Be Persuasive! Automatic Transformation of Virtual Agent's Head and Facial Behavior

Afef Cherni[1], Roxane Bertrand[2] and Magalie Ochs[3]
[1]*Aix-Marseille Univ., IN2P3, CNRS, France*
[2]*Aix-Marseille Univ., LPL, CNRS, France*
[3]*Aix-Marseille Univ., LIS, CNRS, France*

Keywords: Multimodal Cues, Persuasion, Embodied Conversational Agent, Machine Learning Methods, Mathematical Convolution.

Abstract: The persuasiveness of a virtual agent refers to its ability to influence, persuade, or motivate users to take specific actions or adopt certain attitudes or beliefs. Virtual agents can use its multimodal capabilities, including non-verbal cues to enhance their persuasiveness. In this paper, we present a new tool called THRUST (from neuTral Human face to peRsUaSive virTual face) to automatically generate the head movements and facial expressions of a persuasive virtual character. This tool is based on a machine learning approach from a human videos corpus to identify the non-verbal persuasive cues. A convolution-based model then transforms neutral non-verbal behavior to a persuasive non-verbal behavior simulated on a virtual face. Videos generated by the tool have been evaluated through a subjective perceptive study with about 90 participants. The results show that the virtual agent's head and facial behaviors generated by the THRUST tool are perceived as persuasive, thus validating the proposed approach.

1 INTRODUCTION

One challenge facing the field of intelligent virtual agent research is the automatic generation of behaviors for embodied conversational agents, especially concerning social and emotional dimensions. In this article, we focus on the generation of *persuasive* virtual agent's behavior. According to (Miller, 2013), the *persuasion* can be defined as *"any message that is intended to shape, reinforce or change the responses of another or others"*. As highlighted in (Burgoon et al., 1990) and (Chidambaram et al., 2012), the persuasiveness of a message does not solely depend on its content but also largely relies on the multimodal components, encompassing different verbal, vocal, and mimo-gestural levels (such as facial expressions, gestures, and pitch). This article specifically focuses on the social cues conveyed through non-verbal signals, such as facial expressions and head movements, that can be expressed by a virtual agent to enhance persuasiveness. However, this article does not address the argumentative aspects related to persuasion, such as identifying arguments to convince, arranging the order of argument presentation, and responding to arguments raised by the persuadee.

Our ultimate objective is to create an Embodied Conversational Agent (ECA) with persuasive capabilities to encourage the elderly population to engage in physical activity. In the field of intelligent virtual agents, numerous persuasive virtual agents have already been created (e.g. (Lisetti et al., 2013; Petukhova et al., 2017; Nguyen et al., 2007)). The main method for modeling persuasive behavior involves identifying behavioral cues that have an impact on perceived persuasiveness, and integrating these cues into virtual agents. The literature emphasizes certain human behavioral cues related to persuasion, such as body movements (Burgoon et al., 1990) and prosody (Petukhova et al., 2017). In the domain of virtual agents, empirical research has shown the importance of certain verbal and non-verbal cues in enhancing the persuasiveness of the virtual agent (Ghazali et al., 2018; Chidambaram et al., 2012). However, as far as we know, there is currently no existing multimodal behavioral model that generates the behavioral cues that a virtual agent should display to be perceived as persuasive.

In this article, we present a new software tool called *THRUST: from Neutral Human Face to Persuasive Virtual Face* with its subjective evaluation. The

Cherni, A., Bertrand, R. and Ochs, M.
Be Persuasive! Automatic Transformation of Virtual Agent's Head and Facial Behavior.
DOI: 10.5220/0012429700003636

In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024) - Volume 1*, pages 359-366
ISBN: 978-989-758-680-4; ISSN: 2184-433X

tool is designed to automatically convert a video of a human into a video of a virtual character that exhibits a persuasive non-verbal behavior. Specifically, the tool extracts automatically the human's head movements and facial expressions, applies modifications based on a proposed computational model, and reproduces the resulting head and facial movements on a virtual face. The main focus of the paper is the computational model that transforms the head and facial movements obtained from the human face to persuasive movements that are then mimicked on the virtual face. The model is evaluated through a perceptive study to validate that the generated animations are perceived by users as persuasive.

Developing a persuasive behavior model implies several challenges, and in particular the precise identification of the behavioral cues associated to persuasion. These cues should be modified on virtual agent to simulate a persuasive behavior. For this purpose, we investigate, in a first step, the relevant behavioral cues of persuasion. We use machine learning techniques to explore the cues of persuasion in a human video corpus. In particular, we explore the POM corpus (Park et al., 2014), which is, to the best of our knowledge, the only multimedia corpus with annotations of perceived persuasiveness. The POM corpus contains web videos of individuals discussing diverse topics in front of a camera. In our machine learning approach, we pay a particular attention to the *interpretability* of the model to be able to identify features that can be easily understood and replicated on virtual agents. Our aim in this research is not to create a classification model to assess persuasiveness but to use machine learning to identify the relevant features of behavioral persuasiveness. Based on the identified persuasive behavioral cues, the THRUST tool converts the cues extracted from the human face into persuasive ones. From the POM corpus, we propose a dictionary to establish reference points that reflect persuasive non-verbal behavior. A convolution-based model, based on this dictionary, is integrated in the THRUST tool to compute the persuasive behavior of the virtual agent .

This paper is organized as follows. In Section 2, we discuss theoretical and empirical research works that explore the behavioral cues related to persuasion. In Section 3, we introduce an overview of the architecture of the THRUST tool. Section 4 details the machine learning framework and Section 5 the convolution-based model. In Section 6, we present the implementation and the evaluation of the tool. We conclude in Section 7.

2 RELATED WORK

Several research studies, particularly in the human-human interaction field, have explored the efficiency of certain behavioral cues. For instance, (Burgoon et al., 1990; Petukhova et al., 2017; Miller et al., 1976), have emphasized the importance of various multimodal behavioral cues. In the present article, we focus on non-verbal cues related to persuasion. As highlighted by(Burgoon et al., 1990), gestures, body movements, smiles and facial expressions are important non-verbal cues that enhance persuasiveness.

At the interactional level, several works studied the positive impact of mimicry on persuasion (Tanner and Chartrand, 2006). In this article, we analyse corpora of monologue excluding the possibility of studying the interactional level. Other contextual elements, such as the appearance of the persuader (Burgoon et al., 1990), may impact the perceived persuasion. In this article, given the size of the considered corpus and the lack of contextual variability, as a first step, we do not consider the influence of the context. Based on the research showing the importance of face and head movements for persuasion (Burgoon et al., 1990), we consider in our study the facial expressions through the study of action units and the head movements. These behavioral cues considered as features of the learned models are presented in more details in Section 4.3.

In the Intelligent Virtual Agent domain, to generate automatically the behavior of a virtual agent, two main approaches are identified in the literature. The first approach relies on rule-based systems that exploit linguistic information from the text and the meaning of gestures, facial expressions or head movements to determine the appropriate signals to express (e.g. (Cassell, 2001; Marsella et al., 2013)). Rule-based approaches remain very limited, given the variability of human expressions across modalities. In a much more recent approach, machine learning methods are used to automatically generate co-verbal gestures (e.g. (Chiu and Marsella, 2014)), facial expressions and body movements from speech (e.g. (Habibie et al., 2021)) or from speech and text to take into account both acoustic and semantic information (e.g. (Ahuja et al., 2020; Kucherenko et al., 2020)). Most studies are based on deep neural networks (e.g. (Chiu and Marsella, 2014; Hasegawa et al., 2018; Kucherenko et al., 2020)) and, more recently, on the use of GAN architectures (e.g. (Ahuja et al., 2020; Habibie et al., 2021)). Compared to existing works, the originality of the work presented in this article is: (1) we generate non-verbal behavior, not from speech or text, but from a video of a human with a neutral

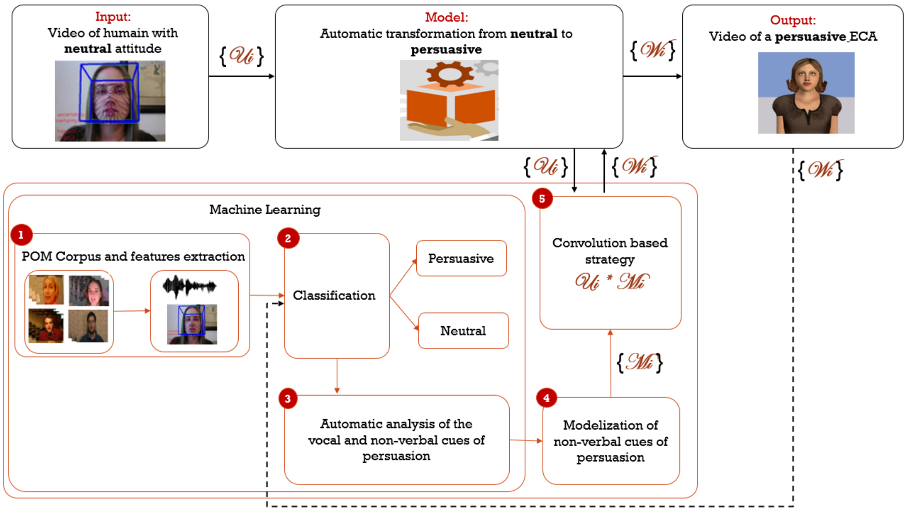
Input:
Video of humain with neutral attitude
$\{\mathcal{U}_i\}$
Model:
Automatic transformation from neutral to persuasive
$\{\mathcal{W}_i\}$
Output:
Video of a persuasive ECA
$\{\mathcal{U}_i\}$
$\{\mathcal{W}_i\}$
$\{\mathcal{W}_i\}$
Machine Learning
1
POM Corpus and features extraction
2
Classification
Persuasive
Neutral
3
Automatic analysis of the vocal and non-verbal cues of persuasion
4
Modelization of non-verbal cues of persuasion
$\{\mathcal{M}_i\}$
5
Convolution based strategy
$\mathcal{U}_i * \mathcal{M}_i$

Figure 1: Global architecture of the system to automatically transform a neutral human video to a persuasive virtual character video. *Input*: a video of a human with automatic extraction of head and facial movements using OpenFace. *Model*: a computational model to automatically transform neutral non-verbal features to persuasive non-verbal features *Output*: a video of a virtual character replicating the behavior of the human but with persuasive head and face movements.

attitude; (2) we generate the facial and head movements whereas most of the existing models consider the body and head movements and (3) unlike existing works that do not allow the generation of socio-emotional behaviors, we propose the automatic generation of *persuasive* behavior.

From a *machine leaning perspective*, few research works have investigated persuasion. The main work has been conducted by Park *et al.* (Park et al., 2014; Park et al., 2016; Nojavanasghari et al., 2016) on the Persuasive Opinion Multimedia (POM) corpus consisting of 1000 movie review videos obtained from a social multimedia website called ExpoTV.com. As proposed by Park *et al.* (Park et al., 2014; Park et al., 2016; Nojavanasghari et al., 2016), we use machine learning algorithms to explore persuasiveness. However, our work differs from the latter in several aspects: (1) contrary to Park *et al.*, in order to obtain *explainable models*, we do not use deep learning methods but "white box" classifiers such as SVM and Random Forest; (2) still in our perspective of interpretability, we consider non-verbal features that can be simulated on a virtual agent[1]; (3) last but not least, our final objective is not to create a prediction model but to explore the non-verbal cues and to use machine learning-based methods in order to create a persuasive artificial agent. In the next section, we present the architecture of the proposed tool to automatically generate a persuasive virtual speaker from a human one.

3 ARCHITECTURE

The THRUST tool takes as input the video of a human and provides as output a video of a virtual agent replicating the same human's behavior but in a persuasive way. The tool is composed of 3 main modules: the *Input* module, the *Model* module, and the *Output* module. The architecture is illustrated Figure 1. We describe each module in the following.

In the *Input* module, the system takes a video of a human speaking in a neutral way. At this step, the OpenFace tool[2] is used to extract the human's head and facial movements. These measures noted as $(\mathcal{U}_i)_{i=1...N}$, where $\mathcal{U}_i$ design the i-th measured feature characterizing the face and head movements, will be saved and used as an input of the *Model* module which transforms them to a set of features $(\mathcal{W}_i)_{i=1...N}$ characterizing the head and face movements of a persuasive speaker as output. For this purpose, a combination of machine learning methods and convolution-based techniques is used. The machine learning methods are employed on an existing corpus to identify the important relevant features of persuasiveness (Step 1, 2, and 3 in Figure 1), which is explained in detail in Section 4. Note that, the resulting learning model is also used as a classifier to automatically determine if the behavior in a video (human or virtual agent) is persuasive (Step 2 in Figure 1). By this way, this learnt model is used to confirm if the transformed fea-

[1] https://github.com/isir/greta/wiki

[2] https://www.cl.cam.ac.uk/research/rainbow/projects/openface/

ture vector $(\mathcal{W}'_i)_{i=1\ldots N}$ is indeed considered as persuasive (as depicted in Figure 1 by the dotted arrow from the ”output” box to the ”classification” box). In the subsequent steps (Step 4 and 5 in Figure 1), a convolution-based method is used to determine how to modify the features to be persuasive (details of these steps are given in Section 5). Finally, in order to deliver the same speech as the original video, but with persuasive head and face movements, the *Output* module uses the embodied conversational agent Greta to simulate the set of variables $(\mathcal{W}'_i)_{i=1\ldots N}$ and generate the video. The vector $(\mathcal{W}'_i)_{i=1\ldots N}$ denotes the value of the head and face movements extracted from the original human video and modified to be persuasive (Figure 1).

4 MACHINE LEARNING FRAMEWORK

4.1 Corpus and Features Extraction

In the step 1 (Figure 1), we consider a specific corpus and extract the features from the video of the corpus. Concerning the choice of the corpus, nowadays, few corpora in the research community are available to study persuasiveness. In this work, we consider the Persuasive Opinion Multimedia (POM) corpus (Park et al., 2014). This corpus is freely available and contains videos of speakers trying to convince on different subjects. POM corpus consisting of 1000 movie review videos obtained from a social multimedia website called ExpoTV.com. It contains different conversational videos cut into a total of 1096 thin slices. Each cut was annotated by different native English-speaking workers of the United States.

Based on the theoretical and empirical research on persuasion presented above (Section 2), we consider the following groups of features: **facial action units** (AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU12, AU14, AU15, AU17, AU18, AU20, AU23, AU24, AU25, AU16, AU28, AU43), **emotions** (Anger, Contempt, Disgust, Joy, Fear, Surprise, Confusion, Frustration), **head movements** (displacement and rotation in (x,y,z) axes, speed of the head movement and its acceleration according to (x,y,z) axis) and **acoustic descriptors** (fundamental frequency f_0, peak slope). For each feature, we computed the mean, median, maximum, minimum, standard deviation and the variance.

4.2 Formalization of the Classification Problem

As illustrated by the step 2 (Figure 1), to identify the importance of the features in the perception of the persuasion, we consider a classification task: based on the features as input, the classifiers have to predict if the features are persuasive as output. As a first step, we consider a binary classification to simplify the learning problem (i.e. prediction if persuasive or not).

4.3 Automatic Analysis of the Vocal and Non-Verbal Cues of Persuasion

In the step 3 (Figure 1), the objective is to compare the performances of the classifiers and then to select the most important features that ensure the highest prediction performances. We propose to experiment different classifiers: the *Naives Bayes* (NB), the *Support Vector Machine* (SVM) and the *Random Forest* (RF). These methods, compared to other neuronal models, are well suited for handling small datasets and have the advantage of interpretability. All experiments were performed with 10-fold cross-validation (CV) where each CV was tested 10 times. In order to estimate the performances of the different classifiers, we compute scores from classifiers returning random predictions, to establish *baselines*. We consider three different strategies: `uniform` (generates predictions uniformly at random) (noted BR), `stratified` (generates predictions with respect to the training set's class distribution) (noted BU0) and `most frequent` (always predicts the most frequent class in the training set) (noted BU1). For each fold of the cross-validation, the random classifiers are fitted on the training set and used to generate predictions on the validation set, for each strategy. Each classifier is trained on 80% of the corpus and tested on 20% of the corpus.

The performances of the classifiers are evaluated though the classical metrics of accuracy and F1 weighted score (to cope with the unbalanced classes). Moreover, we compute the statistical significant differences of the obtained F-scores. The *Student's t-test* is performed to compute the statistical differences between the F1-scores of the classifiers and of the baselines obtained by the k-fold-cross-validation. This test is one of the recommended methods to compare the performance of machine learning algorithms (Dietterich, 1998).

In order to evaluate the importance of each group of features to predict the persuasion, we compute the performance scores of the classifiers considering each group of features and combinations of groups as input. The results show that the emotions do not enable us to obtain significant differences with the baselines. In others words, the emotions are not sufficient to predict persuasion. In the same way, the group of features containing only head movements or only acoustic features leads to performances not significantly different from the baselines. However, the features of the facial expressions provide good performance scores with significant differences with the baselines (with RF, accuracy score $= 0.71$, F1 weighted score $= 0.74$ and p-value < 0.05). Considering combinations of groups of features, the result reveals that the combination of non-verbal and vocal cues improves significantly the accuracy score (with RF, accuracy score $= 0.74$, F1 weighted score $= 0.82$ and p-value < 0.005). These results are in line with the research on persuasion showing the importance of multimodality for perceived persuasion. Finally, the best accuracy score is obtained by combining facial expressions features, head movements and vocal features with a Random Forest classifier, we have an accuracy score $= 0.81$, a F1 weighted score $= 0.72$ and p-value < 0.0005.

In the following steps, since we obtain also good results with the combination of facial expressions features and head movements (accuracy score equals to 0.74 and F1 weighted score equals to 0.82), we focus on these non-verbal cues that we can be simulated on the embodied conversational agent Greta[3].

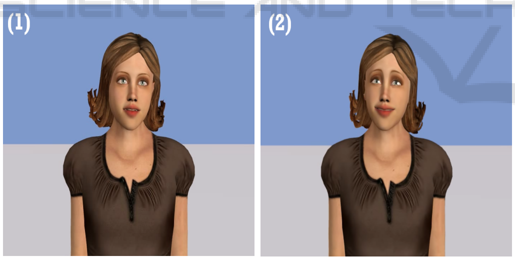

Figure 2: Screenshots at the same time of two animations given as input the same video of human: (1) neutral attitude generated without transformation and (2) persuasive attitude generated by the model of the THRUST tool.

5 CONVOLUTION-BASED MODEL

In this section, we describe the steps 4 and 5 illustrated on Figure 1. These steps consist in computing the non-verbal cues (facial expressions and head movements identified in the previous steps) of the virtual speaker to enhance its persuasiveness. The computation is based on the POM corpus. This corpus contains neutral and persuasive sequences. It is important to note that, in this study, we define a neutral attitude as the act of speaking without attempting to be persuasive. It does not mean that the neutral face is not persuasive.

For each relevant non-verbal feature, we generate a signal that describes its average dynamic across all persuasive or neutral sequences in the POM corpus. For this purpose, we treat each slice (Ambady and Rosenthal, 1992) in the POM corpus as a sample and we use the average value of each non-verbal behavior's dynamic as a reference. The generated signals characterize the typical values of the non-verbal cues associated to a persuasive attitude. Each signal is associated to a non-verbal cue. These signals correspond to references that the non-verbal cues have to follow when we generate the persuasive behavior. The reference values is noted $(\mathcal{M}_i)_{i=1,..,N}$, where $\mathcal{M}_i$ corresponds to the i-th reference of the non-verbal cues indexed with $i = 1,...,N$ (for example, AU1 and AU2 as facial units and head position according (x,y,z) axis). To generate persuasive non-verbal cues, we modify the $\mathcal{U}_i$ value using a convolution product between $\mathcal{U}_i$ and $\mathcal{M}_i$ which involves averaging the $\mathcal{U}_i$ input based on the properties of the reference $\mathcal{M}_i$. A re-sampling step may be necessary at this stage to avoid the issue of size mismatch between $\mathcal{U}_i$ and $\mathcal{M}_i$. It should be noted that our convolution-based strategy is only applied to the non-verbal cues that Greta takes into account, namely head movements along the (x,y,z) axis and specific AUs (AU1, AU2, AU4, AU5, AU6, AU7, AU12). In the next section, we present the implementation and the evaluation of the THRUST tool integrating the models presented above.

6 IMPLEMENTATION AND EVALUATION

The process outlined in Figure 1 has been successfully implemented and is now fully operational. As described in 3, we use OpenFace to extract the features from the input (the video of human with a neutral attitude). Then we developed our *Model* module (Step 1, 2, 3, 4, 5 and 6) to convert neutral human face and head movement into persuasive virtual ones using Python language. The output of our Python script is played with Greta tool. At this final step, we have used the virtual female character *Emma* to create the final output of the THRUST tool (the video of a persuasive Embodied Conversational agent). The entire code of the tool, as well as a tutorial video, is pro-

[3] An open-source platform to create Embodied Conversational Agent: https://github.com/isir/greta/wiki

vided in the GitHub account of the authors as an open-source project[4]. In the following sections, we present an objective and subjective evaluation of the tool.

6.1 Objective Evaluation

The videos generated by our tool show a noticeable difference in terms of facial and head movements between the videos of the virtual speaker without transformation (i.e. the non-verbal cues extracted from the human video are directly replicated on the virtual speaker; called *neutral videos*) and those transformed by the models (i.e. the non-verbal cues extracted from the human video are transformed by the models and replicated on the virtual speaker; called *persuasive videos*). The generated neutral videos appear to have very little movements, while the generated persuasive videos show more eyebrow movements, smiles and head movements. To assess the effectiveness of our tool, we propose an objective evaluation based on the learnt classifier. As described previously (Section 4.2), to construct the THRUST tool, we have developed an accurate classifier for predicting persuasion. We propose to use this classifier to verify that the generated videos are correctly classified as persuasive. For this purpose, we consider the best classifier that was the *Random Forest* one. To evaluate the model objectively, we have generated 24 videos. Four persons have been filmed (2 female and 2 male). We asked to each person to say two different predefined sentences in a neutral way. The sentences are related to our use case, e.g. “Through physical activity you can overcome disabling pain and improve your general well-being”. The two sentences have the same size. The speech production of these sentences lasts around 10 seconds. In total, we have recorded 8 videos of human speaker of 10 seconds. Using the THRUST tool, we have generated two kinds of video using the same virtual speaker (Figure 2):

- videos of a **neutral** virtual speaker that correspond to the replay of the recorded human features on the virtual face;
- videos of a **persuasive** virtual speaker that correspond to the recorded human features transformed by our model and replicate on the virtual speaker.

For each recorded human video, we have generated these two kinds of video. Moreover, we have created **baseline** videos. These videos correspond to the noisy version of the recorded human features on the virtual face (we used a classical additive white Gaussian noise). The baseline videos have been created in order to compare videos with the same amount of movements (persuasive and baseline videos) with the neutral videos with few movements. In total, we had 24 videos to evaluate[5].

The generated videos have been provided to the classifier to evaluate objectively if they were classified as persuasive. Note that only the non-verbal features of the video have been used for the classification. The results show that all the persuasive videos are classified as persuasive whereas the neutral videos and the baseline videos are classified as non persuasive. This first objective evaluation constitutes a first validation step of the proposed tool. However such an evaluation is not sufficient since the classifier and the THRUST tool have been created based on the same data. To complete this evaluation, we present in the next section a subjective evaluation to assess the generated videos with users.

6.2 Subjective Evaluation

The THRUST tool has been evaluative through a perceptive study. We describe the protocol of the experiment and the results of the evaluation in the following.
Videos. In this subjective evaluation, we consider the same 24 generated videos used for the objective evaluation (previous section): 8 videos corresponding to a persuasive virtual speaker, 8 to a neutral virtual speaker and 8 corresponding to baseline with randomly generated animations. As for the objective evaluation, we consider only the non-verbal cues. Consequently, the videos were played without sound to avoid the lip synchronization problem and the impact of the speech on the perception.
Questionnaire. We asked several participants to evaluate the perceived persuasiveness of the virtual speaker, both of the videos before and after the transformation of our model (i.e. the neutral and persuasive videos) and the baseline videos. For this purpose, we consider 2 questions: ***QI*** “Did you find the character persuasive in the video ?” and ***QII*** “Did you find the character animations convincing in the video ” (translation from French). The responses to each question were indicated through a Likert-scale from 1 to 5.
Participants. A total of 89 persons (51 female and 38 male) have participated online to the experiment. They were recruited on French mailing lists. The age of the participants is in average 34.2 ($SD = 11.78$).
Task. Each participant had the task to watch each video and to indicate their perception through the

[4] https://test.i2m.univ-amu.fr/perso/cherni.a/Software.html

[5] Examples of videos: https://www.youtube.com/playlist?list=PL6t9zd1YosSWFwosMBWoPXCPk-0TT0tUB

questions on the virtual speaker's persuasiveness (2 questions). The experiment took place online. The order of the video and of the questions have been counterbalanced to avoid an effect on the results.

Results. The scores of the participants have been analyzed using a Two-Way Repeated Measures ANOVA. We applied a normality test using kewness and Kurtosis. The distribution of each measure is normal. In Table 1, we report the descriptive statics of the results.

We report the results considering each question separately.

Concerning the first question ***QI*** "*Did you find the character persuasive in the video ?*", the results show a significant effect of the agent's behavior on the users' perception, $F(1.48, 89) = 12.66$, and $p < 0.001$. The persuasive virtual speaker ($M = 2.43$, $SD = .08$) has been perceived significantly more persuasive than the baseline virtual speaker ($M = 1.93$, $SD = .08$), $p < 0.001$, and the neutral virtual speaker ($M = 2.18$, $SD = .08$), $p = 0.001$. This result *validates the proposed approach showing that the videos generated by the THRUST tool are perceived significantly more persuasive* than the videos generated without transformation or with randomly generated non-verbal cues. Note that no main effect of users' gender has been found, although women ($M = 2.28, SD = 0.07$) had the tendency to perceive agents more persuasive than men ($M = 2.08, SD = 0.08$), with p-value equal to 0.08. Interestingly, the sentence used for the recording has a significant impact on the perceived persuasiveness $F(1, 89) = 17.41$, $p < 0.001$, whereas the videos had no sound. The videos recorded with the sentence 2 ($M = 2.25$, $SD = 0.06$) were perceived as more persuasive than the videos with the sentence 1 ($M = 2.11$, $SD = 0.06$).

Concerning the question ***QII*** *"Did you find the character animations convincing in the video?"*, the results are coherent with the results for the question ***QI***. In fact, the main effect of the agent's behavior is significant with $F(1.54, 89) = 13.35$ and $p < .001$. The persuasive virtual speaker ($M = 2.51, SD = .08$) has been perceived as more convincing than the baseline ($M = 2.02, SD = .09$) and $p < .001$, and the neutral virtual speaker ($M = 2.27, SD = .08$) and $p = .003$. No main effect of users' gender has been found, although women ($M = 2.37, SD = .07$) had the tendency to find agents more convincing than men ($M = 2.16, SD = .08$) and $p = .06$. The sentence used for the recording has a significant impact on the perceived convincing aspect $F(1, 89) = 29.38$ and $p < .001$. The videos recorded with the sentence 2 ($M = 2.34, SD = .06$) were perceived as more convincing than the videos recorded with the sentence 1 ($M = 2.18, SD = .05$). These results *confirm that the videos generated by the THRUST tool are perceived significantly more convincing* than the videos generated without transformation or with randomly generated non-verbal cues.

Discussion. The results of the perceptive study enable us to validate the videos generated by the THRUST tool. In fact, the animations of the virtual speaker are perceived significantly more persuasive and convincing after the transformation by the THRUST tool. The significant differences with the baseline videos with randomly generated animations show that it is not the fact that the persuasive videos have more facial and head movements compared to the neutral one but it is the animations generated by the tool that allow the perception of persuasiveness. Moreover, we have evaluated the tool considering different human videos as input, both female and male, showing, then, that the tool provides persuasive output whatever human is in the input video. A point of attention is the speech. Even if the videos were played without sound, surprisingly, it appears that what is said has a significant impact on the perceived persuasiveness. A more fine-grained analysis considering a larger set of sentences should be conducted to explain this result. The thrust tool has been evaluated considering only one specific female virtual agent. Videos generated with virtual agents with different appearances should be evaluated to completely assess the efficiency of the tool.

7 CONCLUSION AND PERSPECTIVES

The main goal of the presented work is to develop and validate a new tool that can transform a video of a neutral human face into a video of a virtual agent with a behavior expressing persuasiveness. For this purpose, we have proposed a tool called THRUST (from neuTral Human face to peRsUaSive virTual face) based on machine learning techniques and on a convolution-based model. The tool computes automatically the facial and head movements of a persuasive virtual speaker. The tool has been evaluated through an objective and subjective study. Both evaluations on a set of videos have enabled us to demonstrate that the virtual speaker's behaviors computed by the tool are perceived as persuasive. Even if the evaluation was limited to a specific set of videos, features and virtual agents, these results constitute a first validation of the proposed approach for the automatic generation of persuasive behavior.

Since the THRUST tool is based on the POM corpus, which is the only corpus with persuasion anno-

Table 1: Descriptive statistics on the results of the subjective evaluation. The statistics (M: Mean and STD: standard deviation) are reported for each question asked to the participants (QI and QII) and for each sentence said by the virtual character (Sentence 1 and 2), considering the gender of the participant (women and men) and the condition (baseline, neutral and persuasive - Section 6.1).

		Women						***Men***					
		Baseline		Neutral		Persuasive		Baseline		Neutral		Persuasive	
		M	STD	M	STD	M	STD	M	STD	M	STD	M	STD
Sentence 1	***QI***	1.86	0.06	2.32	0.13	2.37	0.06	1.85	0.07	2.33	0.08	1.98	0.06
	QII	1.94	0.56	2.54	0.14	2.40	0.1	1.97	0.08	2.39	0.15	1.91	0.02
Sentence 2	***QI***	2.26	0.03	2.10	0.03	2.76	0.09	1.79	0.06	2.02	0.14	2.62	0.02
	QII	2.33	0.02	2.20	0.04	2.91	0.13	1.86	0.18	2.00	0.20	2.85	0.03

tations, and was created using the open-source toolboxes Greta and OpenFace, which only use some nonverbal cues, there are certain limitations in the choice of features. We propose on the future work to expand our analysis and include other multimodal features, particularly vocal features, to enhance the persuasive model and to develop an automated artificial agent capable of expressing persuasive speech.

ACKNOWLEDGMENTS

This research was funded by the French National Research Agency as part of the COPAINS project.

REFERENCES

Ahuja, C., Lee, D. W., Ishii, R., and Morency, L.-P. (2020). No gestures left behind: Learning relationships between spoken language and freeform gestures. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 1884–1895.

Ambady, N. and Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological bulletin*, 111(2):256.

Burgoon, J. K., Birk, T., and Pfau, M. (1990). Nonverbal behaviors, persuasion, and credibility. *Human communication research*, 17(1):140–169.

Cassell, J. (2001). H. vilhjilmsson, and t. *BEAT: the Behavior Expression Animation Toolkit*.

Chidambaram, V., Chiang, Y.-H., and Mutlu, B. (2012). Designing persuasive robots: how robots might persuade people using vocal and nonverbal cues. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, pages 293–300.

Chiu, C.-C. and Marsella, S. (2014). Gesture generation with low-dimensional embeddings. In *AAMAS*.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923.

Ghazali, A. S., Ham, J., Barakova, E. I., and Markopoulos, P. (2018). Poker face influence: persuasive robot with minimal social cues triggers less psychological reactance. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 940–946. IEEE.

Habibie, I., Xu, W., Mehta, D., Liu, L., Seidel, H.-P., Pons-Moll, G., Elgharib, M., and Theobalt, C. (2021). Learning speech-driven 3d conversational gestures from video. In *IVA*.

Hasegawa, D., Kaneko, N., Shirakawa, S., Sakuta, H., and Sumi, K. (2018). Evaluation of speech-to-gesture generation using bi-directional lstm network. In *IVA*.

Kucherenko, T., Jonell, P., van Waveren, S., Henter, G. E., Alexandersson, S., Leite, I., and Kjellström, H. (2020). Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *ICMI*.

Lisetti, C., Amini, R., Yasavur, U., and Rishe, N. (2013). I can help you change! an empathic virtual agent delivers behavior change health interventions. *ACM Transactions on Management Information Systems (TMIS)*, 4(4):1–28.

Marsella, S., Xu, Y., Lhommet, M., Feng, A., Scherer, S., and Shapiro, A. (2013). Virtual character performance from speech. In *SIGGRAPH*, pages 25–35.

Miller, G. R. (2013). *On being persuaded: Some basic distinctions.* Sage Publications, Inc.

Miller, N., Maruyama, G., Beaber, R. J., and Valone, K. (1976). Speed of speech and persuasion. *Journal of personality and social psychology*, 34(4):615.

Nguyen, H., Masthoff, J., and Edwards, P. (2007). Persuasive effects of embodied conversational agent teams. In *International Conference on Human-Computer Interaction*, pages 176–185. Springer.

Nojavanasghari, B., Gopinath, D., Koushik, J., Baltrušaitis, T., and Morency, L.-P. (2016). Deep multimodal fusion for persuasiveness prediction. In *ICMI*.

Park, S., Shim, H. S., Chatterjee, M., Sagae, K., and Morency, L.-P. (2014). Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In *ICMI*.

Park, S., Shim, H. S., Chatterjee, M., Sagae, K., and Morency, L.-P. (2016). Multimodal analysis and prediction of persuasiveness in online social multimedia. *ACM TiiS*, 6(3):1–25.

Petukhova, V., Raju, M., and Bunt, H. (2017). Multimodal markers of persuasive speech: Designing a virtual debate coach. In *INTERSPEECH*, pages 142–146.

Tanner, R. and Chartrand, T. (2006). The convincing chameleon: The impact of mimicry on persuasion. *ACR North American Advances*.