An integrated model for predicting backchannel feedbacks

Philippe Blache, Massina Abderrahmane, Stéphane Rauzy, Roxane Bertrand

An integrated model for predicting backchannel feedbacks

Philippe Blache
LPL; CNRS & Aix-Marseille University
blache@ilcb.fr

Massina Abderrahmane
ESI, CNRS & Aix-Marseille University
fm_abderrahmane@esi.dz

Stéphane Rauzy
LPL; CNRS & Aix-Marseille University
stephane.rauzy@univ-amu.fr

Roxane Bertrand
LPL; CNRS & Aix-Marseille University
roxane.bertrand@univ-amu.fr

ABSTRACT

We present in this paper a method for generating in real time a great variability of multimodal backchannel feedbacks, increasing the naturalness of IVAs. The originality of the approach lies in its capacity to generate all types of features into a unique loop.

CCS CONCEPTS

• **Human-centered computing → HCI theory, concepts and models**.

KEYWORDS

Backchannels, embodied conversational agent, multimodal feedback, rule-based model

ACM Reference Format:
Philippe Blache, Massina Abderrahmane, Stéphane Rauzy, and Roxane Bertrand. 2020. An integrated model for predicting backchannel feedbacks. In *IVA '20: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA '20), October 19–23, 2020, Virtual Event, Scotland Uk.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3383652.3423948

1 INTRODUCTION

Backchanneling constitutes one the keys for intelligent virtual agents naturalness, which is closely related to their capacity in generating prompt and adequate feedbacks to the speaker's production. Several predictive models have been proposed, involving different cues from prosody, gestures, lexicon, syntax or semantics. Unfortunately, acquiring all such cues in real time, which is mandatory in the case of conversational agents [9], can be difficult. One simple solution consists in using temporal features [13]. However, we also need to determine backchannels' type: *generic* or *specific* (initially called *continuers* or *assessment*) [1] . These types involve different levels of processing, depending on whether they require or not deep understanding of the speaker's message. Some backchannels (esp. *generic*) are produced almost automatically and can be predicted by low-level cues. Some others require a certain level of semantic processing. The problem is that the production of these two types of backchannels is based on two different predictive mechanisms that are potentially in conflict because applied in parallel. We propose in this paper an approach avoiding this issue by implementing a *single-route backchannel predictive model*, generating appropriate backchannels in real time at a fine-grained level. The method we propose relies on a generic architecture. However, all dialogue systems being dependent from the task, we illustrate the approach with a specific application aiming at training human doctors to break bad news to virtual patients [11]. In this system, the virtual patient is mainly a listener, capable of answering questions, requesting for clarification and producing backchannels.

IVA '20, October 19–23, 2020, Virtual Event, Scotland Uk

ACM ISBN 978-1-4503-7586-3/20/09.
https://doi.org/10.1145/3383652.3423948

2 BACKCHANNELS TYPES

Two BC categories are usually distinguished: *generic* (displaying attention to the speaker) and *specific* (expressing responses to the content of the speaker's production) [1, 2, 15]. They correspond to different communicative functions and can be expressed in different modalities: verbal, visual or multimodal. Our IVA being developed for French, we mainly focus the most frequent BCs given in [**?**]: oui (*yes*) ah oui (*oh yes*), mh, d'accord (*agree*), ok, voilà (*that's it*), non (*no*), oh non (*oh no*), bon (*well*) ah bon (*is that so?*) (note that this list is comparable to that in English). In terms of types, *oui* and *mh* are typically generic whereas *oh non* or *d'accord* are specific.

Visual backchannels on their side correspond to many different types. They are mainly head movement (*nod, jerk, shake, tilt, turn, waggle*), facial expressions (*smile, laughter*) or eyebrow (*frowning, raising*). As noted in [7], they are less disruptive and in a large majority used as *generic* BCs.

Type	Modality	Backchannel
Generic	Verbal	oui, mh
	Visual	nod, smile
	Bimodal	nod+yeah, nod+ok, smile+ok
Specific / agreement	Verbal	oui, ah oui, d'accord, ok, voilà, bon
	Visual	nod, smile
	Bimodal	nod+yeah, nod+ok, nod+ooh
Specific / disagreement	Verbal	non
	Visual	shake
	Bimodal	shake+no, shake+mh
Specific surprise	Verbal	ah bon
	Visual	raising, frowning, tilt
	Bimodal	raising+no
Specific / fear	Verbal	non, oh non
	Visual	frowning, raising, shake
	Bimodal	shake+no, frowning+non

Figure 1: Types of backchannels

Bimodal backchannels (involving both visual and verbal productions) are also very frequent and play an important role [7]. [3] has specifically studied a subset of such bimodals for IVA, associated with their functions: *nod+yeah* (agreement), *shake+no* (disagreement), *smile+ok* (interest), *raise eyebrows+ooh* (understanding), etc. In some cases, bimodality can reinforce the function: for example, bimodal BCs show a stronger agreement than unimodal ones.

BC positions in the discourse also have an influence on their modality [7]: verbal BCs are preferably used during pauses, whereas visual BC are more likely to occur during speech. In terms of temporal features, verbal BCs are preferred soon after the beginning of a turn whereas visual BCs are usually produced later. We propose in this paper, beside the *generic* BC type, to precise the *specific* one by only using certain subtypes: *agreement* and *disagreement* (both being the most frequent in our corpus), completed with two other subtypes fitting with the specific purpose of our study: *surprise* and *fear*. The table 1 summarizes the different backchannels according to their modality and type.

3 BACKCHANNELS PREDICTIVE FEATURES

Many studies showed that the number of BCs has consequence on agent's naturalness. It is then important to identify as many BC triggers as possible for elaborating a natural model. In terms of temporal cues [13] suggests that a BC may occur in average every 6.5 seconds. Another direct placement model consists in generating feedbacks systematically during breaks. These models are very simple, but produce unnatural behaviors of the agent. We have proposed more elaborated approach taking into account the temporal context in the framework of ACORFORMed doctor/patient dialogues [12], based on the duration of the doctor's last silent pause, the duration since the doctor's last silent pause and the duration since last patient feedback. On their side, [13, 16] have presented more detailed mechanisms based on prosodic and duration features:

- After a region of pitch less than the 26th-percentile pitch level, lasting 110ms, after 700ms of speech, with no BC within the preceding 800ms, after 700ms wait.
- After a pause of 400ms preceded by at least 1000ms of speech, where the last 100ms contain a rising or falling pitch of at least 30Hz with no BC within the preceding 1400ms.

Several works [8, 10] propose models based in particular on different prosodic backchannelling cues such as final rising intonation, higher intensity level, higher pitch level, inter-pausal duration, etc.

We propose to use a selection of such features (that can be recognized in real time): *Prosody* (time since IPU began, silent pause for over 200ms), *Discourse* (last dialogue act), *Syntax* (type of the last POS bigram), *Semantic* (semantic type of the last entity).

Moreover, we also identified sequences of multimodal features (mainly POS tags and gestures) that could be used as backchannel predicting cues [14]. From these sequences, several rules have been extracted for generating backchannels:

```
doctor_head_nod ⇒ patient_head_nod
doctor_verb, doctor_head_nod, doctor_noun ⇒ patient_head_nod
doctor_adverb ⇒ patient_head_nod
doctor_medical_vocabulary ⇒ patient_head_nod
```

4 A PREDICTIVE RULE-BASED MODEL

Instead of a classical architecture based on two types of processing (an automatic loop generating low-level BC vs. a deeper one for semantic processing), we propose a unique model integrating both levels. This approach is not only more efficient, but also closer to a parallel human-like processing.

At the technical level, the different techniques now available in dialogue technology give access to many types of linguistic information in real time. More precisely, besides low-level classical cues (such as breaks, turn length, POS, etc.), we can also recognize during the dialogue semantic and discourse-level cues such as dialog acts, slot-based semantic structures as well as the position in the discourse ([2] shows how discourse phases influence BC types, generic BCs being favored when reaching the core of the narration). These different features all contribute to the generation of backchannels by deciding their type and placement. The interest of such an integrated model, on top of its homogeneity and its naturalness, is to avoid conflicts in the generation process, occurring when different backchannels can be generated simultaneously from two different loops.

Our proposal consists in integrating this BC generation model into an embodied conversational agent equipped with a dialogue system. The first question consists in identifying the span of each processing step that is traditionally a temporal window in the signal (e.g. each 3 seconds), an inter-pausal unit (i.e. between pauses of at least 200ms), a turn, etc. Embodied dialogue systems mainly have two input streams: the *audio* signal and its *transcription* (note that, in spite of the important role it can play, we put aside at this stage the gestural speaker's behavior in order to simplify the type of cues to be acquired in real time). The *audio stream* makes it possible to acquire temporal and prosodic features (silent pauses, pitch, IPU duration, etc.) where the *transcription stream* leads to the other linguistic cues. In this second stream, features from different domains can be acquired. First, at the *lexical level*, we have seen that morphosyntax (POS tags) plays an important role: for example, certain POS bigrams such as *V-Adv* favors the production of a *generic* BC. In the same vein, some terms (or more generally some semantic types) can be associated to a *specific* BC (surprise, fear, etc.). We also know that the introduction of new terms in the discourse (new referents) may be associate to a specific BC, indicating for example an agreement. At the discourse level, different studies have shown that the structure of the conversation and more precisely its phasing (opening, closing, argumentation, etc.) can also be associated with specific listener's reactions [2]. This information can be directly obtained thanks to *dialog act (DA) classification* [5, 6] : in the case of our application domain, DAs correspond to dialogue phases. It is thus possible to trigger specific BCs upon phase change (e.g. when the doctor starts announcing a bad new). Finally, *semantics* plays a central role in generating specific BCs: many listener's reactions are triggered upon instantiation of the semantic structure, which is represented in task-oriented dialogues by a common ground (CG) [4]. In such approaches, the CG consists in a set of structured frames made of different variables. The construction of the semantic structure consists in instantiating these variables during the discourse, which can elicit specific BCs (related to the content itself or to the global process of CG construction). We propose to implement a semantic-based BC generation by associating CG slots to specific BCs. For example, a BC expressing *fear* can be triggered when the slot "*urgency*" is instantiated.

The implementation of a unique loop for generating BC requires an appropriate segmentation of the input stream in order to take advantage of all different types of information we want to bring into the model. The processing span need to be large enough to capture semantic and discourse-level information, but not too long in order to allow real-time reactions. This means that neither arbitrary short segments nor entire IPUs returned by the speech recognition system

(that can last in some cases more than 20 seconds) are adequate for our model. We propose an intermediate segmentation approach for answering these needs consisting in segmenting the input flow by using discourse markers (*mais, donc, puis, alors*, etc.) that indicate approximately a change between different discourse units. The list of such items being closed and very short, the mechanism simply consists in checking whether a marker appears. In such case, the current segment is parsed and the linguistic processing tools extract in real time the different features such as POS tags, semantic types of the substantives, dialogue acts and slot instantiation of the common ground in order to generate adequate BCs. Beside them, the different audio features are kept updated, in particular the duration since the last feedback, the indication of the current state of the production (speech or silent pause), the duration of the speech since the last pause, the duration of the pause, etc. Note that in the type of data we are working on, most of BCs are triggered by linguistic cues, only few of them being generated by temporal features. As a consequence, we give them a priority in the algorithm, temporal cues being only used as secondary triggers. Concretely, during the speech, we first look at linguistic BC cues whereas during a pause, the BC generator is mainly based on pause duration. In the case where none of these contexts trigger any BC, then the most general temporal cue (duration since the last BC) is applied. The algorithm 1 presents the general mechanism for controlling BC generation.

Algorithm 1 Backchannel generation

```
if current_mode == speech then
    BC_type ← analyze_slot_filling(current_segment)
end if
if (BC_type is empty) then
    BC_cues ← analyze_segment(current_segment)
    BC_type ← apply_rules(BC_cues)
    if (BC_type is empty and last_BC_time > 6sec) then
        BC_type ← select_BC_type(current_mode, default)
    end if
end if
BC ← BC_select(BC_type, current_mode)
```

This algorithm renders possible to process all types of BC generation in a single loop. It distinguishes different situations, hierarchically ordered (avoiding a concurrent processing generating possible conflicts). At the higher level, we have seen that slot filling (i.e. instantiation of the semantic structure) may directly triggers backchannels. In such case, the BC type associated with the slot becomes the current BC type value (e.g. agreement, surprise, etc.). If no BC comes from slot filling, then the different BC cues are extracted from the analysis of the current segment (e.g. dialogue acts, semantic type, POS sequences, etc.). These cues serve as input to a BC type identification function, based on a set of production rules. Note that the cues as well as the rules can be applied to the audio signal as well as the transcription (in other words, takes into account prosody, discourse, syntax etc.). Figure 2 gives examples of such rules. Given the BC type to be generated and the current mode (pause or speech), the last step consists in generating the BC itself, by choosing among a list of possible candidates (as described in table 1). Note that this list is in a probability space which also depends on the current state (visual or bimodal BC will be preferred during speech where verbal BCs will be favored during pauses).

Level	Cue	BC type	Description
Prosody	`elapsed_time_pause > 200ms`	`generic`	
Discourse	`new_referent`	`specific`	Use of a new term
Syntax	`POS == [V,N | V,Adv]`	`generic`	The last previous POS
Semantics	`medical_term`	`generic`	When using a medical term
Semantics	`positive_emotion`	`agreement`	a positive emotion term triggers an agreement
Semantics	`negative_emotion`	`disagreement`	negative emotions trigger disagreement
Discourse	`DA == bad news`	`fear`	When the phase becomes "bad news"
Discourse	`DA == social interaction`	`generic`	After a social interaction

Figure 2: BC generation rules

5 CONCLUSION

We have presented in this paper a method for generating backchannels which is, at the difference with other approaches, integrated into a single loop. This approach takes advantage of different NLP technologies for a real time acquisition of different linguistic features involved into a high-level BC generating model. We have implemented this model into a virtual agent equipped with a dialogue system. The resulting application offers a very reactive artificial listener, capable of generating a large variety of multimodal BCs. This system is currently under evaluation.

REFERENCES

[1] J. Bavelas, L. Cates, and T. Johnson. 2000. Listeners as co-narrators. *Journal of Personality and Social Psychology* 79, 6 (2000).
[2] R. Bertrand and R. Espesser. 2017. Co-narration in French conversation storytelling: A quantitative insight. *Journal of Pragmatics* 111 (2017).
[3] E. Bevacqua, S. Pammi, S. J. Hyniewska, M. Schröder, and C. Pélachaud. 2010. Multimodal Backchannels for Embodied Conversational Agents. In *IVA-2010*.
[4] P. Blache. 2017. Dialogue management in task-oriented dialogue systems. In *International Workshop on Investigating Social Interactions with Artificial Agents*.
[5] Z. Chen, R. Yang, Z. Zhao, D. Cai, and X. He. 2018. Dialogue Act Recognition via CRF-Attentive Structured Network. In *SIGIR*.
[6] A. Stolcke et al. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics* 26, 3 (2000).
[7] G. Ferré and S. Renaudier. 2017. Unimodal and Bimodal Backchannels in Conversational English. In *Workshop on the Semantics and Pragmatics of Dialogue*.
[8] A. Gravano and J. Hirschberg. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech and Language* 25, 3 (2011).
[9] S. Kopp, H. van Welbergen, R. Yaghoubzadeh, and H. Buschmeier. 2014. An Architecture for Fluid Real-time Conversational Agents: Integrating Incremental Output Generation and Input Processing. *JMUI* 8 (2014).
[10] R. Meena, G. Skantze, and J. Gustafson. 2014. Data-driven models for timing feedback responses in a Map Task dialogue system. *Computer Speech and Language* 28, 4 (2014).
[11] M. Ochs, D. Mestre, G. Montcheuil, J.-M. Pergandi, J. Saubesty, E. Lombardo, D. Francon, and P. Blache. 2018. Training doctors' social skills to break bad news: Evaluation of the impact of virtual environment displays on the sense of presence. *Journal on Multimodal User Interfaces* (2018).
[12] B. Penteado, M. Ochs, R. Bertrand, and P. Blache. 2019. Evaluating Temporal Predictive Features for Virtual Patients Feedbacks. In *IVA-2019*.
[13] R. Poppe, K. P. Truong, D. Reidsma, and D. Heylen. 2010. Backchannel strategies for artificial listeners. In *IVA-2010*.
[14] C. Porhet, M. Ochs, J. Saubesty, G. de Montcheuil, and R. Bertrand. 2017. Mining a Multimodal Corpus of Doctor's Training for Virtual Patient's Feedbacks. In *International Conference on Multimodal Interaction (ICMI)*.
[15] J. Tolins and J. Fox Tree. 2014. Addressee backchannels steer narrative development. *Journal of Pragmatics* 70 (2014).
[16] N. Ward and W. Tsukahara. 2000. Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics* 32, 8 (2000).