Effects of Vocal Mimicry on human-computer interaction in the context of personal assistants

Bilge Mutlu
Madison, WI
mutlu@wisc.edu

Yue Sun
Madison, WI
ysun276@wisc.edu

Andrew Geng
Madison, WI
ageng@wisc.edu

John Balis
Madison, WI
balis@wisc.edu

ABSTRACT
Mimicry can be uncanny when it comes from sources other than humans. In the context of human-computer interaction, there is unexplored territory regarding the effects of computer-human voice mimicry. In our study, we explored the effects of voice mimicry on affinity, listening comprehension, and influence in the context of human interaction with personal assistants. The results observed in our study indicate a possible negative correlation between agent influence and voice mimicry. Additional investigation will be needed in order to confirm our observations.

Author Keywords
Mimicry; Personal Assistants; Voice; Speech; Conversational Agents

BACKGROUND
Voice has become a critical interaction feature in many commercial products in the last few years. Well-known commercially visible conversational agents such as Amazon Echo and Google Home heavily rely on their speech functionality to interact with users. These systems help the user to schedule, search, recommend and complete other daily tasks using voice instruction. Conversational agents are also commonly known as virtual intelligent personal assistants (Porcheron, Fischer, Reeves and Sharples, 2018). Conversational assistants provide an ideal platform for vocal HCI research, as they are extremely prevalent, relevant to modern inquiry in the field of HCI, and provide a primarily vocal interface. Conversational agents are an area under unprecedented emphasis and have the potential to influence interaction between humans and computers along the lines of quality of information transmission, trust, and affinity. Although there are numerous previous studies regarding voice features of personal assistants, voice mimicry of the user's own voice is still an uncharted area. Our study will focus on how of user voice mimicry affects the interaction between users and personal assistants.

Research into mimicry of human aspect and behavior by computational agents has been an extensive and a well defined genre within HCI. One of the foundational research relating to this computational mimicry was Nass, Kim and Lee (1998) study to measure the difference in influence between receiving negative evaluation from an audio-visual image of oneself on a computer screen, and that of receiving the same criticism from a image of a different face (p. 148). This study found that users received criticism paired with an image of their own face were more likely to accept the accuracy and objectivity of the criticism. (p. 152). However, they did not generalize the results to using one's own voice and in a non-negative contexts within their study.

Although we found no previous study exploring the effects of voice mimicry on human and computer interaction directly, the influence of voice similarity has been shown to have significant effect within other studies. In Dahlback, Wang and Nass's (2007) study, American and Swedish participants listened to some guiding tourism information on a website presented in English with either an American or Swedish accent. The result showed that not only did the participants prefer accents that are similar to their own, but also they viewed the same-accents speakers as being more knowledgeable than the different-accent speaker even if when the latter speaker actually provide better-known information (Dahlback, Wang and Nass, 2007, p. 1553). Dahlback presents a theory of similarity attraction which demonstrates that people are more inclined to trust others who have similar personality, ethnic background, accent etc. (Byrne and Clore, 1986). These findings inspired our study to explore these principles by elevating similarity to the level of voice mimicry, and how, under these conditions, similar attraction influences users'comprehension level.

Another study, which also follows along the same vein of computational mimicry, conducted by Nass and Lee (2001), found that participants who communicated with a computer voice matched to their own personality, reported higher affinity to the interface in accordance with similar-attraction theory (p. 171). However, the study was not conducted with the concept of vocal mimicry as a dependent variable.

ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

Contradictory to some of the studies proposed above is the effect of the uncanny theory. Uncanny valley theory suggests an overly humanistic agent could prove to have a negative effect on users and cause a general dislike towards the agent (Mori, M., Macdorman, K., & Kageki, N., 2012). The bases of this theory has been shown to occur in several publications and the theory itself is a widely known and popular theory. Conceptually simple and logical, this theory has proven to be observed already to some extent in Mitchell, et al publication and is has potentially similar consequences in vocal mimicry (Mitchell, W. J et al, 2011). However the publication failed to mention specifics cases under text-to-speech systems and personal assistants, thus making it inconclusive for a generalizable case (Mitchell, W. J et al, 2011).

The absence of existing research makes a compelling case for further exploration the effects of voice mimicry on learning comprehension, reported affinity, and agent influence. In order to determine the effects of voice mimicry on human-agent interaction, we tested for three categories of variable: Listening comprehension, subject relationships to and perceptions of personal assistants, and influence of agent suggestion.

Our proposed conceptual hypotheses are the following three observations. *"Voice mimicry Will increase the user's listening comprehension compared to using a random voice." "Voice mimicry will increase user affinity towards the conversational agent." "Voice mimicry will increase influence of the given agent's suggestion."*

METHODS

Subjects

8 pilot participants (8 male) are recruited from Human and Computer Interaction Research courses and Undergraduate Penguin Lab at University of Wisconsin Madison. Pilot participants are randomly assigned to either the voice mimicry group or the control group. Subjects in the treatment group interacted with a personal assistant which communicated with an approximation of their own voice, while subjects in the control group interacted with a voice randomly selected from the remaining pool of other subjects voices.

Procedure

The experiment consisted of two sessions. For the first session, all pilot participants were welcomed to a quiet room in HCI lab. After they signed the consent form, researchers instructed pilot participants to record their voices by a text-to-speech voice mimicry application called Lyrebird. They were asked to read and record around 40 sentences -some examples being "The log file reveals that nobody ever read the end user license agreement during installation"- for generating the voice mimicry system paired with the recorder's voice. After the recording session, pilot participants will be asked to wait around 15 minutes during which researchers will generate the audio files used in the next session.

For the second session, the pilot participants were first guided to the conference room of HCI lab where a Google Mini was set up. For the listening comprehension test, we directed pilot participants to ask some questions to an app we created for

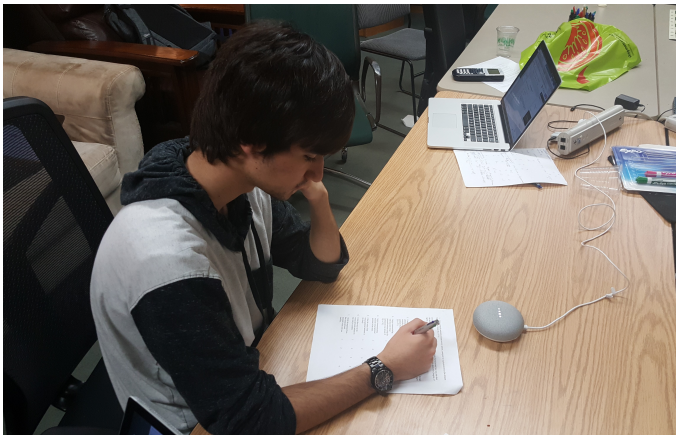


Figure 1. A Researcher Demonstrates a testing session

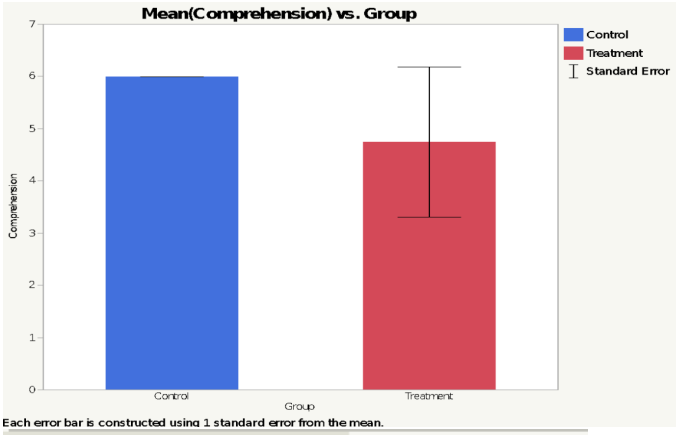
Google Mini using Dialogflow, with the goal of learning about the Dutch Painter Johannes Vermeer. The pilot participants were asked to answer some questions based on the information they just listened to after talking to the personal assistant. Influence was measured based on user-ranking of the importance of 12 objects in a hypothetical survival scenario, before and after receiving advice from a Google Mini. The program on the Google Mini explained the usefulness 6 of the 12 objects and advised against choosing the other 6. The explains of the objects were prewritten using previously found desert survival scenario forms used by Bradner and their study team (Bradner, 2002). After the desert survival scenario test, pilot participants were asked to fill in a questionnaire to measure their affinity level toward the Google Mini they just communicated with. At the end of the study, researchers would debrief the study to pilot participants and answer any lingering questions.

Measurement

Listening comprehension performance was measured as the number of correct answers out of twelve on a multiple choice test applied to the subjects. Questions like "What was the name of Johannes Vermeer's father?"and "How many children did Johannes'wife give birth to?"were presented in the questionnaire.

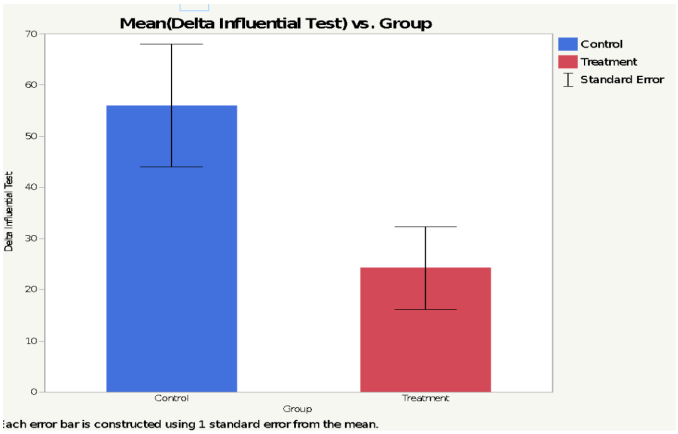
For the desert survival scenario test, we measured influence as the sum of the absolute values of the differences of the numerical ranking of each tool between the pilot participants'original data and recovered data after they communicating with the personal assistant.

The affinity questionnaire consists of one Yes/No question and six Likert scale questions (1 is Strongly Disagree and 5 is Strongly Agree) regarding the role of personal assistants in their life and their subjective experience with the session. These questions ranged from traditional likert scale questions such as "Did you enjoy interacting with the personal assistant?". The testing scenario and model of our likert scale questionnaire was heavily modeled off of previous literatures used by other studies within HCI (Bock, et al 2016, Nass, C., & Lee, K. M 2001, Warmbrod, J. R. 2014).



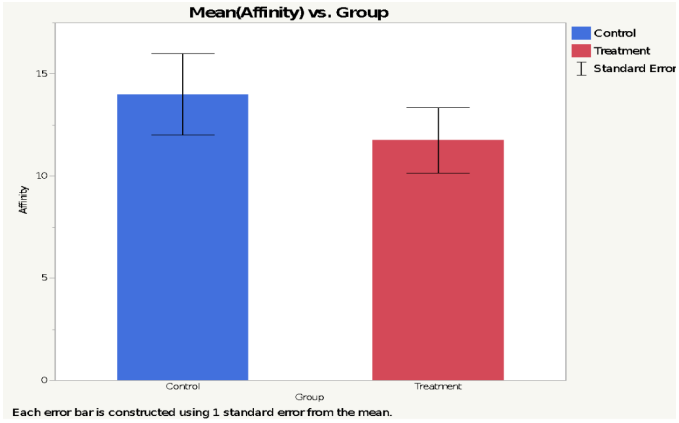
Mean	5.1666667
Std Dev	2.3166067
Std Err Mean	0.9457507
Upper 95% Mean	7.5977963
Lower 95% Mean	2.735537
N	6

Figure 2.



Mean	34.833333
Std Dev	21.994696
Std Err Mean	8.9792972
Upper 95% Mean	57.915352
Lower 95% Mean	11.751315
N	6

Figure 3.



Mean	12.5
Std Dev	3.0166206
Std Err Mean	1.2315302
Upper 95% Mean	15.665749
Lower 95% Mean	9.3342508
N	6

Figure 4.

RESULTS

The data collected through our testing models were collected and tested with the control being compared to the treatment group. The 3 comparison models are showcased in figures 2, 3, and 4.

In figure 2, we conducted an analysis of variance test on the listening comprehension scores we obtained from the control and from the treatment. The results failed to show a significant effect of voice mimicry on comprehension ($F(1, 5) = 0.337$, $p > 0.05$) (Please view figure 2 for additional information).

In figure 3, we conducted an analysis of variance test on the influence test scores we obtained from the control and from the treatment. The results failed to show a significant effect of voice mimicry on the level of influence ($F(1, 5) = 0.089$, $p > 0.05$). However, given an alpha value of 0.1 then we would obtain a significance. As a result of this observation, we feel there would be a significance shown given we had a larger sample size than 8 and if we removed an observable outlier which skewed our distribution. Although this is inconclusive we will continue on in the discussion section with the hypothetical that we did indeed view a significance.

In figure 4 we conducted an analysis of variance test on the affinity level scores we obtained from the control and from the treatment. The results failed to show a significant effect between the voice mimicry and the affinity level ($F(1, 5) = 0.697$, $p > 0.05$) (Please view figure 4 for additional information).

DISCUSSON

Our study was conducted in response to the question posed by Nass, C., and etc (Nass, C., & Lee, K. M., 2001). Within

this paper, one of the proposed questions stated within the discussion for further research, was a proposed study of vocalized mimicry and its role in interactions between agents and humans. We hope that our conclusion has shed some light into the matter and will prompt additional studies to be conducted with the hope of providing a more conclusion answer to the role of vocal mimicry in human-computer interaction.

One of the faults within our study was the limited sample size presented in our control and treatment groups. Our sample size was not large enough to allow us to generalize the results to a larger population. This means that, although correlation was observed, it cannot be definitively proclaimed that there was a significant enough correlation as to be indicative of general population trend. Future studies on the effects of voice mimicry on human-computer interaction should be set on obtaining a large enough sample to conclusively demonstrate the influence of vocalize mimicry on a more generalized population. Other noticeable populational problem is the location of the study. We used a convenience sampling methodology for obtaining the test subjects, and our study population is limited to college students at UW Madison. Additional studies including participants from other regional populations would be necessary in order to provide generalizable results.

As with all scientific pursuits, additional studies must be done in order to correct and make certain of our observations. This is especially important with this study, as it is heavily interaction-based; differing settings and conditions could potentially alter our study results. This does not mean that our study should be ignored due to its limitations, but rather we hope that our study will spark additional interest in the field of HCI regarding human-agent interactions based on vocal mimicry.

This study showcased possible correlation between vocal similarities and affinity to personal assistants. This poses a logical first question: what specific factors influence affinity and influence the most? In other words, boiling down the study proposed to altering the specifics factors which causes the most attractiveness to personal assistants. Some parameters that could potentially be studied are measurements includes frequency, accented speeches, and dictions. These could provide extremely fascinating results which would allow researchers to pinpoint specific vocal properties that would allow for a more dynamic and interactive experience for the user when communicating with a personal assistant.

Another possible area to explore is the potential impacts on interaction when an agent voice shares properties with individuals whose voices are commonly heard, these voices could be of celebrities, politicians, or even close friends and families. Studying how people communicate with a personal assistant that shares commonality with someone of frequent contact could given valuable insight into the how humans could better communicate with agents.

Based on the sample data that we have collected in our exploratory study, we can see some interesting correlations. Our data indicates a possible positive correlation between vocal mimicry and agent influence. This is potentially very interesting as it is very much in opposition to our hypothesis as we hypothesized a given position correlation due to similar attraction theory. There are a couple of potential explanations that our team has looked into as being a possible causes.

The first possible explanation for this phenomenon is uncanny valley theory. This is a generalized theory which states how a potentially overly humanistic agent could prove to have a negative effect onto the given user interaction and cause a general distaste towards the agent (Mori, M., Macdorman, K., & Kageki, N., 2012). Translated in terms of our study, the uncanny valley effect would be observed through a observability low affinity towards the given agent and a generally lower influence on the given user. We believe that it is highly possible that this phenomenon is occurring within our given samples generated. This can be observed through a general, although minimal, trend in affinity and obvious down trend in the influence. As stated before, we are not able to concretely claim such a phenomenon, however we do see correlations which pinpoint a possible uncanny valley effect. Additional studies need to be conducted in order to definitively claim a uncanny valley phenomenon.

There are several important implications if the proposed uncanny valley model was truly occurring. First off, this is incredibly important phenomenon to observe as it means technology has reached the point of being uncanny thus indicating that the given technology is very close to reaching a point of being indistinguishable compared to a given real vocal sample. However this also infers that there is still a gap that the user will realized within their interactions with the personal assistant. But, regardless of its limitations, this is a very exciting time as judging for our collected observations, vocal mimicry technology will reach a point of extreme relevance in the near future.

The uncanny valley effect also gives indications towards current models of technologies that have traditionally collected vocal samples as their text-to-speech system, as the current models essentially mimics the control group as proposed in our given study which has definabily stronger affinity and influence (Mitchell, W. J., al etc). However it is important to realize the complications of being complacent with the given state of personal assistants. Although our study failed to show a positive correlation between vocal mimicry and affinity, attentiveness, and influence measurement models, there are an infinitesimal quantity of additional parameters that our study did not cover. In general, many of our test subjects expressed degrees of fascination with many recognizing their own voice and becoming inherently more interested in the study. To some extent, we believe our study did not measure all the degrees of affinity, attentiveness, and influence as we still believe, due to our end of study question and answers sessions after the given test sessions and the general interest for our given research for the participants that observed their own voice.

ACKNOWLEDGMENTS

Special Thanks to the lyrebird team and Alex for their gracious allowance for us to use their proprietary technology. You can find them at lyrebird.ai or contact them through contact@lyrebird.ai.

Also, most importantly, a very special thanks the UW Madison Legal department for their tireless efforts in this process. We have great respect to their team and their lawyers, and we wish them the best in their future endeavors.

REFERENCES

1. Byrne, D., Clore, G. L., & Smeaton, G. (1986). The attraction hypothesis: Do similar attitudes affect anything? Journal of Personality and Social Psychology,51(6), 1167-1170. doi:10.1037//0022-3514.51.6.1167

2. Nass, C., & Lee, K. M. (2001). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. Journal of Experimental Psychology: Applied,7(3), 171-181. doi:10.1037//1076-898x.7.3.171

3. DahlbÃd'ck, N., Wang, Q., Nass, C., & Alwin, J. (2007). Similarity is more important than expertise. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI 07. doi:10.1145/1240624.1240859

4. Nass, C., Kim, E., & Lee, E. (1998). When my face is the interface. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI 98. doi:10.1145/274644.274667

5. Mori, M., Macdorman, K., & Kageki, N. (2012). The Uncanny Valley [From the Field]. IEEE Robotics & Automation Magazine,19(2), 98-100. doi:10.1109/mra.2012.2192811

6. Mitchell, W. J., Szerszen, K. A., Lu, A. S., Schermerhorn, P. W., Scheutz, M., & Macdorman, K. F. (2011). A Mismatch in the Human Realism of Face and Voice Produces an Uncanny Valley. I-Perception,2(1), 10-12. doi:10.1068/i0415

7. Warmbrod, J. R. (2014). Reporting and Interpreting Scores Derived from Likert-type Scales. Journal of Agricultural Education, 55(5), 30-47. doi:10.5032/jae.2014.05030

8. Bock, B. C., Lantini, R., Thind, H., Walaska, K., Rosen, R. K., Fava, J. L., . . . Scott-Sheldon, L. A. (2016). The Mobile Phone Affinity Scale: Enhancement and Refinement. JMIR MHealth and UHealth,4(4). doi:10.2196/mhealth.6705

9. Nass, C., & Lee, K. M. (2001). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. Journal of Experimental Psychology: Applied,7(3), 171-181. doi:10.1037//1076-898x.7.3.171