



Automatic Detection of Degree of Trust from Speech

Lara Gauder^{1,2}, Leonardo Pepino^{1,2}, Pablo Riera¹, Silvina Brussino^{3,4}, Jazmín Vidal^{1,2}, Agustín Gravano⁵, Luciana Ferrer¹

Instituto de Investigación en Ciencias de la Computación (ICC), CONICET-UBA, Argentina
Departamento de Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires (UBA), Argentina
Facultad de Psicología, Universidad Nacional de Córdoba (UNC), Argentina
Instituto de Investigaciones Psicológicas, CONICET-UNC, Argentina
5- Escuela de Negocios, Universidad Torcuato Di Tella, Argentina

Introduction

¿How do we define trust?

Trust is based upon the trustor's perception of the trustee's ability, benevolence and integrity [Meyer et al.].

¿How does it affect us?

- Linguistic aspects
- Paralinguistic aspects

Introduction

¿Why collect a new dataset?

No corpus is available with annotations with variances in the trust level and large enough to allow statistical analysis or machine learning experiments.

New protocol

We design and implement a protocol that consists of a web application where the subjects interact with different Virtual Assistants while being induced to present different degrees of trust towards them.

Task: Fill out a form with the help of a Virtual Assistant.



Answer

Subject interaction with the Virtual Assistant

¿A qué temperatura hierve el agua? Respuesta según el Asistente Virtual Escribí acá la respuesta usando la información dada por el Asistente Virtual. I ¿Cuánta confianza tenés en la respuesta dada por el Asistente Virtual? Muy inseguro/a O O O O Muy seguro/a

Respuesta propia

Escribí acá la respuesta usando tus conocimientos del tema y/o los del Asistente Virtual.







Pregunta 1 de 18

Presentation of the Virtual Assistant

Presentación del Asistente Virtual

El Asistente Virtual que vas a evaluar ya fue probado por muchos usuarios.

Les preguntamos cuánto confiaban en la capacidad del Asistente para responder preguntas, y en promedio le asignaron 1.4 estrellas

(★ ☆ ☆ ☆ ☆). ¡Ahora nos interesa saber tu opinión!

¡Comenzar a evaluar!

Example "low-score condition"

Deseo interrumpir mi participación.

Presentación del Asistente Virtual

Ya terminamos con la evaluación de prueba. A continuación, te presentamos un Asistente Virtual.

El Asistente Virtual que vas a evaluar ya fue probado por muchos usuarios.

Les preguntamos cuánto confiaban en la capacidad del Asistente para responder preguntas, y en promedio le asignaron 4.9 estrellas

🛉 ★ ★ 🌪). ¡Ahora nos interesa saber tu opinión!

¡Comenzar a evaluar!

Example "high-score condition"

Deseo interrumpir mi participación.

Annotations

Encuesta intermedia Presentación del Hasta ahora, ¿cuánta confianza te genera la capacidad del asistente virtual para responder las preguntas? **Asistente Virtual** Nada de confianza 🔹 🔹 🔹 🛣 式 Total confianza El Asistente Virtual que estás evaluando recibió Ya terminamos con la evaluación de prueba. A continuación, te un promedio de 4.9 estrellas (🛨 🛨 🛨 🛨 👉 por parte de otros usuarios. presentamos un Asistente Virtual. ¿A qué atribuís que te genere menos confianza que el promedio? El Asistente Virtual que vas a evaluar ya fue probado por muchos Respondió bien todas las preguntas usuarios. Les preguntamos cuánto confiaban en la capacidad del Asistente para responder preguntas, y en promedio le asignaron 4.9 ¿Cuánta confianza tenés en la respuesta dada por el Asistente Virtual? estrellas Ahora nos interesa saber tu opinión! Muy inseguro/a O O O O O Muy seguro/a (+ Comenzar a evaluar! Siguiente Deseo interrumpir mi participación

<u>Expected</u>: based on the Virtual Assistant's abilities

Self-reported: provided by subjects



<u>Perceived</u>: annotated by third-party listeners based only on the subjects' speech

Corpus: Trust-UBA

Database statistics

	In-lab	Remote
Amount subjects	50	34
Age	24.26 years (stdev 4.1)	30.32 years (stdev 12.08)
Amount audios	2950	1980
Mean duration per audio	3.97 seconds (stdev 1.71)	4.17 seconds (stdev 2.63)
Mean duration per session	49 minutes (stdev 12)	56 minutes (stdev 39)

Corpus: Trust-UBA

Protocol effectiveness



mean('high-score condition') - mean('low-score condition')

Machine learning experiments

Condition Prediction



Feature extractor

Random Forest

RF

Series Level



Feature Extraction

We focused on features used to measure hyperarticulation

- Syllable rates with and without pauses
- Pause to speech ratio
- Pitch features (7)
- Range and energy end slope
- Features from the first 2 formants. (4)

In total, we calculated 16 features

Machine Learning Methodology

Data splitting

- Subjects recorded at university
- Subset of 19 subjects with at least 12 questions asked before the first virtual assistant error.
- Leave One Speaker Out (LOSO)

Data balancing

- Balancing of questions per condition during training
- Use of 10 different seeds (for undersampling) and score averaging at evaluation. Normalization
- By subject and question
- Correction of the imbalance of questions by series.

Model and evaluation

- We trained Random Forests
 - **500** trees
 - Max Depth: 20
 - Gini impurity
 - $\circ ~\sqrt{N_F}$ features per split
- We didn't explore hyperparameters or model variants because we did not have a held-out set and data was scarce
- Threshold in p(c|x) = 0.5
- Bootstrapping in test set with N = 1000
- Normalized cross-entropy and accuracy

Results: Model performance



Results: Feature importance

Recursive feature elimination

- Syllable rate including pauses
- Pitch Final Slope
- Pitch median



Η

L

Results: Reported trust vs Random forest scores



Conclusions

- The collected dataset will be opensource.
- Preliminary results show that the developed protocol is effective for eliciting trust or distrust in a virtual assistant.
- Expert annotations show a low agreement. This is a cue of the problem difficulty.
- Preliminary results are not applicable in real-word scenarios.
- Preliminary results using machine learning seem to show that some speech features have information about the induced bias.
- The machine learning model was not successful in generalizing to different experimental conditions.
- Therefore, it is necessary to collect more data in more diverse conditions with more subjects.





Thanks for your attention!

Lara Gauder *mgauder@dc.uba.ar*

Leonardo Pepino *lpepino@dc.uba.ar*