AFRL-AFOSR-CL-TR-2020-0008

Automatic Detection of Degree of Trust from Speech

Luciana Ferrer
INSTITUTO DE INVESTIGACION EN CIENCIAS DE LA COMPUTACION
Intendente Guiraldes 2160, Pabellon 1, Ciudad Unive
Ciudad de Buenos Aires, C1428BGA
AR

09/18/2020
Final Report

Air Force Research Laboratory
Air Force Office of Scientific Research
Southern Office of Aerospace Research and Development
U.S. Embassy Santiago, AV. Andrews Bello 2800 Santiago, Chile

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
18-09-2020	Final	01 Nov 2017 to 30 Apr 2020

4. TITLE AND SUBTITLE
Automatic Detection of Degree of Trust from Speech

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA9550-18-1-0026

5c. PROGRAM ELEMENT NUMBER
61102F

6. AUTHOR(S)
Luciana Ferrer, Agustin Gravano

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
INSTITUTO DE INVESTIGACION EN CIENCIAS DE LA COMPUTACION
Intendente Guiraldes 2160, Pabellon 1, Ciudad Unive
Ciudad de Buenos Aires, C1428BGA AR

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/SOARD
U.S. Embassy Santiago
Av. Andres Bello 2800
Santiago, Chile

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/AFOSR IOS

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-AFOSR-CL-TR-2020-0008

12. DISTRIBUTION/AVAILABILITY STATEMENT
A DISTRIBUTION UNLIMITED: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
We proposed to study the problem of detecting the degree of trust a subject has on their interlocutor during a dialogue, based on linguistic and paralinguistic information present in their speech. We focused on human-computer conversations, where we wish to assess the degree of trust the human has on their virtual interlocutor. To this end we proposed to: (1) collect a database of human-computer dialogues, and (2) investigate methods for automatic detection of the degree of trust. During the project, we have successfully accomplished both tasks. The database was collected, curated and annotated and used for the automatic detection of the degree of trust. Our experiments show that it is possible to detect a proxy for the level of trust with a performance significantly better than random, using features that have been previously used to detect speech directed to at risk speakers, like infants, non-native speakers or people with hearing difficulties. To our knowledge, these are the first experiments that indicate that it is possible to automatically detect trust using the voice of the trustee.

15. SUBJECT TERMS
dialog analysis, trust, speech detection

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			MONTES, DANIEL
Unclassified	Unclassified	Unclassified	SAR		19b. TELEPHONE NUMBER *(Include area code)* 571-289-5167

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Automatic Detection of Degree of Trust from Speech

Final Report

Award Number: FA9550-18-1-0026

PI: Luciana Ferrer

CoPI: Agustín Gravano

Summary

We proposed to study the problem of detecting the degree of trust a subject has on their interlocutor during a dialogue, based on linguistic and paralinguistic information present in their speech. We focused on human-computer conversations, where we wish to assess the degree of trust the human has on their virtual interlocutor. To this end we proposed to: (1) collect a database of human-computer dialogues, and (2) investigate methods for automatic detection of the degree of trust.

During the project, we have successfully accomplished both tasks. The database was collected, curated and annotated and used for the automatic detection of the degree of trust. Our experiments show that it is possible to detect a proxy for the level of trust with a performance significantly better than random, using features that have been previously used to detect speech directed to at-risk speakers, like infants, non-native speakers or people with hearing difficulties. To our knowledge, these are the first experiments that indicate that it is possible to automatically detect trust using the voice of the trustee.

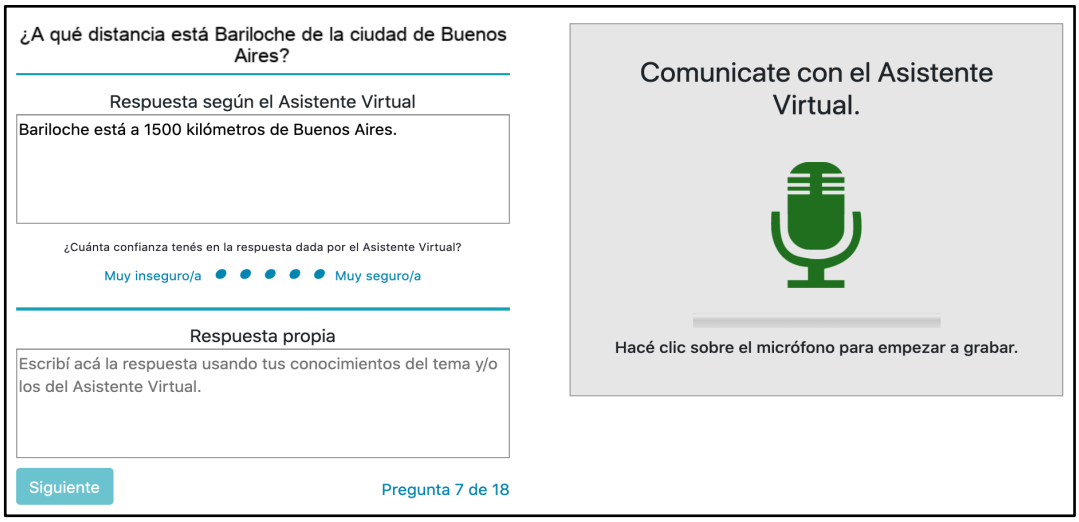
Protocol for collecting a trust dataset

The protocol for collecting the dataset consists of an interactive session where the subject is asked to respond a series of questions with the help of a virtual agent. The interaction with the agent is speech-based: subjects record their questions using a push-to-talk-like interface and the agent responds using a synthesized voice. The responses of the system are designed to elicit mini-dialogues, for example, by being incomplete, demanding further clarification from the subject.

Section structure and initial bias. At the beginning of a series of questions, 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). These two conditions are central in our protocol 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. Subsequently, the agent will respond the questions according to the declared abilities, making no mistakes when the subject was told the agent was very good and making frequent mistakes when the subject was told the agent was bad.

Types of factual questions. Each series 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 days are there in a week?"*) 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 longest rivers in Argentina?"*). 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 questions allow us to manipulate the subjects' varying degree of trust in the VA's skills.

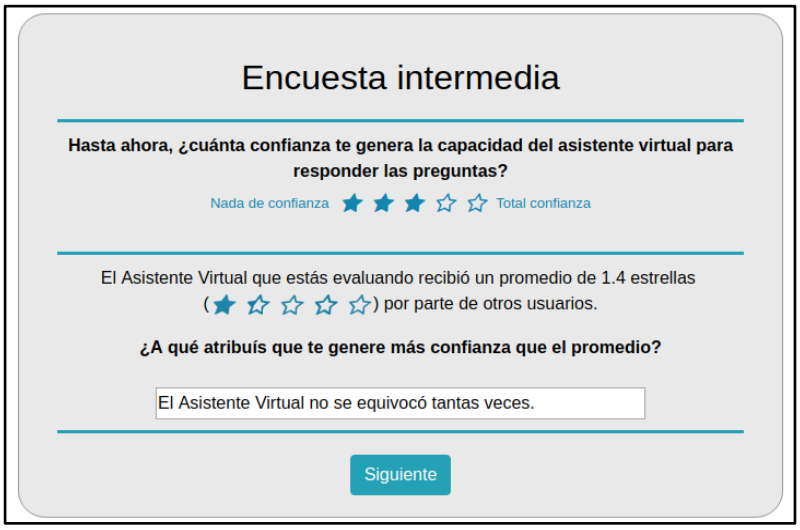
User interface. Throughout the session, the subject is asked to respond the questions in two ways: (1) based only on what the agent said, and (2) based on what he or she believes to be correct. If the subject responds the same in both cases, we can assume that they trusted the agent. Otherwise, if the answers are different, we can assume they did not trust the agent. Furthermore, we also ask the subject to rate the agent's answer with a number of stars from 1 to 5. These two pieces of information are designed to be redundant as a way of checking consistency of the subject's responses. The following figure shows a screenshot of the user interface.



Screenshot of the user interface.

Surveys. At the beginning of a session, subjects are asked to complete a few surveys for demographic information (gender, age, birthplace, first and second languages), for personality type (15 dimensions mapped to the big five personality types), and for their degree of familiarity with, and trust in, virtual assistants and other digital systems.

To assess the progress of the subjects' degree of trust throughout the series, they are required to complete simple *evaluation surveys* after questions 6, 12 and 18. The first question, "*So far, how confident are you in the system's ability to answer questions?*", is answered in a 5-level Likert scale presented using a 5-star metaphor, as seen in the top part of the following figure. Only after answering this question, subjects are reminded that the current VA received an average of X stars by other users (X=4.9 and X=1.4 in the high- and low-score conditions, respectively), and are required to explain in a few words why their score was higher or lower than the average. These things are intended to reinforce the high- or low-score condition.



Screenshot of the evaluation survey.

After completing a series of 18 questions and the third evaluation survey, subjects are required to rate how useful they found the VA, their degree of frustration with it, and how much they trusted it. They also report the extent to which they felt the following emotions and sentiments during the interaction: active, afflicted, attentive, tired, decided, disgusted, distracted, enthusiastic, inspired, uneasy, nervous, and fearful. All questions are answered using a 5-point Likert scale. The purpose of these surveys is to further monitor and understand the subjects' behavior during their interaction with the VA.

Implementation details. The study interface was implemented online, to allow for data collection both in a controlled laboratory, and remotely over the Internet. We built the VA dialogue system with the OpenDial toolkit.[1] We synthesized the VA's responses using Microsoft's publicly available speech synthesizer, with a female Spain Spanish voice.[2] The subjects' utterances were transcribed using Google's publicly available automatic speech recognition system.[3]

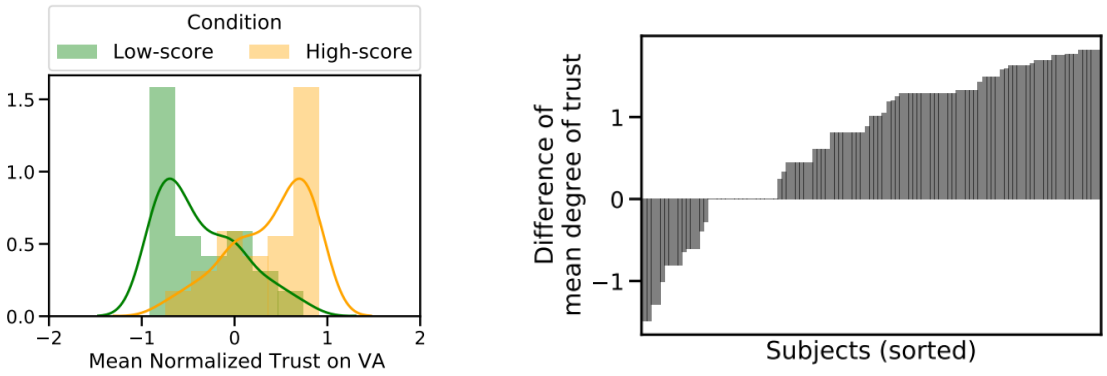
The Trust-UBA dataset

Using this protocol we collected a speech corpus in Argentine Spanish, which we call Trust-UBA. The dataset was curated and subsequently annotated by ten expert listeners. This dataset will soon be made available free of charge for research use. We describe the corpus next; more details can be found in the paper recently submitted to Interspeech [2].

Subjects. Subjects were recruited via ads on social media, emails to student mailing lists, and posters at the University campus. 50 subjects participated at the University, in a controlled, silent environment (we call these the ***in-lab*** subjects). This group was asked to solve two series of questions (one in each condition) and received a small monetary compensation for their time. The remaining 110 subjects participated over the Internet (these are the ***remote*** subjects). In this case we had no control of the environment, which could result in poorer recording quality and lower concentration levels. This group was required to finish at least one series of questions and were included in biweekly draws for a small monetary prize as compensation.

Statistics. From the 160 volunteers, 83 were female, 76 were male, 1 did not reply. The mean age was 27.4 (SD 9.2). Of these, 108 subjects completed two series of questions (50 in-lab, 58 remotely), one in each study condition (high- or low-score); the remaining 52 subjects (all remote) completed just a series in one condition. All subjects reported Spanish as their first language; all but 2 in-lab subjects, and all but 6 remote ones were born in Argentina. Thus, the collected speech is overwhelmingly in Argentine Spanish. The collected data consists of 8493 short audios, with a mean duration of 4.7 seconds (SD 2.1). Subjects that completed two series contributed on average with 62.4 audios (SD 15.7); subjects that completed just one condition, 33.6 (SD 8.1). 740 audios had to be excluded due to technical problems, such as network communication errors.

Protocol effectiveness. We found clear evidence that the protocol succeeded in influencing subjects into the desired mental state of either trusting or distrusting the agent's skills. The left plot in the following figure compares the distributions of the subject-normalized mean trust scores reported in each condition. We observe a clear effect of condition type, with the low-score condition yielding significantly lower trust scores (LME model, p~0). The right plot shows the differences across conditions of the trust scores reported by individual subjects. Based on this evidence, we conclude that the protocol succeeded in inducing the vast majority of subjects into either trusting or distrusting the VAs as intended.



Overall trust in the VA's skills reported by subjects. Left: Histograms of means of normalized trust scores per condition. Right: Differences of mean trust scores from individual speakers across conditions.

1 http://www.opendial-toolkit.net
2 https://azure.microsoft.com/en-us/services/cognitive-services/text-to-speech
3 https://cloud.google.com/speech-to-text

Perceptual annotation of trust. An additional research question in the current project is whether humans are capable of telling solely from the speech signal whether the speaker trusted or distrusted the VA's skills. We gathered a team of psychology researchers and practitioners – who, as experts in human behavior, we considered would be good candidates for succeeding in this task. Ten female expert annotators were asked to listen to a pair of sequences of audios from each in-lab subject (they thus listened to 50 pairs of audios).

Each such sequence was formed by audios produced by a subject in each of the two conditions (low- and high-score). Specifically, these audios were the first recordings in each of the final six questions in a series – during which we expect the trust/distrust effect to be at its maximum level. The six audios in a sequence were merged into a wav file, separated by a simple tone. Each pair of sequences (corresponding to the low- and high-score conditions) was presented to annotators on a web page, in random order. For each pair, they had to select which audio corresponded to utterances directed at the less trustworthy VA. All annotators were paid for this task. We examined inter-annotator agreement using Fleiss' *Kappa* measure, which yielded a value of 0.116. This is interpreted as "slight" agreement above chance. We also conducted a permutation test, which confirmed that this slight agreement is indeed significantly not random ($p \approx 0$). This suggest that the annotators did perceive certain speech cues related to trust, albeit faint and unreliable ones.

Thus, the expert listeners only had a very small level of agreement on the task of indicating which of two series of questions from a certain user corresponded to the highly-rated virtual assistant. On the other hand, as we will describe in the next section, an automatic system trained to perform this same task using all audios in each series is able to detect the correct label with an accuracy significantly better than random, indicating that the speech from the user is indeed affected by their level of trust toward the virtual assistant.

Development of a state-of-the-art emotion detection system

The second task of the project was the automatic detection of trust using state-of-the-art speech processing technology. Since the task of detecting trust from speech had not yet been explored in the literature, we proposed to implement the best emotion detection system available in the literature as a baseline system for trust detection. We believed this to be a good baseline, since the information about the label for both tasks should be encoded in similar features, like the prosody, lexical and syntactic information.

For this reason, in the first half of the project, while the protocol for the collection of the trust database was being developed, we implemented a series of approaches for emotion detection. Our work on the task of emotion detection was published in two papers, one investigating the difficulties involved in evaluating emotion detection systems due to the lack of reliable ground truth labels [5], and another one describing a state-of-the-art method combining both text and acoustic information [3].

Evaluation of emotion detection systems. Human agreement for the task of labeling speech utterances with emotion information is usually low, especially for natural speech, where emotions could be ambiguous or subtle. For this reason, datasets of emotional speech are generally labeled by several human annotators. The common practice in speech emotion recognition literature is to summarize the multiple labels provided by the annotators for a sample into a single one by choosing the majority label. The problem with this approach is that a significant proportion of samples may not be assigned a majority label. These samples are usually ignored for system evaluation, along with any samples initially labeled by the annotators as being from emotions other than the emotions of interest for the specific dataset. This implies that the estimation of emotion recognition performance is incomplete; we do not know how the system will behave when presented with those ambiguous samples, which will certainly appear in practice. In [5], we analyzed the effects that these samples have in system performance and proposed different ways to use the multiple labels available from the annotators during evaluation and to assess system performance without discarding any samples.

State-of-the-art emotion detection system. After studying the literature on emotion detection, we implemented a variety of systems using different features and modeling techniques. The most successful approach was one based on DNNs which merged information extracted from the acoustic signal and from the transcription of the speech in the signal. We proposed to use contextualized word embeddings to represent the information contained in the transcription. The word embeddings were modeled with a time-delay DNN with a pooling layer that computes statistics over all time steps (words, in the case of the text-based system) to generate fixed-length outputs. The acoustic system was based on low-level features describing the spectral

characteristics of the signal over short windows and modeled using a similar architecture as for the text-based system. We also proposed and compare different strategies to combine the audio and text modalities. All our approaches were evaluated on several emotion datasets, focusing mostly on two of the most popular ones: IEMOCAP and MSP-PODCAST. We found that fusing acoustic and text-based systems was beneficial on both datasets, though only subtle differences were observed across the evaluated fusion approaches.

The experience obtained from our work on emotion detection was essential for the development of the trust detection system described next. Yet, we were not able to directly use the same approaches, since they required a significantly larger amount of data to work well than that available in our Trust-UBA dataset. Further, text-based approaches were not possible in this dataset, since the protocol was designed so that very similar phrases were used under different levels of trust. Instead, as we will see, we decided to use simple, high-level features, motivated by the literature and simple modeling techniques that are robust when only small amounts of training data are available. Nevertheless, the knowledge and infrastructure developed during our work on emotion detection was essential for the work on trust detection and allowed a fast turnaround on the results once the Trust-UBA dataset was ready.

Development of a trust detection system

Once the Trust-UBA corpus was curated and annotated, we started the research on whether the level of trust that was elicited on the subjects could be automatically detected from their speech.

Selection of a proxy for the trust level. As in most emotion datasets, where the ground truth label is not known, the actual level of trust each subject in the Trust-UBA dataset is experiencing at every point in time cannot be known with certainty. Nevertheless, the protocol was designed to have several ways of estimating such label. The first way was to assume that the level of trust was determined by the initial score (low or high) for the VA given to the user at the beginning of each series of question. The second way was to use the reports given by each user during the task. The third way was to use the annotations given by the expert listeners. As mentioned above, the scores given by expert listeners had very low agreement and, hence, could not be reliably used as labels. The scores provided by the user during the session correlated heavily with the score provided by the protocol at the beginning of the series of question, indicating that this score did affect the trust that the user had on the system. Hence, both the VA score or the scores provided by the user for every question or in the intermediate surveys could be used as proxies for the trust level that the user experienced during the task. Yet, using the scores provided by the user introduces some difficulties since people tend to use the 1 to 5 scale differently depending on their personality or initial bias. For this reason, for this initial exploratory work, we decided to use the condition of each series (low-score or high-score) as a proxy for the trust level. Future work will include the use of some function of the subject's scores as target label.

Acoustic features per utterance. We implemented a series of features based on those that have been found in previous literature to be useful for characterizing speech directed to at-risk listeners like infants, non-native speakers and people with hearing difficulties. We assumed that such features might also be useful for detecting lack of trust toward the virtual assistant. The characteristics found in the literature for speech directed to at-risk listeners include more frequent and longer pauses, slower speech rate, clearer differentiation of vowel space with respect to formant values, and increased pitch and expansion of pitch range. Guided by these findings from the literature, we designed a series of features aimed at capturing these characteristics. All features can be extracted automatically from the speech signal without the need for manual labeling, except for the transcriptions. The manual transcriptions are used to obtain automatic time-alignments for each word and these alignments are used to discard the silent regions and to obtain timing features like the speech rate and pause-to-speech ratio. We also extract pitch-, energy- and formant-based features by computing statistics (range, median, slopes) over the low-level pitch, energy and formant signals. As a result, we obtained a relatively small number (16) of features for each utterance from each subject.

Data selection and normalization. For the experiments, we selected only the subjects that completed two series of questions: one with a low-score VA and one with a high-score VA. Further, we selected only the subjects that recorded their sessions at the school laboratory, since the remote sessions had challenging recording conditions that are likely to obscure the effect we aim to capture. We leave the experimentation on the rest of the subjects for future work.

As we concluded from the results of the annotation effort, the effect of trust in speech is very subtle. This effect is easily obscured by other factors like the speaker identity and the type of question. The effect of these

two factors is mitigated by doing subject- and question-dependent normalization. That is, before modeling, the features are normalized per subject and then by question identifier where the question identifier determines the text presented to the subject in each screen and, hence, also guides what the subject will say to the VA.

Another factor that could greatly affect the speech characteristics is whether the automated VA worked well for the subject. In some cases, the VA made errors in understanding due to errors in the automatic transcription or in the dialogue system, asking the subject to repeat the question. We observed that, in these cases, subjects tended to speak noticeably slower in an attempt to avoid further errors. The VA errors occurred randomly for both low- and high-score conditions introducing an undesired effect in the data. For this reason, for our experiments we discard all utterances within a question-series that come after the first time the VA made an error. Subjects that were left with too few utterances after this filtering were discarded. The results described below were obtained on the 19 subjects that had at least 12 question left per series after this filtering.

Modeling method. The features were modeled using random forests, a method that is known for its flexibility and robustness. We explored two different tasks: estimating the trust level for each question or for each question-series. In both cases, we only use the first utterance from the subject for each question in the protocol. Subsequent utterances within each question were usually very short (yes/no or short clarification responses). For the series-level case, the features extracted for each question in the series were summarized by taking the 25, 50 and 75 quantiles. Hence, in this case, 16x3 features are used as input to the random forest.

We obtained the results doing leave-one-speaker-out cross-validation. That is, we trained a different model for each speaker using all other speakers and applied that model to all the samples from the test speaker. After a full round through the speakers, we pooled the resulting scores and computed different performance metrics: cross-entropy, accuracy and area under the ROC curve.

Results. We show that a system can learn to perform the task of detecting the condition (low vs high score) for a question-series with an accuracy of up to 76%, where a random baseline would have a 50% accuracy. These results suggest that the speech produced by a user of a VA is affected by their level of trust toward the system. In particular, these results are significantly better than those obtained using the majority label across the 10 expert annotators, which had a 50% accuracy, equal to the random baseline. This difference can be due to a number of factors. First, annotators were only presented with the last third of each question series while the system was provided all questions in the series. Further, annotators might have been confused by the effect of VA errors (we did not know about this issue when the data was provided to the annotators, so we did not do any filtering). Finally, the system was able to learn from all subjects in a supervised way since it had access to the target label during training. Annotators, on the other hand, were not trained for the task and used only their intuition to solve the problem. We believe that, if annotators were trained to solve the task and were given the same data as the system, they might be able to solve the task as well or better than the system. More details on these experiments and results can be found in the paper recently submitted to the 2020 Interspeech conference [1].

We would like to note that these results should be only interpreted as a *preliminary* analysis suggesting that the proposed features contain useful information about the task. This is because the Trust-UBA dataset was collected under very controlled conditions that are not necessarily representative of most use cases. Further, the normalization per speaker and question identifier done in our experiments imposes conditions on the use case that are probably too restrictive. Hence, further data collection is needed to confirm the findings in this work in a less controlled setting.

Conclusions

The project has resulted on the first database for research on the effect of trust in the speech of the trustee. We hope this database (and the protocol developed to collect it) will be useful for the community to enable research on this topic. Further, the dataset allowed us to obtain what we believe is the first evidence in the literature that it is possible to automatically detect the level of trust from the speech of the trustee. As discussed above, though, these should be considered preliminary results since the dataset was collected under a very controlled environment and the experiments performed under some strong assumptions that normalization was possible, which may not be reflective of most use cases. Given the knowledge obtained from this project, we believe it is possible to move forward to more challenging and realistic scenarios.

Publications

We have written five papers under this project, three of them have already been published and two are under review.

[1] L. Pepino, P. Riera, L. Gauder, A. Gravano, and L. Ferrer, **"Detecting distrust towards the skills of a virtual assistant using speech"**, submitted to Interspeech 2020

[2] L. Gauder, P. Riera, L. Pepino, S. Brussino, L. Ferrer, and A. Gravano, **"Trust-UBA: A Corpus for the Study of the Manifestation of Trust in Speech"**, submitted to Interspeech 2020

[3] L. Pepino, P. Riera, and L. Ferrer, and A. Gravano, **"Fusion Approaches for Emotion Recognition from Speech Using Acoustic and Text-Based Features"**, in Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020

[4] L. Gauder, A. Gravano, L. Ferrer, P. Riera, and S. Brussino, **"A protocol for collecting speech data with varying degrees of trust"**, in Proc. SMM19, Workshop on Speech, Music and Mind 2019, 2019

[5] P. Riera, L. Ferrer, A. Gravano, and L. Gauder, **"No Sample Left Behind: Towards a Comprehensive Evaluation of Speech Emotion Recognition Systems"**, in Proc. SMM19, Workshop on Speech, Music and Mind 2019, 2019

Further, we are currently under the process of writing a journal paper summarizing all our conclusions from this project.