Persuasive Synthetic Speech:

Voice Perception and User Behaviour

Mateusz Dubiel[1], Pilar Oplustil Gallegos[2], Martin Halvey[1], Simon King[2]
mateusz.dubiel@strath.ac.uk,P.S.Oplustil-Gallegos@sms.ed.ac.uk,
martin.halvey@strath.ac.uk,Simon.King@ed.ac.uk
[1]Dept. Computer and Information Sciences, University of Strathclyde, UK
[2]Centre for Speech Technology Research (CSTR), University of Edinburgh, UK

ABSTRACT

Previous research indicates that synthetic speech can be as persuasive as human speech. However, there is a lack of empirical validation on interactive goal-oriented tasks. In our two-stage study (online listening test and lab evaluation), we compared participants' perception of the persuasiveness of synthetic voices created from speech in a debating style vs. speech from audio-books. Participants interacted with our Conversational Agent (CA) to complete 4 flight-booking tasks and were asked to evaluate the voice, message and perceived personal qualities. We found that participants who interacted with the CA using the voice created from debating style speech rated it as significantly more truthful and more involved than the CA using the audio-book-based voice. However, there was no difference in how frequently each group followed the CA's recommendations. We hope our investigation will provoke discussion about the impact of different synthetic voices on users' perceptions of CAs in goal-oriented tasks.

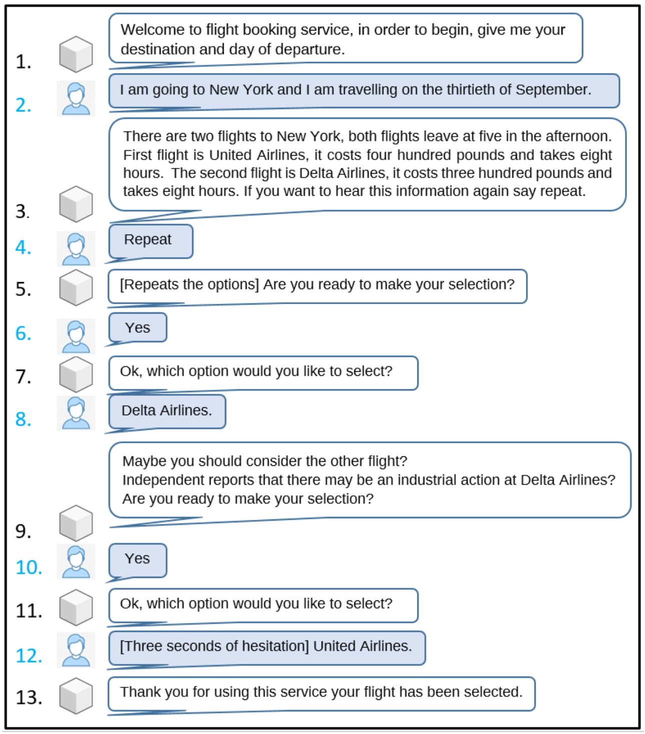
KEYWORDS

Speech Synthesis, Speech Perception, User Behaviour

ACM Reference Format:
Mateusz Dubiel[1], Pilar Oplustil Gallegos[2], Martin Halvey[1], Simon King[2]. 2020. Persuasive Synthetic Speech:: Voice Perception and User Behaviour. In *1st International Conference on Conversational User Interfaces (CUI 2020), July 22–24, 2020, Bilbao, Spain.* ACM, New York, NY, USA, 9 pages. https://doi.org/https://doi.org/10.1145/3313831.XXXXXXX

1 INTRODUCTION

The development and evaluation of a persuasive Conversational Agent (CA) is a long-standing problem. Historically, the focus of research has been mostly on text rather than speech. Applications of persuasive CAs include: the legal domain [19], intelligent tutoring [51], and car sales [1]. Now that synthetic voices are already in widespread use and may even become indistinguishable from human speech [18, 24, 32], more research on the impact on their perception and user behaviour is required. Potentially, persuasive synthetic speech could have a positive impact in, for example, a self-help CA that provides coaching or counselling. But the negative implications are also obvious, such as a CA that manipulates purchasing decisions to the user's disadvantage. As present-day commercial CAs are recently becoming equipped with voice-ordering and product search skills (e.g. itinerary planning, grocery shopping, flight search) their potential to affect users' choices increases. In the current study, heeding Rogers et al.'s call for the HCI community to reflect on the process of development of future voice technology and its impact on society [33], we investigated perceptions of a persuasive synthetic voice and its impact on user behaviour in a voice-only search task.



1. Welcome to flight booking service, in order to begin, give me your destination and day of departure.

2. I am going to New York and I am travelling on the thirtieth of September.

3. There are two flights to New York, both flights leave at five in the afternoon. First flight is United Airlines, it costs four hundred pounds and takes eight hours. The second flight is Delta Airlines, it costs three hundred pounds and takes eight hours. If you want to hear this information again say repeat.

4. Repeat

5. [Repeats the options] Are you ready to make your selection?

6. Yes

7. Ok, which option would you like to select?

8. Delta Airlines.

9. Maybe you should consider the other flight?
Independent reports that there may be an industrial action at Delta Airlines?
Are you ready to make your selection?

10. Yes

11. Ok, which option would you like to select?

12. [Three seconds of hesitation] United Airlines.

13. Thank you for using this service your flight has been selected.

Figure 1: A sample interaction between a participant and the CA. The attempt at persuasion on line 9 is successful and the participant changes their original flight selection.

We seek to answer the following research questions:

(1) What is the difference in perception between a persuasive synthetic voice and an expressive but non-persuasive synthetic voice?
(2) Do users more often follow the recommendations of the persuasive voice, in an interactive goal-oriented scenario?

CUI 2020, July 23–24, 2020, Bilbao, Spain
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-6708-0/20/04…$15.00
https://doi.org/https://doi.org/10.1145/3313831.XXXXXXX

To answer these research questions, we performed an interactive user study in two stages. In both stages, we employed synthetic voices created using a state-of-the-art data-driven method, described in Background Section. In the first, preliminary stage (online listening test), we asked participants to evaluate the persuasiveness of synthetic voices created from two datasets: one which contains intentionally-persuasive speech and another one which contains readings of books, presumed not to be intentionally persuasive. The goal of the first stage was to select the single most persuasive synthetic voice per dataset, thus providing a persuasive voice for the second stage, and strong baseline voice for comparison. In the second stage, (Interactive Evaluation), participants undertook a series of search tasks, interacting with a Conversational Assistant (CA) to achieve the goal of selecting a flight. In each task, the CA attempted to persuade participants to change their original selection by providing counter-arguments. We evaluated the persuasiveness of the CA via: (1) questionnaires adapted from [40] covering perception of the message, voice and personal qualities of the speaker; (2) the number of times a participant followed the recommendation of the CA.

The aim of our two-step evaluation method is to provide additional contextual information and, in turn, offer greater ecological validity than using listening tests alone.

2 BACKGROUND

2.1 Text-to-Speech (TTS)

Text-to-Speech (TTS) automatically converts text into synthetic speech [46] and has diverse applications such as assistive communication [50], screen readers [17], or spoken interfaces for products like Apple Siri [6]. A TTS system must generate synthetic speech for any input text. It should usually sound as similar as possible to natural speech (‘natural'), and be as comprehensible as possible (‘intelligible'): cf. [46, Chapter 17.2].

2.2 Building a TTS system from data

We employ a state-of-the-art method [44][1] in the TTS systems used in our study. The key component of the system is a neural network-based model trained on a dataset of paired audio waveforms and their phonemic transcriptions [2] [53]. The choice of data directly affects the resulting synthetic voice: the model's output will closely resemble the characteristics of the original speech data in speaker identity, accent, style, and all other acoustic properties, both segmental [3] and prosodic [4].

During the training phase, the model learns to regress from a sequence of input phonemes to the corresponding audio (represented by its spectrogram[5]). Once trained, the model can generate output from arbitrary input sequences that were not seen during training. Such models require a large quantity of training data, and this frequently necessitates using speech from multiple speakers. Speaker labels are added to the model's input during both training and generation. Thus, both the message (by providing the text) *and* the speaker identity of the output speech can be controlled. In order to generate synthetic speech with specific properties, such as sounding persuasive, one needs to use that type of data to train the model.

2.3 Persuasive TTS - Prerequisites

In the process of building our persuasive TTS, we chose to rely on methods developed in the area of *expressive TTS* ([5, 16, 37]). There are multiple definitions of the concept of expressivity or styles in TTS. However, in general, it means a *non-neutral* voice which is not constant in all its characteristics (i.e. speech rate, intonation, pronunciation, etc.). Although initial approaches to obtain expressive speech involved signal processing to manipulate audio (e.g. increasing or decreasing fundamental frequency given a rule), these methods negatively impacted the quality of the synthesised speech, and as such, they were progressively replaced with data-based methods [37]. In our study, we assume that given paired text and audio data from persuasive speakers, we can apply a data-based method to build a TTS system that generates speech with persuasive characteristics. The first step in the process is to identify the qualities that make a speaker sound persuasive.

2.4 Qualities of Persuasive Speaker

In order to be persuasive, a speaker needs to be perceived as knowledgeable, truthful, accurate, powerful and trustworthy. These qualities, as summarised by Ketrow [23], are manifested in the following paralinguistic cues: fast speech (speaking rate 150-200 words per minute); fluent speech (few pauses, no unnatural hesitations); use of greater pitch variation; volume and stress emphasis variety (dynamic speech). In a similar vein, Strangert and Gustafson [42, 43] explored the qualities that make a good speaker, focusing on pitch dynamics, fluency and speech rate. They found that by manipulating the pitch range of speech samples, speakers were perceived as more truthful, expressive and involved. On the contrary, speakers with lower pitch dynamics and narrower pitch range were perceived as insecure, hesitant and monotonous.

Dynamic speech (i.e., with high variance of pitch) is also linked to a speaker being perceived as more attractive [13–15, 35] and trustworthy [3] [28]. The positive impact of a high pitch range and fast speaking rate was also found in a banking context [7] where subjects who listened to an advertisement delivered in a quick and dynamic voice were more likely to buy the product, and in a rhetorical appeal [31] where speakers with a wider pitch range and faster speaking rate were perceived as more appealing to the audience. Schirmer et al. [36] found a positive relationship between the valence of vocal expressions and rated trustworthiness. The age of speaker rather than social status was also found to impact their perceived trustworthiness, with younger participants (aged 19-27) perceived as more trustworthy. A higher pitch range had additional credence while fast speech was found to be more persuasive than slow speech. However, it is yet to be explored how pitch variance interacts with other acoustic features in creating the effect of persuasiveness. In our study, we will explore the impact of pitch variance and speaking rate on the perceived persuasiveness in an interactive search scenario. Our approach goes beyond traditional

[1] https://github.com/oliverwatts/ophelia (last accessed: 12th May 2020)

[2] A symbolic representation of speech sounds.

[3] Vowels and consonants of the language.

[4] Suprasegmental features of language, such as intonation, speech rate, intensity.

[5] A representation of the frequencies of sound speech over time.

methods of synthetic speech evaluation that are mostly limited to listening tests.

Once the desired speech qualities of a persuasive speaker have been identified, we can then look for data that contains them and proceed to building our persuasive TTS system. The next crucial step is evaluation, which is required to assess if listeners are able to perceive these characteristics in the generated speech output.

2.5 Evaluating Synthetic Speech

Currently, the most prevalent and reliable evaluation technique in TTS is to use listening tests, e.g., [21]. Participants are asked to judge the synthetic speech, often with reference to 'gold standard' natural speech samples, output from other TTS systems, and a baseline. In many studies, the baseline is crucial to test the null hypothesis. In particular, we need a non-persuasive baseline TTS system to be able to test for a significant difference between the baseline and the persuasive system.

In typical evaluations, participants rate speech samples in terms of naturalness, similarity to the original speaker, and sometimes other dimensions [48]. Most commonly, isolated sentences are presented one at a time. Despite such listening tests being the standard way to evaluate TTS, a fair criticism is the lack of ecological validity, notably that listeners make their judgements without any context [25, 29]. Arguably (but rarely actually investigated), providing context could lead to different ratings. For example, prosodically-different versions of the same text may be rated differently in context compared to when presented in isolation. Improved ecological validity should provide perceptual evaluation results that more closely apply to the eventual application (i.e. context in which synthetic voice will be used). These considerations are particularly important for expressive speech. We expect that a persuasive voice has a greater effect in a context that requires a speaker to be persuasive, such as making recommendations or debating. Here, our ecological validity concerns led us to design the main evaluation as a goal-oriented task (more detailed explanation of the experimental setup is presented in 'Stage 2 - Interactive Evaluation' Section).

2.6 Listeners' Perceptions of Natural and Synthetic Speech

It is important to ask whether the perception of synthetic speech resembles the perception of human speech, such that characteristics that are perceived as persuasive in natural speech will also be considered persuasive if they feature in synthetic speech. Over the last 20 years, the naturalness and quality of computer generated speech has been steadily improving. In a 1999 study by Stern et al. [40] the natural human voice was generally perceived more favourably than the synthesised voice. They, however, found that there was 'little evidence to suggest that there is a difference between natural human speech and synthetic TTS speech in degree of persuasiveness, in terms of attitude and shift of a topic.' [40, p.594]. Later, in 2004, Stern et al. found that arguments presented by human voice *are* more persuasive than by synthetic voice [39]. Participants disliked unnatural voices and, as a result, gave lower ratings to such speakers, the message and the effectiveness of the argument. But later again, in the followup study, Stern et al. [41] found that synthetic speech was rated nearly as positively as a human voice when the listener knew that the source was a computer: 'We are equally comfortable with computers who speak in either human or computer synthesised speech.' [41, p.51]. On balance, TTS appears to be as effective as human speech in persuading listeners. However, since social perception of TTS is affected by its current use and the context in which technology is expected to be used in the future (cf. [29]), it is important to evaluate it in conditions that meet this requirement. In our study we evaluate our TTS empirically via interactive user evaluation.

3 METHODOLOGY

The methodology has two stages: (1) Voice Selection Stage: where we build a persuasive voice and an expressive voice (our baseline), and (2) Interactive Evaluation Stage: where both voices are evaluated in an ecologically-valid way. We evaluate persuasiveness of a CA in a goal-oriented scenario as: (1) it approximates the current capabilities of commercial CAs (e.g. flight search skills such as KAYAK [22] or SkyScanner[38]) and (2) it offers an opportunity to test persuasiveness of a CA by offering flight suggestions that participants can follow or ignore. Our approach goes beyond traditional evaluation of synthetic voices which is limited to listening tests composed of sentences without a context. Our investigation is more ecologically valid as CAs are evaluated in their intended context of use.

3.1 Stage 1 - Voice Selection

The goal of the Voice Selection Stage is to select the most persuasive speaker per dataset for further comparison in the Interactive Evaluation State. In the Stage 1: we first create synthetic voices for the speakers in both the IBM Debater (debating speech) [30] and LibriTTS (audiobooks) [52] datasets, then we select candidates for the listening test with matching prosodic qualities, and finally evaluate the selected speakers in an online-listening test to select the two best voices for further evaluation in Stage 2. Below, we explain the whole process step-by-step.

3.1.1 *Building a TTS.*

First, in order to create a TTS system with persuasive qualities we need a speech corpus that is assumed to have such qualities. The most appropriate publicly available speech corpus that fulfilled the above criterion which we found was the IBM Debater dataset [30] [6]. The corpus contains about 19 hours of American English speech from professional debaters, more males than females, elicited by asking speakers to argue in favour or against a controversial topic (e.g. 'Social media brings more harm than good', 'Gambling should be banned' etc.). We assumed that this 'debating style' must include persuasive features as the speakers are trying to convince an imagined audience of their position on the topic. The dataset contains long monologues (average duration 4 minutes) which we chunked into the shorter samples required to train our systems. The chunking was based only on the text, which we parsed with NLTK[7] so that, by defining a simple grammar of noun phrases and verb phrases we could split it into sentence-like units. The Gentle aligner [8] was then employed to force-align this

[6] https://www.research.ibm.com/artificial-intelligence/project-debater/ (last accessed: 12th May 2020)

[7] https://www.nltk.org/ (last accessed: 12th May 2020)

[8] https://github.com/lowerquality/gentle (last accessed: 12th May 2020)

chunked text with the audio, which could then be consecutively segmented to match, resulting in a collection of text-audio pairs. Since the data in the corpus is mostly-comprised of male speakers, our initial TTS system we created exhibited substantially worse quality for female voices. Therefore, in order to achieve better audio quality, for the rest of the work we used only male speakers. Once female voices were discarded, we were left with 12 hours of speech from a total of 8 speakers. This decision also enabled us to control for gender variable in our user evaluations. To build a strong baseline for comparison, we required data of equivalent quality that was not elicited from speakers intending to be persuasive. In our judgement, the fairest comparison would be to use expressively-spoken audio-books. These were taken from LibriTTS [52] [9] - an extensive dataset of 585 hours of speech from a total of 2,456 speakers. Free audio-books are a common source of training data for TTS [52]. Although vast quantities of these are available, they have to be selected with caution: the speakers are rarely professional, they record themselves in variable environments, and the amount of data per speaker is highly unbalanced (ranging from many hours to just a couple of minutes per speaker). We used a subset of the LibriTTS train-clean-100 partition of comparable size to the persuasive data above, selecting male speakers with the most data until we had about 12 hours of speech. This came from 27 American English speakers. We obtained phonemic transcriptions for both datasets using Festival [8] with the CMU American English lexicon [26].

3.1.2 Speaker Pre-selection. Initially there were 12 speakers in the IBM Debater dataset and 27 in the LibriTTS dataset. Since comparing such a large number of speakers would not be feasible, we ran a pre-selection to shortlist voices with best quality. We began with IBM Debater, the TTS system with persuasive qualities. We first scraped text from web pages with advice and recommendations on air-travel. This step was taken to create a bank of persuasive sentences for synthesis. The sentences were selected based on their relevance to the task - i.e. persuading a listener. In total, we selected one hundred sentences and synthesised them for all the IBM Debater speakers. Some of the example sentences were: 'It is recommended to book your flight via a travel agent', or 'Budget airlines are more likely to go on strike than more expensive carriers'. Whilst the state-of-the-art model that we are using for TTS generally produces a high-quality output, like many sequence-to-sequence models, it makes occasional pronunciation errors that reduce intelligibility. Solving this problem is out of scope for the current study, so we instead screened all samples for intelligibility (using informal listening by the authors) and eliminated sentences with obvious errors (mispronunciations, unnatural pauses, background noise, etc.). We identified 'AM', 'daniel' and 'eitan' as the three highest quality speakers synthesised from the IBM Debater model.

To choose three corresponding speakers from the baseline model (LibriTTS), we ran a prosodic analysis on all the synthetic speech samples synthesised by both models (3 selected speakers from IBM Debater and 27 speakers from the LibriTTS corpus). We computed the standard deviation of fundamental frequency and mean sentence duration - using them as rough indicators of pitch variation and speech rate, respectively (Figure 2). This step was taken to distinguish 3 pairs of speakers who were matched based on the similarity of their prosodic characteristics. In a two-dimensional space (fundamental frequency and mean sentence duration), we calculated the Euclidean distance between every LibriTTS speaker and the three selected IBM Debater speakers which enabled us to pair the most similar speakers. The selected LibriTTS speakers were '412', '5322' and '3664' as matches for 'AM', 'daniel' and 'eitan', respectively. The description of the prosodic characteristics of the synthesised speech by the selected speakers for both data sets is presented in Table 1. We decided to focus on 3 pairs of voices as we considered it to be a manageable number for our online listening test.

Table 1: Attributes of synthesised speech of the speakers selected from the LibriTTS and IBM Debater dataset. The speakers are ordered based on similarity of their prosodic characteristics.

Dataset	Speaker ID	Words per min. (per speaker)	Mean Pitch (per speaker)
LibriTTS	412	~201	116Hz (SD=26)
IBM Deb.	AM	~200	144Hz (SD=30)
LibriTTS	5322	~200	173Hz (SD=123)
IBM Deb.	eitan	~215	171Hz (SD=118)
LibriTTS	3664	~244	215Hz (SD=169)
IBM Deb.	daniel	~243	278Hz (SD=175)

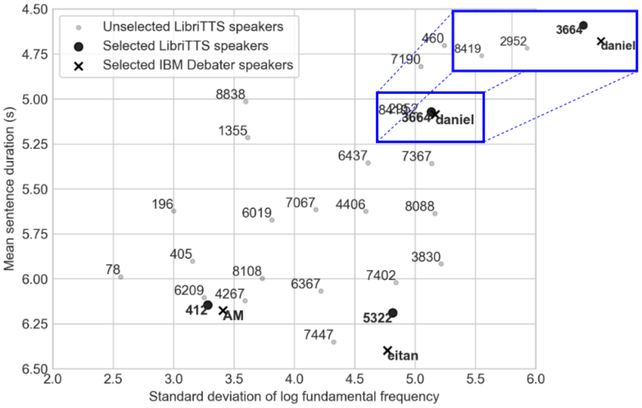


Figure 2: Comparison between the IBM Debater and LibriTTS speakers selected during the Voice Selection Stage based on their mean sentence duration (s) and standard deviation of log fundamental frequency. We used the log to compress the x axis and make it comparable to the y axis through Euclidean distance. Note: for improved readability, blue rectangles provide a closeup view of the selected area.

3.1.3 Listening Test. As a reminder, the goal of the Voice Selection Stage is to select the two most persuasive voices (one per dataset), to be carried forward to the Interactive Evaluation Stage, where they will be used in a goal-oriented task (booking a flight). Having selected three speakers for each dataset, we moved to an online listening test to distinguish the best voice for each dataset. The

[9] https://ai.google/tools/datasets/libri-tts/ (last accessed: 12th May 2020)

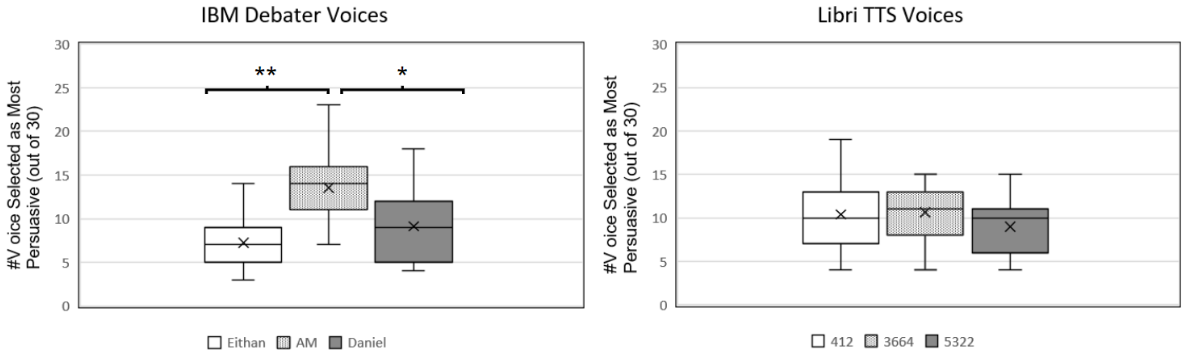


Figure 3: Listening test results for IBM Debater voices (left) and LibriTTS voices (right). The plot shows how many times on average a voice was selected as the most persuasive by 30 participants. Speakers in LibriTTS are identified by numbers (i.e. 412, 3664 and 5322). ’*’ indicates $p < 0.05$ and ’’ indicates $p < 0.005$.**

test was conducted via Amazon Mechanical Turk. In the listening test, the crowd-sourced participants were asked to rank groups of synthetic speech samples (grouped by dataset) in order of perceived persuasiveness. We did not compare the voices between the datasets as we wanted to carry out such a comparison in an ecologically-valid way (explained in ‘Interactive Evaluation’ Section).

Our hypotheses for the listening test were: (H1) There would be a statistically significant difference between the 3 speakers from IBM Debater dataset considering persuasiveness, where ‘daniel’ would be selected as the most persuasive (as it is the fastest and most varied voice); (H2) speaker 3364 would be considered as the most persuasive LibriTTS speaker, as its prosodic characteristics match ‘daniel’. The null hypothesis is therefore that there will be no statistical differences in the perception of the speakers for either of the datasets.

We ran an online listening test using the crowd-sourcing platform Amazon Mechanical Turk by giving a link to a Qualtrics survey [10]. The test was approved by Ethics Board of Department of Computer and Information Sciences, University of Strathclyde (application no.970). Participants were presented with 30 groups of test samples in total; for each group participants had to rank three samples of the same sentence, either generated by the three IBM Debater speakers or by the three LibriTTS speakers. The samples were based on 15 sentences selected from our ‘persuasive sentence bank’ (described in ‘Speaker Pre-selection’ subsection). The reference text was shown above the samples, and the instruction was: ‘Rank the samples from 1 (the most persuasive synthetic voice) to 3 (the least persuasive synthetic voice).’ While at the beginning of the listening test participants were instructed that ‘Your rating should reflect your perception of the synthetic voice i.e. how convincing and persuasive you find the speaker/ how likely are you to follow their recommendations.’ To avoid ordering effect, we randomised the sequence of samples and the displaying order of speakers. Based on pilot experiments, we informed participants that the average task completion time was 25 minutes. We asked the participants to use headphones and provided a sample to regulate the audio level. In order to ensure that listening tests were performed diligently and with due attention, we put several quality control measures in place. To prevent participants from skipping questions, we implemented a script to ensure that participants could not skip or play multiple samples at once. To check our listeners attention, we implemented 3 transcription tasks that were interleaved with the questions. Only the responses of participants who transcribed the sentences were used in the analysis. Moreover, to ensure high quality responses, during the recruitment, we exercised additional precautionary measures. Participants (crowd workers) were only permitted to take the listening test if they: (1) were from The United States[11], and were native English speakers; (2) had HIT (Human Intelligence Task) acceptance rate of 95% and (3) had at least 500 HITs approved. The aim of introducing constraints (2) and (3) was to reduce the risk of recruiting individuals who would not complete the study up to the required standards.

Results: In total 30 people took part (15 M and 15 F). The average age of participant was 32 years (SD=8.88). Participants took 21 minutes and 40 seconds on average to complete the task. Participants were paid 5 dollars for their participation. The results of the listening test are presented in Figure 3. Having confirmed that our data was normally distributed, we have run a paired samples T-Test. For IBM Debater dataset, the results revealed that ‘AM’ voice was perceived as significantly more persuasive than ‘eithan’ voice ($p < .001$) and ‘daniel’ voice ($p < .05$). Therefore, we reject the first hypothesis, as the speaker ‘AM’ was selected as the most persuasive, being the one with the lower speech rate and lower mean pitch of the the IBM Debater speakers. When it came to LibriTTS voices, we have not found any statistically significant differences, and therefore, we reject the second hypothesis. Therefore, we have rerun the comparison based on aggregated ratings (combining scores 1 and 2). The followup comparison revealed that speaker ‘3664’ was significantly more persuasive than speaker ‘5322’ (the lowest rated speaker). Based on this outcome we selected ‘3664’ as our baseline voice for LibriTTS dataset. Speaker ‘3664’ was the fastest and had the broadest pitch range out of the pre-selected TTS voices.

[10] link to the survey: https://preview.tinyurl.com/persuasivetts (last accessed: 12th May 2020)

[11] Since the majority of speakers in both data-sets were speakers of American English, we constrained our pool of participants to MT workers based in the US, as they were most likely to be accustomed to such accents.

3.2 Stage 2 - Interactive Evaluation

Following the results of the listening test, we selected 'AM' (IBM Debater) and '3664' (LibriTTS) as two voices to be compared in our interactive evaluation. In the evaluation we measured how the type of synthetic voice impacts the perceived persuasiveness of the speaker, and their willingness to follow their recommendations in a goal-oriented task. For that purpose, we created a prototype of a CA and used it in a Wizard of Oz (WOZ) set up [10] (a similar setup was implemented in previous research e.g. [12],[11],[47]).

Our hypotheses for the interactive evaluation were: (H1) There would be a statistically significant difference between the perception of both voices, where 'AM' (IBM Debater voice) will be perceived as more persuasive in terms of voice and personal qualities; (H2) participants would be more likely to follow recommendations of the CA using the persuasive voice. The null hypothesis was therefore that there would be no statistical differences in the perception of speaker nor for likelihood to follow its recommendations.

3.2.1 Procedure. The ethical approval for the experiment was gained in the same application as the listening test. The participants were recruited via advertisements posted at notice boards of the main campus of University of Strathclyde. In order to prevent priming, we did not recruit participants who took part in our listening test. The experiment consisted of 3 stages: (1) a pre-interaction questionnaire, (2) a series of four interactive search tasks (each followed by questionnaire on CA's voice, message and personal qualities), and (3) a semi structured interview (not presented in the current version of the paper due to space constraints). The experiment took place in a lab of our research facility. The experiment was based on between-group design with each group interacting with a different type of CA (IBM Debater Voice and LibriTTS voice).

Upon arrival, participants were briefed about the experiment and asked to fill in a pre-evaluation questionnaire that contained questions on demographics and decision making styles. We used the 'Rational and Intuitive Decision Styles Scale' proposed by Hamilton and Mohammed [20], with the aim to investigate if there is a correlation between participants' decision making style and their willingness to follow the CA's recommendations. The next stage was a goal-oriented task which consisted of four search scenarios, where participants interacted with the CA to book a flight. After each scenario, participants filled in a questionnaire. Following Stern et al. [40], we used a questionnaire to evaluate the perceived persuasiveness of the CA in terms of its (1) voice, (2) message and (3) perceived speaker qualities. Finally, having completed all scenarios, participants were invited to an informal interview, where we asked them questions about their flight selections and informed that the CA was operated by a human. As already noted, due to the space constrains, the results of the semi-structured interviews are not analysed in the current paper. However, we provide examples of participants' feedback that offer additional insights into their perception of CAs.

During the search task, participants were asked to interact with the CA to find and select a one way flight to a destination provided in a search scenario (e.g. 'You will be travelling to Boston on the tenth of November'). There were four search scenarios in total, each with a distinct destination and travel date. In order to avoid the ordering effect we used Latin Square Design [4] to rotate the scenarios. For all of the scenarios, there were two available flight options. For each option, travel time, departure time, and fare class were all the same. The options differed in terms of price and airline to make them easily distinguishable for the participants so that they could develop their preference and make their decision. We advised the participants that they should not speak when the CA is speaking (to replicate the functionality of current state-of-the-art systems that do not support 'barge-in' [12]). We did not provide the participants with any specified budget or required arrival time but, instead, we asked them to make their selection based on their interpretation of the information provided by the CA. Our motivation for choosing a flight-booking task was to provide a goal-oriented task that was familiar to participants and offered an opportunity to apply persuasive attempts.

The interaction with the CA is presented in the flow chart illustrated in Figure 4. For each scenario, the CA initiated the interaction by welcoming the participant and prompting them to provide flight details (Introduction Phase). Once the participant provided the details, the CA presented two flights and prompted the participant to make their selection. Following participant's choice, regardless of the option selected, the CA tried to persuade them to change their decision by mentioning a certain problem with the selected airline that could potentially affect the journey (Persuasion Phase). The problems used in scenarios were: 'industrial action', 'bad financial situation of airline', 'problems with the fleet', or 'recent flight cancellations.' Our CA recommendation strategy was applied for consistency – to ensure that, in each scenario, each participant is exposed to a persuasive attempt. In response to the attempt, the participant could either stick with their original choice, or follow the CA's recommendation (Selection Phase). The interactions were mediated via wireless speaker. The wizard (person controlling the CA), situated out of participant's sight, remotely controlled a prompt console to play responses to participant queries in synthetic voice. All participant interactions were audio recorded. An example interaction between the participant and CA is shown in Figure 1.

Since, as demonstrated by Rosselli et al. [34], the presence of affective stimuli (emotional appeal) can decrease attention to the content of a persuasive communication, in our study, we focused on rational messages. Specifically, when trying to persuade the participant to change their flight, the CA referred to a news story that explained a problem with the airline originally chosen by the participant. However, CA did not provide an interpretation of the news story, nor try to appeal to participant's emotions. Instead, the CA suggested that the participant should reconsider their choice by saying: 'Maybe you should consider another flight?', and follow it up with a justification, such as: 'The Guardian reports that there may be industrial action at Delta Airlines.' For all of the persuasive arguments, we have used modal verbs to indicate a potential, rather than certain, problem. We decided to limit the number of available flight options to two per scenario, so as not to overload participants working memory, and to exercise a high degree of control of the simulated scenario. Prior research from cognitive psychology and information retrieval shows that 2 options are an optimal choice when a participant needs to retain several attributes of the object in mind and reason about them [2, 49].

[12] A situation when a user starts speaking during a CA prompt.

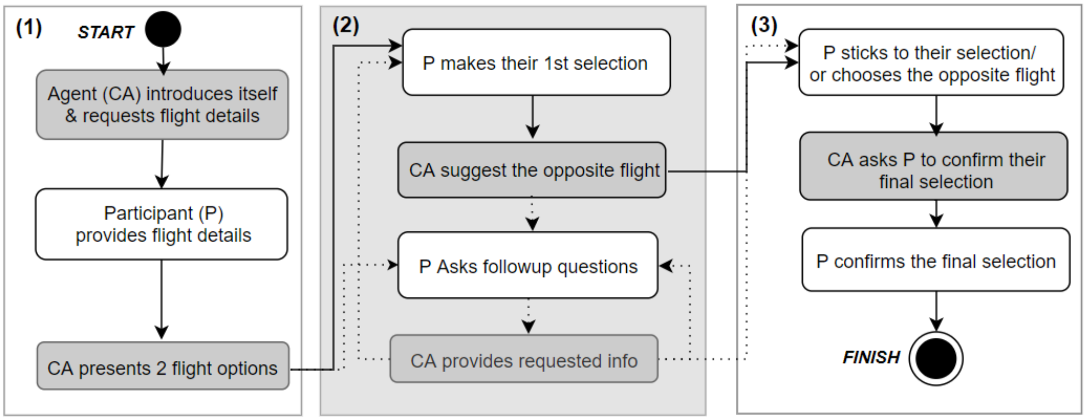

(1)
START
Agent (CA) introduces itself & requests flight details
Participant (P) provides flight details
CA presents 2 flight options
(2)
P makes their 1st selection
CA suggest the opposite flight
P Asks followup questions
CA provides requested info
(3)
P sticks to their selection/ or chooses the opposite flight
CA asks P to confirm their final selection
P confirms the final selection
FINISH

Figure 4: Interaction flow between the Conversational Agent (CA) and Participant (P). The interaction comprises of three stages: (1) Introduction - where initial information about available flights is presented, (2) Persuasion - where CA tries to convince P to change their choice and (3) Confirmation where P communicates their final choice to the CA. Note: solid arrows (–>) indicate the shortest possible path in the conversation while dotted arrows (- ->) indicate optional additional actions.

After each interactive scenario, participants were asked to complete a questionnaire that evaluated their perceptions of the CA's Voice (loudness, depth, liveliness and speed), Message (clarity, simplicity, attention grabbing, convincingness) and Personal Qualities (knowledgeableness, truthfulness, involvement, power and accuracy.) The questionnaire was based on Stern et al.'s study [40]). All of the criteria under evaluation were measured on a 7-point Likert scale, where 0 indicates the worst score and 6 indicates the best score (due to space restrictions we are unable to provide the questionnaire in full).

Results: In total, 26 participants (all Native English speakers) took part in the interactive evaluation (14 M and 12 F; M age = 26, SD = 8). 15 participants reported to have used a CA before. The most frequently used CAs was Amazon Echo (8). For the decisions style pre-study questionnaire [20]), the average score per participant was 3.4/4 for rational items (where 4 indicates the highest rationality), and 1.9/4 for intuitive items (where 4 indicates the highest intuitiveness). Since the majority of our data was not normally distributed, for statistical analysis, unless otherwise stated, we used a Mann Whitney U-Test [27]. We report effect size estimates using Cohen's [9] categories of 'small' (r = .1), 'medium' (r = .3), and 'large' (r = .5). Our comparison focuses on participants' perception of (1) Speaker, (2) Voice, and (3) Message (presented in Figure 5).

We observed statistically significant differences regarding perception of the **Speaker**, with AM being perceived as both more Truthful (Z = -3.049, p = .002, r = .3) and more Involved (Z = -2.923, p = .003, r = .3) than 3664. No statistically significant differences were observed for how Knowledgeable (Z = -1.080, p = .28, r = .1), Powerful (Z = -.328, p= .743, r < .1) or Accurate (Z = -.833, p = .405, r < .1) the speaker is perceived. In terms of **Voice**, participants rated 3664 voice both as Louder (Z = -3.347, p = .001, r = .3) and Deeper (Z = -2.011, p = .044, r = .2). We did not observe any statistical differences with regards to perceived Speed (Z =- .584, p = .559, r < .1) or Liveliness (Z = -.442, p= .659, r <. 1) of the voice. When it comes to **Message**, we observed that participants perceived the message as Simpler when presented by speaker 3664 as compared to speaker AM (Z = -2.298, p = .02, r= .2). No significant differences were observed for Capturing Attention (Z = -.897, p = .37, r = .1), Clarity (Z = -.054, p = .957, r < .1) or with regards to how Convincing the message was perceived by participants (Z = -.278, p = .781, r < .1). Given the results, we cannot fully accept nor reject our H1 (i.e. that type of synthetic voice will affect perceived persuasiveness of the speaker), since we found a statistically significant difference for two out of five aspects that determine perceived persuasiveness of a speaker.

When we look at participants' behaviour, there is no difference between the groups. Since the behaviour is considered in binary categories, i.e. following recommendation or not, we used Cochran Q Test [45]. Both for '3664' and 'AM', participants followed the recommendation 21 out of 52 instances (~ 40%, Q = 0, p. = 1). Therefore, based on this result, we reject our H2 (i.e. that the type of synthetic voice will affect participant's likelihood to follow speaker's recommendations). Although during the interviews, small number of participants (4/26) reported that they felt that the CA was trying to 'manipulate them', the sequence of the scenarios did not appear to have a significant impact on the likelihood of the participant following CA's recommendation (Table 2, top). Overall, Scenario 2 seems to be the least persuasive (Table 2, bottom).

Table 2: Number of times that participants followed the CA's recommendation based on a scenario-completion sequence (top) and a scenario-type (bottom)

	Task Completion Sequence				
CA	**1**	**2**	**3**	**4**	**Overall**
3664	6/13	4/13	6/13	5/13	21/52
AM	5/13	5/13	6/13	5/13	21/52

	Scenario Type				
CA	**1**	**2**	**3**	**4**	**Overall**
3664	5/13	3/13	5/13	8/13	21/52
AM	8/13	3/13	6/13	4/13	21/52

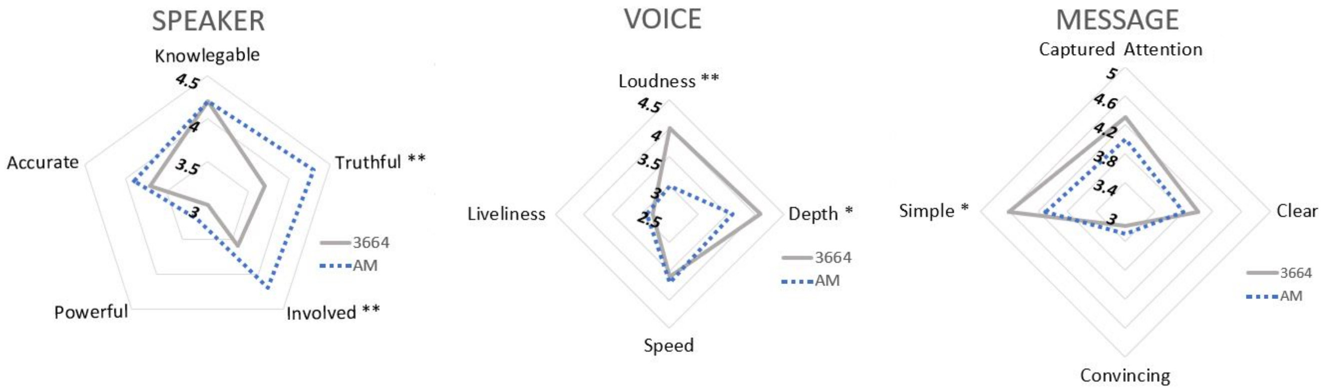


Figure 5: Perception of Speaker, Voice and Message. Note: ’’ indicates p < 0.001, and ’*’ indicates p < 0.05**

4 DISCUSSION

Our study provided us with qualitative and quantitative feedback to answer our research questions: **RQ1:** What is the difference in perception between a persuasive synthetic voice and an expressive but non-persuasive synthetic voice? and **RQ2:** Do users more often follow the recommendations of the persuasive voice, in an interactive goal-oriented scenario? We found that synthetic speech generated from models trained on a persuasive speech dataset was generally perceived as more persuasive than the baseline (**RQ1**). This was particularly the case for speaker’s Truthfulness and Involvement. However, this perception of persuasiveness did not translate into changes in user behaviour, at least in the goal-oriented task we employed (**RQ2**).

Prosody: With regards to the prosodic qualities of a synthetic voice, contrary to our assumption based on literature review, we found that a slower speech rate was linked to higher Trustworthiness and Involvement. Despite the synthetic voice created from speaker ‘AM’ (IBM Debater dataset) being both slower and narrower in pitch range than speaker ‘3664’ (the LibriTTS dataset), it outperformed the baseline in terms of perceived speaker qualities. This could have been caused by the nature of our interactive task, which required dynamic interaction between CA and participants and the fact that participants needed to actively reason about the options presented to them. Thus, slower presentation of the options (as indicated during the interviews) could have given more time to make a decision, and consequently felt less rushed. However, more research involving different task types and scenarios would be required to validate this assertion.

Acoustic Qualities: Interestingly, although AM is perceived as more Truthful and Involved it scores significantly lower in terms of Depth and Loudness. This could be the caused by the fact that IBM Debater dataset had considerably less data than Libri TTS, which in turn translated to inferior audio quality of persuasive voice. However, it did not seem to impact the perceived Liveliness and Speed for which no statistically significant differences were observed. Again this is surprising because 3664 was both faster (244 wpm vs. 200 wpm) and had more varied pitch - attribute associated with liveliness (215Hz, SD = 169 vs. 144Hz, SD = 30). It seems that ’superior audio’ quality could have also impacted the perceived delivery of the message which was considered as presented significantly simpler for 3664 than AM.

TTS Evaluation: We propose some tentative implications evaluation of persuasive TTS: (1) It is important that synthetic speech is tested in the context of its intended usage and not in an ecologically-implausible listening test - this provides a fuller understanding of the impact that the specific voice can have on a user. (2) Designers should be mindful that the pitch range and pace of speaking are not the only determinants of perceived voice persuasiveness. (3) Longitudinal evaluation may be necessary to obtain more accurate results - ideally with participants using a CA in their home setting.

5 LIMITATIONS AND FUTURE WORK

We are mindful about limitations of our study. Firstly, since most participants self-rated as highly-rational decision makers, this group could potentially be more difficult to persuade than the general population. Secondly, due to shortage of female voice data, we only used male voices. Since the IBM Debater dataset is relatively small compared to datasets typically used to train state-of-the-art TTS models; the overall quality of the synthesis in our study was lower than what could be achieved with more data. However, as mentioned in Discussion, this did not seem to affect participants’ perceptions of persuasive qualities of the speaker. Nevertheless, a persuasive voice with superior audio quality may have been ranked even higher for Truthfulness and Involvement. The IBM Debater dataset was the best available choice in terms of speaking style and we assumed that a debating style would translate into persuasiveness. Finally, our evaluation was limited to one-off interaction and different results could have been obtained for long term CA use. As future research we aim to address these limitations and see if our findings generalise to different domains.

6 CONCLUSIONS

In this work, we have evaluated the impact of persuasive synthetic speech on participants’ perception and behaviour in a goal-oriented task. We saw that even though participants perceived a synthetic voice as truthful and involved (characteristics of persuasive speakers) that did not directly translate into following recommendations given by the same voice. Our findings are preliminary and further experiments in different domains would be required to assess their validity. Finally, our proposed experiment is an important step towards more ecologically-valid evaluation of TTS quality for systems aimed to such application.

REFERENCES

[1] Elisabeth André, Thomas Rist, Susanne Van Mulken, Martin Klesen, and Stefan Baldes. 2000. The automated design of believable dialogues for animated presentation teams. *Embodied conversational agents* (2000), 220–255.

[2] Alan Baddeley. 1992. Working memory. *Science* 255, 5044 (1992), 556–559.

[3] Pascal Belin, Bibi Boehme, and Phil McAleer. 2017. The sound of trustworthiness: Acoustic-based modulation of perceived voice personality. *PloS one* 12, 10 (2017), e0185651. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0185651

[4] James V Bradley. 1958. Complete counterbalancing of immediate sequential effects in a Latin square design. *J. Amer. Statist. Assoc.* 53, 282 (1958), 525–528.

[5] Nick Campbell. 2008. Expressive/affective speech synthesis. In *Springer Handbook of Speech Processing*. Springer, 505–518.

[6] Tim Capes, Paul Coles, Alistair Conkie, Ladan Golipour, Abie Hadjitarkhani, Qiong Hu, Nancy Huddleston, Melvyn Hunt, Jiangchuan Li, Matthias Neeracher, et al. 2017. Siri On-Device Deep Learning-Guided Unit Selection Text-to-Speech System.. In *INTERSPEECH*. 4011–4015.

[7] Jean-Charles Chebat, Kamel El Hedhli, Claire Gélinas-Chebat, and Robert Boivin. 2007. Voice and persuasion in a banking telemarketing context. *Perceptual and motor skills* 104, 2 (2007), 419–437. https://doi.org/10.2466/pms.104.2.419-437

[8] Robert AJ Clark, Korin Richmond, and Simon King. 2004. Festival 2–build your own general purpose unit selection speech synthesiser. (2004). https://tinyurl.com/TTSfestival

[9] J Cohen. 1988. Statistical power analysis for the behavioral sciences, Stat. *Power Anal. Behav. Sci* 567 (1988).

[10] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz studies - why and how. *Knowledge-based systems* 6, 4 (1993), 258–266. https://doi.org/10.1016/0950-7051(93)90017-N

[11] Mateusz Dubiel. 2018. Towards human-like conversational search systems. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*. 348–350.

[12] Mateusz Dubiel, Martin Halvey, Leif Azzopardi, and Sylvain Daronnat. 2018. Investigating how conversational search agents affect user's behaviour, performance and search experience. In *The Second International Workshop on Conversational Approaches to Information Retrieval*. https://tinyurl.com/sigir-cair

[13] DR Feinberg, BC Jones, LM DeBruine, JJM O'Connor, CC Tigue, and DJ Borak. 2011. Integrating fundamental and formant frequencies in women's preferences for men's voices. *Behavioral Ecology* 22, 6 (2011), 1320–1325. https://doi.org/10.1093/beheco/arr134

[14] David R Feinberg, Lisa M DeBruine, Benedict C Jones, and David I Perrett. 2008. The role of femininity and averageness of voice pitch in aesthetic judgments of women's voices. *Perception* 37, 4 (2008), 615–623. https://doi.org/10.1068/p5514

[15] David R Feinberg, Benedict C Jones, Anthony C Little, D Michael Burt, and David I Perrett. 2005. Manipulations of fundamental and formant frequencies influence the attractiveness of human male voices. *Animal behaviour* 69, 3 (2005), 561–568. https://doi.org/10.1016/j.anbehav.2004.06.012

[16] Kerstin Fischer. 2004. Expressive Speech Characteristics in the Communication with Artificial Agents. In *Proceedings of the AISB 2004 Convention*. The Society for the Study of Artificial Intelligence and the Simulation of Behaviour, 1–11. https://tinyurl.com/communicating-artificialagents

[17] Vincent Gaudissart, Silvio Ferreira, Céline Thillou, and Bernard Gosselin. 2004. SYPOLE: mobile reading assistant for blind people. In *9th Conference Speech and Computer*.

[18] Andrew Gibiansky, Sercan Arik, Gregory Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. 2017. Deep voice 2: Multi-speaker neural text-to-speech. In *Advances in neural information processing systems*. 2962–2970.

[19] Thomas F Gordon. 1993. The pleadings game. *Artificial Intelligence and Law* 2, 4 (1993), 239–292.

[20] Katherine Hamilton, Shin-I Shih, and Susan Mohammed. 2016. The development and validation of the rational and intuitive decision styles scale. *Journal of personality assessment* 98, 5 (2016), 523–535.

[21] Vasilis Karaiskos, Simon King, Robert AJ Clark, and Catherine Mayo. 2008. The blizzard challenge 2008. In *Proc. Blizzard Challenge Workshop, Brisbane, Australia*.

[22] Kayak. 2020. (Amazon Echo application software). https://www.amazon.co.uk/KAYAK/dp/B01EILLOXI

[23] Sandra M Ketrow. 1990. Attributes of a telemarketer's voice and persuasiveness. A review and synthesis of the literature. *Journal of Direct Marketing* 4, 3 (1990), 7–21. https://doi.org/10.1002/dir.4000040304

[24] Nicole Kobie. 2018. Google's new voice is as good as your own. *New scientist* 3159 (2018), 9.

[25] Javier Latorre, Kayoko Yanagisawa, Vincent Wan, Bala Krishna Kolluru, and Mark J.F. Gales. 2014. Speech intonation for TTS: Study on evaluation methodology. Singapore.

[26] Kevin Lenzo. [n. d.]. The Carnegie Mellon University pronouncing dictionary.

[27] Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics* (1947), 50–60.

[28] Phil McAleer, Alexander Todorov, and Pascal Belin. 2014. How do you say 'Hello'? Personality impressions from brief novel voices. *PloS one* 9, 3 (2014), e90779. https://doi.org/10.1371/journal.pone.0090779

[29] Joseph Mendelson and Matthew P Aylett. 2017. Beyond the Listening Test: An Interactive Approach to TTS Evaluation.. In *INTERSPEECH*. 249–253.

[30] Shachar Mirkin, Michal Jacovi, Tamar Lavee, Hong-Kwang Kuo, Samuel Thomas, Leslie Sager, Lili Kotlerman, Elad Venezian, and Noam Slonim. 2017. A Recorded Debating Dataset. *arXiv preprint arXiv:1709.06438* (2017).

[31] Michel Nienhuis. 2009. Prosodic Correlates of Rhetorical Appeal. (2009). https://tinyurl.com/RethoricalAppeal

[32] 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).

[33] Jon Rogers, Loraine Clarke, Martin Skelly, Nick Taylor, Pete Thomas, Michelle Thorne, Solana Larsen, Katarzyna Odrozek, Julia Kloiber, Peter Bihr, et al. 2019. Our Friends Electric: Reflections on Advocacy and Design Research for the Voice Enabled Internet. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 114. https://doi.org/10.1145/3290605.3300344

[34] Francine Rosselli, John J Skelly, and Diane M Mackie. 1995. Processing rational and emotional messages: The cognitive and affective mediation of persuasion. *Journal of experimental social psychology* 31, 2 (1995), 163–190. https://doi.org/10.1006/jesp.1995.1008

[35] Klaus R Scherer, Tom Johnstone, and Gundrun Klasmeyer. 2003. Vocal expression of emotion. *Handbook of affective sciences* (2003), 433–456.

[36] Annett Schirmer, Yenju Feng, Antarika Sen, and Trevor B Penney. 2019. Angry, old, male–and trustworthy? How expressive and person voice characteristics shape listener trust. *PloS one* 14, 1 (2019), e0210555. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0210555

[37] Marc Schröder. 2009. Expressive speech synthesis: Past, present, and possible futures. In *Affective information processing*. Springer, 111–126.

[38] Skyscanner Flight Search. 2018. (Amazon Echo application software). .https://www.skyscanner.com/tips-and-inspiration/features/amazon-echo-and-skyscanner-a-match-made-in-travel-heaven

[39] Steven E Stern and John W Mullennix. 2004. Sex differences in persuadability of human and computer-synthesized speech: meta-analysis of seven studies. *Psychological reports* 94, 3_suppl (2004), 1283–1292. https://doi.org/10.2466/pr0.94.3c.1283-1292

[40] Steven E Stern, John W Mullennix, Corrie-lynn Dyson, and Stephen J Wilson. 1999. The persuasiveness of synthetic speech versus human speech. *Human Factors* 41, 4 (1999), 588–595. https://doi.org/10.1518/001872099779656680

[41] Steven E Stern, John W Mullennix, and Ilya Yaroslavsky. 2006. Persuasion and social perception of human vs. synthetic voice across person as source and computer as source conditions. *International Journal of Human-Computer Studies* 64, 1 (2006), 43–52. https://doi.org/10.1016/j.ijhcs.2005.07.002

[42] Eva Strangert and Joakim Gustafson. 2008. Improving speaker skill in a resynthesis experiment. (2008). http://www.diva-portal.org/smash/get/diva2:149676/FULLTEXT01.pdf

[43] Eva Strangert and Joakim Gustafson. 2008. What makes a good speaker? subject ratings, acoustic measurements and perceptual evaluations. In *Ninth Annual Conference of the International Speech Communication Association*.

[44] Hideyuki Tachibana, Katsuya Uenoyama, and Shunsuke Aihara. 2018. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4784–4788. https://arxiv.org/abs/1710.08969

[45] Merle W Tate and Sara M Brown. 1970. Note on the Cochran Q test. *J. Amer. Statist. Assoc.* 65, 329 (1970), 155–160.

[46] Paul Taylor. 2009. *Text-to-speech synthesis*. Cambridge university press.

[47] Alexandra Vtyurina and Adam Fourney. 2018. Exploring the role of conversational cues in guided task support with virtual assistants. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 208. https://doi.org/10.1145/3173574.3173782

[48] 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. Dresden, Germany.

[49] Andi Winterboer and Johanna D Moore. 2007. Evaluating information presentation strategies for spoken recommendations. In *Proceedings of the 2007 ACM conference on Recommender systems*. ACM, 157–160.

[50] Junichi Yamagishi, Christophe Veaux, Simon King, and Steve Renals. 2012. Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction. *Acoustical Science and Technology* 33, 1 (2012), 1–5.

[51] Tangming Yuan, David Moore, and Alec Grierson. 2008. A human-computer dialogue system for educational debate: A computational dialectics approach. *International Journal of Artificial Intelligence in Education* 18, 1 (2008), 3–26.

[52] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. *arXiv preprint arXiv:1904.02882* (2019). https://arxiv.org/abs/1904.02882

[53] Heiga Zen, Keiichi Tokuda, and Alan W Black. 2009. Statistical parametric speech synthesis. *speech communication* 51, 11 (2009), 1039–1064.