# A Large Inclusive Study of Human Listening Rates

**Danielle Bragg**[1] **Cynthia Bennett**[2] **Katharina Reinecke**[1] **Richard Ladner**[1]

[1]Computer Science & Engineering [2]Human Centered Design & Engineering
University of Washington – Seattle, WA
{dkbragg,bennec3,reinecke,ladner}@cs.washington.edu

## ABSTRACT

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 the first large-scale study on human listening rates. 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 blind and low-vision people, 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.

### ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

### Author Keywords

Synthetic Speech; Listening Rate; Human Abilities; Accessibility; Blind; Visually Impaired; Crowdsourcing

## 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 [16] and education [21]. Despite popular focus on the artificial intelligence powering these 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.



*CHI 2018,* April 21–26, 2018, Montreal, QC, Canada



ISBN 978-1-4503-5620-6/18/04…$15.00

DOI: https://doi.org/10.1145/3173574.3174018

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.

While synthetic speech is new to many people using conversational agents, people with visual impairments have a long history of accessing text with audio. The National Library Service has been recording and distributing recorded books to blind and low-vision Americans since the 1930's [43], 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 [10]. Because visually impaired people 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 blind and low-vision 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 [47]. In particular, by requiring participants to travel to the study location, in-lab experiments often exclude people with visual impairments and other disabilities, due to inaccessibility of study locations.

In this paper, we present the first large-scale study on human listening rates, with attention to how visual impairment informs listening rate. Its 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 blind and low-vision 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:

- We conduct the first large, inclusive, online study on human listening rates with 453 volunteer participants, demonstrating the feasibility of attaining volunteer crowdworkers for audio tasks, including people with disabilities.
- 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, visually impaired participants 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.

## 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 visually impaired people 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.

### 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 [46].

Several psychophysical models exist for how the brain converts audio signals to words [2]. Some models center around segmenting sounds into words (e.g., [35, 11]). 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 [34], and TRACE [36] 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., [12]). While our understanding of speech processing has advanced significantly, psychoacoustics is still an active research area.

### 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., [9, 54, 42]). 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., [30]) or speech acoustics (e.g., [60]).

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). 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).

The blind and low-vision community has a longer experience with synthetic speech. Screen readers, which emit synthetic speech, are this group's most popular assistive technology [33]. 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 [18], JAWS [55], NVDA [1], TalkBack [19], VoiceOver [4], and Window-Eyes [39] (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 [10], compared to a normal speaking rate of 120-180 words per minute [40].

### Listening Abilities of People with Visual Impairments

People with visual impairments, and in particular congenitally blind people, 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., [20]), and are more likely to have perfect pitch, the ability to identify absolute sound frequencies (e.g., [23, 53]). Blind

people are also typically better at sound localization [65, 52], and process auditory stimuli faster [50]. Some blind people also use echolocation to understand their surroundings [31]. Experts can even ride bikes without hitting obstacles [38] and achieve spatial resolution comparable to peripheral vision [58]. Blind people excel at high-level cognitive functions as well, including processing words and sentences faster (e.g., [48]), and remembering auditory stimuli (e.g., [28]).

It is possible that these blind "superabilities" result from blind people's brains processing information differently from sighted people's brains [59]. 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 [51, 67, 59]. 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, 26].

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,. [66]). 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., [49]). 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.

### Listening Rate Studies

Past studies on human listening rates are small,[1] and have not always included blind or low-vision listeners (see [14]). More recently a push has been made to include people with visual impairments, given their extensive use of text-to-speech (e.g., [5]). Since then, studies have compared sighted and visually impaired listeners (e.g., [6]). Studies have also compared the intelligibility of speech produced by different mechanisms, including natural speech, formant synthesis, and concatenative synthesis (e.g., [41]), and compared efficiency of single vs multi-track speech (e.g., [22]).

These studies have employed diverse methods for assessing listening rate. Comprehension questions (e.g., [45]), word identification tasks (e.g., [6]), and transcription or repetition tasks (e.g., [57, 5]) have been used. Some studies also use subjective metrics (e.g., [41, 61, 5]). Choice of test materials and questions is important, as using even different lengths of text can lead to different conclusions [14]. 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 blind or visually impaired people can comprehend speech at faster rates (e.g., [57, 41, 61]). However, other studies have found no significant difference between these groups (e.g., [45]). Evidence that other factors, such as age and practice, impacts listening abilities has also emerged (e.g., [57]). 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.

[1]Max participants: visually impaired 36 [57], sighted 65 [45].

### 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., [25, 44, 17]). Past studies have generally focused on visual perception, for example evaluating shape and color similarity [13] or visualization techniques [24]. More recently, the development of crowdsourced transcription systems demonstrates that audio tasks can also be effectively crowdsourced (e.g., Legion:Scribe [32] and Respeak [62]). 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 [47], 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 (e.g., [63, 64, 29, 15]). In addition, experiments conducted on LabintheWild have been shown to accurately replicate the results of controlled laboratory studies [47]. In this work, we extend the space of volunteer-based crowdsourced experiments to include studies with auditory tasks.

## STUDY

To help inform the optimization of speaking rates for conversational agents, we conducted a short (5-10 min) online study on LabintheWild to evaluate the intelligibility of fast, synthetic speech. 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 with visual impairments process higher synthetic speaking rates than sighted people, and if so, to what extent does practice with screen readers account for this superior ability?

We crowdsourced the study to reach a larger participant pool than previous lab experiments, and to facilitate participation by blind and low-vision participants, who often face obstacles to participating in lab studies due to transportation inaccessibility. The online study was made fully accessible to include people with visual impairments and other disabilities, who are often excluded from crowd work [68].

### 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 [27], 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 [37], 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.")

### 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.

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 1a). The recording could only be played once. Once the recording finished, they were given a question about the audio they just heard (Figure 1b-d). Participants answered three practice questions, one for each question type, followed by eighteen questions used to measure listening rate. The set of eighteen 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, they advanced to a faster speed.

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: $\frac{1}{2}$ if correct, 0 else; rhyming: $\frac{5}{6}$ if correct, 0 else; transcription: $max(0, 1 - \frac{dist(s_{target}, s_{guess})}{len(s_{target})})$ where $s_{target}$ is the spoken text, $s_{guess}$ is the transcription, and

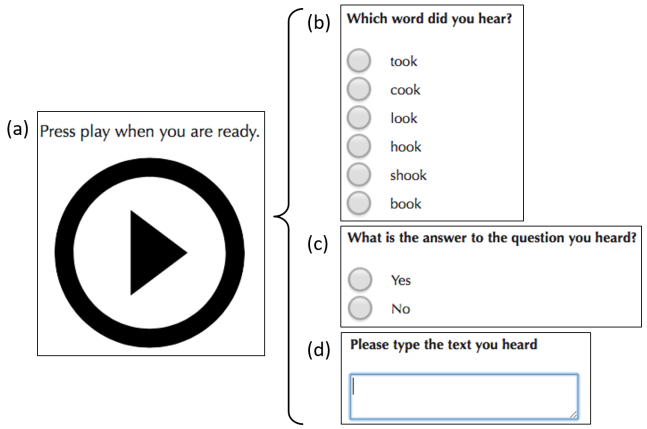


Figure 1: Screen shots of the listening question interface. (a) The prompt for playing a listening question. (b-d) The subsequent question asked about the audio played, (b) for a rhyme test, (c) for a yes/no question, and (d) for transcription.

$dist(a,b)$ is the edit distance between strings a and b.[2] Intuitively, this last quantity approximates the fraction of audio that was transcribed correctly. The threshold for advancing was 4.17 (out of $4.\overline{6}$), 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 help them 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 with visual impairments. This feedback provided education and self-awareness, which served as motivation and compensation for participation.

### 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. 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[3], 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

[2]Our edit distance was Levenshtein distance, and punctuation was removed and capitalization ignored during computation.

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

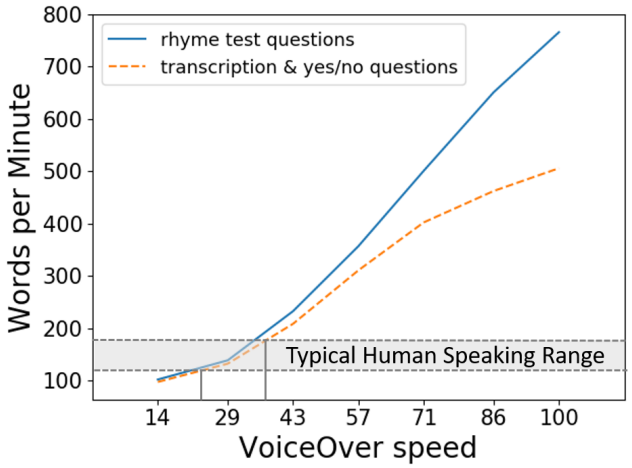


Figure 2: VoiceOver speeds translated into words per minute, for the rhyme test questions (words) and for the transcription and yes/no questions (sentences). Typical human speaking rate 120-180 WPM corresponds to VoiceOver range 24-38.

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: $WPM = \frac{\#\text{ letters}}{5 \text{ x time (min)}}$.

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

### 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 blind and low-vision 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.

### 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., [5]), 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 with visual impairments 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.

### Participants

The study was launched on LabintheWild with IRB approval. It was online for two months, during which 453 participants completed the study. Recruitment occurred through the LabintheWild site, Facebook posts, relevant email lists targeting screen reader users, and word-of-mouth. The completion rate was 74%. Basic participant demographics were:

- Age: 8-80, m=34, sd=15
- Gender: 257 (57%) female, 194 (43%) male, 2 (<1%) other.
- Vision status: 310 (68%) sighted, 143 (32%) visually impaired – 101 (71%) blind, 23 (16%) low-vision, 9 (6%) other, 10 (7%) undisclosed impairment.
- First language: 354 (78%) English, 99 (22%) other.

## 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.

### 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 28.5% of participants falling in this range. The mean Listening Rate was 56.8, which corresponds to 309 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.

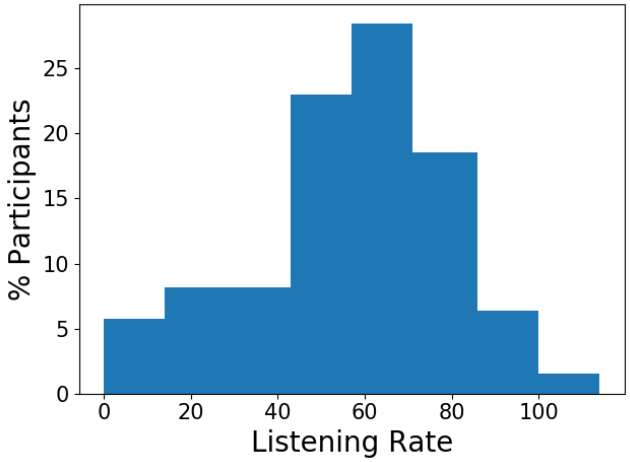


Figure 3: Histogram of Listening Rates for all participants.

### Factors Impacting Listening Rate - Overall Population

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: 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, visually impaired people have the opportunity to use technology and screen readers from a young age, unlike older generations, which could impact Listening Rate. Table 1 provides the results of the linear regression model that minimized information loss.

| Variable | Est. | SE | t | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 54.99 | 3.24 | 17.00 | <.001 | *** |
| VI [yes] | 26.30 | 5.23 | 5.03 | <.001 | *** |
| Age | -0.18 | 0.08 | -2.14 | 0.033 | * |
| Age × VI [yes] | -0.56 | 0.13 | -4.15 | <.001 | *** |
| Native English [yes] | 7.79 | 2.48 | 3.14 | 0.002 | ** |

Table 1: Linear regression predicting Listening Rate for all participants from demographic variables. Abbreviations: VI visually impaired, Est. estimate, SE standard error. Significance codes: *** < .001, ** < .01, * < .05

The model shows that being visually impaired significantly impacts Listening Rate, increasing the predicted rate by 26.30. Age is also significant, though less so, with every year of age, the Listening Rate decreasing by .18. The interaction between age and visual impairment is strongly 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 significant positive effect, increasing Listening Rate by 7.78. This model explains 12% of the variance in people's Listening Rates (multiple and adjusted $R^2 = .12$).

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 60.6 (334 WPM) while for sighted participants it was 55.1 (297 WPM). The difference between these groups is statistically significant ($t(451) = 2.4014$, $p = .0167$).

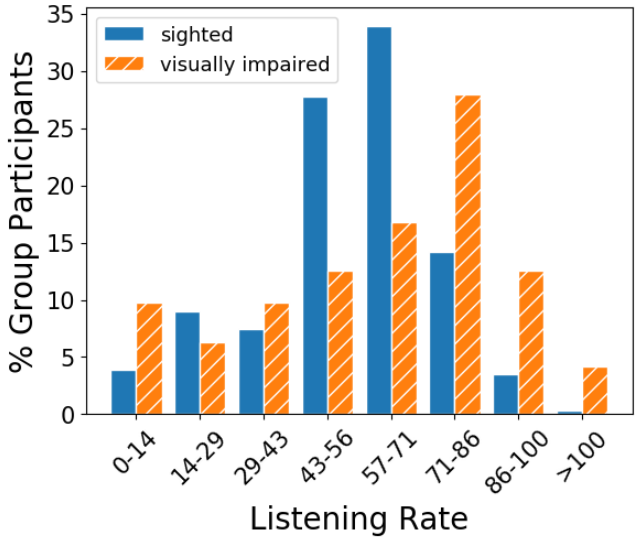


Figure 4: Histogram of Listening Rates, separated into visually impaired and sighted participant groups.

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 visually impaired 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. Furthermore, age correlates significantly ($p < 0.05$) with lower Listening Rates for visually impaired participants ($r = -0.439, p < 0.0001$), but not for sighted participants ($r = -0.094, p = 0.098$).

### Factors Impacting Listening Rate - Visually Impaired Population

To better understand participants with visual impairments, 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,

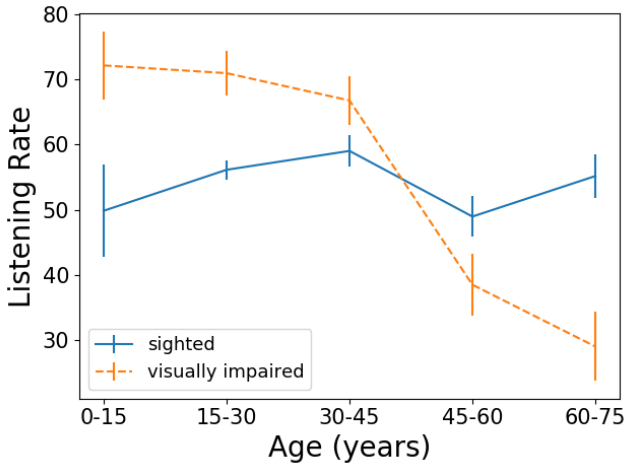


Figure 5: Plot of age vs. Listening Rate, for visually impaired and sighted groups.

and compared models using AIC to determine which factors to include. The same factors were explored: 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 2 provides the model that minimized information loss.

| Variable | Est. | SE | t | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 58.12 | 7.04 | 8.26 | <.001 | *** |
| Age | -0.21 | 0.17 | -1.23 | 0.222 | |
| SR Years | 2.91 | 0.54 | 5.40 | <.001 | *** |
| Age × SR Years | -0.06 | 0.01 | -5.13 | <.001 | *** |

Table 2: Linear regression predicting Listening Rate for participants with visual impairments from demographic variables. Abbreviations: SR screen reader, Est. estimate, SE standard error. Significance codes: *** < .001, ** < .01, * < .05

This model indicates that for every year of screen reader usage, we expect an increase in Listening Rate of 2.91. 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 visually impaired people use screen readers. Age and screen reader years are also correlated ($r = .389, p < .001$), meaning that by including age in the overall model, it captured some information about screen reader usage as well. This model accounts for 34% of the variance in the visually impaired population, (multiple $R^2 = 0.34$, adjusted $R^2 = 0.31$), compared to the overall model's 12%, suggesting that it is a substantially better model for this subpopulation.

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., [66]), 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. As shown in Figure 6, adoption age does correlate with both age ($r = .801, p < .001$) and Listening Rate ($r = -.421, p < .001$). These correlations suggests 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.

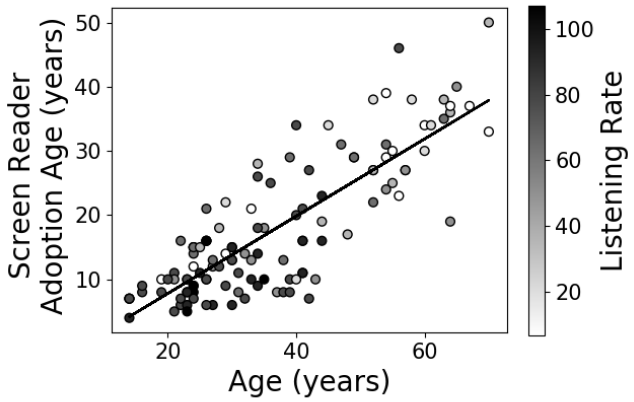


Figure 6: Scatterplot of age (years) vs. screen reader adoption age (years), with Listening Rate encoded in color, for our 123 visually impaired participants who use screen readers.

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 ($t(115) = -.390$, $p = .697$). Our 123 screen reader users reported using screen readers in the following numbers: 95 VoiceOver, 86 JAWS, 74 NVDA, 26 TalkBack, 8 Window Eyes, 8 Other.[4]

**Super-listeners: The Top 1.5%**

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, 7 (1.5%) participants fell into this group. Six out of the seven were blind, representing 4% of visually impaired participants and 6% of blind participants, compared to <1% of the sighted population. To find out more about these super-listeners, we looked at their demographics. This group was generally young (all aged 23-35), blind (all but one), male (all but one), native English speakers (all but one). Note, however, that we cannot infer generalizability of these findings given the small N.

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]Note that the sum exceeds total participants, as many participants used multiple screen readers.

### 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 (86% accuracy), followed by rhyme tests (84%), and transcription (82%).[5] We suspected that question length might impact accuracy due to fatigue, and found statistically significant correlations ($p < 0.05$) between accuracy and question length for transcription ($r = -0.039, p = 0.042$) and yes/no questions ($r = -0.054, p = 0.005$). This difference in intelligibility for different question types and lengths suggests that different speaking rates are appropriate for different auditory interactions, and suggests room to optimize speaking rate for conversational agents based on both the participant and the content.

### 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, 89% of sighted and 88.5% of visually impaired participants answered "no," with no statistical significance between the groups, as computed by a chi-squared independence test ($X^2 < .01, p = .99$). The difference in Listening Rates between the groups who answered "yes" vs. "no" was not statistically significant ($t(451) = -1.09, p = .27$).

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 [7]. 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 visually impaired people typically have higher Listening Rates.

## 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 with visual impairments 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 blind and low-vision 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.

The results also suggest that people with visual impairments might be better at certain jobs than their sighted counterparts, in particular time-sensitive auditory work. For example, visually impaired people 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 [8, 56]. While those blind detectives were recruited under the suspicion that they would do better auditory detective work, this study provides evidence that people with visual impairments are faster listeners, which will hopefully encourage further hiring efforts.

If conversational agents and fast listening are the future, 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.

## 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

[5] Transcription accuracy was computed by the metric from the study procedure, edit distance divided by target string length.

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.

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. For example, 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 with visual impairments outperform their sighted peers. 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 with visual impairments.

## 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 with visual impairments, many of whom had experience with fast, synthetic speech. The study results show that people with visual impairments 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 which can inspire design, as previous research shows. A main takeaway of this project is to not view people with visual impairments 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.


## ACKNOWLEDGEMENTS

We thank Daniel Snitkovskiy for development work on the study, and note NSF grants IIS-1651487 and IIS-1702751.


## REFERENCES


1. NV Access. 2017. NVDA 2017. `http://www.nvaccess.org/`. (2017). (Accessed 2017-09-02).
2. Gerry TM 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*. PhD thesis. Hebrew University, Jerusalem, Israel.
4. Apple. 2017. VoiceOver. `http://www.apple.com/accessibility/mac/vision/`. (2017). (Accessed 2017-09-02).
5. 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.
6. 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 *European AAATE Conference*. Vilamoura, Portugal, 695–699.
7. 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.
8. 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`. (November 2007).
9. 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.
10. Yevgen Borodin, Jeffrey P Bigham, Glenn Dausch, and IV 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.
11. Ronald A Cole and Jola Jakimik. 1980. A Model of Speech Perception. *Perception and Production of Fluent Speech* (1980), 133–163.

12. Delphine Dahan. 2010. The Time Course of Interpretation in Speech Comprehension. *Current Directions in Psychological Science* 19, 2 (2010), 121–126.
13. 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.
14. 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.
15. Wikimedia Foundation. 2001. Wikipedia. `http://www.wikipedia.org/`. (2001). (Accessed 2017-09-03).
16. 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.
17. 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.
18. Google. 2017a. ChromeVox Version 52. `http://www.chromevox.com/`. (2017). (Accessed 2017-09-02).
19. Google. 2017b. TalkBack. `http://play.google.com/store/apps/details?id=com.google.android.marvin.talkback&hl=en`. (2017). (Accessed 2017-09-03).
20. 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.
21. Arthur C Graesser. 2016. Conversations with AutoTutor Help Students Learn. *International Journal of Artificial Intelligence in Education* 26, 1 (2016), 124–132.
22. 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)*. ACM, 3–11.
23. Roy H Hamilton, Alvaro Pascual-Leone, and Gottfried Schlaug. 2004. Absolute Pitch in Blind Musicians. *Neuroreport* 15, 5 (2004), 803–806.
24. 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)*. ACM, 203–212.
25. John J Horton, David G Rand, and Richard J Zeckhauser. 2011. The Online Labyoratory: Conducting Experiments in a Real Labor Market. *Experimental Economics* 14, 3 (2011), 399–425.
26. Kirsten Hötting and Brigitte Röder. 2009. Auditory and Auditory-Tactile Processing in Congenitally Blind Humans. *Hearing Research* 258, 1 (2009), 165–174.
27. Arthur S. House, Carl Williams, Michael H.L. Hecker, and Karl D. Kryter. 1963. Psychoacoustic Speech Tests: A Modified Rhyme Test. *The Journal of the Acoustical Society of America* 35, 11 (1963), 1899–1899.
28. Tim Hull and Heather Mason. 1995. Performance of Blind Children on Digit-Span Tests. *Journal of Visual Impairment & Blindness* 89, 2 (1995), 166–169.
29. Stack Exchange Inc. 2008. Stack Overflow. `http://stackoverflow.com/`. (2008). (Accessed 2017-09-03).
30. Kenzo Ishizaka and James L Flanagan. 1972. Synthesis of Voiced Sounds From a Two-Mass Model of the Vocal Cords. *Bell Labs Technical Journal* 51, 6 (1972), 1233–1268.
31. Andrew J Kolarik, Silvia Cirstea, Shahina Pardhan, and Brian CJ Moore. 2014. A Summary of Research Investigating Echolocation Abilities of Blind and Sighted Humans. *Hearing Research* 310 (2014), 60–68.
32. 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)*. ACM, 303–304.
33. Jonathan Lazar, Aaron Allen, Jason Kleinman, and Chris Malarkey. 2007. What Frustrates Screen Reader Users on the Web: A Study of 100 Blind Users. *International Journal of Human-Computer Interaction* 22, 3 (2007), 247–269.
34. William D Marslen-Wilson. 1987. Functional Parallelism in Spoken Word-Recognition. *Cognition* 25, 1 (1987), 71–102.
35. William D Marslen-Wilson and Alan Welsh. 1978. Processing Interactions and Lexical Access During Word Recognition in Continuous Speech. *Cognitive Psychology* 10, 1 (1978), 29–63.
36. James L McClelland and Jeffrey L Elman. 1986. The TRACE Model of Speech Perception. *Cognitive Psychology* 18, 1 (1986), 1–86.
37. Chris McKinstry, Rick Dale, and Michael J Spivey. 2008. Action Dynamics Reveal Parallel Competition in Decision Making. *Psychological Science* 19, 1 (2008), 22–24.
38. Helena Merriman. 2016. The Blind Boy Who Learned to See with Sound. `http://www.bbc.com/news/disability-35550768`. (February 2016).

39. GW Micro. 2017. Window-Eyes. http://www.gwmicro.com/Window-Eyes/. (2017). (Accessed 2017-09-02).

40. Norman Miller, Geoffrey Maruyama, Rex J Beaber, and Keith Valone. 1976. Speed of Speech and Persuasion. *Journal of Personality and Social Psychology* 34, 4 (1976), 615–624.

41. 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)*. Saarland University Saarbrücken, Germany, 677–680.

42. Eric Moulines and Francis Charpentier. 1990. Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones. *Speech Communication* 9, 5-6 (1990), 453–467.

43. NYU. 2015. Beyond Braille: A History of Reading by Ear. http://www.nyu.edu/about/news-publications/news/2015/january/mara-mills-blind-reading.html. (January 2015).

44. Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. Running Experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5, 5 (2010), 411–419.

45. 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)* (2015), 147–151.

46. Ville Pulkki and Matti Karjalainen. 2015. *Communication Acoustics: An Introduction to Speech, Audio and Psychoacoustics*. John Wiley & Sons.

47. 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)*. ACM, 1364–1378.

48. Brigitte Röder, Lisa Demuth, Judith Streb, and Frank Rösler. 2003. Semantic and Morpho-Syntactic Priming in Auditory Word Recognition in Congenitally Blind Adults. *Language and Cognitive Processes* 18, 1 (2003), 1–20.

49. Brigitte Röder and Frank Rösler. 2003. Memory for Environmental Sounds in Sighted, Congenitally Blind and Late Blind Adults: Evidence for Cross-Modal Compensation. *International Journal of Psychophysiology* 50, 1 (2003), 27–39.

50. Brigitte Röder, Frank Rösler, Erwin Hennighausen, and Fritz Näcker. 1996. Event-Related Potentials During Auditory and Somatosensory Discrimination in Sighted and Blind Human Subjects. *Cognitive Brain Research* 4, 2 (1996), 77–93.

51. Brigitte Röder, Oliver Stock, Siegfried Bien, Helen Neville, and Frank Rösler. 2002. Speech Processing Activates Visual Cortex in Congenitally Blind Humans. *European Journal of Neuroscience* 16, 5 (2002), 930–936.

52. Brigitte Roder, Wolfgang Teder-Salejarvi, Anette Sterr, Frank Rosler, and others. 1999. Improved auditory spatial tuning in blind humans. *Nature* 400, 6740 (1999), 162.

53. David A Ross, Ingrid R Olson, and John C Gore. 2003. Cortical plasticity in an early blind musician: an fMRl study. *Magnetic resonance imaging* 21, 7 (2003), 821–828.

54. 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)*. 279–282.

55. Freedom Scientific. 2006. JAWS 18. http://www.freedomscientific.com/. (2006). (Accessed 2017-09-02).

56. Claire Soares. 2007. Move over Poirot: Belgium Recruits Blind Detectives to Help Fight Crime. http://www.independent.co.uk/news/world/europe/move-over-poirot-belgium-recruits-blind-detectives-to-help-fight-crime-5337339.html. (November 2007).

57. 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)*. ACM, 211–218.

58. Santani Teng, Amrita Puri, and David Whitney. 2012. Ultrafine Spatial Acuity of Blind Expert Human Echolocators. *Experimental Brain Research* 216, 4 (2012), 483–488.

59. Hugo Théoret, Lotfi Merabet, and Alvaro Pascual-Leone. 2004. Behavioral and Neuroplastic Changes in the Blind: Evidence for Functionally Relevant Cross-Modal Interactions. *Journal of Physiology-Paris* 98, 1 (2004), 221–233.

60. 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.

61. Jürgen Trouvain. 2007. On the Comprehension of Extremely Fast Synthetic Speech. *Saarland Working Papers in Linguistics (SWPL)* 1 (2007), 5–13.

62. 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)*. ACM, 1855–1866.

63. Luis von Ahn. 2013. Duolingo: Learn a Language for Free While Helping to Translate the Web. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI)*. ACM, 1–2.

64. Luis Von Ahn and Laura Dabbish. 2004. Labeling Images with a Computer Game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 319–326.

65. 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. *Current Biology* 14, 19 (2004), 1734–1738.

66. 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.

67. 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. *Journal of Neuroscience* 20, 7 (2000), 2664–2672.

68. 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)*. ACM, 1682–1693.