Can Computer-Generated Speech Have an Age?

Earl W. Huff, Jr.
Clemson University
Clemson, SC 29630, USA
earlh@clemson.edu.edu

Richard Pak
Clemson University
Clemson, SC 29630, USA
richpak@clemson.edu

Brodrick Stigall
Clemson University
Clemson, SC 29630, USA
bstigal@clemson.edu

Kelly Caine
Clemson University
Clemson, SC 29630, USA
caine@clemson.edu

Julian Brinkley
Clemson University
Clemson, SC 29630, USA
jbrinkl@clemson.edu

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

ACM ISBN 978-1-4503-6819-3/20/04
http://dx.doi.org/10.1145/3334480.3383082

Abstract
The present study examines whether computer-generated speech is perceived to have an age, and if so, whether we can manipulate the perceived age of the voice. We conducted an experimental study with 51 participants where each computer-generated voice had different age-related characteristics such as the speed rate of the voice and frequency of the pitch. Participants listened to vehicle reviews presented by computer-generated voices with age-related characteristics we manipulated and then evaluated the age of the voice. Results show that we can change the perceived age of a computer-generated voice by manipulating the age-related characteristics of the voice. This work contributes to communities of HCI researchers interested in voice user interfaces (VUIs), conversational agents, and age stereotypes.

Author Keywords
computer-generated speech; voice user interfaces; conversational agents

CCS Concepts
•Human-centered computing → Natural language interfaces; •Social and professional topics → Age;

Introduction

Voice User Interfaces (VUIs) are an increasingly prevalent type of conversational agent capable of simulating human conversation absent, or in concert with, a graphical representation of a human [6, 14]. Coupled with advances in voice recognition technology, the use of voice as a modality in human-computer interaction has the potential to become a widely used method of human-computer interaction [10]. Apple's Siri, for instance, enables smartphone users to complete tasks like placing calls and setting alarms using hands-free and eyes-free interaction while smart speakers like Amazon's Echo and Google Home support control of home automation systems using voice-based interaction. Regardless of the device type, voice-based interaction may allow for a more natural interaction in executing tasks [6, 11], increase task completion rates, reduce time and effort and improve a user's overall satisfaction with a system [12].

Determining Age from Voice

Humans have the ability to discern a person's age from characteristics of human voices [4, 16, 18]. Those characteristics include rate of speech [18] and frequency of the tone of a voice [4, 18]. Other characteristics such as formants, especially F1, [17], voice quality including glottal chink, [22], and intensity [19] can also affect perceived age. In this work, we chose to focus exclusively on pitch and speech rate. Voices that use a faster rate of speech and higher frequency tone are perceived as belonging to a younger person and voices that use a slower rate of speech and lower frequency tone are perceived as belonging to an older person. However, no work we are aware of has investigated whether humans perceive computer-generated speech to have an age based on differences in pitch and speech rate, and if so, whether the perceived age of the computer-generated speech can be manipulated. For example, if we manipulate a computer-generated voice to have a high frequency tone and fast rate of speech, will it be perceived as "young"?

Determining Age in a Computer-Generated Speech

Prior research has explored how perceived gender of a voice affects listeners [8, 9]. In that work, researchers found that gendered computer-generated speech resulted in listeners treating the gendered speech with behaviors stereotypical for that gender. However, we do not know whether stereotypes regarding age (e.g. speed, competence, responsiveness), prevalent in human-human interactions, persist in human-machine dialogues. The presence of such stereotypes has significant implications for the design of VUIs. If the types of age-related stereotypes common in human-human interactions exist in human-machine interactions, expectations regarding the veracity of information and the speed of response, for instance, may inform the design of VUIs.

In the present study we explore how characteristics of a voice can affect perceived age. Within the context of a car review system, where the reviews are presented by computer-generated speech with different age-related characteristics (e.g. speech rate, pitch), participants listened to reviews of vehicles and then evaluated the reviews. In the auto retail industry, we know that a number of stereotypes exist, including cultural [20], gender [3], and age stereotypes [1]. These stereotypes exist from the perspective of both the consumers and salespeople. Salespeople may employ different sales tactics or offer a different experience based on the physical characteristics of the customer [20]. There are also several stereotypes that are prevalent for buyers of particular brands of vehicles [13, 21, 23]. Using a car review system provides an environment to simulate a car-buying experience between a potential consumer and the reviewer (salesperson). The purpose of the research

Voice	Rate	Pitch	Age
F	0.96	-12.4	O
C	1.04	2.4	Y

Table 1: Voice settings for car reviewers old (O) and young (Y). Both settings were applied to Google Text-to-Speech using the Wavenet voice profile (F,C).

is to investigate whether or not age can be elicited from a computer-generated voice. Results of the study will provide evidence about whether and how age may be attributed to computer-generated voices. The results obtained contribute to the growing body of knowledge in human-machine interactions, in particular, understanding how perceived age can influence the design of digital humans and conversational dialogue systems.

Method

The study is a two-group experimental design where voice of the car reviewer (old or young) is a within-subjects variable.

Stimuli

Voice development

For this study, the attributes of voice associated with age that we manipulated are speed and pitch [4, 18]. We adjusted Google Text-to-Speech (TTS) [5] and Apple TTS [2] voices to created several sample voices exhibiting younger and older adult voice traits. The values for rate of speech and frequency of pitch are adjustments to the stock voice (e.g., speech rate 1.04 is 1.04 times the stock rate). We produced a series of MP3 audio files of sized voices reading a small portion of a car review.

Half of the voices were designed to sound “old” (i.e., lower pitch and slower speech) and the other half were designed to sound “young” (i.e., faster pitch and speech). We then asked our team of researchers to listen to each audio file and rank the voices based on the quality of the voices in terms of how they represented the intended perceived age. Each team member provided rankings of the voice groups, one for the older-sounding voice and one for the younger-sounding voice. From reviewing the rankings, the two voice settings chosen were Wavenet-F (speed: 0.96, pitch: -12.4) for the old voice and Wavenet-C (speed: 1.04, pitch: 2.4) for the young voice (see Table 1). We used only female voices to eliminate any effects of perceived gender on the perceived age of the voice.

Car Reviews

We developed the car review scripts using a multi-phase process. The first phase involved collecting car reviews. To ensure the consistency of the language used in the reviews, we searched for multiple car reviews conducted by the same reviewer. A total of 40 reviews conducted by seven authors were found. We then considered each review in terms of quality, level of detail, and whether the vehicle reviewed would fit our stereotype criteria. We chose four car reviews from one author, two representing stereotypically ’old’ vehicles and two representing stereotypically ’young’ vehicles (see sidebar for example script).

The assessment of vehicles being considered for older consumers and younger consumers were based on a literature search of consumer perceptions of certain auto manufacturers [13, 21, 23]. Additionally, we searched online forums and auto journal articles that identified vehicles that are ideal for young and older adults.

Car Review Assessment

The car review assessment consists of seven items, all on a 10-point Likert scale, that measure the overall perceived quality of the review. We used existing questions from the works of Morishima, Bennet, Nass, Lee, Moon, and Green [7, 8], which investigated the presence of gender stereotypes in VUIs. We measured quality using the questions, “What was the quality of the review you just heard?”, “How much did you like the review?”, and “How trustworthy was the review?” We measured credibility using three items that asked about the credibility, reliability, and trustworthiness of the review. We measured appropriateness using the ques-

Sample review script read by old and young computer generated voices:
"...The E-Pace is Jaguar's second all-new SUV in as many model years. Built off a platform that underpins the Range Rover Evoque and Discovery Sport from Jaguar's Land Rover affiliate, the E-Pace slots below the F-Pace in Jaguar's nascent SUV lineup. All-wheel drive and a turbocharged four-cylinder engine are standard; a higher-output turbo four-cylinder is optional..."

tion "How appropriate or inappropriate was the voice for this particular vehicle?" Finally, we measured buying intention using the question, "Based on the review, how likely would you be to buy this vehicle?"

Car Reviewer Assessment
The car reviewer assessment consists of seven items, all but one on a 10-point Likert scale, that measure the quality of the voice conducting the review. Additionally, the last question asks participants to guess the perceived age of the voice. The seven measurable items were adopted from studies conducted by Nass, Moon, and Green [9]. We measure competence of the reviewer using three items asking about how competent, informative, and knowledgeable the reviewer is. We measure informativeness using four items with two of the items asking "How [adjective] was the reviewer?" using the adjectives helpful and sophisticated. The other two items ask how well did the reviewers explain the details about each of the two vehicles.

For this study, we present only findings about the perceived ages of the computer-generated voices. Findings related to ratings of the quality/credibility of the reviews, appropriateness of the voice and other variables will not be discussed in this paper but will be reported in future work.

Participants
Fifty one (23 male and 28 female) participants were recruited through the Amazon Mechanical Turk (MTurk) to participate in the study. We controlled for the age of the participants by only enrolling older adults (age 60 and up). We also limited the inclusion of participants to car owners, since the topic of the review was about cars.

A common concern with performing experiments on Amazon MTurk is the quality of the data due to 1) workers with low or no reputation or approval ratings and 2) inattentive workers [15]. To improve the quality of data, we restricted the sample to workers with a 95% approval rating and having completed at least 5,000 jobs. Additionally, we added four attention check questions (ACQs) in our survey. Workers who answered three or more ACQs incorrectly were not compensated for their assignment and their data was excluded from analysis.

Procedure
Participants provided informed consent prior to entering the portal to complete the experiment by reading the consent document and following the link to the portal. Participants were asked to complete a Completely Automated Public Turing test to tell Computers and Humans Apart (CAPTCHA) test to ensure that bots were not completing the study. After completing the CAPTCHA, they completed a short demographic questionnaire. Next, they read the task instructions, performed a sound test to make sure they can hear the reviews read to them, and clicked "Next" to begin the experiment.

Participants were first presented with four car reviews, two of each read by eacy car reviewer voice (young and old). One at a time, they listened to a car review. Each page was designed to prevent the participants from clicking "Next" for 60 seconds to ensure the participant listened to the review. After hearing a review, participants completed a car review assessment to rate the overall quality of the review. After completion of each review, participants were asked to complete a car reviewer assessment, one for each car reviewer voice, to rate the overall quality of the voice reviewer's performance.

Next, we asked participants a qualitative question about what they thought the study was investigating. Participants who were able to uncover the true intent of the study (the effect of age of voices on car reviews). We excluded data

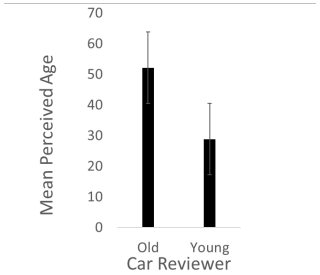

Figure 1: Mean perceived age of and standard error of computer-generated voices.

for these participants, but still compensated them. Finally, participants were taken to a page where they were debriefed on the true nature of the study. The entire study took less than half an hour, and participants were compensated $1.50.

Findings

The findings from this study will help to answer our research question: *can people perceive age from an artificially generated voice? If so, can we manipulate the perceived age using characteristics such as speed and pitch?* To determine whether the participants' perception of age differed between the two computer-generated voices, we will compare the mean differences of the perceived age of the two car reviewer voices.

Participant Demographics

While 51 participants completed the study, we excluded one because the perceived age they reported were below the boundary case (car reviewers should be at least 18 years of age for inclusion). Observations from 50 participants (23 males and 27 females) were used for analysis. Age of participants ranged from 50 to 78 years with a mean of 63.5 and standard deviation of 6. In terms of career, the majority (44%) of participants work in business or industry and 34% work in other non-specified career areas. For educational attainment, 86% of participants attended a college or university with 16.3% earning a graduate degree, 34.9% earning a 4-year degree, 23.3% earning a two year degree and the remaining 25.5% having attended some college but not receiving a degree.

Perception of Age

The study included two different computer-generated voices, one "young" and one "old". The "young" voice had the age-related characteristics of faster speech and higher pitch speech (see Table 1). Each participant listened to two reviews, one by each voice (young and old), which were counterbalanced across participants. For the characteristically-old voice, the perceived age ranged from 30 to 80 years with a mean of 52.08 and a standard deviation of 11.26. For the characteristically-young voice, perceived age ranged from 20 to 50 years with a mean of 28.78 and a standard deviation of 5.4. Observing the differences between the characteristically old and young voices, the mean difference was 23.3 with a standard deviation of 14. Although the distribution of the old voice was normal and the young voice was slightly not normal, the distribution of the differences was normal. Figure 1 presents a bar chart of the mean perceived ages of the two voices with standard errors.

A paired sample t-test revealed there was a significant difference between the means of the ages of the two voices, $t(49) = 11.729$, $p < .001$, $r = 0.859$.

Discussion

We find that people who listen to computer-generated speech distinguish between young sounding and old sounding voices. Perceptions of age of a voice can be manipulated by altering the age-related characteristics of the voice. Specifically, computer-generated voices with a lower rate of speech and a lower frequency pitch are perceived as older, while voices with a higher rate of speech and higher frequency pitch are perceived as younger. Since humans have the ability to discern the age of a human from the characteristics of the human voices [4, 16, 18], it is not surprising that when we manipulate these same characteristics in a computer-generated voice, humans also perceive age differences. However, this is the first time that this result has been established empirically. This result is also similar to the work that found that people perceive computer-generated voices to have a gender [9].

Conclusion and Future Work

The perceived age of a computer-generated voice can be manipulated using differences in age-related characteristics of voices. This result means that people perceive that computer-generated voices have an age, and that the age is affected by the same age-related characteristics as in humans. Knowing that people perceive a computer-generated voice to have an age is the first step in more elaborate investigations about how the perceived age of a voice affects peoples' interactions with VUIs. For example, future work should explore whether, similar to work on the perceived gender of computer-generated voices [9], perceived age affects the social rules and stereotypes associated with age.

Acknowledgements

We would like to thank the reviewers who provided helpful suggestions that improved the paper, the participants who completed the experiment, and lab members who provided input.

REFERENCES

[1] 2016. Millennial Car Buyers Busting Gender Stereotypes. (2016). https://www.fi-magazine.com/322693/millennial-car-buyers-busting-gender-stereotypes

[2] Apple. n.d. Text to Speech - TTS. (n.d.). https://apps.apple.com/us/app/text-to-speech-tts/id1183892228

[3] CJ PonyParts. 2019. Men vs Women Car Buying | Car Buying By Gender. (2019). https://www.cjponyparts.com/resources/men-vs-women-car-buying

[4] W. Endres, W. Bambach, and G. Flösser. 1971. Voice Spectrograms as a Function of Age, Voice Disguise, and Voice Imitation. *The Journal of the Acoustical Society of America* 49, 6B (June 1971), 1842–1848. DOI:http://dx.doi.org/10.1121/1.1912589

[5] Google. n.d. Cloud Text-to-Speech - Speech Synthesis. (n.d.). https://cloud.google.com/text-to-speech/

[6] Earl W Huff, Naja A Mack, Robert Cummings, Kevin Womack, Kinnis Gosha, and Juan E Gilbert. 2019. Evaluating the Usability of Pervasive Conversational User Interfaces for Virtual Mentoring. In *International Conference on Human-Computer Interaction*. Springer, 80–98.

[7] Y Morishima, C Bennett, C Nass, and KM Lee. 2002. Effects of (Synthetic) Voice Gender, User Gender, and Product Gender on Credibility in E-Commerce. *Unpublished manuscript, Stanford, CA: Stanford University* (2002).

[8] Clifford Nass and Kwan Min Lee. Does Computer-generated Speech Manifest Personality? An Experimental Test of Similarity-attraction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2000) *(CHI '00)*. ACM, 329–336. DOI: http://dx.doi.org/10.1145/332040.332452 event-place: The Hague, The Netherlands.

[9] 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.

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

[11] Americo Talarico Neto, Renata Pontin M. Fortes, and Adalberto G. da Silva Filho. 2008. Multimodal Interfaces Design Issues: The Fusion of Well-Designed Voice and Graphical User Interfaces. In *Proceedings of the 26th Annual ACM International Conference on Design of Communication (SIGDOC '08)*. Association for Computing Machinery, New York, NY, USA, 277–278. DOI: http://dx.doi.org/10.1145/1456536.1456597

[12] Sharon Oviatt. 2002. *Multimodal Interfaces*. L. Erlbaum Associates Inc., USA, 286–304.

[13] Joe Parker. 2018. Bucking the Buick stereotype. (2018). https://www.northfulton.com/online_features/automotive/bucking-the-buick-stereotype/article_06938218-5e1a-11e8-81cb-ffb71eda6aa3.html

[14] Cathy Pearl. 2016. *Designing Voice User Interfaces: Principles of Conversational Experiences*. " O'Reilly Media, Inc.".

[15] Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. 2014. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior research methods* 46, 4 (2014), 1023–1031.

[16] Paul H Ptacek and Eric K Sander. 1966. Age recognition from voice. *Journal of speech and hearing Research* 9, 2 (1966), 273–277.

[17] Ulrich Reubold, Jonathan Harrington, and Felicitas Kleber. 2010. Vocal aging effects on F0 and the first formant: A longitudinal analysis in adult speakers. *Speech Communication* 52, 7-8 (2010), 638–651. DOI: http://dx.doi.org/10.1016/j.specom.2010.02.012

[18] Sara Skoog Waller and Mårten Eriksson. 2016. Vocal Age Disguise: The Role of Fundamental Frequency and Speech Rate and Its Perceived Effects. *Frontiers in psychology* 7 (nov 2016), 1814. DOI: http://dx.doi.org/10.3389/fpsyg.2016.01814

[19] Elaine T. Stathopoulos, Jessica E. Huber, and Joan E. Sussman. 2011. Changes in acoustic characteristics of the voice across the life span: Measures from individuals 4-93 years of age. *Journal of Speech, Language, and Hearing Research* 54, 4 (2011), 1011–1021. DOI:http://dx.doi.org/10.1044/1092-4388(2010/10-0036)

[20] Bozi Tatarevic. 2016. How Salespeople Stereotype New Car Buyers. (Sept. 2016). https://www.thetruthaboutcars.com/2016/09/salespeople-stereotype-car-buyers/

[21] Warrenty Direct. 2017. Car Brand Stereotypes Infographic – Warranty Direct Blog. (2017). https://blog.warrantydirect.co.uk/2017/03/07/car-brand-stereotypes-infographic/#.XL9TiehKiUk

[22] Ralf Winkler. 2007. Testing the relevance of speech rate, pitch and a glottal chink for the perception of age in synthesized speech using formant synthesis. In *International Speech Communication Association - 8th Annual Conference of the International Speech Communication Association, Interspeech 2007*, Vol. 3. 2097–2100.

[23] Stephanie Yang. 2017. 15 Cars And Common Stereotypes About Who Drives Them. (2017). https://www.theclever.com/15-cars-and-common-stereotypes-about-who-drives-them/