**Figure 1: Hands-free system, microphone array enclosed within yellow box**


**Figure 2: Camera setup to enable remote Wizard of Oz experiments**

# Empathetic conversations in the car

**Hiroshi Mendoza**
Stanford University
Stanford, USA
hmendoza@stanford.edu

**Julia Alison**
Stanford University
Stanford, USA
jalison@stanford.edu

**Shannon Wu**
Stanford University
Stanford, USA
shannwu@stanford.edu

**James Landay**
Stanford University
Stanford, CA, USA
landay@stanford.edu

**Pablo E. Paredes**
Stanford University
Stanford, CA, USA
paredes@cs.stanford.edu

## ABSTRACT

UPDATED—February 13, 2019. Mental health is a world-wide problem impacting 400 million people and costing 1 trillion annually [1], but currently, providers of mental health therapies and services are not adequately meeting the needs of the patient population. In this work, we discuss user interactions with voice chat agents and limitations of current system designs in facilitating a coherent user experience with these agents. This is especially crucial, if voice agents need to convey the conversational flow required in applications such therapeutic chat bots. Additionally, the use of back-channels, or filler phrases or sounds such as uh-huh or ah, could provide users a more intuitive experience with therapeutic chat bots. Through a series of user experience tests, we present a preliminary data set of human-intuited back-channel insertions and turn-taking detection for future use in developing user interactions with conversational agents.

*CHI'19, May 2019, Glasglow, UK*

**ACM Reference Format:**
Hiroshi Mendoza, Julia Alison, Shannon Wu, James Landay, and Pablo E. Paredes. 1997. Empathetic conversations in the car. In *Proceedings of ACM CHI Conference (CHI'19).* ACM, New York, NY, USA, 6 pages. https://doi.org/https://doi.org/10.1145/3290607

## INTRODUCTION

Although there exist multiple barriers to accessing mental healthcare, the stigma of seeking such help remains a significant obstacle to many patients (Eisenberg 2010) [6]. As such, this unmet need calls for new models of care delivery. As demonstrated by the Woebot, autonomous chat agents or chatbots may offer a novel, promising means of accessing mental health therapy (Fitzpatrick 2017) [3]. Although the body of research on therapeutic chat agents is growing, there still exist many unanswered questions on both the implementation and user experience of these chat agents. Existing chatbots are either limited to interacting via typing or leverage primitive natural speech processing that is error-prone and disrupts the user experience. For example, there exists little work in refining turn-taking, that is, detecting when a human is finished speaking in voice chat agents.

We explore therapeutic micro-chat agents, operating in automobiles during commutes, which can turn an otherwise unused period of time in a users day to build healthy coping mechanisms for stress. As the technologies for both chat agents and self-driving automobiles advance, the commute offers a new opportunity to intervene and provide mental health therapy without disturbing the userfis everyday routine. We believe that the commute may in the future serve as a period for de-stressing or preparation for the day. Furthermore, as self-driving car technology improves, the commute will prove to be an invaluable time to engage in exercises developing resilience and healthy behaviors in a convenient manner. In this paper, we present a wizard of oz user interaction study (n=12) in which participants interacted in conversations with a voice chat agent in order to explore back-channel and turn-taking methods for use in future designs of voice conversational agents. Images of the experimental setup can be seen in Figures 1 and 2.

## RELATED WORK

Several iterations of chatbots designed for mental health therapy have existed over the years. ELIZA was one of the first therapeutic chatbots, implemented in the 1960s to deliver non-directive therapy drawn from the Rogerian tradition (Weizenbaum 1966) [8]. After conducting a broad survey of studies on use of chatbots for mental health intervention, Hoermann et al. (2017) [5] concluded that engagement with these chatbots led to significant and sustained improvements in mental health outcomes relative to control groups. Outcomes were comparable, but not superior to, those following more traditional forms of treatment like in-person therapy.

Studies like Fitzpatrick et al. (2017) [3] specifically highlight that such interventions are most effective when participants feel comfortable sharing their feelings and motivations with the conversational agent, and they argue that such comfort is facilitated by natural and human-like conversational patterns. This led us to focus on the natural patterns of human speech, which we would seek to mimic with our own chat agent. In most normal conversations, humans naturally process a variety of cues to assess when their conversational partner is about to begin, or has completed, expressing a thought. The ebb and flow is important in making conversation appear natural, and, as informed by the above section, may therefore be critical in determining the effectiveness of the intervention. A crucial feature in achieving a natural conversational style with the user is the incorporation of appropriate turn-taking within the bots' speaking patterns.

Our project is not the first to deal with turn-taking cues in verbal interactions with chatbots. Hjalmarsson (2011) [4] found that increasing the number of speaker cues that had been labeled to indicate turn-yielding or turn-holding increased participants' reaction times in labeling the speakerfis subsequent behavior. Thus, including these cues can make listeners more comfortable evaluating how the conversation is likely to proceed. While this finding is important in that it suggests that human listeners pay attention to syntactic and prosodic cues in labeling conversational actions, our research mostly concerns the ability of bots to incorporate this information.

One example of these techniques employed in a working model is Erica, the bot built and presented by Lala et al. (2017) [7]. Erica utilizes a flexible approach to turn-taking; upon a pause in the conversation, the model produces the likelihood score that Erica should take a turn and speak. The likelihood score then produces one of four responses: silence, generating a back-channel, generating a filler, or beginning to speak. The first two options presumably incite the user to continue speaking, while the latter two indicate the bot's willingness to begin speaking. Additionally, Edlund and Heldner's 2005 study [2] attempted to use online prosodic data to artificially generate a human-like conversation. They sought to eliminate awkward silences caused by extremely long silence thresholds for bots to speak by identifying appropriate and inappropriate turn-taking opportunities based on prosodic information. Results indicated that in many cases, pauses within the dialogue were in between speakers holding control of the conversation rather than ceasing to speak.

These results suggest that the direction of identifying prosodic and lexical cues for chatbot turn-taking prediction is a rich area for future research, as the models are currently ill-defined and incomplete. Thus, our research seeks to contribute in this area by collecting data to be labeled for prosodic and lexical features and by testing user responses to a technique for promoting natural turn-taking that is widely discussed in the literature, back-channeling.

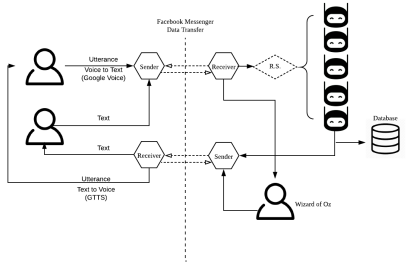


**Figure 3: System Diagram**

Table 1: Experimental Design

| | | IV1: Turn-Taking Detection Method | |
|---|---|---|---|
| | | Level 1: Push-to-Talk | Level 2: Hands-Free |
| IV2: Back-channeling | Level 1: Back-channeling present | Push-to-talk with back-channeling | Hands-free speech with back-channeling |
| | Level 2: No back-channeling present | Push-to-talk without back-channeling | Hands-free speech without back-channeling |

**Figure 4: 2x2 Experimental Design**

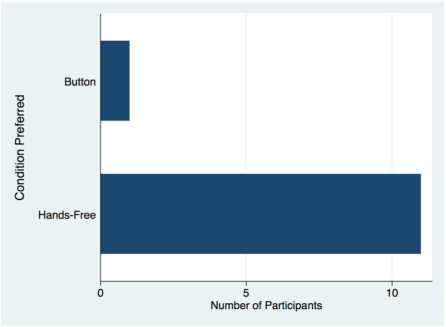


**Figure 5: Hands-free vs Button preference**

## SYSTEM DESIGN

*Positive thinking Chatbot.* The chatbot used in the experiment was from the PopBots suite of micro-chat agents, developed from an earlier design iteration to create a series of quick, therapeutic micro-conversational chatbots. The Positive Thinking chatbot, which prompts the user to attempt to find a positive aspect of a recent, stressful event, was selected for the study for its relatively predictable conversation flow.

*System Architecture.* For this particular study, the user interacted with the chatbot via Wizard of Oz techniques; a researcher on the team delivered pre-recorded audio clips of the chatbot script to the car when the user ended their turn to talk. The chatbot begins the conversation with a greeting and a prompt for the user to describe a stressful experience in the recent past. Based on the user's answers, the researcher would play certain scripts pre-written according to a standardized flow. The content of the Positive Thinking Chatbot was developed according to standard clinical procedure. Figure 3, shows the system diagram of the software architecture. The system contains a Linux computer in the car that is connected to the web via a LTE connection. A 7 array microphone array was used to capture the user's speech. The system allowed for a fully automated approach where the chat bot output would be converted to speech and the input would be converted to text with as shown in the figure. The automated approach was not used since we wanted to more accurately control the flow of the conversations and insert back-channel responses.

*Push to Talk versus Hands Free Speech.* We presented the users with two implementations of the same chatbot conversation. The push-to-talk implementation required the participant to push a handheld button to ″tell″ the chatbot when a human was speaking, taking their turn, and release the button when the chatbot was free to respond. The button was wired to a light attached the car's dashboard such that pressing the button would turn the light on, allowing the researcher to view, through the GoPro feed, when the participant was ″taking their turn″ in the conversation. Once the participant released the button-indicating that they had finished talking, the light would turn off, signifying to the researcher that the chatbot should take their turn and respond.

The hands-free speech implementation allowed the participant to converse with the chatbot as though the chatbot were a human. Using the audio feed, the researcher would intuit when the participant finished their turn and appropriately pipe the chatbotfis response to the carfis input.

Participants were asked to engage in conversations with the chat agent while driving around campus while the research team monitored the conversation and indicated when the chatbot should respond to the participant's comment. 12 participants were recruited and randomly assigned to conditions which automatically determined the order of hands-free and button tests, and also whether each test would include back channels as seen in Figure 4.

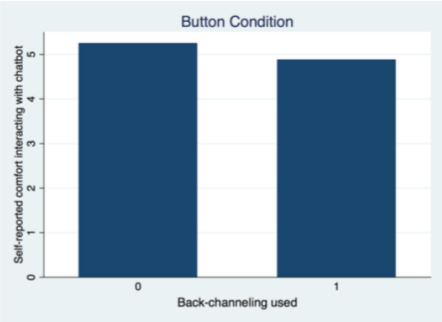


**Figure 6: Button Condition**

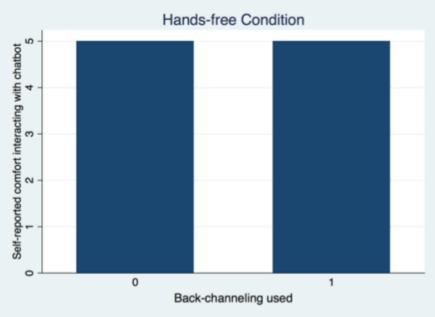


**Figure 7: Hands-free Condition**

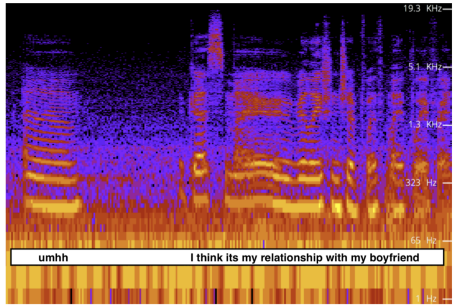


**Figure 8: Log normalized speech frequency at various points in the conversation**

## RESULTS

11 of 12 individuals preferred using hands-free speech to interact with the chatbot (Figure 5). The hands-free speech option ″felt more natural,″ ″felt safer [while] driving,″ and was ″less distracting″ for the driver. We attribute this sentiment to the unwieldy implementation of the push-to-talk button and acknowledge that participants may feel uncomfortable and stressed while having to handle a cumbersome button while attempting to safely drive.

However, in both the hands-free and button conditions, back-channeling does not appear to have meaningfully changed participants′ reported level of comfort interacting with the chatbot (Figure 6 and 7). Back-channeling as a speech pattern is not meant to be particularly noticeable in conversation, so these results at least indicate that the back-channeling, as implemented, was not especially distracting.

The data set collected through this series of experiments is insufficient to perform substantial quantitative analysis. Though qualitative reports indicate that users are comfortable and open to interacting with chatbots, many more conversations would need to be collected before one may use the collection as a training data set. However, this series of experiments provides a model for future researchers to collect the necessary quantity of information, and it allows researchers to initiate the analysis discussed below.

*Prosody.* Information on prosody of user's responses can provide crucial features for identifying turn-taking in conversation. Once clear points at which the participant gives up or takes back control of the conversation are labeled, researchers can use prosody analysis techniques to generate features for the user's speaking pitch and frequency at those points. Once the model is developed, AI-based chatbots will be able to detect similar patterns in ongoing conversations and to therefore predict turn-taking. At that point, the bot may be able to mimic and predict human turn-taking patterns, independent of researcher's prompts.

As shown in Figure 8, we include a series of figures on the log scale of a participant's speech frequency over time. Calculations and visualizations such as these will contribute to the prosody data set which can be used by future iterations of the chatbot.

## DISCUSSION AND FUTURE WORK

*Exploring the Mental Model of Chatbots.* Once a data set of adequate size is collected, future work could compare the turn-taking behaviors of human-to-human conversations with the turn-taking behaviors of human-to-chatbot conversations. Of particular interest is a deeper examination of turn-taking behavior of participants who felt comfortable engaging with the chatbot compared to the turn-taking behavior of participants who felt uncomfortable. As mentioned, some participants chose to engage the chatbot as though it were human while others did not; future work analyzing the turn-taking behavior between these two cohorts could unveil useful insights into how best to model a natural conversation

with an autonomous chat agent, even with individuals who did not conceptualize those chat agents as human conversational partners. Our assumption that a more human-like chatbot would improve user experience appears to be consistent with our findings, but further experimentation would be required to validate such insights.

Therapeutic chat agents have historically proven to be a efficacious method of providing quick and convenient access to mental health therapy when an individual is unable to seek professional help. Our preliminary experimental results demonstrate that though chatbots for mental health are still a nascent idea in the medical and technological fields, users are open to the idea of sharing such personal information with chat agents, indicating a promising future for research in this space. At present, the bulk of the work lies within improving the speech processing capabilities of conversational agents, as the user experience depends heavily on the bot's functionality. There currently exists relatively little research on the implementation of voice-based therapeutic chatbots, so while there still exist many functional challenges before researchers may fully devote their attention to the content and user experience of these chat agents, the future of these mental health providers proves to be promising.

## ACKNOWLEDGMENTS

We thank all the volunteers, publications support, staff, and authors who wrote and provided helpful comments on previous versions of this document. As well the staff at VAIL for supporting this research.

## REFERENCES

[1] D. Chisholm. 2016. Scaling-up treatment of depression and anxiety: a global return on investment analysis. . *The Lancet Psychiatry* 3, 5 (2016), 415–424. https://doi.org/10.1016/S2215-0366(16)30024-4

[2] Heldner M. Edlund, J. 2005. Exploring Prosody in Interaction Control. *Phonetica* 62, 2 (2005). https://doi.org/10.1159/000090099

[3] Fitzpatrick. 2017. Delivering Cognitive Be- havior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. . *JMIR Mental Health* 4, 2 (2017), e19. https://doi.org/10.2196/mental.7785

[4] A. Hjalmarsson. 2011. The additive effect of turn-taking cues in human and synthetic voice. *Speech Communication* 53, 1 (2011), 23–35. https://doi.org/10.1016/j.specom.2010.08.003

[5] Milne DN Calvo RA Hoermann S, McCabe KL. 2017. Application of Synchronous Text-Based Dialogue Systems in Mental Health Interventions: Systematic Review. *J Med Internet Res* 19, 8 (2017), e267. https://doi.org/10.2196/jmir.7023

[6] Eisenberg D. Hunt J. 2010. Mental health problems and help-seeking behavior among college students. *J Adolesc Health* 46, 1 (2010), 3–10. https://doi.org/10.1016/j.jadohealth.2009.08.008

[7] Milhorat P. Inoue K. Ishida M. Takanashi K. Kawahara T. Lala, D. 2017. Attentive listening system with backchanneling, response generation and flexible turn-taking. *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue* (2017), 127–136. https://doi.org/10.1177/089443939201000402

[8] J. Weizenbaum. 1966. ELIZAâĂŤa computer program for the study of natural language communication between man and machine. . *Commun. ACM* 9, 1 (1966), 36–45. https://doi.org/10.1145/365153.365168