SMM19, Workshop on Speech, Music and Mind 2019
14 September 2019, Vienna, Austria

A protocol for collecting speech data with varying degrees of trust

Lara Gauder[1,2], *Agustín Gravano*[1,2], *Luciana Ferrer*[2], *Pablo Riera*[2], *Silvina Brussino*[3,4]

[1]Departamento de Computación, FCEyN, Universidad de Buenos Aires (UBA), Argentina
[2]Instituto de Investigación en Ciencias de la Computación (ICC), CONICET-UBA, Argentina
[3]Facultad de Psicología, Universidad Nacional de Córdoba (UNC), Argentina
[4]Instituto de Investigaciones Psicológicas, CONICET-UNC, Argentina
{mgauder,gravano,lferrer,priera}@dc.uba.ar, silvina.brussino@unc.edu.ar

Abstract

This paper describes a novel experimental setup for collecting speech data from subjects induced to have different degrees of trust in the skills of a conversational agent. The protocol consists of an interactive session where the subject is asked to respond to a series of factual questions with the help of a virtual assistant. In order to induce subjects to either trust or distrust the agent's skills, they are first informed that the agent was previously rated by other users as being either good or bad; subsequently, the agent answers the subjects' questions consistently to its alleged abilities. These interactions will be speech-based, with subjects and agents communicating verbally, which will allow for the recording of speech produced under different trust conditions. Ultimately, the resulting dataset will be used to study the feasibility of automatically predicting the degree of trust from speech. This paper describes a preliminary experiment using a text-only version of the protocol in Argentine Spanish. The results show that the protocol effectively succeeds in influencing subjects into the desired mental state of either trusting or distrusting the agent's skills. We are currently beginning the collection of the speech dataset, which will be made publicly available once ready.

Index Terms: Trust/distrust; Speech corpus; Mental state; Spoken dialogue system; Automatic detection.

1. Introduction

An increasingly important aspect of a conversational agent is its ability to dynamically monitor the user's mental state, including their engagement, satisfaction, and emotions in general [1, 2, 3]. Among these, and especially for virtual assistants, tracking the user's degree of *trust* in the system's skills may be critical for the success of the interaction. Hypothetically, if at some point the user starts displaying cues of distrust and the system can effectively detect such cues, then the dialogue manager could choose to act in consequence for regaining the user's trust. The main goal of this research project is the automatic detection of trust from speech, and in this paper, we focus on the collection of a speech trust dataset.

Trust has been a topic of study for decades among researchers from several disciplines like psychology, sociology, anthropology, economics and political science. One important area of research has been the search for the factors that explain trust. Mayer et al. consider trust to depend on the trustor's perception of the trustee's *ability, benevolence and integrity* [4]. Other factors that might explain trust have been proposed, including the propensity to trust [5], contextual and situational factors [6, 7], among others. Further, trust is dynamic, and it can be created or destroyed during a conversation [8, 9].

The nature of trust has been described both as rational or cognitive [10, 11] and as emotional or affective [12], or a combination of both [13]. In either case, we hypothesize that the degree of trust affects and is affected by linguistic aspects (including the form and content of discourse) and paralinguistic aspects (including the intonation, pitch, speech rate and voice quality) of the trustor and trustee speech. The main goal of our research project is to study to what extent the trustor's degree of trust can be predicted from their speech characteristics using fully automatic methods.

Little research has been done with the goal of measuring or predicting trust directly from speech signals. Elkins and Derrick studied pitch variations in the trustor's speech in the context of human-computer interactions in the form of interviews [14]. They found that the subject's pitch was related to their degree of trust and that the relationship was time-dependent. In another work, Waber et al. studied paralinguistic aspects in medical conversations between nurses [15]. They found that the emphasis used by an outgoing nurse when talking to an incoming nurse was significantly related to the degree of trust that the outgoing nurse reported having on their colleagues during an initial interview. The results in these works support our hypothesis that valuable information can be extracted from speech to predict the degree of trust.

To our knowledge, no speech corpus is available with annotations of varying degrees of trust, large enough to allow for statistical analyses and machine learning experiments. Our initial approach for building a trust speech dataset consisted in annotating an existing dialogue corpus. We conducted an informal pilot study in which three of the authors labeled 40 inter-pausal units[1] extracted randomly from the Argentine Spanish Games Corpus, a collection of task-oriented dialogues in which pairs of subjects played a series of computer games designed to require cooperation and communication [16]. The three raters were asked to indicate their agreement with the statement, *"The current speaker trusts his/her interlocutor's skills to perform the task"*, using a 7-level Likert scale. The resulting average pairwise inter-rater correlation coefficient was lower than 0.1, which was consistent with the raters' unanimous opinion that the task was too hard, if not impossible, for an external listener. We repeated this procedure with full conversations rather than isolated IPUs (to allow raters to become more familiar with the speakers' speech), and also with IPUs extracted from the Switchboard corpus [17] (to consider different domains), but the inter-rater agreements did not improve. A plausible conclusion from these pilot tests is that external annotations of the

[1]In dialogue, an inter-pausal unit (IPU) is a speech segment from a single speaker, surrounded by silence and with no overlapping speech from the interlocutor.

10.21437/SMM.2019-2

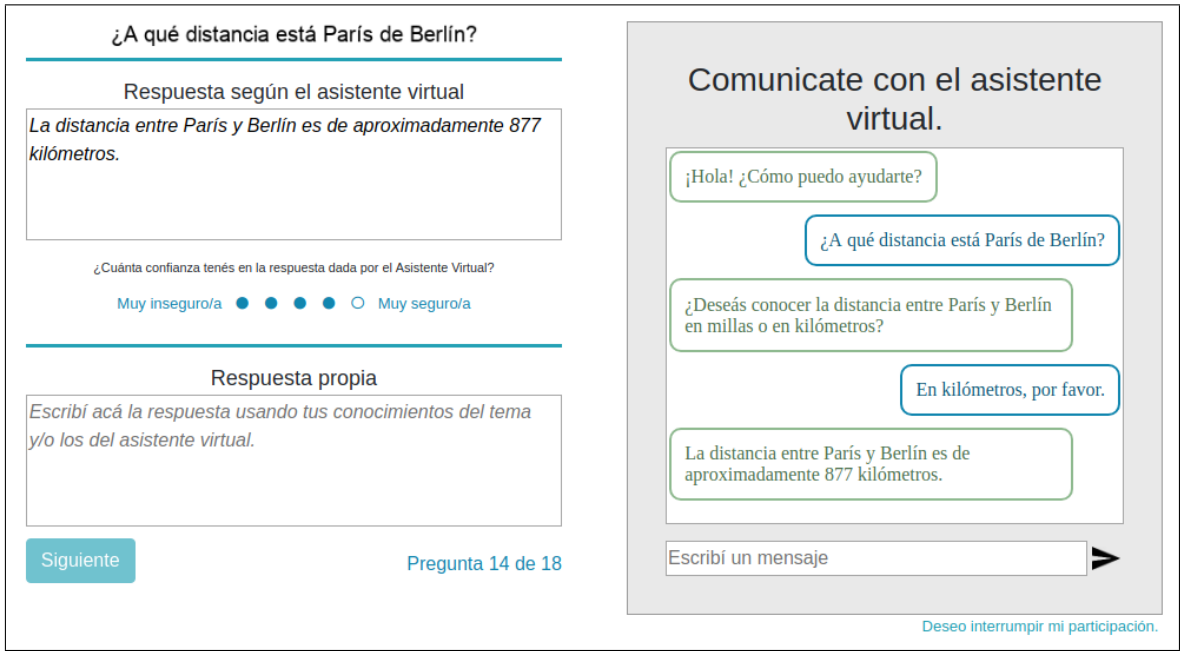

Figure 1: *Screenshot of the user interface for "What is the distance between Paris and Berlin?". Subjects enter in the form on the left, from top to bottom: the virtual assistant's response, their confidence in the assistant's response, and their own response. They interact with the assistant using the chat window on the right.*

degree of trust are not reliable. It is also possible that these corpora do not contain substantial variation in the degree of trust experienced by the speakers, because of the nature of the tasks and topics they talk about. In either case, we must follow an alternative approach for creating a trust speech corpus.

Consequently, we designed and implemented a novel experimental setup for collecting speech from subjects that are *induced* to have different degrees of trust in the skills of a conversational agent. This paper describes the protocol in detail (Section 2), as well as a preliminary experiment, conducted using a text-only version of this setup in Argentine Spanish (Section 3). The results show that our protocol effectively succeeds in influencing subjects into the desired mental state of either trusting or distrusting the agent's skills. In the next stage, we will proceed to put into production a speech-based version of the protocol for recording speech with varying degrees of trust.

2. Experimental Setup

The protocol consists of an interactive session where the subject must respond to a series of questions with the help of a virtual assistant (VA). This section describes the protocol in detail.

2.1. Session structure and initial bias

The subjects' task is to respond a series of factual questions with the help of a VA. For each question, subjects are required to **(1)** interact with the VA via a text chat window to find the answer to the question; **(2)** transcribe the answer given by the VA; **(3)** rate their confidence in the response given by the VA using a 5-level Likert scale; and **(4)** enter their own answer, based on what they believe to be correct (this may or may not match the VA's response). Figure 1 shows a screenshot of the user interface: the current factual question is shown at the top left of the screen; below that is a form in which subjects must enter the VA's response, their confidence in the VA's response, and their own response. On the right lies the chat window used to interact with the VA.

At the beginning of the session, subjects are told that the VA they will interact with was previously rated by other users with either a very high or very low score: 4.9 and 1.4 out of 5 stars, respectively (these two values were chosen empirically based on pilots tests). These two conditions are central in our setup and are meant to bias the user toward either trusting or distrusting the VA's skills. We refer to them as the *high-score* and the *low-score* conditions. With this setup we intend to benefit from a well-studied cognitive bias called *anchoring* or *previous-opinion* bias, in which a person's decision-making process is influenced by an initial piece of information offered to them, such as a house valuation made by another broker, or a patient diagnoses made by another doctor [18, 19].

Subsequently, the quality of the responses given by the VA is consistent with the informed abilities, making no mistakes in the high-score condition, and making some mistakes in the low-score condition. This is intended to reinforce in the subject the feeling that the former system is good, and the latter is bad.

2.2. Types of factual questions

Each session contains 18 factual questions, 6 of which we classify as *easy* and 12 as *difficult*. Easy questions are about topics that should be obviously known by anyone (e.g., *"How many minutes are in an hour?"*) and are used to generate the feeling in the subject that the VA actually works. Difficult questions, on the other hand, were selected so that their correct answers would likely be unknown to most people (e.g., *"What are the three brightest stars in the sky?"*). Thus, for difficult questions subjects should depend on the VA's responses. Furthermore, from the subjects' perspective, difficult questions make the task more challenging and interesting; but from our part, these ques-

tions will allow us to assess the subjects' varying degree of trust in the VA's skills (this is explained in detail in Section 3).

Difficult questions may be answered correctly or incorrectly by the VA, as a reinforcement of the corresponding initial bias presented to the subject: in the low-score condition, 6 of the 12 difficult questions are answered incorrectly; in the high-score condition, all 12 difficult questions are answered correctly. Importantly, no easy questions are ever answered incorrectly, since we found in pilot tests that doing so typically caused unnecessary frustration in the subjects, along with an irreversible feeling that the VA is useless. For that reason, incorrect answers to difficult questions were chosen to trigger a sense of *doubt* in the subjects; even though they may not know the correct answer, they should feel that the VA's answer is wrong, without seriously hurting its reputation. For example, for the question, *"What is the distance between Barcelona and Madrid?"*, the VA's incorrect answer is 1000 km (it is actually 504 km).

Questions can also be divided into two types, depending on the length of the interaction they are expected to trigger. Some questions and answers were prepared for forcing an exchange of several conversational turns. For example, after the subject asks *"What is the melting temperature of aluminum?"*, the VA may ask what measurement unit it should provide the answer in (Celsius or Fahrenheit degrees). Using this strategy, we force subjects to have longer interactions with the VA and produce more dialogue acts (i.e., not only questions but also answers). This will be useful for collecting more speech data in the future.

2.3. Evaluation survey

To assess the progress of the subjects' degree of trust throughout the session, they are required to complete a simple survey after questions 6, 12 and 18. The first question in the survey is *"So far, how confident are you in the system's ability to answer questions?,"* and is answered in a 5-level Likert scale presented using a 5-star metaphor, as seen in the top part of Figure 2. Only after answering this question, subjects are reminded that the current VA received an average of X stars by other users (as explained above, $X = 4.9$ in the high-score condition, and $X = 1.4$ in the low-score condition) and are required to explain in a few words what they attribute any difference to between the two ratings (bottom part of Figure 2).

Figure 2: *Screenshot of the evaluation survey.*

The purpose of this survey is twofold. First, we measure the subjects' current degree of trust. Second, and only afterward, we refresh the anchorage bias introduced at the session beginning. The required textual explanation is intended to make subjects more aware of the bias.

2.4. Chat window and web interface

Subjects interact with the VA via a text chat window, in which they ask for the necessary information to answer each question. The rightmost part of Figure 1 shows the chat window, which looks like a traditional online chat. The experiment interface was designed to be available online. This will allow for a massive data data collection in the next stage of our project, in which interaction will be speech-based.

We built the VA dialogue system using the OpenDial toolkit [20]. We made simple modifications to the toolkit's source code to integrate it into our application. The dialog system was implemented with a separate 'dialogue domain' for each question – i.e., a separate set of rules to trigger the system responses. We built pattern-matching rules to cover the potentially many ways in which subjects may formulate their sentences. For example, a subject may type *"how far is mardel from bsas?"*; thus, we have rules for matching 'mardel' and 'bsas' with 'Mar del Plata' and 'Buenos Aires', respectively. All of these rules are deterministic: we did not use OpenDial's support for probabilistic rules.

3. Results and Analysis

In this section, we analyze the results of our preliminary experiments. Our present goal is to gather different sorts of evidence regarding whether the protocol succeeds in influencing subjects into trusting or distrusting the conversational agent's skills. A total of 15 volunteers participated (8 female, 7 male; mean age 31.2, stdev 9.7); 7 subjects took place in both study conditions (high-score and low-score) and 8 in just one condition. Thus, in total there were 11 participants in each condition.

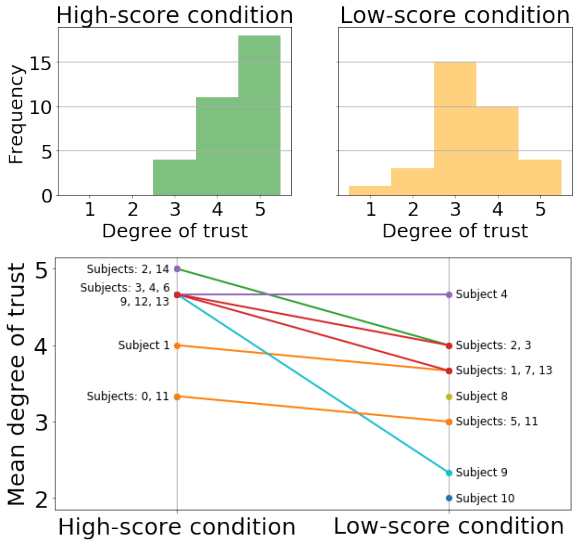

Figure 3: *Reported evaluation of VAs. Top: Histogram of the responses in each condition. Bottom: Comparison of responses from individual speakers in each condition.*

3.1. Reported evaluation of the VAs

After questions 6, 12 and 18, subjects answered a brief evaluation survey (see Section 2.3), which gives us an explicit, conscious assessment of their degree of trust in the VA's skills. Figure 3 shows the histograms of responses to these questions in each of the study conditions (high-score and low-score). We observe marked differences in both distributions, with mean val-

ues of 3.39 for the low-score VAs and 4.42 for the high-score VAs. A linear mixed-effects model with subject response as the dependent variable, condition (low or high) as a random effect, and subject identity as a fixed effect (to control for inter-subject differences) reveals a highly significant effect of study condition on subject responses ($\chi^2(1) = 21.17$, $p \approx 0.0000042$). Figure 3 also shows the mean responses reported by each subject individually, with lines connecting the means of the 7 subjects who participated in both conditions. Again, the differences between conditions are clear (with the sole exception of subject 4, who reported the same confidence in both conditions).

So far, these results validate the existence of an anchoring bias effect of the two conditions on the subjects' ratings. However, note that the bias may only affect how they do the evaluation itself (e.g. how they interpret the 5-level Likert scale), but not how they actually feel about the VA. Therefore, our next step is to attempt to measure trust in two alternative ways, to further assess if our experimental setup can effectively influence the subjects' degree of trust in the VA's skills.

3.2. Reported confidence in the VAs' answers

Subjects were required to indicate in a 5-level Likert scale their confidence in each response given by the VA. Our second measure of trust comes from the confidence reported for the *difficult questions* that the VA answered *correctly*. Typically, subjects were unlikely to know the answers to these questions, and were thus unsure of the accuracy of the system's answers. Therefore, we assume that the subjects' overall degree of trust in the VA's skills is expected to show up in the reported confidence levels to such questions. We only include questions answered correctly by the VA in order to be able to compare across conditions.

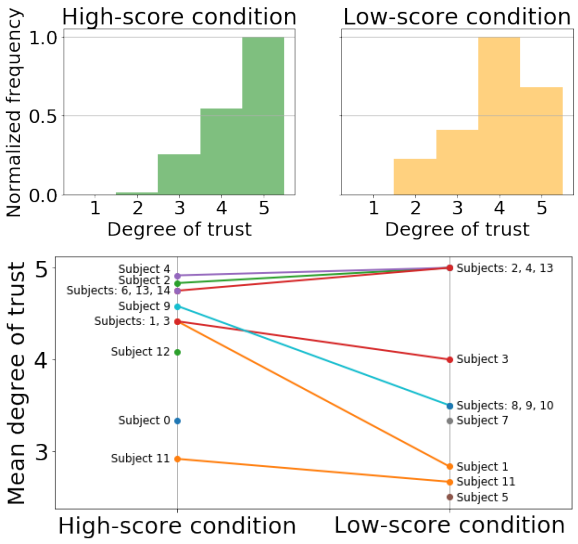

Figure 4: *Reported confidence in the VA's answers. Top: Histogram of the responses in each condition. Bottom: Comparison of responses from individual speakers in each condition.*

Figure 4 shows the histograms of answers in either study condition. As with the responses to the survey, here we also observe marked differences in both distributions, with mean values of 3.71 for the low-score VAs and 4.34 for the high-score VAs. A linear mixed-effects model with subject responses as the dependent variable, condition (low or high) as a random effect, and subject identity as a fixed effect reveals a significant effect of condition ($\chi^2(1) = 12.84$, $p \approx 0.00034$). Figure 4 also shows the mean confidence values reported by each subject, with lines connecting the means of the 7 subjects who took part in both conditions. In this case, the differences between conditions are less clear than with survey responses (e.g. subjects 2, 4 and 13 show similar levels), but in general lower values are observed in the low-score condition.

3.3. Comparing answers from VAs and subjects

For each question, subjects were required to enter both the VA's and their own answer. By design, subjects were unlikely to know the answers to the *difficult* questions, and were thus forced to either guess their responses or (remarkably) *trust the system*. Our third measure of trust consists then in comparing the responses given by VAs and subjects. The higher the proportion of equal responses, the higher the subject's trust in the system.

In the high-score condition, for 91.7% of the correctly-answered difficult questions, subjects used exactly the system's answer as their own (save for minor editions, e.g. in punctuation and capitalization). In contrast, this happened 83.3% of the times in the low-score condition. This difference is notable, in the sense that even when subjects had little idea about how to answer some questions, they chose to respond something else (i.e. make up an answer) more often for the VA in the low-score condition, suggesting that they indeed trusted it less.

4. Discussion

The results of our preliminary experiments indicate that the proposed protocol manages to influence subjects into a particular mental state. One of the virtual assistants was introduced as having received low user ratings and subsequently answered some questions inaccurately during the session. The other virtual assistant had almost perfect user ratings and made no mistakes. When interacting with the former agent, subjects 1) reported lower overall ratings, 2) reported lower confidence in the system's answers, and 3) reused the system's responses as their own less frequently than when interacting with the latter system. These represent three types of evidence supporting the hypothesis that subjects effectively trusted and relied more on the abilities of the former system's to perform this task.

These results are clear for at least half of the subjects in our experiment; for others, however, we found scarce or no evidence of an effect on their degree of trust towards the agent. A more thorough analysis of such cases will be necessary for deciding if their data should be discarded or not (e.g., we could decide to discard those subjects who always reported the same confidence level to all questions). However, for at least half of subjects the protocol seems to have worked, and this is sufficient for moving on to the next stage of our project: putting into production a speech-based version of the protocol for collecting a speech dataset with varying degrees of trust. When ready, this dataset will be made publicly available.

Once the data collection is complete, we will be able to address the two main questions of this project: whether speakers exteriorize trust through their speech, and whether we might be capable of detecting trust automatically with machine learning techniques. Finally, we will also formalize the pilot study reported in the introduction, to further assess the feasibility of reliably annotating trust using external listeners, on this new dataset and on other existing speech corpora.

5. Acknowledgements

This material is based upon work supported by the Air Force Office of Scientific Research under award no. FA9550-18-1-0026.

6. References

[1] J. Kiseleva, K. Williams, A. Hassan Awadallah, A. C. Crook, I. Zitouni, and T. Anastasakos, "Predicting user satisfaction with intelligent assistants," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 2016, pp. 45–54.

[2] S. Sano, N. Kaji, and M. Sassano, "Prediction of prospective user engagement with intelligent assistants," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2016, pp. 1203–1212.

[3] J. Kiseleva and M. de Rijke, "Evaluating personal assistants on mobile devices," in *Proceedings of the 1st International Workshop on Conversational Approaches to Information Retrieval (CAIR'17)*, Tokyo, Japan, 2017.

[4] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An integrative model of organizational trust," *Academy of management review*, vol. 20, no. 3, pp. 709–734, 1995.

[5] J. B. Rotter, "A new scale for the measurement of interpersonal trust," *Journal of personality*, vol. 35, no. 4, pp. 651–665, 1967.

[6] R. Gulati, "Does familiarity breed trust? The implications of repeated ties for contractual choice in alliances," *Academy of management journal*, vol. 38, no. 1, pp. 85–112, 1995.

[7] L. G. Zucker, "Production of trust: Institutional sources of economic structure," *Research in organizational behavior*, 1986.

[8] D. E. Zand, "Trust and managerial problem solving," *Administrative science quarterly*, pp. 229–239, 1972.

[9] M. A. Korsgaard, H. H. Brower, and S. W. Lester, "It isn't always mutual: A critical review of dyadic trust," *Journal of Management*, p. 0149206314547521, 2014.

[10] J. S. Coleman, *Foundations of Social Theory*. Harvard University Press, 1990.

[11] R. Hardin, *Trust*. Malden, MA: Polity Press, 2006.

[12] A. H. Miller, "Political issues and trust in government: 1964-1970," *American Political Science Review*, vol. 68, no. 03, pp. 951–972, 1974.

[13] G. Möllering, *Trust: Reason, Routine, Reflexivity*. Emerald Group Publishing, 2006.

[14] A. C. Elkins and D. C. Derrick, "The sound of trust: Voice as a measurement of trust during interactions with embodied conversational agents," *Group Decision and Negotiation*, vol. 22, no. 5, pp. 897–913, 2013.

[15] B. Waber, M. Williams, J. Carroll, and A. Pentland, *A voice is worth a thousand words: The implications of the micro-coding of social signals in speech for trust research*. Cheltenham: Edward Elgar Publishing Limited, 2012.

[16] P. Brusco, J. M. Pérez, and A. Gravano, "Cross-linguistic study of the production of turn-taking cues in american english and argentine spanish," *Proc. Interspeech 2017*, pp. 2351–2355, 2017.

[17] J. Godfrey and E. Holliman, "Switchboard-1 release 2 ldc97s62," *DVD. Philadelphia: Linguistic Data Consortium*, 1993.

[18] D. L. Sackett, "Bias in analytic research," in *The Case-Control Study Consensus and Controversy*. Elsevier, 1979, pp. 51–63.

[19] A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases," *science*, vol. 185, no. 4157, pp. 1124–1131, 1974.

[20] P. Lison and C. Kennington, "OpenDial: A toolkit for developing spoken dialogue systems with probabilistic rules," *Proceedings of ACL-2016 System Demonstrations*, pp. 67–72, 2016.