2018

End-to-End Listening Agent for Audiovisual Emotional and Naturalistic Interactio

El Haddad, Kevin

Escola das Artes, Universidade Catolica Portuguesa

http://dx.doi.org/10.7559/citarj.v10i2.424

End-to-End Listening Agent for Audiovisual Emotional and Naturalistic Interactions

Kevin El Haddad
TCTS Lab, Polytechnic Faculty of the University of Mons,
Mons, Belgium
---
kevin.elhaddad@umons.ac.be

Yara Rizk
Department of Electrical and Computer Engineering, American University of Beirut,
Beirut, Lebanon
---
yar01@mail.aub.edu

Louise Heron
Department of Psychology, University of Bath,
Bath, UK

L.Heron@bath.ac.uk

Nadine Hajj
Department of Electrical and Computer Engineering, American University of Beirut,
Beirut, Lebanon
---
njh05@aub.edu.lb

Yong Zhao
VUB-NPU Joint AVSP Research lab, Vrije Universitert Brussel, Brussels, Belgium & Northwestern Polytechnical University, Xi'an, China.
---
yzhaol@etrovub.be

Jaebok Kim
Human Media Interaction group, University of Twente,
Enschede, Netherlands
---
j.kim@utwente.nl

Ngô Trọng Trung
University of Eastern Finland Faculty of Science and Forestry / School of Computing
---
trung.ngotrong@helsinki.fi

Minha Lee
Human-Technology Interaction group - Technical University of Eindhoven,
Eindhoven, Netherlands
---
m.lee@tue.nl

Marwan Doumit
National Public Radio (NPR),
Washington D.C., USA

Payton Lin
Research Center for Information Technology Innovation, Academia Sinica,
Taiwan

Yelin Kim
Department of Electrical and Computer Engineering, University at Albany, State University of New York,
Albany, NY,USA
---
yelinkim@albany.edu

Hüseyin Çakmak
TCTS Lab, Polytechnic Faculty of the University of Mons,
Mons, Belgium
---
huseyin.cakmak@umons.ac.be

ABSTRACT

In this work, we established the foundations of a framework with the goal to build an end-to-end naturalistic expressive listening agent. The project was split into modules for recognition of the user's paralinguistic and nonverbal expressions, prediction of the agent's reactions, synthesis of the agent's expressions and data recordings of nonverbal conversation expressions. First, a multimodal multitask deep learning-based emotion classification system was built along with a rule-based visual expression detection system. Then several sequence prediction systems for nonverbal expressions were implemented and compared. Also, an audiovisual concatenation-based synthesis system was implemented. Finally, a naturalistic, dyadic emotional conversation database was collected. We report here the work made for each of these modules and our planned future improvements.

KEYWORDS

Listening Agent; Smile; Laughter; Head Movement; Eyebrow Movement; Speech Emotion Recognition; Nonverbal Expression Detection; Sequence-to-Sequence Prediction Systems; Multimodal Synthesis; Nonverbal Expression Synthesis; Emotion Database; Dyadic Conversation Database.

ARTICLE INFO

Received: 10 November 2017
Accepted: 31 July 2018
Published: 08 November 2018

https://dx.doi.org/10.7559/citarj.v10i2.424

1 | INTRODUCTION

This project is part of the eNTERFACE'17 Workshop. eNTERFACE is a multidisciplinary workshop focusing on multimodal interfaces. It gathers, every year, researchers from around the world to work on different projects for a month. 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 to build a virtual agent which recognizes and takes into account various nonverbal expressions and reacts to the user by generating naturalistic feedbacks. Here, we consider a context of dyadic interactions. And since we focus on nonverbal expressions, the subject of the discussion is not relevant as the goal of the agent is to react to nonverbal expressions with nonverbal expressions. 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. Thus, we rather focus on the speaker's nonverbal and paralinguistic behaviors to predict and generate the agent's nonverbal behavior in real-time.

Figure 1 shows the overall workflow of our project which is a basic pipeline of a human-agent interaction system. Our agent will be built on **recognition**, **prediction** and **synthesis** modules. **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 is then generated by the **synthesis** module. This latter's output is rendered on a human-like avatar (Figure 1). In parallel to the development of these modules, we collected a naturalistic emotional dyadic conversation database as explained later in this paper.

The project we propose is inspired by some of the work found in the literature such as (El Haddad, Cakmak, Gilmartin, Dupont, & Dutoit, 2016). In that work, the authors developed some of the modules previously described here. Indeed, in that work, an audiovisual (AV) concatenative synthesis system and a prediction system are presented. Both were built to create a listening agent. The prediction system is a Conditional Random Field (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.

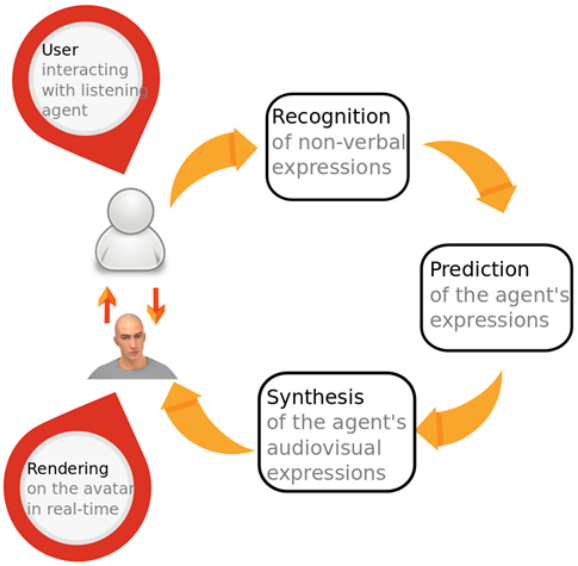


Figure 1 | Modules constituting the listening agent.

A first step towards a fully functional listening agent is to first build a recognition module that would feed a prediction system such as the one mentioned above. For this we utilize both audio and visual signals for emotion recognition using machine learning techniques, such as temporal (Kim & Provost, 2016) or deep learning models (Kim, Provost, & Lee, 2013), (Kim, 2017a, 2017b).

We also compare several sequence prediction systems, other than the CRF to predict the agent's expressions of attentiveness when reacting to a speaker/user. In (El Haddad, Cakmak, Gilmartin, Dupont, & Dutoit, 2016), only smiles, laughs and their intensity levels were considered, here we intend to consider more expressions as will be seen in what follows.

In the project we work on improving the synthesis system mentioned above by making it more generic so that it becomes easier to use with the ability to generate more variate set of expressions.

The expressions considered in this project are: laughs and smiles and their intensity dimensions, head movements (nodding, shaking and tilting) and eyebrow movements (raise and frown), for they frequently occur in dyadic interactions. These expressions are a part of all previously mentioned modules. Depending on the module concerned, the audio, video and motion capture signals will be considered.

In what follows, we detail the work done during in this work for each of the previously mentioned modules.

2 | RECOGNITION

In this section we describe the detection of nonverbal and paralinguistic events occurring during an interaction with a user. This was split in two main tasks:

1. Multimodal Emotion recognition (MER) which predicts arousal and valence values for incoming sentences.

2. Expression detection which detects a list of conversational nonverbal expressions.

Since our agent should work in noisy environments and with "in the wild" data, we chose, for this module, to work with the RECOLA and SEWA databases which meet our requirements.

2.1 MULTIMODAL EMOTION RECOGNITION

The state-of-the-art techniques for MER are based on deep learning (Trigeorgis, et al., 2016). A multimodal system was built with a late fusion approach. On one side, the spectrograms of the audio cue were used to train a system similar to the one described in (Kim, et al., 2017b). As shown in Figure 2, a convolutional neural network (CNN) was used to extract descriptors from the data. The extracted features are then fed to a long short-term memory network (LSTM) which is expected to learn dynamic features since the data is time-dependent. On the other side, we used the pre-trained VGG-16 network to extract features from the visual cue which was the subjects face images cropped using the OpenFace tool (Baltruvsaitis, Robinson, & Morency, 2016). Finally, both networks were connected to a Fully Connected (FC) neural network, and both arousal and valence tasks were simultaneously trained in a multi-task learning fashion (Kim, et al. 2017a).

Due to the difference between the data in both datasets, the system was trained and tested on each database separately. Both contain continuous annotations of the valence and arousal. Considering the relatively limited amount of data available in each database, the valence and arousal values were discretized similar to (Tadas Baltrušaitis, 2016) and the problem of regression was turned into a binary classification problem.

By the end of the workshop we evaluated our model using the RECOLA dataset. We conducted 5-fold speaker-independent cross validations and obtained

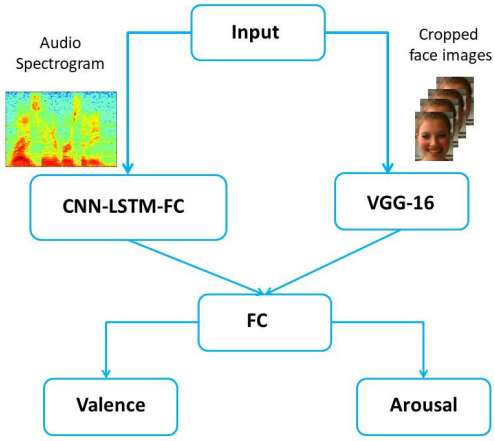


Figure 2 | Deep multitask learning for multimodal speech emotion recognition.

un-weighted accuracy of approximately 60% for arousal and 50% of valence classifications.

2.2 EXPRESSION DETECTION

For this task, we chose rule-based approaches instead of deep learning based ones due to both the limited amount of data available and the considerable variance of the nonverbal conversation expressions intra- and inter-speakers. Of the four expressions mentioned in the introduction, only three were considered and extracted from the video recordings of the human subject: head movement, eyebrow movement and smiling. Although some preliminary work on laughter was undertaken, results concerning it will be reported in future work. The workflow to extract these expressions is shown in Figure 3.

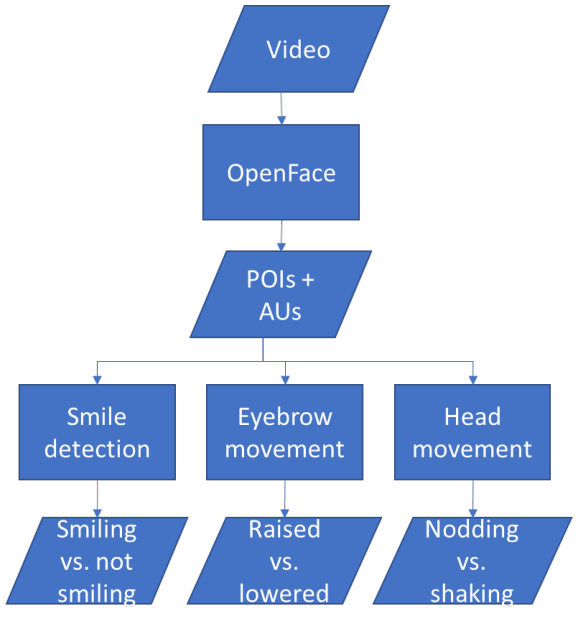


Figure 3 | Facial Expression Detection Workflow.

The video processing is done using mainly OpenFace to extract action units (AUs) and the coordinates of points of interests (POIs) or landmarks in the face.

Smile Detection

Smile detection is based on the values of AU6 (cheek raiser) and AU12 (lip corner puller). If at least one of the AUs was detected in the frame, then a smile was detected and its intensity is based on the intensity of the detected AU. Figure 4-a) plots the intensity of the smiles detected in a video recording with over 8000 frames. Figure 4-b) shows one of the frames from the video where the smile of the subject is detected.

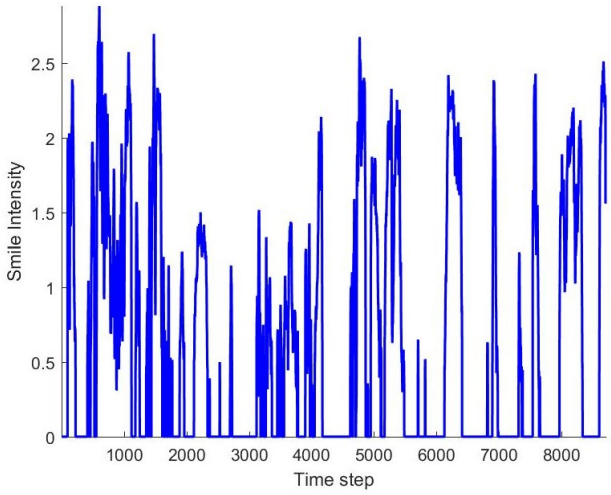
Eyebrow Movement Detection

Eyebrow movement detection is based on the values of AU1 (inner brow raiser), AU2 (outer brow raiser) and AU4 (brow lowerer). This rule based approach determines whether the subject has the eyebrows raised or lowered by comparing the intensities of these three AUs. Specifically, if AU1 or AU2 has a greater intensity than AU4, then the subject is determined to have raised eyebrows. Otherwise, the subject is considered to have lowered eyebrows. If all three AUs have an intensity of zero, then the subject's eyebrows are in neutral position (neither raised nor lowered).

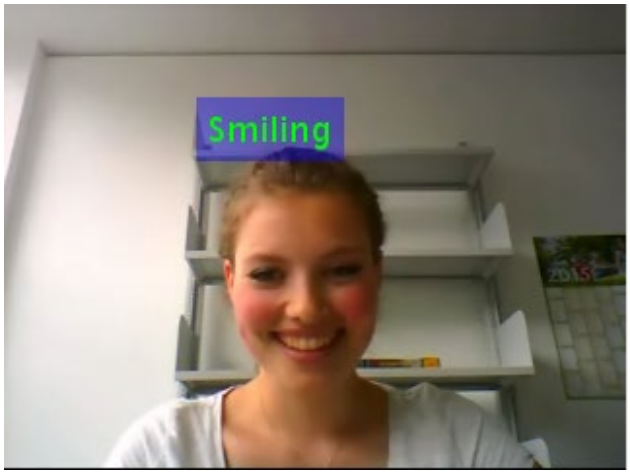
Head Movement Detection

Head movement detection distinguishes between three possible states: head nodding, head shaking or neither. A rule based approach is adopted that tracks the changes of two POIs located between the eyes to identify the type of head movement. Head shaking is characterized by significant oscillation of the x-coordinate of POIs with minimal changes in the y-coordinates while head nodding is characterized by y-axis oscillations with minimal changes in the x direction. If both coordinates exhibit large changes, we consider that the whole head was displaced due to the subject moving in the video; the subjects were not required to stay still. Table 1 summarizes the adopted algorithm to detect head movements.

Figure 5-a) and Figure 5-b) show instances of when head shaking and nodding were detected, respectively (the black line).



a) Smiling intensity.



b) Example of a detected smile.

Figure 4 | Smile detection.

Laughter Detection

Systems can be found in the current literature such as (Neuberger & Beke, 2013), but they usually are not efficient in noisy environments if considering audio and video modalities only, especially for low intensity level laughs. Future work will focus more on building a robust laughter detection system by merging several databases and using deep learning methods that can handle the variance of laughs.

2.4 FUTURE WORK

Concerning the smiles, eyebrow movements and head movements, although the rule-based and AU-based approaches presented are rather simplistic and are not optimal for all the expressions considered here, they were the most efficient we could implement given the time and resources available during this workshop. In the future we plan on developing more robust machine learning based detection systems for these expressions.

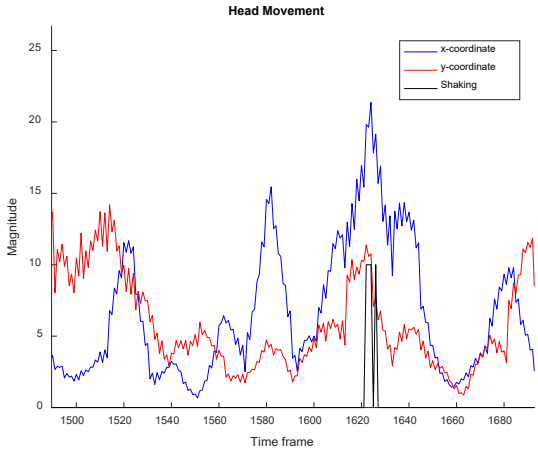
Table 1 | Algorithm to detect head movement.

1.	Define movement thresholds: • x_threshold = 10% x average of the difference between the maximum and minimum x values • y_threshold = 10% x average of the difference between the maximum and minimum y values
2.	Compute movement in horizontal direction over 10 frames x_movement = max(sum(abs(x(t) – x(t-1))
	Compute movement in vertical direction over 10 frames y_movement = max(sum(abs(y(t) – y(t-1))
3.	If x_movement > x_threshold AND y_movement < y_threshold Then head shaking is detected
	If x_movement < x_threshold AND y_movement > y_threshold Then head nodding is detected
	Otherwise, no movement in detected

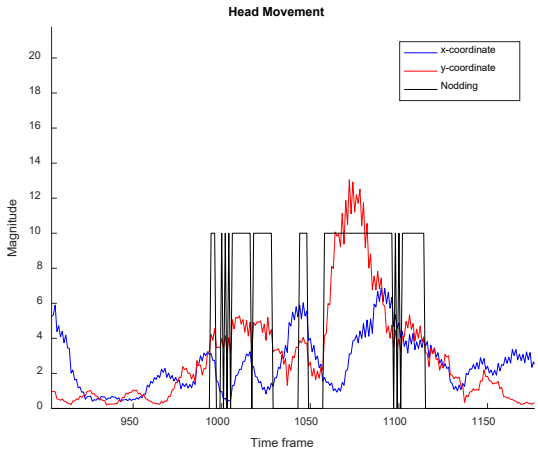
Concerning the MER system, we intend to merge the SEWA and RECOLA databases to increase the amount of data available to train our systems and improve our results. We also intend to leverage other datasets such as IEMOCAP (Busso, et al., 2008) and IMPROV (Busso, et al., 2017) to take advantage of the potential of deep learning.

3 | PREDICTION

The goal of this module is to foresee the listener's expressions given that of the speaker. The problem is tackled as a one-step-ahead prediction task where the system generates the appropriate responses on a frame-by-frame basis. For that, data is split into two subsets: one corresponding to the listener and the other pertaining to the speaker. While the former is treated as the output, the latter is regarded as the set of predictors for the module.



a) Head shaking detected.



b) Head nodding detected.

Figure 5 | Head movement detection.

3.1 DYADIC CONVERSATION DATA

Data of nonverbal expressions of a listener to a speaker were needed. We chose the Cardiff Conversation Database (CCDB) because it contains dyadic conversations during which the interlocutors discuss in a naturalistic way. Although it already contained annotations of some nonverbal expressions, these were not well suited for our task. We therefore re-annotated smiles and laughs in three different intensity levels each, as well as head movements (nodding, shaking, tilting) and eyebrow movements (raise and frown for left, right or both eyes). The annotations were made using the ELAN annotation software. For this project, only 10

conversations (20 videos) were used to train the systems.

3.2 PREDICTION SYSTEMS

To predict the responses of the agent a set of algorithms were tested:

- Linear regression
- Naïve Bayes classification
- Decision tree
- Fuzzy inference system
- Recurrent neural networks

We will ultimately evaluate the generated expressions subjectively, since our goal is to obtain adequate reactions, and not to copy a listener's reactions from a dataset. For this, we will synthesize the predicted sequence of expressions via the synthesis module and evaluate the synthesized expressions in the context of their generation (as reactions to a speaker/user). For the current study, these models were evaluated based on three performance measures: the accuracy, training error and testing error (mean squared error). Table 2 summarizes our results. Below is a detailed description of each of the tested algorithms with the corresponding assumptions and/or preprocessing.

Linear regression

In this model, the dependent variables (listener's expressions) are explained as a linear combination of the regressors (speaker's expression). A multivariate regression framework is employed to predict each of the four expressions as a function of that of the speaker; for example, the agent's head movement is a result of a weighted linear combination of the speaker's head movement, eyebrow movement, smile and laughter. The problem is reduced to learning the optimal weights for each predictor/output combination that minimize the mean squared error on the labeled training data. Due to the categorical nature of the regressor matrix, a conversion to a numerical format (one-to-one mapping) is required. The output is converted to a categorical representation using a threshold-based rule.

Naïve Bayes Classification

The Naïve Bayes classifier is a probabilistic technique that constructs conditional probability using Bayes' theorem assuming naïve (i.e. independent) features. The strength of this method lies in its simplicity requiring only a linear number of parameters in terms of predictor/output pairs. In our model, the parameters are learned using a maximum likelihood algorithm. The use of a probabilistic formulation is possible when using categorical features and hence no numerical conversion is needed for this model. Figure 6 shows a sample of predicted expressions over a period of 10,000 frames. As can be seen in Figure 6, the independence assumption is highly erroneous in our case (it is fair to assume that a person's smiling and nodding simultaneously are correlated).

Decision Tree Learning

A decision tree is a prediction technique that models observations of a feature as a branch and outputs as leaves. While considered to be a non-robust method, it presents the advantage of simple learning

Table 2 | Systems Prediction Performance.

	Accuracy (%)					Training Error					Testing Error				
	H	L	S	E	All	H	L	S	E	All	H	L	S	E	All
Linear Regression	94.7	90.8	93.9	93.2	75.4	0.09	0.09	0.27	0.07	0.13	0.12	0.12	0.24	0.07	0.14
Naïve Bayes Classification	95.4	96.7	93.6	98.1	84.7	0.09	0.05	0.34	0.07	0.14	0.13	0.15	0.29	0.07	0.16
Regression tree	95.6	96.3	93.9	98.0	84.7	0.09	0.09	0.27	0.07	0.13	0.12	0.12	0.24	0.07	0.14
Fuzzy inference system	94.8	92.3	92.9	96.3	78.4	0.09	0.09	0.28	0.06	0.13	0.11	0.12	0.24	0.069	0.13
Recurrent Neural Networks	95.6	96.3	93.9	98.0	84.7	0.09	0.08	0.27	0.06	0.12	0.11	0.12	0.23	0.06	0.13

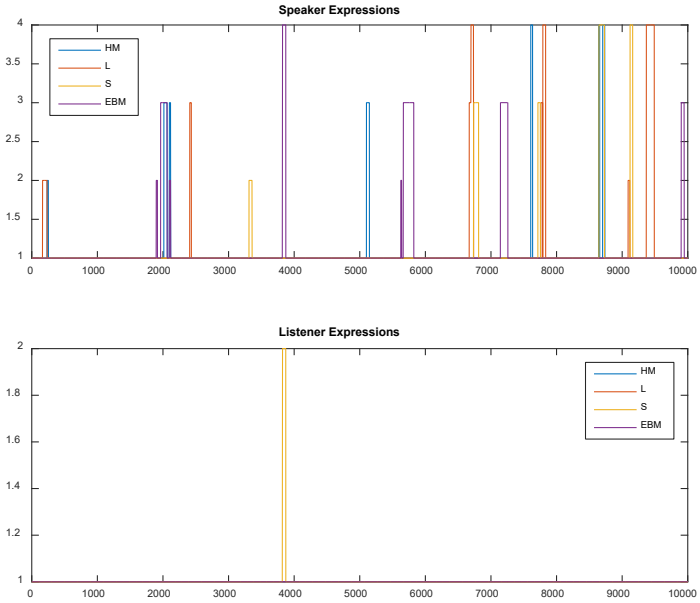


Figure 6 | Sample of predicted expressions using Naive Bayes Classification. Each value of an expression represents one of the possible values that expression can take.

by maximizing the information gain of every attribute and subsequently eliminating those with no discriminative ability.

Fuzzy Inference System

A fuzzy inference system models a prediction problem by mapping attributes to fuzzy sets. Each instance is hence characterized by its probability to belong to a particular set rather than a crisp 0 or 1 mapping in classical sets. Outputs or decisions are then made based on a defuzzification (i.e. mapping from a probabilistic representation to a crisp one) of the result of inference rules, which describe the input-output mapping using a set of logic rules. In our case, a speaker's expressions are described as belonging to 4 fuzzy sets (the 4 expressions) with the probability of belonging modeled as Gaussian distributions. A similar mapping is performed for the output.

Recurrent Neural Networks

Recurrent nets are desirable for their ability to robustly model time series prediction problems such as the one tackled in this module. These models include recurrent connections (at the input layer, hidden layers, output layer or any combination of these) that serve as a time varying real-valued activation function. Hence, this allows the network to exhibit time dependency properties. These models suffer from a lengthy and computationally expensive training, rendering their implementation on limited computational resources challenging. In our module, we employ a fully recurrent neural network with two hidden layers each composed of 100 neurons and trained using gradient descent.

3.3 FUTURE WORK

A next step would be to evaluate objectively all the models tested here in a similar way as in (El Haddad, Cakmak, Gilmartin, Dupont, & Dutoit, 2016). This means that each system will be used to predict the reactions of the agent to a speaker and these expressions will be synthesized using the synthesis module. The synthesized output will undergo subjective tests to evaluate the relevance of the expressions chosen by the system with respect to the speaker to which the agent is reacting.

4 | SYNTHESIS

To generate audiovisual nonverbal expressions, we relied on the concatenation system described in (El Haddad, Cakmak, Gilmartin, Dupont, & Dutoit, 2016). It was improved here by implementing a python-based animation of facial expressions, a **search algorithm** for queried expressions, a **facial normalization technique** to use different types of data and the same interpolation technique as in our previous work.

4.1 SYSTEM OVERVIEW

This system relies on a dataset of AV expressions from which the best suited expressions are picked based on the requirements of a query and concatenated together to form a full sequence. The parameters controlled from the query are currently:

- The type of the expression (laughter, smiles, head nodding, head tilting, head shaking, raise or frown eyebrow left or right)
- The intensity of the expression
- The duration of the expression

The expressions in the AV dataset were manually annotated according to these parameters.

4.2 AUDIOVISUAL CONCATENATION SYSTEM

To concatenate two facial expressions, the starting and ending frames of an expression and the one that succeeds it, respectively, are likely to present discontinuities even for the same expression. To achieve a smooth transition, we used a linear interpolation approach in (El Haddad, Cakmak, Gilmartin, Dupont, & Dutoit, 2016) on the extracted face landmarks. Given two sequences of facial expressions A and B, where B should be concatenated to the end of A. Thus, we apply the interpolation between the last frame a of A and the first frame b of B, yielding the interpolated frame f defined as below:

$$f = w1 * a + w2 * b$$

where $w1$ and $w2$ denote the interpolation weight, and $w1 + w2 = 1$. For a smooth transition, we create more frames by interpolating between a and f, f and b, until the transition looks natural and smooth.

Concerning the audio cue, only laughter is expressed audibly and is therefore concatenated to silence. No smoothing interpolation is needed, but concatenation and truncation are used to control the length of the silence signal.

4.3 RENDERING

The animation is composed of visual and audio cues which are used and should ultimately be rendered on the avatar shown in Figure 7 (Çakmak, El Haddad, & Pulisci, 13-14 June 2016). The visual cue is controlled by facial landmarks extracted from the

Figure 7 | Human-like avatar.

facial expressions in the dataset. The OpenFace tool was used to automatically extract these facial landmarks. This serves as a first and easy visualization of the generated expressions. Since the landmarks defined to control the avatar are not the same as the ones extracted by OpenFace, ultimately a mapping will be made to control the avatar with the OpenFace landmarks.

4.4 EXPRESSION QUERY AND DATASET

The dataset contains segmented expressions of the facial landmarks and the audio (either laughter or silence) separately. The expressions are stored in separated files, the names of which contain the expression's parameters information (type, intensity and duration)

The goal is to receive a query of a sequence of expressions from the prediction module along with the duration and intensity required for each expression and use this query to pick the best suited ones from the dataset.

To have expressions coming from different subjects and therefore maximize the amount of expressions contained in the dataset, we use a landmark normalisation approach. Instead of using the raw landmark coordinate values, we use the movement of the landmark coordinates from one frame to the next (by subtraction). These differences are applied to a reference frame containing initial landmarks for a certain face.

The audio cue normalisation is still an ongoing work. Ultimately voice conversion techniques should be used to transform all the different voices in the dataset to one target voice.

The dataset of expressions for the synthesis module currently contains only a few examples recorded for the purpose of testing the system.

4.5 FUTURE WORK

In the near future we intend to build an extensive dataset of annotated expressions for the synthesis module.

Several voice conversion techniques will be compared to normalize the audio cue of the dataset. Techniques range from simple signal processing methods to bring the fundamental frequency and spectral values, to certain predefined values, to more recent deep learning based techniques such as autoencoders.

Finally, we will work on controlling more parameters such as the social functionality of the expressions, obtaining intermediate intensity levels through interpolation and the possibility of combining several expressions, such as smiling while nodding for instance.

In future work, parametric and deep learning based synthesis systems will also be considered.

5 | DATA COLLECTION

The eNTERFACE workshop gave us the opportunity to collect our own dyadic conversation database. For this, interlocutors took turns as speakers and listeners, the latter asking questions to the former about memories of emotional states. Questions were on negative (i.e. -guilt and shame) and positive (i.e. -pride and compassion) emotions. These emotions are considered to be moral emotions (Haidt, 2003).

5.1 SETUP AND EXPERIENCE DESCRIPTION

Procedure

Participants were asked to read the informed consent form first, which clearly stated that their participation was voluntary and unpaid. The consent form stated that moral emotions will be discussed by the participants. At the start of the experiment, participants were told that they will randomly be assigned to the role of speaker or listener, then switch roles. The speaker answers questions about moral emotions, whilst the listener listens to the speaker's answers and asks any follow up questions as necessary. The instructions were purposively vague to ensure that the dyadic interaction was as natural as possible.

The order in which participants discussed the emotions was randomised: listeners chose one of the two moral emotion options (positive, negative)

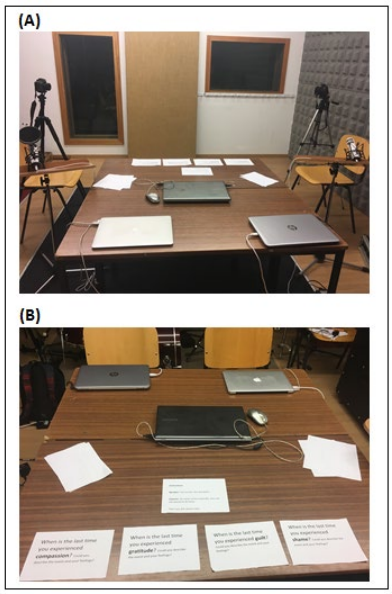


Figure 8 | Pictorial depiction of the experimental setup from (A) side and (B) aerial views.

from question prompts on a table. Each interaction started with the listener asking a question that varied in the emotion category: "When was the last time you experienced gratitude/compassion/guilt/shame? Can you describe the event and your feelings?" The speaker responded to each question. The interaction lasted until the interlocutors both indicated to experimenters that the conversation was finished.

Experimental Setup: Video/Audio Acquisition

Video and audio were recorded in a soundproof room at the Catholic University of Porto, Portugal. Two Canon Cameras: EOS 550D and EOS 6D were used to record the interactions. Camera A (beside the Speaker 1, recorded Listener 1/Speaker 2) and

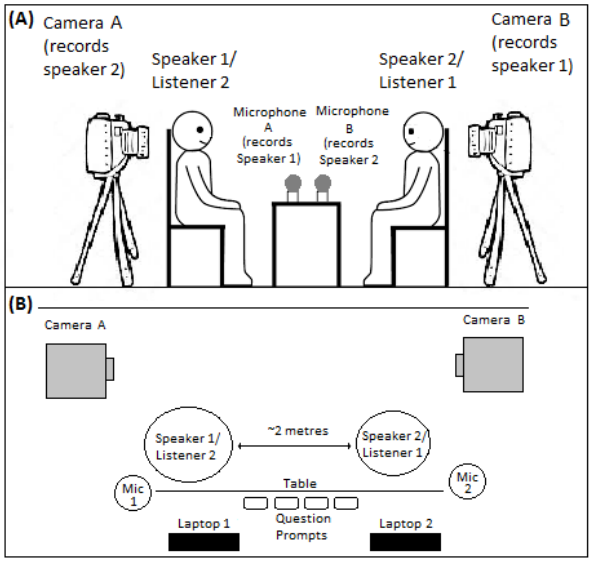


Figure 9 | (A) A side on view of the experimental setup from a third party observer and (B) the setup from the participants' perspective.

Camera B (beside the Speaker 2, recorded Speaker 1/Listener 2). The camera angle and distance were tailored to each participant, ensuring that the head to torso area was captured. The distance between speaker and listener remained constant. Two Rode Podcaster USB microphones on pop shield shock mounts recorded speaker and listener audio. Laptops were attached to microphones for audio recordings and were also used for pre- and post-experimental questionnaires (see Figures 8 and 9 for diagrams of the experimental setup). Two experimenters were in the room and started and stopped video and audio recordings.

5.2 CONTENT

This database was designed to mimic, as much as possible, real world conversations – by being unscripted and containing a mixture of nationalities and familiar/unfamiliar individuals.

Database Demographics

As summarized in Figure 10, the database contains 42 participants (21 pairs), 32 males and 10 females. Participants were largely students and professors (age range: 20 - 48). There were 14 male-male pairs, 3 female-female pairs and 4 male-female pairs, of which 11 pairs knew each other beforehand. The database comprises of 14 nationalities.

Questionnaire Data

Demographic information (e.g., age, gender) was collected. Further, to obtain a richer dataset, mood and personality measurements were also collected.

Participants completed pre- and post- experiment mood evaluation by filling in the Positive and Negative Affect schedule (Watson, 1988). Then after the experiment, they completed additional stress and personality measures: Stress Response Scale (Suzuki, 1998) and the Big Five Inventory (John, 1999).

File Contents of the Database

Our database contains audio, video, excel and ELAN files.

1. **Audio.files (wav)** These contain the audio content of speakers and listeners during the interaction.

2. **Video files (MPEG-4)**. These contain both visual and audio content. An example of the footage captured using this setup is in Figure 11.

3. **Excel Files of Demographic/mood/ personality data.** These files contain all pre- and post- recording data, including, pre- and post-mood scores and post- personality data.

4. **ELAN Files**. These are ongoing and contain annotations of video files for non-verbal cues, including, smiles, nods and laughs.

Ongoing Annotations

Annotations of nonverbal expressions were initiated during the last week of eNTERFACE. These annotations concern:

- Smiles: 3 different intensity levels
- Laughs: 3 different intensity levels
- Head movements: nodding, tilting and shaking
- Eyebrow movements: raise and frown for the left, right or both eyes.

The fully annotated database could be useful for all other modules since it contains naturalistic emotional conversations along with their nonverbal expressions.

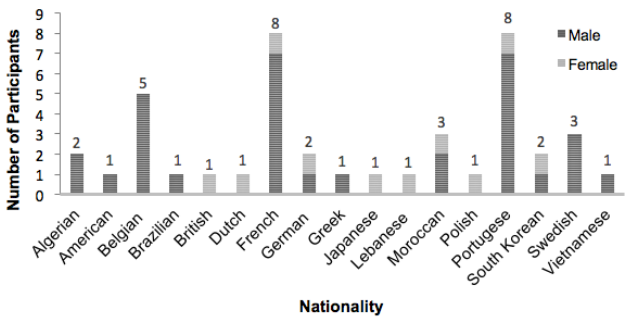


Figure 10 | Participant gender and nationality distribution in the database.

Figure 11 | An example of captured video stimuli.

6 | CONCLUSION

We were successfully able to build the foundations of an end-to-end expressive listening agent system. Based on these foundations we hope to bring, in the near future, a set of tools for each of the components described above that should be useful to anyone interested in affective agents and willing to contribute to our work.

REFERENCES

Aubrey, A., Marshall, D., & Rosin, L. (2013). Cardiff conversation database (ccdb): A database of natural dyadic conversations. 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 277–282). Portland, OR, USA: IEEE.

Baltruvsaitis, T., Robinson, P., & Morency, L.-P. (2016). Openface: an open source facial behavior analysis toolkit. IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 1--10). IEEE.

C. Busso, M. B. (vol. 42, no. 4, pp. 335-359). IEMOCAP: Interactive emotional dyadic motion capture database. Journal of Language Resources and Evaluation, December 2008.

C. Busso, S. P. (January-March 2017). MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception. IEEE Transactions on Affective Computing, vol. 8, no. 1, pp. 119-130.

El Haddad, K., Cakmak, H., Gilmartin, E., Dupont, S., & Dutoit, T. (2016). Towards a listening agent: a system generating audiovisual laughs and smiles to show interest. 18th ACM International Conference on Multimodal Interaction (ICMI 2016) (pp. 248-255). Tokyo, Japan: ACM, New York, NY, USA.

Haidt, J. (2003). The moral emotions. *Handbook of Affective Sciences*, *11*, 852-870.

Hüseyin Çakmak, K. E. (13-14 June 2016). A real time OSC controlled avatar for human machine interactions. Workshop on Artificial Companion Affect Interaction. Brest, France.

Kim, J., Englebienne, G., Truong, K. P., and Evers, V. (2017b). "Deep Temporal Models using Identity Skip-Connections for Speech Emotion Recognition." In: Proceedings of ACM Multime- dia, pp. 1006–1013.

Kim, J., Englebienne, G., Truong, K. P., and Evers, V. (2017a). "Towards Speech Emotion Recognition "in the wild" using Aggregated Corpora and Deep Multi-Task Learning." In: Proceedings of INTERSPEECH, pp. 1113–1117.

Kim, Y., & Provost, E. M. (2016). Emotion spotting: discovering regions of evidence in audio-visual emotion expressions. 18th ACM International Conference on Multimodal Interaction (ICMI 2016) (pp. 92-99). Tokyo, Japan: ACM, New York, NY, USA.

Kim, Y., Provost, E. M., & Lee, H. (2013). Deep learning for robust feature generation in audiovisual emotion recognition. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013) (pp. 3687-3691). Vancouver, BC: IEEE.

Tadas Baltrušaitis, P. R.-P. (2016). OpenFace: an open source facial behavior analysis toolkit. IEEE Winter Conference on Applications of Computer Vision (WACV). Lake Placid, NY, USA: IEEE.

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B. W., & Zafeiriou, S. (2016). Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5200-5204). Shanghai, China: IEEE.

BIOGRAPHICAL INFORMATION

Kevin El Haddad is a teaching assistant and Ph.D. student 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. His research interests include machine learning, affective computing, human-agent interactions and signal processing. He interned at Disney Research helping to develop human-agent interaction systems and lead 2 previous eNTERFACE projects.

Yara Rizk is a PhD student enrolled in the electrical and computer engineering department at the American University of Beirut (AUB). Prior, she has received her BE in computer and communication engineering from AUB, Lebanon, in 2012. Her research interests span robotics, multi agent systems, machine learning, classification, clustering,

and artificial intelligence. Rizk has attended a technical internship (2013-2014) at Intel in Hillsboro, Oregon, USA.

Louise Heron is a PhD student working in the Psychology department at the University of Bath, United Kingdom. Previously, she completed an MSc in Research Methods in Psychological Science at the University of Glasgow in 2015 and an MA (Hons) in Psychology at the University of Dundee in 2014. Her research interests focus on face and voice perception, in particular, on what social trait information can be gleaned from faces and voices. She also explores methodological approaches to studying face/voice perception, drawing on psychology and computer vision perspectives.

Nadine Hajj is a PhD candidate in the department of Electrical and Computer Engineering at the American University of Beirut. She obtained her M.E. in Electrical and Computer Engineering from the American University of Beirut in 2013 and her B.E. from the Lebanese University in 2010. From 2011 to 2012 she joined Intel corporation as a graduate research intern. Her research interests include computational intelligence, deep learning and computational neuroscience.

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 University in 2010 and 2013 respectively. His research focuses on facial expression synthesis with machine learning methods.

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.

Trung Ngô Trọng is a Ph.D. candidate at the University of Eastern Finland (UEF), where he is currently focusing on the development of semi-supervised learning algorithm applied to signal processing. He graduated from Hanoi University of Science and Technology, with bachelor degree in computer engineering, and received his master in information technology from UEF. His research interests include semi-supervised learning, deep learning, probabilistic modeling and causal inference. He received best paper award at IWSDS 2018.

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 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).

Marwan Doumit is a Graduate Student in Computer Science, with an emphasis on theoretical computing, at George Mason University, and a Software Engineer and part of the R&D department of ICES Corporation (Fairfax, Virginia., USA). His work focuses mainly on full stack software development. Marwan interests include machine learning and data analysis in general and image processing in particular.

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.

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.

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.