A multimodal model for predicting feedback position and type during conversation

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

Speech Communication 159 (2024) 103066

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

A multimodal model for predicting feedback position and type during conversation

Auriane Boudin [a,b,c,*], Roxane Bertrand [a,c], Stéphane Rauzy [a,c], Magalie Ochs [b,c], Philippe Blache [a,c]

[a] Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France
[b] Aix Marseille Univ, CNRS, LIS, Marseille, France
[c] Institute of Language, Communication and the Brain (ILCB), France

ARTICLE INFO

Keywords:
Feedback
Multimodality
Linguistic interaction
Statistical model
Corpus study

ABSTRACT

This study investigates conversational feedback, that is, a listener's reaction in response to a speaker, a phenomenon which occurs in all natural interactions. Feedback depends on the main speaker's productions and in return supports the elaboration of the interaction. As a consequence, feedback production has a direct impact on the quality of the interaction.

This paper examines all types of feedback, from generic to specific feedback, the latter of which has received less attention in the literature. We also present a fine-grained labeling system introducing two sub-types of specific feedback: *positive/negative* and *given/new*. Following a literature review on linguistic and machine learning perspectives highlighting the main issues in feedback prediction, we present a model based on a set of multimodal features which predicts the possible position of feedback and its type. This computational model makes it possible to precisely identify the different features in the speaker's production (morpho-syntactic, prosodic and mimo-gestural) which play a role in triggering feedback from the listener; the model also evaluates their relative importance.

The main contribution of this study is twofold: we sought to improve 1/ the model's performance in comparison with other approaches relying on a small set of features, and 2/ the model's interpretability, in particular by investigating feature importance. By integrating all the different modalities as well as high-level features, our model is uniquely positioned to be applied to French corpora.

1. Introduction

To understand the mechanisms of conversation, one must know how interlocutors exchange information in the perspective of mutual understanding. Several studies (Clark, 1996; Garrod and Pickering, 2004; Pickering and Garrod, 2021 among others) have shown that interlocutors exchange information by adapting their linguistic behavior and progressively aligning to each other, adopting similar productions, planning their turn-taking, providing feedback, and implementing transitions between topics, etc. Interlocutors do so by using signals produced by their partners to predict the type and the time of their production during the interaction. This prediction mechanism is critical for the alignment of interlocutors' linguistic representations (Pickering and Garrod, 2013; Pickering and Garrod, 2021; Gandolfi et al., 2023). Feedback (Schegloff, 1982), also called backchannels (Yngve, 1970), is one of the most important phenomena for studying alignment as well as the quality of the interaction. Feedback is usually described as a brief signal produced by the listener in response to the speaker's discourse. Feedback can be verbal (*yes, ok, etc.*), vocal (e.g., *mhm*), and/or gestural (head movements, eyebrows movements, smiling). By using feedback, listeners give speakers information about their awareness, comprehension and appreciation (Allwood et al., 1992; Bunt, 1994), helping them in the elaboration of the dialogue. Dialogue studies underline the crucial role of feedback for updating shared knowledge (*common ground*) (Clark, 1996; Horton, 2017) and promoting alignment. Following (Bavelas et al., 2000), we distinguish between two types of feedback: *generic* and *specific*. The former are items such as *mh* or nods which express interest and understanding, while the latter are more complex (verbal utterances, tone of voice, eyebrows movements, etc.) and involve an evaluative function.

* Corresponding author at: 5 Avenue Pasteur, 13100, Aix-en-Provence, France.
E-mail address: auriane.boudin@univ-amu.fr (A. Boudin).

https://doi.org/10.1016/j.specom.2024.103066
Received 15 June 2022; Received in revised form 15 March 2024; Accepted 23 March 2024
Available online 24 March 2024

Several studies have shown that feedback can be triggered by different multimodal cues (also called *inviting cues* or *inviting features*) from the speaker's production (Koiso et al., 1998; Ward and Tsukahara, 2000; Allwood and Cerrato, 2003; Gravano and Hirschberg, 2011). Previous research in the domain of feedback prediction has provided valuable insights, focusing primarily on *verbal* (Okato et al., 1996; Ward and Tsukahara, 2000; Cathcart et al., 2003; Skantze, 2017) or *gestural* feedback (Morency et al., 2010; Ozkan and Morency, 2010). Only a few studies have taken into account both gestural and verbal feedback prediction (Fujie et al., 2004; De Kok et al., 2010). Moreover, existing models usually only focus on the most general type of feedback, such as *mh, yeah,* or nodding. Recently, some studies have examined more complex feedback (Kawahara et al., 2016; Ortega et al., 2020; Jang et al., 2021). These models aim at predicting how and when feedback is produced.

This study focuses on all types of feedback (generic and specific) and their forms (verbal, vocal and gestural), based on multimodal cues from the main speaker. By bringing these aspects together, we aim to provide a more comprehensive understanding of feedback and the conditions under which a listener produces it. Alongside this theoretical and descriptive goal, the prediction of feedback is also crucial in the perspective of human-machine interaction: the production of appropriate feedback plays a central role in the perception of artificial agents (Poppe et al., 2010; Poppe et al., 2011; Truong et al., 2011; Glas and Pelachaud, 2015).

Modeling feedback involves different issues. One is building corpora enriched with multimodal annotations of both the listener's feedback and the main speaker's inviting features. Additionally, we lack a precise taxonomy for feedback classification. In this perspective, we propose refining the description (and the prediction) of specific feedback by introducing a novel taxonomy taking into consideration the stance of the main speaker's discourse (positive/negative) and the information structure of the feedback scope (given/new). Previous studies underline the significance of the semantic and pragmatic context in feedback production (Allwood et al., 1992; Prévot and Gorisch, 2014). The speaker provides cues about their stance, referring to their explicit or implicit affective treatment of the event (Stivers, 2008). These cues help the listener to react, implementing alignment and affiliation (Stivers, 2008; Ruusuvuori and Peräkylä, 2009). Furthermore, (Allwood et al., 1992) argues that the information status (*new* or *given*) is a key parameter that influences the context of feedback production. Overall, feedback exhibits distinct features when responding to a positive/negative stance, new/given information, etc. By proposing a classification system based on the valence of the main speaker's speech and the information structure of the feedback scope, we take into account important aspects that may influence the way feedback is produced by the listener.

The first goal of our article involves the prediction of the potential sites for feedback. Our model aims to capture feedback-inviting features associated with each type of feedback. The second goal is to classify feedback according to general types (generic and specific) and sub-types (negative/positive; given/new). To the best of our knowledge, this is the first approach attempting to predict both feedback positions and precise types.

As multimodality is a key aspect of language (Wildfeuer et al., 2020), we explore this dimension by bringing together as many modalities as possible. Of course, some modalities can play a more important role than others, a role which can also depend on the context. In a natural environment, visual, acoustic and verbal cues interact. Our goal is to explore the relative importance of different cues from different modalities in a natural context.

Adopting this approach leads to two main contributions: 1/ using a wide set of features improves the performance of the model, in comparison with other approaches relying on a small set, and 2/ examining the importance of features improves the interpretability of the model. This latter contribution comes from the fact that we use linguistically motivated features, making the interpretation transparent.

The paper is organized as follows: in Section 2, we present the definitions of the types of feedback and the inviting cues which have been identified in the literature. Section 3 outlines the main feedback predictive models proposed in the literature, distinguishing two types of tasks: continuous prediction in 3.1 and non-continuous prediction in 3.2. Section 4 presents the French multimodal corpora PACO and Cheese! (Amoyal et al., 2020; Priego-Valverde et al., 2020), with some supplementary annotations we conducted and the extracted features. Our feedback taxonomy is presented in Section 5, along with feedback annotation and an analysis of the components of feedback by type. Our models are then presented in Sections 6 and 7, with the description of the method we used, and our results and discussions for each type of task (continuous prediction and feedback type classification). Finally, Section 8 discusses our main results and presents our perspectives.

2. Feedback and predictive cues: theoretical background

Before delving into the taxonomy and modeling aspects of our study, we will further define the notion and the nature of feedback and identify the different types of verbal and non-verbal features from the main speaker's productions that have been proposed in the literature as predictive cues for feedback production.

2.1. Feedback functions

The notion of *feedback* was introduced in the perspective of underlining the collaboration between a speaker and an interlocutor during an interaction (Yngve, 1970; Schegloff, 1982). Feedback is multimodal and can take on different forms, including nodding, smiling, laughter, short verbalizations, facial expressions, eyebrow movements, hand gestures, etc. In his influential contribution, (Schegloff, 1982) highlights the different functions of feedback. Firstly, feedback can play the role of *continuers*, showing an interlocutor's interest and comprehension of the communicative situation. Feedback can also show disinterest, for example when a listener repeats the same item many times without any change. Secondly, feedback can also express a reaction (surprise, disgust, happiness) related to the semantic content. In this case, items are more variable, generally longer and more often lexicalized than *continuers* (e.g., *"oh my god," "oh wow," "really"*) and perform different functions such as *assessment, acknowledgment*, etc.

In another important contribution, (Bavelas et al., 2000) introduced a typology of feedback responses, pointing out the high level of contributions of all participants during interactive narration. The authors showed how such linguistic and para-linguistic behavior on the part of the listener has an impact on the speaker's productions. According to the authors, interlocutors can produce two types of feedback: **generic** and **specific** feedback, fairly close to the continuer – assessment distinction of Schegloff (1982). Generic responses preferentially occur in the first part of the narration, within the set-up phase and the construction of common ground. Such responses mostly include nodding and/or short vocalizations (e.g., *"yeah," "mhmh," "okay"*). Specific responses are closely connected to the semantic context. They tend to occur later in the narration, once common ground has been established. Therefore, the interlocutor has enough information to react using a particular item (wincing, exclamations, or rising tones) that can show surprise, amusement, enthusiasm, etc. Several studies (Stivers, 2008; Tolins and Fox Tree, 2014; Bertrand and Espesser, 2017) have confirmed the relevance of this typology.

2.2. Feedback predictive cues

Producing an appropriate response requires the identification of specific cues. Several studies have been carried out to establish these inviting cues for feedback responses. In the following section, we summarize the results on the main features of each modality.

Prosodic cues: (Ward, 1996) was one of the first studies to focus on

Upon detection of:
(P1) a region of pitch less than the 26th-percentile pitch level and
(P2) continuing for at least 110 milliseconds,
(P3) coming after at least 700 milliseconds of speech,
(P4) providing you have not output back-channel feedback within the preceding 800 milliseconds,
(P5) after 700 milliseconds wait,
you should produce back-channel feedback.

Fig. 1. (Ward and Tsukahara, 2000) hand-crafted rules for feedback generation.

Table 1
Summary of feedback and no-feedback predictive cues listed in the state of the art according to feature type: **Prosodic, Lexico-syntactic** and **Mimo-gestural** feature types and language investigated: **English** (EN), **Japanese** (JP), **Spanish** (SP), **Slovak** (SL), **French** (FR), **Dutch** (DU).

Cue	Feedback	No Feedback
Prosodic	Falling (flat/rise-fall) (JP, DU) Rising (high/low-rise (EN, SP, SL, FR, DU) Flat intonation (SP, FR) High peak of energy (JP) Final vocalic lengthening (EN, SP, SL) Low pitch regions (EN, JP) Long IPU duration (EN, SP, SL) High speech rate (EN, SP, SL) Low noise-to-harmonics ratio (EN) Pause > 400 ms (EN, JP, DU) High intensity mean (EN)	Flat intonation (EN, SL, JP) Low peak of energy (JP) Short duration of the final phoneme (JP)
Lexico-syntactic	POS bigram: Det-NN; Adj-NN; NN—NN (EN, FR) Connective close (EN, JP) Disfluencies (EN, JP) Final particles (EN, JP) Adverbs (FR) Determiners; Interjections; Pronouns (EN)	Determiners (FR) Interjections (FR) Conjunctions (FR) Speech markers (FR)
Mimo-gestural	Speaker looks at the interlocutor (EN, FR) Nods (EN)	

the prosodic level of feedback-inviting cues, using English and Japanese. The author demonstrated the importance of a low-pitch region in the speaker's speech for a listener to identify the appropriate place to generate feedback. This period lasts at least 110 ms. In addition, (Ward and Tsukahara, 2000) proposed rules for predicting feedback by identifying five conditions, as described in Fig. 1. Moreover, the authors assumed that low-pitch regions are associated with no new information, which tends to favor the production of generic feedback.[1]

(Koisoet al., 1998) investigated the role of intonation, energy, and duration cues. They found falling, flat-fall, and rise-fall F0 patterns, a late decrease in energy, and high peaks of energy to be relevant predictive cues in Japanese. However, they also found that a flat F0 pattern, a short duration of the final phonemes, no decrease in energy, and a low peak of energy discouraged feedback production.

(Cathcart et al., 2003) found that speaker pauses longer than 600 ms indicated a relevant place to produce feedback in English. They obtained better results by associating pauses with the three most frequent POS trigrams. Other studies also found that pauses longer than 500 ms (Terrel and Multu, 2012) in English and 400 ms (Poppe et al., 2010; Truong et al., 2010) in both English and Dutch favor the production of feedback. The latter study obtained better results by associating pauses with low-pitch regions in English and with rising or falling pitch in Dutch.

Feedback predictive cues found in English by (Gravano and Hirschberg, 2011) were final-rising intonation (high rise and low rise) over the last 200–300 ms of the previous IPU (Inter Pausal Unit), IPUs with longer duration, a lower noise-to-harmonics ratio in the last 500–1000 ms of the previous IPU and a higher pitch and intensity mean.

A more recent study by (Brusco et al., 2020) confirmed the previous results for English, Argentine Spanish and Slovak. The authors found that the following prosodic cues predicted feedback: longer IPU duration, final-word lengthening, higher speech rate, high-rising final intonation, plateau intensity level (only in Spanish), and the noise-to-harmonics ratio.

Morpho-syntactic cues: (Ward and Tsukahara, 2000) highlighted the role of utterance endings, completion of the grammatical clause, clause connectives, disfluency markers, and sentence final particles (e.g., *"you see"*) that seemingly favored feedback production in both English and Japanese. Moreover, they noted that these elements often co-occur with low-pitch regions.

(Bertrand et al., 2007) indicated that gestural feedback in French appears most of the time after a noun, a verb, or an adverb, but rarely after a determinant. Similar results on vocal feedback were obtained for English by (Gravano and Hirschberg, 2011): determinant-nouns, adjective-nouns, and noun-nouns seemed to be significantly more present before feedback. Finally, (Ozkan and Morency, 2012) found that nouns and verbs play a significant role in predicting feedback in English. They also found that pronouns, interjections and determiners are significant, which contrast with the findings of (Bertrand et al., 2007), who did not identify a significant role for determiners and interjections in French.

Mimo-gestural cues: (Allwood and Cerrato, 2003) investigated gestural feedback. They found a particular form of alignment between interlocutors during an interaction: they tended to reproduce head movements produced by the speaker as gestural feedback. Speaker nodding also appeared to be important in (Terrell and Mutlu, 2012; Stivers, 2008) for bimodal feedback produced in overlap. (Ozkan and Morency, 2012) also found that eyebrow movement, gaze, and nodding are good feedback predictors.

In (Poppe et al., 2010; Terrell and Mutlu, 2012), mutual gaze between interlocutors stands out as the most important cue to feedback. (Ferré and Renaudier, 2017) also argue that gaze is an important cue for bimodal and visual feedback. Speakers almost always look at their interlocutors before producing visual or bimodal feedback.

Summary: Table 1 summarizes the main different feedback-predictive cues listed in the literature according to their modality and the language investigated. We can see a lack of a comprehensive model of feedback predictive cues due to the different phenomena (low pitch regions, final contours, etc.) and languages taken into consideration. This makes these studies difficult to compare. However, some similar findings, especially prosodic features, seem to play an important role. The contribution as well as the interaction of the relative features are not yet very clear. For example, as shown by (Poppe et al., 2010), performances were reduced when including gaze, whereas only taking into consideration pause and pitch strategy gave better results. Conversely,

[1] We distinguish between the generic and specific terms of feedback for intelligibility reasons, according to the descriptions from the studies we have cited.

Table 2
Summary of the literature on feedback prediction with objective evaluation. The Language column refers to: **English** (EN), **Japanese** (JP), **Dutch** (DU). The Method column refers to the algorithm: **Rule-based (RB), Conditional Random Fields (CRF), Hidden Markov Model (HMM), Deep Neural Network (DNN), Long Short-Term Memory (LSTM), Latent Mixture of Discriminative Experts (LMDE)**. The Type column refers to the feedback studied: **Generic**, and/or **Specific**; the Modality column refers to the feedback modality studied: **Verbal** and/or **Gestural**. The Features column refers to the type of feature: **Prosodic (P), Morpho-syntactic (M), Gestural/Visual (G), Auto-regressive (A)**.

Paper	Language	Method	Type	Modality	Features
(Ward and Tsukahara, 2000)	EN/JP	RB	Generic	Verbal	P
(Cathcart et al., 2003)	EN	RB	Generic	Verbal	P M
(Truong et al., 2010)	DU	RB	Generic/ Specific	Verbal	P
(Ozkan and Morency, 2010)	EN	CRF	Generic	Gestural	P M G
(Morency et al., 2010)	EN	CRF HMM	Generic	Gestural	P M G
(De Kok et al., 2010)	DU	CRF	Generic	Verbal/ Gestural	P M G
(Ozkan and Morency, 2012)	EN	LMDE	Generic	Gestural	P M G
(Mueller et al., 2015)	EN	DNN	Generic	Verbal	P
(Ruede et al., 2019)	EN	LSTM	Generic/ Specific	Verbal	P M A

Table 3
Summary of the literature on feedback classification in an offline fashion. The Language column refers to: **English** (EN), **Japanese** (JP), **French** (FR) and **Korean** (KO). The Prediction column refers to the classification task: **Feedback** (FB), **Turn Taking** (TT), **Turn Taking Willingness** (TTW), **Waiting** (W), **Feedback Form** (FF), **Feedback Type** (FT). The Location column describes the site where classification is performed. The Features column refers to the type of features used: **Prosodic** (P), **Morpho-syntactic** (M), **Gestural** (G), **Contextual** (C), **Lexical** (L), **Sentiment** (S), **Acoustic** (A).

Study	Language	Algorithm	Prediction	Location	Features
(Kitaoka et al., 2006)	JP	C4.5	FB/TT/W	Pauses	P M
(Meena et al., 2014)	EN	J48	Hold/ Response	IPU end	P M C
(Kawahara et al., 2016)	JP	Binary classifier	FF	Boundary end	P M
(Skantze, 2017)	EN	RNN & LSTM	Hold/Shift & Short/ Long Utterances	Pauses & Speech Onset	P M
(Ishii et al., 2021)	JP	Adam optimizer	FB/TT/ TMW	IPU end	A G L
(Jang et al., 2021)	KO	LSTM & KoBert	FT	FB interval	P L S
(Liu et al., 2022)	FR	SVM & LSTM	FB	Frame	P M G

(Truong et al., 2011) did not find significant results for rising and falling pitch, but gaze appeared to be the most relevant indicator. Lastly, laughter and smiling have not often been investigated in the literature as feedback-predictive cues. It is important to note that all of these studies investigate feedback-inviting features only for generic feedback.

3. Two different ways of predicting feedback: existing approaches

We found two types of methods for predicting feedback in the literature. One type involves temporal prediction and consists in identifying whether or not feedback may occur at each time-step (for example every 40 ms). We refer to this type as "*continuous prediction*", summarized in Table 2. The second type consists in studying what happens at specific positions, such as pauses, and predicting whether the next event after the pause will be feedback, a turn change, or a turn hold. We refer to this type as "*non-continuous prediction*", summarized in Table 3. The studies on these methods are summarized in Tables 2 and 3. They focus on methodological differences in feedback definition and feature selection.

3.1. Continuous feedback prediction

The continuous prediction of feedback has been explored using either rule-based or machine-learning methods.

Rule-based methods. In their seminal paper, (Ward and Tsukahara, 2000) proposed predicting generic-verbal feedback in English and Japanese by means of prosodic features based on five rules presented in Fig. 1. This study was based on an audio corpus. From the different algorithms tested, the one based on low-pitch cues provided considerably better results than their baseline (random prediction).

(Cathcart et al., 2003) compared performances of several rule-based models, based on five types of rules applied to a HCRC MapTask corpus to predict generic feedback. The first baseline model generated feedback every seven words. The second type of model tested was based on the three or ten most frequent trigrams of POS that preceded feedback. The third type of model generated feedback according to different pause durations (from 400 ms to 1.5 s). Finally, (Cathcart et al., 2003) tested the combination of both POS and pauses (the 10 most frequent trigrams followed by a pause of at least 900 ms; 3 trigrams followed by a pause of at least 600 ms). The best model was the one with 3 trigrams and a pause duration of 600 ms. (Poppe et al., 2010) proposed testing three rule-based models that jointly and separately used prosodic features (pitch and pause) and visual feature (gaze). They focused only on generic feedback. These models were compared to three baselines (a copy of human feedback, Ward's rules described above, and random generation). The feedback (nodding and/or *"uh-huh"*) was produced by a virtual agent, and a subjective evaluation was conducted by human observers. Average scores between 0 and 100 were obtained, corresponding to how natural the feedback produced by the virtual agent was judged to be. The prosodic model provided the best result, while only gaze gave lower scores, similar to the performance of the random strategy. (Truong et al., 2010) used pitch and pause information to design a rule-based model of feedback. They used the Dutch IFADV corpus, and evaluated their prediction on short vocal feedback (interjections, laughs, short evaluations). This type of feedback can fit the role of either generic or specific. As a baseline, they computed the rule-based model of (Ward and Tsukahara, 2000). Results showed that the "pitch & pause strategy" provides higher precision but a lower recall than the rules of (Ward and Tsukahara, 2000), but performance is slightly better with the pitch & pause strategy.

Machine learning methods. Almost all recent research is now based on *machine learning techniques* to predict generic feedback. Most of them focused on gestural feedback (nodding) (Morency et al., 2010; Ozkan and Morency, 2010; de Kok et al., 2014), while some also investigated bimodal feedback (De Kok et al., 2010; Ruede et al., 2019). A large number of these studies made use of the probabilistic sequence model (Morency et al., 2010; De Kok et al., 2010; Ozkan and Morency, 2012). This method gives an output of a sequence of probabilities, using these probabilities, a threshold can be used to trigger the prediction (feedback). This threshold can then be adjusted to match with a level of expressiveness (the quantity of feedback produced). Thus, different

types of listening behavior can be modeled. In parallel with the use of more sophisticated algorithms, the models take into account features from different information levels: prosodic features have been gradually supplemented with lexico-syntactic, mimo-gestural and temporal cues.

(Morency et al., 2010) used sequential probabilistic models to predict one type of gestural feedback (nodding) based on a set of multimodal features (prosody, spoken words, and eye gaze). CRF and HMM were used in order to produce distinct peaks of probability across time that could be associated with feedback opportunity. In another study, (Ozkan and Morency, 2010) established a new feature-selection method called *self-features* by looking at the influences of several features individually. Various prosodic, lexical, syntactic and visual features were taken into consideration. The importance of the features was computed for each listener, and ultimately a consensus was established on the best features. The authors then ran a CRF algorithm to evaluate the prediction of nodding feedback. The consensus of self-features improved performances compared to a baseline that included input from all the features. Different encodings of features were tested (e.g., binary encoding and ramp function encoding that linearly decrease), but nouns, determinants, eye gaze, and lowness stood out as features which significantly improved the model. (Ozkan and Morency, 2012) predicted feedback nods with a new probabilistic model, the *Latent Mixture of Discriminative Experts* (*LMDE*) model that automatically learned a mapping between multimodal observations and a sequence of labels (nods). A large set of multimodal features was used (including prosodic, visual, lexical, morpho-syntactic features). The best model was the one using all types of features.

Deep learning techniques have also been used in other studies. (Mueller et al., 2015) used prosodic features (pitch and power) to predict generic verbal feedback (e.g., *"yeah," "um-hum," "uh-huh"*) on the audio-only Switchboard corpus. In (Ruede et al., 2019), the authors used a *Long Short-Term Memory* (LSTM) model to predict short forms of verbal feedback. They only predicted when the feedback would be triggered, without making a distinction between the feedback function or type. More than half of the items of feedback were generic interjections (e.g., "*yeah,*" "*um-hum,*" "*uh-huh,*" "*right*"), but no specification was given regarding the other forms of feedback. According to their definition of feedback, we believe that at least some of the other forms of feedback could fall under the specific category. Features were extracted automatically (pitch slopes, pause triggers, fundamental frequency variation, Mel-Frequency Cepstral Coefficient, and word history with word2vec) from the audio-only Switchboard corpus.

3.2. Non-continuous feedback prediction

Current research dealing with interaction modeling is mostly focused on the organization of the interaction and often attempts to predict, individually or simultaneously, either turn-changing, turn-holding, or feedback. For a complete review, see (Skantze, 2021). Few studies have focused on feedback type classification (Kawahara et al., 2016; Ortega et al., 2020; Jang et al., 2021). We distinguish these studies from those presented above, since the questioning and the methods differ. In this section, we present some studies that perform prediction only at a specific moment of the interaction (at the end of an utterance or during the speaker's pauses) and do not provide continuous decisions.

(Kitaoka et al., 2006) investigated response timing during non-overlapping speech. Silent pauses by the speaker were classified into 3 categories, corresponding to listener behavior: *making feedback, turn-taking*, or *waiting*. A C4.5 decision tree was used, based on prosodic and morpho-syntactic features and duration information (e.g., duration of preceding utterance, elapsed time from the end of the previous utterance). (Meena et al., 2014) developed a Response Location Detection system (RLD) in an offline and online fashion from a human-machine corpus. In the training data, a spoken dialogue system (the listener) followed the instructions of a human user (the speaker) to perform a MapTask. The dialogue system could produce an acknowledgment, a clarification request, a repetition, or a guest response. Each speaker's IPUs were annotated into *Response*, meaning that a response was given after the given IPU, or *Hold*, meaning that no response was given. Several models were tested (Naïve Bayes, a J48 decision tree classifier, SVM, Voted Perceptron). The classification into the Response and Hold categories was performed at the end of each speaker's IPU. The importance of each category (prosodic, contextual, syntactic features) was tested. The best offline results were obtained with the J48 classifier when all types of features were used. Both trained and evaluated models performed better than a majority class baseline.

(Ishii et al., 2021) investigated multitask learning in order to improve feedback predictive models. Acoustic, linguistic and visual multi-modal features were extracted. At the end of every IPU, the model would predict either *feedback, turn-changing* or *turn-management willingness* (composed of 4 willingness behaviors: *turn-holding, turn-yielding, turn-grabbing, listening*). A comparison was made between single-task models and multi-task models. The authors used both speaker and listener signals as features. On the one hand, the performance results decreased when feedback and turn-changing were predicted jointly. On the other hand, the best performances were obtained when feedback and turn-management willingness were combined with both interlocutor features. The authors concluded that adding turn-changing to feedback prediction does not improve prediction performances. However, predicting both feedback and turn-management willingness does improve performance. When predicting only feedback, taking into account both speaker and listener features outperformed the model considering only the speaker features. To our knowledge, this is the first study that demonstrates a significant improvement in performances by adding listener information.

(Kawahara et al., 2016) predicted the lexical form of feedback from morpho-syntactic and prosodic features. Short instances of verbal feedback (generic and specific) from a counseling corpus were annotated. The task was to predict the form of the feedback according to four categories (*un, un-un, un-un-un* or assessments).[2] Eight participants discussed their personal troubles with a counselor in dyadic interactions. Only two professionals were hired to perform the role of a counselor. In order to treat the problem of variability, the annotations were augmented: three human annotators annotated the "acceptable" feedback forms after each IPU or boundary. A label was kept if all three annotators had selected the same form. The model was finally extended in order to predict one more category: *no feedback*. This prediction was made after each IPU.

(Skantze, 2017) employs Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) models to capture speaker activity based on various input features, including previous speech activity (vocal activity), prosodic features (pitch, power, spectral stability), and Part-Of-Speech information from both participants. This predictive model extends beyond the current frame, considering the probability of speech activity for the next 60 frames (3 s). The combined use of RNN and LSTM in Machine Learning proves effective in handling long-range dependencies within the context. The model is subsequently applied to two distinct tasks. In the first task, the model predicts which of the two participants will initiate a speech activity following pauses (hold or shift) of at least 500 ms in duration. This task aims to anticipate the speaker's behavior after a significant pause. In the second task, the model predicts, for each turn-taking event, whether the associated speech activity will be a long utterance or a short one. This category encompasses various vocal productions, including vocal feedback.

More recently, (Jang et al., 2021) conducted a multitask learning task to jointly predict feedback categories (no-feedback, continuers, understanding, empathic responses) and a sentiment score. Only verbal feedback was considered in this study. Empathic responses can fit the role of specific. They used data gathered from doctor/patient psychiatric

[2] *un* is a common form used to produce continuers in Japanese.

counseling sessions. This type of data is ideal for studying feedback prediction: within the dataset, 84 % of the doctor utterances were annotated as feedback. Prosodic (Mel-Frequency Cepstral Coefficients) and lexical information were used as features. Finally, the sentiment score was computed, based on a dictionary containing associations between sentiments and words (strong positive, positive, neutral, negative, strong negative). Using the sentiment feature allowed the authors to outperform previous studies by (Kawahara et al., 2016; Ortega et al., 2020), used as baselines. The authors showed the importance of using sentiments words in order to improve feedback category prediction, specifically for empathic responses.

In the perspective of studying feedback development in children, (Liu et al., 2022) used a child-adult corpus in French (Bodur et al., 2021) to compare the prediction of feedback between child-adult and adult-adult. Different models were also computed to test the importance of visual (head movement, gaze, eyebrow movement, smiling, laughter) vocal (pitch, MFCC, voice quality, energy, pauses) and verbal features (POS and word probabilities). The prediction was done on a balanced set of randomly selected feedback frames and no-feedback frames, using a context window of two seconds before each frame. As a baseline, they used a model trained to predict feedback only with speech/non-speech information. As a result, all unimodal models and the multimodal model outperformed the baseline. The three modalities have similar contributions to the prediction of feedback instances.

In sum, some studies focus on the prediction of potential sites of feedback (continuous prediction), while others predict feedback types. In the remainder of this paper, we present our dataset, methodological choices, and two models improving the prediction of feedback in conversation: the first model involves the level of prediction of feedback position and the second the level of prediction of feedback type.

3.3. Main issues in predicting feedback

In this section we identify the main factors which can complicate feedback prediction arising from our state of the art.

Firstly, feedback production is highly variable due to its multimodal and multifunctional nature. Feedback can be of different types (e.g., generic, specific), occur at different moments, and take on many different forms. As a consequence, predicting all possible types of feedback at any moment of the interaction is a complex task, only rarely addressed in the literature. A solution to this problem consists in studying feedback production at a specific moment in the conversation, typically after pauses. A similar approach has been proposed in (Skantze, 2017) for identifying whether, after a turn taking, the speech activity will correspond to a long or short utterance (feedback falling into this second category). Interestingly, this work explores feature sets of the two speakers and their impact on the model. We propose in our approach to extend this type of investigation with more modalities. Moreover, the method proposed in (Skantze, 2017) corresponds to a classification task after a given position (turn taking).

The second issue involves the acquisition and multimodal annotations of the data (for both feedback-inviting features and feedback itself). As manual annotation is a time-consuming task, automatic annotations can accelerate the process of annotation. Some annotations remain difficult to interpret, however (e.g., openSmile, openFace, etc.), as they do not directly refer to precise labels; these annotations need to be manually corrected in order to include accurate and high-level features.

Moreover, when using machine learning techniques, feature encoding, window of feature extraction and features used are key to finding the right balance between the quantity of data needed, and the quantity of features. It is thus of great importance to represent relevant information in the most efficient way. Adding more features does not necessarily imply better performances (Hastie et al., 2009).

Another difficulty is the unbalanced nature of feedback production. In conversations, the number of positions where feedback is produced is far lower than positions where no feedback is produced, leading to an unbalanced dataset that creates difficulties for continuous prediction both to train and evaluate models.

The last and most important issue lies in capturing the most appropriate locations to produce feedback when a lot of variability in feedback production is observed. Interlocutors produce feedback at appropriate moments in response to the main speaker's signals. Nonetheless, a listener can be more or less expressive and decide to exploit many feedback opportunities or just a few. This variability in feedback production is raised in several studies and referred to as the *expressiveness problem* (Morency et al., 2010). This variability can also be found in the form of the feedback produced (a listener can decide to produce generic feedback by nodding or expressing an interjection). Studies that focus on one particular type of feedback implicitly reduce the number of feedback opportunities analyzed.

In the following study, we consider all of these issues in our methodology, from the taxonomy of feedback to the implementation of the model. Although all these points render the comparison of all of the above-mentioned studies almost impossible, we believe these studies to be complementary. They provide information about the nature of particular feedback in particular types of conversations.

4. Dataset and features extraction

In this section, we present the PACO—Cheese! corpus used in our study and the supplementary annotations performed. We then present the extraction of the features before providing descriptions of the data.

4.1. PACO—Cheese! corpus

Our dataset comprises two corpora of natural conversations in French: PACO (Amoyal et al., 2020) (https://www.ortolang.fr/market/corpora/paco) and Cheese! (Priego-Valverde et al., 2020) (https://www.ortolang.fr/market/corpora/cheese), referred as PACO—Cheese!. This dataset contains 7 h of audio-visual recordings. Participants (dyads) were sat face-to-face. They were instructed to first read a short story and then speak freely together. PACO—Cheese! is composed of 26 interactions, each lasting between 15 and 20 min. The participants of Cheese! knew each other beforehand whereas the participants in PACO were meeting for the first time. The corpus is enriched with different annotations. The corpus was manually transcribed. This transcription was then automatically aligned with the signal using the SPPAS system (Bigi, 2012; Bigi, 2015) (http://www.sppas.org/) which segments the transcriptions into phonemes, syllables and IPUs. The MarsaTag analyzer (Rauzy et al., 2014) (https://www.ortolang.fr/market/tools/sldr000841/v1#!) was then applied to extract lemmas and POS. Moreover, smiles were annotated semi-automatically thanks to SMAD (https://github.com/srauzy/HMAD) (Rauzy and Amoyal, 2020; Amoyal and Priego-Valverde, 2019) manually corrected, on 4 levels (S1, S2, S3 and S4) and neutral faces were annotated with the S0 label.

We further enriched the existing annotations with prosodic annotations (see Section 4.2.1), nods annotations and feedback annotations (see Section 5) on a subset of 13 dyads (half from Cheese! and half from PACO).

Nodding occurs frequently during speech production from the main speaker and also represents the most common type of gestural feedback (Schegloff, 1982; Allwood and Cerrato, 2003; Stivers, 2008). Nodding is a vertical movement of the head that can be carried out in a single form (one movement from bottom to top or vice versa) or in a plural form (a sequence of several nods). Nodding occurrences were annotated semi-automatically in our corpus: we first applied an automatic extraction step followed by a manual correction. The automatic annotation was performed with the HMAD open-source tool (Rauzy and Goujon, 2018). Technically, detecting nods consists in looking for sinusoidal movements of the pitch angle locating the head pose and

returning the temporal interval during which the movement is performed. As input, HMAD uses a front-view video treated by OpenFace software (Baltrusaitis et al., 2018) and outputs the annotation in an ELAN format. The following step consists in correcting the ELAN output manually (Sloetjes and Wittenburg, 2008).

4.2. Feature extraction

In this section, we present the different features we used and how we extracted them according to the first level (feedback position) and the second level (feedback type) of prediction, presented in Sections 6 and 7.

4.2.1. Prosodic features

We investigated prosodic features by integrating the intonation patterns given by a sequence of tones. Tones represent an intermediary level between low-level acoustic features (e.g., pitch) and phonological interpretations. In some cases, at the end of a sequence potentially bounded by a pause, tone patterns correspond to a final intonation contour. The intonation contour could be correlated with the introduction of new information and possibly carried an important part of the interactional meaning. Several studies have also shown that this final contour might be a good predictor for feedback occurrences. We examined tone patterns (encoded in our approach by n-grams) to compare the influence of tones taken separately or by sequence. Tone extraction was done automatically using the pitch modeling tool MOMEL-INTSINT (Hirst, 2007; Hirst, 2022) in a two-step process. The first step consists in modeling the f0 based on a sequence of transitions between successive points on the curve (called anchor points). This step (corresponding to the calculation of MOMEL) is based on the relationship between the median, minimum, and maximum values of each speaker's pitch range. The *Octave-Median Scale* is used to compare speakers with different pitch ranges (for example, males versus females). In a second step, the MOMEL anchor points are automatically encoded into an alphabet of tonal symbols *T(op), B(ottom),* and *M(id)* referring to absolute values and *H(igher), L(ower), S(ame), U(pstepped),* and *D(ownstepped)* referring to relative values. This encoding provides intonation patterns represented by the key/midpoint and the span of the speaker's pitch range. To extract the features from the tone annotation, for the first level of prediction, we extracted the 3 last tones produced before each time-span of 40 ms. Next, we extracted only n-grams of tones that were produced more than a given occurrence threshold to limit the number of features. We used two thresholds of occurrence (500 and 800). If an n-gram was present more often than the given threshold, it was then selected as a feature.[3] For the second level of classification, we tested a window of extraction of 2 s and the 3 last n-grams. The thresholds of occurrence used were 10, 15, 20, 50, and 100. Prosodic models considered silent pauses to be a relevant boundary cue of the intonational phrase. An intonation contour associated with a pause could reinforce the end of this major prosodic unit that could be associated with the end of a discursive unit, creating conditions that favor feedback. We also encoded the duration of silent pauses (200–400 ms, 400–600 ms, 600–1200 ms, 1200 ms and longer). Speech rate was also considered (the number of tokens produced by the main speaker in the two previous seconds). Finally, we also automatically annotated when the main speaker was speaking for the second level of prediction, feature that we called *overlap*. This feature indicated if the main speaker was speaking when feedback occurred.

4.2.2. Morpho-syntactic features

On the lexico-syntactic level, we included POS as well as lexico-semantic information. POS are often employed in the literature for predicting feedback. Some POS (adverbs, for example (Bertrand et al., 2007)) have been shown to play a role in favoring the occurrence of feedback. POS can also provide important information at the discourse level: for example, discourse markers reveal the discourse structure and are often associated with transitions between discourse units (that may correspond to a listener's reactions). We thus included them in the model sequences of POS encoded with n-grams. POS were automatically extracted from the transcription with the MarsaTag tagger (Rauzy et al., 2014). Based on the morpho-syntactic information, MarsaTag also inserts two optional categories between the POS – strong and weak punctuation – which correspond respectively to the written counterpart of the end of a sentence and to the end of a clause or phrase. Although there is no punctuation in speech, these punctuation tags are useful and help segment the speech flow into units of varying degrees of completion. POS features were extracted and encoded as described above in the same way as the tone n-grams. Syntactic punctuation was extracted in a window of 2 s before each time-span.

In addition to POS, we also extracted lexico-semantic information about word polarity (positive, negative) and aspect (concreteness) on the basis of word lists given in (Bonin et al., 2018). These features were extracted in a window of 2 s before each time-span. We used a binary encoding (presence or absence) for the first level of prediction. For the second level of prediction, we counted the number of tokens which occurred since the last feedback occurrence, resulting in a numerical encoding. This information is important in particular when studying specific feedback (emotion, surprise, introduction of new discourse referents, etc.).

4.2.3. Mimo-gestural features

The introduction of gestures as a feature completed the multimodal description: nodding, laughter, and high intensity smiles (S3 and S4, referred to as *smiles*) were taken into account in the prediction.

For the first level of prediction, two windows were tested to extract nods, laughs, and smiles. Multimodal features were extracted in the previous time-span of 40 ms in the case of the first window. For the second window, they were extracted in a previous window of 2 s. The second level of prediction used mimo-gestural features which were extracted in a previous window of 2 s.

4.2.4. Auto-regressive features

In addition to the context of the speaker's productions, it is crucial to consider what was produced previously by the listener: the prediction of feedback also depends on previous feedback occurrences. We integrated this information into the model by means of "*auto-regressive*" features implementing the fact that the source of information does not come only from the speaker but also from the listener. A system (in our case, a participant in the interaction) is able to self-regulate according to their memory of previous actions.

Concretely, auto-regressive features encode information about the amount of time that has elapsed since the last feedback occurrence. In order to produce binary features bearing precise information, we encoded the information into 5 classes, depending on the elapsed time since the last feedback occurrence (0–2 s, 2–5 s, 5–10 s, 10–20 s, 20 s and more since the last feedback occurrence).

Finally, we also added a categorical feature indicating the type of the previous feedback for the second level of prediction (0 = generic; 1 = positive-given; 2 = positive-new; 3 = negative-given; 4 = negative-new).

We would like to clarify one point concerning the source of this type of information. For an online prediction engine, auto-regressive features must be computed dynamically based on the previous predictions already made by the engine. In our case, however, our predictive model worked in an offline mode during the evaluation step. This meant in particular that all events were already established. We knew that the feedback produced guided and supported the course of the interaction and could simultaneously modify the signal of the speaker and the feedback-inviting cues produced that were processed by the listener. At

[3] Note that the count is realized on the sliding window data-frame, so when an n-gram is produced, it is annotated several times.

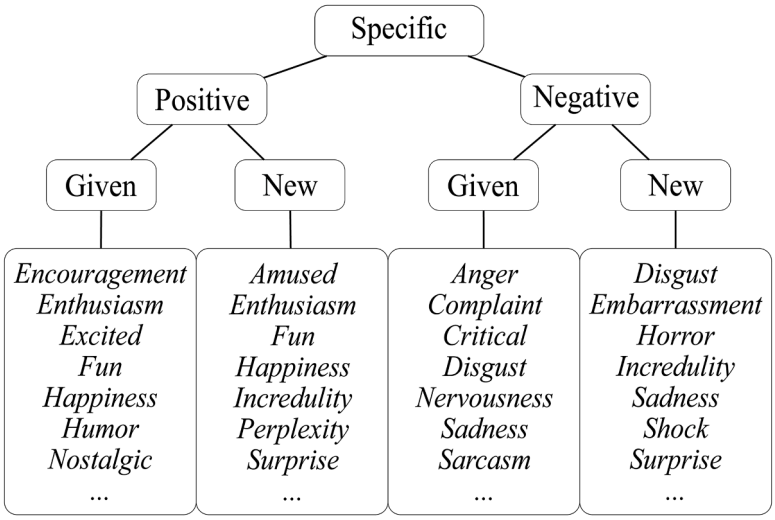

Fig. 2. Specific feedback classification scheme. The 1st level contains the classification of **Positive/Negative** feedback. The 2nd level contains **Given/New** feedback. The 3rd level contains some examples of attitudes per specific feedback type.

this stage of our study, we evaluated the predicted feedback based on the observed feedback, anchored in an unchangeable context. Thus, for this study, auto-regressive features were computed for the observed feedback, instead of the predicted feedback.

5. Towards a more precise description of feedback

Here we introduce a taxonomy for classifying feedback into five types: *generic, positive-new, positive-given, negative-new*, and *negative-given*. In this section we first provide the motivations for our feedback taxonomy, then we present our feedback annotations and a statistical analysis of the components of the feedback according to type. Finally, we present the most frequent set of feedback components observed in our data by type.

5.1. Feedback taxonomy

Our proposed taxonomy is built upon the distinction between generic and specific feedback that stand out clearly through their distinctive components and functions (Bavelas et al., 2000). However, specific feedback encompasses a wide range of attitudes and reactions. (Allwood et al., 1992; Bunt, 2012; Prévot and Gorisch, 2014) emphasize the significance of semantic and pragmatic aspects in feedback production. They demonstrate that the meaning and the function of feedback is dependent on the previous utterance (or scope) of the feedback. Moreover, (Allwood et al., 1992) argue that three parameters of the feedback scope are highly important: the type of speech act (e.g., statement, offer, request, etc.), the factual polarity (affirmative or negative utterance), and the information status (*new* or *given* information).

(Stivers, 2008) also points out the importance of the previous context of feedback to study alignment and affiliation between interlocutors. The author defines alignment as an adaptation to the current activity; listeners mainly use feedback to show their alignment. *Stance* is the main speaker's explicit or implicit affective treatment of the events being communicated, such as something sad, horrible, funny, exciting, etc. Main speakers give clues about their stance to help listeners to react in a preferred way (preferred response) (Sacks et al., 1974; Jefferson, 1978). Thus, the listener can potentially show affiliation, defined as the way that *"that the hearer displays support of and endorses the teller's conveyed stance"* (p.35) (Ruusuvuori and Peräkylä, 2009) analyze the way that the speaker and listener manage their facial expressions and talk to express their stance related to their discourse. Main speakers have been found to use facial expressions and lexical markers to show their stance. The authors illustrate cases where listeners demonstrate affiliation by adopting markers similar to the speaker's stance.

We propose a taxonomy for specific feedback according to two main characteristics of the contextual discourse: **the main speaker's stance** (positive/negative) and the **information structure of the feedback scope** (given/new). A feedback instance is first classified as *generic* or *specific*. Next, we define the specific feedback as *positive* or *negative*, based on the main speaker's stance expressed in the speech rather than on the feedback components themselves, with the assumption that the feedback characteristics (components and function) will be highly dependent on the polarity of the main speaker's stance, through alignment (Pickering and Garrod, 2021) and affiliation mechanisms (Stivers, 2008; Ruusuvuori and Peräkylä,. 2009). The second characteristic concerns the elaboration of common ground between participants. Common ground plays a central role in helping conversation progress (Clark, 1996; Horton, 2017). Feedback can demonstrate that *new* information has been correctly processed and instantiated in the common ground. *New* feedback allows the main speaker to monitor the listener's understanding. Alternatively, the main speaker may refer to *given* information (i.e., shared knowledge, shared experience or general knowledge). In this case, the main speaker needs to ensure that the listener has correctly retrieved this knowledge from their common ground. In both cases, feedback is potentially expected from the main speaker before continuing (or adjusting the speech in case of inappropriate feedback).

Consequently, we argue that all feedback, whatever its modality, can be classified as either *generic, positive-new, positive-given, negative-new*, or *negative-given*. We chose not to include inappropriate or disapproving feedback in our taxonomy; our data presents very few cases of this types of feedback, which is consistent with the literature. (Prévot and Gorisch, 2014) also note that *disapproval* feedback occurs rarely, and annotations from the Switchboard corpus from (Figueroa et al., 2022) of 1504 instances of feedback show that *non-understanding, disagree* and *disapproval* functions represent only 8.98 % of the feedback annotated. Lastly, inappropriate feedback is not always easily identifiable at first glance and can only be characterized as such by a precise discursive analysis (Bertrand and Priego-Valverde 2017).

In what follows, we define the different types of feedback and give the criteria for distinguishing between generic/specific, positive/negative, and given/new sub-types.

Generic vs. Specific: Generic feedback is a consistent phenomenon which mainly takes the form of a nod, an interjection, a smile or a combination/repetition of these components. The functions of generic

feedback are more restricted than those of specific feedback, and simply show comprehension and encourage the main speaker to continue speaking. Since generic feedback is homogeneous in form (limited to a closed list of realizations (Prévot et al., 2016)) and in function, it does not need a more detailed taxonomy. Conversely, specific feedback is more context-dependent and related to semantic interpretation. This type of feedback can be composed of various visual and/or vocal components (marked intonation, longer lexicalizations, laughter, eyebrow movements, smiles, head movements or facial expressions). Specific feedback can convey different attitudinal/emotional values that we represent by a finer-grained classification using two levels of sub-classes, illustrated in Fig. 2.

Positive vs. Negative: We first annotated feedback according to what the interlocutor was reacting to. Did the interlocutor react to something positive or to something negative? Did the speaker talk about an experience or an event that they were evaluating negatively by expressing criticism, sadness, hunger, etc.? Or on the contrary, was the speaker evaluating something positively with joy or humor? We note that we did not consider negative feedback as feedback that was incorrectly produced, showed disagreement, or was rejected by the main speaker, as other studies have done in the past. This positive versus negative aspect of feedback represents the stance associated with the semantic content produced by the main speaker and to which the listener responds.

Given vs. New: The second level of distinction involves the information structure, namely, given or new, of the information to which the listener is reacting. The question is whether the listener already knows this information from something expressed earlier in the conversation, in a previous conversation or from a domain of knowledge shared between the participants (e.g., two students in linguistics will have common knowledge about linguistics). This level of feedback definition allowed us to distinguish feedback that showed new instances in the common ground rather than that which was reacting to pre-existing common ground.

Fig. 2 illustrates our proposed specific feedback classification and provides examples of associated attitudes, showing that only two subtypes cover a wide range of attitudes.

5.2. Feedback annotation on PACO—Cheese!

We applied our taxonomy to annotate feedback in a corpus involving 13 dyads (26 speakers), which represents approximately 3.6 h of recording. The current study is based solely on this part of the corpus. A total of 2377 items of feedback were obtained (1206 generic and 1171 specific, including 416 positive-given, 548 positive-new, 115 negative-given, and 92 negative-new).

Feedback was annotated by two students trained to perform the task and by one of the authors. The difficulty of the task arose first in the identification of feedback in the whole conversation and secondly in the selection of the feedback type. In order to train the annotators, annotations were first carried out on the same 12 speakers by the three annotators. Regular meetings were set up between them to discuss disagreements and choose the best category by consensus. Prior to discussions, the Fleiss Kappa was 0.25 for the annotations of the 12 speakers, corresponding to a fair level of agreement. After the training phase, 3 additional speakers were annotated by the 3 annotators, without consultation between them. The Fleiss Kappa obtained rose to **0.65**, corresponding to substantial agreement. Given the final level of agreement obtained, the remaining speakers to be annotated were then divided between the 3 annotators.

The frequency of the feedback observed in our dataset firstly confirms the literature (Poppe et al., 2010): feedback is a frequent and consistent phenomenon. Our data show a frequency of **10.78 items of feedback per minute** (see Table 8), with quite a high standard deviation ($\sigma = 3.02$). The less expressive interlocutors tended to produce half as many items of feedback compared to the more expressive interlocutors: this observation also confirms the problem of individual variability. The mean duration of feedback was 1.27 s (minimum duration = 400 ms, maximum duration = 6.52 s).

A Welch two sample *t*-test was conducted type-by-type on the feedback frequency between Cheese! (8 dyads) and PACO (5 dyads) to ascertain whether the degree of acquaintance between the participants has an impact on feedback frequency. We did not find any differences, except for the positive-given type (p-value < 0.05). Cheese! speakers produced twice as many positive-given type occurrences, which can be explained by the fact that they had more common ground to refer to. However, the examination of variations in the participants' degree of acquaintance falls outside the purview of this study. Thus, further exploration of this aspect is not pursued as it appears to have negligible influence on the quantity and frequency of feedback in our findings. We did not explore other speaker characteristics such as gender, as our subset of 26 participants contained only 3 males, or age, as all speakers were students of a similar age.

Distribution between generic and specific feedback is roughly equivalent. The types of specific feedback were more often positive than negative. Overall, the most frequent type of feedback was the positive-new type, the least common was the negative-new type.

We note that **47 %** of the feedback was produced in *verbal overlap* (i.e., both speaker and listener were speaking simultaneously). There was an equal amount of generic and specific feedback produced in verbal overlap. If we look at all the feedback, **76 %** was produced *in overlap* with the speaker's speech, either partially or totally.

5.3. Analysis of the feedback components

Our analysis investigates the verbal, vocal, and gestural components of feedback itself, as well as their combinations. Our hypothesis is that there are different sets of feedback components for each type and subtype, with certain components being more frequently employed based on feedback type, main speaker's stance, and information status.

In the gestural modality, we identified specific gestures such as **low-intensity smiles** (S1 and S2), **high-intensity smiles** (S3 and S4), **head nods**, and grouped **eyebrow movements** (raised and lowered). A more fine-grained analysis of the feedback components was produced by annotating eyebrow movements (raised and lowered) during each feedback instance by three raters (Fleiss Kappa = 0.70 computed on a subset of 30 min). In terms of the verbal and vocal modality, we indicate whether the feedback was expressed through **speech** and/or **laughter**. As there was a wide range of tokens produced, we categorized frequent utterances into three groups: **continuer tokens** (e.g., *"mh," "ouais – yeah," "ok,"* etc.), **prototypical given tokens** (e.g., *"c'est ça - that's right," "exactement - exactly"*), and **prototypical new tokens** (e.g., *"c'est vrai" – really," "sérieux - seriously"*). Additionally, we recorded the usage of **positive** and **negative** tokens from the list of (Bonin et al., 2018).

We now turn to the relationship between the types of feedback and the set of feedback components. We proceeded as follows: for a given contrast between two types (e.g., generic versus specific), we explored whether the presence or absence of a component (e.g., laughter, continuer-token, nodding) favored one of the two types. We used Logistic Regression to assess the statistical significance and the sign of the effect for each component. We then tested this significance for the three exclusive levels of the taxonomy: 1) between *generic* and *specific* 2) between *positive* and *negative* and 3) between *given* and *new*.

Table 9 presents the significance of each component according to the type. In comparing generic and specific feedback, we observe that *nodding* and *continuer-tokens* are significant for **generic** feedback. Conversely, all other components lean towards **specific** feedback, except for *negative tokens*, which do not show significance for any type.

Positive feedback is characterized by *laughter, nodding,* and *positive tokens*, while **negative** feedback is characterized by *eyebrow movements, speech,* and *given tokens*. Finally, *high-intensity smiles, eyebrow movements,*

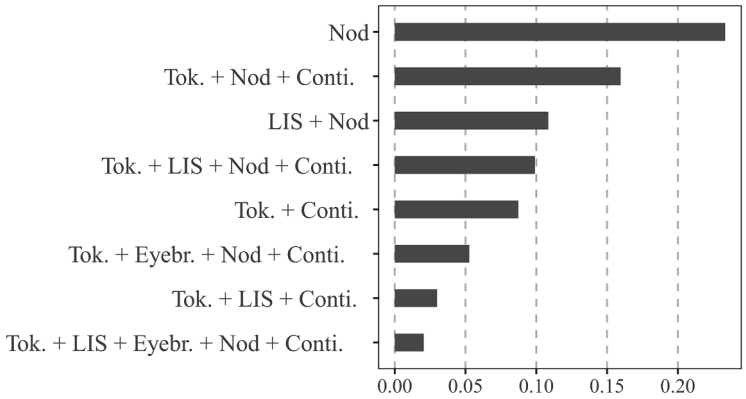

Fig. 3. The most frequent combinations of components used to produce generic feedback, determined by their ratio (combination frequency/total number of feedback). These combinations represent 80 % of generic feedback. The different components include: **Tokens** (Tok.), **Low-Intensity Smile** (LIS), **Eyebrow movement** (Eyebrw.), **Continuer** (Conti.), **Nodding**.

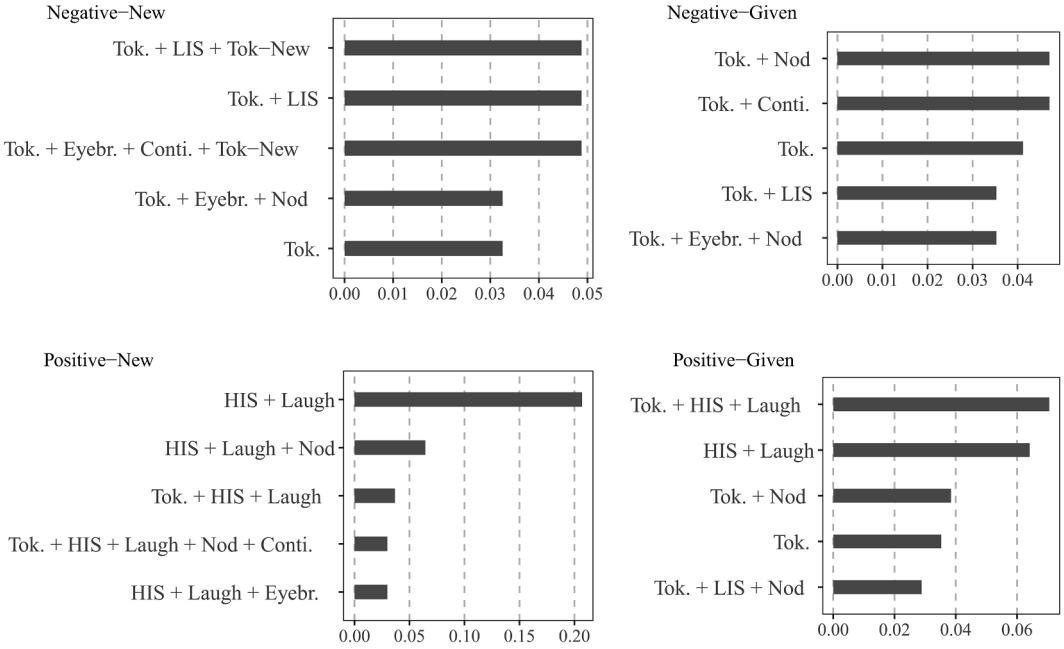

Fig. 4. The five most frequent combinations of components used to produce specific feedback: **Negative-New, Negative-Given, Positive-New, Positive-Given** based on their ratio (frequency of the combination/total number of feedback). The different components include: **Tokens** (Tok.), **Low-Intensity Smile** (LIS), **High-Intensity Smile** (HIS), **prototypical new token** (Tok-New), **prototypical given token** (Tok-Giv.), **Eyebrow movement** (Eyebrw.), **Continuer token** (Conti.), **Nodding, Laughter** (Laugh).

and *given tokens* are indicative of feedback of the **given** type while *laughter* and *new tokens* are significant components of feedback of the **new** type.

Contrary to our expectations, *negative tokens* do not hold significance in predicting negative feedback, possibly due to their low frequency (accounting for only 2 % of the feedback observed).

For each feedback type, we computed all the possible combinations of the feedback components as the ratio of the combination (frequency of the combination/the number of feedback instances). Fig. 3 presents the most common combination of feedback components for generic feedback, representing 80 %. Fig. 4 presents the five most frequent combinations of negative-new, negative-given, positive-new, and positive-given.

Among the different types of feedback, the most frequent combinations in **positive-new** feedback consistently involve *high-intensity smiles* and *laughter*. It also appears that the most frequent combinations in positive-new feedback are the *least verbalized* compared to other types. This could indicate that high-intensity smiling and/or laughter is the preferred response in the context of positive-new feedback. Conversely, **positive-given** feedback comprises two combinations with *no smile* and one with a *low-intensity smile*. In contrast, we observe that for the most common set of feedback components of **negative** feedback, there is no high-intensity smiling or laughter, but sometimes a *low-intensity smile*. High-intensity smiling and laughter are not excluded during negative feedback production, but they appear to be rare. Moreover, we observe that *eyebrow movement* is mainly associated with **negative** or **new** feedback. Eyebrow movement can indicate a negative stance and/or surprise. The **negative-given** feedback seems to be the most neutral in terms of co-speech gestures, and is essentially realized with *speech* (token), *nodding*, and/or *continuer token*. Finally, *prototypical new tokens* are very frequent when feedback is of the **negative-new** type.

Table 4
Summary of prosodic features and their extraction according to the task: 1st level of prediction and 2nd level of prediction. The table presents the **feature name**, its **encoding**, its **window of extraction** before the time-span (**3 last elements (3 last), overlap with the left border of the time-span (LB)** and **2 s before** (2 s)), and the **threshold of occurrence for n-grams**.

Level of prediction	Features	Encoding	Extraction window(s)	Threshold of occurrence
1st	Number of tokens	Count	2s	500, 800
	N-grams of tones	Binary	3 last	–
	200–400 ms Pauses		LB	
	400–600 ms Pauses			
	Pause 600–1200 ms Pauses			
	1200 ms + Pauses			
2nd	Number of tokens	Count	2s	–
	N-grams of tones	Binary	3 last; 2s	0, 20, 40, 50, 1000
	Overlap		FB	–
	200–400 ms Pauses		LB	
	400–600 ms Pauses			
	600–1200 ms Pauses			
	1200 ms + Pauses			

Table 5
Summary of the morpho-syntactic features and their extraction according to the task: 1st level of prediction and 2nd level of prediction. The table presents the **feature name**, its **encoding**, its **window of extraction** before the time-span (**3 last elements (3last), 2 s before** (2 s)**, time since last feedback occurrence (Last FB)**), and **the threshold of occurrence for n-grams**.

Level of prediction	Features	Encoding	Extraction window(s)	Threshold occurrences
1st	N-gram of POS	Binary	3 last	500–800
	Positive words		2s	–
	Negative words			
	Concrete words			
	Discourse markers			
	Punctuation			
2nd	N-grams of POS	Binary	3 last; 2s	10, 20, 40, 50, 1000
	Positive words	Count	Last FB	–
	Negative words			
	Concrete words			
	Discourse markers			
	Punctuation	Binary	2s	

6. First level of prediction: predicting feedback positions

This section presents the methodology implemented for the continuous prediction of feedback (every 40 ms) as well as data sampling, data preprocessing, feature selection, the construction of the model, results and a discussion.

6.1. Data sampling

We defined two timescales to predict the precise locations of feedback occurrences. The first was the sampling rate, the second, the timescale of the extracted feature.

This question of temporal segmentation is addressed quite differently from one study to another. The main difficulty is that predicting feedback location requires us to examine all possible positions, which means applying high frequency sampling. As a consequence, the set of positions where no feedback occurs is by far the largest.

A previous study by (Boudin et al., 2021) proposed that all time locations corresponding to the end of verbal or non-verbal units be considered (e.g., end of a word, pauses, laughter, smiling, etc.). Each observation generated an entry in the data table. The value of the dependent variable was set at 0 for encoding absence of feedback and at 1 for observed feedback. The task therefore used a standard binary classification. In our case, the main difficulty arises from the unbalanced class distribution: "*no feedback*" events are by far the most frequent. The problem in this case is that a model predicting only no-feedback, whatever the position, would still lead to good accuracy.

Hereafter in the current study, we chose a sampling rate of 40 ms (i.e., a time step) which was in accordance with the frame rate of the captured video. This sampling frequency made it possible to register rapid changes inherent to the interaction and determine the accuracy within which the feedback onset was predicted. A second timescale specified the size of the time window from which the features were extracted. This timescale duration in practice depends on the features considered (as described in Tables 4–7). This in mind, our input data was sampled at increasing time-spans (i.e., the series of time incremented by the given time step). For each time-span, the presence of the observed

Table 6
Summary of mimo-gestural features and their extraction according to the task: 1st level of prediction and 2nd level of prediction. The table presents the **feature name**, its **encoding**, its **window of extraction** before the time-span (**overlap with the left border of the time-span (LB)** and **2 s before** (2 s)).

Level of prediction	Features	Encoding	Extraction window(s)
1st	Nods	Binary	LB: 2s
	Smiles		
	Laughs		
2nd	Nods		2s
	Smiles		
	Laughs		

Table 7
Summary of auto-regressive features and their extraction according to the task: 1st level of prediction and 2nd level of prediction. The table presents the **feature name**, its **encoding**, its **window of extraction** before the time span (**2/5/10/20 s** before **and time since Last Feedback (Last FB)**).

Level of prediction	Features	Encoding	Extraction window(s)
1st	Last FB 0–2s	Binary	2s
	Last Fb 2–5s		5s
	Last FB 5–10s		10s
	Last FB 10–20s		20s
	Last FB 20s		- s
2nd	Last FB type	Discrete	Last FB

Table 8
Feedback frequency per minute and per type: **Generic, Positive-given, Positive-new, Negative-given, Negative-new, All specific Feedback,** and **All Feedback**.

Feedback type		Frequency per minute
Generic		**5.56 (± 2.00)**
Specific	Positive-given	1.88 (± 1.00)
	Positive-new	2.44 (± 1.58)
	Negative-given	0.58 (± 0.43)
	Negative-new	0.44 (± 0.31)
	Total Specific	**5.14 (± 2.58)**
Total Feedback		**10.78 (± 3.02)**

Table 9
Results of the three logistic regressions to investigate: 1) **Generic** vs. **Specific**, 2) **Positive** vs. **Negative**, 3) **Given** vs. **New** according to the different feedback components: **Low-Intensity Smile** (LIS), **High-Intensity Smile** (HIS), **Laughter, Eyebrow movement** (raised or lowered), **Nodding, Token, Continuer Token, Negative Token, Positive Token, Given Token** and **New Token**. The "*Proportion*" column indicates the proportion of the feedback component observed in our data. The model 1) considers all feedback, whereas the model 2) and 3) consider only specific feedback.

Component	Proportion	Generic	Specific	Positive	Negative	Given	New
LIS	28.37 %		***				
HIS	27.29 %		***			**	
Laughter	19.24 %		***	***			***
Eyebrows	25.39 %		***		**	***	
Nodding	66.18 %	***		**			
Token(s)	68.56 %		***		**		
Continuer Token(s)	43.47 %	***					
Negative Token(s)	2.03 %						
Positive Token(s)	3.59 %		***	**			
Given Token(s)	3.5 %		***		*	***	
New Token(s)	12.37 %		***				***

feedback received Boolean encoding, i.e., 0 before the feedback location and 1 from the feedback onset until the end of the feedback.

6.2. Preprocessing

The model used each interaction from the corpus as input twice, once with participant A considered as the speaker (features were therefore extracted from their signal) and participant B as the listener (the one who produced feedback, the binary dependent variable). The same interaction was then used a second time with the participants roles reversed. A preprocessing step to discriminate the role of speaker and listener was thus necessary: without it, the features from the complete signal produced by an interlocutor would be processed regardless of whether the interlocutor was in speaker or listener position. This would introduce a bias in the learning process, and at the same time some features selected by the model would indicate a listener position (e.g., interjections such as "*mmh*," "*okay*," etc.), and would therefore favor "no-feedback" prediction. (Ruede et al., 2019) tried to overcome this problem by only selecting periods where the speaker speaks for at least 5 s and the listener has been silent for at least 5 s. In our study, we only examined frames in which the speaker had produced more tokens than the listener within the 2 previous seconds (unless the listener produced feedback). As a result, we mainly retained frames in which the speaker held the floor, and we took into account all the feedback, whether it was produced in overlap or not. The dataset size was then drastically reduced to strictly relevant information: the initial dataset contained 660,334 frames (each 40 ms), and 326,674 frames after preprocessing. This method also allowed us to reduce the problem of unbalanced classes by reducing the number of no-feedback frames.

6.3. Feature engineering

Different combinations of features and extraction of features were tested in the present study. N-grams of tones and of POS represent an important part of our features. They were selected according to their distribution in the dataset. Their occurrence was calculated on a sliding window containing the three last POS and the three last tones produced by the speaker before the end of the time-span. The n-gram was included in the feature set when it occurred more often than a given threshold in the dataset, as indicated in Section 4.2.

Several tests were conducted for feature extraction in order to represent information in the most efficient way and to reduce the multidimensionality of the data. Firstly, we tried different selection thresholds (500, 800) for POS and tone n-gram selection. Secondly, we reduced the number of features by clustering them. The different bigrams and trigrams of tones were clustered into intonational patterns (*falling, rising, flat* intonation and their combinations, e.g., *rising-falling-rising*, referred to below as *tone patterns*). On the basis of the pitch range calculated by MOMEL-INTSINT, we encoded the span with two additional features: *small* or *large* span. Finally, in addition to using detailed morpho-syntactic information (i.e., the POS category plus its features such as *Ppd* encoding a dative personal pronoun), we also simply used the POS label (referred to below as clustered-POS vs. detailed-POS). Clustered-POS were selected with a frequency threshold of 500 occurrences.

We tested several feature combinations, based on these different types of encoding, for POS and tones, without modifying other feature encodings:

- Clustered-POS & tones 800 (total of 420 features)
- Clustered-POS & tone patterns (total of 273 features)
- Detailed-POS 500 & tones 500 (total of 524 features)
- Detailed-POS 800 & tone patterns (total of 255 features)
- Detailed-POS 800 & tones 800 (total of 402 features)

We also tested the performances of the model with mimo-gestural features extracted immediately preceding each time-span (with an overlap on the left border of the time-span), and when they were extracted in a previous 2-second window.

6.4. Building the predictive model

In this study, we used Logistic Regression (*Logit*) to build the model. Logit, among other possible techniques, provides the probability of feedback realization instead of a simple classification. This characteristic is interesting in the perspective of implementing the model in a human-machine communication system, making it possible to introduce variability in feedback production. *Logit* also provides the possibility of evaluating feature importance, which is essential when interpreting the results. Finally, *Logit* is also appropriate when dealing with small datasets, being less prone to overfitting. Usually, the probability of producing a given type of feedback (or the probability that the feedback is produced at a given time location) is modeled by the equation:

$$logit(p) = ln\left(\frac{p}{1-p}\right) = a_0 + a_1x_1(t) + \ldots + a_jx_j(t) + \ldots \quad (1)$$

where $x_j(t)$ are the predictors which depend on the time location t, and can adopt binary, categorical, or continuous types. In a first step, the parameters of the *Logit* model (i.e., the a_i coefficients) are estimated based on the training sample. An analysis of the result makes it possible to decide which predictor contributes significantly to the prediction. The model is finally built with the set of relevant predictors and a probability p is attributed to any combined values of the predictors.

Concretely, the position of the feedback produced by a listener is predicted as follows. The system is a two states automaton, the first state q_0 corresponds to no-feedback positions, the second state q_1 to places where feedback can occur. The system is initialized at time $t_0 = 0$ on the q0 state and evolves at each time step Δt following the probability given

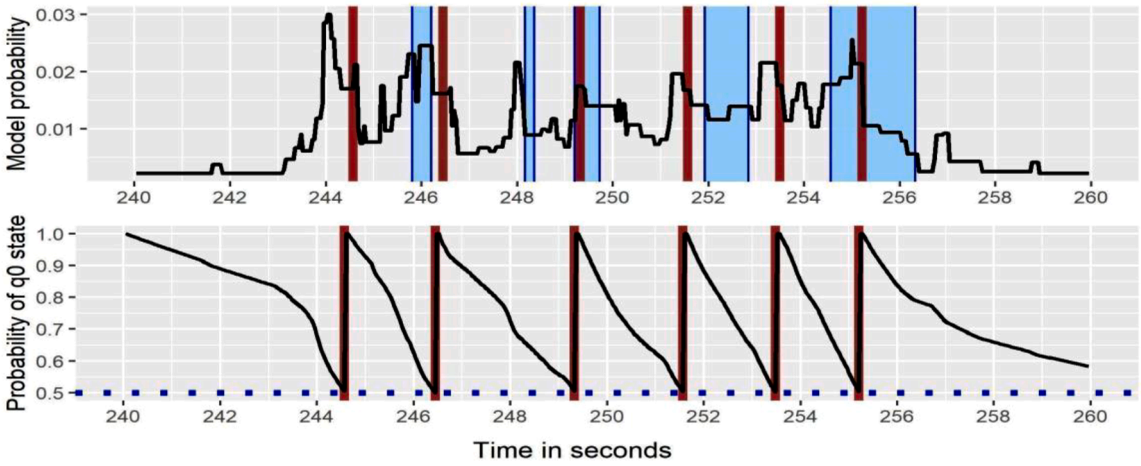

Fig. 5. Illustration of the method to predict feedback location. From the measurements along the timeline of the selected features, the model predicts the variation in the probability that feedback occurs (the black curve on the top panel). The cumulative probability of the system to remain in the q_0 no-feedback state (the black curve on the bottom panel) is computed. Once this probability falls below 1/2, feedback is generated (the vertical red lines) and the cumulative probability is reinitialized for the next prediction. Predicted feedback locations are close to maxima of the Logit model probability. Blue areas identify the true locations of observed feedback for this segment of interaction.

by the *Logit* model, i.e., the probability $p(t)$ that the feedback onset occurs within the time interval $[t, t + \Delta t)$. This probability depends on the values of the features at time t: the features characterizing the production of the speaker and the auto-regressive features of the listener which specify in particular the current state of the system (q_0 or q_1) but also other features such as the time elapsed since the production of the last occurrence of listener feedback. As the system progresses, we compute the cumulative probability that the system remains in the q_0 state. This probability is given by the product of the individual model probability starting at the time the last feedback occurrence was produced, $P_r(q_0) = \Pi_i(1 - p(t_i))$. Once the cumulative probability falls below 1/2, feedback is generated. Note that this procedure assumes that the feedback production is well modeled by a Poisson-like process. In particular, the probability to produce feedback within a time interval of duration Δt is in average proportional to the interval duration Δt (in first approximation if the temporal variations of $p(t)$ are neglected). The procedure is illustrated in Fig. 5. Other methods that associate the positions of the feedback with the local maxima of the probability curve can be considered as well (Morency et al., 2010).

In order to obtain relevant features and their importance, we first trained a global model with all the features for the five combinations described in Section 6.3. For each feature, the model returned an associated coefficient as presented in Eq. (1). The coefficients of the Logit model are computed with one coefficient a_j per feature $x_j(t)$ without interaction terms between the different features. The amplitude of the coefficient marks the strength of the contribution for the given feature.[4] A positive coefficient indicates that the presence of the feature enhances the probability for the listener to produce feedback whereas a negative coefficient decreases this feedback probability. The Logit model also provides the confidence level interval associated with each coefficient estimate. A feature is selected if its estimate is significantly different from 0 within the confidence error bars. In practice this makes it possible to discard features which are not relevant for feedback prediction.

In order to avoid potential overfitting problems, we evaluated the performance of the proposed models by using a Monte Carlo cross-validation procedure. For the 26 participants (i.e., 13 dyads) of our gold standard, the gold standard was split into two subsamples: a training dataset containing 80 % of the gold standard used to compute the fitting parameters of the model and a test dataset containing the remaining 20 % of the corpus which was set aside for the evaluation task. We generated 50 random partitions of the gold standard. For each partition, the test dataset consists of 26 time intervals (one per gold standard participant) of 20 % of the interaction duration and with a random starting time boundary. The 52 remaining intervals (2 intervals per participant, preceding and following the test segment) form the training dataset. For each partition, the model parameters are estimated on the training data and a confusion matrix is computed for the test sample. The confusion matrices of the 50 random partitions are merged at the end. This evaluation procedure was also applied in order to measure the relative contribution of each feature combination (e.g., grouped by modalities).

6.5. Evaluation

Several methods have been used in the literature to objectively and subjectively evaluate predictive feedback models; for a full description, see (de Kok & Heylen, 2015). Because prediction is done at each time stamp of 40 ms, a window of evaluation (referred as margin of error) is necessary. We compared the results using two methods found in the literature. The first, introduced by (Ward and Tsukahara, 2000) used a window of evaluation that takes the onset of a predicted instance of feedback and the onset of the observed feedback in the corpus. A window spanning 500 ms is used around the onset of the observed feedback. If the time of the predicted feedback fits within this window, the prediction is considered as correct. This method of evaluation has been used and adapted in several other studies (Truong et al., 2010; Ozkan and Morency, 2012; Mueller et al., 2015; Ruede et al., 2019).

The second method, by (Morency et al., 2010; Ozkan and Morency, 2010) is to take the whole feedback interval observed in the data as the window of evaluation. Predicted feedback is considered as correct when a peak of probability falls within an interval corresponding to an observed instance of feedback in the corpus.

To evaluate our models, we computed recall, precision, and f-scores using three different margin of error windows (MoE): the one proposed by (Ward and Tsukahara, 2000), 500 ms before or after the corresponding feedback onset (referred as MoE-onset-500), that we also extended to 1000 ms (referred as MoE-onset-1000), and the last MoE is the one that uses the observed feedback time-span (referred as MoE-time-span).

[4] A direct comparison of the amplitude of the coefficients is herein meaningful since almost all the features are encoded as binary variables (with an identical range of values).

Table 10
F-score, Precision and **Recall** for potential feedback sites with Logit. The **Regular Baseline** predicts feedback every 5 s, based on the frequency of feedback observed in the data. The **Random Baseline** predicts feedback at random places by respecting the proportion of feedback/no-feedback instances observed in the data. Results of the same models are presented with 3 margins of errors MoE-onset-500 ms and MoE-onset-1000 ms and MoE-time-span .

Feedback prediction	MoE-onset				MoE-time-span		
	Window	**F-score**	Precision	Recall	**F-score**	Precision	Recall
Logit	500ms	**0.28** (± 0.018)	**0.23** (± 0.016)	**0.34** (± 0.022)	**0.37** (± 0.019)	**0.31** (± 0.019)	**0.46** (± 0.022)
	1s	**0.49** (± 0.022)	**0.40** (± 0.021)	**0.63** (± 0.027)			
Regular baseline	500ms	0.21 (± 0.016)	0.16 (± 0.013)	0.31 (± 0.026)	0.24 (± 0.019)	0.18 (± 0.018)	0.37 (± 0.030)
	1s	0.38 (± 0.018)	0.28 (± 0.015)	0.58 (± 0.029)			
Random baseline	500ms	0.19 (± 0.013)	0.15 (± 0.012)	0.26 (± 0.020)	0.19 (± 0.014)	0.14 (± 0.012)	0.29 (± 0.020)
	1s	0.34 (± 0.002)	0.28 (± 0.002)	0.43 (± 0.026)			

Table 11
F-score, Precision and **Recall** for feedback prediction for the final model (with all types of features) and per type of feature: **Auto-regressive, Morpho-syntactic, Prosodic, Mimo-gestural**, and interaction between **Prosodic/Morpho-syntactic/Mimo-gestural** features. A margin of error of ± 500 ms around the feedback onset is used (MoE-onset-500).

Modality	MoE-onset-500		
	F-score	Precision	Recall
All	**0.28** (± 0.016)	**0.23** (± 0.015)	**0.34** (± 0.019)
Auto-regressive	0.24 (± 0.016)	0.20 (± 0.016)	0.30 (± 0.017)
Morpho-syntactic	0.21 (± 0.015)	0.18 (± 0.020)	0.24 (± 0.017)
Prosodic	0.19 (± 0.017)	0.16 (± 0.016)	0.22 (± 0.017)
Mimo-gestural	0.18 (± 0.020)	0.16 (± 0.015)	0.21 (± 0.017)
Prosodic/Morpho-syntactic/ Mimo-gestural	0.22 (± 0.016)	0.19 (± 0.015)	0.25 (± 0.020)

6.6. Results & discussion

The best model was obtained by using the combination of detailed-POS-500 and tones-500, with mimo-gestural features extracted just before the time-span.[5] For the target prediction (feedback), features from each subset of features were selected. The model finally consisted in 112 significant features, including 53 that favored *feedback* and 59 that favored *no-feedback*. Feature coefficients are presented in Fig. B (see Appendix B) ranked by the importance of their contribution to the model. The strength of the contribution for each feature is traced herein by the amplitude of the coefficient estimate.

Concerning those features which discourage feedback production, only lexico-syntactic and prosodic features stand out as being significant. Table 10 presents the results (f-score, precision and recall) obtained with our model with our three MoE. We compared these results with two baselines. The "*Regular*" baseline generates a feedback occurrence every 5.56 s. This value corresponds to the frequency of feedback observed in our data (10.78 items of feedback per minute). The "*Random*" baseline generates a certain number of feedback items based on the mean frequency per minute but at random places.

Table 11 compares results (using MoE-onset-500) of the model trained and tested with all the features vs. models trained and tested with only one type of feature. As anticipated in our hypothesis, the best performance was reached when using all features (f-score = 0.28). In terms of relative contribution (without interaction), auto-regressive features lead to an f-score of 0.24, with morpho-syntactic features having an f-score of 0.21, prosodic features an f-score of 0.19 and finally mimo-gestural features an f-score of 0.18. Regarding the importance of auto-regressive features, we also tested our model with only features from the speaker (prosodic, morphosyntactic and mimo-gestural), giving an f-score of 0.22. Auto-regressive features are thus important but need to be combined with other types of information.

Our results confirm two hypotheses. Firstly, taking into consideration all modalities outperforms other feature combinations, in particular unimodal ones. Secondly, the use of detailed linguistic features also improves the results. Finally, the model reports linguistic information in the significant features that match with our expectations. Moreover, the level of accuracy of the features greatly improves the interpretability of the model and the understanding of feature importance. We also evaluated the performance of the model at predicting the time location for each type of feedback. We computed the recall ratio for each type which ranges from 0.309 for negative-new to 0.386 for positive-new (with a value of 0.352 for all types together). These slight differences are not statistically significant, however, and more data will be required in order to investigate this effect.

As presented in Table 11, *auto-regressive* features play an important role in feedback prediction, by providing information from the context. In our corpus, feedback is produced on average every 5 s. However, the first two auto-regressive features correspond to a shorter time-span than 5 s (last feedback between 0 and 2 and last feedback between 2 and 5 s). One explanation is that feedback could occur close to other feedback depending on the dynamics of the interaction. Among other parameters, the dynamics relate to the type of activities (such as storytelling, explanations, etc.) within which the occurrences of feedback can be more concentrated in a short time-span (< 5 s.) (see (Stivers, 2008) or (Bertrand and Espesser, 2017)) than in speech sequences which exhibit more symmetrical speaking times between both participants (less feedback in this case). In this way, as given from the different studies in the literature, durations longer than 5 s are also significant. Consequently, even if feedback is based on a cycle that varies in regularity, the regular baseline shows significantly lower results, indicating that predicting feedback every 5 s is not sufficient. Information about the time that has elapsed since the last occurrence of feedback allows the listener to balance their production of feedback. Moreover, the listener has to adapt this cycle according to the needs of the interaction. As the results by modality of features show, auto-regressive features alone are not enough to predict feedback, but they do complete information from other modalities.

The three mimo-gestural features stand out as being relevant: *smiles, nods,* and *laughs*.

The 5 trigrams of tones selected by the model (*SMS, BUU, DUU, LUU, DDT*) that are then relevant for feedback prediction mainly correspond to a rising pattern. The latter seems to be the most powerful within the prosodic features. Among the 5 n-grams, 4 correspond to a rising pattern

[5] Note that differences in results are not of great importance between the two windows of extraction for the mimo-gestural features.

Table 12
Examples of feature combinations observed in the dataset with the probability of obtaining feedback, expressed in **Pmean**, given by the Logit. When no feature has been produced, the Pmean is 0.025. Examples from **1 to 6** present **feature combinations that increase the probability of obtaining feedback**. Examples from **7 to 9** present **feature combinations that decrease the probability of obtaining feedback**. N-grams of POS are indicated by a dash ("-"). N-grams of tones are indicated by a period ("."). Abbreviations are detailed in Tables A.1 and A.2 (see Appendix A).

Id	Features	Pmean
1	*Last FB 0–2 s + Smile + Punctuation + d-Nc*	0.046
2	*Last feedback 2–5 s + Smile*	0.038
3	*Last feedback 10–20 s + Smile + Punctuation + Af + Laugh*	0.037
4	*Last feedback 10–20 s + Pause 600–1200* ms *+ Nod + Af + Punctuation*	0.035
5	*Last feedback 0–2 s + Smile + Nod + Punctuation + I + Va + Vm*	0.033
6	*Last feedback 0–2 s + Vm-R-Pause + Nod + Punctuation*	0.031
	No features	0.025
7	*Last FB 0–2 s + Vm + Punctuation + Ppn-Vm*	0.017
8	*Last feedback 2–5 s + Nc-S-D + Pause*	0.010
9	*Last feedback 0–2 s + R-I + M.L + Punctuation*	0.004

Table 13
F-score, Precision and **Recall** for **Generic/Specific** feedback prediction for the final model (with all types of features) and per type of features: **Auto-regressive, Morpho-syntactic, Prosodic, Mimo-gestural**. The ground truth used is the type of the observed feedback. The **Baseline** predicts randomly generic/specific feedback based on the distribution observed in the data.

Features modality	F-score	Precision	Recall
All	**0.62** (± 0.030)	0.63 (± 0.036)	0.61 (± 0.035)
Auto-regressive	**0.58** (± 0.029)	0.59 (± 0.033)	0.58 (± 0.031)
Morpho-syntactic	**0.58** (± 0.024)	0.53 (± 0.025)	0.64 (± 0.031)
Prosodic	**0.59** (± 0.021)	0.57 (± 0.061)	0.61 (± 0.035)
Mimo-gestural	**0.25** (± 0.029)	0.74 (± 0.047)	0.15 (± 0.021)
Baseline	**0.48** (± 0.029)	0.48 (± 0.030)	0.49 (± 0.036)

Table 14
F-score, Precision and **Recall** for **Positive/Negative** feedback prediction for the final model (with all types of features) and per type of features: **Auto-regressive, Morpho-syntactic, Prosodic, Mimo-gestural**. The ground truth used is the type of the observed feedback. The **Baseline** predicts randomly positive/negative based on the distribution observed in the data.

Features modality	F-score	Precision	Recall
All	**0.36** (± 0.049)	0.26 (± 0.040)	0.61 (± 0.092)
Auto-regressive	**0.28** (± 0.038)	0.20 (± 0.033)	0.45 (± 0.058)
Morpho-syntactic	**0.33** (± 0.048)	0.21 (± 0.032)	0.76 (± 0.011)
Prosodic	**0.31** (± 0.029)	0.19 (± 0.022)	0.82 (± 0.054)
Mimo-gestural	**0.31** (± 0.035)	0.19 (± 0.025)	0.92 (± 0.043)
Baseline	**0.17** (± 0.061)	0.17 (± 0.061)	0.17 (± 0.068)

Table 15
F-score, Precision and **Recall** for **Given/New** feedback prediction for the final model (with all types of features) and per type of feature: **Auto-regressive, Morpho-syntactic, Prosodic**. The ground truth used is the type of the observed feedback. The **Baseline** predicts randomly given/new based on the distribution observed in the data.

Features modality	F-score	Precision	Recall
All	**0.62** (± 0.037)	0.63 (± 0.039)	0.61 (± 0.054)
Auto-regressive	**0.55** (± 0.041)	0.66 (± 0.040)	0.47 (± 0.048)
Morpho-syntactic	**0.59** (± 0.034)	0.63 (± 0.040)	0.56 (± 0.043)
Prosodic	**0.64** (± 0.034)	0.57 (± 0.037)	0.74 (± 0.064)
Baseline	**0.45** (± 0.035)	0.45 (± 0.040)	0.46 (± 0.045)

often associated with a large span in case of the occurrence of an absolute point (B or T). The *SMS* trigram corresponds to a flat intonation in a medium pitch. However, the first tone is of the type *Same* (in other words, at the same level of the tone preceding the trigram). As a consequence, the intonation variation cannot be determined and could be falling or rising depending on the value of the previous tone. The interpretation of this last trigram remains therefore difficult to establish.

In terms of the other prosodic features, we found that *silent pauses lasting between 400 and* 600 ms and *600*–1200 ms favor feedback production. This is consistent with different results from the literature: (Cathcart et al., 2003) for a duration of at least 400 ms and (Ward and Tsukahara, 2000) for a duration between 600 ms and 1200 ms. Finally, *punctuation* (see Section 4.2.2 for the definition of syntactic punctuation in speech transcription) and the *number of tokens* produced by the speaker are also relevant in predicting feedback.

As for morpho-syntax, 36 n-grams of POS were selected for feedback prediction. Fifteen n-grams correspond to the end of a chunk, corresponding to a short non-hierarchical syntactic unit (Abney, 1991).

Table A.1
The encoding of the French unitary POS (Parts-of-Speech).

POS	Type	Example	POS	Type	Example
Af	Qualitative adjective	petit	Rq	Other adverb of negation	que
A-	Other adjective	aucune	R-	Other adverb	facilement
Cc	Coordinating conjunction	et	Sa	Preposition "à"	à
Cs	Subordinating conjunction	lorsque	Sd	Preposition "de"	de
I	Interjection	hein	SP	Preposition + Determiner	du
D-	Determiner	le	S-	Other preposition	dans
Nc	Common noun	chapeau	Van-	Auxiliary *avoir* - infinitive	avoir
Np	Proper noun	Baudelaire	Vapp	Auxiliary *avoir* - present participle	ayant
Nk	Cardinal noun	huit	Vaps	Auxiliary *avoir* - past participle	eu
Ppn	Personal pronoun – nominative	je	Va–	Other form of auxiliary *avoir*	ai
Ppj	Personal pronoun - accusative	la	Ven-	Auxiliary *être* - infinitive	être
Ppd	Personal pronoun - dative	lui	Vepp	Auxiliary *être* - present participle	étant
Ppo	Personal pronoun - oblique	moi	Veps	Auxiliary *être* - past participle	été
Pr	Relative pronoun	qui	Ve–	Other form of auxiliary *être*	suis
Pd	Demonstrative pronoun	ce	Vmn-	Main verb - infinitive	aimer
Pi	Indefinite pronoun	personne	Vmpp	Main verb - present participle	finissant
Pt	Interrogative pronoun	quel	Vmps	Main verb - past participle	payé
Px	Reflexive pronoun	se	Vm–	Other form of main verb	partirent
Ps	Possessive pronoun	le mien	U	Unknown or foreign word	zarbi
Pk	Cardinal pronoun	quinze	Wd	Strong punctuation	.
Rn	Particle of negation	ne	Wm	Weak punctuation	,
Rd	Adverb of negation	pas	#	Pause	

Table A.2
The encoding of the MOMEL anchor points of the MOMEL-INTSINT (Hirst, 2022).

Absolute Value	Encoding	Relative Value	Encoding
Top	T	Higher	H
Bottom	B	Lower	L
Mid	M	Same	S
		Upstepped	U
		Downstepped	D

Table B
Examples of bigrams and trigrams that can indicate the end of a chunk extracted from PACO—Cheese!. and their translation into English. Bigrams and trigrams are highlighted in bold when there is an equivalence between French and English. N-grams of POS are indicated by a dash ("-"). Abbreviations are detailed in Tables A.1 and A.2 (see Appendix A).

N-grams	Examples	English Translation
Vm-R	*Je me suis dit* ***c'est trop***	*I thought it* ***was too much***
Rd-Nc	*J'en ai* ***pas besoin***	*I don't need it*
Vm-I	*On* ***verra hein***	*We'll* ***see huh***
Af-I	*J'étais* ***présente hein***	*I was* ***there huh***
Sd-Nc	*J'ai peur qu'on ait beaucoup* ***de boulot***	*I'm afraid we've got a lot* ***of work***
R-Af	*Les calanques de Marseille c'est* ***vraiment beau***	*Marseille's calanques are* ***really beautiful***
D-Nc	*Pour apprendre* ***la langue***	*To learn* ***the language***
R-I	*Il l'avait enfermé dedans tout* ***simplement hein***	*He simply (R) locked it inside huh (I)*
Ppn-Ppj-Vm	*C'est pareil parce que* ***je la voyais***	*It's the same because I (Ppn) saw (Vm) her (Ppj)*
Vm-S-Nc	*Ouais je voulais pas* ***aller en master***	*Yeah I didn't want to* ***go into a master's*** *[program]*

Table B (see Appendix B) shows examples extracted from our corpus. Three of these n-grams end with an interjection. In French, interjections including phatic markers are often used to involve the listener at the end of a sentence in order to accentuate the previously-given information. Moreover, punctuation that is also a significant feature can mark either the end of a phrase or the end of a sentence. These results confirm that feedback tends to occur at the end of a sentence. However, we also found significant features that suggested the occurrence of feedback in the middle of a sentence. Our results can be explained by the fact that we considered all different types of multimodal feedback as well as not only generic feedback, as it is the case in most studies, but also specific feedback. These results are also consistent with our observation in 5.2, showing that a large quantity of feedback (76 %) is produced during the speaker's speech production.

Fifteen n-grams contain an *adverb* or a *qualifying adjective* (including 7 which also correspond to the end of the chunk). These categories play the role of information modifiers. They complete existing information, which has generally just been given, and provide new information. Feedback here can be an explicit mark of grounding.

As for the n-grams that are not explained by the two previous points, 4 contain a *common noun* (N) and 2 contain an *interjection*. The literature has shown that common nouns encourage feedback production. Interjections could again correspond to phatic or discourse markers and could also punctuate the previous proposition. Few studies have looked at feedback produced in overlap; part of our results is explained by the fact that we took into account both feedback produced during speech and non-speech. In this case, the POS we studied are those corresponding to the end of syntactic units. Our model learned to predict feedback both in overlap and non-overlap contexts. Relevant n-grams can thus occur in the middle or at the end of a syntactic unit, but only in specific contexts, for example, closing a sequence, or bringing new information to the listener.

Individual features are crucial to help verify the linguistic

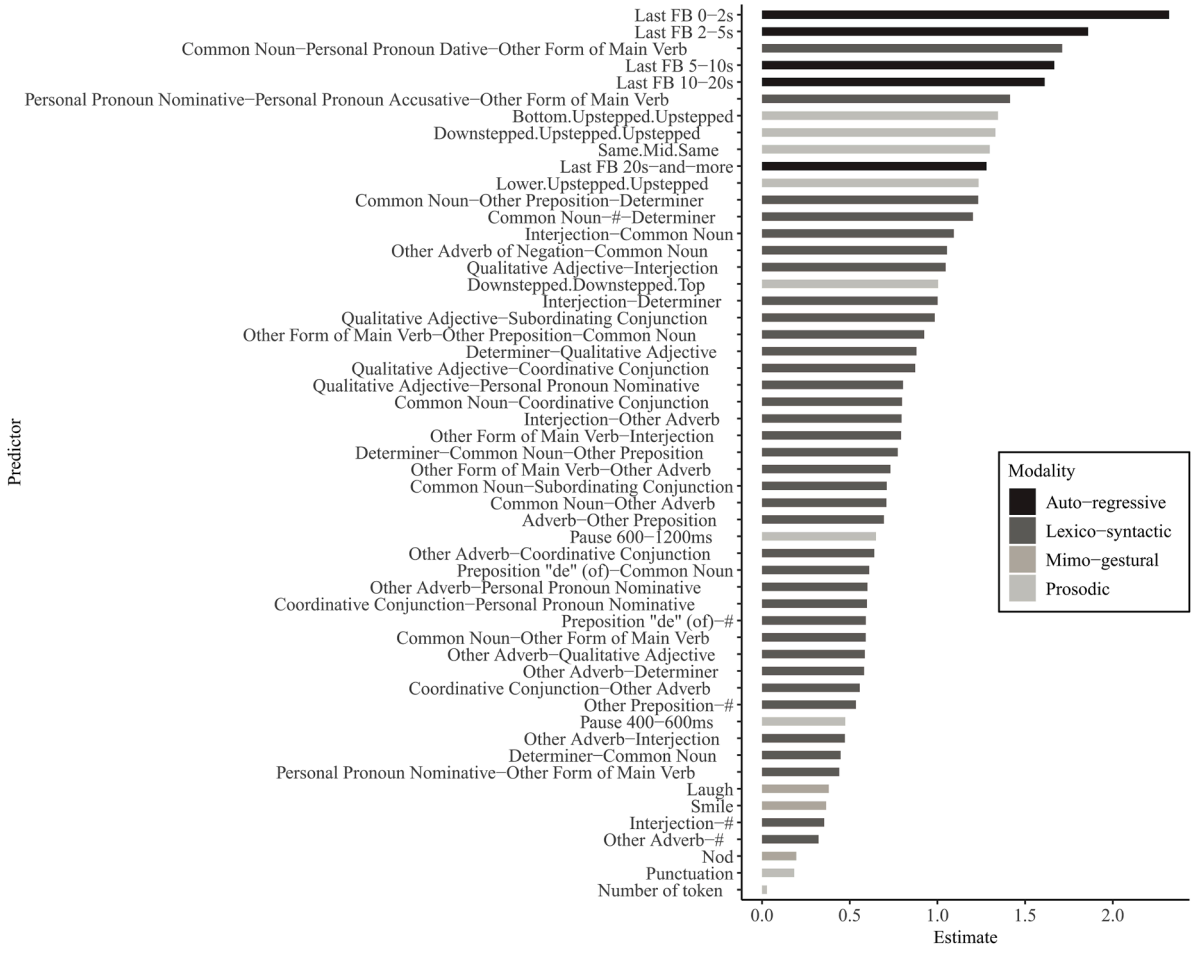

Fig. B. Feature importance for **feedback** prediction. N-grams of POS are indicated by a dash ("-"). N-grams of tones are indicated by a period (".."). Abbreviations are detailed in Tables A.1 and A.2 (see Appendix A). The **Estimate** corresponds to the estimated coefficients from the logistic regression model, showing the contribution of each feature to the prediction of the dependent variable.

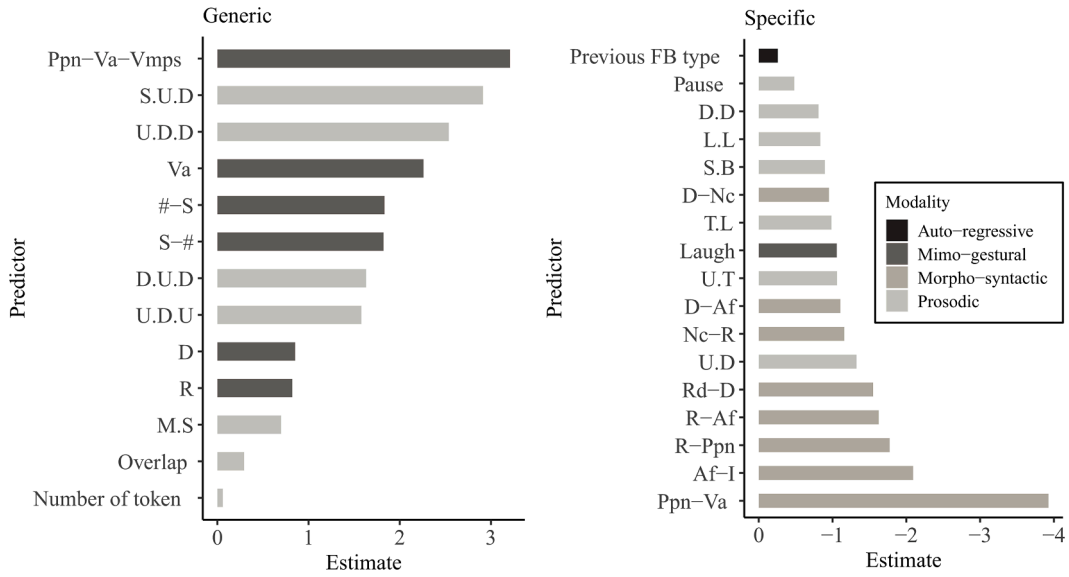

Fig. C. Feature importance for the **Generic** (left panel) and **Specific** (right panel) feedback. N-grams of POS are indicated by a dash ("-"). N-grams of tones are indicated by a period (".."). Abbreviations are detailed in Tables A.1 and A.2 (see Appendix A). The **Estimate** corresponds to the estimated coefficients from the logistic regression model, showing the contribution of each feature to the prediction of the dependent variable.

consistency of our model. However, one feature alone is not informative enough to trigger feedback, whereas a combination of features may have a strong impact on the probability of its production. The Logit model makes it possible to compute for every such combination of features the probability of obtaining feedback at a time *t*. This probability, expressed in *pmean*, can be higher or lower than the probability of obtaining feedback when no features are present in the preceding window. In our dataset, the *pmean* when no features are present is 0.025, as presented in Table 12. The top part this table shows examples of combinations that increase this probability, while the lower part illustrates those decreasing the probability. As explained above, the context is highly important, but taking into account only auto-regressive features is not sufficient. For instance, if the last item of feedback is produced within the previous 2 s (Examples 1 and 9), the associated features could completely change the *pmean*. In the configuration of the combination *Last feedback 0–2 s, Smile, Punctuation, Determinant-Common Noun*, the probability that feedback occurs is almost doubled even if other feedback has just been produced. In contrast, *Last feedback 0–2 s, Adverb-Interjection, ML, Punctuation* is associated with a low probability of producing feedback.

7. Second level of prediction: predicting feedback types

After a first step aiming at predicting positions where feedback may potentially occur, we aim at predicting the type of the feedback in a second step. Recalling that the type concerns the generic/specific distinction proposed in (Bavelas et al., 2000) and the two sub-types for specific feedback (positive/negative, given /new), we hypothesized that the different types/sub-types could be triggered by different inviting cues. We proposed two models for feedback types and sub-types.

7.1. Data sampling and model building

The dataset for feedback type classification is a table containing, for each observed feedback occurrence: the dependent variable (binary encoding, 1 for a feedback type and 0 for its opposite type) and the predictors, in binary type (e.g., presence of a given POS n-gram, presence of a pause, …) or continuous type (e.g., number of tokens).

Most of the predictors are the same as those presented in the previous section. Unlike the prediction of feedback position, nods, laughs, and smiles are extracted in a window of 2 s before the onset of the feedback. Positive, negative, and concrete tokens were counted since the last feedback occurrence. The previous feedback type was implemented (with the following encoding: 0 for generic, 1 for positive-given, 2 for positive-new, 3 for negative-given, and 4 for negative-new) and was the only auto-regressive feature considered here. Auto-regressive features that used the time elapsed since the last feedback occurrence were not kept in this second level of classification, since we considered these features important for the position of the feedback. Overlap was implemented with binary encoding, where 0 indicated that the main speaker was not speaking, and 1 indicated that the feedback was produced in total overlap. We also tested different thresholds of extraction for n-grams of POS and of tones (10, 15, 20, 50, 100). Finally, as in 6.3, we tested the clustering of n-gram of tones and of POS (a total of 164 features for clustered features and 273 for detailed features).

The entire dataset was used for the generic/specific classification. For positive/negative and given /new classification, only feedback of specific types were retained.

The question of imbalanced classes is less problematic for this classification than for feedback prediction: the proportion of generic/specific feedback was roughly equivalent (1206 generic and 1171 specific). However, the sub-types of specific feedback were slightly imbalanced, especially for polarity: 964 positive and 207 negative feedback occurrences. The distribution of given/new sub-types was more balanced: 640 new and 531 given feedback occurrences. In the next section, we used the following encoding for the binary dependent variables: generic/1, specific/0; positive/1, negative/0; new/1, given/0. In this second level of prediction, the type of the feedback was predicted by using the following classifier equation to convert the *Logit* probability into a binary response:

$$\text{if}(p > p_{\text{threshold}})\{\text{response} = 1\}\text{else}\{\text{response} = 0\} \quad (2)$$

where $p_{\text{threshold}}$ was chosen in such a way that the proportion of feedback of type 1 corresponds to the one observed. This threshold value was estimated based on the training corpus. The model thus provides us with a binary prediction which depends on the predictor values in input.

For each model, a cross-validation was obtained by running a Monte Carlo cross-validation (on 50 trials with a 80 %−20 % ratio for the training versus the evaluation sample). A baseline was computed for each level of classification that randomly predicted class 0 or 1 following

the observed distribution.

7.2. Results & discussion

Results for the 3 models predicting the type of feedback showed that all the best performances were obtained with the features that used detailed n-gram of POS and n-gram of tones. We thus present only these results.

Generic vs. Specific: The best occurrence threshold of n-grams for generic/specific classification is 10. In the case of generic/specific classification, the model returns 30 significant features. Thirteen features favor the *generic* class using auto-regressive, morpho-syntactic, and prosodic modalities. 17 features were selected for the *specific* class from all modalities. Performances are presented in Table 13 and feature importance in Fig. C (see Appendix C). The resulting f-score was **0.62**. The *t*-test confirmed that our model is more efficient than the baseline (f-score = 0.48), and always better with all features types ($p \leq 0.01$).

Several interesting observations can be made from the comparison of the features selected in the generic/specific classification. Regarding n-grams of POS, 5 n-grams out of 6 selected for **generic** prediction correspond to a middle position of a chunk (*#-S; S-#; Va; R; D*). Generic feedback is said to often correspond to a listening function allowing the main speaker to hold the floor. Feedback can therefore be produced in overlap or in the middle of a chunk. In the same way, the feature overlap favors the prediction of generic feedback.

On the contrary, 5 n-grams out of 8 for the prediction of **specific** feedback correspond to the end of a chunk or utterance (*D-Nc; d-Af; Nc-R; R-Af; Af-I*). The n-gram *Ppn-Va* corresponds to a clitic followed by the auxiliary "*avoir*" (have), and precedes a past participle also corresponding to new information. Finally, 6 of these n-grams contain an *adverb* or a *qualifying adjective* (*Nc-R; d-Af; Af-I, Rd-D, R-Af, R-Ppn*) which play the role of modifying information. As shown in (Bavelas et al., 2017), specific feedback tends to occur after new information. This observation is reinforced by the *pause* features that also favor specific feedback.

At the prosodic level, results show that n-grams of tones selected for predicting **generic** feedback are more often produced in a medium or a small pitch range, or with a flat intonation (*SUD, UDU, MS, UDD, DUD*). Overall, we observed small variations. No points appeared in a very high or very low absolute peak (*B* or *T*). Conversely, for the prediction of **specific** feedback, the tones selected showed changes of larger amplitude with tones located in a high pitch and one in a low pitch (*LL, TL, UT, SB*). Finally, two tones showed a falling intonation (*LL, DD*).

Nodding and smiling did not seem to play a role in predicting the type of feedback. Only *laughs* were selected by the model. As expected, the presence of laughter favors the production of **specific** feedback.

In terms of auto-regressive features, the *type of the previous feedback occurrence* was also selected to predict **specific** feedback. The interpretation of this observation remains difficult since we are using a discrete encoding. A possibility is that we added these features in accordance with observations made from our data. Specific feedback can be produced immediately after generic feedback when the semantic content allows for it. In the same way, many items of specific feedback are produced several times depending on the importance of the information given and can be produced with the same type as the previous one.

Positive vs. Negative: The best occurrence threshold of n-grams for positive/negative classification was 20. For the second level of classification, positive/negative classes, 14 features were selected by the model. Seven features from morpho-syntactic, prosodic and mimo-gestural modalities were selected in favor of positive class. Eight features were selected for the negative category, from each auto-regressive, morpho-syntactic and prosodic modality.

Performances are presented in Table 14. The f-score of **0.36** is very encouraging and outperformed the baseline (f-score = 0.17). For this task as well, taking modalities separately led to a considerable drop in f-score (in particular for auto-regressive features). Despite the low scores for the prediction of polarity, features selected by the model tended to support our hypothesis. Indeed, *laughs* and *positive tokens* favored **positive** feedback and *negative tokens* seemed significant in predicting **negative** feedback. We also note that *pauses (400–*600 ms*)* favor **positive** feedback where *overlap* favors **negative** feedback. One possible explanation is that negative feedback occurs less in speech but shows an immediate reaction. Low scores for this task can be partially explained by the highly unbalanced dataset which contained almost 5 times more positive feedback than negative. Future studies with wider datasets including different types of interactions (e.g., confrontation, debate) are needed in order to decide whether this type of classification for specific feedback is relevant for this task.

Given vs. New: The best occurrence threshold of n-grams for new/given sub-types classification was 20. Thirty features were considered as significant by the model without any mimo-gestural features, with 8 for the type *new* from morpho-syntactic and prosodic modalities and 6 features for the type *given* from auto-regressive, prosodic and morpho-syntactic modalities. Performances in Table 15 showed a higher f-score with only prosodic features. A *t*-test confirmed that this difference was significant (p-value ≤ 0.001). Here, our hypothesis concerning multi-modality was not confirmed: prosodic information alone provided the best f-score. In all cases, prediction always outperformed the baseline (f-score = 0.45).

By considering only the prosodic features selected by the model with respect to the results, we observed that the n-gram of tones selected for the **new** type showed a rising and a high pitch span (*HT*). Conversely, the selected tone for the **given** type showed a small amplitude with a small falling (DUD). These results are in line with the findings of (Gravano and Hirschberg, 2006). They observed that *new* nouns are typically produced with a higher pitch and after a longer pause compared to *given* nouns. When examining the bigrams of selected POS by the model for the new type, we also identified *common noun-pause* and *adverb-common noun* bigrams.

In this section, we first classify feedback as generic or specific and subsequently, we further classify specific feedback according to two essential features of the discourse, namely the stance of the main speaker and the information status. Moreover, feedback that is not in line with the speaker's stance and/or the information status may require a redirection of the conversation or a repair. Furthermore, in the context of human-machine interactions, it is essential for a virtual agent to not only provide feedback at the appropriate time but also adapt it to the context of the conversation, to prevent user disengagement.

The presented results are encouraging and outperform our baseline models. However, it is important to acknowledge the need for additional investigations to classify sub-types of feedback with the availability of larger and more balanced datasets.

8. Conclusion

This study is the first to propose a method for feedback prediction and classification of both generic and specific feedback from vocal/visual and gestural modalities. We introduce a new taxonomy for specific feedback, including speaker stance and information structure characteristics, opening the way to a precise prediction of all feedback sub-types. We identify different features, from different modalities (prosody, morpho-syntax, lexicon, acoustics, gestures) as predictive cues. We use *Logistic Regression (Logit)*, a machine learning technique which is particularly appropriate for feature selection and for unbalanced datasets (i.e., *feedback* vs. *no-feedback*). Moreover, the Logit method offers the possibility to evaluate feature contributions, which can be interpreted in the light of linguistic hypotheses.

This study confirmed our main hypothesis on the interest of involving high-level and interpretable multimodal features in the perspective of building a classification method distinguishing *feedback/no-feedback, generic/specific, positive/negative*, and *given/new* types. All models outperformed the different baselines. The importance of

multimodal features was confirmed (for 3 out of 4 of the models). Moreover, our results show that all of the models performed better when more detailed features were used (clustered features vs. detailed features) confirming that high-level linguistic features are more relevant in predicting feedback.

The main results of the first level of prediction demonstrated that feedback is triggered by different features: rising and large-span intonation for the prosodic level, end of chunks, interjections, adverbs, qualifying adjectives, and common nouns for the morpho-syntactic level, and all mimo-gestural features (nodding, laughter and high intensity smiling). We also showed the importance of auto-regressive features, in interaction with other features.

Our results for the second level of prediction showed that predictive features vary depending on the type of feedback (generic/specific, positive/negative, given/new):

- POS n-grams indicating a middle of a chunk and light intonational variation favor a generic type.
- POS n-grams indicating the end of a chunk, a modifier (adverb, qualifying adjective), large intonational variation and laughter favor a specific type.
- Positive tokens, laughter and pauses favor a positive sub-type.
- Negative tokens and overlap favor a negative sub-type.
- Pronoun-Verb bigrams, interjections and small intonation falling favor the given sub-type.
- Common nouns and rising with a high pitch pan favor the new sub-type.

Furthermore, in terms of feedback description, we have shown that different sets of feedback components emerge between the *generic, positive-new, positive-given, negative-new* and *negative-given* types. The analyses show that generic feedback reactions are clearly distinguished from specific reactions by their greater use of nodding and interjections. In the case of specific feedback, laughter and smiling are favored for positive types, while negative feedback is characterized by eyebrow movements and feedback produced with more verbalizations than positive feedback. Concerning the information status, frequent expressions were found for the given type and for the new type. Eyebrow movements and laughter were also more characteristic of the new type than the given type. The next step towards a complete predictive feedback model would be to predict the feedback form based on an even more detailed description, including the different attitudes that can be realized in each sub-category (e.g., surprise), as illustrated in the last level of Fig. 2.

In terms of the efficiency of our model, let us recall that we are predicting the *possible* positions, which means that in our reference corpus, not all these positions are filled by feedback. We assume that during natural interactions, interlocutors balance their feedback production in order to provide optimal support to the main speaker in a sufficient but non-excessive manner. Feedback production is thus highly variable in the sense that the position but also the type of feedback to be produced may highly differ. This variability in the timing and form renders feedback evaluation difficult (Morency et al., 2010).

Our results are also difficult to compare given that data, features and evaluation differ from one study to another. Corpora vary in terms of modalities (face-to-face conversations, phone calls, videoconferencing, etc.), tasks (free conversation, task-oriented dialogue, narration, etc.), relationships between participants, or language typology. Most studies consider only a subset of feedback (only generic and/or only verbal feedback), implicitly reducing the number of feedback instances and their variability.

Our objective evaluations thus need to be completed with perceptual experiments, such as those proposed in (Fujie et al., 2004; Kitaoka et al., 2006; Huang et al., 2010a, Huang et al., 2010b, de Kok et al., 2014). In particular, we observed considerable improvement by using a margin of error of $\pm$ 1000 ms (f-score = 0.49). We will verify the validity of this window by looking for the accuracy of the feedback using a perceptual experiment presenting the participant with feedback produced within or outside the evaluated window.

Several perspectives for improvement can be put forward. First, we are currently working on integrating new features and feedback components into the model, including gaze direction (Bavelas et al., 2002; Ozkan and Morency, 2010; Hjalmarsson and Oertel, 2012), blinking (Hômke et al., 2017), eyebrow movements, and facial expressions. We also plan to work on different feature aggregation solutions. A second line of improvement involves the use of embedding and associated learning techniques, moving towards neural networks and the integration of semantics. Next, we will also integrate the question of the modality of the feedback (vocal, visual and bimodal) by studying whether it plays a role in the model (in other words, whether some features can be associated with different feedback modalities) and more generally whether it is possible to predict feedback modality together with its type. (Truong et al., 2011) showed that different inviting cues reside between vocal and visual modalities. In this study, our focus was the site and type of feedback, but we could also include the question of modality in our future work. Finally, it is crucial to encompass feedback prediction in conjunction with speech turns, as proposed by (Kitaoka et al., 2006; Skantze, 2017; Ishii et al., 2021), to achieve a comprehensive exploration of the dynamics of interactive communication.

At a more general level, it is interesting first of all to consider factors such as gender, age and relationship between participants. Therefore, our results serve as a solid basis for studying and comparing feedback production according to different social factors. Second, feedback-inviting features are similar across languages, a hypothesis confirmed by the literature on feedback production in different languages. The different prosodic and syntactic structures across languages lead to different feature combinations for predicting feedback. To the best of our knowledge, this study is the first to propose a continuous prediction of feedback in French, thereby contributing to our understanding of the differences in feedback production across languages and cultures. In the same perspective, future research on inappropriate, disagreement, and misalignment feedback would be welcome, although these types of feedback appear infrequently in interactions.

CRediT authorship contribution statement

Auriane Boudin: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Roxane Bertrand:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Stéphane Rauzy:** Writing – review & editing, Writing – original draft, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Conceptualization. **Magalie Ochs:** Writing – original draft, Supervision, Methodology, Funding acquisition, Conceptualization. **Philippe Blache:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

This work, carried out within the Institut Convergence ILCB (ANR-16-CONV-0002), benefited from the support of the French government,

managed by the French National Agency for Research (ANR).

Appendix A. POS and Tone abbreviations

Table A.1, Table A.2

Appendix B. First Level of Prediction: Predicting Feedback Position

Table B, Fig. B

Appendix C. Second Level of Prediction: Predicting Feedback Type

Fig. C

References

Abney, S.P., 1991. Parsing By Chunks. In: Berwick, R.C., Abney, S.P., Tenny, C. (Eds.), Principle-Based Parsing. Studies in Linguistics and Philosophy, 44. Springer, Dordrecht, pp. 257–278.

Allwood, J., Cerrato, L., 2003. A study of gestural feedback expressions. In: First nordic symposium on multimodal communication. Copenhagen, pp. 7–22..

Allwood, J., Nivre, J., Ahlsén, E., 1992. On the semantics and pragmatics of linguistic feedback. J. Semant. 9 (1), 1–26.

Amoyal, M., Priego-Valverde, B., 2019. Smiling for negotiating topic transitions in French conversation. GESPIN-Gesture and Speech in Interaction. Paderborn, Germany.

Amoyal, M., Priego-Valverde, B., Rauzy, S., 2020. In: Paco: a corpus to analyze the impact of common ground in spontaneous face-to-face interaction. European Language Resources Association, Marseille, France, pp. 628–633.

Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.-P., 2018. Openface 2.0: facial behavior analysis toolkit. In: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). IEEE, pp. 59–66.

Bavelas, J.B., Coates, L., Johnson, T, 2000. Listeners as co-narrators. J. Pers. Soc. Psychol. 79 (6), 941.

Bavelas, J.B., Coates, L., Johnson, T, 2002. Listener responses as a collaborative process: the role of gaze. J. Commun. 52 (3), 566–580.

Bavelas, J., Gerwing, J., Healing, S., 2017. Doing mutual understanding. calibrating with micro-sequences in face-to-face dialogue. J. Pragmat. 121, 91–112.

Bertrand, R., Espesser, R., 2017. Co-narration in french conversation storytelling: a quantitative insight. J. Pragmat. 111, 33–53.

Bertrand, R., Priego-Valverde, B., 2017. Listing practice in French conversation: from collaborative achievement to interactional convergence. Discours. Revue de linguistique, Psycholinguistique Et informatique. A journal of linguistics, Psycholinguistics and Computational Linguistics (20).

Bertrand, R., Ferré, G., Blache, P., Espesser, R., Rauzy, S., 2007. Backchannels revisited from a multimodal perspective. Auditory-visual Speech Processing. Hilvarenbeek, pp. 1–5.

Bigi, B., 2012. Sppas: a tool for the phonetic segmentations of speech. In: In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey, pp. 1748–1755.

Bigi, B., 2015. Sppas-multi-lingual approaches to the automatic annotation of speech. *The Phonetician*. J. Int. Soc. Phonet. Sci. 111, 54–69. ISSN: 07416164.

Bodur, K., Nikolaus, M., Kassim, F., Prévot, L., Fourtassi, A., 2021. Chico: a multimodal corpus for the study of child conversation. In: Companion Publication of the 2021 International Conference on Multimodal Interaction (ICMI '21 Companion). Association for Computing Machinery, New York, NY, USA, pp. 158–163.

Bonin, P., Méot, A., Bugaiska, A., 2018. Concreteness norms for 1,659 french words: relationships with other psycholinguistic variables and word recognition times. Behav. Res. Method. 50 (6), 2366–2387.

Boudin, A., Bertrand, R., Rauzy, S., Ochs, M., Blache, P., 2021. In: Ekštein, K., Pártl, F., Konopík, M. (Eds.), Text, Speech, and Dialogue, TSD 2021. Lecture Notes in Computer Science, 12848. Springer, Cham, pp. 537–549.

Brusco, P., Vidal, J., Beňuš, S., Gravano, A., 2020. A cross-linguistic analysis of the temporal dynamics of turn-taking cues using machine learning as a descriptive tool. Speech. Commun. 125, 24–40.

Bunt, H., 1994. Context and dialogue control. Think Q. 3 (1), 19–31.

Bunt, H., 2012. The semantics of feedback. In: Brown-Schmidt, S., Ginzburg, J., Larsson, S. (Eds.), Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL 2012). Paris, France, pp. 118–127.

Cathcart, N., Carletta, J., Klein, E., 2003. A shallow model of backchannel continuers in spoken dialogue. Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics. In: (EACL '03), 1. Association for Computational Linguistics, USA, pp. 51–58.

Clark, H.H., 1996. Using Language. Cambridge university press.

De Kok, I., Ozkan, D., Heylen, D., Morency, L.-P., 2010. Learning and evaluating response prediction models using parallel listener consensus. In: International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction, pp. 1–8 pages.

de Kok, I., Poppe, R., Heylen, D, 2014. Iterative perceptual learning for social behavior synthesis. J. Multimod. User Interface. 8 (3), 231–241.

Ferré, G., Renaudier, S., 2017. Unimodal and bimodal backchannels in conversational english. Proceedings of the 21st Workshop on the Semantics and Pragmatics of Dialogue - Full Papers. SEMDIAL, Saarbrucken, Germany, pp. 27–37.

Figueroa, C., Adigwe, A., Ochs, M., Skantze, G., 2022. In: Annotation of Communicative Functions of Short Feedback Tokens in Switchboard. European Language Resources Association (ELRA, Marseille, France, pp. 1849–1859.

Fujie, S., Fukushima, K., Kobayashi, T., 2004. A conversation robot with back-channel feedback function based on linguistic and nonlinguistic information. In: Proc. ICARA Int. Conference on Autonomous Robots and Agents, pp. 379–384.

Gandolfi, G., Pickering, M.J., Garrod, S., 2023. Mechanisms of alignment: shared control, social cognition and metacognition. Philosoph. Transact. Roy. Soc. B 378 (1870), 20210362.

Garrod, S., Pickering, M.J., 2004. Why is conversation so easy? Trend. Cogn. Sci. (Regul. Ed.) 8 (1), 8–11.

Glas, N., Pelachaud, C., 2015. Definitions of engagement in human-agent interaction. In: International Conference on Affective Computing and Intelligent Interaction (ACII). Xi'an, China, pp. 944–949.

Gravano, A., Hirschberg, J., 2006. Effect of genre, speaker, and word class on the realization of given and new information. In: *Ninth International Conference on Spoken Language Processing*. Pittsburgh, USA.

Gravano, A., Hirschberg, J., 2011. Turn-taking cues in taskoriented dialogue. Comput. Speech. Lang. 25 (3), 601–634.

Hömke, P., Holler, J., Levinson, S.C., 2017. Eye blinking as addressee feedback in face-to-face conversation. Res. Lang. Soc. Interact. 50 (1), 54–70.

Hastie, T., Tibshirani, R., Friedman, J.H., Friedman, J.H., 2009. The Elements of Statistical learning: Data mining, inference, and Prediction, 2. Springer, New York, pp. 1–758.

Hirst, D.J., 2007. A praat plugin for momel and intsint with improved algorithms for modelling and coding intonation. In: Proceedings of the XVIth International Conference of Phonetic Sciences. Saarbrücken, pp. 1233–1236.

Hirst, D., 2022. A multi-level, multilingual approach to the annotation of speech prosody. Jonathan Barnes & Stefanie Shattuck-Hufnagel. Prosodic Theory and Practice. MIT Press, pp. 117–149 pages.

Hjalmarsson, A., Oertel, C., 2012. Gaze direction as a backchannel inviting cue in dialogue. Proc. of the IVA 2012 Workshop on Realtime Conversational Virtual Agents (RCVA 2012). Santa Cruz, CA, USA.

Horton, W.S., 2017. Theories and approaches to the study of conversation and interactive discourse. In: Schober, M.F., Rapp, D.N., Britt, M.A. (Eds.), The Routledge Handbook of Discourse Processes, 2nd ed. Routledge Press, pp. 22–68.

Huang, L., Morency, L.-P., Gratch, J., 2010. Learning backchannel prediction model from parasocial consensus sampling: a subjective evaluation. In: Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., Safonova, A. (Eds.), Intelligent Virtual Agents, Lecture Notes in Computer Science, 6356. Springer, Berlin, Heidelberg. IVA 2010.

Ishii, R., Ren, X., Muszynski, M., Morency, L.-P., 2021. Multimodal and multitask approach to listener's backchannel prediction: can prediction of turn-changing and turnmanagement willingness improve backchannel modeling?. In: Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents (IVA '21). Association for Computing Machinery, New York, NY, USA, pp. 131–138.

Jang, J.Y., Kim, S., Jung, M., Shin, S., Gweon, G, 2021. Bpm mt: enhanced backchannel prediction model using multi-task learning. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3447–3452.

Jefferson, G., 1978. Sequential aspects of storytelling in conversation. Studies in the Organization of Conversational Interaction. Academic Press, pp. 219–248.

Kawahara, T., Yamaguchi, T., Inoue, K., Takanashi, K., Ward, N.G., 2016. Prediction and generation of backchannel form for attentive listening systems. Interspeech, pp. 2890–2894 pages.

Kitaoka, N., Takeuchi, M., Nishimura, R., Nakagawa, S., 2006. Response timing detection using prosodic and linguistic information for human-friendly spoken dialog systems. Inform. Med. Technol. 1 (1), 296–304.

Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., Den, Y., 1998. An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs. Lang. Speech. 41 (3–4), 295–321.

Liu, J., Nikolaus, M., Bodur, K., Fourtassi, A., 2022. Predicting backchannel signaling in child-caregiver multimodal conversations. In: Companion Publication of the 2022 International Conference on Multimodal Interaction (ICMI '22 Companion). Association for Computing Machinery, New York, NY, USA, pp. 196–200.

Meena, R., Skantze, G., Gustafson, J., 2014. Data-driven models for timing feedback responses in a map task dialogue system. Comput. Speech. Lang. 28 (4), 903–922.

Morency, L.-P., de Kok, I., Gratch, J., 2010. A probabilistic multimodal approach for predicting listener backchannels. Auton. Agent. Multi. Agent. Syst. 20 (1), 70–84.

Mueller, M., Leuschner, D., Briem, L., Schmidt, M., Kilgour, K., Stueker, S., Waibel, A., 2015. Using neural networks for data-driven backchannel prediction: a survey on input features and training techniques. In: Kurosu, M. (Ed.), Human-Computer Interaction: Interaction Technologies. Springer International Publishing, Cham, pp. 329–340.

Okato, Y., Kato, K., Kamamoto, M., Itahashi, S., 1996. Insertion of interjectory response based on prosodic information. In: Proceedings of IVTTA '96. Workshop on Interactive Voice Technology for Telecommunications Applications. Basking Ridge, NJ, USA, pp. 85–88.

Ortega, D., Li, C.-Y., Vu, N.T., 2020. Oh, jeez! or uh-huh? a listener-aware backchannel predictor on asr transcriptions. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain, pp. 8064–8068.

Ozkan, D., Morency, L.-P., 2010. Concensus of self-features for nonverbal behavior analysis. In: Salah, A.A., Gevers, T., Sebe, N., Vinciarelli, A. (Eds.), Human Behavior Understanding, Lecture Notes in Computer Science, 6219. Springer, Berlin, Heidelberg, pp. 75–86. HBU 2010.
Ozkan, D., Morency, L.-P., 2012. Latent mixture of discriminative experts. IEEE Trans. Multimed. 15 (2), 326–338.
Pickering, M.J., Garrod, S., 2013. An integrated theory of language production and comprehension. Behav. Brain Sci. 36 (4), 329–347.
Pickering, M., Garrod, S., 2021. Understanding dialogue: Language use and social interaction. Cambridge University Press.
Poppe, R., Truong, K.P., Reidsma, D., Heylen, D., 2010. Backchannel strategies for artificial listeners. In: Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., Safonova, A. (Eds.), Intelligent Virtual Agents, Lecture Notes in Computer Science, 6356. Springer, Berlin, Heidelberg, pp. 146–158. IVA 2010.
Poppe, R., Truong, K.P., Heylen, D., 2011. Backchannels: quantity, type and timing matters. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (Eds.), Intelligent Virtual Agents, Lecture Notes in Computer Science, 6895. Springer, Berlin, Heidelberg, pp. 228–239. IVA 2011.
Prévot, L., Gorisch, J., 2014. Crossing empirical and formal approaches for studying french feedback items. In: *Proceedings of 11th Conference on Logical Engineering and Natural Language Semantics (LENLS)*. Tokyo, Japan.
Prévot, L., Gorisch, J., Bertrand, R., 2016. A cup of cofee: a large collection of feedback utterances provided with communicative function annotations. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 3180–3185.
Priego-Valverde, B., Bigi, B., Amoyal, M, 2020. In: cheese!": a corpus of face-to-face french interactions. a case study for analyzing smiling and conversational humor. European Language Resources Association (ELRA, Marseille, France, pp. 467–475.
Rauzy, S., Amoyal, M., 2020. Smad: a tool for automatically annotating the smile intensity along a video record. In: HRC2020, 10th Humour Research Conference, Commerce. Texas, United States.
Rauzy, S., Goujon, A., 2018. Automatic annotation of facial actions from a video record: the case of eyebrows raising and frowning. In: Workshop on" Affects, Compagnons Artificiels et Interactions", WACAI 2018 pages 7–pages.
Rauzy, S., Montcheuil, G., Blache, P., 2014. Marsatag, a tagger for french written texts and speech transcriptions. In: Second Asian Pacific Corpus linguistics Conference. Hong Kong, China, pp. 220–220.
Ruede, R., Müller, M., Stüker, S., Waibel, A., 2019. Yeah, right, uh-huh: a deep learning backchannel predictor. Advanced Social Interaction With Agents. Springer, pp. 247–258 pages.
Ruusuvuori, J., Peräkylä, A., 2009. Facial and verbal expressions in assessing stories and topics. Res. Lang. Soc. Interact. 42 (4), 377–394.
Sacks, H., Schegloff, E.A., and Jefferson, G. (1974). A simplest systematics for the organization of turn taking for conversation. In *Language 50*, pages 696–735.
Schegloff, E.A., 1982. Discourse as an interactional achievement: some uses of 'uh huh'and other things that come between sentences. Analyz. Discour.: Text Talk 71, 71–93.
Skantze, G., 2017. In: Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks. Association for Computational Linguistics, Saarbrücken, Germany, pp. 220–230.
Skantze, G., 2021. Turn-taking in conversational systems and human-robot interaction: a review. Comput. Speech. Lang. 67, 101178.
Sloetjes, H., Wittenburg, P., 2008. In: Annotation by category-elan and iso dcr. European Language Resources Association (ELRA, Marrakech, Morocco.
Stivers, T., 2008. Stance, alignment, and affiliation during storytelling: when nodding is a token of affiliation. Res. Lang. Soc. Interact. 41 (1), 31–57.
Terrell, A., Mutlu, B., 2012. A regression-based approach to modeling addressee backchannels. In: *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Seoul, South Korea, pp. 280–289.
Tolins, J., Fox Tree, J.E., 2014. Addressee backchannels steer narrative development. J. Pragmat. 70, 152–164.
Truong, K.P., Poppe, R., Heylen, D, 2010. A rule-based backchannel prediction model using pitch and pause information. In: Proceedings of Interspeech 2010. International Speech Communication Association (ISCA), pp. 3058–3061.
Truong, K.P., Poppe, R., de Kok, I., Heylen, D., 2011. A multimodal analysis of vocal and visual backchannels in spontaneous dialogs. In: Proceedings of Interspeech 2011. International Speech Communication Association (ISCA), pp. 2973–2976.
Ward, N., Tsukahara, W., 2000. Prosodic features which cue backchannel responses in english and japanese. J. Pragmat. 32 (8), 1177–1207.
Ward, N., 1996. Using prosodic clues to decide when to produce back-channel utterances, ICSLP'96, volume 3,. IEEE, pp. 1728–1731.
Wildfeuer, J., Pflaeging, J., Bateman, J., Seizov, O., Tseng, C., 2020. Multimodality: Disciplinary Thoughts and the Challenge of Diversity. De Gruyter, Berlin, Boston.
Yngve, V.H., 1970. On getting a word in edgewise. In: Campbell, M.A. (Ed.), Papers from the Sixth Regional Meeting. Chicago Linguistics Society. Department of Linguistics, University of Chicago, Chicago, pp. 567–578.