Investigating Credibility of Voice-Controlled Intelligent Personal Assistants Based on Linguistic Mimicry and User Agreeableness

Shyamal H Anadkat
University of Wisconsin-Madison
Madison, Wisconsin
anadkat@wisc.edu

Sanjay Rajmohan
University of Wisconsin-Madison
Madison, Wisconsin
rajmohan@wisc.edu

Hiroshi Shu
University of Wisconsin-Madison
Madison, Wisconsin
hshu5@wisc.edu

Karim Nassef
University of Wisconsin-Madison
Madison, Wisconsin
nassef@wisc.edu

ABSTRACT
As voice-controlled intelligent personal assistants (IPAs) become more capable, the question of how much users trust their judgement and find them credible is becoming more important. In this study, we explore the perceived credibility of voice-controlled IPAs in relation to linguistic mimicry and try to answer the questions of whether users find voice-controlled IPAs more credible when they employ the same speech habits and nuances as the user and if results differ depending on where users fall on an IPIP "Agreeableness" scale. We developed a recommender system built on Google DialogFlow in which we tested whether or not users tended to follow the recommendations given by the IPA more when it employed linguistic mimicry specific to that user. We controlled with a stock version of the recommender system which employed no linguistic mimicry. Preliminary results found that overall, users who scored higher on the Agreeableness scale tended to accept the recommendation of the IPA, even more so when the IPA employed linguistic mimicry. However, the study needs to be done with more participants to ensure that this trend is accurate and not the result of a type-1 error.

AUTHOR KEYWORDS
intelligent personal assistants; credibility; voice-assistants; linguistic mimicry; agreeableness

ACM CLASSIFICATION KEYWORDS
H.5.m. Information Interfaces and Presentation (e.g. HCI).

INTRODUCTION
Voice-controlled intelligent personal assistants (IPAs) are becoming pervasive, and as a result HCI professionals should be more aware about the issues pertaining to their perceived credibility. Credibility is of utmost importance when personal assistant agents, which are a subset of computer products, act as knowledge sources as well as decision aids [1]. This is especially true in the context of recommender agents. Previous research has investigated how human-human credibility markers can apply to human-computer interactions. For example, literature has shown that certain types of linguistic mimicry in human-agent conversations can potentially correlate to the establishment of trust [7]. Moreover, subjects have shown to prefer interaction with a computer agent possessing traits similar to their own personality, and were indeed more satisfied during such interaction [3]. We plan to extend these developments to personal voice assistant agents and cover the research gap by determining the interaction effects of agreeableness in users' personality and IPA's linguistic mimicry on the credibility of the recommendations provided by the agent. Out of the five basic dimensions of personality, we will be looking at incorporating the spectrum of agreeableness. To explore this research question, we survey the agreeableness of the participants, and setup an interaction task with a personal assistant that is programmed to mimic the subjects' linguistic style. We use Google Home Mini, and the DialogFlow framework to build a custom conversational agent used for the task.

Our findings suggest that users overall tend to follow recommendations from the personal assistant agent which employs linguistic mimicry, and users who are more agreeable tend to follow the agent's recommendation as well. In addition, the influence of personality (agreeableness) was greater than the linguistic mimicry manipulations. The results of this work can be used to better understand the conversational aspects employed by personal assistants, and help build credibility with potential users of the system. In addition, it can also help develop and design recommender systems which are more reliable and are amiable with the user's personality.

RELATED WORK

Credibility of IPAs

There is a scattered array of research into IPAs and linguistic mimicry. Interestingly, this is a field that has started to get a lot of attention in recent times. The first relevant piece of literature to consider is about PTIME [2], a semi-automated calendar scheduling system which was used to measure human perception of credibility with its users. Its relevance stems from the quantitative model they built to measure credibility. This started our conversation on how credibility could be assessed.

Applications of Linguistic Mimicry

There have been multiple studies that have used linguistic mimicry as a dependent variable, using it as a measure to assess exogenous variables. A prime example of this is literature which looked at linguistic mimicry in text-based conversation. Here, linguistic mimicry was the dependent variable, and the study concluded that conversation partners who share a high level of trust tend to employ a high level of linguistic mimicry [7]. While their research focuses on employing mimicry in chat based conversation, we are interested in extending the scope of linguistic mimicry over conversational agents which are voice-controlled.

Linguistic Mimicry in Natural Language IPAs

One cannot talk about linguistic mimicry without talking about the popularization of natural language in IPAs, especially on the user-facing side with Siri, Google Assistant, and Alexa. There have been numerous studies looking at the way users interact with these IPAs in a more fluid manner. Subsequently, previous studies have shown the use of implicit language indicators (as opposed to explicitly stating every individual fact to the IPA), which contribute to a smoother interaction [9]. Our study aims to build on this by recognizing that not all users imply information in the same way. Adaptive linguistic mimicry can be a helpful tool in tailoring an IPA's speech style to every user's mannerisms, as opposed to settling for a compromise that is the mean of all users, where which no user is fully mimicked.

Setting Objective Standards for Evaluating Linguistic Mimicry

In order to use linguistic mimicry effectively, there has to be a standardized objective approach to its manipulation. In the documentation for the Linguistic Inquiry and Word Count engine, Pennebaker et al presented an extensive list of speech elements in a table. Our study built on these elements of speech detailed within to build linguistic mimicry models in Google DialogFlow. Study administrators had copies of the table in front of them while the study was being conducted, marking where users were displaying elements of speech which could be later coded into the IPA [5].

DESIGN GOALS

The main concerns of designing the interaction flow for the personal voice assistant would be that the responses provided by it have to be smooth, have the right context with regards to user responses and should give enough time for the users to respond. Recognizing the difference in pronunciation of words of the users by the personal assistant is important. If the IPA could not understand the user response or the response was not related to the IPA's prompt, the IPA needs to ensure that a default prompt is provided which tells the user that the IPA could not hear what the user said and ask them to repeat their response again. Designing an interaction where the IPA provides different answers depending on the user responses is crucial and should be accommodated by the system. For instance, the IPA should not give a negative response when the user says something positive. The system should provide recommendations based on the responses given by the user and the recommendations given must be such that they should not be biased in any way. In other words, the recommendations should not favor one choice over the other to assure that users do not always pick the same choice. Since Google Assistant cannot capture filler words such as *"uhm"* and *"uh"*, and it is very difficult to effectively identify all of the linguistic word types and instances utilized by a user including assent, insight, positive emotion and so on, session 1 was conducted in which users are asked some general questions and output variables necessary to perform linguistic mimicry manipulations were retrieved from the answers they provide.

Google Assistant has inbuilt Natural Language Processing (NLP) and the voice that narrates the prompts closely resembles that of a human, thus making the interaction smooth. To provide enough time for a user to respond, Google Assistant waits for a few seconds before asking the prompt, and repeating the sequence until a response is received. When the user says something that does not correlate with the prompt given by the IPA or the response given is misinterpreted, the IPA provides a default response of "*I'm sorry, I couldn't hear that*" and asks the prompt again for the user to respond one more time. However, due to the limitations of DialogFlow, the interaction system terminates when the user response does not match the context of the prompt asked three times in a row. The interactions are designed in a way that the IPA provides different responses to the user that aligns with the answers that they provide. For instance, the IPA asks the user *"How's the day treating you?"* and if the user says something positive such as *"It's good"*, the IPA will capture the positive context and take the positive branch of the interaction and its response would be *"I'm glad to hear that your day is good*". If user response is negative such as *"It's pretty bad"*, the IPA will take the negative branch of the interaction and its response will be *"I'm sorry to hear that..."*. The choices presented to the users have very similar ratings and costs, and the recommendations given by the IPA describe the advantages of both the choices while recommending one solely based on the user responses provided to some of the questions asked previously in the interaction. The names of the choices, the towns that they come from and their genres are all made up so that users who might have a preference to names or genres would not be biased while picking a choice. For the Google Assistant to perform linguistic mimicry for a particular user, the data retrieved from session 1 is used to manually change the prompts that are a part of the interaction. For example, if the user used "*uhm"* and *"okay"* often during session 1, these words will be used in some of the prompts similar to the following: "*I see"* would be changed to *"Uhm, okay"*. Care was taken to ensure that the same mimicry words do not appear very frequently during the interaction.

SYSTEM DESCRIPTION

The study involved target participants interacting with a voice assistant (Google Home Mini) to accomplish a task. The assistant required to ask some questions regarding the subject's preferences, understand the context of the response, and carry on the interaction.

We used the Google Home Mini as the intelligent personal assistant agent. Custom test apps were developed and deployed onto the Google mini using DialogFlow (https://dialogflow.com/; formerly Api.ai) which is an end-to-end intuitive development suit for building conversational agents and interfaces. DialogFlow is powered by Google's Machine Learning and NLP. The custom agent we used is essentially a container that contains intents, entities, and other responses we intend to deliver to the user. In order to process user input and fulfill the desired action, we utilized DialogFlow's built-in fulfillment functions.

STUDY

In this work, we developed a wedding planner simulation on the Google Home to study the credibility of voice assistants based on subjects' agreeableness and the linguistic mimicry employed by personal assistant agent. We describe the protocol and apparatus below.

Procedure

We invited 11 participants, aged between 18 to 65, to engage in session 1 and session 2 of the study (described below). The participants were recruited from across Madison, WI through flyers, university email blasts, and word of mouth. Participants were compensated $10/hour for their time.

We followed the between participants design. Half of the recruited participants were randomly assigned to be exposed to linguistic mimicry manipulations and the rest half weren't. The participants who weren't exposed to linguistic mimicry did not participate in session 1.

The experiment took place in the Human Computer Interaction Lab in the Computer Sciences facility. Participants were briefed upon arrival, but weren't instructed in detail about the research goal.

Session 1: Interview Study

The participants who were exposed to linguistic mimicry manipulations underwent the first session of the study. Session 1 involved a 10-15 mins long casual recorded conversation with the subject based on pre decided questions. These questions were mostly open-ended and required no specific responses. The intent of of this session was just to gather linguistic mimicry clues based on subjects' responses. The users' responses were later transcribed and checked against the Linguistic Mimicry and Word Count (LIWC) manual in order to identify their selection of words and output information (depicted in Figure 1).

Some categories of output variables and linguistic dimensions that were considered for manipulations from the LIWC2015 include Psychological Processes (Positive and negative emotions, anxiety, anger, sadness etc), Cognitive processes (insight, causation, tentative, certainty, etc), Informal language (netspeak, assent, nonfluencies, and fillers), summary language variables and other phrases [6].

LIWC Results

Details of Writer: 21 year old Female
Date/Time: 1 May 2018, 7:34 pm

LIWC Dimension	*Your Data*	*Personal Texts*	*Formal Texts*
Self-references (I, me, my)	7.87	11.4	4.2
Social words	9.09	9.5	8.0
Positive emotions	4.35	2.7	2.6
Negative emotions	0.61	2.6	1.6
Overall cognitive words	7.26	7.8	5.4
Articles (a, an, the)	4.97	5.0	7.2
Big words (> 6 letters)	12.53	13.1	19.6

The text you submitted was 1309 words in length.

Figure 1: LIWC results for one of the participants after Session 1

There was usually a gap of couple days before session 2 for the participants who participated in session 1. During this time, we reviewed the session 1 interview and programmed a custom agent in Dialogflow which employed linguistic mimicry for the respective subject.

Session 2: Interaction with custom IPA

Figure 2: Researcher piloting Session 2 of the study. The handout is a visual reference to the shortlisted florist and wedding band descriptions.

Session 2 involved the subjects interacting with a custom Google home agent programmed specifically for them. The subjects for whom linguistic mimicry was not enabled had one common agent without any manipulations. The custom agent was embedded with responses that matched subjects' choice of specific words, phrases and LIWC output variables, so as to employ linguistic mimicry. Few of the linguistic mimicry dimensions that were considered included Lexical (repetition of words/phrase, connecting words etc), Syntactic (mimicking the phrase/sentence structure), and emotion-related [8].

```
// ******************* //
// set after session 1 //
// *******************//
var negation_mimicry = '';   //example: no, not, never
var quatifier_mimicry = '';  //example: few, many
var anxiety_emotion = '';    //worried, fearful, nervous
var positive_emotion = '';   //example: love, nice, sweet
var negative_emotion = '';   //example: hurt, ugly, nasty

var nonfluency = '';        // example: er, hm, uhmm
var fillers = '';           // blah, Imean, youknow
var assent = '';            // agree, OK, yes
var insight_mimicry = '';   // think, know, consider
var inclusive_mimicry = ''; // and, with, include
var causation_mimicry = ''; // because, hence, effect
```

Figure 3: Sample variables for Linguistic Mimicry manipulations configured in the DialogFlow fulfillment

The actual task for session 2 consisted of the participant deciding a florist and a wedding band for a hypothetical person's wedding. A wedding scenario was chosen as it gave the flexibility of making unbiased recommendations, and involved more participant involvement. The agent initially made small talk with the participant, asking questions regarding their wedding preferences and experiences. Later, the participants were presented with descriptions of two shortlisted florists and two shortlisted bands. The agent then made balanced recommendations based on subjects' preferences. The participant had to decide on a florist and band respectively. Manipulations were embedded throughout the dialogue, even when the agent furnished recommendations.

Upon concluding the session 2, participants were asked to complete the NASA Task Load Index (TLX), System Usability Survey (SUS) questionnaire, and a custom designed likert scale questionnaire. In addition, we conducted semi-structured interviews to know more about the participants' experience. We finally de-briefed participants about the study, and answered any further questions they might have.

Based on the notion of similarity/attraction theory, we posit to see an interaction effect between subjects who are highly agreeable, subjects who are not agreeable, and IPAs that incorporate various linguistic mimicry features. We expect to see this effect for both the groups: subjects who are highly agreeable and subjects who aren't agreeable.

DISCUSSION

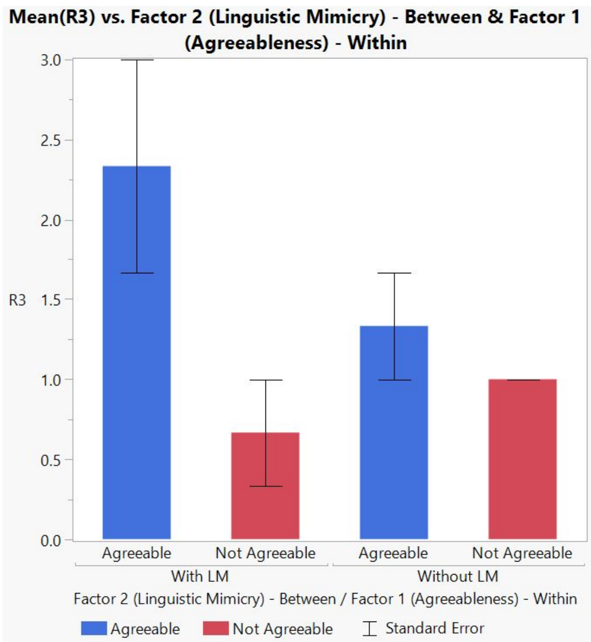

Figure 4: Mean (R3) vs Factor 2 (Linguistic Mimicry) & Factor 1 (Agreeableness)

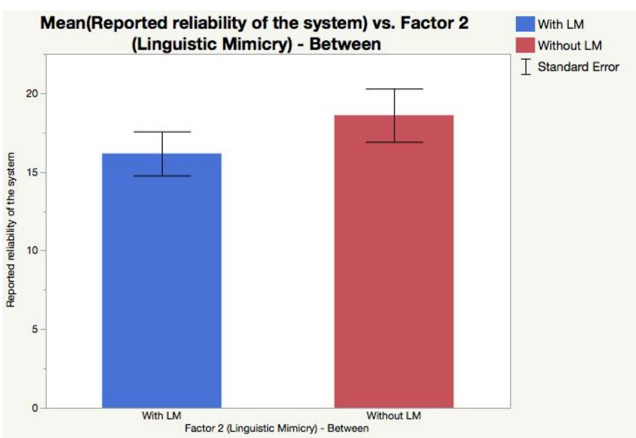

Figure 5: Mean (Reported reliability) vs Factor 2 (Linguistic Mimicry)

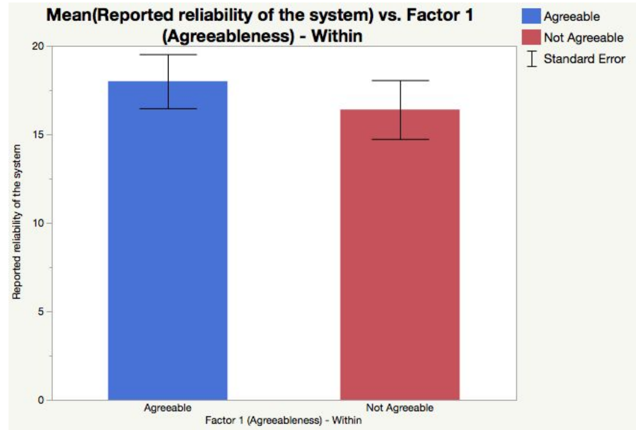

Figure 6: Mean (Reported reliability) vs Factor 1 (Agreeableness)

All of the recruited participants successfully finished the study. Our system seemed to be viable for deployment as indicated by the mean SUS score of 74.54 according to our users. Overall, we found that participants tend to follow the recommendation given by the system equipped with linguistic mimicry. Participants who scored higher in the IPIP agreeableness survey (indicating a more agreeable personality compared to those who scored low in IPIP test) are shown to be more inclined to trust the recommendation given by the DialogFlow agent.

Narrative Description

We recorded the final result as R3. It was calculated by assigning a weight of one for the florist choice and a weight of two for the band choice. We interpreted it this way because we based the system's florist recommendation output on linguistic variables received during the conversation between the participant and the system. Subsequently, for the band recommendation, we deliberately picked the opposite band from users' initial choice as the system's recommendation. Hence, we placed more weight on the wedding band recommendation than the florist one. For each recommendation that the user followed or went against, we assigned a value of either 1 or 0 respectively to the measure in R3.

Our analysis found a significant effect of linguistic mimicry on R3. $F(1,10)=2.66$, $p=0.618$. The R3 result employed with linguistic mimicry in the recommendation system ($M=1.500$, $SD=1.225$) was significantly higher than the result without the application of linguistic mimicry ($M=1.2$, $SD=0.447$). We also found that the personality (agreeableness) of the participants (agreeable vs non-agreeable) has some extent of effect on R3. $F(1,10)=4.65$, $p=0.059$. The R3 result for more agreeable people ($M=1.833$, $SD=0.983$) was significantly higher than the result for non-agreeable people ($M=0.800$, $SD=0.447$). Based on the F ratio result we get from the 11 participants, Factor 2 (agreeableness) had more influence on R3 than Factor 1 (Linguistic Mimicry).

After the study, participants were informed to fill out the post study questionnaire that measures the credibility and satisfiability of our system. We asked them to provide a score on a scale of one to five to agree with statements such as *"I would trust the IPA to help plan my wedding"*. Surprisingly, we found out that participants tend to trust the system without linguistic mimicry (Figure 5) even though in general, they tend to follow the system's recommendation that had linguistic mimicry enabled.

As we expected, more agreeable users tend to trust the recommendation system (Figure 6). An interesting observation is the interaction effect between attitude towards the linguistic mimicry IPA and agreeableness. One could, on the flipside, interpret the results as that less agreeable people will tend to feel more negatively towards an IPA which mimics their speech.

LIMITATIONS

The main limitation of this study is that the results were not deemed statistically significant ($P=0.618$). However, observing the charts, one can see that there does seem to be a correlation between agreeableness and statistical significance, with whether or not users embraced the suggestions of the IPA. These results present a solid framework to be built upon on a larger scale of participants, something that was not possible under the time constraints under which this paper was written. Increasing the number of participants would make it easier to draw a conclusion. Given these promising initial findings, one could say that this is an avenue worth exploring.

While our experiment was extensive, there is definitely room for improvement and to built up on in the future. Linguistic mimicry is powerful by itself, but oftentimes, it was clear to users that the IPA was going out of its way to seem relatable (with esoteric phrases like "dope!"). This is symbolic of the bigger issue that even if we modelled each test subjects' speech patterns perfectly, the execution would have bottlenecked this achievement. In this study, we did not manipulate important characteristics of speech that extend past linguistic mimicry, elements like inflection and pitch. An experiment which is able to control for those variables could lead to more conclusive results.

Another potential limitation could be the presence of the study administrators in the room. While study administrators were overwhelmingly hands-off and only intervened when absolutely necessary, it could be argued that the interaction between their presence as well as the fact that this was a new, unfamiliar system to the user, changed their behavior in relation to how they would act were there no study administrators around. One way to implement this could be to simply have a panic button or trigger phrase which would alert the administrator if attention was needed, but otherwise having them left out of the study room entirely.

ACKNOWLEDGEMENTS

We thank Professor Bilge Mutlu for his guidance through this study. We also thank all the participants for their time and consideration in fulfilling our study, as well as our acquaintances and friends for helping us through the pilot studies.

REFERENCES

[1] Fogg, B. J., & Tseng, H. (1999). The Elements of Computer Credibility. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 80–87). New York, NY, USA: ACM. https://doi.org/10.1145/302979.303001

[2] M. Berry, Pauline & Gervasio, Melinda & Peintner, Bart & Yorke-Smith, Neil. (2011). PTIME: Personalized Assistance for Calendaring. ACM TIST. 2. 40. 10.1145/1989734.1989744.

[3] Nass, C., Moon, Y., Fogg, B. J., Reeves, B., & Dryer, D. C. (1995). Can computer personalities be human personalities? International Journal of Human-Computer Studies, 43(2), 223–239. https://doi.org/10.1006/ijhc.1995.1042

[4] Pauline M. Berry, Melinda Gervasio, Bart Peintner, and Neil Yorke-Smith. 2011. Ptime. *ACM Transactions on Intelligent Systems and Technology*2, 4 (January 2011), 1–22.

[5] Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (n.d.). *The Development and Psychometric Properties of LIWC2007*(Publication).

[6] Pennebaker, James W, Ryan L Boyd, Kayla Jordan, and Kate Blackburn N.d.*The Development and Psychometric Properties of LIWC2015: 26.*

[7] Scissors, L. E., Gill, A. J., & Gergle, D. (2008). Linguistic mimicry and trust in text-based CMC. *Proceedings of the ACM 2008 Conference on Computer Supported Cooperative Work - CSCW 08*. doi:10.1145/1460563.1460608

[8] Swaab, R. I., Maddux, W. W., & Sinaceur, M. (2011). Early words that work: When and how virtual linguistic mimicry facilitates negotiation outcomes. Journal of Experimental Social Psychology, 47(3), 616–621. https://doi.org/10.1016/j.jesp.2011.01.005

[9] Vtyurina, A., & Fourney, A. (2018). Exploring the Role of Conversational Cues in Guided Task Support with Virtual Assistants. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI 18*. doi:10.1145/3173574.3173782