The POTUS Corpus, a database of weekly addresses for the study of stance in politics and virtual agents

Thomas Janssoone, Kevin Bailly, Gael Richard, Chloé Clavel

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1546–1553
Marseille, 11–16 May 2020

The POTUS Corpus, a database of weekly addresses for the study of stance in politics and virtual agents

Thomas Janssoone*, Kévin Bailly*, Gaël Richard†, Chloé Clavel†
*Sorbonne Universités, UPMC Univ Paris 06, CNRS UMR 7222, ISIR,F-75005 Paris France
{name}@isir.upmc.fr
†Télécom ParisTech,Université Paris-Saclay,Paris, France
{firstname.lastname}@telecom-paris.fr

Abstract

One of the main challenges in the field of Embodied Conversational Agent (ECA) is to generate socially believable agents. The common strategy for agent behaviour synthesis is to rely on dedicated corpus analysis. Such a corpus is composed of multimedia files of socio-emotional behaviors which have been annotated by external observers. The underlying idea is to identify interaction information for the agent's socio-emotional behavior by checking whether the intended socio-emotional behavior is actually perceived by humans. Then, the annotations can be used as learning classes for machine learning algorithms applied to the social signals.
This paper introduces the POTUS Corpus composed of high-quality audio-video files of political addresses to the American people. Two protagonists are present in this database. First, it includes speeches of former president Barack Obama to the American people. Secondly, it provides videos of these same speeches given by a virtual agent named Rodrigue. The ECA reproduces the original address as closely as possible using social signals automatically extracted from the original one. Both are annotated for social attitudes, providing information about the stance observed in each file. It also provides the social signals automatically extracted from Obama's addresses used to generate Rodrigue's ones.

Keywords: Multi-modal Social Signal, Signal Processing, Embodied Conversational Agent, Audio Video Corpus, POTUS

1. Introduction

ECAs can improve the quality of life in our modern digital society. For instance, they can help soldiers to recover from Post Traumatic Stress Disorder or help a patient to undergo treatment [Truong et al., 2015] if they are empathetic enough to provide support. The main challenge relies on the naturalness of interactions between Humans and ECAs. With this aim, an ECA should be able to express different stances towards the user, as for instance dominance for a tutor or friendliness for a companion.

To give ECAs the capacity to express emotions and interpersonal stances is one of the main challenges in their design [Vinciarelli et al., 2009b]. However, this field of research is thriving as more and more databases are available for the processing of social signals [Vinciarelli et al., 2012]. These databases are mainly audiovisual and provide mono-modal or multi-modal inputs to Machine Learning methods [Pentland, 2004, Rudovic et al., 2017, Baltrušaitis et al., 2017]. Features such as prosodic descriptors or activations of facial muscles, potentially labeled as Action Units (AUs see Figure 1), are extracted to recognize a social expression (emotion, stance, behavior ...). Data is usually labeled by an external annotator who rates his/her perception of the ongoing interaction (*e.g* the levels of valence, arousal, antagonism, tension ...). These annotations provide different classes useful for supervised machine learning algorithms that link them to features extracted from the audio-video files. Thanks to the latest computer-vision algorithms, these features can now be automatically computed if the data quality is good enough. Solutions for these automatic extractions are proposed in part 3.2.. The final purpose of this corpus is the study of interpersonal stance defined by Scherer [2005] as the "characteristic of an affective style that spontaneously develops or is strategically employed in the interaction with a person or a group of persons, colouring the interpersonal exchange in that situation (e.g. being polite, distant, cold warm, supportive, contemptuous)". Indeed, the scheduling of non-verbal signals can lead to different interpretations. Keltner [1995] illustrates the importance of this multi-modality dynamics: a long smile shows amusement while a gaze down followed by a controlled smile displays embarrassment. The way to obtain proper annotations of interpersonal stance is detailed in part 3.4..

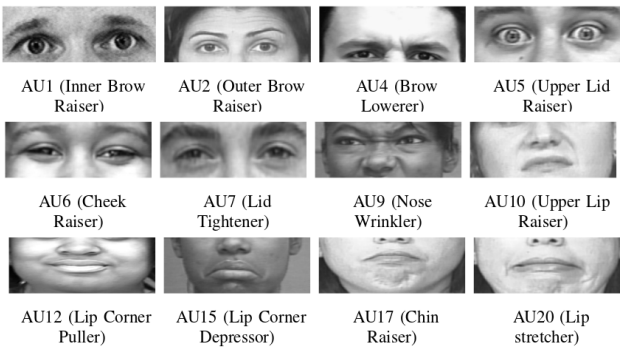

Figure 1: Facial Action Unit locations, images are obtained from http://www.cs.cmu.edu/~face/facs.htm

An efficient methodology of corpus study for ECA generation is proposed by Cassell [2007] with the development cycle *Study-Model-Build-Test*. This circular methodology begins with data collection. Their interpretation allows elaborating a formal model that can be implemented on a platform of virtual agents. This result is then evaluated by humans to correct it or to initiate a new collection of data if too

many failures are noticed. A corpus is usually composed of a set of files, generally multimedia, containing interactions of humans (sometimes with one or several agents and sometimes with a human alone) that are annotated to describe the observed communicative behavior. This methodology comes from the social sciences that developed this process. It allowed the development of a rich literature, based on observation.

Rehm and André [2008] nuance this by underlining that proposed corpus and models did not take into account the particular constraints due to generation. For instance, the multi-modality or the synchronization of the different signals involved is rarely the focus of these studies. Yet, for the design of ECAs, these pieces of information are essential. This explains why new corpora have still to be built for the synthesis of conversational agents. For this specific task, Rehm and André [2008] indicate that two strategies can be applied for corpus-based synthesis: either by directly cloning behaviors from the corpus corresponding to the desired affect (”cloning”-strategy), or by extracting rules that will construct a common model for the generation (”rule-extraction”-strategy). The ”rule-strategy” is widely used with machine-learning algorithms applied on the social signals events to extract sequences [Chollet et al., 2013], temporal rules [Janssoone et al., 2016] or even representations with deep-learning [Dermouche and Pelachaud, 2019] for example. Yet the ”cloning” strategy is rarely evaluated.

This work presents the POTUS Corpus, a new database containing videos of political addresses to the American people performed by two protagonists: former president Barack Obama and a virtual agent named Rodrigue. The virtual agent was generated with social signals automatically extracted from the videos with the human. Then, both sets of videos were annotated in stances (friendliness and dominance). This allows us to evaluate the ”cloning” strategy introduced by Rehm and André [2008] and, in the future, also allow to compare it to ”rule-extraction” strategies. Figure. 2 presents the protocol that we proposed and that underlies the building of the POTUS corpus.This corpus will be soon freely distributed to the community here[1].

2. Related work

Many corpora exist to study different affective phenomena by proposing various supports or modalities as well as a diverse set of annotations, more or less rich and precise. The goal of our work is to find relevant information for the synthesis of attitudes in a virtual agent during a face-to-face interaction. A second constraint was to use algorithms to automatically evaluate the different social signals we could control on the ECA.
A first step has been to study existing databases in order to determine which ones can satisfy the established restrictions. One requirement is to have good quality audio and video files in order to be able to extract the characteristics automatically, and annotations in attitude.

[1] https://clavel.wp.imt.fr/corpora/

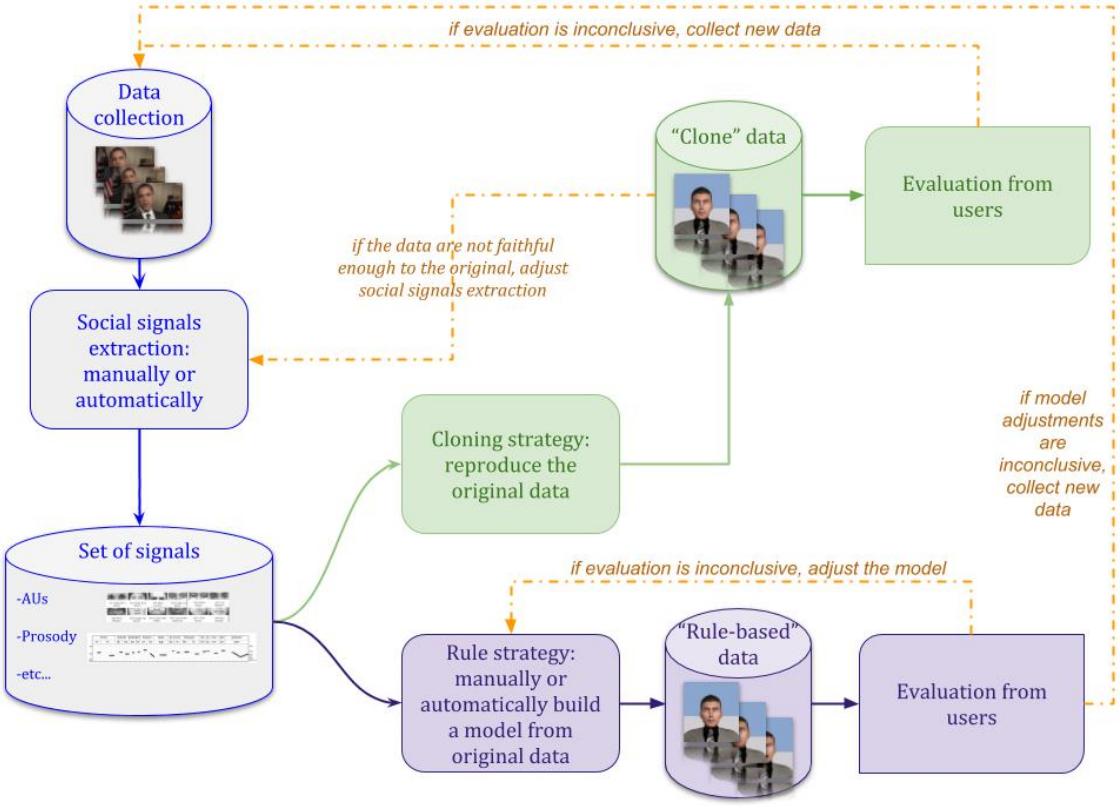

Figure 2: Pipeline sumarizing the two main strategies for synthesis of virtual agents

The main limitation lies in the fact that an attitude needs time to develop. For example, *Gemep* [Bänziger and Scherer, 2010] or *AFEW-VA* [Kossaifi et al., 2017] offer sequences that are too short for the studies. The second challenge concerns the extraction of *action units*. The solution of Nicolle et al. [2016], an AU extractor we used, requires a video of good quality where the face is clearly visible and facing grontwards. Thus, corpora such as the *Canal 9 political debate* [Vinciarelli et al., 2009a] or the *TED database* [Pappas and Popescu-Belis, 2013] have many changes of plan making the algorithm ineffective. Some corpora were also filmed at an angle like *IEMOCAP* [Busso et al., 2008] or the corpus *Tardis* [Chollet et al., 2013] which also disqualifies them, just like *Recola* [Ringeval et al., 2013] where the participants, during a survival task, look at a sheet by tilting their head, which affects the performance of the detector.
The lack of database with "visible-face" interactions and stance annotations led us to build the POTUS Corpus the design of which is detailed in the following section 3..

3. Corpus design

3.1. Weekly addresses of Barack Obama

The limitations of existing corpora, related to the quality of the videos or the provided annotations, motivated the creation of the POTUS Corpus. To find interesting information for the synthesis of the behavior of a virtual agent, videos of good qualities in which the protagonists exaggerate the expression of their emotional state have been looked for. Besides, the automatic detection constraint of *action units* requires the speaker to be facing the camera.
These constraints quickly oriented the research of the audio-video files necessary for the realization of the corpus of speeches made by political personalities. Previous studies, including D'Errico et al. [2013], support this choice: politicians are generally well trained in controlling their images to reach a specific audience. The recent camera-facing speeches thus make it possible to have a satisfactory extraction of social signals with various messages. A first step in the realization of the POTUS corpus is thus focused on weekly addresses of President Obama during his two mandates. During his tenure, President Obama made an address to the American people every Saturday morning called *weekly address* [2]. In each video, President Obama makes a short speech, usually alone in front of the camera, to comment on the news and discuss his actions. The communicative talents of the former president[3] make these videos particularly interesting to study.

3.2. Feature extraction

We choose to focus on a set of social signals composed by *facial AUs activation* (see Fig1), *the head pose* and information such as *dialogic events*. We detail here the process that is used to compute these descriptors.

The 3D head pose is estimated with Intraface [Xiong and De la Torre, 2013], a fully automatic face tracker. Its outputs are the pitch, the yaw and the roll of the head of the actor present in the video. We use these values as descriptor after a moving average smoothing over a 3 frames window and we cluster them in 10 degrees group. We then create events when the head passes from one 10 degrees group to one of the next 10 degrees group. Hence we keep the continuity of the original signal in our new symbolic temporal events.
The Action Units are estimations of facial expressions following the *facial action coding system* introduced by Ekman and Friesen [1978][4], as shown in Figure 1. They were automatically detected using the solution proposed by Nicolle et al. [2016].

3.3. Mirror-Agent generation

For each President Obama's *weekly address*, a similar video with a virtual agent was made which will be titled Agent-Mirror thereafter. For this, the Greta platform[5] was used with Agent Rodrigue. The design of the Agent-Mirror was done as follows.

The audio of the original video has been extracted and added to the Agent-mirror video. From the video, head movements and action units were evaluated respectively as previously explained. However, these automatically extracted data presented noise and using them directly on the agent was not satisfactory. Indeed, there could be unnatural jolts in the AUs or head movements which affected the realism of the rendering. An operation to find adjustment correction parameters was then applied to ensure that the signals displayed by the agent and those detected by the algorithms were coherent.

This calculation step used artificial data: "control" videos where the agent used one or more of its *action units* or performed precise head movements were generated. The algorithms then analyzed the videos to estimate the intensities of the *action units* and head movements. This allows to compare the measured values with the theoretical ones. Corrective parameters to smooth and shift the measured values to the theoretical ones are computed on subsets of values. This allows to reach an acceptable consensus, resistant to noise due to measurements.

For this, an iterative method is applied to find the smoothing and affine transformation parameters to be applied between $signal_{\text{original}}$ and $signal_{\text{extracted}}$. A local smoothing is applied to $signal_{\text{extracted}}$ on a window of size w, giving the $signal_{\text{smoothed}}$. Several time windows are then chosen randomly. On each of these windows, the averages and variances of $signal_{\text{original}}$ and $signal_{\text{smoothed}}$ are compared to find the resetting coefficients. They are then applied to the full $signal_{\text{smoothed}}$ and the error with the $signal_{\text{original}}$ is computed. If the error is less than the previously calculated error, the coefficients are retained. This process is repeated several times to find the optimal coefficients. The different

[2] https://obamawhitehouse.archives.gov/briefing-room/weekly-address

[3] http://www.scienceofpeople.com/2015/02/body-language-leaders-president-obama/
https://www.fastcompany.com/1070311/communicative-power-barack-obama-how-he-became-president-elect
http://www.mediate.com/articles/sharlandA8.cfm

[4] http://www.la-communication-non-verbale.com/2013/05/facial-action-coding-system-6734.html

[5] https://github.com/isir/greta

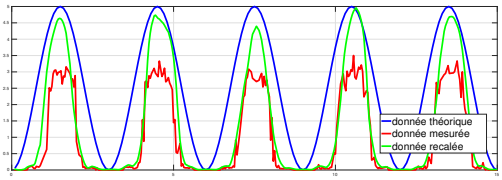

Figure 3: Original(blue) and extracted (red) signals compared that allow to compute recall parameters to obtained the smoothed (green) signals for synthesis.

steps are illustrated in the figure 3.
This method has been applied to correct the values of *action units* and automatically extracted head movements. A video of this process is visible here[6].

Next, the behavior of the mirroring agent is completed with a translation of his head equivalent to the original movement of President Obama as well as random eye blinks according to the process of the Greta platform. Finally, Obama's voice has been added to the Agent-Mirror video with a satisfying lip synchronization thanks to the smoothing of the AUs of the mouth (10,12 15). A sample of the results is visible here [7]. During the design of the cloning strategy, we had to make a few choices on how to reproduce each modality, so we choose to keep the audio from the original files. We also test two other strategies but they were not satisfactory enough for this release of the corpus. The first one was to use prosogram[8] to morph the original voice but we get a very unnatural sound. The second was to use a Text-to-speech system and it is still under development. Furthermore, these methods looked like we went further away from the pure clone strategy we were looking for. Indeed, each one required more choice like what kind of fundamental frequency or pitch to use. We plan to develop and to evaluate these other audio strategies to find how each modality has an impact on the generation of the virtual agent behavior.

3.4. Annotations of the Corpus

3.4.1. Annotation strategy

Argyle [1975] proposes a bi-dimensional representation of interpersonal stances, shown in the figure 4 that we used to build this corpus as a measure of the interpersonal stances perceived in the video. A first dimension represents *Amicality*, a positive value expressing friendliness, a negative value for hostility. The second one is related to *dominance*, a positive value indicating a social superiority, a negative value representing a social inferiority. The present work has a focus set on the scheduling of the multi-modal signals expressed by a protagonist in an intra-synchrony study of his/her stance. Intra-synchrony refers to the study of one individual's multi-modal social signals while the inter-synchrony studies synchrony between two interlocutors.
This model of the interpersonal circumplex suits particularly the precise quantification of socio-emotional phenomena

[6] https://youtu.be/C83s3wvyUF0
[7] https://youtu.be/Yf7Ze7nQNbw
[8] https://sites.google.com/site/prosogram/home

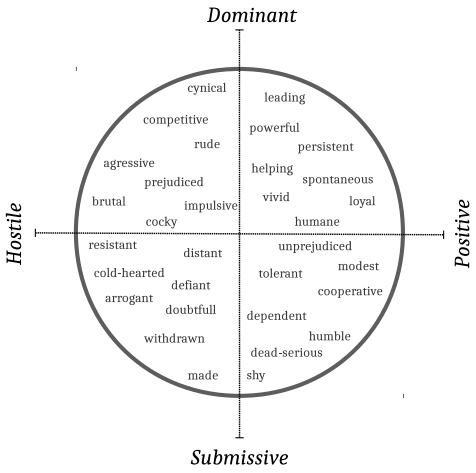

Figure 4: Representation of the interpersonal circomplex, as defined by [Argyle, 1975]

including stance *attitudes*. Locke [2006] illustrates in his review the use of the circumplex for studies in psychopathology as a precise measuring instrument. Several measurement scales are used to allow placement on the circumplex ([Wiggins, 1979, Kiesler, 1996, Horowitz et al., 2006]). Of these, the Interpersonal Adjective Scales-Revised Questionnaire of Wiggins [1979] remains the main measure. It was built to follow the interpersonal circumplex and serves as a reference. It consists of eight scales (or octants) according to eight adjectives (warm, shy, ...). These different scales make it possible to evaluate the attitude of a person without directly asking their judgments along the two axes. This smooths the result and avoids a bias related to the meaning of the words used in the assessment.

Trapnell and Broughton [2006] used these measurements to propose a precise questionnaire to effectively evaluate a social attitude composed of twelve $Likert-5$ scales to do the evaluations. These are based on twelve adjectives corresponding to Wiggins' circomplex for the representation of attitudes: *"Assertive", AS: 90 °, "Dominant", DO: 130 °, "Manipulative", MA: 150 °, "Coldhearted", CO: 180 °, "Aloof", AL: 210 °, "Introverted", IN: 240 °, "Timid", TI: 270 °, "Deferent", DE: 300 °, "Agreeable", AG: 330 °, "Nurturant", NU: 0 °, "Warm", WA 30 °, and "Extraverted", EX: 60 °*. The formulas 1 and 2 make it possible to calculate the scores in Wiggins representation.

$$DOM = (DO + AS + EX) - (DE + TI + IN) \quad (1)$$
$$AMI = (WA + NU + AG) - (MA + CO + AL) \quad (2)$$

Following previous works [Cafaro et al., 2016, Zhao et al., 2016], the *15s slices method* is used to evaluate the social attitudes expressed in the videos. The principle is to cut

each video into 15-ish-second segments that will be annotated independently and randomly. To avoid ruptures in the dialogic events, the slices begin with the start of a sentence and the duration is modulated to end with a pause in the dialogue. This technique of *thin slice*, introduced by Ambady and Rosenthal [1992], was studied and validated by Murphy [2005] as allowing to have coherent judgments on a longer interaction. This makes it possible to observe the evolution of attitudes over relatively short time frames.

3.4.2. Annotation validation on a sub part of the corpus

A first annotation test was carried out to validate the interest of the design of this corpus. To do so, the public of the *Cité des Sciences*, a museum whose purpose is the dissemination of scientific culture, judged the videos during workshops. The results of this study served as proof of concepts for launching a larger annotation campaign. Indeed, important differences were detected according to the videos, indicating differences in the perception of the dominance and the friendliness during the speeches of Barack Obama.

3.4.3. Large scale annotation

Then, the *Crowdflower*[9] platform was used to ask native English speakers to judge these segments in a dozen dimensions. It allows to quickly obtain a large number of annotations. The *thin slices* of the audio-video file were thus presented to users who were then asked to complete a questionnaire. Both the agent part and the Obama part contain 13 addresses cut into a total of 221 thin slices. The agent part was annotated by 54 unique annotators, while 90 proceeded to the Obama part. The difference in the number of annotators is due to the way Crowdflower divide the data into annotation sets and dispatch them to the available contributors of the platform, which we can't control. The dimensions used correspond to Trapnell's IPQ-r evaluation method[10]. The twelve dimensions, also called duodecants, allow a fine assessment of attitudes according to the two dimensions of friendliness and dominance. Each duodecant was evaluated using a five-dimensional Likert scale with five different annotators. The two operations of transposition, presented in the formulas 1 and 2, make it possible to evaluate the attitudes according to the two axes wished, the dominance and the friendliness, as shown in the Figure 7. This method avoids a bias related to the meanings of the words used during the evaluation and thus ensures more robust results. As our annotations are continuous and the annotators did not evaluate the same sets of videos, we used Krippendorff's alpha-reliability coefficient [Krippendorff, 2011] on a ratio scale to evaluate the agreement between annotations. For dominance, we obtain an alpha for Obama of 0.86 and for the agent of 0.83. For friendliness, the alpha is of 0.77 for Obama and 0.84 for the agent. As these coefficients are close to one, they show a good consensus between the annotators, the lowest value for Obama's friendliness may be due to political cleavages.

[9] https://www.crowdflower.com/

[10] www.paultrapnell.com/measures/IPQ-revised.doc

4. Statistical study of the annotations

The proposed analysis here aims to observe the difference between the perception of attitudes according to the observation of the participant: President Obama or the mirroring-Agent. It is motivated by the cognitive model of Brunswik [1956] which emphasizes the importance of outsourcing and assigning social signals during an interaction. Indeed, the Mirror Agent is animated from the videos of President Obama and copies them at best. However, the automatically extracted signals are an approximation of the true values due to computation mistakes and the processing applied to them. Furthermore, only a subset of the signals can be controlled on the agent, adding more differences. Besides, an observer's adjudication between an officer and a former president is likely to be different. This study therefore seeks to quantify these differences to evaluate the influence of personality in the perception of social attitudes. The two cases studied here are 1) the speaker is known and human, 2) the speaker is unknown and a virtual agent.

The grouped annotations for each *weekly address* are shown in Figure 5. From a global point of view, the comparison of average judgments shows slight differences in the perception of friendliness (1.72 on average for Obama against 1.59 for the Mirror Agent) and dominance (1.38 against 1.25). However, an analysis with a Wilcox paired test shows that these differences are not significant. This indicates that, overall, the role and incarnation of non-verbal language does not seem to influence the perception of dominance and friendliness.

A finer analysis of the annotation of each piece of video, each *slice* to use the terminology of Zhao et al. [2016], allows visualizing the differences in the dynamics of perceived

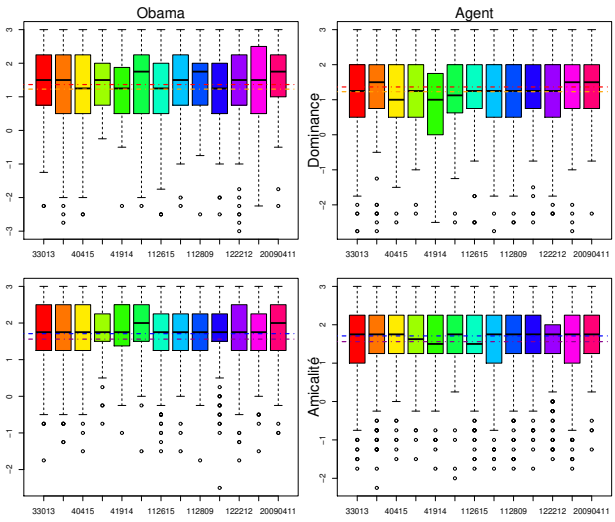

Figure 5: Visualizations of the annotations for each *weekly address*. Dominance is at the top, friendliness down, Pr. Obama left, Mirror Agent right. The dashed lines show the average values for each annotation (red and blue for Obama, orange and purple for the agent). Each color corresponds to a *weekly address*.

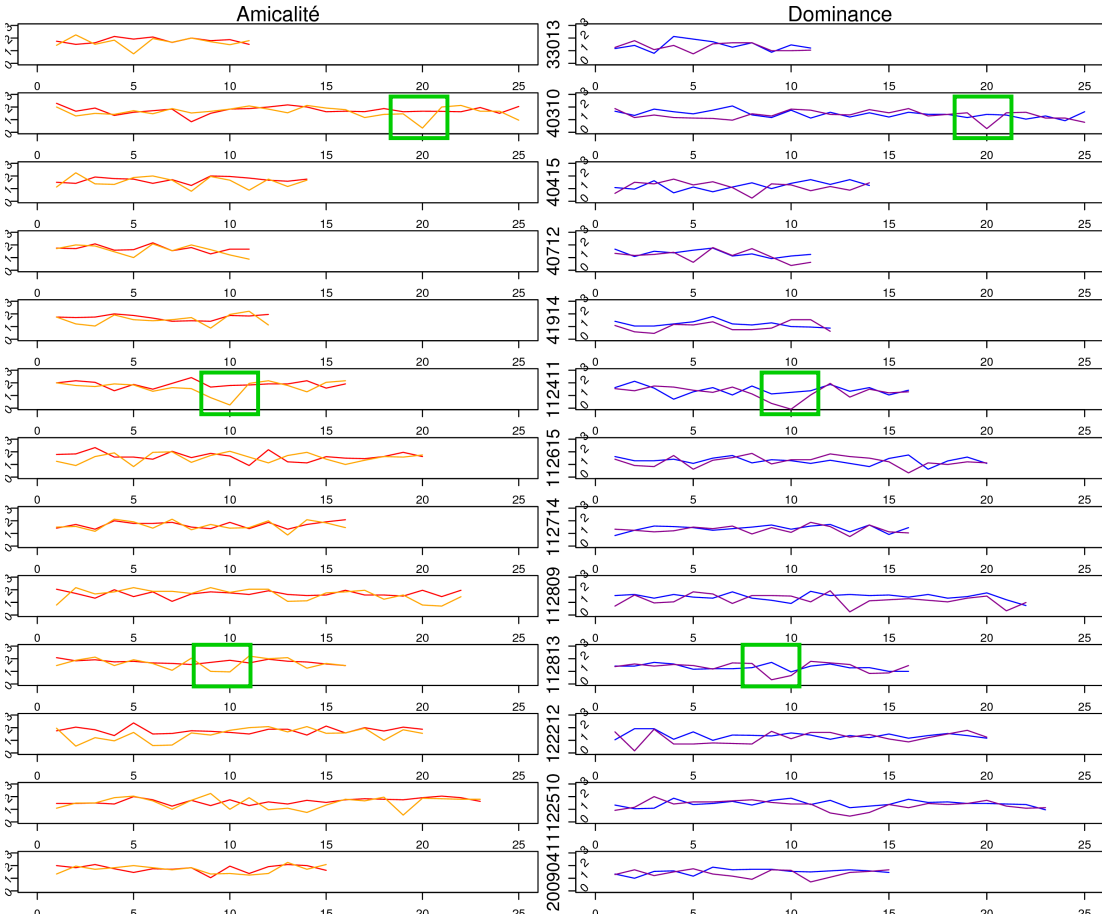

Figure 6: Dynamic annotations for each video, time is counted in *thin slice* annotation. Legend: red: Obama friendliness, orange: mirroring-Agent friendliness, blue: dominance Obama, purple: dominance mirroring−Agent.

attitudes. This can be seen in figure 6 which shows the evolution of annotations in dominance and friendliness, for each *weekly address*, over the slices. This allows observing locally the differences between Barack Obama and the Mirror Agent.

Generally, annotations remain consistent between the human and the mirroring-Agent. Differences can be observed in some places, the most important being in friendliness and dominance at the same time and always in the same direction (sharp decrease of the perception of the agent compared to Obama, framed in green in the figure). An explanation, proposed after viewing the scenes in question, may lie in a disturbing perception of the agent related to the *uncanny valley* Mori et al. [2012]. It is the phenomenon according to which the more an agent resembles a human, the more its imperfections are monstrous, giving a disturbing or even morbid sensation. Moreover, this phenomenon can be reinforced by the inadequacy between the speech pronounced, politically colored and loaded with references including religious, with the voice of Obama and the physical of the virtual agent employed. This phenomenon is therefore still being analyzed. In particular, the use of other agents with this methodology will evaluate the persistence of these differences with different physics.

A statistical analysis was conducted at the level of *slices* to quantify the difference between the videos of President Obama and those of the mirroring-Agent. Following the recommendations of Motulsky [2013], a Wilcoxon matched-pairs test was performed. It did not show a significant difference in friendliness, but an effect was found for dominance (p-value = $0.004 < 0.01$). This can be explained by the stature of Barack Obama, formerly known president, where the mirroring-Agent is a stranger (for now).

5. Conclusion and future work

This paper reports the creation of the *POTUS* corpus which will be soon available here 1. It offers videos of high quality displaying speeches of President Obama to the American people annotated in social attitudes. On top of that is a set of social signals that have been automatically extracted, corrected and adjusted to generate the videos of a mirroring-Agent who copies the former president (see section 3.3.).
The annotations of these mirroring-Agent videos show that the perception of attitudes is similar to the one of the original Obama videos in friendliness and slightly modified in dominance (see section 4.).
As a first step, this makes it possible to validate the processing of automatically extracted signals because the outsourcing/attribution process seems to remain coherent. The ”mirroring” is nevertheless perfectible and could be extended

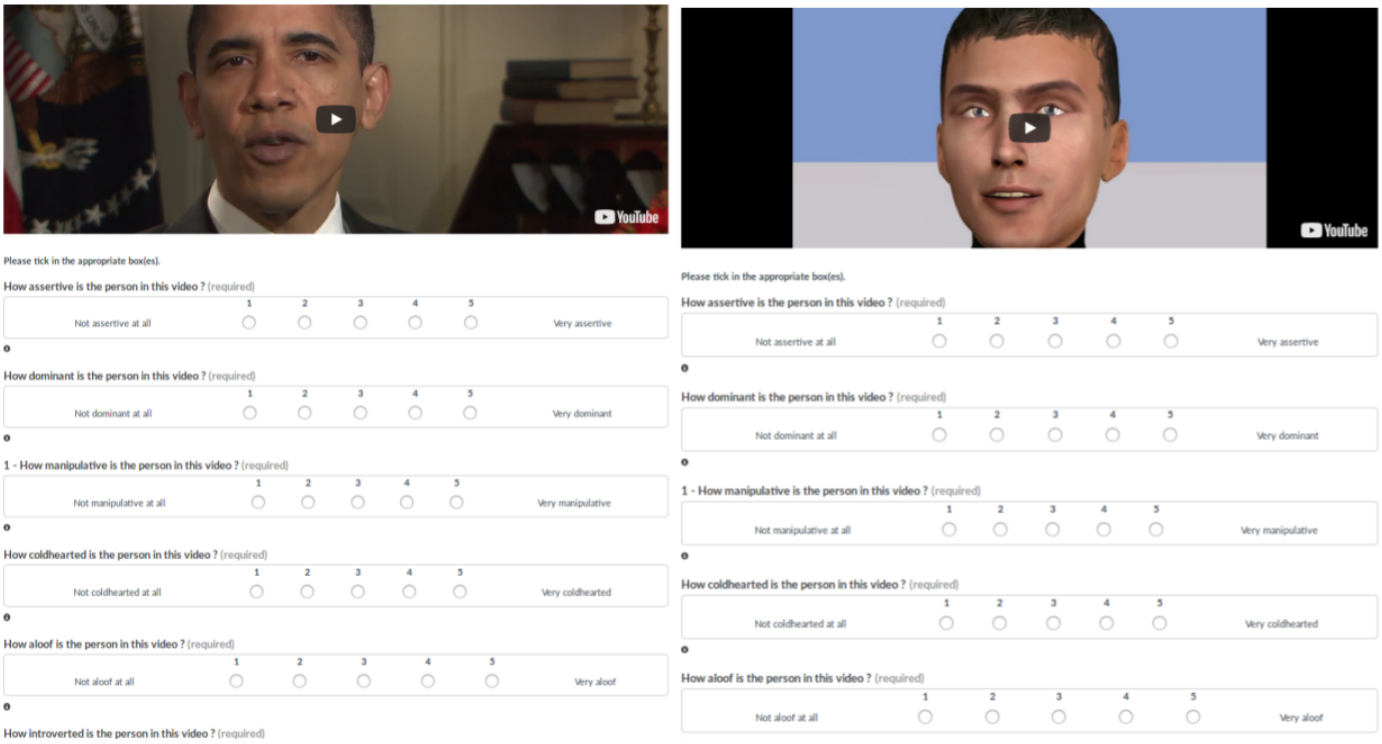

Figure 7: Two snapshots of the *crowdflower* platform screen during the annotation task following the Trapnell questionnaire.

to the voice with similar adjustments on a vocal synthesizer. By providing annotations of the Mirror Agent in the *POTUS* corpus, it is possible to evaluate the two strategies proposed by Rehm and André [2008], ”cloning” and ”rule extraction”. Annotations of the mirroring-Agent give an estimate of the performances for the ”clone” strategy which can then be compared with the performances of rules extracted by models which is a perspective to a short term of this work.

6. Acknowledgement

This work was performed within the Labex SMART supported by French state funds managed by the ANR within the Investissements d´Avenir programme under reference ANR-11-IDEX-0004-0. This work was supported by a grant overseen by the French National Research Agency (ANR-17-MAOI) and by the European project H2020 ANIMATAS (ITN 7659552).

References

N. Ambady and R. Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Journal of Psychological Bulletin*, 1992.

M. Argyle. *Bodily communication*. Methuen Publishing Company, 1975.

T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *arXiv preprint arXiv:1705.09406*, 2017.

T. Bänziger and K. R Scherer. Introducing the geneva multimodal emotion portrayal (gemep) corpus. *Blueprint for affective computing: A sourcebook*, 2010.

E. Brunswik. *Perception and the representative design of psychological experiments*. Univ of California Press, 1956.

C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, Jeannette N. C., S. Lee, and S. S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Journal of Language resources and evaluation*, 2008.

A. Cafaro, H. H. Vilhjálmsson, and T. Bickmore. First impressions in human–agent virtual encounters. *ACM Transactions on Computer-Human Interaction*, 2016.

J. Cassell. *Body language: Lessons from the near-human*. Chicago: University of Chicago Press, 2007.

M. Chollet, M. Ochs, and C. Pelachaud. A multimodal corpus for the study of non-verbal behavior expressing interpersonal stances. In *Proceedings of the Workshop Multimodal Corpora: Beyond Audio and Video, hosted by the International Conference on Intelligent Virtual Agents*, 2013.

S. Dermouche and C. Pelachaud. Engagement modeling in dyadic interaction. In *2019 International Conference on Multimodal Interaction*, ICMI ’19, 2019.

F. D’Errico, R. Signorello, D. Demolin, and I. Poggi. The perception of charisma from voice: A cross-cultural study. In *Proceedings of the Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013.

P. Ekman and W. Friesen. Facial action coding system: a technique for the measurement of facial movement. *Journal of Palo Alto: Consulting Psychologists*, 1978.

L. M Horowitz, K. R Wilson, B. Turan, P. Zolotsev, M. J Constantino, and L. Henderson. How interpersonal motives clarify the meaning of interpersonal behavior: A revised circumplex model. *Journal of Personality and Social Psychology Review*, 2006.

T Janssoone, C Clavel, K Bailly, and G Richard. Using temporal association rules for the synthesis of embodied conversational agents with a specific stance. In *proceedings of the International Conference on Intelligent Virtual Agents*, 2016.

D. Keltner. Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. *Journal of Personality and Social Psychology*, 1995.

D. J Kiesler. From communications to interpersonal theory: A personal odyssey. *Journal of personality assessment*, 1996.

J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic. Afew-va database for valence and arousal estimation in-the-wild. *Journal of Image and Vision Computing*, 2017.

Klaus Krippendorff. Computing krippendorff's alpha-reliability. 2011.

K. D Locke. Interpersonal circumplex measures. *Journal of S. Strack (Ed.), Differentiating normal and abnormal personality*, 2006.

M. Mori, K. F MacDorman, and N. Kageki. The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, 2012.

H. Motulsky. *Intuitive biostatistics: a nonmathematical guide to statistical thinking*. Oxford University Press, USA, 2013.

N. A Murphy. Using thin slices for behavioral coding. *Journal of Nonverbal Behavior*, 2005.

J. Nicolle, K. Bailly, and M. Chetouani. Real-time facial action unit intensity prediction with regularized metric learning. *Journal of Image and Vision Computing*, 2016.

N. Pappas and A. Popescu-Belis. Combining content with user preferences for ted lecture recommendation. In *Proceedings of the 11th International Workshop on Content Based Multimedia Indexing*, 2013.

A. Pentland. Social dynamics: Signals and behavior. In *Proceedings of the 3rd International Conference on Developmental Learning*, 2004.

M. Rehm and E. André. From annotated multimodal corpora to simulated human-like behaviors. *Journal of Modeling Communication with Robots and Virtual Humans*, 2008.

F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *Proceedings of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2013.

O. Rudovic, M. A Nicolaou, and V. Pavlovic. 1 machine learning methods for social signal processing. *Social Signal Processing*, page 234, 2017.

K. R. Scherer. What are emotions? and how can they be measured? *Journal of Social science information*, 2005.

Paul D Trapnell and Ross H Broughton. The interpersonal questionnaire (ipq): Duodecant markers of wiggins' interpersonal circumplex. 2006. URL `http://www.paultrapnell.com/measures/IPQ-revised.pdf`.

K. Truong, D. Heylen, M. Chetouani, B. Mutlu, and A. A. Salah. the international workshop on emotion representations and modelling for companion technologies. In *Proceedings of International Conference on Multimodal Interaction (Workshop)*, 2015.

A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin. Canal9: A database of political debates for analysis of social interactions. In *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, 2009a.

A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Journal of Image and vision computing*, 2009b.

A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schröder. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing*, 2012.

J. S Wiggins. A psychological taxonomy of trait-descriptive terms: The interpersonal domain. *Journal of personality and social psychology*, 1979.

X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013.

R. Zhao, T. Sinha, A. Black, and J. Cassell. Socially-aware virtual agents: Automatically assessing dyadic rapport from temporal patterns of behavior. In *Proceedings of the International Conference on Intelligent Virtual Agents*, 2016.