the variance in people's Listening Rates (multiple and adjusted $R^2 = .06$).

The factors included in our model according to AIC are consistent with those published previously from the smaller deployment. More specifically, the model that fit our large dataset with the best AIC contained the exact same factors as in the prior model based on the smaller dataset. As in the prior model, all factors are also statistically significant, though the level of significance for some variables has shifted. Specifically, the intercept is comparably significant (***); significance increased for age (* to ***) and Native English [yes] (** to ***); and significance decreased for VI [yes] (*** to *) and Age x VI [yes] (*** to *).

The model shows that being visually impaired significantly impacts Listening Rate, increasing the predicted rate by 7.36. Age is also significant, and more so, with every year of age, the Listening Rate decreasing by .21. The interaction between age and visual impairment is also significant, meaning that age has a moderating effect on how much visual impairment boosts the predicted Listening Rate.

Being a native English speaker also has a strong significant positive effect, increasing expected Listening Rate by 9.17. Indeed, the significance of native language is higher than that of vision status. It is possible that the first years of language exposure are difficult to make up for later in life, even with many hours of practice through a screen reader or other text-to-speech software.

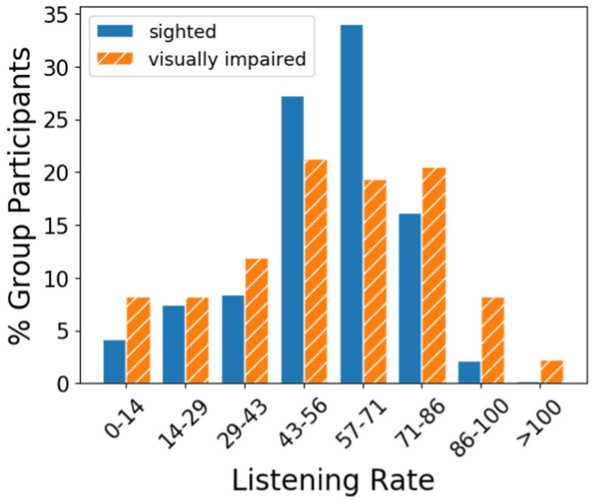

Fig. 4. Histogram of Listening Rates, separated into visually impaired and sighted participant groups.

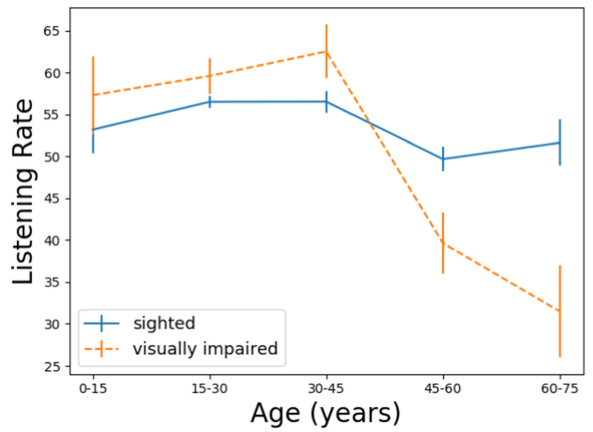

Fig. 5. Plot of age vs. Listening Rate, for visually impaired and sighted groups.

To better understand the difference in Listening Rates between sighted and visually impaired participants, determined significant by our model, we examined the difference in Listening Rate distributions between the two groups. The histograms are shown, side-by-side, in Figure 4. The distribution for visually impaired participants appears shifted to the right. The mean Listening Rate for visually impaired participants was 55.7 (301 WPM) while for sighted participants it was 55.2 (297 WPM).

Given the significance of age and visual impairment as covariates, we explored the relationship of these two variables further. Figure 5 shows the results, in a plot of Listening Rate vs. age for sighted and VI groups. The figure shows that while young (under 45), visually impaired participants typically had the highest Listening Rates, older (over 45), visually impaired participants typically had the lowest Listening Rates. Age correlates significantly ($p < 0.05$) with lower Listening Rates for both visually impaired participants ($r = -0.235, p = 0.0001$) and sighted participants ($r = -0.108, p = 0.0002$).

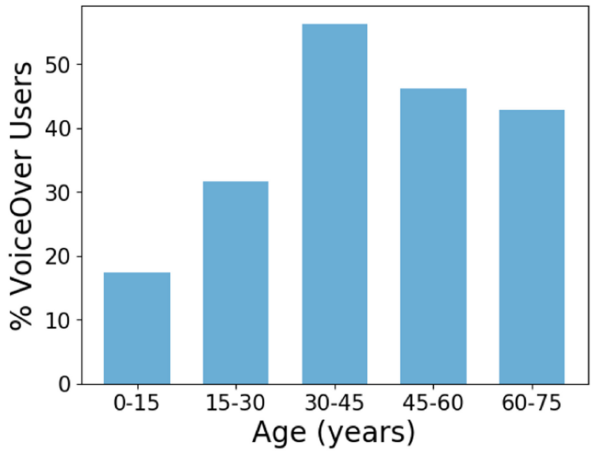

Fig. 6. Plot of age vs. VoiceOver usage for visually impaired participants.

To explore possible reasons for the marked drop in Listening Rate for visually impaired participants over 45 years old, we explored trends in several demographic variables. In particular, we examined VoiceOver usage, because experience with the specific screen reader used in the test may have impacted results, and it is possible that people in different age brackets primarily use different screen readers. Figure 6 provides a plot of the percent of participants in each age group that used VoiceOver, which was used in our study. The trend across age groups mirrors the trend in Listening Rate across ages, with a peak at age group 30–45. The difference in Listening Rate between visually impaired participants who reported using VoiceOver (102 people) compared to those who did not (165 people) is statistically significant ($t(265) = 4.531$, $p < 0.001$).

4.3 Factors Impacting Listening Rate—Visually Impaired Population

To better understand participants who are visually impaired, and in particular why Listening Rate declines with age only among our visually impaired participants, we ran another linear regression to predict Listening Rate with only visually impaired participants. Again, we ran a series of multiple regressions, and compared models using AIC to determine which factors to include. The same factors were explored as for the more general model: age, visual impairment, years of screen reader use, whether they use a screen reader in their daily lives, native language, and education level. We included the interaction between age and years of screen reader use as a covariate to account for correlation between the two, with older participants having more experience. Table 3 provides the model that minimized information loss. The same factors are included in this model, as were in the model formed from our prior smaller deployment.

This model indicates that for every year of screen reader usage, we expect an increase in Listening Rate of 3.08. The negative covariance between age and screen reader usage indicates that with age, having used a screen reader for a longer time has less of an impact. It is likely that years of screen reader use is significant to the model for visually impaired participants, but not for the overall population, because a significantly higher percentage of people who are visually impaired use screen readers. Age and screen reader years are also correlated ($r = .472$, $p < 0.001$), meaning that by including age in the overall model, it captured some information about screen reader usage as well. This model accounts for 26% of the variance in the visually impaired population, (multiple $R^2 = .26$, adjusted $R^2 = 0.25$), compared to the overall model's 6%, suggesting that it is a substantially better model for this subpopulation.

Table 3. Linear Regression Predicting Listening Rate for Participants Who Are Visually Impaired from Demographic Variables

Variable	Est.	SE	t	Pr(>|t|)	
(Intercept)	54.74	3.60	15.18	<0.001	***
Age	−0.18	0.11	−1.57	0.119	
SR Years	3.08	.37	8.36	<0.001	***
Age × SR Years	−0.06	0.01	−7.08	<0.001	***

Abbreviations: SR screen reader, Est. estimate, SE standard error.
Significance codes: ***<0.001, **<0.01, *<0.05.

To further analyze the relationship between age and screen reader years, we examined screen reader adoption age. Given prior work suggesting that listening abilities are most adaptable at a young age (e.g., Reference [79]), combined with the lack of screen reader availability when older generations were young and the possibility of becoming visually impaired later in life, we hypothesized that early adoption might differ across age, along with Listening Rate. It is also possible that most of any improvement in listening rate comes in the first few years of screen reader use, making the age at which this learning curve is tackled particularly important. Adoption age does correlate with both age ($r = .734, p < 0.001$) and Listening Rate ($r = -.346, p < 0.001$). These correlations suggest that age in and of itself might not account for the decline in Listening Rates for visually impaired participants. Rather, lack of exposure at a young age to screen readers and fast speaking rates might account for older generations' lower performance.

4.4 General Screen Reader Usage

In total, 138 participants reported using a screen reader. Interestingly, this group included 17 sighted participants, in addition to 121 visually impaired participants. The breakdown in screen reader choice for our participants is listed in Table 4.

We further analyzed the impact of screen reader use among visually impaired participants, who comprise the vast majority of this group. Whether these participants used a screen reader did have a significant impact on performance ($t(266) = -3.694, p = 0.0002$), as shown in Figure 7. However, among people who used a screen reader, choice of screen reader did not significantly impact Listening Rate. In particular, experience with VoiceOver, which was used in the study, compared to other screen readers was not statistically significant ($t(129) = -.370, p = 0.712$).

We also examined demographics and comments from sighted participants who reported using screen readers. The majority of these participants (12) did not report using one of the mainstream screen readers we listed on the form, but instead selected "Other." Given the tiny percent of sighted participants who indicated that they used screen readers (1% of sighted participants vs. 45% of visually impaired participants), it is possible that these participants mistakenly interpreted this question as referring to general text-to-speech software. It is also possible that these participants used screen readers due to other disabilities, such as dyslexia. However, the general comments that these participants left did not reveal any comorbidities or details about their reported screen reader usage.

4.5 Super-listeners: The Top .6%

Our study identified a group of elite listeners, who answered all questions correctly at the highest available speed (VoiceOver speed 100), achieving a Listening Rate over 100. Specifically, 9 (0.6%) participants fell into this group we call super-listeners. Six of the nine were blind, representing 2.2%

Table 4. Participants Who Reported Using a Screen Reader

	Total	VI	Sighted
Screen Reader Users	138/1409 (9.8%)	121/268 (45.1%)	17/1141 (1.5%)
VoiceOver	106 (76.8%)	102 (84.3%)	4 (23.5%)
JAWS	92 (66.7%)	92 (76.0%)	0 (0%)
NVDA	80 (58.0%)	80 (66.1%)	0 (0%)
TalkBack	27 (19.6%)	26 (21.5%)	1 (5.9%)
Window Eyes	9 (6.5%)	7 (5.8%)	2 (11.8%)
ChromeVox	9 (6.5%)	7 (5.8%)	2 (11.8%)
Other	23 (16.7%)	11 (9.1%)	12 (70.6%)

The number of participants in each category is provided, and the percentage of the demographic that this number comprises is provided in parentheses. Note that the sum exceeds total participants, as many participants used multiple screen readers.

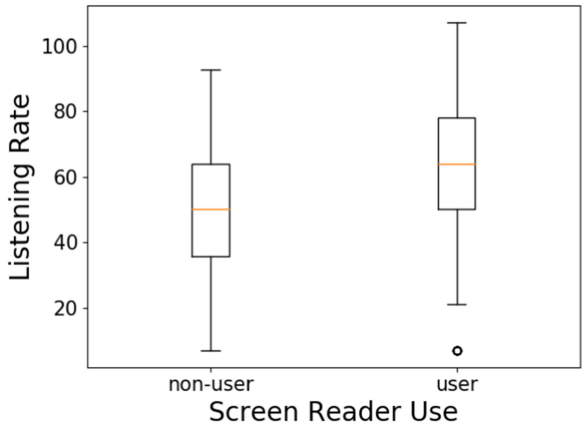

Fig. 7. Boxplot comparing performance of people who were screen reader users to those who were non-users. The orange line shows the median, the box contains the middle two quartiles (the middle 50%), and the whiskers extend an additional 1.5 times the interquartile range in each direction.

of visually impaired participants and 5.6% of blind participants, compared to 0.3% of the sighted population. To find out more about these super-listeners, we looked at their demographics (shown in Table 5. This group was generally young (all aged 18–35), blind (all but three), male (all but two), used English as their primary language (all), and were native English speakers (all but one). Note, however, that we cannot infer generalizability of these findings given the small N. Interestingly, although 17 sighted participants reported using screen readers (above), the three sighted super-listeners were not self-reported screen reader users.

Because this population's Listening Rate exceeds the range of speeds currently available on popular screen readers and many of these "super-listeners" are blind, they might benefit from an expanded set of speeds available on screen readers and text-to-speech software more generally. Additionally, if screen readers provided an expanded set of speeds, these people would be able to practice at speeds over 100, which could result in even higher Listening Rates.

4.6 Great-listeners: The Top 4.0%

Our study identified a group of elite or nearly elite listeners, who answered all questions correctly at the highest or second-highest available speed, achieving a Listening Rate of 86 or higher. This

Table 5. Demographics for the Nine Super-listeners (Top .6% Performers)

	VI						Sighted		
	S1	S2	S3	S4	S5	S6	S7	S8	S9
Gender	Male	Male	Male	Male	Male	Female	Male	Male	Female
Age	24	24	35	26	23	24	18	31	20
Vision	blind	blind	blind	blind	blind	blind	sighted	sighted	sighted
Uses screen reader	yes	yes	yes	yes	yes	yes	no	no	no
First language	English	English	English	English	English	Spanish	English	English	English
Primary language	English	English	English	English	English	English	English	English	English
Multinational	No	No	Yes	No	No	No	No	No	No

Table 6. Demographics for Great-listeners (Top 4.0% Performers)

	56 Total		28 VI		28 Sighted	
Gender	25 female	31 male	12 female	14 male	13 female	15 male
Age	26.7 mean	7.0 SD	28.5 mean	7.3 SD	24.9 mean	6.5 SD
Vision	28 VI	28 sighted	20 blind	5 low-vision	0 VI	28 sighted
Uses screen reader	25 yes	31 no	24 yes	4 no	1 yes	27 no
First language	47 English	9 other	23 English	5 other	24 English	4 other
Primary language	53 English	3 other	27 English	1 other	26 English	2 other
Multinational	8 yes	48 no	5 yes	23 no	3 yes	25 no

SD stands for standard deviation. Three participants who identified as visually impaired did not identify with either blind nor low-vision. Two entered "other" and one entered "undisclosed." Other first languages included Chinese (2), Spanish (4), German (1), Danish (1), and unspecified (1). Other primary languages included German (2) and Danish (1).

group includes the group of 8 super-listeners in the prior section. Of 1,409 participants, 56 (4.0%) are classified as great-listeners, compared to 28 (10.4%) of the 268 visually impaired participants and 20 (18.5%) of the 108 blind participants. As can be seen in Table 6 the 56 great-listeners were equally divided into visually impaired and sighted, in contrast to the only 19% being visually impaired in the entire study. Genders were about equally balanced for great-listeners regardless of being visually impaired or not. The great-listeners were on average slightly younger than the average participant (see Table 1), and the sighted great-listeners were slightly younger than those who are visually impaired. Interestingly, one great-listener identified as sighted and being a screen reader user. This is not unusual, because some sighted people with certain reading related disabilities use screen readers on a regular basis.

4.7 Impact of Text Complexity

To shed light on how conversational agents can adapt not only to users, but also to content, we analyzed the intelligibility of various content used in our study. Since equal numbers of questions from each of the three types were asked at each speed, we used accuracy as a metric for intelligibility. We found yes/no questions to be easiest (84.5% accuracy), followed by rhyme tests (83.7%), and transcription (82.9%). Recall that transcription accuracy was computed by the metric from the study procedure, edit distance divided by target string length. These metrics and thresholds for progression were designed to make it difficult to progress to higher speeds by guessing. The comparable accuracy across all question types suggests that these thresholds and metrics were appropriately designed.

We also suspected that question length might impact accuracy due to fatigue, and so examined the relationship between question length and accuracy in more depth. Figure 8 shows the relationship between question length and question accuracy. Question length was computed as the number of characters in the question. Accuracy is computed for each question type according to

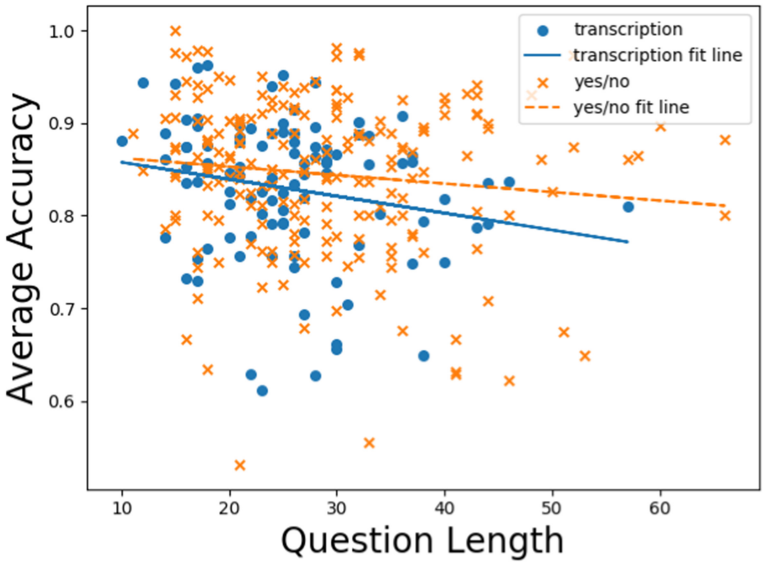

Fig. 8. Scatterplot of question length vs. accuracy for transcription and yes/no questions. Accuracy for transcription questions is computed as the edit distance between the answer and target over the target string length. Fit lines are provided for each, which minimise the squared error.

our metrics defined in our study procedures—whether or not correct for yes/no questions, and 1 − (edit distance)/(target length) for transcription questions. The average number of characters per question for each question type was: transcription 25.77 (std 8.27), yes/no 28.61 (std 10.47), and MRT 3.60 (std .56). The average word length (characters per word) for each question type was: transcription 5.87 (std 1.78), yes/no 5.40 (std 1.35), and MRT 3.60 (std .56). We found statistically significant correlations ($p < 0.05$) between accuracy and question length for transcription ($r = -0.052, p < 0.001$) and yes/no questions ($r = -0.025, p = 0.021$).

The two questions with the lowest accuracies (lowest x's on Figure 8) represent the questions "Do candles give milk?" with 53% accuracy and "Does jam make a good rocket fuel?" with 56% accuracy. It is possible that these low accuracies are due to the questions sounding similar to alternative questions with the opposite answer, e.g., "Do *cows* give milk?" rather than "Do *candles* give milk?" The relatively low frequency with which these questions were randomly selected also increased the likelihood of them having uncharacteristically low accuracies—we had 32 answers for "Do candles give milk?" and 45 answers for "Does jam make a good rocket fuel?" compared to an average of 52 answers/question.

Differences in intelligibility for different question types and lengths suggests that different speaking rates may be appropriate for different auditory interactions, and suggests room to optimize speaking rate for conversational agents based on both the participant and the content.

4.8 Impact of Retakes

Of our participants, a small percentage (5%) self-reported that they were retaking the study, or that they had already participated. To better understand the possible impact of retaking the study, we examined this set of participants. This group skewed male (48% of retakers vs. 41% of first-time takers), and sighted (87% of retakers vs. 80% of first-time takers). Interestingly, retaking the study was significantly correlated with *lower* performance ($r = -0.054, p = 0.043$). Specifically, the

average (mean) Listening Rate for repeat takers was 50.73 (std dev 22.54), compared to 55.60 (std dev 20.40) for first-time takers. Given that practice typically improves performance, it seems that lower-performing participants more often opted to retake the study, perhaps hoping for a better score the next time around.

4.9 Possible Confounding Variables

Because the study was run online, the environment and setup could vary between participants. To help control for these possible confounding variables, we recorded the participant's device when available, and asked them at the end of the study if they experienced any issues with environmental noise. Comparing groups with different setups and audio quality revealed no statistically significant differences, suggesting that environmental differences did not systematically skew results.

Specifically, the difference in the distributions of devices used by sighted vs. visually impaired groups was not statistically significant, as computed by a chi-squared independence test ($X^2 = 30.32, p = 0.60$). When asked if they experienced interfering environmental noise, 91.4% of sighted and 87.4% of visually impaired participants answered "no," with no statistical significance between the groups, as computed by a chi-squared independence test ($X^2 = 3.408, p = 0.065$). The difference in Listening Rates between the groups who answered "yes" vs. "no" was not statistically significant ($t(1407) = .014, p = 0.999$).

It is possible that using a screen reader affected participants' memory of the original audio. To minimize the experiential difference, the interface design was simple, and all participants chose setups that best suited their abilities. Still, as reported above, visually impaired participants typically took longer to answer questions, likely due to screen reader usage [9]. The time required to navigate using a screen reader places the question audio farther in the past, making it harder to remember than it was for non-screen reader users. Despite this disadvantage, visually impaired participants significantly outperformed their sighted peers, reaffirming the finding that people who are visually impaired typically have higher Listening Rates.

While we largely account for variation in difficulty across question types through our accuracy metrics and thresholding, we did not control for variation in question difficulty within a single question type. As explored earlier, question length may impact difficulty or average response accuracy. So for example, a participant who was unlucky enough to receive very long or otherwise difficult questions, in particular early on in the test, may receive a lower score than otherwise warranted. Conversely, a participant who receives disproportionately short or easy questions early may achieve a high score not reflective of their actual abilities.

4.10 Qualitative Feedback

To better understand participants' experiences of the study, we conducted an analysis of participants' qualitative feedback. To that purpose, at the end of the study, we asked participants if they had any feedback they wanted to share with the researchers, and gave them a textbox to enter it. We used an open coding process to analyze their feedback [16, 20]. First, two researchers independently identified the main themes present in the responses, and labeled each response with these themes. Final labels were created for each response by iterating and discussing any disagreements until consensus was reached.

Of 1,409 participants, 205 (14.5%) provided feedback. Of these participants, 64 were visually impaired (23.9% of 268 total visually impaired participants), and 140 were sighted (12.3% of all 1141 sighted participants). Participants who are visually impaired were more likely to have feedback.

4.10.1 Themes. The identified themes were grouped according to the study component to which they referred: Study Criticism (related to the study procedures), Interface Criticism

Table 7. Thematic Analysis of Participants' Open Feedback for Both Sighted and VI Participants

Study Component	Theme	64 VI	161 Sighted
Study Criticism	Speed	11 (17.2%)	41 (29.3%)
	Voice Choice	12 (18.8%)	15 (10.7%)
	Questions	6 (9.4%)	22 (15.7%)
	Confounding Variables	7 (10.9%)	2 (1.4%)
	Confusion	2 (3.1%)	7 (5%)
Interface Criticism	Lag	3 (4.6%)	3 (2.1%)
	Audio	2 (3.1%)	4 (2.9%)
Overall Experience	Enjoyment	21 (32.8%)	25 (17.9%)

Some participants expressed more than one criticism, so the numbers may not sum to the total in each group.

(specific to technical interface components), and Overall Experience. The themes are summarized in Table 7, and discussed in more detail below.

Study Criticism - Many participants provided feedback or criticism on the study design. Themes that emerged surrounding the study design were: speed (of the question text), voice choice (used for question text), questions (content and design of the questions), confounding variables (e.g., concurrent noises or hearing loss), and confusion (e.g., about how the speed progressed over the course of the study).

Interface Criticism - Many participants also provided feedback or criticism of the web interface built for the study. Themes in this group were: lag (delay between interaction such as clicking and response) and audio (e.g., seemingly missing audio).

Overall Experience - Participants also reflected on their overall experience with the study. Only one theme emerged here: enjoyment (positive or negative).

4.10.2 Differences by Vision Status. To better understand experiential differences between participants are visually impaired, and hearing participants, we analyzed and present results for these groups separately.

Sighted Participants - Unlike their peers who are visually impaired, sighted participants took more issue with the voice speed used during the study (29.3% vs. 17.2%). In particular, they were commonly offended by or disagreed with the test's (too-low) assessment of their listening rate. For example, one participant wrote, "I'm insulted that it spoke slowly." This mismatch between how well they thought they performed and the study's progression to slower speeds also led many participants to question the study methodology, rather than question their own listening abilities. For instance, one participant remarked, "don't understand why the screen reader not speeding up for me."

Sighted participants also more frequently criticized the questions used in the study (15.7% vs. 9.4%). In particular, many participants wanted an "I don't know" answer choice, as they felt they were randomly guessing in many cases. Some participants also criticised the lack of context for some questions, in particular the rhyme test, which consisted of a single isolated word.

Visually Impaired Participants - Participants with visual impairments commented more frequently on the study's voice choice (18.8% vs. 10.7%), and had richer feedback about the voice. While sighted participants' feedback on the voice was limited to commenting on the the artificial quality of the voice, participants who are visually impaired discussed the difference between different screen readers, and even differences between specific voices within particular screen readers.

Some participants recognized that the study used VoiceOver, and some the particular VoiceOver voice used, for example one remarking "Hello voice of Alex."

Participants who are visually impaired also generally enjoyed the study more than their sighted peers (32.8% vs. 17.9%). It is possible that their relative lack of frustration with the results (discussed above) contributed to this difference. In addition, the subject matters (listening to content on a screen reader) was closer to home for the participants who are visually impaired, likely contributing further to their positive experience. One participant commented, "I love this! Almost like a typing test, but a hearing test for screen reader users!"

Participants who are visually impaired also more frequently pointed out possible confounding variables (10.9% vs. 1.4%). Examples of confounds that participants mentioned included background noises, speaker/device quality, simultaneous audio alerts (e.g., notifications), and hearing loss. The participants did not generally call these factors "confounds", but rather discussed how they could have interfered with a participant's performance. It is likely that unlike sighted participants, participants who are visually impaired were primed to identify these confounds, because they experience them in daily life while using screen readers or otherwise relying on audio information.

Commonalities - Across both groups, there was minimal criticism of the interface (<5% for both interface criticism themes across both groups). However, participants who are visually impaired were more sensitive to issues with lag.

4.10.3 Great-listeners. It is interesting to look at some of the comments that great-listeners had after completing the study. In our group of 56 great-listeners 11 (19.6%) provided feedback, as opposed to 11.4% of all participant who provided feedback. Of the 11 comments, 7 were from visually impaired participants. Six of the 7 expressed enjoyment of the experience. Of the 7, two had a study criticism related to confounding variables and one had interface criticism related to lag. Of the 4 sighted participants, one expressed enjoyment, one had study criticisms related to the yes/no questions, and two had an interface criticism related to lag and audio.

5 DISCUSSION

This work provides the first large, inclusive, online study on the intelligibility of fast, synthetic speech. Our large recruitment demonstrates the availability of volunteers for audio tasks, providing scalability for workflows based on human auditory work, such as real-time captioning. Our large number of participants who are visually impaired highlights the importance of inclusive design. We suggest that future large-scale studies and crowdwork platforms make their platforms and tasks accessible. Online studies and crowdwork could be important ways for visually impaired people to contribute to research as they may have fewer barriers to participating, for example not needing transportation to the study.

Based on the data collected, we presented models of human listening rates, which inform opportunities for conversational agents to tailor speaking rates to users. Overall, we found synthetic speech to be intelligible much faster than normal human spoken rates, suggesting there is room to optimize speaking rate for most users. Visually impaired participants typically understood faster speeds than sighted participants. For this user group, age is nuanced by how much experience they have using synthetic speech, suggesting that with practice and early exposure, the general population might achieve fast listening rates, and save themselves listening time. We also found that content impacts intelligibility at fixed speaking rates, indicating an opportunity for conversational agents to adapt speaking rate to both content and user.

Our analysis of qualitative feedback revealed an over-confidence in many sighted participants. Despite making mistakes on listening questions, sighted participants often complained that the

test incorrectly lowered the speed. Instead of questioning their own listening abilities, they criticized the study. None of the participants who are visually impaired provided such complaints or criticisms. It is possible that because sighted participants are not used to relying solely on their hearing, they underestimate the challenge that relying on listening poses—for example, they may not be aware of just how difficult it is to distinguish rhyming words from one another. In contrast, through their lived experience over years, participants who are visually impaired have likely relied on their hearing much more frequently, including through screen reader use. These experiences would have served not only as practice to hone their listening skills, but also as a humbling experience whenever they made mistakes, an experience their sighted peers never had. Alternatively, it is possible that participants who are visually impaired have uniquely tested the limitations of their auditory skills more consciously, for example by attempting to maximize productivity by setting their screenreader speed to the max that they could understand.

The results also suggest that people who are visually impaired might be better at certain jobs than their sighted counterparts, in particular time-sensitive auditory work. For example, people who are visually impaired might make the best real-time transcribers, stenographers, or translators. Given that many blind people are fast listeners and blind unemployment is high (as in many disabled communities), it might make sense to recruit and train blind workers for these jobs. A precedent exists in Belgium, where blind people were recruited to join the police detective force, and use their superior auditory skills to decipher wiretaps [10, 69]. While those blind detectives were recruited under the suspicion that they would do better auditory detective work, this study provides evidence that people who are visually impaired are faster listeners, which will hopefully encourage further hiring efforts.

If conversational agents and fast listening are the future, then it could be useful to build online training tools to help people become faster listeners. Based on our study results that early adoption of screen readers correlates with faster listening rates, practice during childhood might be particularly effective. Practice during adulthood could also benefit people who become visually impaired later in life (which is more common than congenital blindness), who lack experience with the fast, synthetic speech of screen readers. Tasks similar to those in our study could be used, though the process could also be gamified to engage young children, similar to typing games that teach the player to type faster.

6 LIMITATIONS AND FUTURE WORK

Our study design faces several limitations. The study was run online, so we could not supervise the procedures and were only able to recruit relatively tech-savvy people. Our questions also had limitations. We did not test long passages, and our rhyming tests consisted of individual words devoid of any context, which might not represent real-world use cases of fast synthetic speech. We tested a single synthetic voice, rather than multiple voices. The maximum speed was also capped at the maximum VoiceOver rate. Some participants answered all questions correctly at that rate, so we had no way of measuring their limits. However, this work demonstrated that crowdsourced studies can effectively recruit small elite subpopulations, suggesting that online studies can effectively evaluate the limits of human abilities in future work.

Another potential limitation is the use of binary search to find an appropriate speed for the participant. We chose binary search for its efficiency, so that the study would be short and incentivize broader participation and higher completion rates. However, it is possible that the user experience of binary search was not optimal. In particular, fluke mistakes or over-performance at early speeds has a strong impact on the end speed. During binary search, the speed also jumps around, which several participants commented on, and seems to have been a confusing experience for some participants that made them doubt the ability of our study to gauge their listening rate.

It would be great to exploring alternative methods for efficiently tuning the speed in a smoother fashion.

It is also possible that confounding variables impacted our results. In particular, it is possible that people use screen readers or general text-to-speech software for reasons besides visual impairment. For example, people who experience difficulties reading including dyslexia may consume text auditorally, and this experience may have contributed to higher listening rates. It is also possible that participants had other disabilities that impacted their experience, including hearing impairments, which may have contributed to lower listening rates. In addition, background noise or audio quality more generally may be important to study and control for in future studies. For example, it may be valuable to understand the impact of different levels and types of noise, as well as speaker or device quality. Examining the impact of such variables would be interesting future work.

Ultimately, we envision a world where conversational agents dynamically adapt to their users and surroundings. Such a system could take into consideration a person's baseline listening rate. It could also consider the content being spoken, and information about the surroundings, including background noise level, and whether the user is multitasking while they are listening. For example, a GPS system might speak more slowly during rush-hour traffic, or a screen reader might speed up for easy passages. To dynamically adapt to the user and environment, future studies on people's listening rates that manipulate various parameters are needed.

In particular, exploring the impact of more parameters on intelligibility will be needed to make conversational agents that intelligently adapt speaking rates. In terms of the synthetic voice used, various parameters of synthesis may impact intelligibility. Our study was based on a single synthesized voice (VoiceOver's Alex), but did not explore other synthesized voices that vary in terms of speaker gender, speech clarity, prosody, and other factors. It is possible that different synthesized voices may work better or worse in particular use cases, for example against particular background noises, or for particular users. In terms of the human user, there might be a difference between a person's maximum intelligible rate, which we measured, and their comfortable listening rate. In other words, people might prefer slower rates than what is physically possible. They also might fatigue after listening at a high rate for an extended period of time, needing the conversational agent to adapt. Consequently, maximal sustainable speeds might be lower than what we measured.

Other potential future work using this study as a model could focus on sound localization, contributing to the development of virtual reality and richer sound systems. Like fast listening, sound localization is a task on which people who are visually impaired outperform their sighted peers (e.g., References [19, 27]). A similarly inclusive, online study could shed light on people's abilities to localize various sounds in various environments, learning from the abilities of people who are visually impaired.

7 CONCLUSION

In this work, we presented the first large-scale study of human listening rates, with the aim of informing the optimization of speech rate for conversational agents. By conducting a volunteer-based online study, we were able to reach a larger participant pool than previous studies. By making it accessible, we also reached a larger number of people who are visually impaired, many of whom had experience with fast, synthetic speech. The study results show that people who are visually impaired are typically the fastest listeners, in particular those exposed to screen readers at a young age. These results suggest that in optimizing conversational agent speech rate, an expanded set of speech rates should be considered, as well as tailoring to the individual user and content.

More importantly, this work demonstrates that people with disabilities have incredible abilities and personal experiences that can inspire design, as previous research shows. A main takeaway

of this project is to not view people who are visually impaired primarily as consumers of assistive technologies; rather, recognize that they can inspire new avenues for human-conversational agent interactions. Recognizing important contributions of blind people beyond their necessary perspective for accessibility improvements is an important step toward further integrating blind people into research and design.

REFERENCES

[1] NV Access. 2017. NVDA 2017. Retrieved September 2, 2017 from http://www.nvaccess.org/.

[2] Gerry T. M. Altmann (Ed.). 1995. *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives*. MIT Press.

[3] Amir Amedi. 2004. *Visual and Multisensory Processing and Plasticity in the Human Brain*. Ph.D. Dissertation. Hebrew University, Jerusalem, Israel.

[4] Shumin An, Zhenhua Ling, and Lirong Dai. 2017. Emotional statistical parametric speech synthesis using LSTM-RNNs. In *Proceedings of the 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 1613–1616.

[5] Apple. 2017. VoiceOver. http://www.apple.com/accessibility/mac/vision/. (Accessed 2017-09-02).

[6] Chieko Asakawa, Hironobu Takagi, Shuichi Ino, and Tohru Ifukube. 2003. Maximum listening speeds for the blind. In *Proc. of the International Conference on Auditory Display (ICAD)*. 276–279.

[7] Tal August and Katharina Reinecke. 2019. Pay attention, please: Formal language improves attention in volunteer and paid online experiments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.

[8] Marialena Barouti, Konstantinos Papadopoulos, and Georgios Kouroupetroglou. 2013. Synthetic and natural speech intelligibility in individuals with visual impairments: Effects of experience and presentation rate. In *Proceedings of the European AAATE Conference*. Vilamoura, Portugal, 695–699.

[9] Jeffrey P. Bigham, Anna C. Cavender, Jeremy T. Brudvik, Jacob O. Wobbrock, and Richard E. Ladner. 2007. WebinSitu: A comparative analysis of blind and sighted browsing behavior. In *Proceedings of the International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS)*. ACM, 51–58.

[10] Dan Bilefsky. 2007. In Fight Against Terror, Keen Ears Undistracted by Sight. http://www.nytimes.com/2007/11/17/world/europe/17vanloo.html?mcubz=1.

[11] Alan Black and Nick Campbell. 1995. Optimising selection of units from speech databases for concatenative synthesis. European Speech Communication Association (ESCA), Madrid, Spain, 581–584.

[12] Yevgen Borodin, Jeffrey P. Bigham, Glenn Dausch, and I. V. Ramakrishnan. 2010. More than meets the eye: A survey of screen-reader browsing strategies. In *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility (W4A)*. ACM, Article 13. 1–10.

[13] Danielle Bragg, Cynthia Bennett, Katharina Reinecke, and Richard Ladner. 2018. A large inclusive study of human listening rates. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.

[14] Dasom Choi, Daehyun Kwak, Minji Cho, and Sangsu Lee. 2020. "Nobody speaks that fast!" An empirical study of speech rate in conversational agents for people with vision impairments. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI'20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376569

[15] Ronald A. Cole and Jola Jakimik. 1980. A model of speech perception. *Perception and Production of Fluent Speech* (1980), 133–163.

[16] Juliet Corbin, Anselm Strauss, et al. 2008. Basics of qualitative research: Techniques and procedures for developing grounded theory. (2008).

[17] Delphine Dahan. 2010. The time course of interpretation in speech comprehension. *Current Directions in Psychological Science* 19, 2 (2010), 121–126.

[18] Çağatay Demiralp, Michael S. Bernstein, and Jeffrey Heer. 2014. Learning perceptual kernels for visualization design. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 1933–1942.

[19] Larisa Dunai, Ismael Lengua, Guillermo Peris-Fajarnés, and Fernando Brusola. 2015. Virtual sound localization by blind people. *Archives of Acoustics* 40, 4 (2015), 561–567.

[20] Robert M. Emerson, Rachel I. Fretz, and Linda L. Shaw. 2011. *Writing Ethnographic Fieldnotes*. University of Chicago Press.

[21] Emerson Foulke and Thomas G. Sticht. 1969. Review of research on the intelligibility and comprehension of accelerated speech. *Psychological Bulletin* 72, 1 (1969), 50–62.

[22] M. Furmankiewicz, A. Sołtysik-Piorunkiewicz, and P. Ziuziański. 2014. Artificial intelligence systems for knowledge management in E-health: The study of intelligent software agents. *Latest Trends on Systems* 2 (2014), 551–556.

[23] Laura Germine, Ken Nakayama, Bradley C. Duchaine, Christopher F. Chabris, Garga Chatterjee, and Jeremy B. Wilmer. 2012. Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review* 19, 5 (2012), 847–857.
[24] Google. 2017. ChromeVox Version 52. http://www.chromevox.com/. (Accessed 2017-09-02).
[25] Google. 2017. TalkBack. Retrieved September 3, 2017 from http://play.google.com/store/apps/details?id=com.google.android.marvin.talkback&hl=en.
[26] Frédéric Gougoux, Franco Lepore, Maryse Lassonde, Patrice Voss, Robert J. Zatorre, and Pascal Belin. 2004. Neuropsychology: Pitch discrimination in the early blind. *Nature* 430 (2004), 309–310.
[27] Frédéric Gougoux, Robert J. Zatorre, Maryse Lassonde, Patrice Voss, and Franco Lepore. 2005. A functional neuroimaging study of sound localization: Visual cortex activity predicts performance in early-blind individuals. *PLoS Biol.* 3, 2 (2005), e27.
[28] Arthur C. Graesser. 2016. Conversations with AutoTutor help students learn. *Int. J. Artif. Intell. Educ.* 26, 1 (2016), 124–132.
[29] João Guerreiro and Daniel Gonçalves. 2015. Faster text-to-speeches: Enhancing blind people's information scanning with faster concurrent speech. In *Proceedings of the International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS'15)*. ACM, 3–11.
[30] Roy H. Hamilton, Alvaro Pascual-Leone, and Gottfried Schlaug. 2004. Absolute pitch in blind musicians. *Neuroreport* 15, 5 (2004), 803–806.
[31] Jeffrey Heer and Michael Bostock. 2010. Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'10)*. ACM, 203–212.
[32] John J. Horton, David G. Rand, and Richard J. Zeckhauser. 2011. The online labyoratory: Conducting experiments in a real labor market. *Exp. Econ.* 14, 3 (2011), 399–425.
[33] Kirsten Hötting and Brigitte Röder. 2009. Auditory and auditory-tactile processing in congenitally blind humans. *Hear. Res.* 258, 1 (2009), 165–174.
[34] Arthur S. House, Carl Williams, Michael H. L. Hecker, and Karl D. Kryter. 1963. Psychoacoustic speech tests: A modified rhyme test. *J. Acoust. Soc. Am.* 35, 11 (1963), 1899–1899.
[35] Tim Hull and Heather Mason. 1995. Performance of blind children on digit-span tests. *J. Vis. Impair. Blind.* 89, 2 (1995), 166–169.
[36] Akemi Iida, Nick Campbell, Fumito Higuchi, and Michiaki Yasumura. 2003. A corpus-based speech synthesis system with emotion. *Speech Commun.* 40, 1–2 (2003), 161–187.
[37] Kenzo Ishizaka and James L. Flanagan. 1972. Synthesis of voiced sounds from a two-mass model of the vocal cords. *Bell Labs Techn. J.* 51, 6 (1972), 1233–1268.
[38] Andrew J. Kolarik, Silvia Cirstea, Shahina Pardhan, and Brian C. J. Moore. 2014. A summary of research investigating echolocation abilities of blind and sighted humans. *Hear. Res.* 310 (2014), 60–68.
[39] Walter S. Lasecki, Raja Kushalnagar, and Jeffrey P. Bigham. 2014. Legion scribe: Real-time captioning by non-experts. In *Proceedings of the International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS'14)*. ACM, 303–304.
[40] Jonathan Lazar, Aaron Allen, Jason Kleinman, and Chris Malarkey. 2007. What frustrates screen reader users on the web: A study of 100 blind users. *Int. J. Hum.-Comput. Interact.* 22, 3 (2007), 247–269.
[41] Qisheng Li, Krzysztof Z. Gajos, and Katharina Reinecke. 2018. Volunteer-based online studies with older adults and people with disabilities. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS'18)*. Association for Computing Machinery, New York, NY, 229–241. https://doi.org/10.1145/3234695.3236360
[42] Qisheng Li, Sung Jun Joo, Jason D. Yeatman, and Katharina Reinecke. 2020. controlling for participantsâ.. viewing distance in large-scale, psychophysical online experiments using a virtual chinrest. *Sci. Rep.* 10, 1 (2020), 1–11.
[43] William D. Marslen-Wilson. 1987. Functional parallelism in spoken word-recognition. *Cognition* 25, 1 (1987), 71–102.
[44] William D. Marslen-Wilson and Alan Welsh. 1978. Processing interactions and lexical access during word recognition in continuous speech. *Cogn. Psychol.* 10, 1 (1978), 29–63.
[45] James L. McClelland and Jeffrey L. Elman. 1986. The TRACE model of speech perception. *Cogn. Psychol.* 18, 1 (1986), 1–86.
[46] Chris McKinstry, Rick Dale, and Michael J. Spivey. 2008. Action dynamics reveal parallel competition in decision making. *Psychol. Sci.* 19, 1 (2008), 22–24.
[47] Helena Merriman. 2016. The Blind Boy Who Learned to See with Sound. Retrieved from http://www.bbc.com/news/disability-35550768.
[48] G. W. Micro. 2017. Window-Eyes. September 2, 2017 http://www.gwmicro.com/Window-Eyes/.

[49] Norman Miller, Geoffrey Maruyama, Rex J. Beaber, and Keith Valone. 1976. Speed of speech and persuasion. *J. Pers. Soc. Psychol.* 34, 4 (1976), 615–624.
[50] Anja Moos and Jürgen Trouvain. 2007. Comprehension of ultra-fast speech–blind vs. "normally hearing" persons. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS'07)*. Saarland University Saarbrücken, Germany, 677–680.
[51] Eric Moulines and Francis Charpentier. 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Commun.* 9, 5–6 (1990), 453–467.
[52] NYU. 2015. Beyond Braille: A History of Reading by Ear. Retrieved from http://www.nyu.edu/about/news-publications/news/2015/january/mara-mills-blind-reading.html.
[53] Gabriele Paolacci, Jesse Chandler, and Panagiotis G. Ipeirotis. 2010. Running experiments on Amazon mechanical turk. *Judg. Decis. Mak.* 5, 5 (2010), 411–419.
[54] Konstantinos Papadopoulos and Eleni Koustriava. 2015. Comprehension of synthetic and natural speech: Differences among Sighted and visually impaired young adults. *Proceedings of the International Conference on Enabling Access for Persons with Visual Impairment (ICEAPVI'15)*, 147–151.
[55] Donatella Pascolini and Silvio Paolo Mariotti. 2012. Global estimates of visual impairment: 2010. *Br. J. Ophthalmol.* 96, 5 (2012), 614–618.
[56] Ville Pulkki and Matti Karjalainen. 2015. *Communication Acoustics: An Introduction to Speech, Audio and Psychoacoustics*. John Wiley & Sons.
[57] Katharina Reinecke and Krzysztof Z. Gajos. 2015. LabintheWild: Conducting large-scale online experiments with uncompensated samples. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW'15)*. ACM, 1364–1378.
[58] Brigitte Röder, Lisa Demuth, Judith Streb, and Frank Rösler. 2003. Semantic and morpho-syntactic priming in auditory word recognition in congenitally blind adults. *Lang. Cogn. Process.* 18, 1 (2003), 1–20.
[59] Brigitte Röder and Frank Rösler. 2003. Memory for environmental sounds in sighted, congenitally blind and late blind adults: Evidence for cross-modal compensation. *Int. J. Psychophysiol.* 50, 1 (2003), 27–39.
[60] Brigitte Röder, Frank Rösler, Erwin Hennighausen, and Fritz Näcker. 1996. Event-related potentials during auditory and somatosensory discrimination in sighted and blind human subjects. *Cogn. Brain Res.* 4, 2 (1996), 77–93.
[61] Brigitte Röder, Oliver Stock, Siegfried Bien, Helen Neville, and Frank Rösler. 2002. Speech processing activates visual cortex in congenitally blind humans. *Eur. J. Neurosci.* 16, 5 (2002), 930–936.
[62] Brigitte Roder, Wolfgang Teder-Salejarvi, Anette Sterr, Frank Rosler, et al. 1999. Improved auditory spatial tuning in blind humans. *Nature* 400, 6740 (1999), 162.
[63] David A. Ross, Ingrid R. Olson, and John C. Gore. 2003. Cortical plasticity in an early blind musician: An fMRl study. *Magn. Reson. Imag.* 21, 7 (2003), 821–828.
[64] Marc Schröder. 2001. Emotional speech synthesis: A review. In *Proceedings of the 7th European Conference on Speech Communication and Technology*.
[65] Diemo Schwarz, Grégory Beller, Bruno Verbrugghe, and Sam Britton. 2006. Real-time corpus-based concatenative synthesis with CataRT. In *Proceedings of the International Conference on Digital Audio Effects (DAFx'06)*. 279–282.
[66] Freedom Scientific. 2006. JAWS 18. Retrieved September 2, 2017 from http://www.freedomscientific.com/.
[67] Celia Scully. 1990. Articulatory synthesis. In *Speech Production and Speech Modelling*. Springer, 151–186.
[68] Christine H. Shadle and Robert I. Damper. 2001. Prospects for articulatory synthesis: A position paper. In *Proceedings of the 4th ISCA Tutorial and Research Workshop (ITRW'01) on Speech Synthesis*.
[69] Claire Soares. 2007. Move over Poirot: Belgium Recruits Blind Detectives to Help Fight Crime. Retrieved from http://www.independent.co.uk/news/world/europe/move-over-poirot-belgium-recruits-blind-detectives-to-help-fight-crime-5337339.html.
[70] Jose Sotelo, Soroush Mehri, Kundan Kumar, Joao Felipe Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio. 2017. Char2wav: End-to-end speech synthesis. In *Proceedings of the International Conference on Learning Representations (ICLR'17) Workshop Submission*. Retrieved from https://openreview.net/forum?id=B1VWyySKx.
[71] Amanda Stent, Ann Syrdal, and Taniya Mishra. 2011. On the intelligibility of fast synthesized speech for individuals with early-onset blindness. In *Proceedings of the International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS'11)*. ACM, 211–218.
[72] Santani Teng, Amrita Puri, and David Whitney. 2012. Ultrafine spatial acuity of blind expert human echolocators. *Exp. Brain Res.* 216, 4 (2012), 483–488.
[73] Hugo Théoret, Lotfi Merabet, and Alvaro Pascual-Leone. 2004. Behavioral and neuroplastic changes in the blind: Evidence for functionally relevant cross-modal interactions. *J. Physiol.-Par.* 98, 1 (2004), 221–233.
[74] Tomoki Toda and Keiichi Tokuda. 2007. A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. *IEICE Trans. Inf. Syst.* 90, 5 (2007), 816–824.

[75] Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. 2000. Speech parameter generation algorithms for HMM-based speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 3. IEEE, 1315–1318.

[76] Jürgen Trouvain. 2007. On the comprehension of extremely fast synthetic speech. *Saarland Working Papers in Linguistics,* Vol. 1, 5–13.

[77] Aditya Vashistha, Pooja Sethi, and Richard Anderson. 2017. Respeak: A voice-based, crowd-powered speech transcription system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'17)*. ACM, 1855–1866.

[78] Patrice Voss, Maryse Lassonde, Frederic Gougoux, Madeleine Fortin, Jean-Paul Guillemot, and Franco Lepore. 2004. Early- and late-onset blind individuals show supra-normal auditory abilities in far-space. *Curr. Biol.* 14, 19 (2004), 1734–1738.

[79] Catherine Y. Wan, Amanda G. Wood, David C. Reutens, and Sarah J. Wilson. 2010. Early but not late-blindness leads to enhanced auditory perception. *Neuropsychologia* 48, 1 (2010), 344–348.

[80] Robert Weeks, Barry Horwitz, Ali Aziz-Sultan, Biao Tian, C. Mark Wessinger, Leonardo G. Cohen, Mark Hallett, and Josef P. Rauschecker. 2000. A positron emission tomographic study of auditory localization in the congenitally blind. *J. Neurosci.* 20, 7 (2000), 2664–2672.

[81] Teng Ye, Katharina Reinecke, and Lionel P. Robert Jr. 2017. Personalized feedback versus money: The effect on reliability of subjective data in online experimental platforms. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 343–346.

[82] Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. 1999. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Proceedings of the 6th European Conference on Speech Communication and Technology*.

[83] Heiga Ze, Andrew Senior, and Mike Schuster. 2013. Statistical parametric speech synthesis using deep neural networks. In *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 7962–7966.

[84] Kathryn Zyskowski, Meredith Ringel Morris, Jeffrey P. Bigham, Mary L. Gray, and Shaun K. Kane. 2015. Accessible crowdwork? Understanding the value in and challenge of microtask employment for people with disabilities. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW'15)*. ACM, 1682–1693.

Received December 2020; revised April 2021; accepted April 2021