From neutral human face to persuasive virtual face, a new automatic tool to generate a persuasive attitude

Afef Cherni, Roxane Bertrand, Magalie Ochs

From neutral human face to persuasive virtual face, a new automatic tool to generate a persuasive attitude

Afef Cherni[1,2], **Roxane Bertrand**[2], **Magalie Ochs**[1]

Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France

Aix Marseille Univ, CNRS, LPL, Aix en provence, France

Summary — In order to motivate the user to change her/his behavior or attitudes, for instance to practice physical activities to improve her/his well-being, virtual agents should have persuasive capabilities. The persuasiveness of the virtual agent not only depends on its speech but also on its non-verbal behavioral cues. In this paper, we propose the 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 from a video of a human. Combining a machine learning approach on a corpus of persuasive human speech and a convolution-based method, we propose a model, based on real data of persuasive human message, that transforms the non-verbal behavior of the human expressed in a video to a persuasive non-verbal behavior replicated on a virtual face.

Keywords— Multimodal cues, persuasion, embodied conversational agent, machine learning methods, convolution.

1 Introduction

A key challenge in intelligent virtual agent research concerns the automatic generation of the embodied conversational agent's behaviors, in particular related to social and emotional dimensions. In this article, we focus more particularly on the social skills of *persuasion*. Persuasion can be defined as "*any message that is intended to shape, reinforce or change the responses of another or others*" [1]. As highlighted in [2, 3], the persuasiveness of a message does not only depend on its content but crucially depends on all the multimodal components involving the different verbal, vocal and mimo-gestural levels (facial expressions, gestures, pitch). In this article, we particularly focus on the social signals expressed by non-verbal cues (facial expressions, head movements), that a virtual agent could express to be more persuasive. We do not consider the argumentative aspects (e.g. the identification of the arguments to convince, the order of the presentation of the arguments, the responses to the arguments of the persuadee). As a first step, we concentrate our study on *non-verbal cues* of persuasion.

The final goal of our project is to develop a *persuasive Embodied Conversational Agent* (ECA) to motivate elderly population to practice physical activity. In the intelligent virtual agent domain, several persuasive virtual agents have already been developed (e.g. [4, 5, 6]). The main approach to model persuasive behavior consists in identifying in the literature the behavioral cues that have an impact on persuasiveness and to integrate these cues in artificial agents. Indeed, the literature highlights some human's behavioral cues related to persuasion, as for instance the body movements [2] or the prosody [5]. In the domain of virtual agents, some empirical research works have shown the importance of certain verbal and non-verbal cues to improve the virtual agent's persuasiveness [7, 3]. However, as far as we know, no multimodal behavioral model to determine the verbal and non-verbal cues that a persuasive virtual agent should express during an interaction with a user, has been yet proposed.

In this article, we propose a new tool called *THRUST: from neuTral Human face to peRsUaSive virTual face*. This tool automatically transforms the video of a human to a video of a virtual character with a persuasive non-verbal behavior. More precisely, the tool extracts automatically the head movements and the facial expressions of the human, modifies them based on a computational model that we proposed, and replay the computed head and facial movements on a virtual face. The main contribution of the paper concerns the computational model that automatically modified the head and facial movements extracted from the human face to persuasive movements replicated on the virtual face.

One first challenge that has to be tackled to create such model is to identify more precisely the cues related to persuasion; i.e. the cues that we will modify to simulate persuasive behavior. For this purpose, in a first step, we propose to explore the relevant behavioral cues of persuasion using machine learning methods applied on a corpus of human videos. We consider the POM corpus [8] which is the only multimedia corpus created with annotations for studying persuasiveness to our knowledge. It contains web videos of speakers talking about different subjects in front of a camera.

Based on machine learning classifiers, we have explored the behavioral cues of persuasion. In this work, special attention has been paid to create explainable models with interpretable features (i.e. features that we can understand contrary to raw data). Indeed, our objective in this first step is not to create a classification model to recognize persuasiveness, but to identify the relevant features that we have to consider to create persuasive behavior on a virtual character.

The second step to create the tool is to transform the relevant signals extracted from the human face to persuasive one. Since the POM corpus contains the real measurements of different non-verbal cues of human expression with neutral and persuasive attitude, we create a dictionary from these real data to define the references reflecting persuasive non-verbal behavior. Based on this dictionary, a convolution-based method has been applied. Note that in this work, we define the neutral attitude as to speak without being persuasive or without making the effort to be so.

This article is organized as follows. In the next section, we present related works, i.e. the theoretical and empirical research works exploring the behavioral cues related to persuasion. We then introduce an overview of our proposed architecture to create persuasive behavior in Section 3. We detail the machine learning framework in Section 4. Then, Section 5 is dedicated to the convolution-based model that we propose

based on the POM corpus. We present the implementation of our new automatic tool to generate a persuasive attitude in Section 6. We conclude in section 7 by presenting the limits of the study and by discussing future works.

2 Related work

The study of the persuasiveness of specific behavioral cues has been the main interest of various works, specially in the context of human-human interaction. The research works show the importance of several multimodal behavioral [2, 5, 9]. In this article, we focus on the non-verbal cues of persuasiveness. [2] proved that gestural and facial activity (e.g. gestures, body movements, facial expressions and smiles) improve the persuasion. At the interactional level, several works studied the positive impact of mimicry on persuasion [10]. 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 [2], 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 [2], 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. [11, 12]). 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. [13]), facial expressions and body movements from speech (e.g. [14]) or from speech and text to take into account both acoustic and semantic information (e.g. [15, 16]). Most studies are based on deep neural networks (e.g. [13, 17, 16]) and, more recently, on the use of GAN architectures (e.g. [15, 14]). Our presented work differs from the existing models on different aspects: (1) we generate non-verbal behavior, not from speech or text, but from a video of a human with a neutral attitude; (2) we generate the facial and head movements whereas most of the existing models consider the body and head movements and (3) we explore the automatic generation of *persuasive* behavior whereas a limitation of the existing work is that the proposed models do not allow for the generation of different social-emotional behaviours.

From a *machine leaning perspective*, few research works have investigated persuasion. The main work has been conducted by Park *et al.* [8, 18, 19] 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.* [8, 18, 19], we use machine learning algorithms to explore persuasiveness. However, our work differs from the latter in several aspects:

- 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,
- still in our perspective of interpretability, we consider non-verbal features that can be simulated on a virtual character[1],
- last but not least, our final objective is not to create a prediction model but to explore the non-verbal cues and use machine learning-based methods in order to create a persuasive artificial agent.

3 General architecture

In this section, we present the general architecture of our system illustrated on the Figure 1.

Input. The system takes as input a video of a speaker talking about a specific topic in a neutral way. The input is not limited to real-time data video, it can be webcam video, recorded video files or sequences of images. The important aspect is to be able to extract the facial landmarks, the head poses, the eye gaze and the facial Action Unit (AUs) from the video. For this purpose, we use the OpenFace tool[2]. We note these N measures as $(\mathcal{U}_i)_{i=1...N}$ where we index them by i from 1 to N and we design by each vector $\mathcal{U}$ the measured feature (for example AU1, AU2, AU12, head position according to x, y or $z-$axis, ...) and i its index in the variables set. In other words, the vector $(\mathcal{U}_i)_{i=1...N}$ represents the values of the features characterizing the head and face movements extracted from the video of the human. These variables set $(\mathcal{U}_i)_{i=1...N}$ represents the input of the "Model" box.

Model: As illustrated in Figure 1, the proposed model takes as input the set of features $(\mathcal{U}_i)_{i=1...N}$ characterizing the face and head movements of a neutral human speaker and produces as output a set of features $(\mathcal{W}_i)_{i=1...N}$ characterizing the head and face movements of a persuasive speaker. To compute how to modify the set of features to be persuasive, we combined a machine learning approach and a convolution-based method. The first steps consist in using machine learning methods on an existing corpus to identify the important relevant features to consider to simulate persuasiveness (Step 1, 2, and 3, Figure 1). These steps are detailed in Section 4. Note that in these steps, we learn a classification model to automatically determine if the face and head movements are persuasive or not. This classification model is also used to verify that the transformed vector of characteristics $(\mathcal{W}_i)_{i=1...N}$ is indeed considered as persuasive (illustrated on the Figure 1 by the dotted arrow from the "output" box to the "classification" box). The second steps consist in modeling a convolution-based method, from the data of the corpus, in order to determine how to modify the features to be persuasive (Step 4 and 5 , Figure 1). These steps are detailed in Section 5.

Output. As output, the system produces a video of a persuasive virtual character. The virtual character replicates the same speech but with persuasive face and head movements. For this purpose, the "Output" box simulates the variables set $(\mathcal{W}_i)_{i=1...N}$ on the embodied conversational agent Greta, and generates the target video. The vector $(\mathcal{W}_i)_{i=1...N}$ represents the value of the head and face movements extracted from the video of the human and transformed to be persuasive.

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

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

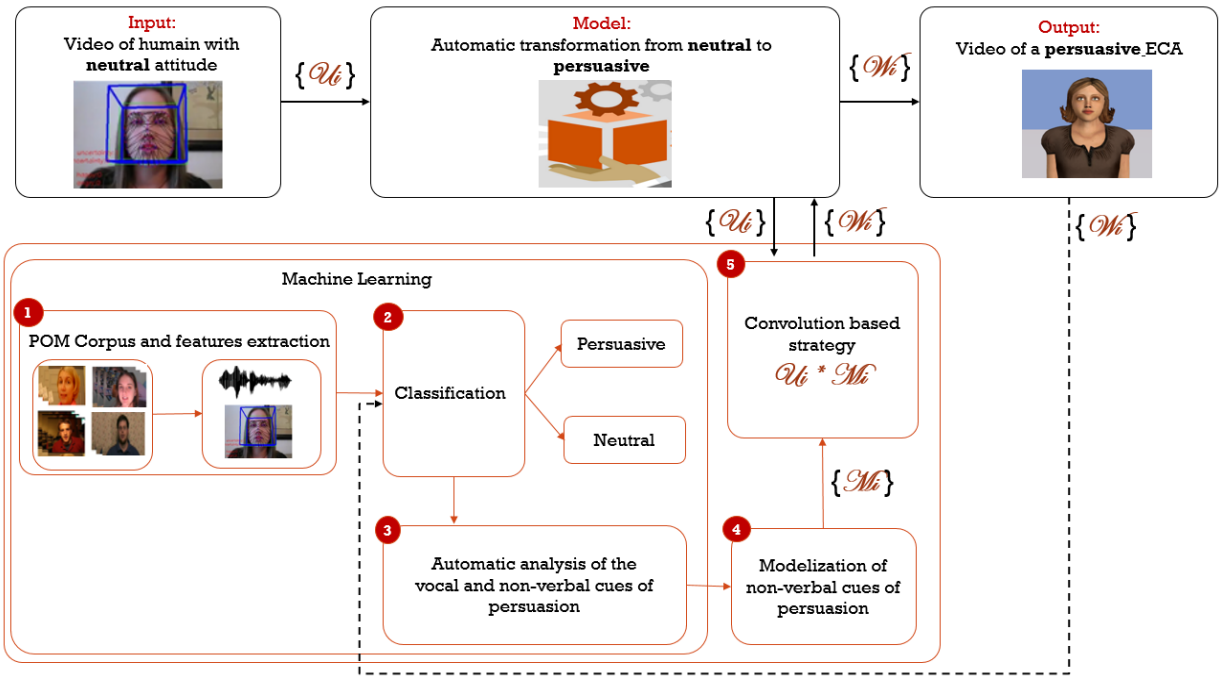

Figure 1: Global architecture of the system to automatically transform a neutral human video to a persuasive virtual character video. *Input*: a video of a human with automatic extraction of head and facial movements using OpenFace. *Model*: a computational model to automatically transform neutral non-verbal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

4 Machine learning framework

In this section, we describe the steps 1, 2 and 3 illustrated on the Figure 1.

4.1 Corpus and features extraction

In the step 1 (Figure 1), we consider a specific corpus and extract the features form 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 [8]. 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. The research works conducted by Park *et al.* on this corpus [8, 19], show that behavioral cues can be used to predict *extreme value* of persuasion. This work has been conducted to automatically classify the persuasiveness of a human speech. In this article, we aim at exploring the corpus to identify the relevant features of persuasiveness, i.e. the non-verbal cues that enable a human to be perceived as persuasive. Then, we propose to determine the set of relevant features to consider to simulate persuasive message. For this purpose, we explore different sets of features and their impact on the performances of classifiers to evaluate their importance.

Based on the theoretical and empirical research on persuasion presented above (Section 2), we consider the following groups of features:

- *Group 1: facial expressions*: facial action units (AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU12, AU14, AU15, AU17, AU18, AU20, AU23, AU24, AU25, AU16, AU28, AU43);
- *Group 2: emotions*: Anger, Contempt, Disgust, Joy, Fear, Surprise, Confusion, Frustration. The emotions are computed based on vocal and non-verbal cues [18].
- *Group 3: head movements*: head position (displacement and rotation in (x, y, z) axes, speed of the head movement and its acceleration according to (x, y, z) axis.
- *Group 4: acoustic descriptors*: fundamental frequency f_0, peak slope.

In order to explore the importance of these features, we propose to compute statistical functions for each feature: the mean, median, maximum, minimum, standard deviation and the variance. Note that even if we do not consider the acoustic features for the simulation on the virtual character, at this step, we consider relevant to evaluate the importance of these acoustic features in comparison to non-verbal ones.

4.2 Formalization of the classification problem

Our objective by using machine learning methods is to investigate the behavioral cues related to persuasion. 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.

Different classification methods could be considered (e.g. binary classification, multi-class classification, or regression). As a first step, we consider a binary classification to simplify the learning problem (i.e. prediction if persuasive or not). The accuracy of the prediction may depend on the chosen definition of the output classes. Indeed, we can choose to predict only extreme values (as proposed in [8]) or to split in two balanced classes without excluding middle values. In order to compare the different approaches, we propose to explore 2 strategies to define the clustering of the two classes:

Strategy 1: This strategy considers the extreme values of the annotated persuasion to create two classes as proposed in [8]: one class clusters the values equal to or greater than 5.5 and the other class clusters the values equal to or less than 2.5 (for values ranging from 1 to 7);
Strategy 2: With the above *Strategy 1*, the classes are imbalanced. Therefore, we propose in *Strategy 2* to explore over-sampling methods to increase the amount of data and to obtain balanced classes. The over-sampling methods generate new samples of the minority class based on the existing dataset, in order to remove class imbalance. For this purpose, we propose to use SMOTE (Synthetic Minority Over-sampling Technique) algorithm [20].

4.3 Automatic analysis of the vocal and non-verbal cues of persuasion

In the step 3 (Figure 1), the objective is to test the clustering strategies proposed in Section 4.2 to compare the performances of the classifiers and then to select the most important features that ensure the highest prediction performances. We consider at this step both vocal and non-verbal cues.

Classifiers: We propose to experiment different classifiers: the *Naives Bayes* (NB), the *System Vector Machine* (SVM) and the *Random Forest* (RF). These methods, among the best classifiers [21], have the advantage, compared with other statistical models such as RNN, to handle high-dimensional data with a high generalization power [22, 23]. They are also well suited for handling small datasets. All experiments were performed with 10-fold cross-validation (CV) where each CV was tested 10 times.
Baselines: 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.
Prediction model: Each classifier will be fitted on the training set (80%) and testing set on 20% of the corpus. This experiment was performed with 10-fold cross-validation (CV) where each CV was tested 10 times. The performances of the classifiers are evaluated though the classical metrics of accuracy and F1 weighted score (to cope with the unbalanced classes for the *Strategy 1*). Table 1 summarizes the performances of the different classifiers. We moreover 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 [24].
Prediction results: In order to evaluate the importance of each group of features (facial expressions, emotions, head movements and acoustic descriptors) to predict the persuasion, we compute the performance scores of the classifiers considering each group of features as input and combinations of several groups of features. Table 1 summarizes the performances of the best classifiers and the significant differences with the baselines considering the different groups of features as input (the scores significantly different from the average scores of the 3 baselines are presented with gray cells on the Table).

Considered features	Classifiers	***Strategy 1***		***Strategy 2***	
		Accuracy score	F1 weighted score	Accuracy score	F1 weighted score
Group 1	SVM	0.64	0.66	0.64	0.73
	RF	**0.71**	**0.74**	0.67	0.67
	NB	0.66	0.63	0.66	0.64
Group 2	SVM	0.54	0.71	0.54	0.96
	RF	0.55	0.58	0.55	0.53
	NB	0.54	0.55	0.54	0.56
Group 3	SVM	0.63	0.65	0.63	0.62
	RF	0.72	0.074	0.72	0.71
	NB	0.57	0.55	0.57	0.45
Group 4	SVM	0.61	0.65	0.61	0.65
	RF	0.69	0.51	0.69	0.51
	NB	0.72	0.62	0.72	0.62
Group 1 + 3	SVM	0.64	0.69	0.64	0.67
	RF	0.69	0.83	**0.74**	**0.82**
	NB	0.64	0.55	0.64	0.50
Group 1 + 4	SVM	0.45	0.55	0.45	0.56
	RF	0.52	0.70	0.54	0.68
	NB	0.63	0.68	0.63	0.68
Group 3 + 4	SVM	0.71	0.72	0.71	0.71
	RF	0.74	0.76	0.74	0.78
	NB	0.54	0.57	0.59	0.65
Group 1 + 3 + 4	SVM	0.65	0.73	0.65	0.73
	RF	0.76	0.74	**0.81**	**0.72**
	NB	0.64	0.56	0.64	0.56
Group 1 + 2 + 3 + 4	SVM	0.63	0.74	0.63	0.74
	RF	0.76	0.76	0.77	0.68
	NB	0.64	0.53	0.64	0.56

Table 1: Performance scores of different classifiers (*Support Vector Machine* (SVM), *Random Forest* (RF), *Naive Bayes* (NB)) using two strategies. We design by ***Group 1*** : facial expressions, ***Group 2*** : emotions, ***Group 3*** : head movements, ***Group 4*** : acoustic descriptors. The highest scores are written in bold and the scores significantly different from the average scores of the 3 baselines are presented with gray cells.

The results (Table 1) show that the emotions (*Group 2*) 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 (*Group 3*) or only acoustic features (*Group 4*) lead to performances not significantly different from the baselines. However, the facial expressions (features of the *Group 1*) provide good performance scores with significant differences with the baselines.

Considering combinations of groups of features, the result reveals than the combination of non-verbal and vocal cues improves significantly the accuracy score. 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.

4.4 Modelization of the non-verbal cues of persuasion

In the previous section, we have considered both verbal and non-verbal features. The results show that the set of features with the facial expressions (AUs), the head movements and the vocal features ensures efficiently the persuasion prediction with

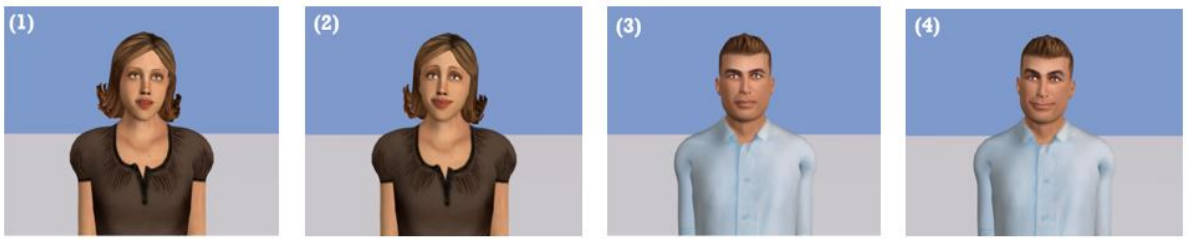

Figure 2: Screenshots at the same time of two studied cases: (1) and (2) played with the virtual female character "Emma" of Greta, (3) and (4) played with the virtual male character "John" of Greta. (1) and (3) present the input data (neutral attitude), (2) and (4) present the output of our model (persuasive attitude).

the highest accuracy score (0.81). However, we obtain similar results considering only facial expressions features (*Group 1*) (accuracy score equal to 0.71 and F1 weighted score equal to 0.74) or the combination of facial expressions features and head movement (*Group 1 + 3*) (accuracy score equal to 0.74 and F1 weighted score equal to 0.82) (Table 1)).

In this step, we focus on the non-verbal cues that can be simulated on the embodied conversational agent Greta: AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU12, AU14, AU15, AU20, AU23, AU25, AU26, AU45. Note that we do not consider the AUs related to lips movements. We plan to use a specific tool to ensure the lips synchronisation on the speech because the AUs are not sufficient to synchronize speech and lips movements. Moreover, since the POM corpus is based on different videos in which the speakers give their point of views or feelings about a particular subject, we propose to avoid the disgust expressions (described in our case by the facial expression AU9 and AU10). Consequently, in order to ensure the transformation from neutral to persuasive non-verbal movements, as a first step, we consider only face and head movements by modifying the following features AU1, AU2, AU4, AU5, AU6, AU7, AU12 (in *Group 1*) and head movement according to $(x, y, z)-$axis (in *Group 3*). Note that we have run our prediction model presented in Section 4.3 with these considered features and we have obtain a good performance scores with an accuracy score equal to 0.63 and F1 weighted score equal to 0.73.

5 Convolution-based model

In the steps 4 and 5 (Figure 1), we define how to transform the non-verbal features (considered as important for persuasion in the previous steps) to increase the persuasiveness of a virtual speaker. These steps are based on the data of the POM corpus. Indeed, for each important non-verbal selected feature, we propose to generate a signal that describes on average its dynamic according to all the sequences of POM corpus classified as persuasive or neutral. More precisely, since each video in POM corpus is annotated with respect to thin slice method [25], we consider each slice as a sample and the average value of each non-verbal behavior dynamic as a reference. After considering all the slices, we obtain a signal that describes the average values that non-verbal cue takes with respect to a persuasive attitude. This makes a reference for each non-verbal cue (i.e. feature) to follow in order to build a persuasive attitude. These references will be noted $(\mathcal{M}_i)_{i=1,..,N}$. We remind that N is the number of considered features, i its index and $\mathcal{M}$ presents the reference of each feature (expression facial units (AU1, AU2, ...), head movements).

Concerning the convolution-based strategy, at this step, we have two types of data: (1) the features set of the neutral human video noted $\mathcal{U}_i$, (2) the reference of each non-verbal cues $\mathcal{M}_i$ build in the previous step based on the POM corpus. In order to adapt the dynamic of the values $\mathcal{U}_i$ and make it persuasive, we apply a convolution product between $\mathcal{U}_i$ and $\mathcal{M}_i$. This step can be considered as an average filtering where the $\mathcal{U}_i$ input follows the specific properties of the $\mathcal{M}_i$ function. Since the size of each $\mathcal{M}_i$ depends on the used corpus, an inadequacy with the input size $\mathcal{U}_i$ may occur. To avoid this problem, a re-sampling treatment of each $\mathcal{M}_i$ according to the size of its corresponding input $\mathcal{U}_i$ is highly recommended at this step. Moreover, we propose to apply the convolution product to the i-th variables ($\mathcal{U}_i$ and $\mathcal{M}_i$) with respect to windows with size w to keep the same level of reference evolution and avoid the outliers. The results are noted $\mathcal{W}_i$. We remind that our convolution-based strategy proposed in this paper is applied only on the non-verbal cues that Greta takes into consideration (head movement according $(x, y, z)-$axis, and specific AUs).

6 Implementation and Evaluation

The entire process of our proposed tool illustrated in Figure 1 has been implemented. The tool is called THRUST: from neuTral Human face to peRsUaSive virTual face. The entire code of the tool was provided in open source on GitHub[3]. We present examples of outputs of the system in Figure 2 considering two different embodied conversational agent of the Greta platform: one female and one male. The resulting videos show that the proposed tool can be used on virtual characters of various appearances. In the Figure 2, we compare (1) the videos replicating directly the features extracted from the human video on an ECA, i.e. without any transformation and (2) the videos after the transformation of the model to create persuasive non-verbal behavior. The videos are available in THRUST channel[4].

The resulting videos show a significant difference of the face and head movements of the two videos. While the movements in the video without transformation is quite stable, in the video after the transformation, we can notice eyebrow movements and smiles. In order to evaluate these results, we first propose an objective method based on the persuasion classifier. Indeed, since we have build an efficient classifier that ensures the persuasion prediction (Section 4.2, Table 1), we propose to use it to test if the output of the model is correctly classified. We use the best identified classifier (*Random Forest*). To evaluate the model objectively, we have generated 50 videos (25 recorded human faces, replayed on virtual characters of the same gender and transformed to 25 videos of persuasive virtual faces). These videos correspond to 5 different speech, each speech produced by 5 different participants (3 female and 2 male) and lasts

[3]https://github.com/CherniAfef/THRUST-Tool

[4]https://www.youtube.com/channel/UC87g8UeHbMJync8n8DjLe8g/videos

around 10 seconds. In total, we have generated 25 videos of *neutral* virtual faces (replay of the recorded human features on a virtual face) and 25 videos of *persuasive* virtual faces (output of the model). Using the classifier, the results show that all the videos transformed by our model are classified as persuasive whereas those before the transformation are classified as neutral. This first objective evaluation constitutes a first validation step of the proposed tool.

The next step is to conduct a subjective evaluation of the generated videos. In a perceptive studies, we plan to ask participants to evaluate the believability and naturalness of the generated behavior (as proposed in [26]) but also the perceived persuasiveness of the virtual characters: both from the video before and after transformation. Different questionnaires would be considered to evaluate the level of persuasiveness such as Godspeed questionnaire [27] with some modifications in order to adapt it to persuasiveness context. Note that the video will be played without sound to avoid the lip synchronization problem and the effects of the speech on the perception.

7 Conclusion and perspectives

In this article, we have proposed a new tool, called THRUST (from neuTral Human face to peRsUaSive virTual face), to automatically transform a video of a neutral human face to a video of a virtual character face expressing persuasive head and face movements. To create such a tool, we have based our work on real videos of human with different levels of persuasiveness. Combining a machine learning approach and a convolution-based model, the proposed tool modify automatically the relevant features of the face expressions and head movements to increase the persuasiveness of the behavior. The objective evaluation of the resulting video shows that the video generated by the tool as in fact automatically classified as persuasive.

The presented work present some limits. We have limited the considered features according to the available corpus (POM corpus, the only corpus that contains persuasion annotations) and given the freely accessible toolbox (Greta and OpenFace used in our work). In a second step, we aim at extending the study to other multimodal features and in particular to vocal ones in order to improve the persuasive model and build an automatic artificial agent able to speak and express itself persuasively.

References

[1] G. R. Miller, *On being persuaded: Some basic distinctions.* Sage Publications, Inc, 2013.

[2] J. K. Burgoon, T. Birk, and M. Pfau, "Nonverbal behaviors, persuasion, and credibility," *Human communication research*, vol. 17, no. 1, pp. 140–169, 1990.

[3] V. Chidambaram, Y.-H. Chiang, and B. Mutlu, "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*, 2012, pp. 293–300.

[4] C. Lisetti, R. Amini, U. Yasavur, and N. Rishe, "I can help you change! an empathic virtual agent delivers behavior change health interventions," *ACM Transactions on Management Information Systems (TMIS)*, vol. 4, no. 4, pp. 1–28, 2013.

[5] V. Petukhova, M. Raju, and H. Bunt, "Multimodal markers of persuasive speech: Designing a virtual debate coach." in *INTERSPEECH*, 2017, pp. 142–146.

[6] H. Nguyen, J. Masthoff, and P. Edwards, "Persuasive effects of embodied conversational agent teams," in *International Conference on Human-Computer Interaction*. Springer, 2007, pp. 176–185.

[7] A. S. Ghazali, J. Ham, E. I. Barakova, and P. Markopoulos, "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)*. IEEE, 2018, pp. 940–946.

[8] S. Park, H. S. Shim, M. Chatterjee, K. Sagae, and L.-P. Morency, "Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach," in *Proceedings of the 16th International Conference on Multimodal Interaction*, 2014, pp. 50–57.

[9] N. Miller, G. Maruyama, R. J. Beaber, and K. Valone, "Speed of speech and persuasion." *Journal of personality and social psychology*, vol. 34, no. 4, p. 615, 1976.

[10] R. Tanner and T. Chartrand, "The convincing chameleon: The impact of mimicry on persuasion," *ACR North American Advances*, 2006.

[11] J. Cassell, "H. vilhjilmsson, and t," *Bickmore. BEAT: the Behavior Expression Animation Toolkit. In Proceedings of SIGGRAPH-Oi*, p. 477486, 2001.

[12] S. Marsella, Y. Xu, M. Lhommet, A. Feng, S. Scherer, and A. Shapiro, "Virtual character performance from speech," in *Proceedings of the 12th ACM SIGGRAPH/Eurographics symposium on computer animation*, 2013, pp. 25–35.

[13] C.-C. Chiu and S. Marsella, "Gesture generation with low-dimensional embeddings," in *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, 2014, pp. 781–788.

[14] I. Habibie, W. Xu, D. Mehta, L. Liu, H.-P. Seidel, G. Pons-Moll, M. Elgharib, and C. Theobalt, "Learning speech-driven 3d conversational gestures from video," in *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, 2021, pp. 101–108.

[15] C. Ahuja, D. W. Lee, R. Ishii, and L.-P. Morency, "No gestures left behind: Learning relationships between spoken language and freeform gestures," in *Findings of the association for computational linguistics: EMNLP 2020*, 2020, pp. 1884–1895.

[16] T. Kucherenko, P. Jonell, S. van Waveren, G. E. Henter, S. Alexandersson, I. Leite, and H. Kjellström, "Gesticulator: A framework for semantically-aware speech-driven gesture generation," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, 2020, pp. 242–250.

[17] D. Hasegawa, N. Kaneko, S. Shirakawa, H. Sakuta, and K. Sumi, "Evaluation of speech-to-gesture generation using bi-directional lstm network," in *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, 2018, pp. 79–86.

[18] S. Park, H. S. Shim, M. Chatterjee, K. Sagae, and L.-P. Morency, "Multimodal analysis and prediction of persuasiveness in online social multimedia," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 6, no. 3, pp. 1–25, 2016.

[19] B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrušaitis, and L.-P. Morency, "Deep multimodal fusion for persuasiveness prediction," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016, pp. 284–288.

[20] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[21] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3133–3181, 2014.

[22] G. Forman and I. Cohen, "Learning from little: Comparison of classifiers given little training," in *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, 2004, pp. 161–172.

[23] C. Salperwyck, V. Lemaire, and D. U. d. P. de Bois, "Impact de la taille de l'ensemble d'apprentissage: une étude empirique," *Confrence Internationale Francophone sur l'Extraction et la Gestion de Connaissance*, 2011.

[24] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural computation*, vol. 10, no. 7, pp. 1895–1923, 1998.

[25] N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis." *Psychological bulletin*, vol. 111, no. 2, p. 256, 1992.

[26] T. Kucherenko, P. Jonell, Y. Yoon, P. Wolfert, and G. E. Henter, "A large, crowdsourced evaluation of gesture generation systems on common data: The genea challenge 2020," in *26th international conference on intelligent user interfaces*, 2021, pp. 11–21.

[27] C. Bartneck, E. Croft, and D. Kulic, "Measuring the anthropomorphism, animacy, likeability, perceived intelligence and perceived safety of robots," 2008.