Are you Smiling When I am Speaking?

Auriane Boudin, Roxane Bertrand, Magalie Ochs, Philippe Blache, Stéphane Rauzy

Are you Smiling When I am Speaking?

Auriane Boudin[1,2], Roxane Bertrand[1], Magalie Ochs[2], Philippe Blache[1], Stéphane Rauzy[1]
[1] LPL, CNRS & Aix-Marseille University
[2] LIS, CNRS & Aix-Marseille University
{auriane.boudin, roxane.bertrand, magalie.ochs, philippe.blache, stephane.rauzy, }@univ-amu.fr

Abstract

The aim of this study is to investigate conversational feedback that contains smiles and laughter. Firstly, we propose a statistical analysis of smiles and laughter used as generic and specific feedback in a corpus of French talk-in-interaction. Our results show that smiles of low intensity are preferentially used to produce generic feedback while high intensity smiles and laughter are preferentially used to produce specific feedback. Secondly, based on a machine learning approach, we propose a two-stage classification of feedback to automatically predict not only the presence/absence of a smile but, also the type of smile according to an intensity-scale (low or high).

Keywords: Conversational feedback, Smile, Laughter, Corpus study, Generic/Specific feedback

1. Introduction

During conversations, interlocutors switch dynamically between the role of speaker and listener. The speaker produces discourse, giving information to the listener who produces *feedback* (referred as FB)[1] to show his/her active listening (Schegloff, 1982) but also to contribute to the elaboration of the current discourse (Bavelas et al., 2000; Horton, 2017). FB promotes alignment between interlocutors, which allows the success of the interaction (Pickering and Garrod, 2013). FB production is also studied in part to render human-machine conversations more efficient (Glas and Pelachaud, 2015).

Following (Bavelas et al., 2000), **generic FB** (e.g. *"mhmh", "okay"*, nod) is used to show understanding, while **specific FB** (e.g. *"oh really", "that's so nice"*) is used to show assessment through diverse attitudes (Schegloff, 1982; Bavelas et al., 2000; Horton, 2017). Both generic and specific FB can be unimodal or bimodal (vocal *and/or* visual). In this work, we focus on FB that contains smiles and laughter (associated with verbalization and/or nods).

In this study, we first propose to explore how smiles and laughter are distributed according to the generic and specific FB dichotomy. Next, we present a two-stage classification to automatically predict smile in FB instances. The 1st stage of classification will predict whether a FB should be realized with a smile or a neutral face. The 2nd stage of classification will predict for FB with a smile, the intensity of the smile (high or low). We make use of the open-access corpus PACO (Amoyal et al., 2020) and Cheese! (Priego-Valverde et al., 2020) to investigate multimodal FB. Through a statistical analysis and a machine learning approach on a conversational corpus, we explore the 2 following hypotheses. (1) High intensity smiles are more salient in the discourse and should be preferentially used to show assessments or specific attitudes rather than understanding. Indeed, specific FB is generally more marked than generic FB. Consequently, **Neutral Faces (NF)** and **Low Intensity Smiles (LI Smiles)** should be preferentially used to produce generic FB while **High Intensity Smiles (HI Smiles)** and laughter should be preferentially used to produce specific FB. (2) Listeners are influenced by the main speaker behavior and tend to align during FB production by adopting similar conversational markers (e.g. same smile intensity). Given the mechanism of alignment, the prediction of the smile and the intensity of the smile during a FB realization should be derived from the smile annotation of the speaker (Heerey and Crossley, 2013). In consequence, we expect to observe an important quantity of FB produced by the listener with a smile intensity similar to the one expressed by the main speaker.

[1](also called *conversational feedback* or *backchannel*)

2. Related Works

Feedback shows the collaboration between a speaker and an interlocutor during interactions (Schegloff, 1982). According to (Bavelas et al., 2000), interlocutors can produce two types of FB: **generic** and **specific**. Generic FB shows understanding and is mostly realized with a nod and/or short vocalizations (e.g. *"yeah", "mhmh"*). On their side, specific FB is closely connected to the semantic content. It occurs once the common ground is established, when the listener has enough information to react with particular elements (wince, exclamation, rising tone) that can show *surprise, amusement, enthusiasm, etc.* (Tolins and Tree, 2014). Specific FB can be realized with variable elements such as *lexicalization, laughter, head movements, eyebrow movements, facial expressions, etc.*.

Following the generic/specific dichotomy, we propose a fine-grained classification for specific FB by adding two sub-levels (Boudin et al., 2021). The 1st level corresponds to the *polarity*: *positive* or *negative*. This polarity refers to the semantic content produced by the main speaker (e.g. a positive FB can respond to a fun story and a negative FB to a critic). The

2nd level concerns the *expected* or *unexpected* aspect of the information given to the listener. The expected/unexpected category refers to the transmission of information. The main speaker can refer to the common ground, i.e. the information already shared with the listener (expected) or she/he can also give new information to the listener (unexpected). These two levels of specific FB allow to classify different attitudes expressed by FB (enthusiasm, happiness, humor, compassion, embarrassment, critic). Within each sub-category (positive-expected, positive-unexpected, negative-expected, negative-unexpected), we infer that FB could be realized with some typical patterns (e.g. a rising intonation, with a smile and raised eyebrows for a positive-unexpected FB). In this work we focus on the different types of smiles used within each type and sub-type of FB.

To our knowledge, there is few systematic studies on smiling and its role as FB. Smiles have been identified as a part of FB form quite early (Brunner, 1979). (Duncan et al., 1979) observe that the listener's FB has a greater probability to be produced with a smile if the speaker is actually smiling. (Allwood and Cerrato, 2003) investigate FB functions and point out that smiles are frequently used to produce acknowledgments and clarification requests. Smiles can also show a reinforcement of a positive attitude. Among few studies, (Jensen, 2015) look at smiles and laughter as FB. In their data 33.3% of smiles and 18.6% of laughter is used as FB.

Note that as far as we know, there are few research works which attempt to predict automatically smiles in FB production. (Kok and Heylen, 2011) predict 3 types of smiles (*amused, polite* and *embarrassed*) during conversation with a Conditional Random Fields (CRF) algorithm. Four models are trained and evaluated to predict smiles and the type of smile. However, the prediction scores remain low, with f-score under 0.20.

(El Haddad et al., 2016) predict smiles and laughter FB with different intensity levels. A CRF model is trained accepting as input features laughter and smiles of different intensities produced both by the speaker and the listener. The predicted FB instances are implemented in a virtual agent and compared with different baselines. A subjective evaluation leads to satisfying and promising results.

3. Corpus & Method

PACO-Cheese! Corpus We used the French Cheese! and PACO corpora. They contain a total of 7 hours of audio-visual recording of 26 dyadic face-to-face interactions, lasting between 15 and 20 minutes. In the current work a subset of 13 dyads (3.6 hours) is used, on which instances of FB have been annotated.

The full set of available annotations is described in (Priego-Valverde et al., 2020; Amoyal et al., 2020; Boudin et al., 2021). Laughter has been manually annotated during the transcription process. In (Amoyal and Priego-Valverde, 2019; Rauzy and Amoyal, 2020; Amoyal et al., 2020), smiles have been annotated with 5 labels from the smile intensity scale (SIS) proposed by (Gironzetti et al., 2016): *S0* (neutral face), *S1* (close mouth smile), *S2* (open mouth smile), *S3* (wide open mouth smile), *S4* (laughing smile). S4 are mainly associated with vocal laughter. Regions between two vocal laughter could also be annotated S4 if the facial posture did not change. On their side, laughter is vocal elements that can be produced with neutral face or smile of lower intensity. Therefore, there is not a one-to-one correspondence between the two entities and we prefer to distinguish both laughter and S4 in the current work.

Instances of FB have been annotated following the 5 labels described in section 2: *generic*, *positive-expected, positive-unexpected, negative-expected* and *negative-unexpected*. They are reactions from one speaker to the other speaker's speech and can be composed of verbalization, nods, laughter and smile or a combination of these elements.

Logistic regression We used a Logistic Regression algorithm (Logit). The Logit models the probability that a FB occurs with a smile (and its associated intensity level). It allows to evaluate the specific contribution of each feature which facilitates the interpretation of the model. The Logit proves also to be relevant when dealing with small datasets. A binary classifier response is obtained from the Logit probability, by applying a probability threshold filter.

While in (Boudin et al., 2021) we aimed at predicting the position and the type of FB, in the current research work we consider that the position and the type of the FB is already known and we focus on the prediction of one element of the FB form: smile.

For that, we propose a two-stage classification where the 1st stage predicts the presence or absence of a smile in the FB form (993 FB with a neutral face and 1372 FB with a smile). The 2nd stage predicts the intensity of the smile (high or low) for the subset of FB containing a smile. In order to obtain balanced classes, the S1 and S2 smiles have been grouped as **LI Smiles** (361 FB with 'S1' and 284 FB with 'S2') and S3 and S4 as **HI Smiles** (188 FB with 'S3' and 539 FB with 'S4'). The dataset is composed of all the annotated FB with the associated smile used to produce it. When different smiles are used to produce the FB, we keep only the smile with the highest intensity. A prediction is correct if the item that composed the FB predicted matches with the item that composed the observed FB. A cross-validation has been obtained by running a Monte Carlo cross-validation (on 50 trials with a ratio 80%-20% for the training versus the evaluation sample) for both models. For comparison, two baseline models are computed that randomly predict the class according to the observed corpus frequency for the 2 distributions: smile/no-smile and LI/HI Smiles. Features are extracted from the speaker signal before the listener's FB. The subset of multi-modal features (a total of 16

features) is based on our previous analysis in (Boudin et al., 2021) to predict the FB type:

- Pause (presence or absence of silent pauses, before FB). Overlap (FB is produced during the speech of the main speaker) - Binary encoding (0: absence, 1: presence).
- Positive, Negative, Concrete tokens (that give potential cues about the FB sub-type) (Bonin et al., 2018) - Categorical encoding: counted since the last FB produced.
- Interjection, Discourse markers, Punctuation (Rauzy et al., 2014). Extracted in a previous window of 2 seconds and binary encoding. Number of tokens in the previous 2 seconds - Categorical encoding.
- Nod, Smiles (S1, S2, S3, S4, S0), Laughter - Extracted in a previous window of 2 seconds ; binary encoding.

4. Results & Discussion

4.1. Laughter and Smiles for FB Production

A total of 2,380 instances of FB was annotated: 1,207 generic and 1,173 specific, including 416 positive-expected, 550 positive-unexpected, 115 negative-expected, 92 negative-unexpected. During the 13 interactions, we report a total of 1215 'S0' , 1014 'S1', 944 'S2', 729 'S3', 798 'S4', among smiles 40% are used inside FB. 1051 laughter has been annotated (including 417 as FB).

20% of FB is produced with more than one intensity of smiles (e.g. *"yeah exactly"* that begins with a S0, continues with S1 and ends in S4). Among these instances of FB with particular smile's pattern, the majority (66%) shows an increasing smile intensity.

Figure 1 presents the smile intensities and laughter used to produce FB according to their generic/specific type[2] Figure 2 details the smiles and laughter for sub-types of specific FB.

All FB: Globally, 42.35% of FB is produced with a Neutral face (NF). S1, S2, S3 and S4 are equally used (27% for S1/S2, 30% for S3/S4). 17.52% of FB is produced with a laughter. Only 9.87% of FB is realized with a smile or a laughter alone. The rest of the time, FB is associated with verbalization, nods or others facial movements. Note that at least 71% of FB annotations in our corpus are multimodal[3].

Generic FB : NF (58.58%) is mostly used to produce generic FB. Regarding FB produced with a smile, the more the intensity of the smile increases, the more its use decreases. Generic FB rarely contains a laughter (1.74%).

[2]When several smiles are used, only the one with the highest intensity is counted.

[3]Our annotations did not contain eyebrow movements, nor other head movements than nods, nor facial expressions. With these annotations, the percentage of multimodal FB would be probably be higher

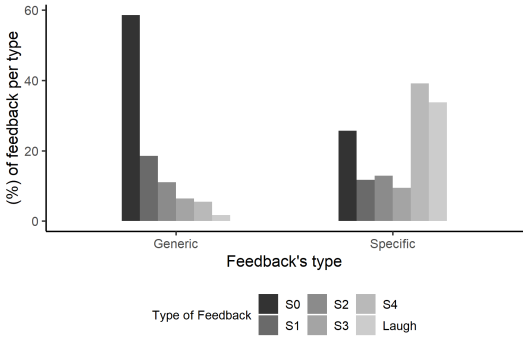

Figure 1: *Generic and Specific FB produced with Neutral face (S0), Smiles according to their intensity level (S1, S2, S3, S4) and laughter. When a FB contains plural smiles, only the highest intensity is kept.*

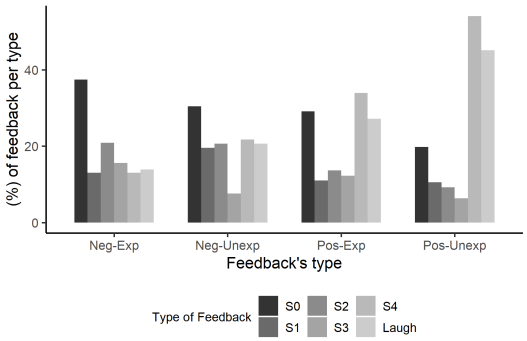

Figure 2: *Specific FB (Positive-Expected, Positive-Unexpected, Negative-Expected, Negative-Unexpected) produced with Neutral face (S0), Smiles according to their intensity level (S1, S2, S3, S4) and laughter. When a FB contains plural smiles, only the highest intensity is kept.*

Specific FB: Only 25.66% of specific FB is produced with a NF. 24.55% are produced with a LI Smile. 48.65% of specific FB is produced with a HI Smile. 33.76% contain a laughter. Laughter and HI Smiles are more present for positive FB, specifically for unexpected ones compared to negative FB. Concerning negative FB, NF is preferentially used, especially for the expected ones. Nonetheless, as we expected, smiles are still present for negative FB since smiles can be used to show embarrassment or compassion.

These observations confirm our 1st hypothesis: NF and LI Smiles are mainly used to produce generic FB whereas LI Smiles and laughter are mainly used to produce specific FB. These observations support our typology of FB, particularly useful to characterize the form of FB.

4.2. Speaker & Listener alignment

There are various ways to evaluate alignment between interlocutors (Rauzy et al., 2022). Herein, we focus on the alignment between the listeners and the speakers by looking at the smiles and laughter produced both as FB (by the listener) and as features (by the speaker in a window of 2s before the FB). For each level of smiles defined above, we compute 3 quantities: the proportion P_{FB} of FB containing the given level among all

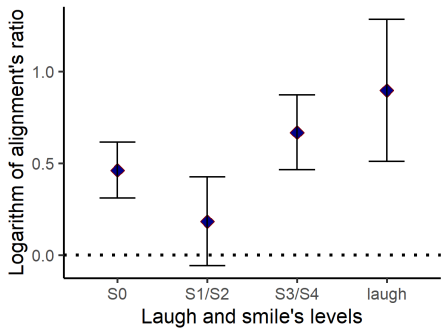

Figure 3: *Logarithm of alignment's ratio for NF (S0), LI Smile (S1/S2), HI Smile (S3/S4) and Laughter between the speaker and the listener.*

the FB, the proportion P_{Feat} of features containing the given level and the proportion $P_{FB/Feat}$ of pair of features/FB containing conjointly the given level. The first two quantities allow to compute the proportion of co-occurrences by chance. We define the alignment's ratio as the ratio of the observed proportion $P_{FB/Feat}$ to the proportion expected by chance $P_{FB} \times P_{Feat}$. Figure 3 presents the logarithm of the alignment's ratio and its associated 2σ standard error bars for NF, LI Smiles, HI Smiles and Laughter.

We observe a significant alignment for all the group except for the LI Smiles, particularly for laughter and HI Smiles. These results confirm our 2nd hypothesis about alignment, except for LI Smiles. Nonetheless, LI Smiles are visually more subtle, which can explain that they are less employed in the alignment strategy.

4.3. Logit

The 1st model predicts smiles in FB. The 2nd model predicts the smile intensity. The performances and the selected features are presented in Table 1 and 2.

Smile prediction: The 1st model provides accurate performances, significantly better than the baseline (t-test provided a p-value < 0.001). All smiles intensity levels are selected as features by the Logit and multimodal features appear significant. To estimate the importance of multi-modality, we test the models with only smiles features. For the 1st model, a t-test (p-value < 0.05) confirms that multimodal features perform better than smile features alone.

Smile intensity prediction: The 2nd fine-grained model gives reliable scores, better than the baseline (t-test provided a p-value < 0.001). Only NF and HI Smiles are selected, which are the most extreme smile intensity. This suggests that the most salient markers produced by the main speaker are the most informative for choosing the smile intensity. Removing the other multi-modal features does not significantly alters the performance obtained when using only smile features. These results suggest that not only smiles but also contextual parameters (speaker activity and semantic polarity) are relevant to decide whether a FB should be produced with a smile or not. Once the listener has decided if a smile will compose his/her FB, NF and HI

Pred	F	P	R
Smile	**0.72**	0.83	0.64
Smile Baseline	0.57	0.57	0.57
Intensity	**0.66**	0.72	0.61
Intensity Baseline	0.48	0.48	0.48

Table 1: F-score (F), Precision (P) and Recall (R) for the two predictive models and their baseline.

Pred	Features
Smile	S4, S3, S1, S0, S2, Overlap, Laughter, Pause, Positive Token
Smile intensity	S4, Overlap, Discourse marker, S0, S3

Table 2: Features selected by the *Logit* for the two classification tasks: smile/non smile and LI Smile/HI Smile prediction. Features presented are those selected by the Logit and ranked by their order of importance.

Smile are sufficient enough to choose the smile intensity, through mechanisms of alignment. Finally, these results are in line with our 2nd hypothesis, indicating that the smiles from the speaker are a good predictor of the smiles produced by the listener.

5. Conclusion

In this work, we focused on smiles and laughter as conversational FB in French face-to-face conversation. The data reveal that neutral faces (NF), Low Intensity Smiles (LI Smiles) and High Intensity Smiles (HI Smiles) are used to produce both generic and specific FB. Nonetheless, some trends emerge. Our analysis highlights that generic FB is preferentially produced with NF and LI Smiles, while specific FB, especially positive FB, are preferentially produced with laughter and HI smiles. The same behavior is observed for unexpected FB. For negative FB the trend in the different intensity of smiles stays unclear and need deepest investigations. To better understand it, we could analyse the smiles functions (e.g. embarrassment, compassion, showing sympathy) (Hoque et al., 2011; Mazzocconi et al., 2020). Alignment between the speaker and the listener is measured for NF, HI Smiles and laughter. Laughter is the behavior that is the most reproduced by the listener when it is produced by the speaker. Finally, we presented a hierarchical classifier method to predict smiles and their intensity for FB production, that obtains reliable performances. The model also indicates that the smile intensity features play an important role in the prediction which confirms our results on alignment. The current work come along with a larger project about the prediction of the FB position and the type of FB. Ultimately, it will provide a complete model including the prediction of localization, types and multimodal component of FB allowing the implementation in an effective dialog system, see for example (El Haddad et al., 2016).

6. Acknowlodgments

Research supported by grants ANR-16-CONV-0002 (ILCB) and the Excellence Initiative of Aix-Marseille University (A*MIDEX).

7. Bibliographical References

Allwood, J. and Cerrato, L. (2003). A study of gestural feedback expressions. In *First nordic symposium on multimodal communication*, pages 7–22. Copenhagen.

Amoyal, M. and Priego-Valverde, B. (2019). Smiling for negotiating topic transitions in french conversation. In *GESPIN-Gesture and Speech in Interaction*.

Amoyal, M., Priego-Valverde, B., and Rauzy, S. (2020). Paco: A corpus to analyze the impact of common ground in spontaneous face-to-face interaction. In *Language Resources and Evaluation Conference*.

Bavelas, J. B., Coates, L., and Johnson, T. (2000). Listeners as co-narrators. *Journal of personality and social psychology*, 79(6):941.

Bonin, P., Méot, A., and Bugaiska, A. (2018). Concreteness norms for 1,659 french words: Relationships with other psycholinguistic variables and word recognition times. *Behavior research methods*, 50(6):2366–2387.

Boudin, A., Bertrand, R., Rauzy, S., Ochs, M., and Blache, P. (2021). A multimodal model for predicting conversational feedbacks. In *International Conference on Text, Speech, and Dialogue*, pages 537–549. Springer.

Brunner, L. J. (1979). Smiles can be back channels. *Journal of personality and social psychology*, 37(5):728.

Duncan, S., Brunner, L. J., and Fiske, D. W. (1979). Strategy signals in face-to-face interaction. *Journal of Personality and Social Psychology*, 37(2):301.

El Haddad, K., Çakmak, H., Gilmartin, E., Dupont, S., and Dutoit, T. (2016). 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 '16, page 248–255, New York, NY, USA. Association for Computing Machinery.

Gironzetti, E., Attardo, S., and Pickering, L. (2016). Smiling, gaze, and humor in conversation: A pilot study. In Leonor Ruiz-Gurillo, editor, *Metapragmatics of Humor: Current research trends*, pages 235 – 254.

Glas, N. and Pelachaud, C. (2015). Definitions of engagement in human-agent interaction. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 944–949. IEEE.

Heerey, E. A. and Crossley, H. M. (2013). Predictive and reactive mechanisms in smile reciprocity. *Psychological Science*, 24(8):1446–1455.

Hoque, M., Morency, L.-P., and Picard, R. W. (2011). Are you friendly or just polite?–analysis of smiles in spontaneous face-to-face interactions. In *International Conference on Affective Computing and Intelligent Interaction*, pages 135–144. Springer.

Horton, W. S. (2017). Theories and approaches to the study of conversation and interactive discourse. In *The Routledge handbook of discourse processes*, pages 22–68. Routledge.

Jensen, M. (2015). Smile as feedback expressions in interpersonal interaction. *International Journal of Psychological Studies*, 7(4):95–105.

Kok, I. d. and Heylen, D. (2011). When do we smile? analysis and modeling of the nonverbal context of listener smiles in conversation. In *International Conference on Affective Computing and Intelligent Interaction*, pages 477–486. Springer.

Mazzocconi, C., Tian, Y., and Ginzburg, J. (2020). What's your laughter doing there? a taxonomy of the pragmatic functions of laughter. *IEEE Transactions on Affective Computing*, pages 1–1.

Pickering, M. J. and Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and brain sciences*, 36(4):329–347.

Priego-Valverde, B., Bigi, B., and Amoyal, M. (2020). "cheese!": a corpus of face-to-face french interactions. a case study for analyzing smiling and conversational humor. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 467–475.

Rauzy, S. and Amoyal, M. (2020). SMAD: a tool for automatically annotating the smile intensity along a video record. In *HRC2020, 10th Humour Research Conference*.

Rauzy, S., Montcheuil, G., and Blache, P. (2014). Marsatag, a tagger for french written texts and speech transcriptions. In *Second Asian Pacific Corpus linguistics Conference*, pages 220–220.

Rauzy, S., Amoyal, M., and Priego-Valverde, B. (2022). A measure of the smiling synchrony in the conversational face-to-face interaction corpus PACO-CHEESE. In *SmiLa Workshop, Language Resources and Evaluation Conference*.

Schegloff, E. A. (1982). Discourse as an interactional achievement: Some uses of 'uh huh'and other things that come between sentences. *Analyzing discourse: Text and talk*, 71:71–93.

Tolins, J. and Tree, J. E. F. (2014). Addressee backchannels steer narrative development. *Journal of Pragmatics*, 70:152–164.