Evaluating Temporal Predictive Features for Virtual Patients Feedbacks

Bruno Elias Penteado, Magalie Ochs, Roxane Bertrand, Philippe Blache

Evaluating Temporal Predictive Features for Virtual Patients Feedbacks

Bruno Elias Penteado*
brunopenteado@usp.br
Institute of Mathematical and Computer Sciences,
University of São Paulo
São Carlos, Brazil

Magalie Ochs
magalie.ochs@lis-lab.fr
Aix Marseille Univ, Université de Toulon, CNRS, LIS
Marseille, France

Roxane Bertrand†
roxane.bertrand@univ-amu.fr
Laboratoire Parole & Langage, Aix-Marseille Université
Aix-en-Provence, France

Philippe Blache†
philippe.blache@univ-amu.fr
Laboratoire Parole & Langage, Aix-Marseille Université
Aix-en-Provence, France

ABSTRACT

One key challenge to create believable embodied conversational agents (ECA) is to produce engaging behavior - and feedbacks (short verbal, vocal and gestural reactions produced when hearing the main speaker) play an important role. In this paper we propose a machine learning-based model for multimodal feedbacks. The goal is to learn, from a corpus of human-human interactions, when a virtual agent should display a feedback along with its type. And to be feasible, an important aspect is to be able to process them in real time, using reliable features. For this purpose, we used random forests with different features, using annotated corpora of task-oriented interactions. Our case study is the context of training doctors to break bad news to a patient (played by an actor or by the ECA). The performance of the method highlights the capacity to predict verbal and non-verbal feedbacks based on a small number of features characterizing temporal information, in particular, the silence and the position of the last feedback.

CCS CONCEPTS

• **Computing methodologies → Artificial intelligence**; *Intelligent agents*; Feature selection.

KEYWORDS

virtual agent; ECA; machine learning; backchannel prediction

ACM Reference Format:
Bruno Elias Penteado, Magalie Ochs, Roxane Bertrand, and Philippe Blache. 2019. Evaluating Temporal Predictive Features for Virtual Patients Feedbacks. In *ACM International Conference on Intelligent Virtual Agents (IVA '19), July 2–5, 2019, PARIS, France.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3308532.3329426

IVA '19, July 2–5, 2019, PARIS, France

ACM ISBN 978-1-4503-6672-4/19/07.
https://doi.org/10.1145/3308532.3329426

1 INTRODUCTION

In the intelligent virtual agent domain, several machine learning models have been proposed to automatically determine the feedbacks of virtual agents during an interaction, using human-human interaction datasets as training corpora and most commonly based on verbal and prosodic features [12, 20]. These approaches suppose an accurate system to automatically recognize speech and prosody. That makes the overall model's performance dependent on the individual performances of speech and prosody recognizers. As a consequence, one challenge remains to identify features that could be easily and accurately recognized during a human-machine interaction for predicting virtual agents' feedbacks in real time.

Several works have been conducted to implement into *ECA* the possibility to generate appropriate feedbacks during an interaction (e.g. [5, 19] - both of them express verbal and nonverbal feedbacks during interactions - e.g. "mhm" or head nodes). In [9] logistic regression was applied to predict verbal feedback in the context of simulations of counseling sessions (n=8), using prosody and linguistic features from the dialogues, in a 4 binary-classes approach after the end of each IPU (accuracy: 64.3%, precision, recall, and F1-score: 0.643), with a low recall for verbal feedbacks. Ruede et al. [17] applied LSTM networks to detect feedbacks based on acoustic features (power and pitch), in different time windows, in the context of telephone conversations (n=2348), with best results of precision=0.305 and recall=0.488 (F1-score: 0.375). Meena et al. [11] elicited prosodic, contextual, and syntactic features, with different combinations of machine learning algorithms. The context was an artificial task of a user describing a route to a computer, with 10 participants (2272 IPUs) and two possible outcomes (feedback, no feedback) at the end of each IPU (accuracy: 84.64%, using Naive Bayes classifier).

In this study, our objective is then to propose a set of features that could be used to learn the feedback model of an individual and, then, be replicated on a virtual agent.

2 FROM SEQUENCES TO TEMPORAL FEATURES LEARNING

2.1 Multimodal corpus

We used an audio-visual corpus comprised of real training sessions (avg. 15 min) between doctors, being trained, and patients (trained

actors[1] following a script) for breaking bad news. The corpus consists of 11 videos, with different patient-doctor dyads (119 min). The videos were automatically segmented into IPUs using SPPAS [2] and manually transcribed using PRAAT [3]. The doctors' and patients non-verbal behaviors have been manually annotated using ELAN [18]. The part-of-speech (POS) tags were automatically identified using MarsaTag [16].

Different gestures of both doctors and patients have been annotated: head movements, posture changes, gaze direction, eyebrow expressions, hand gestures, and smiles. More details on the corpus are presented in [13]. Three annotators - paid graduate students in linguistics - coded the corpus, with 5% double-checked for validation (Cohen's Kappa=0.63).

2.2 Segmentation and sequence extraction

Turn-taking is of great relevance to improving human-machine interaction [6] and often used to used to segment conversations. As such, we consider Inter-Pausal Units (IPU) as relevant for measuring speech production of speakers in a dialog. It corresponds to the speech of each speaker bounded by silent pauses. Thus, the videos were segmented into IPUs, supposing that the end of an IPU can be a potential completion of a turn where the interlocutor can provide a feedback response. In our corpus, there are 3882 doctor's IPUs, from the 11 dyads.

Each doctor's IPU may contain zero or many feedbacks from the patient. Thus, we considered the patient's feedbacks inside the IPUs to build *sequences* of signals by going through all the IPUs and considering : 1) all the doctor's signals that started during the IPU timespan, 2) all the doctor's signals that started before and ended after the IPU, 3) all patient's signals that started during the IPU and all patient's signals that occurred until 1s after the end of the IPU (derived from [7]). As some IPUs can be very long, we limited the total number of verbal tokens to the last five ones (*cf.* [15]) – the non-verbal signals did not present this limitation. As a result, we had a database of sequences, composed by a set 5547 sequences.

Each sequence, used here as unit of analysis, is therefore composed by a set of ordered doctor's verbal and non-verbal signals that can end by a patient's feedback. Based on studies on feedbacks, both in linguistics and in the domain of virtual agents [4, 6, 10], we have focused on the following feedback types: head movements, hand gestures, eyebrows movements, smiles, posture changes and gaze direction. Concerning the verbal feedbacks, we base our work on [14] that have constituted a list of French verbal expressions frequently used as feedback (e.g. "oui" (yes), "hmm", "euh", "d'accord" (ok), "non" (no) ...). Note that a sequence may have "no feedback" at all. Following Bertrand et al. [1], we only retained nouns, verbs, adjectives, and adverbs. We also added medical specific terms (identified based on a dictionary of French medical terms[2]). In this work, only the category of the words were used (nouns, verbs, adjectives, adverbs, medical terms), not its actual value.

2.3 Features creation

2.3.1 Feature representation. To capture the nuances regarding the temporal aspects, we developed a structured model, considering the following summarized features: 1) the last doctor's silent pause, 2) the first and last signals in the sequence. We chose these relational features because the sequences had variable lengths (avg. 4 signals, sd: 4.1), rendering difficult to model absolute positions properly. In some cases, where there is only one signal in the sequence, both carry the same values. For each of these features, we collected: its duration, the relative difference between its start and the feedback's start, the relative difference between its end and the feedback start, and the label of the signal (eg.: 'head movement', 'posture', etc.). The silent pause is an important cue for feedbacks [7] and it is also relatively a simple feature to detect automatically, in a reasonably controlled environment, so we also used it. In addition, we also considered information regarding the occurrence of the last patient's feedbacks, in two conditions: the last feedback that already ended and the last feedback that started. Both of them may refer to the same feedback, but there are sometimes, where there is multimodal feedback, where two or more feedbacks are elicited in conjunction (i.e. feedback may have started close to the reference feedback, but ended after). These features were selected due to the empirical observation that some feedbacks occur more spaced than others, which may occur in 'bursts', i.e., multiple times in a short period of time.

2.3.2 Features selection. A single feedback category is associated with each sequence (multiclass classification): verbal feedback, hand gesture, gaze direction, head movement, posture, eyebrows expression, smile, no feedback. In order to select the set of optimal features, we applied an iterative backward selection procedure, based on the accuracy, by removing the feature with the lowest attribute importance value given by the random forest algorithm in each iteration, until no more improvement is observed in the accuracy.

The optimal set of features comprised five features:

(1) The duration (in seconds) of the last doctor's silent pause;
(2) Time (s) since the last doctor's silent pause ended, relative to the start of the feedback;
(3) Time (s) since the last doctor's silent pause started, relative to the start of the feedback;
(4) Time (s) since the last patient's feedback started;
(5) Time (s) since the end of the last patient's feedback;

3 FEEDBACK PREDICTION BASED ON TEMPORAL FEATURES

We used three baselines to evaluate the model built in the previous section. Firstly, a baseline which always outputs the majority class (head movement, 48,4% of instances). In the second case, we consider only the presence or absence, encoded as 0 or 1, of the categories of doctor's signals in the sequence (11 categories). The third baseline considered the same modalities as columns, but for each one of them we considered continuous values for i) the duration of the signal, ii) the relative timespan from its start and the start of the sequence's feedback and iii) the relative timespan from its end and the start of the sequence's feeback (a total of 33 columns). The modalities not present in the sequence received null values.

For classification, Random Forest was applied equally to all cases (except for the majority class baseline) since it performs well in similar tasks, with multivariate data from symbolic time sequences

[1] For ethical reasons, it is not possible to videotape real breaking bad news situations.
[2] https://www.vocabulaire-medical.fr/

[8]. The proposed model and the baseline were cross-validated by 10 folds, independently of the dyads.

The corpus was very unbalanced. In this regard, we also tested a balanced version of the dataset, by considering a technique of cost-sensitive classifier, justified because we did not want to oversample or downsample the original corpus since it is already small.

Table 1 shows the performance of the proposed method and the baselines over the multimodal corpus. The selected features, in conjunction with the random forest classifier, improved significantly the results. The balancing of the corpus did not show improvements on the performance metrics, evidencing the ability of random forests to handle the unbalanced dataset.

Table 1: Cross-validation evaluation of the proposed method and the baseline ones on the human-human corpus.

	Accuracy	Precision	Recall	Kappa
Majority class	0.484	-	-	0
Presence-based	0.447	0.303	0.447	0.01
Rel. timespans	0.478	-[3]	0.478	0.03
Prop. model	**0.705**	**0.700**	**0.705**	**0.55**
Prop. method + balancing	0.518	0.593	0.518	0.45

The results show that the model proposed in this study can capture the variability of the feedbacks with reasonable accuracy (an improvement in accuracy of 57.7% over the presence-based algorithm and 45.6% over the majority class classifiers). This a good result considering the number of classes to be predicted (eight). However, some feedbacks were better predicted than others. In particular, the smile feedback and the 'no feedback' need further improvements in future works.

4 DISCUSSION

The main contribution of this study is to provide a set of simple and robust features for the prediction of virtual patient' feedbacks in breaking bad news. The selected features do not require complex processing for real-time detection, which depends a lot on the accuracy of the recognition algorithms and their execution time.

Although not directly comparable, the performance of the method proposed outperforms the existing ones, in simpler settings of classification. For example, Kawahara et al [9] reported precision and recall values of 0.643 to predict, using more complex linguistic and prosodic features for five classes. Meena et al [11] obtained 84.64% of accuracy in a binary classification (feedback or not) in an artificial task, using a large set of prosodic, syntactic and contextual features. The same classes were predicted, in the context of telephone conversations, with an F1-score of 0.375. These studies used the IPUs segmentation for the prediction of the feedback (binary or multiclass), using verbal feedbacks at the end of it. In this work we used a more fine-grained method, considering not only the IPUs but sequences inside it, which capture verbal and non-verbal cues which occur while the doctor is speaking. In this setting, we obtained an accuracy of 70.5% (precision: 0.7 and recall: 0.705, F-1 score: 0.691) in our multimodal corpus of natural interactions in the context of doctor-patient dialogue. To our knowledge, this is the first study that explores the temporality of the cues to predict the feedbacks. As future work, we plan to integrate the model into the virtual agent to test its performance during interactions, i.e. to ensure that the predictive model provides believable feedbacks behavior to the agent according to the user involved in an interaction with the virtual agent.

ACKNOWLEDGMENTS

This work has been funded by the French National Research Agency project ACORFORMED (ANR-14-CE24-0034-02) and supported by grants ANR-16-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI) and ANR-11-IDEX-0001-02 (A*MIDEX) and also STIC AMSUD-CAPES, 18-STIC-03, project EMPATIA. We also would like to thank Patricia Jaques Maillard and Seiji Isotani for the collaboration in this project.

REFERENCES

[1] Roxane Bertrand, Gaëlle Ferré, Philippe Blache, Robert Espesser, and Stéphane Rauzy. 2007. Backchannels revisited from a multimodal perspective. In *Auditory-Visual Speech Processing 2007, AVSP 2007*. 9.

[2] Brigitte Bigi. 2012. SPPAS: a tool for the phonetic segmentations of Speech. In *8th Intl. Conf. on Language Resources and Evaluation*. Istanbul, Turkey, 1748–1755.

[3] Paul Boersma and David Weenink. 2001. PRAAT, a system for doing phonetics by computer. *Glot International* 5, 9/10 (2001), 341–345.

[4] Lawrence J. Brunner. 1979. Smiles can be back channels. *Personality and Social Psychology* 37, 5 (1979), 728–734. https://doi.org/10.1037/0022-3514.37.5.728

[5] J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjálmsson, and H. Yan. 1999. Embodiment in Conversational Interfaces: Rea *(CHI '99)*. New York, NY, USA, 520–527.

[6] Starkey Duncan. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology* 23, 2 (1972), 283–292.

[7] Agustín Gravano and Julia Hirschberg. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech & Language* 25, 3 (2011), 601–634.

[8] Mathieu Guillame-Bert and Artur Dubrawski. 2014. Learning temporal rules to forecast events in multivariate time sequences. In *2nd Workshop on Machine Learning for Clinical Data Analysis, Healthcare and Genomics. NIPS*. Canada.

[9] Tatsuya Kawahara, Takashi Yamaguchi, Koji Inoue, Katsuya Takanashi, and Nigel G. Ward. 2016. Prediction and Generation of Backchannel Form for Attentive Listening Systems. In *INTERSPEECH*. San Francisco, USA, 2890–2894.

[10] Stefan Kopp, Jens Allwood, Karl Grammer, Elisabeth Ahlsen, and Thorsten Stocksmeier. 2008. Modeling Embodied Feedback with Virtual Humans. In *Modeling Communication with Robots and Virtual Humans*. 18–37.

[11] Raveesh Meena, Gabriel Skantze, and Joakim Gustafson. 2014. Data-driven models for timing feedback responses in a Map Task dialogue system. *Computer Speech & Language* 28, 4 (2014), 903–922.

[12] Louis-Philippe Morency, Iwan Kok, and Jonathan Gratch. 2010. A Probabilistic Multimodal Approach for Predicting Listener Backchannels. *Autonomous Agents and Multi-Agent Systems* 20, 1 (Jan. 2010), 70–84.

[13] Chris Porhet, Magalie Ochs, Jorane Saubesty, Grégoire de Montcheuil, and Roxane Bertrand. 2017. Mining a Multimodal Corpus of Doctor's Training for Virtual Patient's Feedbacks. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI '17)*. ACM, New York, NY, USA, 473–478. https://doi.org/10.1145/3136755.3136816

[14] Laurent Prévot, Brigitte Bigi, and Roxane Bertrand. 2013. A quantitative view of feedback lexical markers in conversational French. In *Proceedings of the SIGDIAL 2013 Conference*. Metz, France, 87–91.

[15] Laurent Prévot, Jan Gorisch, and Sankar Mukherjee. 2015. Annotation and Classification of French Feedback Communicative Functions. In *Conference on Language, Information and Computation*. Shanghai, China, 302–310.

[16] Stéphane Rauzy, Grégoire Montcheuil, and Philippe Blache. 2014. MarsaTag, a tagger for French written texts and speech transcriptions. In *Proceedings of Second Asian Pacific Corpus linguistics Conference*. Hong Kong, 220.

[17] Robin Ruede, Markus Müller, Sebastian Stüker, and Alex Waibel. 2017. Enhancing Backchannel Prediction Using Word Embeddings. In *INTERSPEECH*. 879–883.

[18] Han Sloetjes and Peter Wittenburg. 2008. Annotation by category: ELAN and ISO DCR.. In *IV Intl Conf on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco, 816–820. http://www.lrec-conf.org/proceedings/lrec2008/.

[19] Kristinn R. Thórisson. 1996. *Communicative Humanoids - A Computational Model of Psychosocial Dialogue Skills*. Ph.D. Dissertation. MIT, USA.

[20] K. P. Truong, R. W. Poppe, and D. K. J. Heylen. 2010. A rule-based backchannel prediction model using pitch and pause information. In *Proceedings of Interspeech*. Makuhari, Japan, 3058–3061.