Choice of Voices: A Large-Scale Evaluation of Text-to-Speech Voice Quality for Long-Form Content

Julia Cambre* **Jessica Colnago***
Carnegie Mellon University, Pittsburgh, PA, USA
{jcambre, jcolnago}@cmu.edu

Jim Maddock*
Northwestern University, Evanston, IL, USA
maddock@u.northwestern.edu

Janice Tsai **Jofish Kaye**
Mozilla Corporation, Mountain View, CA, USA
{jtsai, jofish}@mozilla.com

ABSTRACT
The advancement of text-to-speech (TTS) voices and a rise of commercial TTS platforms allow people to easily experience TTS voices across a variety of technologies, applications, and form factors. As such, we evaluated TTS voices for *long-form content*: not individual words or sentences, but voices that are pleasant to listen to for several minutes at a time. We introduce a method using a crowdsourcing platform and an online survey to evaluate voices based on listening experience, perception of clarity and quality, and comprehension. We evaluated 18 TTS voices, three human voices, and a text-only control condition. We found that TTS voices are close to rivaling human voices, yet no single voice outperforms the others across all evaluation dimensions. We conclude with considerations for selecting text-to-speech voices for long-form content.

Author Keywords
voice quality; text-to-speech; TTS; voice interface; synthesized speech; long-form; listening experience

CCS Concepts
•**Human-centered computing** → **Natural language interfaces;** *Sound-based input / output;*

INTRODUCTION
"I'm sorry, Dave, I'm afraid I can't do that," said HAL, the voice interface-based computer in the movie *2001: A Space Odyssey*. In 1968, the movie's creators imagined that by 2001, we would live in a world where computers would be speaking to us with human-like voices. Instead, it would not be until 2011, after almost a half century's worth of work, that we would start to see the mainstream adoption of synthesized voices, primarily in voice assistants (VAs) such as Apple's Siri, Amazon's Alexa, and the Google Assistant [43]. Now, this capability to create speech-enabled products is available to anyone via commercial TTS platforms.[1] The epoch of everyday interaction with TTS voices is here.

Concurrently, with the availability of audio-based online content (e.g. news articles, audiobooks, and podcasts), the consumption of spoken audio has increased [18]. The percentage of the United States population aged 12 and older listening to online audio continues to increase, with an estimated 67%, or 167 million people, listening to online audio content monthly in 2019. This audio consumption is driven by increases in listening for both podcasts (51% of the population) and audiobooks (50% of the population) in the United States [18]. There are similar audio-consumption trends outside the US [45]. In this vein, new technology allows people to listen to articles using TTS voices [39,51]. Furthermore, some publishers, such as The Atlantic,[2] Medium,[3] and Wired[4] are making professionally narrated versions of articles available for people's listening pleasure [42]. It can be assumed that more online content will be provided by TTS voices, as long as the experience can be a pleasant and enjoyable one. A key dependency for this growth, especially for long-form content, will be the voice of the narration [23, 27].

Despite robust tests and metrics that exist for assessing the quality of TTS voices at the word, sentence, and paragraph level [49], and significant work around evaluation of voices generated from a single dataset [5], there is a gap in understanding how to consistently evaluate TTS voices reading long-form content such as news articles or audiobooks, which require a voice that is pleasant to listen to for several minutes at a time. In this paper, we contribute a method that will inform others in selecting a voice for long-form content. We demonstrate its use through a large-scale evaluation of 18 TTS voices, three human voices, and a text-only control condition.

* Authors contributed equally to this paper and completed this work during internships at Mozilla.

CHI '20, April 25–30, 2020, Honolulu, HI, USA.

ACM 978-1-4503-6708-0/20/04
https://doi.org/10.1145/3313831.3376789

[1] These platforms include Amazon Polly: https://aws.amazon.com/polly/ and Google Cloud Text to Speech: https://cloud.google.com/text-to-speech/.
[2] The Atlantic Audio Articles. https://www.theatlantic.com/podcasts/audio-articles/
[3] Medium Audio. https://medium.com/topic/audio
[4] Wired. https://www.wired.com/

BACKGROUND

We build upon and contribute to a long history of HCI and speech synthesis research on voice quality and how it shapes interaction dynamics. Though we focus on evaluating text-to-speech (TTS) voices for long-form content, here, we consider a broader range of related work that has informed our own.

Voice shapes the user experience

The voice that an interface takes on has a profound effect on the user experience. Over several decades of research in HCI, speech technology, communication, and related fields, a wide range of studies have demonstrated how the quality and characteristics suggested by a voice—whether it is a pre-recorded human voice or a text-to-speech (TTS) voice—changes the dynamics of an interaction [9, 36, 46].

Prior work found that users are quick to infer features like personality traits from a system's voice, and often show "similarity attraction" towards a voice interface that mirrors their own in introversion/extroversion. In one study, researchers created introvert and extrovert versions of a TTS voice by manipulating parameters like the voice's volume, speaking rate, and frequency range [37]. In the experiment, participants listened to book descriptions and were asked to evaluate the book and voice. Participants who listened to voice that matched their own in introversion or extroversion were more likely to perceive it as likeable and credible, and expressed a stronger intention to buy the book [37]. A more recent study in the design of a car navigation interface also found that drivers preferred personality traits that matched their own [7]. A large body of work has found similar effects for other voice features such as gender and accent (see [36] and [11] for a review), suggesting, for example, that voices may activate gender stereotypes [32, 38, 48] and influence the perceived credibility of a voice-based agent [1, 16], and that these perceptions may change depending on how human-like a voice sounds [3]. Importantly, how these voice features affect the user experience is highly dependent on contextual factors such as the domain or functionality of the system, its culture of use, and its embodiment [9, 46], highlighting the value of finding a voice that is well-aligned with a particular use case.

Furthermore, overall voice quality can significantly impact the user experience. While synthesized voices impose a higher cognitive load on listeners relative to natural speech [21], TTS voices that have lower intelligibility (i.e. that are more difficult to understand) are even more taxing [20]. These quality concerns can also affect user's satisfaction. In one study, participants consistently rated voices as more likeable the more human-like (or natural) the voices sounded [4]. Another study compared the experience of talking with a human partner and with two popular voice assistants, Alexa and Siri, to understand perceptions of humanness [17]. Participants remarked that the voices of Alexa and Siri were easy to understand (highly intelligible), but lacked a sense of expressiveness or emotion, particularly in comparison to a human speaker. By contrast, in a different study, participants who rarely use voice assistants felt that Siri's voice was human-like and reflected attempts to incorporate cultural cues and personality [14].

Using a voice that is too human-like may backfire due to a mismatch of expectations and functionality. In studies of users' experiences with today's common voice assistants, researchers found that the human-like names, personalities, and voices the assistants take on mislead users into expecting human-like intelligence [14, 31]. In light of this, some have argued for using intentionally robotic-sounding voices to better match the technology's capabilities [2, 34, 35]. Given this paper's focus on evaluating voices that narrate long-form content (e.g. for an article or audiobook), we use the *expressivity* and *naturalness* of human speech as our baseline. However, other applications of TTS that take on more agent-like qualities (e.g. engaging a user in conversation) may require other evaluation criteria.

Approaches to evaluating voice quality

Within the speech synthesis community, TTS voices are typically evaluated using a subjective listening test in which listeners are presented with samples of synthesized speech and asked to rate them along dimensions such as clarity and overall quality of the experience [5,19,26,49]. While there are no well-defined best practices around TTS evaluation, a recent review of speech synthesis evaluation [49] notes that the current state-of-the-art is largely driven by 1990s standards established to evaluate telephone-based systems. From this tradition, perhaps the most common subjective evaluation measure is the Mean Opinion Score (MOS), which asks participants to rate their overall impression of a voice on a scale from 1 to 5 [5, 12, 50]. We see MOS as a useful benchmark metric given its simplicity and widespread use; but as others have suggested [12, 24, 49], we believe it is not sufficient for evaluating quality for long-form listening purposes. Our evaluation of the voices in this paper incorporates the MOS as one among several measures used to evaluate voice quality.

Other strategies for evaluating synthesized speech focus on its intelligibility (i.e., how easy it is to accurately hear the content of the speech). Tests of intelligibility often present listeners with sentences that are semantically meaningless, and ask them to transcribe what they heard as accurately as possible [49]. In recent years, however, synthesized speech technology has improved to the point that intelligibility is nearing ceiling performance [21, 26, 49], suggesting that these tests of intelligibility are perhaps no longer necessary or informative. In this study, we instead consider other behavioral measures of voice quality such as *comprehension* and *perception of voice speed*; this shift allows us to consider whether the words come across clearly, as well as whether the voice affects users' attention and understanding across a longer listening experience.

While these evaluation methods are common practice, it has been noted that current approaches are limited, and concerns have been raised about the rigor and reliability of speech evaluation studies [50]. In an analysis of papers from the 2014 Interspeech conference, a major conference within the language technology community, researchers found that the majority of TTS evaluation studies used sample sizes that were too small to draw reliable conclusions: 60% of studies based their conclusions on ratings from fewer than 20 participants. However, according to their re-analysis of a large-scale listening evalu-

ation data set (the Blizzard Challenge 2013), they found that studies involving a MOS test need at least 30 participants and a varied set of test sentences to yield reliable conclusions [50].

Another concern relates to the ecological validity of TTS evaluation practices. In most cases, evaluations of synthesized voices lack context [33, 50], are tested in unrealistically ideal circumstances with high quality listening equipment and no noise [26], and involve generic tasks that do not reflect the real-world use cases [33]. In one attempt to address this issue, researchers explored an alternative approach to TTS evaluation by immersing participants in an interactive scenario in which they engage in dialogue with an embodied virtual agent [33]. Their study found that it is feasible to evaluate TTS voices in an interactive, realistic context in which participants only hear one voice. However, the effect sizes are smaller compared to a traditional listening test in which each participant hears all voices with unrelated content.

These concerns around context-specificity underscore the importance of evaluating voice quality for the specific use-case of long-form content. Long-form content poses distinct challenges from TTS voices intended for brief dialogue or notifications—the context of a sentence may affect how it should be read, where the emphasis should be placed, etc.

Our work is not the first to address the specific challenges of evaluating TTS voices for long-form content. The Blizzard Challenge [5], an annual competition within the speech synthesis community, brings together researchers to develop and evaluate voices generated based on a common underlying audio and text dataset (typically recordings of audiobooks). Participants in the challenge have developed automated models to evaluate synthetic voices [40] based off the same corpus of audio and text, and designed a questionnaire with 11 scales specifically tailored to audiobook listening [24]. More generally, the approaches that the Blizzard Challenge takes towards voice evaluation for long-form content are complementary to our approach in this work: the Blizzard Challenge is largely focused on setting a level playing field by using a common set of voice training data and isolating which particular speech synthesis techniques are most promising. The challenge structures their listening test as a within-subjects task (each participant listens to all voices in a given challenge year with a variety of sentences or paragraphs) with a participant population that includes volunteer and paid student listeners as well as many speech experts. In contrast, we consider the space of commercially available voices and evaluate them in a more naturalistic scenario, listening to a full long-form article with a given voice, using a study population of paid Mechanical Turk workers.

Finally, recent work considered three ways of presenting listeners with long-form content while keeping the tested speech in context and minimizing demands on cognitive load [12]. They found that content presentation style affects results on measures like the MOS, with listeners yielding different MOS ratings when they evaluate individual sentences compared to evaluating the paragraph as a whole.

Our work builds upon these efforts at voice evaluation through a large-scale evaluation of voices—exceeding the recommended number of listeners per voice [50] and comparing across a wide range of human and synthetic voices. We set our study in a realistic use-case (listening to a long-form article) and leverage a varied set of evaluation metrics that goes beyond MOS, introducing other measures relevant to long-form listening such as voice speed and comprehension.

Voice	WPM	Gender	Date	n	M:F
Human 3	182	M	19-Apr	55	1.1
Human 1	137	M	19-Apr	49	1.8
Judy W 1	146	F	19-Jul	41	0.9
Judy W 2	163	F	19-Jul	51	0.9
Google C	172	F	18-Aug	47	1.4
Windows 2	159	M	19-Apr	38	1.5
Mac Default	174	M	19-Apr	66	1.1
Polly Matthew	192	M	18-Aug	46	1.2
Polly Sally	173	F	18-Aug	52	1.2
Judy GL 1	146	F	19-Jul	56	1.4
Windows 1	162	F	19-Apr	46	1.6
Voicery Nichole	177	F	18-Aug	57	1.9
Human 2	183	F	19-Apr	49	2.0
Polly Joanna	187	F	18-Aug	50	2.1
Google A	176	M	18-Aug	51	2.3
Nancy 2	200	F	19-Apr	41	3.0
Judy GL 2	163	F	19-Jul	50	0.9
Nancy 1	174	F	19-Jan	43	1.7
LJ Speech	145	F	18-Aug	50	1.3
Android UK	153	M	19-Jan	48	1.8
iOS	189	F	19-Jan	51	1.4

Table 1. Summary statistics for each voice condition, listed in descending order by Mean Opinion Score (see Table 2). Columns represent the voice name (anonymized for human voices), words per minute, gender of the voice, date (month and year) that the voice was captured, number of participants assigned to the voice condition who completed the full task on MTurk, and the gender ratio (male:female) of participants.

METHOD

In this research, we developed and implemented a method to evaluate text-to-speech (TTS) voices reading long-form content. We compared 18 TTS voices, three human voices, and a text-only control condition.

Selecting Voices

We selected 18 voices from a variety of platforms: commercially available TTS platforms,[5] desktop and mobile operating systems,[6] and some proprietary voices made by Mozilla.[7] In general, we chose the default male and female, US-based voices for each platform/OS, except where indicated. For the commercially available TTS platforms, we used their developer tools to generate the audio files. For the OS-based voices, we used built-in screen-reader software to generate the audio. Lastly, for human voices, we selected members of our research team that offered a range of characteristics of interest, in particular gender and accent (Human 3, British accent). See Table 1 for details on the voices we studied, and refer

[5] Amazon Polly, Google, and Voicery.
[6] Android, iOS, Mac, and Windows.
[7] Judy, LJ Speech, and Nancy.

to https://ttschoice.github.io for voice clips and additional features.

Comparison

We compared 18 computer-generated TTS voices and three human voices on two high-level aspects: how pleasant the listening experience was, and how it impacted listening comprehension. We also included a control condition where participants read the content, rather than listened.

We conducted a survey-based between-subjects experiment using Amazon Mechanical Turk (MTurk) from mid-2018 to mid-2019. We used MTurk to reach a large sample of participants across the United States. We recruited 1095 U.S.-based MTurkers who were paid, on average, $2.50 for an approximately 10-minute task. We selected MTurkers who had completed 1,000 HITs and had an approval rate of over 95%. The survey was approved by Mozilla's review process and is available as an online appendix.

We developed a reproducible survey-based methodology to evaluate voice quality for long-form content. The survey consisted of 4 activities: (1) listening to the audio recording, (2) answering questions about the quality of the listening experience, (3) answering questions about the article to measure comprehension, and (4) answering demographic questions.

Listening to Audio

At the start of the survey, participants were randomly assigned to one of the voice (or text-only, control) conditions. We presented them with an audio clip with the assigned voice reading an article. The article was kept constant across voices to avoid variation due to article characteristics, like length, structure, or topic. When generating the audio clips, we only used the plain text of the article, keeping its existing punctuation and paragraph breaks—we did not introduce SSML tags to add emphasis, correct pronunciation errors, or otherwise alter the generated audio. The human-read recordings were voiced by members of our research team, none of whom have had professional voice acting experience. We include samples of each voice at https://ttschoice.github.io.

We selected the article "Reduce Your Stress in Two Minutes a Day" [22] as it is a well structured article, not too long to be tiresome but long enough to provide a range of different speech patterns (909 words), and its content is politically neutral. Note that participants were not provided with the text or URL for the article, except in the text-only control condition, where they were provided with the text of the article. Participants were able to pause the audio and move on to the next section of the survey at any time—the last value on the audio counter was considered their "audio completion time." However, we required that participants listened to at least 10 seconds of the audio to be considered in our analysis. This led us to eliminate 5 of our 1090 participants from our analysis.

Listening experience

Participants rated their listening experience on a 5-point Likert item question, from "Excellent" to "Very poor," and selected how likely they were to listen to their favorite book, magazine, or podcast using that voice on a scale of 0 to 10. An 11-point scale was, as opposed to a 5-point scale, as we believed this variable warranted a finer grained response, and because 0-10 feels like a more natural scale in this case. To help us better understand what makes some voices better than others, participants rated the voices' speed (5-point Likert from "Much too fast" to "Much too slow") and select which voice characteristics were true (binary option): voice is monotone, sounds natural, is easy to comprehend, lacks emotion or personality, and allows the listener to focus on the content. In the control condition, we only asked about their reading experience.

Text comprehension

Participants answered six text comprehension questions that focused on specific sections of the text. These questions were multiple choice and generated based on the content of the article. For example:

> What were the sources of stress for Bill?
>
> - His tense relationship with his in-laws
> - His job at a major tech company
> - His constant need for success
> - His relationship with his wife

We piloted the questions with a small group to test clarity and difficulty before being deployed. Our text-based control condition allows us to compute a baseline of the general comprehension of the article.

Demographics

To understand the characteristics of our participants and compare across groups, we asked demographic questions that covered age and gender. We further asked participants about the type of device they used to complete the survey and, when applicable, if they used headphones to listen to the audio clip, as it could impact their listening experience and comprehension.

DATA & ANALYSIS OVERVIEW

We analyzed data from 1090 participants who completed our study and passed the check for minimum amount of audio listened. Approximately 41% (n=444) self-identified as female, 58% (n=630) as male, and 1.5% (n=16) told us that neither of these categories described them, or they declined to state. Our participants mainly had headphones on while answering the survey (71%, n=769), with 25% (n=268) not wearing headphones, and about 5% (n=53) who did not specify. We aimed at having approximately 50 participants for each voice condition. The random assignment of participants to voice conditions led the groups to have between 38 and 66 participants, with a median of 50. Table 1 shows the number of participants who completed each condition and the gender breakdown between these groups. We did not find statistically significant differences when comparing our groups based on age and gender (One way ANOVA, $F = 0.965$, $df = 21$, $p > 0.05$ and Pearson's Chi-squared test, $\chi^2 = 43.167$, $df = 42$, $p > 0.05$ respectively).

Listening Experience

For our general reporting of participants' overall experience, we collapsed "Good" and "Excellent" into a single "Positive"

category, and "Poor" and "Very Poor" into a single "Negative" category. We reported the percentage of participants in each condition who gave each voice a certain rating, allowing us to compare ratings across conditions with varying numbers of participants. Collapsing these ratings into positive and negative categories allows for clearer explanation of these ratings at a descriptive level. There is a clear conceptual difference between "good," "neutral," and "poor," while the differences between "good" and "excellent" are less clear. This is exacerbated by the between-subjects study design, as different participants will likely have different conceptions of "good" and "excellent."

We generated a Mean Opinion Score (MOS) for each voice by converting each experience rating to a numeric value between 1 and 5 (1 = very poor, 5 = excellent) and computing the average for each voice condition. We chose a 5 point scale given its standard presentation and use in related work. Finally, we computed and reported the median score of whether participants would listen to a given voice again. We tested the independence of responses across each voice condition using a Kruskal-Wallis rank sum test for experience ratings and whether participants would listen again, and a one-way ANOVA for MOS rankings.

To better understand what contributed to overall listening experience, we asked participants whether they agreed with 5 statements about voice quality: (1) The voice was easy to comprehend, (2) The voice was monotone, (3) The voice sounded natural, (4) The voice lacked emotion/personality, and (5) I could focus on the content. We reversed responses to questions 2 and 4 in order to make each statement positive. Using Principal Component Analysis (PCA) we determined that these questions reduced to two factors,[8] which roughly correspond to a voice's *clarity* and a voice's *quality*. Questions 1 and 5 contributed to the clarity factor, while questions 2, 3, and 4 contributed to the quality factor. All of the items loaded onto a single factor with a loading of .6 or higher. These factors explained 49% of the variance.

To understand the relationship between demographics, voice features, and listening experience, we constructed an ordinal logistic regression model, where our dependent variable is the listener's experience rating, and one listener-response is our unit of analysis. For this analysis we used the original 5-point experience rating scale instead of collapsing our positive and negative categories.

Model 1 predicts listening experience using voice speed ($Speed_v$), voice type (TTS_V), voice gender ($Male_V$), participant gender ($Male_L$), and participant headphone use ($Headphones_L$). $Male_V$ and $Male_L$ are binary variables that indicate that the voice or the listener identified as male, respectively. TTS_V is a binary variable that indicates that the voice was generated by a TTS algorithm, not read by a human. $Headphones_L$ is a binary variable that indicates that the participant used headphones to listen to the audio.

We interacted voice gender with participant gender to understand whether listeners of certain genders prefer certain TTS voice genders. We included a second order polynomial term for voice speed to account for a curviliniar relationship between speed and listening experience (e.g., there might be a "just right speed" where both faster and slower voices provide a worse listening experience). More formally:

$$\begin{aligned} Rating_{Experience} = \beta_1 &+ \beta_2 TTS_V \\ &+ \beta_3 Speed_V + \beta_4 Speed_V{}^2 + \\ &+ \beta_5 Male_V + \beta_6 Male_L \\ &+ \beta_7 Headphones_V \\ &+ \beta_8 Male_V * Male_L + \varepsilon \end{aligned}$$

Sixteen of our 1090 participants did not report their gender or chose "These choices do not describe me." We ran our experience model with 2 additional variables $NonBinary_L$ and $NonBinary_L * Male_V$, but neither variable was statistically significant and our point estimates remained stable. We therefore omitted $NonBinary_L$ and $NonBinary_L * Male_V$ from the model to ease comprehension.

Voice Speed

Similar to the experience ratings, we collapsed speed ratings into 3 categories: "too fast," "just right," and "too slow." Again, we reported the percentage of participants in each condition who gave each voice a certain rating, allowing us to compare ratings across conditions with varying numbers of participants.

We constructed a simple ordinal logistic regression model to explore the relationship between participants' speed ratings and the speed of the voice. We predicted a participant's speed rating with a single independent variable: voice speed, measured in words per minute. Since our outcome variable is non-monotonic, the center of the scale represents a positive response while the two edges represent negative responses, we transformed our speed ratings into a monotonic variable by collapsing it into three categories. "Too fast" and "too slow" become "poor", while "much too fast" and "much too slow" become "very poor". As in Model 1, we included a second order polynomial term for voice speed to account for a curviliniar relationship between speed and listening experience:

$$\begin{aligned} Rating_{Speed} = \beta_1 &+ \beta_2 Speed_V \\ &+ \beta_3 Speed_V{}^2 + \varepsilon \end{aligned}$$

Comprehension

We computed reading comprehension grades based on how many of the comprehension questions each participant answered correctly. These grades were normalized based on the number of questions each participant would be able to answer given the time at which they paused the audio playback. We conducted a one-way ANOVA to determine whether grades across voice conditions were statistically significant, and Tukey Honest Significant Differences to determine whether differences in grades were statistically significant between TTS voice, human voice, and text-only conditions.

RESULTS

In this research, we aimed to understand how different TTS voices compared when participants listened to long-form online content. Specifically, we analyzed participants' subjective

[8] A scree plot suggested two factors were optimal.

Voice	MOS	(sd)	Listen Again	Quality	(sd)	Clarity	(sd)	Grade	(sd)
Human 3	4.2	(0.7)	8	84	(26)	78	(33)	5.6	(2.4)
Text only	4.0	(0.7)						5.3	(1.7)
Human 1	3.9	(0.7)	7	63	(42)	67	(40)	5.1	(2.6)
Judy W 1	3.9	(1.0)	7	48	(41)	59	(42)	5.1	(2.3)
Judy W 2	3.7	(0.9)	6	44	(41)	51	(42)	5.2	(2.5)
Google C	3.7	(0.9)	7	40	(38)	55	(39)	4.3	(2.6)
Windows 2	3.7	(1.0)	7	44	(40)	39	(37)	4.1	(2.8)
Mac Default	3.7	(1.0)	7	42	(38)	39	(37)	4.2	(2.1)
Polly Matthew	3.6	(0.9)	5	29	(35)	43	(42)	4.5	(2.7)
Polly Sally	3.5	(0.9)	4	29	(34)	52	(42)	5.4	(2.6)
Judy GL 1	3.5	(1.0)	3	30	(39)	47	(44)	5.2	(2.3)
Windows 1	3.5	(1.0)	5	42	(42)	49	(40)	4.3	(2.7)
Voicery Nichole	3.5	(0.9)	6	39	(40)	46	(41)	4.9	(2.6)
Human 2	3.4	(0.9)	3	76	(33)	58	(41)	5.9	(2.1)
Polly Joanna	3.4	(1.0)	4	31	(37)	51	(42)	5.0	(2.7)
Google A	3.4	(0.9)	4	38	(38)	47	(39)	5.1	(3.0)
Nancy 2	3.4	(0.9)	2	29	(33)	51	(45)	5.0	(2.3)
Judy GL 2	3.3	(1.0)	3	25	(33)	48	(45)	6.1	(2.3)
Nancy 1	3.3	(1.0)	3	23	(31)	55	(41)	5.5	(2.4)
LJ Speech	3.2	(1.1)	2	30	(37)	20	(34)	4.6	(2.6)
Android UK	2.9	(1.1)	1	22	(30)	25	(41)	5.6	(2.2)
iOS	2.8	(1.1)	0	10	(19)	26	(37)	5.2	(2.3)

Table 2. Experience ratings for each voice condition. Columns represent Mean Opinion Score (1–5 where 5 is best), the median score of whether participants would listen to a given voice again (0–10), mean quality (0–100), mean clarity (0–100), and the mean comprehension grade (0–10).

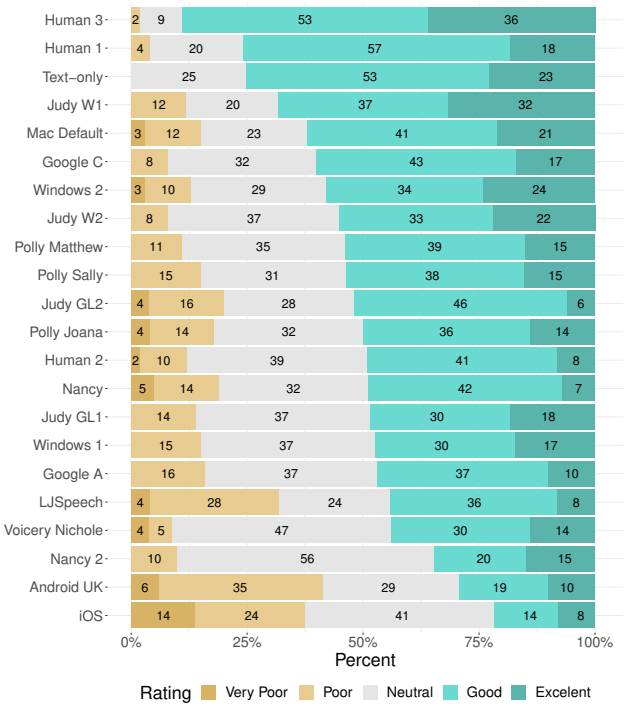

Figure 1. Percentage of positive, neutral, and negative listening experience ratings for each voice, ordered by positive ratings.

responses to TTS voices. While prior work has focused on overall quality and intelligibility (e.g. whether listeners could understand the words spoken by a synthetic voice), we further aim to understand whether listeners found different voices more pleasant to listen to, and listeners' ability to comprehend and retain information from an article.

Listening Experience

Participants reported a wide range of listening experiences to various TTS voices. Overall, Judy W1 and Mac Default received the highest percentage of positive ratings of any TTS voice (68% and 62%), while Android UK and iOS had the highest percentage of negative ratings (42% and 37%). Using a Kruskal-Wallis rank sum test, we determined that these differences in opinions were statistically significant ($\chi^2 = 113.06, df = 21, p < .001$). We report the listening experience ratings for all human and TTS voices in Figure 1.

By converting our listening experience ratings to Mean Opinion Scores (MOS) we see a slight change in the rankings. This difference is due to the collapse of the two positive categories (and negatives) into a single variable. We determined that these differences in rankings are statistically significant (Wilcoxon signed rank test, $Z = 231, p < .001$). Two human voices, Human 1 and Human 3, still outperformed all TTS voices (MOS equal to 4.2 and 3.9, respectively). Of the TTS voices, Judy W1 still received the highest ranking (MOS = 3.9), but Mac Default fell to 5th for TTS voices (MOS = 3.7) behind Judy W2, Google C, and Windows 2.

The difference between the percent of positive ratings and MOS ranking indicates that Mac Default listeners were more polarized. Mac Default received both a higher percent of positive and negative ratings than other TTS voices (e.g. Judy W2), and fewer neutral ratings. Android UK and iOS both received the lowest MOS (2.9 and 2.8, respectively). These

Variable	Odds Ratio	Coef	S.E.	P-Value
$Speed_V$	0.12	-2.09	1.83	0.253
$Speed_V^2$	0.03	-3.64	1.84	0.048*
TTS_V	0.44	-0.81	0.17	0.000*
$Male_L$	1.00	-0.00	0.14	0.973
$Male_V$	1.54	0.43	0.19	0.025*
$Headphones_L$	1.48	0.39	0.13	0.003*
$Male_V * Male_L$	0.73	-0.31	0.25	0.208
$Very.poor\|Poor$	0.02	-4.11	0.30	0.000*
$Poor\|OK$	0.13	-2.06	0.23	0.000*
$OK\|Good$	0.67	-0.41	0.22	0.062
$Good\|Excellent$	4.17	1.43	0.22	0.000*

Table 3. Ordinal Logistic regression on experience ratings ($Rating_{Experience}$). Raw coefficients and odds ratios are reported with standard errors and p values. * values indicate that coefficients are significant to $p < .05$. Values below the horizontal rule (e.g. $Very.poor|Poor$) represent intercepts.

differences between MOS scores are statistically significant (one-way ANOVA, $F = 6.163$, $df = 21$, $p < .001$) and a full report of MOS and other quality metrics are listed in Table 2.

By comparing human and TTS voices, we found that the highest rated TTS voices still performed slightly worse than some human voices. For example, Judy W1 received 21% fewer positive ratings than the highest rated human reader. On the other hand, Judy W1 performed better than the lowest rated human voice, receiving 19% more positive ratings and the same percentage (12%) of negative ratings. Ten TTS voices received a higher percentage of positive ratings than the lowest rated human voice, and eight TTS voices received fewer negative ratings than the lowest rated human voice.

The highest rated TTS voices also performed slightly worse than the text-only condition. While 76% of readers rated the text-only condition as a positive experience, only 68% listeners rated the best TTS voice, Judy W1, as positive. No readers gave the text-only condition a negative rating, but 12% of listeners gave Judy W1 a negative rating. On the other hand, Human 1 had the same percentage of positive ratings as the text-only condition, and listening to Human 3 was perceived as more positive than reading the text (89% vs 76%).

We observed similar results as to whether participants would listen to a specific voice again on a 0 to 10 scale (Table 2). The top TTS voices (Judy W1, Mac Default, Google C, and Windows 2) received a median rating of 7, while the lowest rated TTS Voices (Android UK and iOS) received median ratings of 1 and 0. The highest rated human voice still performed slightly better than the highest rated TTS voices. A Kruskal-Wallis test determined differences across voice conditions to be statistically significant ($\chi^2 = 142.14$, $df = 20$, $p < .001$).

Demographics

Model 1 aims to understand the relationship between voice features, listener characteristics, and listening experience. We present a full regression table in Table 3. Due to the ordinal logistic regression model specification, raw coefficients represent log odds. For interpretability, we present both raw coefficients and the odds ratio for each variable. We calculate a p-value for each coefficient by comparing the t-value against the standard normal distribution.

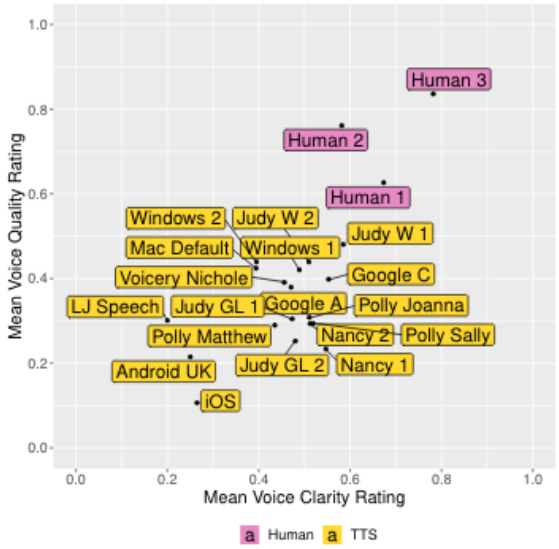

Figure 2. Comparison of clarity and quality ratings.

Results from Model 1 indicate that voice speed, whether the voice was human or TTS, voice gender, and whether the listener was wearing headphone are related to differences in experience ratings.[9] Participants were 54% *more* likely to give a higher experience rating to a male voice than a female voice. They were also 56% *less* likely give a higher rating to a TTS (non-human) voice than a human voice. Participant were also 48% more likely to give a voice a higher rating when listening to the audio through headphones then those not wearing headphones. These results control for several exogenous factors. For instance, male voices were more likely to receive positive ratings, regardless of whether the voice was generated by a human reader or a TTS algorithm.

Neither the participant's self-identified gender nor the interaction between participant gender and voice gender were statistically significant. This indicates that we did not observe a relationship between the gender of the participant and experience ratings (e.g. men and women were both equally as likely to assign a positive rating). We also did not observe a relationship between participant gender, voice gender, and experience ratings (e.g. men were not more likely to assign positive ratings to male voices).

Clarity and Quality

We examined the correlation between voice clarity and voice quality, and noticed that the correlation between these two constructs is relatively low: 0.20 on the full set of TTS and human voices, and 0.10 after we removed the human voices. This indicates some voices are easily understood but still unpleasant, other voices are pleasant to listen to but are not easy to understand, with a few excelling along both dimensions—generally the human voices. For instance, of the 18 TTS voices, Voicery

[9]These coefficients were all statistically significant to $p < .05$

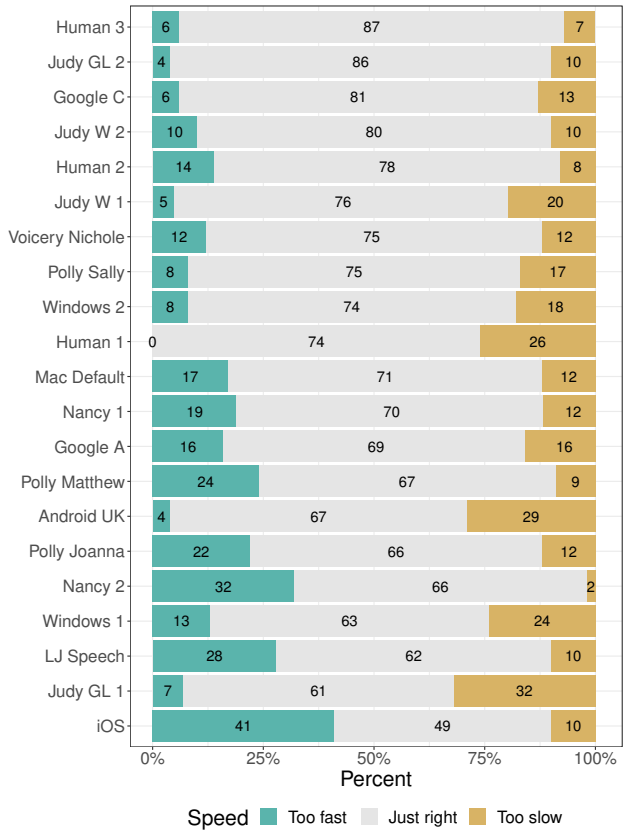

Figure 3. Percentage of "too fast," "just right," and "too slow" Speed ratings for each voice ordered by "just right" scores.

Nichole is ranked fifth in voice quality but eight in clarity, and Windows 2 is ranked second in clarity, but tenth in quality. Conversely, Judy W1 was ranked first in both quality and clarity, while iOS was ranked lowest ranking voice for quality and third lowest in clarity, respectively. In both categories, human voices performed better than TTS voices, though one human reader was ranked slightly behind Judy W1 in clarity. We conducted a Wilcoxon signed rank test and found rank differences between clarity and quality to be statistically significant ($Z = 196, p < .05$).

Voice Speed

Overall, most voices were rated "just right," and voices rarely received ratings skewed heavily towards "too fast" or "too slow." Again, the best paced human voice, Human 3, received slightly more "just right" ratings than the best TTS voice under the speed metric, Judy GL2. Notably, while Judy W1 received the most positive listening experience ratings, it is the fourth TTS voice in speed ratings (Figure 3).

We observed a relationship between voice speed (measured in words per minute) and user experience ratings, as well as voice speed and user speed ratings. However, we see that in both Models 1 and 2, the coefficient is fairly small (increase of 3% and 1%, respectively). Nevertheless, the statistically significant squared term indicates a curvilinear relationship and the negative signs indicate that this relationship is concave.

Variable	Odds Ratio	Coef	S.E.	P-Value
$Speed_V$	0.73	-0.31	2.13	0.883
$Speed_V^2$	0.01	-4.86	2.12	0.022*
$Very.poor \vert Poor$	0.03	-3.56	0.19	0.000*
$poor \vert Just.right$	0.40	-0.91	0.07	0.000*

Table 4. Ordinal Logistic regression on speed ratings ($Rating_{Speed}$). Raw coefficients and odds ratios are reported with standard errors and p values. * values indicate that coefficients are significant to $p < .05$. Values below the horizontal rule (e.g. $Very.poor \vert Poor$) represent intercepts.

When we omitted the quadratic voice speed term from Models 1 and 2, using a single linear term instead, the voice speed variable was not statistically significant. Generally, these results indicated that there is a "just right speed" in the range of 163 to 177 words per minute (WPM) where both faster and slower voices provide a worse listening experience. We present the results of Model 2 in Table 4.

Audio completion

During the first portion of our survey we tracked when people stopped listening to the audio clip, moving on to the listening experience questions in this survey. While this data is not perfectly clean in a simulated environment, as we address in the limitations section, this data allows us to have an idea when a voice became "too unbearable" for the listener that they chose to stop it. We calculated a mean for listening completion for each voice. As expected, given our data collection environment (MTurk) we see fairly high percentages of audio completion for most voices. The 5 participants that did not make the 10 seconds cut where distributed among: iOS Female, Google A, Google C (x2), and Polly Matthew. However, we also observe a few voices with lower audio completion percentages: LJ Speech (76%), Polly Sally (84%), and Google A (87%). We also observe a wide range of standard deviation, from no variation whatsoever for Nancy 2, to 33% for Polly Sally. LJ Speech had a median completion rate of 83% and was the only voice with a median completion rate different than 100%, potentially indicating it was the least pleasant voice to listen to.

Comprehension

The mean comprehension grade for each voice, which ranged from 0 to 10, was relatively low. The minimal observed value was 4.1 for Windows 2 and the maximum was 6.1 for Judy GL2 (see Table 1 for all scores). The text-only condition had a mean of 5.3. We found these differences to be statistically significant (one-way ANOVA, $F = 2.491$, $df = 2$, $p < 0.001$) even when we removed the human voices and text-only condition. We used Tukey Honest Significant Differences to compare TTS voice comprehension grades to the text-only condition and to the human voices. We did not observe a statistically significant difference between TTS comprehension grades and the text-only grade (Tukey multiple comparisons of means, $diff = 0.353$, $p_{adj} = 0.573$), but the difference between TTS comprehension grades and human voices was statistically significant (Tukey multiple comparisons of means, $diff = -0.586$, $p_{adj} < 0.05$).

While we observed a statistically significant relationship between voice and comprehension, additional work in this space

must be done. The range of comprehension scores for our subset of voices was relatively small, which may be a result of the content of the article that we selected. Nevertheless, there was no statistically significant difference in comprehension between participants who listened to a TTS voice and participants who read the article. There is a small difference between the mean grade for the different voices and the text-only grade, 5 and 5.3 respectively. This is an indication that people listening to an article may be able to understand as much as they would by reading long-form content.

LIMITATIONS

There are several limitations with our research. For a start, preferences for listening to long-form content may well vary based on culture, the subject of the content itself, the physical situation in which the listener is listening, and the different languages of the listener [5, 49, 50]. Our evaluation was centered around a single article on a casual topic [22]. As such, we cannot evaluate the interaction between content and voices, and how that impacts perceived quality. We also primarily (although not exclusively) tested voices with American accents reading an article in English. Furthermore, our study population was U.S. based, which could bias their preferences towards specific accents. Additionally, a MTurk population is one that generally optimizes for hourly pay, which could influence how they listened to the audio at the start of the survey. We did not collect information about participants first-language or English fluency, nor ask whether participants were visually impaired, or which assistive technologies (e.g. screen readers) or TTS-based services they might already be familiar with or use. We assumed that listening environments were broadly similar, whereas differences in frequencies and volume of background noise could make certain voices more or less intelligible than others [13, 29]. Refinements to the survey design may also yield richer data (albeit with a potential trade-off of increasing participant fatigue). Measuring the voice characteristics as Likert as opposed to binary could be one such improvement that we chose not to incorporate as an attempt to reduce participant fatigue. All of these are opportunities for future work.

DISCUSSION

Whether listening to an article, audiobook, or podcast, the voice of long-form content can have a profound effect on enjoyment, understanding, and willingness to keep listening. Which voice yields the best listening experience? The evidence from our large-scale study of 18 TTS and three human voices identifies clear trends in top performing voices, yet also echoes prior work in voice design [34, 36, 46, 49] in highlighting that the best voice for a given application depends on its context. The patterns which emerged in our analysis raise several insights which we believe can inform future work on how HCI researchers and practitioners might select a voice for a long-form listening use case.

While our analysis allows us to make broad generalizations that certain voices performed better than other voices for reading the article we chose under the circumstances we studied, our results do not conclusively identify any particular voice as the "optimal" voice for long-form content. At a high level, we found that voices such as Human 3, Judy W1, and Google C ranked highly across several measures such as speed, quality, and desire to listen to other content using that voice. However, none of these voices consistently outperformed all other voices across all of the quality dimensions we studied. For example, Human 3 received the highest rating on Mean Opinion Score, willingness to listen again, clarity, and quality, yet was ranked third in comprehension.

Perhaps unsurprisingly, we found that human voices still largely outperform TTS voices. On almost all quality dimensions we studied, including Mean Opinion Score (MOS), overall positive quality ratings, clarity, quality, and voice speed, two of the human voices consistently received higher ratings than all TTS voices. However, several of the TTS voices consistently performed better than one of the human voices (Human 2).

While using natural (recorded) human speech has traditionally been considered preferable to using synthesized speech [36], these results suggest that there are indeed situations where a high-quality TTS voice may be preferable over certain human voices. The relatively small differences in quality ratings between the highest performing TTS voices and the top human voices also reflects the increasing sophistication of today's speech synthesis technology. Computerized voices are nearing or exceeding certain human speakers, and TTS voices may soon reach parity with human speech in naturalness, expressivity, and so on, making them an even more viable option for long-form listening than they are at present.

More generally, the variation that we observe between how voices ranked across the quality dimensions in this study underscores that no single metric is sufficient for evaluating long-form speech. A voice that is, overall, enjoyable to listen to might make it more difficult to comprehend and absorb content (such as the Windows 2 TTS voice), or could score highly in clarity but read too slowly to be rated as pleasurable to listen to (such as Human 1).

Selecting a voice for any application, whether long-form or otherwise, will necessarily require navigating these trade-offs. In this study, we selected a long-form article that we believed to be as neutral as possible. However, other contexts may benefit from a voice that privileges certain dimensions over others. For instance, designers may want to optimize for clarity and comprehension in choosing a voice for a textbook. On the other hand, a voice intended for a children's storybook may want to use a highly expressive voice that emulates a character's persona, even if listeners would not wish to listen to it again in a different context. Likewise, selecting a voice that speaks more rapidly (higher words per minute) may be more fitting for an action novel. Listeners may wish to chose one voice for their commute on a subway train with large amounts of background noise, and a different one for listening while doing the dishes, and yet another voice for listening in bed at night. Clarity when a voice is further sped up is another important factor: it is common for blind smartphone users to listen at several hundred words a minute [6], which may require attention to specific voice factors outside of the scope of this study.

Prior work focuses on holistic quality measures such as the Mean Opinion Score (MOS) and intelligibility. We believe that the ability to collect and differentiate voices based on these more nuanced features is a key strength of our methodology, and a crucial step towards designing speech interfaces with TTS voices that are context-appropriate [2, 9, 34, 46]. Simultaneously, we also acknowledge that the voice dimensions we analyze and discuss here are by no means exhaustive. Because the primary audience for this work is system designers seeking a TTS voice for long-form content, we focused on the voice features that are most salient to end-users. However, other communities—for example, those in the speech synthesis community—may find other, more technical details regarding the voices (e.g. the synthesis model) useful in contextualizing these results. As such, we provide additional information regarding these voices where feasible at https://ttschoice.github.io, and invite further analysis and annotations from both HCI and technical voice technology researchers on these and other voices.

Social and ethical considerations

Voice is a powerful medium for persuasion. The results of our analysis and the implications of the technology we describe in this paper have important social and ethical consequences. One such concern centers on gender and representation. While the topic has received attention recently from both the press [25, 28, 44] and HCI research community [47] within the context of virtual assistants, we find similar concerns in studying the space of TTS voices. In selecting voices to analyze for this study, we aimed to include a diverse, balanced set of gender identities; however, to our knowledge, there are no publicly available human-like TTS voices that present a non-binary gender identity at present.[10] Building inclusive, representative TTS voices remains an important and urgent area for future work. The findings in this paper also reveal gender issues which we find disheartening, yet unsurprising in light of prior work: male voices in this study were significantly more likely to receive a higher quality rating than female voices, controlling for factors such as the voice's speed, whether it was text-to-speech or human, and the self-identified gender of the listener. This result is consistent with prior studies, which show that people (regardless of their own gender) generally perceive male voices—and deeper female voices—to be more knowledgeable, competent, and trustworthy [8, 30, 38]. This suggests to us that there is also a long road ahead in confronting or challenging gender expectations in TTS systems.

An implicit argument we have made in this paper—that the quality of TTS voices should improve to the point that they approximate the naturalness and expressivity of human speech—also has important social and ethical consequences. Improving the quality of TTS voices to improve the long-form listening experience undoubtedly has tremendous potential for social good by making a broader range of content accessible, increasing learning opportunities, and more. However, the same technology may also have unintended consequences. Overly realistic TTS voices, such as the voice of Google Duplex, have already created an outcry after seemingly deceiving people into thinking they are speaking with another human [15], and in some cases, can already convincingly replicate a specific individual's voice identity (i.e., a "deepfake") to problematic effect [10]. In its application to long-form content, developing TTS voices that are equivalent to or better than humans also risks displacing jobs for workers in professions like voice acting. While using a TTS voice for long-form content with full disclosure and knowledge on the part of the listener does not carry the same consequence as a deepfake, keeping these broader effects on the voice ecosystem in mind will be critically important moving forward.

[10] One promising project in this space is Q (https://www.genderlessvoice.com/), a prototype voice designed to be genderless. The Q development team reached out to us, but as Q is not a TTS voice, they were unable to provide us with a recording of the article in Q's voice for analysis.

IMPLICATIONS FOR FUTURE WORK

Even if we were to identify a single "optimal" voice for long-form listening, we openly acknowledge that such a finding would be of little practical use. In many ways, this paper represents a snapshot in time of many (but by no means all) popular, commercially available TTS voices from mid-2018 to mid-2019. However, speech technology is advancing so rapidly that the state-of-the-art is likely outpacing our own publishing cycles [41,49]. The top voice for long-form content on the market by the time this paper is published may not have existed at the time of our research and analysis. While the rankings of the particular voices we have presented in this paper will have a limited shelf-life, they have allowed us to identify that no single metric will capture all of the value of a specific voice. Instead, we suggest that practitioners and researchers looking to select a voice for their own purposes will first need to decide which set of metrics they are most interested in and then test different voices. The method detailed in this paper provides a wide range of such metrics that we hope can aid researchers and practitioners in selecting an appropriate voice. Moving forward, we also see value in periodically replicating the methodology and analysis presented here with the latest TTS voices to benchmark the community's longitudinal progress in voice quality for long-form content.

ACKNOWLEDGEMENTS

We would like to thank our participants and colleagues, particularly Abe Wallin, Tamara Hills, Ian Bicking, Kelly Davis, Eren Gölge, Eitan Isaacson, Michael Feldman, and Alan Black. We also thank our reviewers for their time and valuable feedback.

REFERENCES

[1] Sean Andrist, Micheline Ziadee, Halim Boukaram, Bilge Mutlu, and Majd Sakr. 2015. Effects of Culture on the Credibility of Robot Speech: A Comparison between English and Arabic. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction - HRI '15*. ACM Press, Portland, Oregon, USA, 157–164. DOI: http://dx.doi.org/10.1145/2696454.2696464

[2] Matthew P. Aylett, Selina Jeanne Sutton, and Yolanda Vazquez-Alvarez. 2019. The Right Kind of Unnatural:

Designing a Robot Voice. In *Proceedings of the 1st International Conference on Conversational User Interfaces (CUI '19)*. ACM, New York, NY, USA, 25:1–25:2. DOI: http://dx.doi.org/10.1145/3342775.3342806 event-place: Dublin, Ireland.

[3] Alice Baird, Stina Hasse Jørgensen, Emilia Parada-Cabaleiro, Nicholas Cummins, Simone Hantke, and Björn Schuller. 2018a. The Perception of Vocal Traits in Synthesized Voices: Age, Gender, and Human Likeness. *Journal of the Audio Engineering Society* 66, 4 (2018), 277–285.

[4] Alice Baird, Emilia Parada-Cabaleiro, Simone Hantke, Felix Burkhardt, Nicholas Cummins, and Björn Schuller. 2018b. The Perception and Analysis of the Likeability and Human Likeness of Synthesized Speech. In *Proc. Interspeech 2018*. 2863–2867. DOI: http://dx.doi.org/10.21437/Interspeech.2018-1093

[5] Alan W Black and Keiichi Tokuda. 2005. The Blizzard Challenge - 2005: Evaluating corpus-based speech synthesis on common datasets. In *Proc. Interspeech 2005*. 77–80.

[6] 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 (CHI '18)*. ACM, New York, NY, USA, Article 444, 12 pages. DOI:http://dx.doi.org/10.1145/3173574.3174018

[7] Michael Braun, Anja Mainz, Ronee Chadowitz, Bastian Pfleging, and Florian Alt. 2019. At Your Service: Designing Voice Assistant Personalities to Improve Automotive User Interfaces. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 40:1–40:11. DOI: http://dx.doi.org/10.1145/3290605.3300270 event-place: Glasgow, Scotland Uk.

[8] Alison Wood Brooks, Laura Huang, Sarah Wood Kearney, and Fiona E. Murray. 2014. Investors prefer entrepreneurial ventures pitched by attractive men. *Proceedings of the National Academy of Sciences* 111, 12 (2014), 4427–4431. DOI: http://dx.doi.org/10.1073/pnas.1321202111

[9] Julia Cambre and Chinmay Kulkarni. 2019. One Voice Fits All? Social Implications and Research Challenges of Designing Voices for Smart Devices. *To appear in Proc. ACM Hum.-Comput. Interact.* CSCW (2019).

[10] Catherine Stupp. 2019. Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case. *The Wall Street Journal* (Aug. 2019). https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402

[11] Leigh Clark, Philip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Justin Edwards, and Benjamin R Cowan. 2019a. The State of Speech in HCI: Trends, Themes and Challenges. *Interacting with Computers* iwz016 (Sept. 2019). DOI: http://dx.doi.org/10.1093/iwc/iwz016

[12] Rob Clark, Hanna Silen, Tom Kenter, and Ralph Leith. 2019b. Evaluating Long-form Text-to-Speech: Comparing the Ratings of Sentences and Paragraphs. *arXiv:1909.03965 [cs, eess]* (Sept. 2019). http://arxiv.org/abs/1909.03965 arXiv: 1909.03965.

[13] Martin Cooke, Catherine Mayo, and Cassia Valentini-Botinhao. 2013. Intelligibility-enhancing speech modifications: the hurricane challenge.

[14] Benjamin R Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. "What Can I Help You with?": Infrequent Users' Experiences of Intelligent Personal Assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '17)*. ACM, New York, NY, USA, 43:1–43:12. DOI: http://dx.doi.org/10.1145/3098279.3098539

[15] Alex Cranz. 2018. Uhh, Google Assistant Impersonating a Human on the Phone Is Scary as Hell to Me. (May 18, 2018). https://gizmodo.com/uhh-google-assistant-impersonating-a-human-is-scary-as-1825861987

[16] Nils Dahlbäck, QianYing Wang, Clifford Nass, and Jenny Alwin. 2007. Similarity is More Important Than Expertise: Accent Effects in Speech Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. ACM, New York, NY, USA, 1553–1556. DOI: http://dx.doi.org/10.1145/1240624.1240859 event-place: San Jose, California, USA.

[17] Philip R. Doyle, Justin Edwards, Odile Dumbleton, Leigh Clark, and Benjamin R. Cowan. Mapping Perceptions of Humanness in Speech-Based Intelligent Personal Assistant Interaction. In *MobileHCI 2019: 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM. DOI:http://dx.doi.org/10.1145/3338286.3340116 arXiv: 1907.11585.

[18] Edison Research and Triton Digital. 2019. *The Infinite Dial 2019*. Marketing report.

[19] Maxine Eskenazi, Gina-Anne Levow, Helen Meng, Gabriel Parent, and David Suendermann. 2013. *Crowdsourcing for speech processing: Applications to data collection, transcription and assessment*. John Wiley & Sons.

[20] Avashna Govender and Simon King. 2018a. Measuring the Cognitive Load of Synthetic Speech Using a Dual Task Paradigm. In *Proc. Interspeech 2018*. 2843–2847. DOI:http://dx.doi.org/10.21437/Interspeech.2018-1199

[21] Avashna Govender and Simon King. 2018b. Using Pupillometry to Measure the Cognitive Load of Synthetic Speech. In *Proc. Interspeech 2018*. 2838–2842. DOI: http://dx.doi.org/10.21437/Interspeech.2018-1174

[22] Greg McKeown. 2013. Reduce Your Stress in Two Minutes a Day. *Harvard Business Review* (Nov. 2013). https://hbr.org/2013/11/reduce-your-stress-in-two-minutes-a-day

[23] Iben Have and Birgitte Pedersen. 2013. Sonic Mediatization of the Book: Affordances of the Audiobook. *MedieKultur: Journal of media and communication research* 29 (03 2013), 18. DOI: http://dx.doi.org/10.7146/mediekultur.v29i54.7284

[24] Florian Hinterleitner, Georgina Neitzel, Sebastian Möller, and Christoph Norrenbrock. 2011. An evaluation protocol for the subjective assessment of text-to-speech in audiobook reading tasks. *Proceedings of Blizzard Challenge* (2011).

[25] Katharine Schwab. 2019. The real reason Google Assistant launched with a female voice: biased data. *FastCompany* (Sept. 2019). https://www.fastcompany.com/90404860/the-real-reason-there-are-so-many-female-voice-assistants-biased-data

[26] Simon King. 2014. Measuring a decade of progress in text-to-speech. *Loquens* 1, 1 (2014), 006.

[27] Sara L. Knox. 2011. Hearing Hardy, talking Tolstoy : the audiobook narrator's voice and reader experience. (2011). http://handle.uws.edu.au:8081/1959.7/543239

[28] Marianne LaFrance. 1989. The quality of expertise: implications of expert-novice differences for knowledge acquisition. *ACM SIGART Bulletin* 108 (1989), 6–14. DOI:http://dx.doi.org/10.1145/63266.63267

[29] B. Langner and A. W. Black. 2005. Improving the understandability of speech synthesis by modeling speech in noise. In *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, Vol. 1. I/265–I/268 Vol. 1. DOI:http://dx.doi.org/10.1109/ICASSP.2005.1415101

[30] Eun Ju Lee, Clifford Nass, and Scott Brave. 2000. Can Computer-generated Speech Have Gender?: An Experimental Test of Gender Stereotype. In *CHI '00 Extended Abstracts on Human Factors in Computing Systems (CHI EA '00)*. ACM, New York, NY, USA, 289–290. DOI:http://dx.doi.org/10.1145/633292.633461 event-place: The Hague, The Netherlands.

[31] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16* (2016), 5286–5297. DOI: http://dx.doi.org/10.1145/2858036.2858288

[32] C. McGinn and I. Torre. 2019. Can you Tell the Robot by the Voice? An Exploratory Study on the Role of Voice in the Perception of Robots. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 211–221. DOI: http://dx.doi.org/10.1109/HRI.2019.8673305

[33] Joseph Mendelson and Matthew P. Aylett. 2017. Beyond the Listening Test: An Interactive Approach to TTS Evaluation. In *Proc. Interspeech 2017*. 249–253. DOI: http://dx.doi.org/10.21437/Interspeech.2017-1438

[34] Roger K Moore. 2017a. Appropriate Voices for Artefacts: Some Key Insights. In *1st International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots*.

[35] Roger K. Moore. 2017b. Is Spoken Language All-or-Nothing? Implications for Future Speech-Based Human-Machine Interaction. In *Dialogues with Social Robots: Enablements, Analyses, and Evaluation*, Kristiina Jokinen and Graham Wilcock (Eds.). Springer Singapore, Singapore, 281–291. DOI: http://dx.doi.org/10.1007/978-981-10-2585-3_22

[36] Clifford Nass and Scott Brave. 2005. *Wired for speech: How voice activates and advances the human-computer relationship.* MIT press.

[37] Clifford Nass and Kwan Min Lee. 2001. Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of experimental psychology: applied* 7, 3 (2001), 171.

[38] Clifford Nass, Youngme Moon, and Nancy Green. 1997. Are machines gender neutral? Gender-stereotypic responses to computers with voices. *Journal of Applied Social Psychology* 27, 10 (1997), 864–876. DOI: http://dx.doi.org/10.1111/j.1559-1816.1997.tb00275.x

[39] Casey Newton. 2018. Pocket redesigns its mobile apps to emphasize listening. (Oct. 11, 2018). https://www.theverge.com/2018/10/11/17961564/pocket-redesign-listening-amazon-polly

[40] Christoph R Norrenbrock, Florian Hinterleitner, Ulrich Heute, and Sebastian Möller. Towards perceptual quality modeling of synthesized audiobooks-Blizzard Challenge 2012. Proceedings of the Blizzard Challenge, 2012. http://festvox.org/blizzard/bc2012/Norrenbrock_etal_Blizzard_workshop_2012_final.pdf

[41] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016).

[42] Sarah Perez. 2017. Audm turns long-form print journalism into professionally narrated digital audio. (July 14, 2017). https://techcrunch.com/2017/07/14/audm-turns-long-form-print-journalism-into-professionally-narrated-digital-audio/

[43] Victoria Petrock. 2019. Voice Assistant Use Reaches Critical Mass. (August 15, 2019). https://www.emarketer.com/content/voice-assistant-use-reaches-critical-mass

[44] Quentin Hardy. 2016. Looking for a Choice of Voices in A.I. Technology. *The New York Times* (Oct. 2016). https://www.nytimes.com/2016/10/10/technology/looking-for-a-choice-of-voices-in-ai-technology.html

[45] Falk Rehkopf. 2019. Audio is the new video: Will podcasts take off in Europe? https://www.ubermetrics-technologies.com/blog/audio-is-the-new-video-will-podcasts-finally-take-off-in-europe/. (Feb. 2019). Accessed: 2019-3-19.

[46] Selina Jeanne Sutton, Paul Foulkes, David Kirk, and Shaun Lawson. 2019. Voice As a Design Material: Sociophonetic Inspired Design Strategies in Human-Computer Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 603:1–603:14. DOI: http://dx.doi.org/10.1145/3290605.3300833 event-place: Glasgow, Scotland Uk.

[47] Marie Louise Juul Søndergaard and Lone Koefoed Hansen. 2018. Intimate Futures: Staying with the Trouble of Digital Personal Assistants through Design Fiction. *Proceedings of the 2018 on Designing Interactive Systems Conference 2018 - DIS '18* (2018), 869–880. DOI: http://dx.doi.org/10.1145/3196709.3196766

[48] Benedict Tay, Younbo Jung, and Taezoon Park. 2014. When stereotypes meet robots: The double-edge sword of robot gender and personality in human–robot interaction. *Computers in Human Behavior* 38 (Sept. 2014), 75–84. DOI: http://dx.doi.org/10.1016/j.chb.2014.05.014

[49] Petra Wagner, Jonas Beskow, Simon Betz, Jens Edlund, Joakim Gustafson, Gustav Eje Henter, Sébastien Le Maguer, Zofia Malisz, Éva Székely, Christina Tånnander, and Jana Voße. 2019. Speech Synthesis Evaluation — State-of-the-Art Assessment and Suggestion for a Novel Research Program. In *Proc. 10th ISCA Speech Synthesis Workshop*. 105–110. DOI: http://dx.doi.org/10.21437/SSW.2019-19

[50] Mirjam Wester, Cassia Valentini-Botinhao, and Gustav Eje Henter. 2015. Are We Using Enough Listeners? No!—An Empirically-Supported Critique of Interspeech 2014 TTS Evaluations. In *Proc. Interspeech 2015*. https://www.isca-speech.org/archive/interspeech_2015/papers/i15_3476.pdf

[51] Andy Wolber. 2017. 4 Text-to-Speech apps that will read online articles to you. (April 05, 2017). https://www.techrepublic.com/article/4-text-to-speech-apps-that-will-read-online-articles-to-you/