Annotating Nonverbal Conversation Expressions in Interaction Datasets

Kevin El Haddad, Noe tits, Thierry Dutoit
numediart Institute, University of Mons /31 Boulevard Dolez, Mons, Belgium
{kevin.elhaddad, noe.tits, thierry.dutoit}@umons.ac.be

Abstract

In this paper, we present our work on building a database of Nonverbal Conversation Expressions (NCE). In this study, these NCE consist of smiles, laughs, head and eyebrow movements. We describe our annotation scheme and explain our choises. We finally give inter-rater agreement results on small part of the dataset

1 Introduction

Virtual agent systems like chatbots, virtual assistants, etc. have seen a lot of improvements in the last decades thanks mainly to the progress of artifical intelligence in general and machine learning/deep learning in particular. These systems are becoming more and more part of our daily lives and will become more enchored in it in the near future. It it therefore important that our interactions with them be as comfortable as possible. This is why it is important for them to better understand the human ways of interaction and also to be able to behave in a human-like way.

Nonverbal and paralinguistic expressions form a big part of human-human interactions. They are very frequent and have different important functionalities. It was reported that laughter, for instance, accounted for about 10% of the total verbalizing time(**?**). Other studies also report the importances of these nonverbal expressions in interactions(). But they are yet to be well implemented in human-agent interaction systems.

In this paper we present an ongoing work on building a nonverbal conversation expression dataset. Nonverbal conversation expressions or NCE (El Haddad, 2017) are expressions that come to complement the semantic of a sentence's linguistic content (e.g. emotional speech), or as standalone expressions that are understandable without needing words (e.g. nodding, smiling, affect bursts, etc...) .

The main purpose of the database is to be used to build human-agent interaction systems. Considering the efficiency of artifical intelligence in general and deep learning in particular, the dabase should be oriented, among other things, to deep learning applications.

2 Data Used

In order to answer deep learning systems needs, the ulitmate goal of this work is to obtain a large database of NCE. So the work presented here should be applied on different open-source and available databases of interactions. However, for now, we are using a dataset comprising audio and video recordings of dyadic conversations for which the topic was moral emotions (Heron et al., 2018). Moral emotions are emotions that are ethically relevant (Haidt, 2003) such as (gratitude, aw, empathy, shame, etc...). The setup of this dataset was made in a way to control the listener/speaker roles. Each of the participants was assigned randomly the role of the speaker or listener. The listener was told to ask the speaker predefined questions about moral emotions in the form *"When was the last time you felt ...?"*. The moral emotions in question were: shame, guilt, compassion and gratitude. Then the speaker/listener roles and questions were altered randomly until the questions for all emotions were asked. This way, the dataset provides data of speaker and listener expressions during a naturalistic interaction. The dataset contains 21 sessions (42 speakers) of 14 different nationalities. Each session containing 4 topics, one for each emotion asked. It is worth noting that due to the hardware setup (microphones and cameras) the data contain overlapping speech.

3 NCE Annotation

Intuition

As mentioned previously, the goal of this database it to help building human-agent interaction systems. Therefore, we consider that the data should be useful mainly for detection systems, decision making and generative systems.

So the annotations undertaken here will focus on localizing the start and end times of different NCE as accurately as possible. In this work, the functionality of the NCE are not considered. We consider only the event independently from the social function, intend/purpose of the expressions, situation or context. Two main reasons are behind this choice.

1. Annotating such contextual information would be a lot more challenging, tedious and time consuming than just delimitting the event. Indeed the values to be considered must be decided beforehand and more time will be required for each annotation. Also, such expressions might be dependent on the individual's culture, personality and even on the state of mind at the time of recording. Which are information for which the access is difficult and sometimes impossible especially if our ultimate goal is to obtain enough data for machine learning and deep learning systems.

2. Deep learning systems have already shown their ability to learn internal representations of the data and the task. So we hope that, with enough data such systems can be used to map specific NCE with specific context, situations and subject without requiring such annotation task.

With the NCE time intervals we will be able to train supervised machine-learning classifiers, build expression prediction systems for speakers/listeners and synthesis-by-concatenation systems like in (El Haddad et al., 2016b) and even audiovisual generative systems.

Annotation Scheme

Based on the literature related to several NCE, we consider, here, 4 different NCE in this work: smiles, laughter, head and eyebrow gestures. The criteria we used for this choice is are the fact that: i) they occur frequently in human-human interactions ii) they play a role in dialog strategies and phenomena like mirroring.

Expression	Values
Smiles	subtle, low, medium, high
Laughs	low, medium, high
Head movements	nod, shake, tilt
Eyebrow gestures	left/right/both raise/frown

Table 1: NCE annotation values

Indeed it has been shown in several previous separate work that these 4 expressions answer both of these criteria by happening frequently in dialogs and by being used for mirroring and other functionalities (Paggio and Navarretta, 2011b; Navarretta, 2016; Paggio and Navarretta, 2011a; Aubrey et al., 2013; McKeown et al., 2012; Dupont et al., 2016; Paggio and Navarretta, 2017; El Haddad et al., 2016a).

Each of the above-mentioned expressions will have descriptive values as shown in Table 1 and as detailed in what follows.

Smiles and Laughter: Both of these expressions have been the subject of many studies (El Haddad et al., 2016a). intensity or arousal is very important for both of these expressions. Indeed, in (McKeown and Curran, 2015) presents a study the relationship between laughter intensity and humor.

Concerning the smiles, the definition we are using is not focused on the lips movements alone. Several studies of the smile facial expressions can be found. Most of them agree that the Action Units (AU) corresponding to cheek raising (AU06) and lips spreading (AU12) respectively are important to consider (Ochs et al., 2017; Ekman and Friesen, 1982). But also lower eyelids raised (AU7), lips upside down (AU15) or pressing the lips together (AU24) have also been reported to be linked to smiling. But smiles can occur while speaking or while doing other facial expressions, for example, compressed smiles can be a combination of lips spreading (AU12) with turning the lips upside down (AU15) or pressing the lips together (AU24) (Harris and Alvarado, 2005; Ekman and Friesen, 1982). These facial expressions will therefore be used to determine the occurrence or not of a smile. Then, the smiles are segmented based on their intensity levels. The intensity is itself based on the intensity of the facial expressions used to deter-

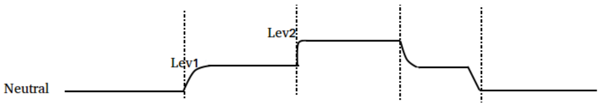

Figure 1: Example of segmentation of different level of smiles based on the intensity levels.

mined it was a smile. We define three different intensity levels (low, medium and high). One of the particularities of our annotation scheme is that we consider the smiles of very low level that seem to last ”all the time”. Chovile did not consider smiles in (Chovil, 1991), as smiles were so overwhelmingly frequently present in the data compared to other expressions. Similarly, many databases neglect these types of smiles. We decided to annotate them because they are part of the interaction and must have an effect since they can be perceived. So, we include a fourth level too: subtle (not related to the term used for micro-expressions). This is to annotated smiles of very low intensity which usually stay for a long period of time (and sometimes not) and to which it is sometimes hard to associate a specific AU or facial expression.

In order to have precise limits between two smiles, we rely on the transitions. Indeed, the work presented in (Schmidt et al., 2003) shows the importance of the speed of the transition from one expression to another. The choice of the intensity is somewhat subjective. The segments will start and end at the beginning of a level and the beginning of the next level respectively. An example is shown in Fig. 1

For laughs, the segments start when an audio, facial expression or body movement related to laughter is observed and stops when a breath intake is perceived whether audibly or visually (from the stomach, face, etc.). If no breath intake is perceived the end of the segment is considered to be when the movement stops.

Finally, we consider that these laughter and smiles cannot overlap: a laughter is not a smile and a smile with one of the movements mentioned above is a laugh.

Head and Eyebrow Movements: For head movements we consider nodding, shaking and tilting: pitch, yaw and roll movements respectively. The segments start and end with the movements. In the case of tilting, the annotations do not include the static head bent on the side after the movement has occurred. Only the movement is annotated. Considering the eyebrow movements, we annotate the raise and frown states of each or both eyebrows. Unlike the head movements, the annotations are not based on the movement only. The segments start when the movement starts and ends when the eyebrow is not perceived as raised or frowned anymore, taking the raised or frown state in between into account.

NCE	**Smiles**	**Laughs**	**HM**	**EM**	**All**
CKC	0.47	0.43	0.48	0.15	0.4

Table 2: Cohen’s Kappa Coefficients (CKC) to estimate inter-rater agreement

4 Inter-rater Agreement

Until now, 27 topics (part of a session) are annotated for the speaker and the corresponding listening in the dataset mentioned above, only 4 of which were annotated by 2 annotators. The total amount of time of 7 minutes and 11 seconds of data. Fig. 2 show examples of the obtained results for smiles, laughter, head and eyebrow movements for each of the annotators with respect to time. The integer values on the ordinate axis correspond to the intensity levels in case of the smiles and laughs (the lower the integer the lower the intensity (0 corresponding to neutral). In the case of head movements they correspond to nod (1), shake (2), tilt (3) and no movement (0). In the case of eyebrow movements 1 corresponds to raised (whether it is both eyebrows or only one), 2 to frown (none in this case) and 0 to no movement. The Cohen’s Kappa Coefficients were calculated to estimate the inter-rater agreement. The results mean values are given in Table 2.

Considering the complexity of the choice making and that part of the annotations were rather subjective an average Cohen’s Kappa of 0.4 is acceptable.

5 Future Work

After the dataset mentioned here is fully annotated we intend to use it to build NCE detection, prediction and generation systems. We also intend to carry on the annotations to other datasets as well.

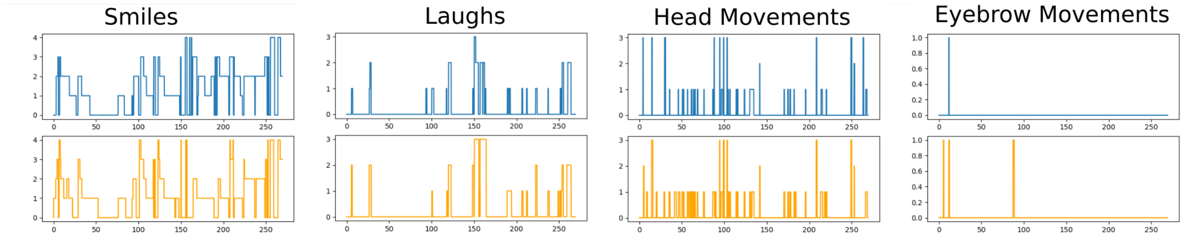

Figure 2: Annotations with respect to time for 2 annotators (blue and orange). The integers (1 to 4) correspond to the different annotation values corresponding to each expression mentioned in Table 1.

References

Andrew J Aubrey, David Marshall, Paul L Rosin, Jason Vandeventer, Douglas W Cunningham, and Christian Wallraven. 2013. Cardiff conversation database (ccdb): A database of natural dyadic conversations. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, pages 277–282. IEEE.

Nicole Chovil. 1991. Discourse-oriented facial displays in conversation. *Research on Language & Social Interaction*, 25(1-4):163–194.

Stéphane Dupont, Hüseyin Çakmak, Will Curran, Thierry Dutoit, Jennifer Hofmann, Gary McKeown, Olivier Pietquin, Tracey Platt, Willibald Ruch, and Jérôme Urbain. 2016. Laughter research: a review of the ilhaire project. In *Toward Robotic Socially Believable Behaving Systems-Volume I*, pages 147–181. Springer.

Paul Ekman and Wallace V. Friesen. 1982. Felt, false, and miserable smiles. *Journal of Nonverbal Behavior*, 6(4):238–252.

Kevin El Haddad. 2017. Nonverbal conversation expressions processing for human-agent interactions. In *Affective Computing and Intelligent Interaction (ACII), 2017 Seventh International Conference on*, pages 601–605. IEEE.

Kevin El Haddad, Hüseyin Cakmak, Stéphane Dupont, and Thierry Dutoit. 2016a. Laughter and Smile Processing for Human-Computer Interactions. In *Just talking - casual talk among humans and machines*, Portoroz, Slovenia.

Kevin El Haddad, Hüseyin Çakmak, Emer Gilmartin, Stéphane Dupont, and Thierry Dutoit. 2016b. 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*, ICMI 2016, pages 248–255, New York, NY, USA. ACM.

Jonathan Haidt. 2003. The moral emotions. handbook of affective sciences.

Christine Harris and Nancy Alvarado. 2005. Facial expressions, smile types, and self-report during humour, tickle, and pain. *Cognition and Emotion*, 19(5):655–669.

Louise Heron, Jaebok Kim, Minha Lee, Kevin El Haddad, Stephane Dupont, Thierry Dutoit, and Khiet Truong. 2018. A dyadic conversation dataset on moral emotions. In *Proceedings of the Workshop on Large-scale Emotion Recognition and Analysis, Face and Gesture*, China.

Gary McKeown, Roddy Cowie, Will Curran, Willibald Ruch, and Ellen Douglas-Cowie. 2012. Ilhaire laughter database. In *Proceedings of 4th International Workshop on Corpora for Research on Emotion, Sentiment & Social Signals, LREC*, pages 32–35. Citeseer.

Gary McKeown and Will Curran. 2015. The relationship between laughter intensity and perceived humour. In *The 4th Interdisciplinary Workshop on Laughter and other Non-Verbal Vocalisations in Speech, Enschede, Netherlands*, pages 27–29.

Costanza Navarretta. 2016. Mirroring facial expressions and emotions in dyadic conversations. In *LREC*.

Magalie Ochs, Catherine Pelachaud, and Gary Mckeown. 2017. A user perception–based approach to create smiling embodied conversational agents. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 7(1):4.

Patrizia Paggio and Costanza Navarretta. 2011a. Feedback and gestural behaviour in a conversational corpus of danish.

Patrizia Paggio and Costanza Navarretta. 2011b. Head movements, facial expressions and feedback in danish first encounters interactions: A culture-specific analysis. In *Universal Access in Human-Computer Interaction. Users Diversity*, pages 583–590, Berlin, Heidelberg. Springer Berlin Heidelberg.

Patrizia Paggio and Costanza Navarretta. 2017. The danish nomco corpus: multimodal interaction in first acquaintance conversations. *Language Resources and Evaluation*, 51(2):463–494.

Karen L Schmidt, Jeffrey F Cohn, and Yingli Tian. 2003. Signal characteristics of spontaneous facial expressions: Automatic movement in solitary and social smiles. *Biological Psychology*, 65(1):49–66.