INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Mirroring to Build Trust in Digital Assistants

Katherine Metcalf, Barry-John Theobald, Garrett Weinberg, Robert Lee, Ing-Marie Jonsson, Russ Webb, Nicholas Apostoloff

Apple Inc., 1 Apple Parkway, Cupertino, CA, 95014, USA

{kmetcalf, bjtheobald, gweinberg, perihare}@apple.com,
{ingmarie, rwebb, napostoloff}@apple.com

Abstract

We describe experiments towards building a conversational digital assistant that considers the preferred conversational style of the user. In particular, these experiments are designed to measure whether users prefer and trust an assistant whose conversational style matches their own. To this end we conducted a user study where subjects interacted with a digital assistant whose response either matched their conversational style, or did not. We found that people strongly prefer a digital assistant that mirrors their "chattiness" and that this preference can be reliably detected.

Index Terms: speech, mirroring, conversation

1. Introduction

Long-term reliance on digital assistants requires a sense of trust in the assistant and its abilities. Therefore, strategies for building and maintaining this trust are required, especially as digital assistants become more advanced and operate in more aspects of people's lives.

In the context of human-human interactions, people use certain behaviors to build rapport with others [1]. One such behavior is "mirroring" [2], which occurs over the course of an interaction when people "reflect" some of their partner's behaviors back to them, e.g. adopting the posture or facial expression of the conversational partner. This phenomena, often created via the process of entrainment, has also been referred to as: mimicry, social resonance, coordination, synchrony, attunement, the chameleon effect, and so on. We hypothesize that an effective method for enhancing trust in digital assistants is for the assistant to mirror the *conversational style* of a user's query. We start by exploring a single aspect of conversational style, the degree of "chattiness." We loosely define chattiness to be the degree to which a query is concise (high information density) versus talkative (low information density).

To test our hypothesis we conducted a user study, the results of which demonstrate that people not only enjoy interacting with a digital assistant that mirrors their level of chattiness in its responses, but that interacting in this fashion increases feelings of trust. Furthermore, we demonstrate that it is possible to extract the information necessary to predict when a chatty response would be appropriate.

The rest of this paper is organized as follows: Section 2 discusses background related to mirroring and building trust. Section 3 provides an overview of the experiments. The user study is described in Section 3.1 and the experiments for classifying query style are provided in Section 4. Finally, Section 5 offers a summary and some suggested future work.

2. Background

People are able to engender trust and camaraderie through behavioral mirroring [3–6], where conversational partners mirror one another's interaction style as they negotiate to an agreed upon model of the world [7–9]. Behavioral mirroring is strongly correlated with and predictive of many qualitative interaction measures [3]. It has been shown that the modalities involved in and the degree of mirroring are both predictive of how natural an interaction will be ranked [6]. Understanding and detecting instances of mirroring has become of increasing research interest to the human computer interaction (HCI), machine learning (ML), and developmental robotics fields. Most of the work focuses on detecting and estimating the degree of non-speech-based mirroring and its effects on how people interact with one another [10–16]. For example, the process of mirroring has been specifically leveraged to improve predictions about turn-taking in multi-person interactions. Such systems typically integrate the previous and current actions of all interactants to predict the next actions of a subset of the interactants [17], e.g. to predict turn transitions [18, 19] and next utterance type [19]. Mirroring has also been proposed as a key tool that developmental robotics may leverage to improve language acquisition through interacting with and observing humans [20]. Mirroring has since been used as a learning technique to develop social robots [20–23]. However, to the best of our knowledge, no work has explored mirroring conversational style as a behavior to be produced by a digital assistant.

3. Experiments

We describe: (1) an interactive Wizard-of-Oz (WOZ) user study, and (2) automatic prediction of preferred conversational style using the queries, responses, and participant feedback from the WOZ. In (1), all interactions between the participants and the digital assistant were strictly verbal; there was no visual realization of the digital assistant's responses.

3.1. User Study

The user study evaluated the hypothesis: *interacting with a digital assistant that mirrors a participant's chattiness will cause a positive change in the participant's opinion of the assistant*. In testing this, we also tested whether people who score as chatty according to our measure of interaction style (**Survey 1** in [24]) are more likely to prefer interacting with an assistant who is also chatty, and furthermore whether people who score as non-chatty will prefer non-chatty assistant interactions. Prospective participants completed a questionnaire designed to assess their chattiness level along with other personality traits (e.g. extrovert vs. introvert) after volunteering, but prior to being selected for the study. This allowed us to somewhat balance participants across

http://dx.doi.org/10.21437/Interspeech.2019-1829

types. After selecting the participants, they each completed a pre-study survey (**Survey 2** in [24]) to determine how they use digital assistants (frequency of use, types of query, style of interaction, trustworthiness, likability, etc.).

The study consisted of three conditions: interacting with (1) a chatty, (2) a non-chatty, and (3) a mirroring digital assistant. In all conditions the digital assistant was controlled by a wizard (*i.e.* the experimenter), and the wizard was instructed to not interact directly with the participants during the study.

In the chatty and non-chatty conditions, the participants were prompted (via instructions displayed on a wall-mounted TV display) to make verbal requests of the digital assistant for tasks in each of the following domains: timers/alarms, calendars/reminders, navigation/directions, weather, factual information, and web search. Prompts were text-based and deliberately kept short to limit the tendency to copy the prompts' phrasing when speaking a query. The assistant's responses were generated for each prompt *a priori* and they did not vary between participants. As an example, a prompt might read "next meeting time", for which the chatty response was, "It looks like you have your next meeting after lunch at 2 P.M."; and the corresponding non-chatty response was simply "2 P.M.". After hearing a response, participants were directed to verbally classify its qualities: *good*, *off topic*, *wrong information*, *too impolite*, or *too casual*, which was recorded (via a keyboard) by the wizard. To counter presentation ordering effects, the order of the prompts was randomly assigned such that half of the participants experienced the chatty condition first, whilst the other half non-chatty. After completing both the chatty and non-chatty conditions, participants answered questions (**Survey 3** in [24]) about their preference for the chatty vs. non-chatty interactions. Participants specified their preferences both within and across the task domains.

All participants interacted with the mirroring assistant after completing both the chatty and non-chatty conditions. The mirroring assistant interactions were designed to be as natural as possible within the confines of a WOZ user study. The wizard for this condition was the same as was used in the chatty and non-chatty conditions, and again, the wizard was instructed to not interact with the participants during the study. Note that in the first two conditions the wizard was not required to rate the chattiness of a query since the type of response is defined *a priori* depending on the condition, and so the responses are not dependent on the conversational style of the user. However, in this condition the role of the wizard was to assign a chattiness score (1, 2, 3, 4, or 5) for each utterance produced by a participant, which was then used to select the chattiness level of the assistant's response.

3.1.1. Eliciting Natural Queries for Mirrored Responses

To guide the formation of natural inputs, participants were asked to imagine an "evening out" scenario, which involved meeting friends, watching a show, and planning dinner. The wizard walked subjects through the scenario, prompting them (via the TV display) to make requests using a set of image-icon pairs, see Figure 1; no word-based prompts were used. The hand-drawn images depicted the imagined stage in the evening, and the icons indicated which of the digital assistant's task-oriented functionalities the participant should take advantage of. For example, early in the scenario walkthrough, an image of a closet of clothes was paired with a weather app icon, see Figure 1(b). The set of possible responses associated with each prompt was fixed across participants, and the chattiness level of the selected response always matched chattiness level assigned to the query by the wizard. The responses associated with Figure 1(b), sorted least to most chatty, were:

1. "74 and clear."
2. "It will be 74 degrees and clear."
3. "It will be a comfortable 74 degrees with sunny skies."
4. "It's supposed to be 74 degrees and clear, so don't bother bringing a sweater or jacket."
5. "Well, my sources are telling me that it's supposed to be 74 degrees and clear. You probably don't need to bother bringing a sweater or jacket."

As in the chatty and non-chatty conditions, participants rated each of the assistant's responses. After completing all interactions in the mirroring condition, participants completed the post-study survey, which was used to measure changes in opinion about the likability and trustworthiness of the digital assistant.

3.1.2. Setup

During the user study we recorded speech and video for each participant. In particular, speech was recorded at 16 kHz mono, 16-bits per sample using a Røde Lavalier lapel microphone and a Focusrite 2i4 USB audio interface. Video (not used here) was recorded in both grayscale and near-infrared at resolution 1088 $\times$ 1088 at 100 frames per second, 8-bits per pixel using Ximea MQ022MG-CM and MQ022RG-CM cameras respectively. In addition we also captured depth information (also not used here) using a PrimeSense RD1.09 sensor.

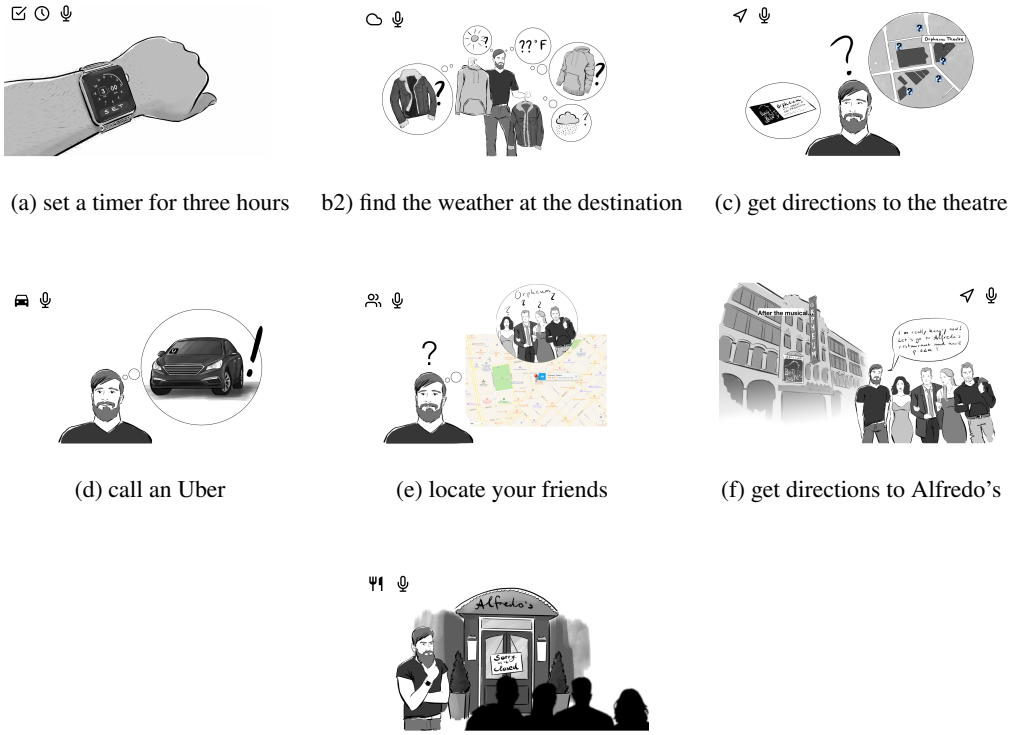
Subjects were asked to sit and face a wall-mounted TV at a distance of approximately 2.5m. This was used to display all text and imagery to the participants, and the cameras were mounted beneath this display to obtain a good view of the participants. The wizard sat behind a dividing screen and used a MacPro to control the digital assistant and drive the display; the same MacPro was used to synchronize all hardware and capture the data using ROS [25].

Table 1: *Significance of the difference in preferences for chatty vs. non-chatty responses per domain. There is no significant difference for timers and factual information.*

Task Domain	F-score	p-value
navigation/direction	6.24	0.02
factual information	2.86	0.12
timers/alarms/calendar	0.08	0.79
weather	5.12	0.04
web search	7.73	0.02

3.1.3. Results and Discussion

In total twenty people (three women and seventeen men) participated in the study, with session durations ranging from 17 to 56 minutes, depending on participants' verbosity. The majority of participants (70%) preferred interacting with the chatty digital assistant. According to participant responses to the personality and interaction style survey (**Survey 1**), 60% of participants were generally chatty and 40% were non-chatty. As more people preferred the chatty response than were identified as chatty, one's own self-reported conversational style does not necessarily predict one's preference for assistant chattiness. However,



(a) set a timer for three hours b2) find the weather at the destination (c) get directions to the theatre

(d) call an Uber (e) locate your friends (f) get directions to Alfredo's

(g) get recommendations for an alternative restaurant

Figure 1: *The prompts for the seven queries used in our mirroring user study.*

in general, the participants identified as chatty preferred the chatty interactions, and those identified as non-chatty preferred the non-chatty interactions.

The effect of mirroring on opinions of likability and trustworthiness was tested using a one-way ANOVA. We compared the participants' trustworthiness ratings of the assistant from the pre-study (mean=4.0, stdev=0.48) and post-study (mean=4.46, stdev=0.31) surveys. Users were asked to rate how much they agreed with the statements about liking vs. not liking interacting with the assistant on a seven-point Likert scale, where one corresponds to strong disagreement, four is neutral, and seven corresponds to strong agreement (see **Survey 1**). The difference in the mean score pre-mirroring and post-mirroring conditions was statistically significant (f-score $= 7.12$, $p \leq 0.01$), meaning that interacting with a style-mirroring assistant had a significant, positive impact on opinions and trustworthiness. Additionally, the task domain had a smaller, but still significant impact on whether subjects preferred chatty or non-chatty responses (f-score $= 2.67$, $p \leq 0.02$). For a specific breakdown by task domain, see Table 1.

Anecdotal evidence from comments in the post-study debrief suggest that participants prefer the assistant in the mirroring conditions. In summary, we conclude that chattiness preferences differ across individuals and across task domains, but mirroring user chattiness increases feelings of likability and trustworthiness in digital assistants. Given the positive impact of mirroring chattiness on interaction, we proceeded to build classifiers to determine whether features extracted from user speech could be used to estimate their level of chattiness, and thus the appropriate chattiness level of a response.

4. Detecting Preferred Interaction Style

We built multi-speaker and speaker-independent classifiers to detect from a query utterance: (1) if the utterance is chatty or non-chatty, and (2) if a chatty vs. non-chatty response would be preferred. The chatty or not classification was based solely on the audio features and did not include a measure of utterance duration nor word count as both require extra semantic and pragmatic information to be useful in detecting chattiness. The chatty vs. non-chatty target label for each utterance was extracted from the survey responses, overall participant chatty vs. non-chatty labels were extracted from **Survey 1**, and response preference labels (*good*, *too casual*, etc.) were extracted from the digital assistant response evaluations obtained in the user-study. Only the utterances from the chatty and non-chatty conditions were included in the data set. Each utterance was assigned two labels, where one indicated whether the speaker was chatty, and the other indicated the response preference for the utterance. From the speech, 95 acoustic features were extracted: the mean, standard deviation, maximum, minimum of the fundamental frequency, energy, the first twelve MFCC's, and the first five formants [26].

Ten classifiers were trained on the two binary classification tasks: logistic regression, naive Bayes, artificial neural network (one hidden layer with 64 units), random forest using Gini, random forest using entropy, SVM, SVM with an RBF, polynomial, linear, and sigmoid kernel. These classifiers were selected for their simplicity due to the small number of data points, and we used the standard implementations in `scikit-learn` [27]. The classifiers were evaluated with

Table 2: *F_1-scores for the utterance chattiness and response chattiness preference binary classification tasks (chatty vs. non-chatty) for both the Leave One Participant Out (top) and the 80/20 evaluation splits (bottom).*

	Log. Reg.	*Naive Bayes*	*ANN*	*RF (Gini)*	*RF (Entropy)*	*SVM*	*SVM (RBF)*	*SVM (Linear)*	*SVM (Poly.)*	*SVM (Sig.)*
Leave One Participant Out										
Detector	0.60	0.71	0.82	0.83	0.83	0.83	0.82	0.84	0.82	0.84
Selector	0.65	0.71	0.79	0.79	0.81	0.83	0.84	0.81	0.83	0.84
80/20 Split										
Detector	0.62	0.74	0.85	0.85	0.85	0.86	0.86	0.86	0.85	0.86
Selector	0.68	0.78	0.82	0.81	0.85	0.85	0.87	0.84	0.86	0.86

both an 80/20 split of training/test data, so we train on samples for all speakers and test on different data from the same speakers (multi-speaker), and a leave-one-participant-out train/test split (speaker-independent). Performance was evaluated according to the F_1-score due to the label imbalance.

4.0.1. Results and Discussion

The classification results are shown in Table 2, which demonstrate that the classifiers performed well for both forms of evaluation split. This is a promising indicator that both a speaker's degree of chattiness and their preference for chatty vs. non-chatty response can be detected reliably. The majority of classifiers had performance comparable to one another and better than chance, with the SVM methods performing best on both types of evaluation split and both classification tasks. These results are encouraging, especially given the small size of the data set. Performance on the 80/20 split indicates that the algorithms do not need many samples from a user to learn. Performance on the leave-one-participant-out split suggests that the models are able to generalize to new speakers. However, access to samples across participants does improve performance for all the classifiers tested.

5. Conclusion and Future Directions

We have shown that user opinion of the likability and trustworthiness of a digital assistant improves when the assistant mirrors the degree of chattiness of the user, and that the information necessary to accomplish this mirroring can be extracted from user speech. Future work include detecting ranges of chattiness rather than the binary labels used here, expanding the participant pool, and using multimodal signals from the videos and depth images to measure the degree to which users appreciate the assistant responses.

6. References

[1] E. Delaherche, M. Chetouani, A. Mahdhaoui, C. Saint-Georges, S. Viaux, and D. Cohen, "Interpersonal synchrony: A survey of evaluation methods across disciplines," *IEEE Transactions on Affective Computing*, vol. 3, no. 3, pp. 349–365, 2012.

[2] R. I. Swaab, W. W. Maddux, and M. Sinaceur, "Early words that work: When and how virtual linguistic mimicry facilitates negotiation outcomes," *Journal of Experimental Social Psychology*, vol. 47, no. 3, pp. 616–621, 2011.

[3] T. Chartrand and J. Bargh, "The chameleon effect: the perception–behavior link and social interaction." *Journal of personality and social psychology*, vol. 76, no. 6, p. 893, 1999.

[4] M. Pickering and S. Garrod, "Toward a mechanistic psychology of dialogue," *Behavioral and brain sciences*, vol. 27, no. 2, pp. 169–190, 2004.

[5] D. Goleman, *Emotional intelligence*. Bantam, 2006.

[6] A. Nenkova, A. Gravano, and J. Hirschberg, "High frequency word entrainment in spoken dialogue," in *Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies: Short papers*. Association for Computational Linguistics, 2008, pp. 169–172.

[7] K. Niederhoffer and J. Pennebaker, "Linguistic style matching in social interaction," *Journal of Language and Social Psychology*, vol. 21, no. 4, pp. 337–360, 2002.

[8] J. Pennebaker, M. Mehl, and K. Niederhoffer, "Psychological aspects of natural language use: Our words, our selves," *Annual review of psychology*, vol. 54, no. 1, pp. 547–577, 2003.

[9] P. Taylor and S. Thomas, "Linguistic style matching and negotiation outcome," *Negotiation and Conflict Management Research*, vol. 1, no. 3, pp. 263–281, 2008.

[10] E. Delaherche and M. Chetouani, "Multimodal coordination: exploring relevant features and measures," in *Proceedings of the 2nd international workshop on Social signal processing*, 2010, pp. 47–52.

[11] D. Messinger, P. Ruvolo, N. Ekas, and A. Fogel, "Applying machine learning to infant interaction: The development is in the details," *Neural Networks*, vol. 23, no. 8-9, pp. 1004–1016, 2010.

[12] S. Michelet, K. Karp, E. Delaherche, C. Achard, and M. Chetouani, "Automatic imitation assessment in interaction," in *International Workshop on Human Behavior Understanding*, 2012, pp. 161–173.

[13] F. Ramseyer and W. Tschacher, "Nonverbal synchrony in psychotherapy: coordinated body movement reflects relationship quality and outcome." *Journal of consulting and clinical psychology*, vol. 79, no. 3, p. 284, 2011.

[14] X. Sun, A. Nijholt, K. Truong, and M. Pantic, "Automatic visual mimicry expression analysis in interpersonal interaction," in *CVPR 2011 WORKSHOPS*. IEEE, 2011, pp. 40–46.

[15] X. Sun, K. Truong, M. Pantic, and A. Nijholt, "Towards visual and vocal mimicry recognition in human-human interactions," in *2011 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 2011, pp. 367–373.

[16] J. Terven, B. Raducanu, M. Meza-de Luna, and J. Salas, "Head-gestures mirroring detection in dyadic social interactions with computer vision-based wearable devices," *Neurocomputing*, vol. 175, pp. 866–876, 2016.

[17] D. Ozkan, K. Sagae, and L. Morency, "Latent mixture of discriminative experts for multimodal prediction modeling," in *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 860–868.

[18] L. Huang, L. Morency, and J. Gratch, "A multimodal end-of-turn prediction model: learning from parasocial consensus sampling," in *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 3*. International Foundation for Autonomous Agents and Multiagent Systems, 2011, pp. 1289–1290.

[19] D. Neiberg and J. Gustafson, "Predicting speaker changes and listener responses with and without eye-contact," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[20] A. Cangelosi, G. Metta, G. Sagerer, S. Nolfi, C. Nehaniv, K. Fischer, J. Tani, T. Belpaeme, G. Sandini, F. Nori, and L. Fadiga, "Integration of action and language knowledge: A roadmap for developmental robotics," *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 3, pp. 167–195, 2010.

[21] A. Galdeano, A. Gonnot, C. Cottet, S. Hassas, M. Lefort, and A. Cordier, "Developmental learning for social robots in real-world interactions," in *First Workshop on Social Robots in the Wild at the 13th Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI 2018)*, 2018.

[22] R. Nakajo, S. Murata, H. Arie, and T. Ogata, "Acquisition of viewpoint transformation and action mappings via sequence to sequence imitative learning by deep neural networks," *Frontiers in neurorobotics*, vol. 12, 2018.

[23] J. B. and G. G., "Deep curiosity loops in social environments," *CoRR*, vol. abs/1806.03645, 2018. [Online]. Available: http://arxiv.org/abs/1806.03645

[24] K. Metcalf, B.-J. Theobald, G. Weinberg, R. Lee, I.-M. Jonsson, R. Webb, and N. Apostoloff, "Mirroring to build trust in digital assistants," *arXiv preprint arXiv:1904.01664*, 2019.

[25] M. Q., K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Ng, "ROS: an open-source robot operating system," in *ICRA Workshop on Open Source Software*, 2009.

[26] D. Jurafsky, R. Ranganath, and D. McFarland, "Extracting social meaning: Identifying interactional style in spoken conversation," in *Proceedings of Human Language Technologies*. Association for Computational Linguistics, 2009, pp. 638–646.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=1953048.2078195