numediart
UMONS RESEARCH INSTITUTE FOR CREATIVE TECHNOLOGIES

inspire
Interaction Sensing and Perception

End-to-End Listening Agent for Audio-Visual Emotional and Naturalistic Interactions

Kevin El Haddad[1], Yelin Kim[2], Hüseyin Çakmak[1], Payton Lin[3], Jaebok Kim[4], Minha Lee[5], and Yong Zhao[6]

[1]TCTS Lab - numediart institute - University of Mons, Belgium
[2]Inspire Lab - University at Albany, State University of New York, USA
[3]Research Center for Information Technology Innovation, Academia Sinica, Taiwan
[4]Human Media Interaction group, University of Twente, Enschede, Netherlands
[5]Human-Technology Interaction group - Technical University of Eindhoven, Eindhoven, Netherlands
[6]VUB-NPU Joint AVSP Research Lab - Vrije Universiteit Brussel, Belgium & Northwestern Polytechnical University, China

Abstract

In this project, we aim at building a listening agent that would react with a naturalistic and human-like behavior and using nonverbal expressions to a user. The agent's behavior will be modeled by and built on three main components: recognizing and synthesizing emotional and non-verbal expressions, and predicting the next expression to synthesize based on the currently recognized expressions. Its behavior will be rendered on a previously developed avatar which will also be improved during this workshop. At the end we should obtain functioning and efficient modules which ideally should work in real-time.

numediart
UMONS RESEARCH INSTITUTE FOR CREATIVE TECHNOLOGIES

inspire
Interaction Sensing and Perception

Objectives

The goal of this project is to build a listening agent that would react to a user using mainly nonverbal expressions. Our ultimate goal is for such a virtual agent to recognize and take into account all nonverbal expressions to then express a convenient and naturalistic feedback. Ideally, the system would run in real time. One of the main challenges of this project is to tease apart the effect of verbal and semantic content in speech. We will rather focus on the speaker's nonverbal and paralinguistic behaviors to predict and generate the agent's nonverbal behavior in real-time.

Figure 1: Human-like avatar on which the synthesis output will be rendered

Fig. 1 shows the overall workflow of our project which is a basic and simple pipeline. Our agent will be built on recognition, prediction and synthesis modules. The recognition will detect/recognize relevant expressions, from which the prediction system will take a decision on what should be the agent's reaction. This reaction will be generated by the synthesis module. This latter's output will be rendered on a human-like avatar(fig. 1). Of course, ideally this whole system should be working in real time. So each of the three previously mentioned modules should be implemented with this idea in mind.

In parallel to the development of these modules, we plan on using the avatar, for the first time, in a data collection experiment. The goal of this experiment is explained at the end of the following section.

Background & Technical Description

The project we propose is inspired from and will be built on some of our previous work. Our work in [1], initiated the creation of some of the modules previously described. Indeed, in that work, an audiovisual (AV) concatenative synthesis system and a prediction system are presented. Both were built with the purpose of creating the listening agent we talked about previously. The prediction system is a CRF that takes as input a sequence of labels from a speaker and predicts the most suitable sequence of expressions for the agent. The synthesis system generates AV smiles and laughs predicted by the CRF.

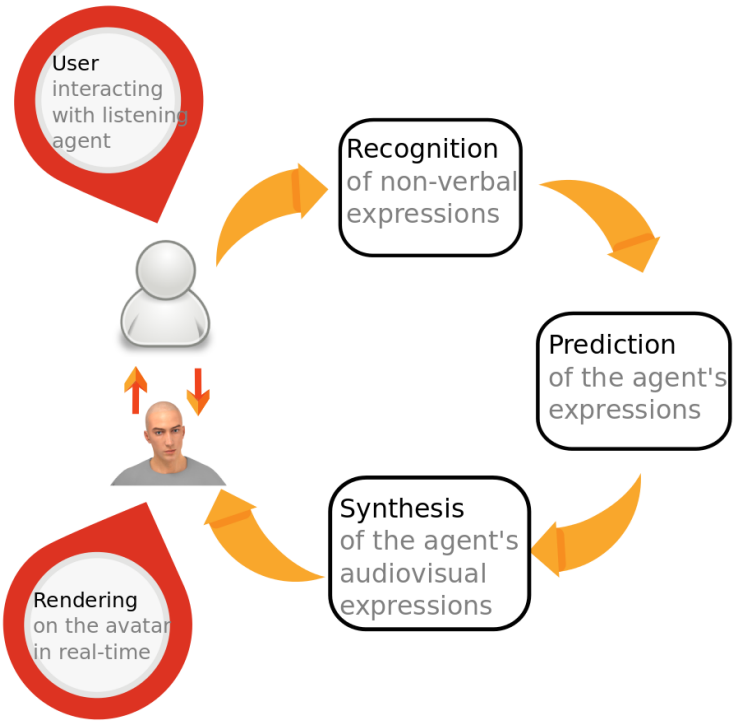

Figure 2: Overall project workflow

Based on the work described in [1], this project will aim at building a full listening agent by first adding a recognition module that would feed the prediction system. We will consider both audio and visual signals for the recognition of emotions using machine learning techniques, such as temporal [2] or deep learning [3] models. Such techniques performances depend on the amount of data available. Such data can be challenging to collect, specifically in the case of nonverbal and emotional expressions. We will attempt to increase the amount of data by utilizing a recently proposed procedure that relies on Deep Learning to synthesize new data sets [4].

The prediction system can be improved by, first, comparing other sequential prediction systems to the CRF we presented in [1] (such as RNN networks, HMM, different CRF configurations, etc...). Also by taking more expressions into account at the input and output of the system.

The synthesis system can be improved from the one in [1] by, first, adding more expressions to be synthesized than just laughs and smiles. Second, by increasing the size of the synthesis dataset. Other systems will also be tested such as the HMM-based parametric synthesis system in [5]. We expect the concatenative approach to give more naturalistic results but a parametric system

would allow a better control over the duration and the intensity level of the expression synthesized.

The work to obtain a real-time working agent has already been initiated in our eNTERFACE'15 project [6]. Indeed a real time controllable avatar has been built and will be used in this project too. The avatar can be controlled by OSC signals, can generate synchronous AV expressions and can have the direction of the face controlled [7]. At the end we should be able to evaluate each of the developed modules separately and also evaluate the system working as a whole. Adequate evaluation experiments will be put in place for this purpose.

Data

The recognition, prediction and synthesis modules defined previously are data driven. In this project, we intend to use available databases to train these modules. We also plan on running a data collection experiment. In this section we cite the databases we intend to use and detail the data collection experiments.

Data Bases
The following list of databases will be considered for each of the recognition, prediction and synthesis modules:

- IEMOCAP [8]: recognition module
- RECOLA [9]:recognition module
- CCDB [10]: recognition and prediction module
- IVADF [11]: recognition and prediction module
- AVLASYN [12]: synthesis module
- Cohn-Kanada(CK,CK+) [13, 14]: synthesis module

Data Collection for Real-life Application
Many emotion recognition and prediction modules are novel in regard to technical solutions, yet appropriate use cases must be considered in real-life contexts. One area of application is in ethical expertise development. Humans are bounded not only by rationality, but also by ethicality [15]. Learning to be ethically competent is relevant for many, if not all, professions [16]. To consider virtual agents as possible partners in ethical expertise development for humans, what is a priority is to study how humans respond to virtual agents on sensitive topics on practical ethics in everyday life. One study suggests that humans may have an easier time disclosing personal information to virtual agents compared to other people, which goes beyond merely economic reasons for using virtual agents [17]. In parallel to the developed modules we intend to take advantage of the eNTERFACE workshop to collect data with the goal to later explore how participants interact with virtual agents that appear to have capabilities

to understand them by giving them nonverbal AV feedback. This will be done through a setup where users will interact with a nonverbal expressive agent. Our collected data will then inform how virtual agents may be adjusted to their potential human counterparts, applied to a possible use-case of ethical expertise training.

Our data collection will proceed as two experiments. In both cases, participants will be asked to reflect on past experiences that induced what is categorized as moral emotions, both positive and negative. While there are a variety of emotions that are morally relevant, we will focus on negatively valenced emotions of shame, embarrassment, and guilt, as well as positively valenced emotions of compassion, gratitude, and elevation [18]. First we will perform a smaller experiment with human-to-human interaction on how people talk about prior experiences with aforementioned six moral emotions. In these interactions, one person will be an active listener and the other will be an active speaker, without further directions on how they should respond to each other. This is done to observe how people naturally react to these emotional experiences. Then, listeners' behaviors will be generalized for virtual agents to mimic, though wholly naturalistic responses by virtual agents will not be an expectation to be fulfilled at eNTERFACE. The second experiment with an updated virtual agent will ask participants to discuss their experiences of six moral emotions. Both experiments are exploratory.

Due to the ethically sensitive topic of the experiments, prior experiment proposals according to the standards of the Code of Ethics of the NIP (Nederlands Instituut voor Psychologen – Dutch Institute for Psychologists) will be completed and submitted at the Technical University of Eindhoven. The experimenters will abide by the recommended standards of the ethics committee.

Work plan

The project will be divided into 6 work packages (WP), 2 of which will be considered only depending on the amount of time left and the number of participants recruited (WP5 and WP6). The members of the team will work on these WP in parallel during the one-month period of the workshop and as shown in fig 3.

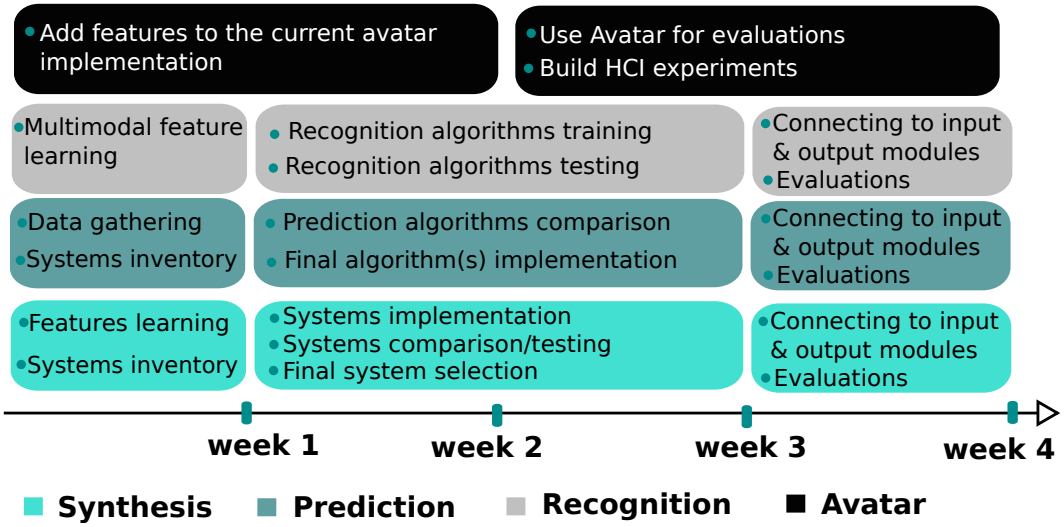

Figure 3: Planned task schedule

WP1: Oh! I know what you mean!
(Recognition)

An emotion and expression recognition module will be developed based on the speech (prosodic and spectral features) and facial (2-D facial points) movements over time. We will consider static recognition systems that use SVM as a baseline, and extend this module to temporal systems using HMM, CNN, and/or CRF.

WP2: What should I do now...mhmmmm..
(Decision Making)

A machine learning based system will be built in order to predict from the recognized expressions from WP1 the expression to synthesize in WP2. This will be done using graphical sequential models such as CRF or HMM and deep learning methods.

WP3: That's how it should look like!
(Synthesis)

Explore synthesis methods for multimodal expressions and a system able to generate audio and facial expressions should be built and work offline. Graphical models (HMM) and deep learning methods will be considered for this task.

WP4: IT'S ALIIIIIVE!!
(Rendering on a Real-Time Avatar)

In this WP, the goal is to make the generated audio and visual features be rendered as real facial and audio expressions on an avatar. So they should also work synchronously and we should be able to control their rendering in real time. A previous real-time working avatar was developed in enterface'15 in Mons but was never put in action [2]. We would thus finally put it into action and improve it as much as needed.

WP5: Do it! NOW!!!
(Implementation in Real-time)

Make sure everything works in real time so that our different WPs can interact with users in a real life experiment.

WP6: Let's test all that!
(Building the Social Experiment)

Build a real scenario to test our real-time working system. Or build test modules that would evaluate the efficiency of WP1,2,3. This WP considers building Wizard of Oz experiments and perception tests.

numediart
UMONS RESEARCH INSTITUTE FOR CREATIVE TECHNOLOGIES

inspire
Interaction Sensing and Perception in real environment

Deliverables & Research Benefits

We expect the out of this project to give the following deliverables:

- D1: Inventory and processing of available databases containing nonverbal expressions for synthesis, recognition and prediction.
- D2: Evaluation and comparisons of several AV synthesis systems for nonverbal expressions.
- D3: Implementation of several AV synthesis systems for nonverbal expressions.
- D4: Evaluation and comparisons of several prediction systems for nonverbal expressions.
- D5: Implementation of several prediction systems for nonverbal expressions.
- D6: Evaluation and comparisons of several recognition systems for nonverbal expressions.
- D7: Implementation of several recognition systems for nonverbal expressions.
- D8: Real-time avatar able to interact with an input synthesis system.

Participants Profile Requirements

We are looking for motivated participants interested in machine learning and virtual agents. They should also be able to work in a team and not afraid of challenges. So please do not hesitate to apply!

The ideal candidate profiles will be:

1. For the recognition, prediction and synthesis modules:
    - Prior knowledge in machine learning.
    - Programming skills required, preferably: Matlab, Python (Java, C++ is a plus).
    - Prior experience in speech/vision processing tools (such as SPTK, CERT, OpenSmile, OpenFace, etc...) would be preferred.
2. For the avatar implementation:
    - Knowledge in 3D creation/animation tools (preferably Blender [19]).
3. For the data collection:
    - General interest in psychology and emotion research

numediart
UMONS RESEARCH INSTITUTE FOR CREATIVE TECHNOLOGIES

inspire
Interaction Sensing and Perception in real environment

Profile Team

Principal Investigators

Dr. Yelin Kim is an assistant professor in the University at Albany, State University of New York (SUNY Albany) since September 2016. She directs the Inspire lab (Backronym: Interaction Sensing and Perception in Real Environment) at SUNY Albany. She received her M.S. and Ph.D. in Electrical and Computer Engineering from the University of Michigan, Ann Arbor. Her Ph.D. thesis was entitled "Automatic Emotion Recognition: Quantifying Dynamics and Structure in Human Behavior." Her work received several awards, including the Best Student Paper Award from ACM Multimedia 2014. Her main research interests are in human-centered and affective computing, multimodal sensing, and computational behavior analysis. Her research builds upon techniques from machine learning, multimodal (speech and video) signal processing, computer vision, and behavioral science. The long-term research goal is to understand human interactions, by using data-driven AI approaches on audio-video recordings of the interactions.

Dr. Hüseyin Çakmak holds a double degree in Aeronautics from the Higher Institute of Aeronautics and Space (ISAE) and in Electrical Engineering from the Polytechnic Faculty of Mons (FPMS). In 2013, he won a FRIA grant to continue with a PhD thesis. In 2016, he finished his PhD on audiovisual laughter synthesis based on a statistical approach. His research interests are audio and visual synthesis and recognition.

Kevin El Haddad is a teaching assistant and Ph.D. candidate at the University of Mons. He holds an M.S. in microsystems and embedded systems from the Lebanese University in 2013. His Ph.D. work currently focuses on the use of nonverbal and affective expressions in human-agent interactions, with a focus on smiling and laughter. His research interests include machine learning, affective computing, human-agent interactions and signal processing. He lead 2 previous eNTERFACE projects.

numediart
UMONS RESEARCH INSTITUTE FOR CREATIVE TECHNOLOGIES

inspire
Interaction Sensing and Perception in real environment

Proposed Team Members

Dr. Payton Lin received the B.S. degree in cognitive science and biology from university of California, San Diego, in 2005, and the Ph.D. degree in biomedical engineering from University of California, Irvine, in 2012. He is currently a researcher at Center for Information Technology Innovation (CITI), Academia Sinica, Tapei, Taiwan. He was a postdoctoral researcher at the Department of Electronic Engineering, City University of Hong Kong, from 2013 to 2014. His research interests include user-centered design, neural networks, computational neuroscience, machine learning, and micro-electromechanical systems.

Jaebok Kim is a Phd student at the Human Media Interaction group, University of Twente. He studied automatic speech recognition and speech emotion recognition during a master's program in Korea Advanced Institute of Science and Technology (M.Sc., 2011, Daejon, Korea) and worked as a research engineer in LG Electronics Advanced Research Institute (2011-2014, Seoul, Korea). His research focuses on automatic analysis of children's speech using machine learning methods.

Minha Lee is a PhD student at the Technical University of Eindhoven, in the Human-Technology Interaction group. She is broadly interested in how people are influenced by morally defining events, as observed through and mediated by technology. She previously graduated from the University of Amsterdam (M.Sc. in Information Science), Pratt Institute in Brooklyn, NY (B.F.A. in Digital Arts), and University of Minnesota - Twin cities (B.A. in Philosophy).

Yong Zhao is a joint-PhD student at the VUB-NPU Joint Audio Visual Signal Processing Research Lab, Vrije Universiteit Brussel and Northwestern Polytechnical University. He received his B.S. degree in computer science and technology, and the M.S. degree in computer application technology from Northwestern Polytechnical Univeristy in 2010 and 2013 respectively. His research focuses on facial expression synthesis with machine learning methods.

References

[1] K. El Haddad, H. Çakmak, E. Gilmartin, S. Dupont, and T. Dutoit, "Towards a listening agent: A system generating audiovisual laughs and smiles to show interest," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ser. ICMI 2016. New York, NY, USA: ACM, 2016, pp. 248–255. [Online]. Available: http://doi.acm.org/10.1145/2993148.2993182

[2] Y. Kim and E. Mower Provost, "Emotion spotting: Discovering regions of evidence in audio-visual emotion expressions," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ser. ICMI 2016. New York, NY, USA: ACM, 2016, pp. 92–99. [Online]. Available: http://doi.acm.org/10.1145/2993148.2993151

[3] Y. Kim, H. Lee, and E. Mower Provost, "Deep learning for robust feature generation in audio-visual emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 3687–3691.

[4] P. Lin, D. Lyu, F. Chen, S.-S. Wang, and Y. Tsao, "Multi-style learning with denoising autoencoders for acoustic modeling in the internet of things (iot)," in *Accepted for publication, Feb. 2017 in Computer Speech and Language*. [Online]. Available: https://www.citi.sinica.edu.tw/papers/yu.tsao/5584-F.pdf

[5] H. Çakmak, "Audiovisual laughter synthesis a statistical parametric approach," Ph.D. dissertation, University of Mons, 2016.

[6] Çakmak Hüseyin, E. H. Kevin, and D. Thierry, "A real time OSC controlled agent for human machine interactions," in *Proceedings of 7th Workshop on Artificial Companion, Affect, Interaction (WACAI 2016)*, Brest, France, 13-14 June 2016.

[7] "Website Laughter," http://tcts.fpms.ac.be/ laughter.

[8] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, p. 335, 2008. [Online]. Available: http://dx.doi.org/10.1007/s10579-008-9076-6

[9] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, April 2013, pp. 1–8.

[10] A. J. Aubrey, D. Marshall, P. L. Rosin, J. Vandeventer, D. W. Cunningham, and C. Wallraven, "Cardiff conversation database (ccdb): A database

of natural dyadic conversations," in *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, June 2013, pp. 277–282.

[11] R. v. Son, W. Wesseling, E. Sanders, and H. v. d. Heuvel, "The ifadv corpus : A free dialog video corpus," in *Proceedings of the sixth international language resources and evaluation (LREC 2008)*, 27 May 2008. [Online]. Available: http://hdl.handle.net/2066/68299

[12] "The av-lasyn database : A synchronous corpus of audio and 3d facial marker data for audio-visual laughter synthesis."

[13] T. Kanade, Y. Tian, and J. F. Cohn, "Comprehensive database for facial expression analysis," in *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000*, ser. FG '00. Washington, DC, USA: IEEE Computer Society, 2000, pp. 46–. [Online]. Available: http://dl.acm.org/citation.cfm?id=795661.796155

[14] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, June 2010, pp. 94–101.

[15] D. Chugh, M. H. Bazerman, and M. R. Banaji, "Bounded ethicality as a psychological barrier to recognizing conflicts of interest," *Conflicts of interest: Challenges and solutions in business, law, medicine, and public policy*, pp. 74–95, 2005.

[16] E. Dane and S. Sonenshein, "On the role of experience in ethical decision making at work: An ethical expertise perspective," *Organizational Psychology Review*, vol. 5, no. 1, pp. 74–96, 2015.

[17] G. M. Lucas, J. Gratch, A. King, and L.-P. Morency, "It's only a computer: virtual humans increase willingness to disclose," *Computers in Human Behavior*, vol. 37, pp. 94–100, 2014.

[18] J. Haidt, "The moral emotions," *Handbook of affective sciences*, vol. 11, pp. 852–870, 2003.

[19] "Blender Website," https://www.blender.org/.